Evaluating LLMs on Bolivian Quechua: A Ground-Truth-Based Approach for Low-Resource Languages

Andy Garcia UAGRM	Niko Witczak UAGRM	Max Baldiviezo IXPANTIA
Sebastian Martinez UAGRM		Joel Brugman UAGRM

With

Apart Research

Abstract

Large Language Models (LLMs) have achieved remarkable performance in widely represented languages such as English and Spanish. However, their ability to understand and generate content in low-resource and indigenous languages remains largely unexplored. This work presents TriloByte, an evaluation framework designed to assess how effectively modern LLMs capture the meaning of vocabulary in Bolivian Quechua, an indigenous language with limited representation in contemporary AI training datasets.

Using a bilingual Quechua–Spanish dictionary as ground truth, we compare dictionary definitions with model-generated explanations through semantic similarity metrics and human evaluation. Our methodology combines lexical resources, embedding-based similarity analysis, and expert judgment to assess adequacy, completeness, fluency, and semantic alignment across multiple language models. The results provide insights into the strengths and limitations of current LLMs when processing underrepresented languages and highlight the challenges of evaluating semantic understanding beyond direct lexical matching.

By introducing a reproducible evaluation framework for Bolivian Quechua, this work contributes to ongoing efforts in AI evaluation, linguistic inclusion, and the development of more reliable language technologies for low-resource language communities.

1. Introduction

The Problem and Its Importance

Large Language Models (LLMs) are primarily trained and evaluated using massive datasets collected from the internet, resulting in a strong bias toward widely represented languages such as English and Spanish. Indigenous and low-resource languages, including Bolivian Quechua and its regional variants, are significantly underrepresented in these training corpora. As a result, modern language models often struggle to accurately understand, generate, and interpret content in these languages.

The consequences of this limitation extend beyond translation quality. When interacting with low-resource languages, LLMs may produce inaccurate definitions, lexical hallucinations, or semantically inconsistent responses. In contexts such as public communication, education, healthcare, or community outreach, these failures can contribute to misinformation, reduce institutional trust, and exclude vulnerable linguistic communities from the benefits of modern AI systems. Understanding and measuring these limitations is therefore an important challenge for both AI Safety and linguistic inclusion.

Context and Threat Model

This work addresses the problem of semantic misalignment, lexical hallucination, and lack of robustness in LLMs when processing low-resource indigenous languages. While much of the current AI Safety literature focuses on toxicity, harmful content, and jailbreak attacks in high-resource languages, comparatively little attention has been devoted to evaluating how language models behave when confronted with underrepresented linguistic communities.

To address this gap, we introduce an empirical evaluation framework based on a bilingual Bolivian Quechua–Spanish dictionary used as ground truth. Rather than focusing solely on direct translation accuracy, our approach evaluates whether model-generated definitions preserve the semantic meaning captured in the reference dictionary. By combining semantic similarity metrics, embedding-based analysis, and human evaluation, we assess the extent to which modern LLMs can accurately represent lexical knowledge in Bolivian Quechua.

Main Contributions

The main contributions of this work are:

- **A Ground-Truth-Based Evaluation Framework:** We design a reproducible methodology that leverages a curated Bolivian Quechua lexical resource as a reference dataset for evaluating semantic understanding in LLMs.
- **Comparative Evaluation of Commercial Language Models:** We benchmark multiple state-of-the-art language models, including Gemini and Claude variants, to analyze their ability to capture the meaning of vocabulary entries from an underrepresented indigenous language.
- **A Semantic Validation Pipeline:** We develop an automated evaluation pipeline that combines dictionary-based ground truth, embedding similarity metrics, and human judgment to assess adequacy, completeness, fluency, and semantic alignment.
- **Insights into Low-Resource Language Processing:** We provide empirical evidence regarding the strengths and limitations of current LLMs when operating in low-resource linguistic settings, contributing to ongoing discussions in AI evaluation, safety, and inclusion.

