
Probing Latent Colombian Identity Inferences in Qwen2.5-7B with Natural Language Autoencoders

Pablo Santiago Potes Velasco
Universidad Autónoma de Occidente
Cali, Colombia
pablo.potes@uao.edu.co

María del Mar García Matabanchoy
Universidad Autónoma de Occidente
Cali, Colombia
maria_d.garcia_m@uao.edu.co

Óscar Julián Pérez Ladino
Universidad Autónoma de Occidente
Cali, Colombia
oscar.perez@uao.edu.co

Jhoan Stevan Mosquera Ortiz
Universidad Autónoma de Occidente
Cali, Colombia
jhoan.mosquera@uao.edu.co

Nicolás Lozano Mazuera
Universidad Autónoma de Occidente
Cali, Colombia
nicolas.lozano_m@uao.edu.co

Gilber Alexis Corrales Gallego
Universidad Autónoma de Occidente
GobLab-UAI
Cali, Colombia
gacorrales@uao.edu.co

Abstract

Large language models may infer demographic attributes from subtle linguistic cues even when those attributes are not explicitly stated. This pilot study examines whether Qwen2.5-7B-Instruct internally represents Colombian identity, socioeconomic status, or stereotype-related information when processing Colombian-Spanish and English prompts. We use Natural Language Autoencoders (NLA) to verbalize residual-stream activations from layer 20 across four positional quartiles per prompt. Our dataset contains 30 prompts arranged as 15 matched Spanish-English pairs, spanning explicit Colombian cues, implicit Colombian cues, and neutral controls. We report descriptive rates and qualitative evidence rather than statistically powered effects, focusing on whether latent nationality or stereotype representations appear before they are verbalized in the model output. This work connects activation-level interpretability with bias evaluation for underrepresented Spanish varieties.

1 Introduction

Large language models infer user attributes—nationality, dialect, socioeconomic background—from cues never stated explicitly, and these inferences shape generation even when un verbalized: instruction-tuned models refuse over 98% of explicit demographic queries yet still condition outputs on inferred attributes [Bouchaud and Ramaciotti, 2025]. Fraser-Taliente et al. [2026] illustrate this directly with Natural Language Autoencoders (NLA): their *language-switching* case study shows a model representing a user as Russian several tokens before any lexical cue, later traced to mislabeled training data. Whether an analogous un verbalized inference occurs for Spanish is unknown, despite evidence it should matter most for the varieties most often misread—Colombian Spanish scores far below Peninsular Spanish in recognition benchmarks ($F1 = 0.282$ vs. 0.723) for reasons tracking training-data composition [Kawasaki, 2026, Mayor-Rocher et al., 2025], yet no study inspects what the model represents internally before producing that output.

We use the open-source NLA for Qwen-2.5-7B on ~ 30 matched Colombian-Spanish/English prompt pairs to probe for latent nationality, socioeconomic, or stereotype representations absent from the model’s final response. Section 2 situates this work; Section 3 details our procedure; Section 4 reports findings.

2 Related Work

Interpreting what a language model represents about its input without supervised probes has converged on two strategies: projecting or patching activations to recover output-relevant information without training [nostalgebraist, 2020, Belrose et al., 2023, Ghandeharioun et al., 2024], and training a reader model—via sparse dictionaries [Cunningham et al., 2024] or a full verbalizer—to translate activations into free text [Chen et al., 2024, Pan et al., 2024, Karvonen et al., 2025]. We adopt the latter, using the Natural Language Autoencoder (NLA) of Fraser-Taliente et al. [2026] as our probing instrument; their *language-switching* case study, where the verbalizer surfaces a latent nationality inference before any lexical cue appears in the prompt, motivates asking whether an analogous un verbalized inference occurs for Colombian Spanish.

