

Probing Language Understanding in Colombian Indigenous Languages: Zero-Shot Translation, Visual Grounding, and Cultural Fidelity

Santiago Arenas Rubén Manrique

Departamento de Ingeniería de Sistemas y Computación

Universidad de los Andes, Bogotá, Colombia

{s.arenasr, rf.manrique}@uniandes.edu.co

Abstract

Language models can generate fluent, confident-sounding text even for languages they barely cover, but fluency alone reveals little about whether that text preserves situated meaning, cultural reference, and communicative intent. We study this question for Colombian Indigenous languages through two complementary probes of language understanding: (1) zero-shot machine translation across five languages (Arhuaco, Inga, Kamentsá, Nasa Yuwe, Wayuunaiki) and (2) multimodal cultural grounding on a community-built Inga image-description dataset that, to our knowledge, is the largest available for Inga. We additionally document and release this culturally grounded multimodal dataset as a reusable resource for the Inga community and future research. We evaluate GPT-5.2, GPT-4o-mini, Gemini 2.5 Flash-Lite, Qwen2-VL-72B-Instruct, NLLB-200, and T5 variants without fine-tuning. GPT-5.2 achieves the highest translation scores (chrF 21.36, BLEU 3.42), yet absolute performance remains uniformly low. No evaluated model is suitable for direct deployment, placing Colombian Indigenous languages among the most underserved in NLP even relative to other low-resource languages. For multimodal cultural grounding, semantically aligned visual context consistently improves preservation of community-provided cultural ideas, while mismatched images cause sharp degradation, a distinction that standard similarity metrics fail to capture. These findings show that fluent generation can mask failures of situated language understanding and motivate purpose-built evaluation for Indigenous language NLP.

1 Introduction

Natural Language Processing has advanced rapidly in recent years, yet this progress remains deeply uneven across languages. Colombia officially recognizes 65 Indigenous languages representing rich

cultural and linguistic diversity, but these languages share a structural disadvantage: scarce digital resources, predominantly oral traditions, and negligible representation in the training data of current models (Lucas et al., 2025).

For these languages, language understanding cannot be reduced to lexical translation or generic semantic similarity. Many expressions are tied to community practices, oral traditions, territorial knowledge, and culturally specific objects. A model that produces fluent Spanish may still fail to preserve the intended meaning, omit culturally central information, or introduce unsupported claims. We therefore treat zero-shot translation and multimodal cultural description as complementary probes of situated language understanding.

Prior NLP work on Colombian Indigenous languages has focused almost exclusively on machine translation, employing reduced Transformer architectures and transfer learning (Salazar et al., 2025), multilingual pretrained models such as NLLB (Prieto et al., 2024; Robles et al., 2024), and dictionary-guided reinforcement learning (Mosquera et al., 2025). Collectively, these studies demonstrate that careful data curation and task-specific adaptation can partially mitigate data scarcity, but they exclusively evaluate fine-tuned or adapted systems.

This distinction has practical consequences. Communities, researchers, and educators who lack the computational resources or annotated data required for fine-tuning are more likely to rely on general-purpose pretrained models, accessed via commercial APIs or open-weight releases, without any specialized adaptation. Understanding what these models can and cannot do under such conditions is a pressing research priority that has not been systematically addressed for Colombian Indigenous languages.

A recent IDB evaluation (Lucas et al., 2025) documented substantial gaps in AI performance on Indigenous languages of the Americas relative

to high-resource languages, but did not cover several Colombian Indigenous languages. In parallel, multimodal approaches have shown benefits for low-resource NLP in other linguistic contexts (Hosain et al., 2025; Das and Singh, 2022; Ahmad et al., 2025), yet their effectiveness in Colombian Indigenous settings has not been investigated.

This paper addresses these gaps through three complementary contributions:

1. **A zero-shot translation probe for language understanding.** We evaluate commercial and open-source pretrained models across five Colombian Indigenous languages without fine-tuning, establishing empirical baselines under realistic low-resource deployment conditions.
2. **A multimodal grounding probe for culturally situated meaning.** Using a curated Inga image-description dataset, we conduct a controlled five-condition modality ablation and introduce an LLM-as-a-Judge protocol that measures preservation of culturally specific information beyond standard lexical and semantic similarity metrics.
3. **An open community-built Inga multimodal dataset.** We document and release a culturally grounded image-description dataset constructed through a community-led collection protocol, including Inga names, free-form cultural descriptions, and Spanish translations verified by a paid Inga-Spanish translator from the community.

Our results show that general-purpose LLMs outperform NLLB-200 and T5 baselines in zero-shot settings, neither of which covers these languages, yet absolute scores remain low across all systems. Semantically aligned visual context consistently improves cultural grounding, while mismatched images cause sharp degradation. Standard metrics such as BERTScore and AlignScore show only weak correlation with cultural awareness, demonstrating that culturally grounded generation requires purpose-built evaluation methods. Together, these results highlight a key risk for human-machine language use: fluent model outputs may appear plausible while failing to preserve situated meaning.

2 Related Work

2.1 Language Understanding, Grounding, and Cultural Meaning

Language understanding in human-machine settings raises a central question for current NLP systems, namely whether fluency and benchmark performance correspond to robust language understanding. This question is especially acute in low-resource Indigenous language settings, where meaning is often tied to communally held knowledge, oral tradition, territorial context, and culturally specific practices. In such cases, misunderstanding may appear not as ungrammatical output, but as fluent text that omits core cultural ideas, introduces unsupported claims, or treats situated meaning as generic description.

Multimodal grounding offers one way to examine whether models use contextual evidence when interpreting culturally specific language. However, visual input does not by itself guarantee understanding. An image can support the intended meaning when it is semantically aligned, or mislead the model when it is irrelevant. Our work therefore evaluates language understanding through both translation and controlled visual grounding, with a cultural awareness score used as a reference-based measure of whether community-provided ideas are preserved.

