LLM-Adapted Colombian Spanish Lexicography: Proficiency Control, Hallucination, and Cultural Distortion

Johnatan E. Bonilla
Humboldt-Universität zu Berlin - Instituto Caro y Cuervo
j.bonilla@hu-berlin.de

Abstract

We evaluate whether open-source LLMs can produce proficiency-graded English adaptations of entries from the *Diccionario de colombianismos* (DiCol), a Colombian Spanish lexicographic resource used in language teaching. Three 7–8B instruction-tuned models—Llama 3.1, Qwen2.5, and Mistral—generate Beginner, Intermediate, and Advanced translations for all 8,252 definitions using structured zero-shot prompts identical across levels except for the target CEFR band. Automated metrics show that Intermediate targeting collapses (73–83% classified as Advanced by vocabulary, $\chi^2 > 705$, $p < .001$) and that Advanced outputs expand 4.9–8.2× relative to the source. Expert annotation of a 360-entry stratified sample ($\kappa = 0.61$–0.68) identifies hallucination in 19% of entries (Fleiss' $\kappa = 0.77$ for cultural preservation categories, 97% unanimity among flagged cases). Hallucination concentrates in the Advanced condition (81%, $\chi^2 = 86.6$, $p < .001$) and is associated with higher expansion ($U = 16{,}662$, $p < .001$, $r = 0.68$), manifesting primarily as generic elaboration and, in a smaller proportion, as Colombia-stereotyping and pragmatic polarity inversion. We discuss these findings through the lens of Venuti (1995)'s domestication framework and describe the observed pattern as *algorithmic domestication*.

1 Introduction

The *Diccionario de colombianismos* (DiCol) (Instituto Caro y Cuervo, 2018), published by Colombia's Instituto Caro y Cuervo, is a reference resource for Colombian Spanish language teaching. It documents over 5,900 headwords, which together with their subentries (collocations, fixed phrases, secondary senses) yield 8,252 individual definitions. Its recent availability through the open-source LEXICC platform (Monsalve Muñoz et al., 2025)[1] raised the question of whether LLMs could help produce English-language adaptations graded by proficiency level—making the resource accessible to a broader audience of learners.

This paper reports the results of that effort. We prompted three open-source 7–8B models to translate every DiCol definition at three CEFR levels (Beginner, Intermediate, Advanced) and evaluated the outputs with automated metrics and structured expert annotation. We address three research questions:

RQ1 (Proficiency control) To what extent do zero-shot 7–8B LLMs produce outputs whose vocabulary and perceived complexity match the target CEFR level?

RQ2 (Hallucination) At what rate and under which conditions does elaboration pressure induce hallucination, and how does it distribute across models and CEFR targets?

RQ3 (Cultural distortion) When hallucination occurs, does the fabricated content exhibit directional bias toward dominant-culture representations of Colombia?

The results reveal systematic failures in proficiency control and, more importantly, a recurring pattern of hallucination under the Advanced condition. When models are asked to produce complex output from short source definitions, they expand substantially and fill the additional space with fabricated content. *Guacal* (a wooden crate for transporting goods, 15 source words) is rendered as "a receptacle [...] colloquially referred to in Colombian vernacular as a 'carriola' "—a term the model invents. *Malamañoso* ("ladrón"—thief, 1 word) becomes a 47-word narrative about urban crime "reflecting crimes common in Colombian cities."

These fabrications are not random. They draw on dominant-culture representations of Colombia—violence, exoticism, romanticized

[1] https://lexicc.caroycuervo.gov.co

Proceedings of the 4th Workshop on Cross-Cultural Considerations in NLP (C3NLP 2026), pages 67–75

tradition—reproducing stereotypes rather than recovering source-culture meaning. Venuti (1995) describes *domestication* as the reduction of cultural foreignness through fluent translation. Qadri et al. (2025) identify omission and simplification as mechanisms of cultural erasure in LLMs. Our data suggest a related pattern: under elaboration pressure, models replace source-culture content with fabricated content reflecting dominant-culture representations—a pattern we describe as *algorithmic domestication*. Our findings are limited to zero-shot prompting of 7–8B models; whether they persist under iterative refinement or larger architectures remains open. Data and annotation are publicly available.[2]

2 Related Work

Translation and domestication. Venuti (1995) argues that Anglo-American translation practice is dominated by fluent domestication: reducing foreign texts to target-culture norms. Lefevere (1992) frames all translation as *rewriting* shaped by institutional constraints. Newmark (1988) provides the taxonomy we draw on: neutralization, functional equivalence, borrowing, and omission.