A parallel line indicates that demographic attributes are linearly decodable from activations, regardless of the verbalization method. Bouchaud and Ramaciotti [2025] report AUC-ROC up to 0.995 for probing gender, race, and socioeconomic status from indirect cues, concentrated in middle layers; Lauscher et al. [2022] and Tang et al. [2023] corroborate this across architectures, while Hu et al. [2026] shows finer-grained attributes are distributed rather than strictly linear. Most relevant to our hypothesis is the *alignment gap*: instruction-tuned models refuse over 98% of explicit demographic queries [Bouchaud and Ramaciotti, 2025] yet still condition generation on stereotype-aligned inferences when the attribute is only implied [Neplenbroek et al., 2025, Tang et al., 2023], consistent with bias persisting in contextualized representations after debiasing [Guo and Caliskan, 2021, Tan and Celis, 2019, Bommasani et al., 2020, Huang et al., 2020, Zhang et al., 2025].

Behaviorally, this latent inference produces measurable disparities for Spanish varieties and non-native English speakers. Peninsular Spanish is consistently best recognized and generated, while Latin American varieties lag substantially—Colombia among the lowest ($F1 = 0.282$ vs. 0.723 for Spain)—tracking training-data composition rather than digital resource volume [Kawasaki, 2026, Mayor-Rocher et al., 2025, Martínez et al., 2025], and English bias-mitigation techniques do not transfer to Spanish [Robles et al., 2025]. English shows analogous effects: anchoring on perceived non-nativeness degrades response quality [Reusens et al., 2024], minoritized dialects receive more stereotyped and condescending outputs [Fleisig et al., 2024], and alignment training widens rather than narrows these gaps [Ryan et al., 2024, Mire et al., 2025, Nayeem and Rafiei, 2026, Kantharuban et al., 2023].

These two literatures remain unconnected: linear-probe studies establish *that* demographic inference diverges from verbalized output, and bias-audit studies establish *that* this divergence harms underrepresented varieties, but none use a training-free, unsupervised verbalizer with controlled explicit/implicit/neutral elicitation to localize *when* a Colombian-identity inference emerges from indirect cues within the prompt itself. This is the gap we target.

3 Method

Overview. [Fraser-Taliente et al., 2026] show that Natural Language Autoencoder (NLA) explanations can reveal that a model internally represents a user as Russian *before* any unambiguous lexical cue and *before* this belief is verbalized. We adapt this finding into a controlled probe of whether Qwen2.5-7B-Instruct internally infers *Colombian* identity from a single implicit cue, using the open-source NLA pair released with [Fraser-Taliente et al., 2026].

Extraction and quartile sampling. Let M be the target model and $h_\ell \in \mathbb{R}^d$ ($d=3584$) the residual-stream activation at layer $\ell=20$, the depth at which the released Qwen2.5-7B NLA pair was trained, and an N -token prompt. A single deterministic forward pass of M yields the full sequence of activations $\{h_\ell[1], \dots, h_\ell[N]\}$, one per token position. From this sequence we select four activations, the *last* token of each of four contiguous quartiles,

$$q_k = \lfloor kN/4 \rfloor, \quad k \in \{1, 2, 3, 4\},$$

so $h_\ell[q_1], \dots, h_\ell[q_4]$ are four distinct, full-dimensional vectors drawn from four different positions in the same forward pass — not a single activation, and not a partition of h_ℓ ’s 3584 dimensions. Querying the AV at every position is infeasible, since each call emits hundreds of tokens, so this four-point subsample keeps the per-prompt AV cost fixed regardless of N [Fraser-Taliente et al., 2026]. The last-token choice additionally keeps each query in-distribution: the AV’s supervised warm-start was trained only on prefix-final activations, and each q_k is, formally, the final token of prefix $x[1:q_k]$. Each $h_\ell[q_k]$ is passed once (no resampling) to the AV, $z_k \sim \text{AV}(\cdot \mid h_\ell[q_k])$, yielding one explanation per quartile.

Design. We construct 15 base scenarios, each realized in Spanish and in English translation (30 prompts), with five scenarios per *explicitness* level: **explicit** (unambiguous Colombian marker, positive control), **implicit** (exactly one subtle cue, analogous to the *vodka*→“Russian” case in [Fraser-Taliente et al., 2026]), and **neutral** (topically matched, no national marker, negative control distinguishing a Colombia-specific effect from generic AV confabulation).

	Explicit ($n=5$)	Implicit ($n=5$)	Neutral ($n=5$)
Spanish	explicit cue	single implicit cue	no cue
English	explicit cue	single implicit cue	no cue

Table 1: 2 (language) $\times$ 3 (explicitness) design, $n=30$ prompts.