2.2 Machine Translation in Colombian Indigenous Languages

Systematic NLP work on Colombian Indigenous languages began with the adaptation of neural architectures to low-resource conditions. Salazar et al. (2025) evaluated reduced Transformer architectures and Finnish-based transfer learning for Spanish-Wayuunaiki and Spanish-Nasa Yuwe, finding that smaller models frequently outperform full-scale Transformers in data-scarce regimes and that linguistic similarity alone does not guarantee consistent transfer gains. Robles et al. (2024) compared custom Transformers, transfer learning, and NLLB for multiple Colombian languages, showing that NLLB achieves superior results when data is carefully curated and that domain mismatch can substantially degrade performance. Prieto et al. (2024) compiled a parallel corpus for four Colombian Indigenous languages and established NLLB-based baselines, emphasizing the critical role of dataset quality over model scale. More recently, Mosquera et al. (2025) in-

roduced dictionary-guided fine-tuning with GRPO reinforcement learning for Spanish–Wayuunaiki, achieving measurable improvements. All these studies evaluate task-specific adapted systems; our work establishes what is achievable in zero-shot settings, reflecting conditions of realistic community deployment.

2.3 Multimodal NLP in Low-Resource Settings

Multimodal approaches have demonstrated consistent advantages for low-resource NLP beyond Colombian Indigenous languages. [Hossain et al. \(2025\)](#) show that fusing multilingual text embeddings with visual representations substantially improves text classification accuracy in low-resource settings. [Das and Singh \(2022\)](#) report similar gains for multimodal sentiment analysis in Assamese, and [Ahmad et al. \(2025\)](#) demonstrate improvements in named entity recognition for Urdu through cross-modal fusion. Whether these gains extend to Colombian Indigenous language settings, and whether visual information can support culturally grounded generation, remained open questions prior to this work.

Recent work has begun to evaluate cultural understanding directly in vision-language models. CulturalVQA ([Nayak et al., 2024](#)) frames cultural understanding as geo-diverse visual question answering over facets such as clothing, food, rituals, and traditions, revealing substantial regional and facet-level disparities in VLM performance. PEARL ([Alwajih et al., 2025](#)) provides a large Arabic multimodal instruction dataset and benchmark with human-in-the-loop cultural annotations, showing that culturally grounded multimodal evaluation requires domain-specific data rather than generic image-understanding benchmarks. Our setting differs in scale and resource conditions. Instead of constructing a broad VQA benchmark, we evaluate whether models preserve culturally specific ideas in community-authored Inga descriptions under controlled modality ablations. We therefore treat our main judge-based metric as a reference-based cultural awareness score rather than as a standalone measure of cultural understanding.

3 Datasets

3.1 Machine Translation

The MT dataset was compiled through multiple research efforts at the FLAG Lab, Universidad de

Language	Pairs Sources
Arhuaco	5,785 Bible, Constitution, Dict., Stories
Inga	4,552 Books, Constitution, Dict., Misc.
Kamentsá	281 Short Stories, Misc.
Nasa Yuwe	4,031 Constitution, Dict., Stories
Wayuunaiki	8,996 Bible, Books, Const., Stories, Misc.

Table 1: Machine translation dataset statistics.

los Andes ([Prieto et al., 2024](#)), with the most recent version developed by Caceres (ongoing work, used with acknowledgment). It contains parallel sentence pairs between Spanish and five Colombian Indigenous languages drawn from heterogeneous source domains including biblical texts, constitutional documents, bilingual dictionaries, literary works, and miscellaneous community texts. Table 1 summarizes dataset statistics.

The datasets differ substantially in size, domain composition, and orthographic standardization, reflecting the heterogeneous state of digitized resources for each language.

3.2 Multimodal Inga Dataset

The multimodal dataset was created by members of the Inga community from the Putumayo region of Colombia and contains 100 instances. The Inga people are a small Indigenous community. The 2005 Colombian National Census reported 15,450 self-identified Inga people ([Departamento Administrativo Nacional de Estadística, 2005](#)). This demographic context is relevant because the number of community members formally able to perform Inga–Spanish translation is limited.

The dataset was constructed through a community-centered collection protocol. Community members were asked to photograph elements they considered culturally relevant, with full freedom to choose the image content. For each image, contributors provided an Inga name indicating the main object or element and, more importantly, a free-form culturally enriched description. The instruction was not to describe only what appears visually in the image, but to explain its meaning for the community. As a result, the captions include cultural context such as cosmovision, community practices, symbolism, and oral tradition, rather than only visual attributes.

We hired and paid one of the few available Inga–Spanish translators from the community to consolidate the data: verifying and refining the Inga captions and translating them into Spanish. No ad-



Field	Content
ESP name	Arcoiris
ESP desc.	For the Inga people, the rainbow is understood as a spirit that may announce positive or negative events for the community. Associated with oral tradition, the <i>día grande</i> , colorful clothing, and textile practices...
Inga suti	Kuichi
Inga desc.	<i>Inga aillukunamanda kami sug iaia uilladus, sugllai alli imakuna tukun-gapa kaskakunamanda...</i>

Figure 1: Instance from the Inga multimodal dataset: image, Inga name (*suti*), and bilingual culturally grounded description.

ditional translation verification was performed because no other trained native Inga translators were available. Each final instance therefore pairs an image with its Inga-language name, Inga-language cultural description, and corresponding Spanish translations. A small development subset was reserved for prompt refinement and excluded from final evaluation. On average, descriptions contain about 68 Inga words and 113 Spanish words, for a total of 18,072 words across all 100 instances. The dataset is released openly for community use and future research.¹ Figure 1 shows a representative instance.