Cultural bias in LLMs. Qadri et al. (2025) define two mechanisms of LLM cultural erasure: omission and simplification. Agarwal et al. (2025) show that AI writing suggestions lead non-Western users to adopt Western styles. Navneet et al. (2025) quantify cultural marker erosion in LLM writing assistance across World English varieties. Hershcovich et al. (2022) identify dominant-culture training data bias as a fundamental cross-cultural NLP challenge. When cultural content is absent from training data, models do not simply leave gaps—they fill them, and the question is *with what*.

Hallucination. Hallucination—generating fluent but unfaithful content—is well documented in summarization (Maynez et al., 2020) and across NLG tasks (Ji et al., 2023). Ji et al. (2023) categorize hallucination as intrinsic (contradicting the source) or extrinsic (adding unverifiable content). Most hallucination research focuses on factual accuracy; the intersection of hallucination and cultural representation has received less attention. Our study documents a domain where extrinsic hallucination produces culturally marked distortion rather than content-agnostic error, connecting the hallucination literature to the cultural bias literature above.

CEFR-based text adaptation. Barayan et al. (2025) find systematic under-simplification at lower readability targets and over-elaboration at higher ones. Alva-Manchego et al. (2025) report that reliable readability control requires multi-iterative processes. Farajidizaji et al. (2024) confirm that without iterative feedback no LLM reliably hits target readability. We evaluate the zero-shot baseline—the default output before any refinement—as a necessary first step.

The DiCol. The DiCol (Instituto Caro y Cuervo, 2018) documents Colombian Spanish with strong sociolinguistic specificity. It is accessible through LEXICC (Monsalve Muñoz et al., 2025), an open-source dictionary writing system. Its online availability postdates all models' training data cutoffs, ruling out contamination.

3 Methodology

3.1 Data and Models

We use all 8,252 Spanish definitions from the DiCol (mean: 12.2 words, SD = 7.8, range 1–49). Three open-source instruction-tuned models in the 7–8B parameter range were deployed via Ollama on an NVIDIA T4 16 GB: **Llama 3.1 8B Instruct** (Meta; ≈15T training tokens), **Qwen2.5 7B Instruct** (Alibaba; >18T tokens), and **Mistral 7B v0.3** (Mistral AI; ≈7.3T tokens).

Each entry was translated at three proficiency levels—Beginner (A1–A2), Intermediate (B1–B2), Advanced (C1–C2)—using a single structured zero-shot prompt template (Appendix A). The template was instantiated separately for each CEFR band; the only difference across conditions was the target proficiency descriptor. The same instructions—including the cultural preservation clause ("preserve the regional connotations and sociolinguistic context")—appear in all three prompts. Generation parameters: temperature 0.7, top-p 0.9, max_tokens 300. Google Translate API v3 served as a non-adapted baseline: it establishes the natural complexity distribution of direct translation without any CEFR instructions, and is not a competing system for the proficiency adaptation task. Its inclusion in the vocabulary table (Table 1) allows readers to assess how the untranslated baseline distributes across vocabulary levels.

[2] https://github.com/johnatanebonilla/dicol_translation

3.2 Automated Metrics

We use two automated metrics as complementary proxies, each with known limitations stated upfront.

Vocabulary level (Oxford 5000). Content words are mapped to CEFR macro-levels using the Oxford 5000 (Oxford University Press, 2020): Beginner (A1–A2), Intermediate (B1–B2), or Advanced (C1+). **Words absent from the list are counted as Advanced**, creating a structural upward bias that is particularly acute for a regional variety lexicon: culturally specific terms (fauna, flora, local foods, social practices) frequently lack English equivalents in the Oxford list and are therefore automatically classified as Advanced regardless of the translation's actual lexical choices. This bias means that the Intermediate collapse rates reported in Section 4 represent an upper bound on the true effect; some portion of the 73–83% Advanced classification may reflect domain-specific OOV inflation rather than genuine complexity failure. We report Oxford 5000 distributions as a directional proxy while foregrounding expansion ratio—which does not depend on vocabulary list coverage—as the primary measure of elaboration pressure. The text-level label is assigned by majority vote; stopwords are excluded.