Structured coding. Each prompt’s four quartile-level explanations (z_1, z_2, z_3, z_4) are expanded into 120 prompt $\times$ quartile units and independently coded by Claude Sonnet for *nationality*, *socioeconomic status*, and *stereotype* mentions. The coding prompt requires (i) a category marked `true` only on explicit textual evidence, (ii) default `false` for explanations consisting solely of generic model self-description (e.g. “*I am Qwen, a large language model created by Alibaba Cloud*”), a known AV failure mode rather than a substantive judgment about the prompt, and (iii) a verbatim supporting quote for every positive label. An automated audit confirms full coverage, flags failed API calls, and rejects any returned quote not found *verbatim* in its source explanation; one of 120 units failed extraction after retries and is excluded listwise ($n=119$).

Because the AV is known to confabulate plausible-sounding but contextually wrong specifics — in our data, substituting Spain, Turkey, or Canada for Colombia in an otherwise on-theme explanation — a positive nationality label conflates two distinct events: representing *some* national identity versus *Colombian* identity specifically. We therefore additionally tag each positive nationality citation for whether it names Colombia (or an unambiguous metonym, e.g. Bogotá, TransMilenio), and report both the unrestricted and the Colombia-specific rate in §4; this distinction is fixed at analysis design time, not introduced post hoc.

Statistical analysis. We report each (group $\times$ quartile) mention rate with a Wilson score interval, valid at this design’s small per-cell counts ($n=9$ – 10) where a normal approximation would not be. Implicit-vs-neutral differences at each quartile are tested with Fisher’s exact test on the corresponding 2×2 table, rather than a χ^2 test whose asymptotic assumptions are unmet here. To summarize *when* nationality is first represented, we take, for each prompt where the category appears at least once, its earliest quartile, and compare distributions between groups with a Mann–Whitney U test; prompts where it never appears are reported as a separate per-group count rather than imputed, since their absence is itself informative. All tests are two-sided and uncorrected for multiple comparisons, consistent with the exploratory scope below.

Scope. With $n=5$ per cell, results are reported as descriptive rates with qualitative evidence rather than as statistically powered effects; we extract from a single layer and a single AV sample per quartile, given hackathon-scale compute.

4 Results

Table 2 reports the rate at which AV explanations mention nationality, for each group and quartile. The implicit-cue group rises monotonically across all four quartiles (0.10→0.78), consistent with the context-accumulation hypothesis; Figure 1 shows this as a sustained positive slope with neither

the early ceiling of the explicit control nor a late collapse. Two cells deviate from monotonicity — explicit at Q2 (0.00) and neutral at Q4 (0.20) — and we resolve both below before turning to the quartile of first appearance (Table 3) and the implicit-vs-neutral comparison (Table 4).

	Q1	Q2	Q3	Q4
Explicit	0.10	0.00	1.00	1.00
Implicit	0.10	0.20	0.50	0.78
Neutral	0.10	0.10	0.60	0.20

Table 2: Nationality mention rate by group and quartile ($n=9-10$ per cell; one unit excluded for failed extraction, §3).

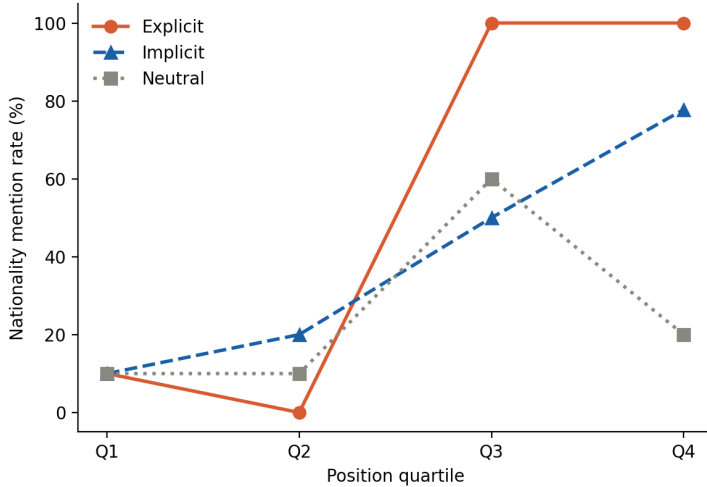


Figure 1: Nationality mention rate by quartile, all three groups. Implicit (blue, triangles) rises steadily from Q1 to Q4; explicit (orange, circles) ceils early; neutral (gray, squares) is non-monotonic, addressed in §4.