4 Experimental Setup

4.1 Models

Table 2 summarizes all evaluated models. We evaluate two categories. The *baselines* are NLLB-200 variants and T5 variants, included because they are the reference systems used in prior work on Colombian Indigenous language MT (Prieto et al., 2024; Robles et al., 2024; Salazar et al., 2025). Neither covers the languages studied here: none of the five is among the 200 languages of NLLB-200, and T5 is an English-only model whose pretraining mixture contains no Indigenous-language translation

¹<https://huggingface.co/datasets/Flaglab/CIL-Dataset>

task (§4.2). We retain both, unmodified, so that our zero-shot numbers remain directly comparable to the configurations reported in that prior work, and we read them as out-of-coverage lower bounds rather than as competitive multilingual MT systems. We deliberately use vanilla T5 rather than mT5 for the same reason: the prior Colombian Indigenous MT literature evaluated T5, and substituting mT5 would break comparability with those published baselines. The *general-purpose LLMs* are GPT-5.2, GPT-4o-mini, Gemini 2.5 Flash-Lite, and Qwen2-VL-72B-Instruct, selected for their multilingual capability, availability, and multimodal support. The cultural grounding experiment uses only the four multimodal-capable LLMs. No fine-tuning is applied to any model.

4.2 Machine Translation Protocol

All models translate from each Indigenous language to Spanish under a single zero-shot condition. Prompts instruct models to translate faithfully, preserving named entities and avoiding explanations, and to return output as structured JSON; the complete template is provided in Appendix A. Temperature is set to 0 for the LLMs; the NLLB and T5 baselines use beam search (Appendix C). Performance is measured with BLEU (Papineni et al., 2002) and chrF (Popović, 2015) (word order 0, i.e. chrF rather than chrF++); full metric configurations are given in Appendix C. Because chrF operates at the character level, it better captures partial lexical and morphological overlap in morphologically rich, low-resource languages where exact word-level matches are rare.

Baseline coverage and source-language configuration. None of the five languages has a language code in the FLORES-200 inventory used by NLLB-200, so no valid source-language token exists for them. In our runs the NLLB tokenizer is therefore instantiated with its default source language and only the target is constrained (`forced_bos_token_id = spa_Latn`): the model receives Indigenous-language input under a source-language assumption that does not hold. T5 is prompted with a task prefix of the form "translate <lang> to Spanish:", a task-language pair absent from its pretraining mixture, which covers only English into German, French, and Romanian. Both baselines are thus evaluated strictly outside their supported language sets. We state this explicitly because it is the concrete form

Model	Provider	Type	Parameters	Multilingual	Multimodal	Access
GPT-5.2	OpenAI	Decoder LLM	Undisclosed	✓	✓	API (commercial)
GPT-4o-mini	OpenAI	Decoder LLM	Undisclosed	✓	✓	API (commercial)
Gemini 2.5 Flash-Lite	Google	Decoder LLM	Undisclosed	✓	✓	API (commercial)
Qwen2-VL-72B-Instruct	Alibaba	Decoder LLM	72B	✓	✓	Open-weight
NLLB-200-distilled-600M	Meta	Encoder-Decoder	600M	✓	×	Open-weight
NLLB-200-distilled-1.3B	Meta	Encoder-Decoder	1.3B	✓	×	Open-weight
T5-small	Google	Encoder-Decoder	60M	×	×	Open-weight
T5-base	Google	Encoder-Decoder	220M	×	×	Open-weight

Table 2: Models evaluated in this study. “Undisclosed” indicates a parameter count not publicly released at the time of submission. Checkmarks indicate native multilingual or multimodal capability. NLLB-200 is multilingual over 200 languages, none of which is among the five evaluated here (§4.2); T5 is English-only.

the resource gap takes for these languages: the standard remedy of “use a multilingual MT model” is unavailable, not merely weak.

4.3 Multimodal Cultural Grounding Protocol

We conduct a controlled ablation across five input conditions that isolate the contribution of textual and visual information:

- **TEXT_NAME**: Inga name only.
- **TEXT_FULL**: Inga name + Inga description.
- **IMG_NAME**: Image + Inga name.
- **IMG_FULL**: Image + Inga name + Inga description.
- **IMG_MISMATCH**: Unrelated image + Inga name.

Models are prompted to produce the Spanish name and a culturally informed Spanish description for each cultural object. The system prompt and per-modality user-message templates are included in Appendix B. All outputs and evaluations are conducted in Spanish to leverage reliable semantic tools unavailable for Inga.

Evaluation Metrics. We report four complementary metrics. BERTScore (Zhang et al., 2020) measures contextual embedding similarity between generated and reference texts. AlignScore (Zha et al., 2023) evaluates factual consistency and entailment alignment. Name F1 and Description F1 measure token-level overlap for the name and description fields respectively.

Cultural Awareness Score. Standard metrics are insufficient for culturally grounded generation, as they do not distinguish culturally specific content from generic fluent output. We therefore introduce

an *LLM-as-a-Judge* evaluation protocol (Zheng et al., 2023), informed by recent multimodal cultural benchmarks (Nayak et al., 2024; Alwajih et al., 2025).

All cultural awareness scores are produced by a single judge model, o3-mini, accessed through Azure OpenAI (API version 2024-02-01) with a fixed prompt and a 4,096-token completion budget; no temperature or reasoning-effort parameter is set, so provider defaults apply. The judge decomposes both the reference and generated Spanish descriptions into sets of core cultural ideas, aligns them by semantic equivalence (allowing paraphrases), and computes the cultural awareness score as

$$\text{cultural_score} = \frac{n_{\text{matched}}}{n_{\text{reference}}}, \quad (1)$$

$$\text{cultural_score} \in [0, 1].$$

The judge additionally records hallucinated ideas (predicted content unsupported by the reference), enabling transparent error analysis. We use a text-only judge because the evaluated object is the generated Spanish cultural description relative to the community-provided Spanish reference, not open-ended visual recognition. Prompts are fixed across all evaluations. The multimodal benchmark comprises 2,000 evaluations (100 instances $\times$ 5 conditions $\times$ 4 models), and we report the accounting in full because it affects how the means should be read. In 3 evaluations the judge itself failed; these yield no score and are excluded, so all reported means are computed over the remaining 1,997 evaluations. In 17 evaluations the model returned no parsable description; these are retained and scored zero, since an empty output preserves none of the reference ideas. In a further 30 evaluations the judge extracted no ideas from the reference, leaving Equation (1) undefined; these are

also recorded as zero. Both conventions bias the reported means downward rather than upward, and the effect is small: excluding the 17 empty generations changes the pooled mean cultural awareness from 0.129 to 0.130.