Expansion ratio and readability. The **expansion ratio** (translation words / source words) is our primary measure of elaboration pressure. Flesch–Kincaid Grade Level (Kincaid et al., 1975) serves as a secondary surface complexity proxy. FK-GL was derived for US Navy training manuals at a 35% cloze-test threshold (Wang et al., 2013)—not for language proficiency assessment. Shah et al. (2022) show that FK-GL and CEFR produce inconsistent results, and Tanprasert and Kauchak (2021) demonstrate susceptibility to surface manipulation. FK-GL and word count correlate at $r = 0.49$–0.71, so a substantial share of FK scores reflects length inflation. We report FK-GL to detect overcompensation (output exceeding the approximate C2 ceiling of FK ≈ 14) while foregrounding expansion ratio for interpretation.

3.3 Human Evaluation

Sampling. A 360-entry balanced sample was drawn using a 3×3 stratified design (3 models $\times$ 3 CEFR levels, 40 entries per cell), without replacement, using a fixed random seed (`random_state=42`).

Raters. Three professional translators served as raters, all native speakers of Colombian Spanish with 5–8 years of English-to-Spanish and Spanish-to-English translation experience and familiarity with the CEFR framework. None had lexicography training or prior exposure to the dataset or research hypotheses.

Protocol. Raters reviewed entries independently and *blind* to the designed level—they saw only the English translation alongside the Spanish source. Each rater assigned: (1) a CEFR level (Beginner / Intermediate / Advanced), or *Mistranslation* or *Language Error*; and (2) a cultural preservation category: *Neutral/Fully Preserved*, *Hallucination* (content fabricated, absent from source), *Added Context* (extra content, not incorrect), or *Minor Loss of Nuance*.

4 Results

4.1 Vocabulary and the Intermediate Collapse

Table 1 presents the Oxford 5000 distributions. Models partially control Beginner vocabulary (22–39%) and reliably produce Advanced vocabulary (95–97%), but the Intermediate condition collapses across all three families: only 9–10% of entries reach the target, with 73–83% classified as Advanced instead ($\chi^2 > 705$, $p < .001$, Cramér's $V = 0.21$–0.30).

At Beginner, Llama achieves 39.2% target accuracy, Qwen 30.1%, Mistral 22.1%—meaningful variation suggesting that training data composition affects low-complexity control. But at Intermediate, all three converge on near-identical failure: 9.0%, 8.6%, and 10.3%, respectively.

The collapse is visible at entry level. *Parcero* ("amigo, compañero"—friend) yields a clean Beginner from Llama: "A friend or companion. Someone you know well" (10 words, Beginner vocabulary). The Intermediate output expands to 37 words with "mutual understanding and trust that transcends mere acquaintance," classified as Advanced. *Bacano* ("simpática o agradable"—nice) shows the same pattern: 10-word Beginner vs. 56-word Intermediate with "warmth and friendliness in social interactions." *Guaro* ("trago de aguardiente"—a shot of aguardiente, 3 source words) yields 13 words at Beginner but 33 at Intermediate, adding "vallenato music" and "social gatherings"—none of which appears in the source. The models do not exhibit

continuous control over complexity but rather a binary regime: low-complexity simplification vs. high-complexity elaboration, with no stable intermediate zone.

Table 1: Oxford 5000 vocabulary classification (%, $N = 8{,}252$). Bold = target-level accuracy. Unk = unlisted words (negligible).

Model	Cond.	Beg	Int	Adv	Unk
GT baseline	—	28.1	12.1	59.5	0.3
Llama 3.1 8B	Beg	**39.2**	11.3	49.4	0.0
	Int	10.7	**9.0**	80.3	0.0
	Adv	0.9	1.7	**97.4**	0.0
Qwen2.5 7B	Beg	**30.1**	13.3	56.2	0.3
	Int	7.8	**8.6**	83.0	0.5
	Adv	1.3	2.4	**95.7**	0.5
Mistral 7B	Beg	**22.1**	9.6	68.2	0.0
	Int	16.3	**10.3**	73.4	0.0
	Adv	2.4	2.7	**94.9**	0.0

4.2 Elaboration Pressure

Table 2 reports expansion ratio, word count, and FK-GL. Under the Advanced condition, all three models inflate output: Llama produces 2.9× more words than at Beginner, Qwen 2.5×, Mistral 1.9× (Kruskal–Wallis $H > 5{,}880$, $p < .001$ for all models). Llama expands source content 8.2× on average at Advanced; Qwen 4.9×; Mistral 5.2×.