Both irregularities trace to non-Colombian confabulation, not noise. We inspected the source quotes behind explicit/Q2 and neutral/Q4 directly. Every Q2 explanation in the explicit group is off-theme confabulation unrelated to the prompt (e.g. “*Marketing de Google*”, “*conferencia Tesla*”) — the same low-anchoring failure mode documented at Q1 (§3), recurring at a second early position for short prompts. The two positive citations behind neutral/Q4 name Spain and paella, never Colombia. Restricting the nationality variable to Colombia-specific mentions (Spain/Turkey/Canada excluded; §3) removes both irregularities: neutral falls to 0.00 at every quartile, while implicit remains strictly above zero from Q2 onward (0.00/0.10/0.20/0.11). The qualitative pattern in Figure 1 therefore understates, rather than fabricates, the implicit-vs-neutral separation; we report the unrestricted rate in Table 2 for comparability with the AV’s overall behavior, and the Colombia-specific rate as the variable of record for testing the hypothesis.

	Mean Q of 1st mention	Median	Never mention (n)
Explicit	2.80	3.0	0/10
Implicit	2.88	3.0	2/10
Neutral	2.50	3.0	4/10

Table 3: Quartile of first nationality mention, computed over prompts where the category appears at least once; the right column reports prompts where it never appears.

Table 3 shows comparable mean onset quartiles across groups ($\approx Q3$) among prompts that ever trigger a mention, but a markedly different *rate* of never triggering one at all: 0/10 for explicit, 2/10 for

implicit, 4/10 for neutral. This count — not the onset quartile — carries most of the between-group signal, and is reported alongside the mean rather than imputed into it.

Quartile	k implicit	n implicit	k neutral	n neutral	p (Fisher)
Q1	1	10	1	10	1.000
Q2	2	10	1	10	1.000
Q3	5	10	6	10	1.000
Q4	7	9	2	10	0.023

Table 4: Fisher’s exact test, implicit vs. neutral, by quartile.

Only Q4 reaches conventional significance ($p=0.023$); given $n \leq 10$ per cell and four uncorrected comparisons, we read this as directional evidence for a late-quartile separation rather than a confirmatory effect (§3), consistent with the exploratory scope of this study.

See Appendix A for a qualitative, case-by-case analysis of individual prompts.

5 Conclusion

Using an unsupervised, training-free verbalizer, we find that Qwen2.5-7B-Instruct’s residual stream comes to represent Colombian identity from a single implicit lexical cue, with the Colombia-specific mention rate rising from 0.00 at Q1 to 0.20 at Q3 while the neutral control remains at 0.00 throughout. This separation reaches conventional significance only at the final quartile ($p=0.023$) and rests on $n=5$ scenarios per cell; we report it as directional evidence that an un verbalized nationality inference can emerge from a single cue, not as a confirmed effect. Two apparent irregularities in the raw mention rate were traced to a known AV failure mode — confabulating a contextually wrong country — and resolved by restricting to Colombia-specific citations, a check we recommend for any reuse of NLA explanations as a regional-identity probe. The natural next step is repeating this design at higher n per cell and at a second residual-stream layer, to test whether the late-quartile separation we observe is a property of this specific depth or holds more generally.

References

- Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.
- Rishi Bommasani, Kelly Davis, and Claire Cardie. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. *Annual Meeting of the Association for Computational Linguistics*, 2020. doi: 10.18653/v1/2020.acl-main.431. URL <https://www.aclweb.org/anthology/2020.acl-main.431.pdf>.
- Paul Bouchaud and Pedro Ramaciotti. Linear socio-demographic representations emerge in Large Language Models from indirect cues. *arXiv.org*, 2025. doi: 10.48550/ARXIV.2512.10065. URL <https://arxiv.org/abs/2512.10065>.
- Haozhe Chen, Carl Vondrick, and Chengzhi Mao. SelfIE: Self-interpretation of large language model embeddings. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.
- Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations (ICLR)*, 2024.
- Eve Fleisig, Genevieve Smith, Madeline Bossi, Ishita Rustagi, Xavier Yin, and Dan Klein. Linguistic Bias in ChatGPT: Language Models Reinforce Dialect Discrimination. *Conference on Empirical Methods in Natural Language Processing*, 2024. doi: 10.48550/ARXIV.2406.08818. URL <https://arxiv.org/abs/2406.08818>.