Self-preference bias. The judge shares a provider with two of the four evaluated systems (GPT-5.2 and GPT-4o-mini), and GPT-5.2 obtains the highest cultural awareness scores, so provider-shared judging is a legitimate concern (Zheng et al., 2023). Three properties of our protocol limit its scope. First, the judge belongs to a different model series than any evaluated system, and in particular is not the model under evaluation. Second, the task is reference-based rather than preference-based: the judge never sees more than one candidate at a time and never ranks systems, scoring each output only against a fixed, human-authored Spanish reference produced by an Inga–Spanish translator from the community. The denominator $n_{\text{reference}}$ is therefore determined by the human reference and is independent of the candidate. Third, and most importantly, the claim our multimodal experiment rests on is a *within-model* contrast across input conditions (IMG_FULL vs. IMG_MISMATCH), which holds the generator fixed and is consequently insensitive to any constant judge preference for a given provider. We nevertheless cannot rule the bias out for the *between-model* ranking: idea matching involves discretion over what counts as a paraphrase, and stylistic affinity between judge and generator could make such matches more permissive. Absolute values for the two OpenAI systems should be read with this caveat, and cross-provider or multi-judge corroboration is required before the between-model ranking is treated as settled.

5 Results

5.1 Machine Translation

Figure 2 shows that GPT-5.2 achieves the highest average BLEU (3.42) and chrF (21.36) across all five languages, followed by Gemini 2.5 Flash-Lite (BLEU 2.28, chrF 19.72). GPT-4o-mini and Qwen2-VL-72B-Instruct show intermediate performance. The NLLB and T5 baselines consistently underperform all general-purpose LLMs. This is expected rather than surprising once coverage is taken into account: neither baseline supports any of the five source languages, so their scores are out-of-coverage lower bounds (§4.2). Their inclusion remains informative, because the absence of

these languages from a 200-language translation model is itself a measure of how underserved they are, and because the general-purpose LLMs are equally unadapted yet substantially better. Despite their relative success, absolute BLEU values remain uniformly low across all systems, confirming that translation into Colombian Indigenous languages is difficult for current pretrained models.

Figure 3 reveals substantial dataset-level variation. chrF provides more differentiated scores than BLEU across all model–language combinations, consistent with the expectation that character-level metrics better capture partial morphological overlap in low-resource morphologically rich languages. Kamentsá yields the highest average BLEU (2.47), but it is also the smallest dataset by a large margin (281 pairs) with limited domain diversity, which likely inflates word-level overlap scores rather than reflecting superior language-level translatability. Arhuaco and Wayuunaiki remain consistently harder under both metrics.

Table 3 illustrates two characteristic failure modes identified in the Arhuaco dataset and includes English glosses/back-translations to make the evidence checkable for non-Spanish readers. GPT-5.2 produces fluent Spanish output but hallucinated content (water, letters, arrival) absent from the reference. Qwen2-VL-72B-Instruct copies the source text verbatim without translating it. These qualitatively distinct failures suggest that different mitigation strategies (faithfulness constraints versus translation capability improvements) may be needed for different model families.

5.2 Multimodal Cultural Grounding

Figure 4 shows that GPT-5.2 consistently achieves the highest cultural awareness scores across all input conditions. For the three models that achieve non-trivial cultural awareness (GPT-5.2, Gemini 2.5 Flash-Lite, Qwen2-VL-72B-Instruct), IMG_FULL yields the highest score and IMG_MISMATCH the lowest or near-lowest. GPT-4o-mini is the exception: its scores remain close to zero in every condition (0.012–0.096), so its ordering is not informative. Where the effect is measurable at all, it is driven by semantically relevant visual content rather than by the mere presence of an image. Pooled across the four models, the modality effect is roughly four to five times larger in cultural awareness (0.24 between the best and worst condition) than in Description F1, BERTScore, or AlignScore (0.04–0.06): the standard metrics are

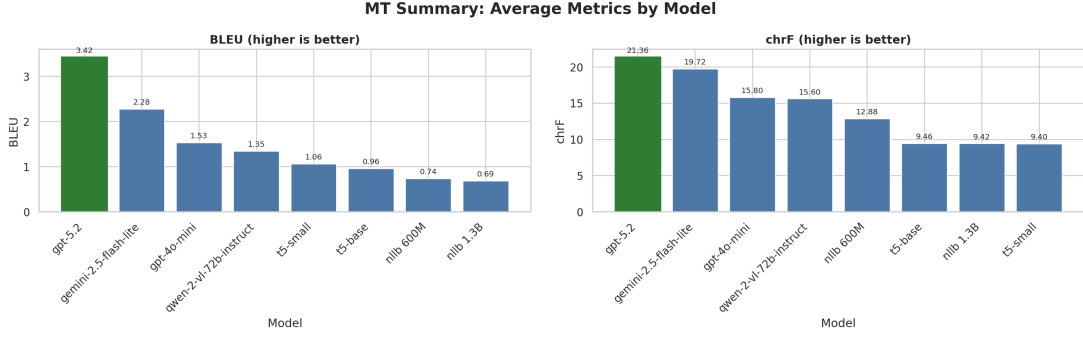


Figure 2: Average MT performance by model (BLEU and chrF). GPT-5.2 achieves the highest scores on both metrics; NLLB and T5 variants underperform general-purpose LLMs across all five languages.