2. Related Work

Previous research has shown that Large Language Models (LLMs) often struggle when working with indigenous and low-resource languages due to the limited availability of training data. A notable example is *AmericasNLI* (Ebrahimi et al., 2021), a benchmark designed to evaluate multilingual language models across ten indigenous languages of the Americas. The study found that model performance on these languages was significantly lower than on high-resource languages, highlighting the challenges faced by underrepresented linguistic communities.

More recent studies have explored methods for improving natural language processing in indigenous languages. For example, Dhawan et al. investigated the use of synthetic data and language-specific preprocessing techniques for Quechua–Spanish and Guaraní–Spanish language tasks. Their findings demonstrated that additional linguistic resources and targeted preprocessing strategies can significantly improve performance in low-resource settings.

While these studies provide valuable insights into multilingual NLP and indigenous language processing, most focus on model training, adaptation, or machine translation performance. In contrast, our work focuses specifically on Bolivian Quechua and evaluates the semantic understanding capabilities of existing LLMs using a bilingual Quechua–Spanish dictionary as ground truth. Rather than proposing a new language model or translation system, we investigate how accurately modern LLMs capture the meaning of lexical entries through semantic similarity analysis and human evaluation.

When and Why Would Someone Use Our Method?

Most existing approaches focus on training new models, improving translation systems, or evaluating multiple languages simultaneously. While these methods are valuable, they often rely on general-purpose benchmarks and may overlook the linguistic characteristics of specific indigenous language variants such as Bolivian Quechua.

Our methodology is particularly useful when researchers or practitioners need to evaluate the semantic reliability of LLMs in a specific low-resource language using a trusted local reference source. By leveraging a curated dictionary and combining embedding-based similarity metrics with human judgment, our framework provides a more targeted and interpretable assessment of model behavior than traditional translation benchmarks alone.

What New Information Does Our Method Provide?

Previous studies have demonstrated that language models face challenges when processing low-resource languages. However, there is limited empirical evidence regarding how well modern LLMs understand lexical meaning in Bolivian Quechua.

Our work contributes a concrete evaluation framework that combines dictionary-based ground truth, semantic similarity analysis, and human evaluation. This approach enables the identification of semantic inconsistencies, lexical misunderstandings, and limitations in model-generated definitions. Furthermore, it provides insights that can support future research on indigenous language inclusion, AI evaluation methodologies, and the development of more reliable language technologies for underrepresented linguistic communities.

3. Methods

Methodology and Workflow

Our approach consists of an evaluation pipeline designed to measure how effectively Large Language Models (LLMs) capture the semantic meaning of vocabulary entries from Bolivian Quechua. Rather than focusing solely on translation accuracy, the framework evaluates whether model-generated definitions preserve the meaning represented in a bilingual Quechua–Spanish dictionary used as ground truth.

The workflow is composed of the following stages:

1. **Ground Truth Construction:** We extracted and structured lexical entries from a Bolivian Quechua–Spanish dictionary into a machine-readable format. This resource served as the reference dataset against which model-generated definitions were evaluated.
2. **Definition Generation:** For each lexical entry, the evaluated LLMs were prompted to generate definitions or explanations of the corresponding terms. Responses were collected and stored for subsequent analysis.
3. **Semantic Similarity Analysis:** Generated definitions were transformed into embeddings and compared against the corresponding dictionary definitions using semantic similarity metrics. This process provided an automated estimate of semantic alignment between the model output and the ground-truth definition.

4. **Human Evaluation:** In addition to automated similarity analysis, human evaluators assessed the generated definitions according to three criteria: adequacy, completeness, and fluency. This allowed us to compare automated metrics with human judgment.
5. **Comparative Analysis:** Results from semantic similarity metrics and human evaluation were aggregated and compared across multiple models to identify strengths and weaknesses in their understanding of Bolivian Quechua vocabulary.

Models, Datasets, and Tools

- **Ground Truth Dataset:** A structured bilingual Quechua–Spanish lexical resource containing dictionary definitions used as the reference source throughout the evaluation process.
- **Evaluated Models:** We evaluated multiple commercial language models, including Gemini Flash, Gemini Flash Lite, and Claude Haiku, through their respective APIs.
- **Tools and Environment:** The evaluation framework was implemented in Python, using standard data-processing libraries, embedding generation tools, and automated scripts for similarity analysis and result aggregation.