- Kit Fraser-Taliente, Subhash Kantamneni, Euan Ong, Dan Mossing, Christina Lu, Paul C. Bogdan, Emmanuel Ameisen, James Chen, Dzmitry Kishylau, Adam Pearce, Julius Tarnig, Alex Wu, Jeff Wu, Yang Zhang, Daniel M. Ziegler, Evan Hubinger, Joshua Batson, Jack Lindsey, Samuel Zimmerman, and Samuel Marks. Natural language autoencoders produce unsupervised explanations of LLM activations. *Transformer Circuits Thread*, 2026. URL <https://transformer-circuits.pub/2026/nla/index.html>.
- Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.
- Wei Guo and Aylin Caliskan. Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 122–133. ACM, 2021. doi: 10.1145/3461702.3462536. URL <http://dx.doi.org/10.1145/3461702.3462536>.
- Shiyue Hu, Ruizhe Li, and Yanjun Gao. Race, Ethnicity and Their Implication on Bias in Large Language Models. *medRxiv*, 2026. doi: 10.64898/2026.01.04.26343415. URL <https://www.medrxiv.org/content/medrxiv/early/2026/01/05/2026.01.04.26343415.full.pdf>.
- Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. Reducing Sentiment Bias in Language Models via Counterfactual Evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 65–83. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.findings-emnlp.7. URL <http://dx.doi.org/10.18653/v1/2020.findings-emnlp.7>.
- Anjali Kantharuban, Ivan Vulić, and Anna Korhonen. Quantifying the Dialect Gap and its Correlates Across Languages. *Conference on Empirical Methods in Natural Language Processing*, 2023. doi: 10.48550/ARXIV.2310.15135. URL <https://arxiv.org/abs/2310.15135>.
- Adam Karvonen, James Chua, Clément Dumas, Kit Fraser-Taliente, Subhash Kantamneni, Julian Minder, Euan Ong, Arnab Sen Sharma, Dean Wen, Owain Evans, and Samuel Marks. Activation oracles: Training and evaluating LLMs as general-purpose activation explainers. *arXiv preprint arXiv:2512.15674*, 2025.
- Yoshifumi Kawasaki. Digital Linguistic Bias in Spanish: Evidence from Lexical Variation in LLMs. *arXiv.org*, 2026. doi: 10.48550/ARXIV.2602.09346. URL <https://arxiv.org/abs/2602.09346>.
- Anne Lauscher, Federico Bianchi, Samuel Bowman, and Dirk Hovy. Socioprobe: What, When, and Where Language Models Learn about Sociodemographics. *Conference on Empirical Methods in Natural Language Processing*, 2022. doi: 10.48550/ARXIV.2211.04281. URL <https://arxiv.org/abs/2211.04281>.
- Gonzalo Martínez, Marina Mayor-Rocher, Cris Pozo Huertas, Nina Melero, María Grandury, and Pedro Reviriego. Spanish is not just one: A dataset of Spanish dialect recognition for LLMs. *Data in Brief*, 63:112088, 2025. ISSN 2352-3409. doi: 10.1016/j.dib.2025.112088. URL <http://dx.doi.org/10.1016/j.dib.2025.112088>.
- Marina Mayor-Rocher, Cristina Pozo, Nina Melero, Gonzalo Martínez, María Grandury, and Pedro Reviriego. It’s the same but not the same: Do LLMs distinguish Spanish varieties? *Proces. del Leng. Natural*, 2025. doi: 10.48550/ARXIV.2504.20049. URL <https://arxiv.org/abs/2504.20049>.
- Joel Mire, Zubin Trivadi Aysola, Daniel Chechelnitsky, Nicholas Deas, Chrysoula Zerva, and Maarten Sap. Rejected Dialects: Biases Against African American Language in Reward Models. *North American Chapter of the Association for Computational Linguistics*, 2025. doi: 10.48550/ARXIV.2502.12858. URL <https://arxiv.org/abs/2502.12858>.
- Mir Tafseer Nayeem and Davood Rafiei. Which English Do LLMs Prefer? Triangulating Structural Bias Towards American English in Foundation Models. 2026.