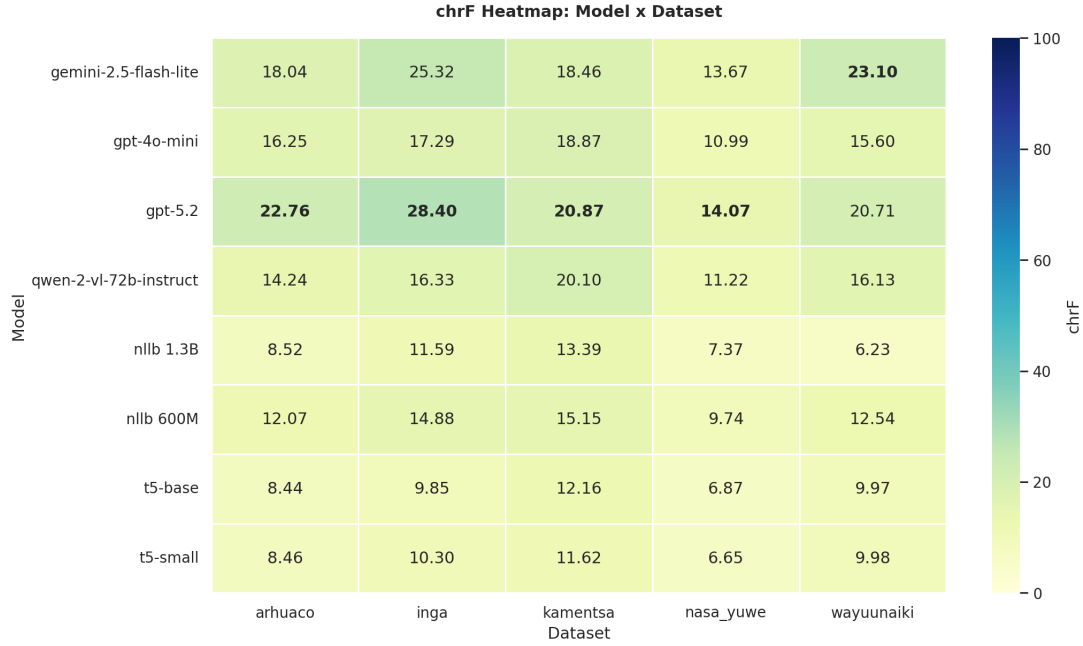


Figure 3: chrF scores by model and dataset. GPT-5.2 achieves the highest per-language scores in Arhuaco (22.76), Inga (28.40), Kamentsá (20.87), and Nasa Yuwe (14.07). Gemini 2.5 Flash-Lite leads in Wayuunaiki (23.10).

largely blind to the manipulation.

Figure 5 presents the inter-metric correlation structure. The cultural awareness score shows moderate correlation with Description F1 (0.46), indicating that it is partly related to content overlap, as expected for a reference-based metric. However, its weaker correlation with BERTScore and AlignScore suggests that idea-level decomposition captures omissions and unsupported cultural claims that are not fully reflected by embedding-based similarity. This motivates using the LLM-as-a-Judge protocol as a complementary evaluation axis. A model can achieve moderate BERTScore or AlignScore while still omitting critical cultural knowledge or introducing unsupported cultural claims.

6 Discussion

Translation as a probe of understanding. General-purpose LLMs outperform the NLLB-200 and T5 baselines in zero-shot translation across all five Colombian Indigenous languages. This comparison must be read carefully: NLLB-200, despite being massively multilingual, includes none of these five languages among its 200 supported codes, and T5 is English-only, so both operate entirely out of coverage (§4.2). The conclusion is therefore not that scale beats task-specific training in general, but that for languages excluded even from flagship multilingual MT resources, general-purpose instruction-following LLMs are currently the stronger zero-shot option available to communities, while still falling far short of usable quality.

	Text
Source	<i>ema wa'nun kugise' sabana cresco awiri jewrwa, uya unkuwaséykumey, eyméy nannundi ey a'gwaku' nunno izari, kartiya a'chunákuma gun...</i>
Reference	el año anterior en las regiones de sabana cresco y jewrwa, al informarse sobre el desarrollo de esta traducción se apreció su conveniencia e importancia <i>EN gloss: The previous year, in the regions of Sabana Cresco and Jewrwa, after learning about the development of this translation, its convenience and importance were recognized.</i>
GPT-5.2	Entonces, en la sabana de Crespo, Awiri Jewrwa, se reunieron; y así, ellos dijeron que el agua no estaba allí; la carta la enviaron para que él la viera; y luego, cuando llegaron, no encontraron nada allí. <i>EN back-translation: Then, in the Crespo savanna, Awiri Jewrwa, they gathered; and so they said that the water was not there; they sent the letter so that he would see it; and later, when they arrived, they found nothing there.</i>
Qwen2-VL	<i>ema wa'nun kugise' sabana cresco awiri jewrwa, uya unkuwaséykumey, eyméy nannundi ey a'gwaku' nunno izari, kartiya a'chunákuma gun...</i> <i>EN note: identical to the Arhuaco source text above; no Spanish translation was produced.</i>

Table 3: Qualitative MT example (Arhuaco), with English glosses/back-translations added for verifiability. GPT-5.2 hallucinates content not present in the reference; Qwen2-VL copies the source without translation.

However, the uniformly low absolute scores across all evaluated systems, whether commercial or open-source, indicate that no current model is suitable for direct deployment in real community contexts. This positions Colombian Indigenous languages among the most underserved even within the low-resource NLP landscape, where performance gaps relative to high-resource and even other minority languages remain severe. The two qualitatively distinct failure modes (hallucination of fluent but unfaithful content and source copying) are not randomly distributed across models and likely require different mitigation strategies. The chrF-BLEU divergence is also methodologically important, since chrF provides more informative differentiation in these morphologically rich, low-resource language settings, and should be treated as the primary metric in future evaluations.

These errors illustrate a central risk in human-machine language understanding: fluency can mask misunderstanding. In our qualitative examples,

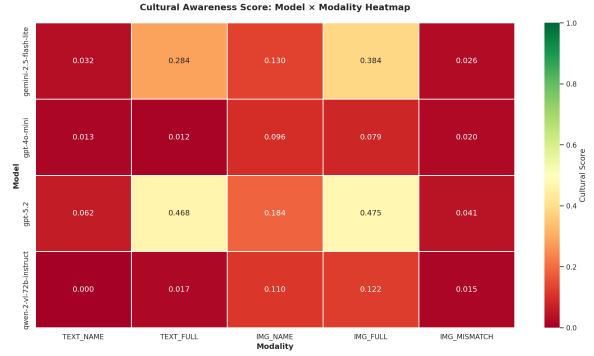


Figure 4: Cultural awareness score by model and input modality. IMG_FULL yields the highest scores for three of the four models; for GPT-4o-mini, whose scores are uniformly low, IMG_NAME is marginally higher. IMG_MISMATCH produces the lowest or near-lowest scores throughout, confirming that gains depend on semantically aligned visual context.

models can generate coherent Spanish while failing to preserve the source meaning or while introducing unsupported events. Such outputs are especially risky in Indigenous language settings, where users may lack independent tools or sufficient support to verify model behavior.