All models exceed the approximate C2 ceiling (FK ≈ 14) at Advanced, though up to 57% of FK-GL variance is explained by word count alone (R^2: Qwen 0.57, Llama 0.24, Mistral 0.16), confirming that much of the apparent complexity is length inflation.

Overcompensation follows two strategies. Llama generates through *verbosity*: 46.6 words per entry at Advanced. *Llave* ("amigo, compañero"—friend, 2 source words) becomes 49 words at Intermediate, including "trustworthy and familiar [...] partner in social activities." Qwen generates through *lexical elevation*: shorter entries (26.4 words) with low-frequency substitutions. *Reversazo* ("retorno a una decisión"—reversal, 7 words) becomes "laden with the nuanced implications of reconsideration [...] characteristic of Colombia's dynamic social fabric" (27 words).

4.3 Expert Annotation

Table 3 reports pairwise Cohen's κ for the five-category CEFR scheme (3 levels + Mistranslation + Language Error). Agreement is substantial ($\kappa = 0.61$–0.68, 82–85% raw agreement).

Table 2: Word count, expansion ratio, and FK Grade by model and condition. GT baseline: 11.2 words, FK 11.2. † = FK exceeds C2 ceiling (≈ 14; see Kincaid et al. 1975).

Model	Cond.	Words		Exp	FK Grade	
		M	SD		M	SD
Llama	Beg	16.2	8.9	2.5×	9.8	4.0
	Int	35.4	19.9	6.8×	17.6†	5.2
	Adv	46.6	17.1	8.2×	27.0†	5.7
Qwen	Beg	10.4	5.5	1.1×	10.6	4.9
	Int	14.9	8.9	2.4×	14.5	5.3
	Adv	26.4	13.5	4.9×	21.7†	6.5
Mistral	Beg	16.3	11.2	1.8×	13.5	5.3
	Int	14.2	10.1	2.2×	13.8	5.3
	Adv	31.6	17.9	5.2×	20.4†	5.5

Table 3: Inter-rater agreement ($N = 360$; 5 CEFR categories).

Pair	Agree%	κ
R1–R2	85.3	0.675
R1–R3	82.5	0.609
R2–R3	85.0	0.651

Mistranslation and language error. Raters identified 12 entries as Mistranslation (≥2 agree; 11 unanimous) and 9 as Language Error (≥2; all unanimous). Mistranslations concentrate in Advanced (9/12) and span several failure modes: semantic contamination (*guagua*, "nursing infant" → "Colombian tenderloin"—a culinary term with no referent in any Spanish variety), register distortion (*salchipapa*, street food described as "premium rapid cuisine"), factual invention (*ovejo*, "male sheep" → "dominant or aggressive individual"), and domain confusion (*queso siete cueros*, a layered pulled-curd cheese → "a roll made by rolling out dough"). At Beginner, the same terms are typically translated correctly (*guagua* → "baby"; *ovejo* → "male sheep"). The errors emerge under elaboration pressure, not from inability to access the basic meaning.

Language Errors are exclusively Qwen (9/9) and consist of mid-sentence Chinese-character contamination—a multilingual training artifact. *Llantero* ("tire repairman") is rendered as "Individual who specializes in the 修复与维护汽车轮胎…" This is a data-leakage issue unrelated to the cultural patterns discussed below.

Blind ratings vs. designed level. Table 4 cross-tabulates the majority-vote blind label against the designed condition. At Intermediate, raters recover

the intended level in 87.5% of cases ($\chi^2 = 67.5$, $p < .001$). At Beginner, 73.3% are perceived as Intermediate ($\chi^2 = 32.0$, $p < .001$). At Advanced, only 45.8% are recognized; this rate does not differ from chance ($p = .36$). Figure 1 shows the convergence: all three raters independently assign Intermediate as the dominant label (63–69%), despite the balanced design ($N = 120$ per condition).

Table 4: Majority-vote CEFR label vs. designed condition (%, $N = 120$ per condition). MT = Mistranslation, LE = Language Error.