Design Decisions and Justification

- **Dictionary-Based Ground Truth:** We selected a curated bilingual dictionary as the reference source in order to provide a consistent and reproducible basis for evaluation.
- **Semantic Evaluation Beyond Exact Matching:** Rather than relying solely on lexical overlap, we employed embedding-based similarity measures to capture semantic relationships between generated and reference definitions.
- **Human-in-the-Loop Validation:** Because semantic understanding cannot be fully captured through automated metrics alone, human evaluators were included to assess adequacy, completeness, and fluency.

Lessons Learned

- **Automated Metrics Alone Are Insufficient:** Embedding similarity provides useful information regarding semantic alignment, but it does not always capture nuances identified by human evaluators.
- **Challenges of Low-Resource Languages:** Even advanced LLMs exhibit difficulties when working with underrepresented languages, particularly when definitions contain culturally specific or regionally dependent concepts.
- **Importance of Human Evaluation:** Human judgment remains a critical component when assessing semantic understanding in indigenous and low-resource languages, where linguistic variation and contextual meaning play a significant role.

.

4. Results

Overview of Model Performance

This section presents the results obtained from evaluating multiple Large Language Models on Bolivian Quechua lexical definitions. The evaluation combines embedding-based semantic similarity metrics with human judgments in order to assess how accurately each model captures the meaning represented in the ground-truth dictionary.

Quantitative Results

The evaluated models exhibited notable differences in their ability to produce semantically aligned definitions. Among the tested systems, Gemini Flash Lite achieved the highest agreement with the reference definitions, followed by Gemini Flash and Claude Haiku.

Table 1. Relative Model Performance

Model	Relative Agreement (%)
Gemini Flash Lite	~32
Gemini Flash	~25
Claude Haiku	~10

These results indicate that even state-of-the-art language models experience significant challenges when processing lexical knowledge from an underrepresented indigenous language such as Bolivian Quechua.

Semantic Similarity and Human Evaluation

To validate the effectiveness of the automated evaluation procedure, embedding-based similarity scores were compared against human judgments. Human evaluators assessed generated definitions according to adequacy, completeness, and fluency.

The comparison revealed a moderate positive correlation between automated similarity measurements and human evaluation, with a Spearman correlation coefficient of approximately 0.48. This result suggests that semantic similarity metrics capture meaningful aspects of lexical understanding, although they do not fully replace human judgment.

Key Observations

Several important observations emerged from the evaluation:

- Model performance varied substantially across lexical entries, indicating inconsistent understanding of Bolivian Quechua vocabulary.
- Embedding-based similarity metrics provided a useful approximation of semantic quality but failed to capture all nuances identified by human evaluators.
- Even the best-performing models demonstrated limitations when dealing with low-resource indigenous languages.
- The results highlight the importance of combining automated evaluation methods with human assessment when auditing language models in underrepresented linguistic contexts.

5. Discussion and Limitations

Discussion

The results indicate that current Large Language Models still face significant challenges when processing low-resource indigenous languages such as Bolivian Quechua. Although some models demonstrated a moderate ability to generate semantically related definitions, performance remained inconsistent across lexical entries. This suggests that linguistic knowledge acquired during pretraining is often incomplete when applied to underrepresented languages.

The comparison between embedding-based similarity metrics and human evaluation also revealed important insights. While semantic similarity scores provide a useful approximation of lexical understanding, they do not fully capture all aspects of meaning identified by human evaluators. The observed correlation suggests that automated evaluation can support large-scale auditing efforts, but human judgment remains essential when assessing semantic quality in indigenous language settings.

Limitations

Several limitations should be considered when interpreting the results. First, the evaluation relies on a single bilingual Bolivian Quechua–Spanish dictionary as the reference source. Although this resource provides a consistent ground truth, it cannot fully represent the linguistic diversity of Quechua spoken across different regions and communities.

Additionally, the study focuses on lexical definitions rather than broader language understanding tasks such as dialogue generation, reasoning, or long-form translation. Therefore, the findings should not be interpreted as a complete assessment of overall model capabilities in Bolivian Quechua.