Why two probes. Our two probes are motivated by the claim that surface adequacy and situated meaning can come apart, and our measurements support this quantitatively. Within the multimodal probe the dissociation is direct. Pooling the four models, the cultural awareness score varies by 0.24 across input conditions (from 0.03 under IMG_MISMATCH to 0.27 under IMG_FULL), whereas Description F1, BERTScore, and AlignScore vary by only 0.04–0.06 over the same conditions. The same manipulation that changes how much community-provided cultural content survives leaves standard similarity metrics almost unmoved, and cultural awareness correlates only moderately with Description F1 ($r = 0.46$) and weakly with BERTScore ($r = 0.20$) and AlignScore ($r = 0.18$). Across probes the two views of the same language also diverge descriptively: Inga is the strongest language in the translation probe (chrF 28.40 for GPT-5.2, the highest single cell in Figure 3), yet on the Inga multimodal data that same model recovers about a quarter of the reference cultural ideas when averaged over the five conditions (0.25), and under half even in its best condition (0.48 under IMG_FULL). Relative competence at producing Spanish text from Inga input does not carry over to preserving what that

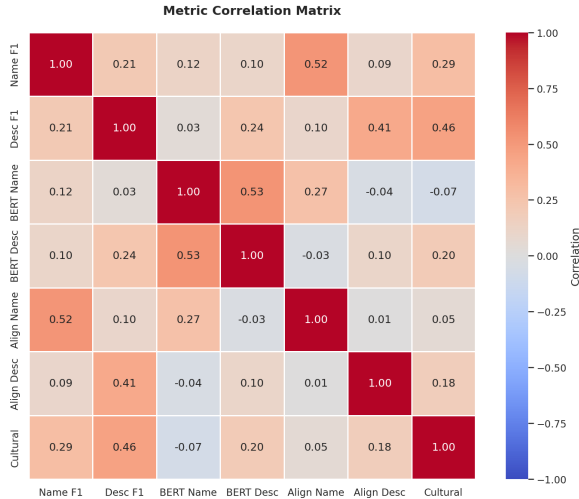


Figure 5: Correlation matrix across evaluation metrics. The cultural awareness score shows moderate correlation with Description F1 (0.46) and weak correlation with BERTScore and AlignScore, indicating that it captures complementary aspects of model behavior.

text is about. We report the latter as a descriptive comparison rather than a correlation, since the two probes draw on disjoint item sets with no shared instances (see Limitations).

Multimodal grounding and cultural awareness.

The modality ablation provides controlled evidence of visual grounding. IMG_FULL outperforms all text-only conditions, while IMG_MISMATCH falls to the level of the minimal TEXT_NAME condition, and below it for the two strongest models. This extends findings from multimodal low-resource NLP (Hossain et al., 2025) to Colombian Indigenous language settings and further demonstrates that the improvement is content-specific rather than a generic multimodal effect. Standard metrics (BERTScore, AlignScore) do not capture this modality sensitivity, reinforcing the need for a reference-based cultural awareness metric that explicitly targets culturally specific content and unsupported claims.

7 Conclusion

We have presented the first systematic zero-shot evaluation of pretrained language models as probes of situated language understanding for Colombian Indigenous languages. General-purpose LLMs, led by GPT-5.2, outperform out-of-coverage NLLB-200 and T5 baselines in zero-shot translation across five languages, yet absolute performance remains low across all systems, establishing an empiri-

cally grounded floor for future work. No evaluated model, whether commercial or open-source, achieves performance levels suitable for direct use in real community settings. This reflects a gap that is stark even by low-resource standards; Colombian Indigenous languages remain among the most data-scarce and underrepresented in current NLP research. To our knowledge, the dataset employed here is the largest image–description dataset for Inga available to date. For cultural awareness, our controlled modality ablation and LLM-as-a-Judge protocol demonstrate that semantically aligned visual context substantially improves preservation of community-provided cultural ideas, while mismatched images cause sharp degradation. Standard similarity metrics underestimate the challenge of culturally grounded generation in Indigenous language settings. These findings motivate (1) expanded multimodal corpora for additional Colombian Indigenous languages, (2) fine-tuning and retrieval-augmented studies using our zero-shot results as baselines, (3) community-aware evaluation protocols that involve native speakers and cultural experts in assessing model outputs, and (4) bidirectional evaluation, including generation into these languages, paired with community-led human assessment.

Limitations

The multimodal dataset contains 100 instances, limiting statistical generalizability. Cultural awareness is evaluated via Spanish proxy, since reliable Inga-language evaluation tools are unavailable; the judge-based score therefore approximates fidelity to the community-provided reference rather than directly measuring cultural understanding in Inga. MT datasets vary substantially in domain and size, meaning observed performance differences partially reflect dataset characteristics rather than pure language-level difficulty.

One-directional evaluation. We evaluate only the Indigenous→Spanish direction. Generation *into* these languages is the harder and, for community use, often the more consequential direction, and our results say nothing about it: a model may achieve non-trivial comprehension of Indigenous input while being unable to produce well-formed Indigenous output at all. The asymmetry is not incidental. Reference-based automatic metrics would have to be applied to morphologically rich languages with unstable orthographic standardisation,

where character-level overlap is a poor proxy for adequacy, and credible assessment would require native-speaker judgement at a scale our resource constraints could not support (§3.2). We therefore report the direction we could evaluate credibly and leave Spanish→Indigenous benchmarking, with community-led human evaluation, to future work.