	Majority-vote label				
Designed	**Beg**	**Int**	**Adv**	**MT**	**LE**
Beginner	**24.2**	73.3	0.0	2.5	0.0
Intermediate	5.0	**87.5**	4.2	0.0	3.3
Advanced	0.0	42.5	**45.8**	7.5	4.2

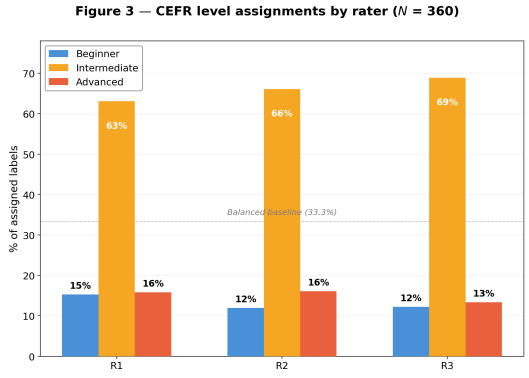

Figure 1: CEFR distributions by rater ($N = 360$; $N = 120$ per designed condition). Intermediate dominates regardless of the designed level.

4.4 Cultural Preservation and Hallucination

Table 5 presents cultural preservation by model and CEFR condition. Of 360 entries, 68 (18.9%) were flagged as Hallucination by $\geq$2 raters, 67 unanimously (98.5%). Inter-rater agreement on the four categories is substantial (Fleiss' $\kappa = 0.77$). Hallucination concentrates in the Advanced condition (55/68, 80.9%; $\chi^2 = 86.6$, $p < .001$, Cramér's $V = 0.49$) and distributes across models without a dominant contributor (Llama 21, Mistral 22, Qwen 25). Hallucinated entries are more expanded: median 4.2× vs. 1.1× for non-hallucinated (Mann–Whitney $U = 16{,}662$, $p < .001$, rank-biserial $r = 0.68$). This association does not establish directionality: expansion may facilitate hallucination, but hallucination may also drive expansion, or both may arise from the interaction between the complexity target and instruction-following under resource constraints.

Table 5: Cultural preservation by model and CEFR condition ($N = 360$; majority-vote label, $\geq$2 raters; Fleiss' $\kappa = 0.77$). NP = Neutral/Preserved; HAL = Hallucination; AC = Added Context; MN = Minor Loss of Nuance. †One entry in Qwen Beginner received a majority-vote label of *Significant Erasure*, outside the four-category scheme; it is excluded from the cell count and reported in the footnote.

Model	**Cond.**	**NP**	**HAL**	**AC**	**MN**
Llama 3.1	Beg	34	0	4	2
	Int	34	3	3	0
	Adv	12	18	10	0
Qwen 2.5	Beg	31	1	3	4†
	Int	33	6	1	0
	Adv	18	18	4	0
Mistral 7B	Beg	26	2	11	1
	Int	37	1	0	2
	Adv	17	19	1	3
Total		**242**	**68**	**37**	**12**

4.5 Qualitative Analysis of Hallucination

To understand how hallucination manifests, we examined all 68 flagged entries. Generic elaboration is the most frequent pattern ($N = 48$), within which 27 involve cultural filler; Colombia-stereotyping accounts for 10 entries, and polarity inversion for 4. The classification below was performed by the authors based on source–target comparison; it is descriptive and no inter-rater reliability was computed.

Generic elaboration. The most frequent pattern: models expand short definitions with verbose but culturally neutral paraphrasing. *Carajito* ("persona inmadura y descortés"—immature and rude, 4 words) becomes 62 words under Mistral Advanced, including "a lack of maturity and refinement in one's behavior [...] disregard for social norms and decorum." *Nerdo* (a student with good grades who may be mocked, 22 source words) acquires "the societal tension between scholarly aptitude and peer acceptance"—a sociological commentary absent from the dictionary. *Raimundo y todo el mundo* ("todo tipo de personas"—all kinds of people, 4 words) expands to 32 words: "individuals from all walks of life, encompassing a diverse array of socio-economic backgrounds."

Cultural filler. A subset of elaboration adds culturally vacuous phrases tied to Colombia. *Trompo*

de poner ("someone blamed for others' errors," 15 source words) becomes 40 words ending with "a perpetual object of criticism in Colombian culture." *Ojo de pescado* (a photography term for fisheye lens, purely technical) becomes "typical in some Colombian photographic traditions"—the model attributes cultural significance to a universal technical term. *Patijunto* (knock-kneed, a physical description) acquires "a specific posture typically associated with certain traditional dances." Phrases like "vibrant cultural richness," "deeply rooted in Colombian cultural ethos," and "dynamic social fabric" recur across entries and models as templated padding.