- Vera Neplenbroek, Arianna Bisazza, and Raquel Fernández. Reading Between the Prompts: How Stereotypes Shape LLM’s Implicit Personalization. *Conference on Empirical Methods in Natural Language Processing*, 2025. doi: 10.48550/ARXIV.2505.16467. URL <https://arxiv.org/abs/2505.16467>.
- nostalgebraist. Interpreting GPT: The logit lens. LessWrong, 2020. URL <https://www.lesswrong.com/posts/AcKRB8wDpdAN6v6ru/interpreting-gpt-the-logit-lens>.
- Alexander Pan, Lijie Chen, and Jacob Steinhardt. LatentQA: Teaching LLMs to decode activations into natural language. In *International Conference on Learning Representations (ICLR)*, 2024.
- Manon Reusens, Philipp Borchert, Jochen De Weerd, and Bart Baesens. Native Design Bias: Studying the Impact of English Nativeness on Language Model Performance. *IJCNLP-AACL*, 2024. doi: 10.48550/ARXIV.2406.17385. URL <https://arxiv.org/abs/2406.17385>.
- Melissa Robles, Catalina Bernal, Denniss Raigoso, and Mateo Dulce Rubio. SESGO: Spanish Evaluation of Stereotypical Generative Outputs. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2025. doi: 10.48550/ARXIV.2509.03329. URL <https://arxiv.org/abs/2509.03329>.
- Michael J. Ryan, William Held, and Diyi Yang. Unintended Impacts of LLM Alignment on Global Representation. *Annual Meeting of the Association for Computational Linguistics*, 2024. doi: 10.48550/ARXIV.2402.15018. URL <https://arxiv.org/abs/2402.15018>.
- Y. Tan and Elisa Celis. Assessing Social and Intersectional Biases in Contextualized Word Representations. *Neural Information Processing Systems*, 2019.
- Raphael Tang, Xinyu Zhang, Jimmy Lin, and Ferhan Ture. What Do Llamas Really Think? Revealing Preference Biases in Language Model Representations. *arXiv.org*, 2023. doi: 10.48550/ARXIV.2311.18812. URL <https://arxiv.org/abs/2311.18812>.
- Renhan Zhang, Lian Lian, Zhen Qi, and Guiran Liu. Semantic and Structural Analysis of Implicit Biases in Large Language Models: An Interpretable Approach. In *2025 4th International Conference on Artificial Intelligence, Internet of Things and Cloud Computing Technology (AIoTC)*, pages 699–703. IEEE, 2025. doi: 10.1109/aiotc66747.2025.11198661. URL <http://dx.doi.org/10.1109/AIoT66747.2025.11198661>.

A Technical appendices and supplementary material

A.1 Cross-Lingual Divergence in Internal Representations

The analysis of the model’s latent thoughts reveals a marked divergence in the contextualization of neutral prompts depending on the language. This variation introduces geographic and migratory biases into the representation space that are not explicit in the original input. Below, three representative cases illustrating this behavior are detailed:

Case C01: Health and Employment

Neutral prompt: *How can I enroll in private health insurance if I have just started working?* (ES/EN variants).

Latent representation: In Spanish, the model internally generates premises such as “*If you are starting to work in Spain...*” and anticipates terms like “*migratory*” or “*from the United States*”. In English, the model mentions Canada superficially, without linking it to a transitional status.

Analysis: There is a strong implicit association in Spanish between labor/health insertion and emigration, biasing the interpretation towards a context of migratory vulnerability.