Single-judge design. All cultural awareness scores come from one judge model from a single provider, which is also the provider of two of the four evaluated systems (§4.3). We did not measure judge self-consistency across repeated runs, nor agreement between the judge and human Inga speakers, so the metric is characterised by its construction rather than empirically validated. Multi-judge ensembles, cross-provider judges, and human-agreement studies are necessary before judge-based cultural scores are used for system selection.

No cross-probe instance-level analysis. The two probes use disjoint item sets (parallel-corpus sentences versus community-authored image descriptions) and per-instance translation scores were not retained, so we cannot test at the instance level whether translation adequacy predicts preservation of cultural content. The cross-probe relationship reported in §6 is descriptive.

Finally, commercial model outputs may vary with provider updates, affecting long-term reproducibility.

Ethical Statement

The multimodal Inga dataset was created by members of the Inga community of Putumayo, Colombia. Community contributors selected culturally relevant images themselves and provided culturally grounded descriptions rather than purely visual captions. A community member with Inga–Spanish translation expertise was hired and paid to verify and refine the Inga captions and translate them into Spanish; no further translation validation was possible because of the limited availability of trained native translators. The dataset is released openly for community use and future research. The machine translation dataset was compiled from academic publications in which Indigenous community members received formal acknowledgment and compensation as part of the original research process, with institutional oversight. No model fine-tuning was performed; all evaluations used commercially avail-

able or publicly released models. Results should not be interpreted as measures of cultural competence for deployment without community involvement.

Acknowledgements

We are especially grateful to Manuel Muyuy Chasoy for his support in the translation and construction of the multimodal dataset. We also thank the FLAG Lab at Universidad de los Andes for the machine translation corpora, and Caceres for sharing the most recent version of the parallel data.

References

- Hussain Ahmad, Qingyang Zeng, and Jing Wan. 2025. [A benchmark dataset and a framework for Urdu multimodal named entity recognition](#). *IEEE Access*, 13:100904–100919.
- Fakhraddin Alwajih, Samar M. Magdy, Abdellah El Mekki, Omer Nacar, Youssef Nafea, Safaa Taher Abdelfadil, Abdulfattah Mohammed Yahya, Hamzah Luqman, Nada Almarwani, Samah Aloufi, Baraah Qawasmeh, Houdaifa Atou, Serry Sibae, Hamzah A. Alsayadi, Walid Al-Dhabyani, Maged S. Al-shaibani, Aya El aatar, Nour Qandos, Rahaf Alhamouri, and 18 others. 2025. [Pearl: A multimodal culturally-aware Arabic instruction dataset](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23048–23079, Suzhou, China. Association for Computational Linguistics.
- Ringki Das and Thoudam Doren Singh. 2022. [A multi-stage multimodal framework for sentiment analysis of Assamese in low resource setting](#). *Expert Systems with Applications*, page 117575.
- Departamento Administrativo Nacional de Estadística. 2005. Censo nacional de población 2005. Technical report, Departamento Administrativo Nacional de Estadística (DANE).
- Md Rajib Hossain, Sadia Afroze, Asif Ekbal, Mohammed Moshikul Hoque, and Nazmul Siddique. 2025. [MultiModFuseNet: Advancing multimodal text classification for low-resource languages through textual-visual feature fusion](#). *Knowledge-Based Systems*, 328:114085.
- Miguel Lucas, Alejandro Burgueño, César Buenadicha Sánchez, Miguel Carazas, Smeldy Ramirez Rufino, César Said Rosales Torres, Daniel Korn, Hiwot Tesfaye, and Gretchen Deo. 2025. [El desempeño de la inteligencia artificial en el uso de lenguas indígenas americanas](#). Technical report, Inter-American Development Bank.
- Manuel Mosquera, Melissa Robles, Johan Rodríguez, and Rubén Manrique. 2025. [Improving low-resource translation with dictionary-guided fine-tuning and](#)

RL: A spanish-to-wayunaiki study. *Preprint*, arXiv:2508.19481.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Karolina Stanczak, and Aishwarya Agrawal. 2024. [Benchmarking vision language models for cultural understanding](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5769–5790, Miami, Florida, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [BLEU: A method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Maja Popović. 2015. [chrF: Character n-gram F-score for automatic MT evaluation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

Juan Prieto, Cristian Martinez, Melissa Robles, Alberto Moreno, Sara Palacios, and Rubén Manrique. 2024. Translation systems for low-resource colombian indigenous languages, a first step towards cultural preservation. Technical report, Universidad de los Andes.

Melissa Robles, Cristian A. Martínez, Juan C. Prieto, Sara Palacios, and Rubén Manrique. 2024. [Preserving heritage: Developing a translation tool for indigenous dialects](#). *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 1200–1203.

Ivan Salazar, Ruben Manrique, and Bernardo Pereira Nunes. 2025. [Machine translation strategies for low-resource colombian indigenous languages](#). *SN Computer Science*, 6(6).

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. [AlignScore: Evaluating factual consistency with a unified alignment function](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [BERTScore: Evaluating text generation with BERT](#). In *International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhonghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, et al. 2023. [Judging LLM-as-a-judge with MT-Bench and chatbot arena](#). *Preprint*, arXiv:2306.05685.

A Machine Translation Prompt

```
# Commercial MT Prompt Template
```

```
## System Prompt
```

```
```text
You are a professional translator for
Colombian Indigenous languages and Spanish
.
Translate faithfully from SOURCE_LANG to
TARGET_LANG.
Rules:
1) Preserve meaning, tone, and named entities.
2) Do not add explanations, notes, or extra
fields.
3) If a term has no direct equivalent, keep
the original term and continue naturally.
4) Output only valid JSON with exactly one key:
{"translation":"..."}.
```

```
User Prompt Template
```

```
```text
SOURCE_LANG: {source_lang}
TARGET_LANG: {target_lang}
TEXT:
{source_text}
```

```
Return JSON only.
```
```

```
Example Variables
```

```
- `source_lang`: `ingá`
- `target_lang`: `spanish`
- `source_text`: row text from source column
(`ing`, `arh`, `kam`, `nas`, `way`)
```