Colombia-stereotyping. In a smaller number of cases, the filler draws on specific stereotypes. *Chamba* ("deep wound") acquires "Colombia's turbulent history and its indelible mark on the human spirit." *¡Epa!* ("warning") is linked to "security-conscious cultural ethos [...] where heightened vigilance is practically ingrained." *Caminar* ("pursuing someone romantically," 9 source words) becomes "a suitor's dogged determination, often observed in traditional romantic courtship practices." The elaborations reproduce the dominant image of Colombia available in English-language training data: violence, exoticism, and romanticized tradition.

Pragmatic polarity inversion. In four cases, the model inverts the pragmatic force. *Mamerto* (derogatory: "naive communist sympathizer") is rewritten across all three models at Advanced as laudatory ("commitment to egalitarian principles"). *Ahuevarse* ("to become cowardly") becomes "to muster up courage." This pattern is consistent with RLHF alignment rewarding positive framing, though we cannot isolate this effect without comparing base vs. instruction-tuned variants—an experiment we did not conduct.

5 Discussion

5.1 Expansion and Hallucination

The concentration of hallucination in Advanced (81%) co-occurs with the highest expansion ratios. Source definitions average 12.2 words; at 8× expansion, a model must generate ∼100 words from a 12-word source. The statistical association is large ($r = 0.68$). At Beginner, where expansion is minimal (1.1–2.5×), hallucination occurs in only 3 of 120 entries (2.5%).

This relationship is correlational. We did not independently manipulate output length (e.g., via fixed token budgets), and alternative explanations—including prompt-specific effects or model-level instruction-following limitations—cannot be ruled out. The uniformity across three architecturally distinct model families provides some evidence against purely model-specific explanations, but a controlled experiment varying token budgets would be needed to establish directionality.

The role of the prompt. The cultural preservation instruction is identical across all three CEFR levels (Appendix A). Because this instruction is constant, any effect it induces should be uniformly distributed across conditions. The strong skew toward Advanced (81%) suggests that hallucination emerges from the interaction between the preservation instruction and the high-complexity generation target, rather than from either factor alone. The prompt triggers the expansion; the training data determines what fills it.

5.2 Hallucination and Domestication

Venuti (1995) describes domestication as the reduction of cultural foreignness through fluent translation. Qadri et al. (2025) distinguish omission from simplification. Our data suggest a related pattern. The model does not merely omit or simplify Colombian cultural content—it replaces it with fabricated content reflecting dominant-culture representations. *Guacal* is not omitted but renamed with an invented term ("carriola") and framed as "an architectural testament to Colombian rural craftsmanship." *Mamerto* does not lose its political charge through simplification; it is rewritten as praise.

The directionality of fabrication is notable. Unlike generic hallucination, which tends to be content-agnostic (Ji et al., 2023), the fabricated content here draws systematically on the dominant-culture image of Colombia available in English-language training data—violence, exoticism, romanticized rural life. A fisheye lens becomes "typical in Colombian photographic traditions"; a knock-kneed person is linked to "traditional dances"; a deep wound evokes "Colombia's turbulent history." The fabrications are not only false but *directional*: they consistently reproduce the same narrow set of cultural associations. This directional bias is what connects the pattern to Venuti's domestication framework and distinguishes it from standard extrinsic hallucination. We use the term *algorithmic*

domestication to describe this observation at a phenomenological level: under elaboration pressure, fabricated content draws on dominant-culture representations rather than recovering source-culture meaning, and the directionality is consistent across architecturally distinct model families. We are careful not to claim that this constitutes a novel mechanism. It may be a domain-specific manifestation of known extrinsic hallucination dynamics (Ji et al., 2023), amplified by the mismatch between the cultural specificity of the source and the English-centric training distribution. What the data establish is the *pattern*: directional, culturally marked fabrication concentrated at high elaboration targets. Whether the theoretical apparatus of Venuti's domestication framework maps onto an underlying generative mechanism—or whether "algorithmic domestication" names a heuristically useful analogy—remains an open empirical question requiring controlled experiments that independently vary output length, cultural source density, and training data composition.