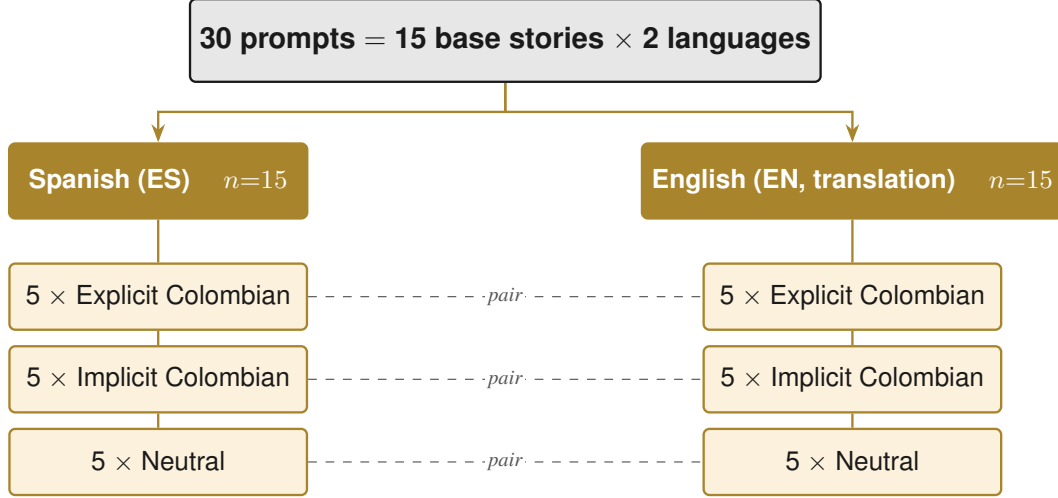


Figure 2: **Dataset structure.** Each of the 15 base stories is realized in Spanish-English pairs; the explicitness factor (explicit / implicit / neutral) varies *between* stories, yielding $n=5$ per cell.

Case C02: Urban Transportation

Neutral prompt: *What time does the last metro train run on weekends?*

Latent representation: Processing in Spanish forces an immediate geographic anchoring, generating the anticipation: “*What time is the metro in Barcelona?*”. In English, the internal context is more evenly distributed among various global metropolises (New York, London, Tokyo).

Analysis: This demonstrates a Hispanic-centric localization bias that reduces the model’s spatial generalization when faced with generic urban queries in Spanish.

Case C03: Cultural References

Neutral prompt: *What is the traditional recipe for a vegetable soup served in many cultures?*

Latent representation: In Spanish, internal activations are directed towards Ibero-American references, anticipating “*Mexico*” or “*paella*”. In contrast, in English, the search space is oriented towards the Northern Hemisphere, mentioning Italian and Turkish food, and holidays like *Thanksgiving*.

Analysis: A cultural preconditioning is evident in the pre-generation layers, where the language restricts the scope of what the model considers a “generic culture”.

Case A01: Health System and Employment

Original context: Query about enrollment in EPS (public health insurance) and the Colombian health system when starting a new job.

Latent representation: In Spanish, the model assumes the user is a “foreigner or newcomer,” injecting a migratory bias absent in the prompt. In English, the search space shifts towards health insurance in Canada, Germany, or Turkey, linking it to residency procedures.

Analysis: A strong implicit association between employment/health and migration is evident. Processing in Spanish projects local migratory vulnerability, while English completely internationalizes the query, losing the original geographic relevance.

Case A03: Gastronomy and Cultural References

Original context: Query about a traditional Colombian dish (*ajiaco*).

Latent representation: In Spanish, the model recognizes the local traditional gastronomic context, although it generalizes by mixing it with other dishes (*sancocho*, *arepas*). In English, *ajiaco* undergoes a drastic shift towards “Peruvian, Mexican food” and invents concepts like “Colombian paella.”

Analysis: This demonstrates a clear homogenization in English, where Latin American cultural identities are mixed and become interchangeable within the model’s latent space.