```
Recommended Generation Parameters
```

```
- `temperature = 0`
```

## B Multimodal Cultural Grounding Prompt

You are a bilingual specialist in the Inga language and culture from the Putumayo region of Colombia.

Your task is to analyze the provided information about a traditional Inga cultural item and produce a culturally informed Spanish name and description.

Instructions:

- Carefully study the information provided (name, description, and/or image, depending on the test).
- Produce:
  - Name of the item in Spanish, if identifiable.
  - A detailed description that incorporates relevant cultural context of the Inga Indigenous people of Colombia, when applicable.
- If cultural elements are not clearly identifiable from the provided information

, do NOT invent them. Describe possible cultural relevance cautiously.

- Do not speculate beyond what can be reasonably inferred from the provided information.
- Always respond ONLY in valid JSON format.
- Do not include explanations outside the JSON.
- The Spanish description ("desc\_esp") must be approximately 112 words long. Do not write significantly more or fewer words than this target.

JSON structure:

```
{
 "nombre_esp": "ONLY the name in Spanish, or
 null if no name can be identified",
 "desc_esp": "Detailed description
 integrating the provided information
 with relevant cultural context of the
 Inga people, considering cosmovision,
 territory (Putumayo), traditional
 practices, symbolism, attire,
 rituality, crafts, or community
 elements as applicable."
}
```

---

<!-- MODALITY PROMPTS -documentation only; not  
part of the system prompt sent to the  
model -->

## Per-Modality User Prompts (Reference)

The templates below document what each  
benchmark modality sends as the \*\*user  
message\*\*.

They are defined in `benchmark/modalities.py`  
and combined with the system prompt above  
at runtime.

---

### TEXT\_NAME -Inga name only (text-only)

**\*\*Model receives:\*\*** The Inga-language name of  
a traditional cultural item.

**\*\*Model must output:\*\*** Spanish name +  
culturally-informed Spanish description in  
JSON.

---

The following is the name of a traditional  
Inga cultural item  
written in the Inga language:

Inga name: <nombre\_inga>

Respond with a JSON object that has exactly  
two string fields: "nombre\_esp"  
(the Spanish name for the item) and "desc\_esp"  
(a Spanish description of  
the item). Do not include any text outside the  
JSON.

---

### TEXT\_FULL -Inga name + Inga description (text-only)

**\*\*Model receives:\*\*** The Inga-language name and  
full Inga-language description of a  
traditional cultural item.

**\*\*Model must output:\*\*** Spanish name +  
culturally-informed Spanish description in  
JSON.

---

The following is the name and description of a  
traditional Inga cultural item  
written in the Inga language:

Inga name: <nombre\_inga>  
Inga description: <desc\_inga>

Respond with a JSON object that has exactly  
two string fields: "nombre\_esp"  
(the Spanish name for the item) and "desc\_esp"  
(a Spanish description of  
the item). Do not include any text outside the  
JSON.

---

### IMG\_NAME -Image + Inga name (vision)

**\*\*Model receives:\*\*** An image of the cultural  
item and its Inga-language name.

**\*\*Model must output:\*\*** Spanish name +  
culturally-informed Spanish description in  
JSON.

---

An image of a traditional Inga cultural item  
is provided.

The item's name in the Inga language is:

Inga name: <nombre\_inga>

Respond with a JSON object that has exactly  
two string fields: "nombre\_esp"  
(the Spanish name for the item) and "desc\_esp"  
(a Spanish description of  
the item). Do not include any text outside the  
JSON.

---

### IMG\_FULL -Image + Inga name + Inga  
description (vision)

**\*\*Model receives:\*\*** An image of the cultural  
item, its Inga-language name, and its Inga  
-language description.

**\*\*Model must output:\*\*** Spanish name +  
culturally-informed Spanish description in  
JSON.

---

An image of a traditional Inga cultural item  
is provided.

The item's name and description in the Inga  
language are:

Inga name: <nombre\_inga>

```

Inga description: <desc_inga>

Respond with a JSON object that has exactly
two string fields: "nombre_esp"
(the Spanish name for the item) and "desc_esp"
(a Spanish description of
the item). Do not include any text outside the
JSON.
...

IMG_MISMATCH -Mismatched image + Inga name
(vision, adversarial)

Model receives: An image that does NOT
correspond to the named item, plus the
Inga-language name of the correct item.
Purpose: Tests whether the model relies on
its cultural/linguistic knowledge rather
than blindly describing the (wrong) image.
Model must output: Spanish name +
culturally-informed Spanish description in
JSON.

...

An image of a traditional Inga cultural item
is provided.
The item's name in the Inga language is:

Inga name: <nombre_inga>

Respond with a JSON object that has exactly
two string fields: "nombre_esp"
(the Spanish name for the item) and "desc_esp"
(a Spanish description of
the item). Do not include any text outside the
JSON.
...

```

**Metric implementations.** BLEU and chrF are computed with sacrebleu at default settings: BLEU with the 13a tokenizer, and chrF with word\_order = 0, i.e. plain character  $n$ -gram F-score and *not* chrF++. BERTScore uses bert-score 0.3.13 with lang="es", which selects the library's default multilingual checkpoint, without baseline rescaling; values are therefore not comparable to rescaled BERTScore figures elsewhere. AlignScore uses the AlignScore-base checkpoint over roberta-base in nli\_sp mode. Name F1 and Description F1 are unigram token-overlap F1 over lowercased, whitespace-tokenised text.

**Judge configuration.** See §4.3.

## C Reproducibility and Configuration Details

**LLM decoding.** The four general-purpose LLMs were queried with temperature 0.0 and a maximum of 256 completion tokens for machine translation and 512 for multimodal generation. No top\_p, top\_k, or generation seed was specified, so provider defaults apply; each request was retried up to three times on transport or parsing failure. No dated model snapshots were pinned, so exact provider-side versions cannot be guaranteed.

**Local baselines.** NLLB-200-distilled-600M and -1.3B were run from the public checkpoints with beam search (num\_beams = 5, early\_stopping, max\_length = 512), forcing spa\_Latn as target; no source-language token was set, for the reason given in §4.2. T5-small and T5-base were run from the google-t5 checkpoints with a natural-language task prefix.