5.3 The Beginner Collapse

Beginner-prompted entries are perceived as Intermediate in 73.3% of cases. *Tinto* ("hot coffee without milk," 8 source words) already produces 8 words at Beginner—adequate—but *llave* ("friend," 2 source words) generates 30 words, including "someone with whom you do things together. A person you like and consider as a good friend." This is verbose for A1–A2 but not recognizably Advanced; raters classify it as Intermediate. The only condition these 7–8B models reliably produce—as perceived by expert readers—is Intermediate. The B-level under-specification identified by Barayan et al. (2025) operates in both directions: models cannot target B1–B2 from above (Intermediate collapses into Advanced vocabulary) or produce output distinguishable from B1–B2 at A1–A2 (Beginner collapses into Intermediate perception).

6 Conclusion

We evaluated LLM-driven CEFR-level adaptation of the DiCol across three 7–8B model families under zero-shot conditions. The Intermediate condition collapses into Advanced vocabulary ($\chi^2 > 705$, $p < .001$), while Beginner is perceived as Intermediate by expert raters in 73.3% of cases. Advanced outputs expand 4.9–8.2$\times$ relative to the source. Expert annotators ($\kappa = 0.61$–0.68; Fleiss' $\kappa = 0.77$ for cultural categories) identify hallucination in 19% of entries (97% unanimity), concentrated in Advanced ($\chi^2 = 86.6$, $p < .001$) and associated with higher expansion ($r = 0.68$). Hallucinated content manifests primarily as generic elaboration and cultural filler, with a smaller proportion of Colombia-stereotyping and pragmatic polarity inversion.

We interpret these findings through Venuti (1995)'s domestication framework: under elaboration pressure, fabricated content draws on dominant-culture representations rather than recovering source-culture meaning—a pattern we describe as *algorithmic domestication*. These findings suggest that deploying LLMs for lexicographic or pedagogical adaptation without external grounding risks systematic cultural distortion, particularly at higher proficiency targets. For practitioners considering LLM-assisted adaptation of regional variety resources, our results indicate that Beginner-level outputs may be usable with editorial review, but Advanced outputs require substantial post-editing or alternative approaches. A natural next step is to evaluate retrieval-augmented generation using the LEXICC database as an external knowledge base, grounding model output in verified lexicographic content. Future work should also test whether few-shot prompting with DiCol examples reduces hallucination, and whether these patterns persist under larger models, iterative prompting, or controlled token budgets.

Limitations

Model and compute scope. Only 7–8B open-source models were evaluated, deployed on an NVIDIA T4 16 GB. Compute constraints prevented evaluation of larger models (70B+) that might exhibit different patterns. Iterative prompting (chain-of-thought, multi-turn refinement) may mitigate hallucination; our study evaluates the default zero-shot behavior as a necessary baseline. We did not evaluate base (non-instruction-tuned) variants, so the contribution of RLHF to polarity inversion remains speculative.

Metrics. The Oxford 5000 counts unlisted words as Advanced, creating upward bias. It is unclear how much of the Intermediate collapse is driven by genuine vocabulary complexity versus domain-specific OOV terms that are automatically classified as Advanced. A regional variety lexicon contains many culturally specific terms (fauna, flora,

foods, social practices) that lack English equivalents in the Oxford list; their automatic classification as Advanced inflates the collapse rate independently of the model's actual vocabulary choices. FK-GL conflates length with complexity and was not designed for proficiency assessment; we mitigate this by foregrounding expansion ratio. Future work could evaluate alternative vocabulary lists (e.g., the English Vocabulary Profile) or semantic similarity metrics between source and target to complement the current proxies.

Annotation. The "Minor Loss of Nuance" category shows low inter-rater stability (5 of 66 entries with all-3 agreement). The qualitative hallucination typology (Section 4.5) was author-assigned without independent reliability verification. Raters had translation but not lexicography experience. Annotation costs limited the human evaluation to a 360-entry sample; scaling to additional raters or the full corpus was not feasible within the resources of this study.

Generalizability. Our findings are specific to Colombian Spanish lexicography under zero-shot 7–8B conditions. Whether they generalize to other varieties, languages, text types, or model scales requires further investigation. Future work should test whether RAG with the LEXICC database itself, or few-shot prompting with DiCol examples, reduces both collapse and hallucination rates.