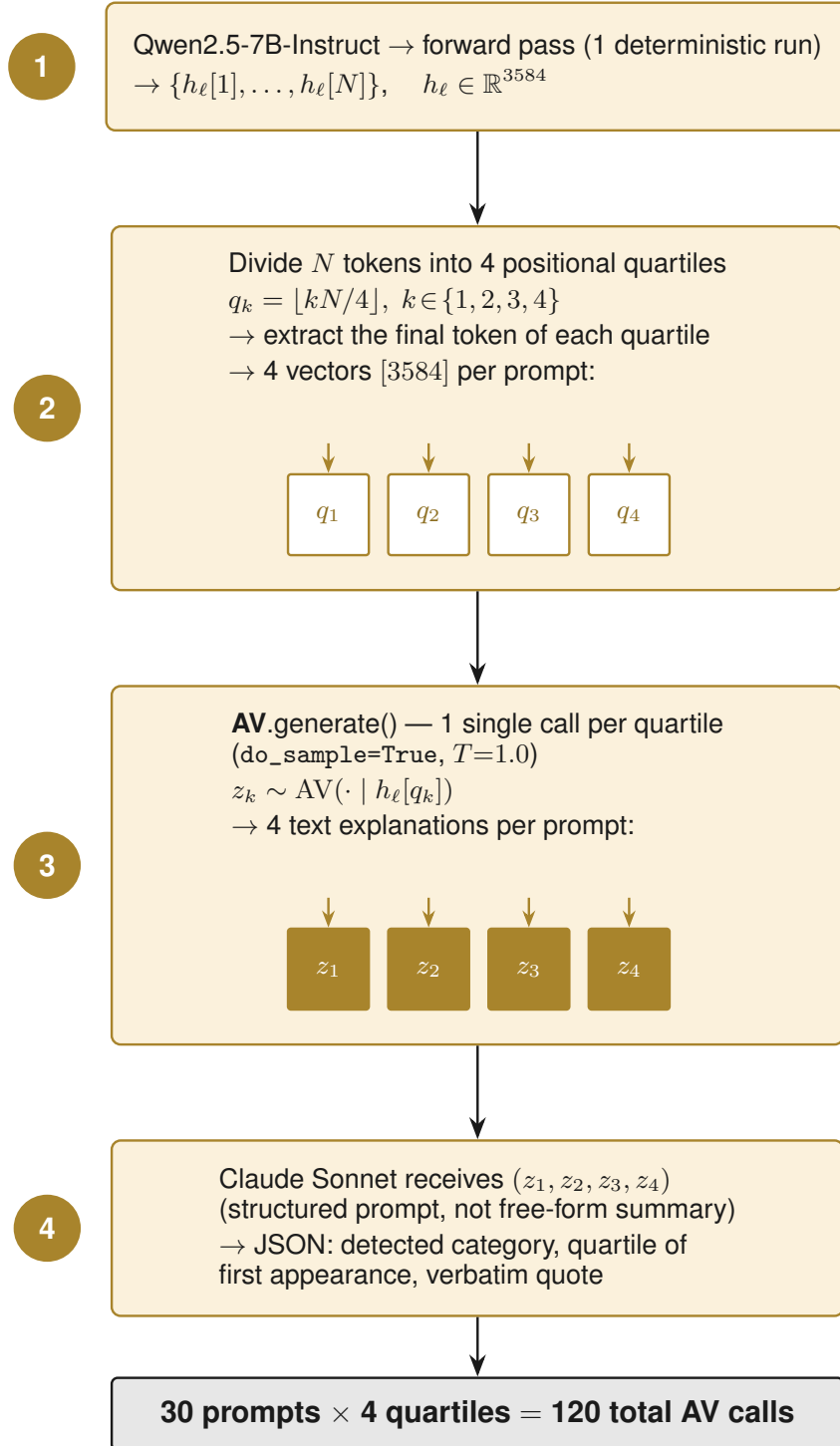


Figure 3: **Per-prompt pipeline.** Each of the 30 prompts passes through the four stages; the resulting vectors $h_\ell[q_k]$ and explanations z_k are indexed by quartile to strictly preserve the positional signal.

Case A05: Education and Financing

Original context: Query about how a university educational loan (*Icetex*) works.

Latent representation: In Spanish, the model tends to transform the concept of a loan into a state “general aid or scholarship.” In English, the model deflects the query toward international scholarships, mentioning the Erasmus program and French universities.

Analysis: This reflects a severe socioeconomic divergence: processing in Spanish associates the loan with basic local assistance and subsidies, whereas English associates it with international academic mobility.