Ethics Statement

This study uses a publicly available lexicographic resource (DiCol). Raters were compensated at professional translation rates and gave informed consent. The stereotypical content generated by models reproduces harmful representations of Colombia; we report it for analytical purposes. The annotated data is released under CC-BY-4.0 with an explicit warning about stereotypical content.

References

Dhruv Agarwal, Mor Naaman, and Aditya Vashistha. 2025. AI suggestions homogenize writing toward Western styles and diminish cultural nuances. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, Yokohama, Japan. ACM.

Fernando Alva-Manchego, Regina Stodden, Joseph Marvin Imperial, Abdullah Barayan, Kai North, and Harish Tayyar Madabushi. 2025. Findings of the TSAR 2025 shared task on readability-controlled text simplification. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility and Readability (TSAR 2025)*, pages 116–130, Suzhou, China. Association for Computational Linguistics.

Abdullah Barayan, Jose Camacho-Collados, and Fernando Alva-Manchego. 2025. Analysing zero-shot readability-controlled sentence simplification. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6762–6781, Abu Dhabi, UAE. Association for Computational Linguistics.

Asma Farajidizaji, Vatsal Raina, and Mark Gales. 2024. Is it possible to modify text to a target readability level? an initial investigation using zero-shot large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9325–9339, Torino, Italia. ELRA and ICCL.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Instituto Caro y Cuervo. 2018. *Diccionario de colombianismos*. Instituto Caro y Cuervo, Bogotá, Colombia.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

J. Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas for Navy Enlisted Personnel. Technical Report RBR 8-75, Naval Air Station Memphis.

André Lefevere. 1992. *Translation, Rewriting, and the Manipulation of Literary Fame*. Routledge, London and New York.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Udiluz del Carmen Monsalve Muñoz, Johnatan E. Bonilla, Ruth Yanira Rubio López, and Andrés Esteban Luna Cortés. 2025. LEXICC: The design and development of an online dictionary writing system. *Lexikos*, 35(1):83–108.

Satyam Kumar Navneet, Joydeep Chandra, and Yong Zhang. 2025. When AI writes, whose voice remains? quantifying cultural marker erasure across world English varieties in large language models. *Preprint*, arXiv:2602.22145.

Peter Newmark. 1988. *A Textbook of Translation*. Prentice Hall International, New York.

Oxford University Press. 2020. The Oxford 5000: The most important words to learn in English. https://www.oxfordlearnersdictionaries.com/wordlists/oxford3000-5000. Word list with CEFR level annotations A1–C1.

Rida Qadri, Aida M. Davani, Kevin Robinson, and Vinodkumar Prabhakaran. 2025. Risks of cultural erasure in large language models. *Preprint*, arXiv:2501.01056.

Muhammad Hammad Hussain Shah, Shahid Nawaz, and Shazia Bukhari. 2022. Lexical comparison between the common European framework of reference for languages and the Flesch-Kincaid. *Pakistan Journal of Society, Education and Language*, 8(2):471–482.

Teerapaun Tanprasert and David Kauchak. 2021. Flesch-kincaid is not a text simplification evaluation metric. In *Proceedings of the First Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 1–14, Online. Association for Computational Linguistics.

Lawrence Venuti. 1995. *The Translator's Invisibility: A History of Translation*. Routledge, London and New York. Second edition 2008.

Lih-Wern Wang, Michael J. Miller, Michael R. Schmitt, and Frances K. Wen. 2013. Assessing readability formula differences with written health information materials: Application, results, and recommendations. *Research in Social and Administrative Pharmacy*, 9(5):503–516.

A System Prompt

A single prompt template was used for all three models and all three conditions. The template was instantiated separately for each CEFR band; the *only* element that varies is the target proficiency descriptor (shown in **bold**). The placeholder `{definition}` was replaced with each DiCol entry's Spanish definition. Temperature: 0.7, top-p: 0.9, max_tokens: 300.

```
You are a specialist translator for
language learners.  Translate the
following Colombian Spanish dictionary
definition into English strictly
targeting the CEFR [A1-A2 | B1-B2 |
C1-C2] proficiency level.

Rules:
- Lexicon & Grammar: Adapt all
vocabulary choices, sentence lengths,
and grammatical complexity to align
exclusively with the requested CEFR
level.
- Cultural Nuance: Preserve the regional
connotations and sociolinguistic
context of the Colombian term.

Output ONLY the translated definition
text.  No labels, no preamble, no
commentary.

Definition: {definition}
```