Another important limitation concerns linguistic variation. Different Quechua-speaking regions may use alternative words, expressions, or meanings that are not captured by the selected dictionary. As a result, some model outputs may be semantically valid despite differing from the reference definitions.

Finally, due to the time constraints of a hackathon environment, the evaluation was conducted on a limited set of lexical entries and human assessments. Future studies with larger datasets and additional evaluators would provide more robust conclusions.

Future Work

This work represents an initial exploration of semantic evaluation for Bolivian Quechua using dictionary-based ground truth and human judgment. Several opportunities exist to extend and strengthen this research.

Future work could expand the dataset by incorporating additional lexical resources, idiomatic expressions, and regional variants of Quechua. Such an expansion would allow for a more comprehensive assessment of model behavior across diverse linguistic contexts.

Another promising direction involves evaluating additional language models, including open-source multilingual systems, to better understand how different architectures perform in low-resource language settings. Future studies could also explore alternative semantic evaluation techniques and stronger embedding models to improve the alignment between automated metrics and human judgment.

Finally, the proposed framework could serve as the foundation for standardized benchmarks targeting indigenous and underrepresented languages. Such benchmarks would contribute to more inclusive AI evaluation practices and support the development of language technologies that better serve diverse linguistic communities.

6. Conclusion

This work presented TriloByte, a framework for evaluating the semantic understanding capabilities of Large Language Models (LLMs) in Bolivian Quechua, an indigenous language that remains significantly underrepresented in modern AI training datasets. By leveraging a bilingual Quechua–Spanish dictionary as ground truth and combining embedding-based similarity analysis with human evaluation, we developed a reproducible methodology for assessing how effectively LLMs capture lexical meaning in a low-resource language setting.

Our findings suggest that, while modern language models demonstrate some ability to generate semantically related definitions, their performance remains limited and inconsistent when dealing with underrepresented linguistic communities. Furthermore, the observed differences between automated similarity metrics and human judgment highlight the importance of incorporating both approaches when evaluating semantic understanding. Beyond the specific models analyzed, this work emphasizes the need for evaluation frameworks that account for linguistic diversity and support the development of more inclusive, reliable, and culturally aware AI systems.

Code and Data

Code repository: https://github.com/Maxbm19/TriloByte

The repository contains the source code, evaluation scripts, dataset files, and materials used to reproduce the TriloByte evaluation pipeline.

Main Contributions

1. We evaluate the performance of modern Large Language Models (LLMs) on Bolivian Quechua, an underrepresented indigenous language with limited representation in contemporary AI training datasets.
2. We introduce a ground-truth-based evaluation methodology that leverages a bilingual Quechua–Spanish dictionary to assess semantic understanding through dictionary definitions rather than relying solely on translation tasks.
3. We develop a reproducible evaluation pipeline that combines embedding-based semantic similarity metrics with human evaluation, enabling the assessment of adequacy, completeness, fluency, and semantic alignment of model-generated definitions.
4. We provide empirical evidence regarding the capabilities and limitations of current LLMs in low-resource language settings, highlighting challenges associated with evaluating semantic understanding in indigenous languages and contributing to ongoing efforts in AI evaluation, safety, and linguistic inclusion.

Appendix C. Evaluation Criteria

Definitions generated by the models were evaluated using both automated and human-centered criteria.

Automated Evaluation

- **Semantic Similarity: Similarity score between the generated definition and the dictionary definition using embedding representations.**

Human Evaluation

- **Adequacy: Degree to which the generated definition captures the intended meaning of the lexical entry.**
- **Completeness: Extent to which relevant information contained in the reference definition is preserved.**
- **Fluency: Grammatical correctness and readability of the generated definition.**

Appendix D. Additional Information

Additional implementation details, dataset statistics, embedding configurations, evaluation parameters, and extended examples are available in the project repository.

Repository:

https://github.com/Maxbm19/TriloByte

LLM Usage Statement

During the development of this project, several artificial intelligence tools were used to support programming, debugging, research, brainstorming, and technical writing tasks. These tools included GPT (OpenAI), Gemini (Google), Claude (Anthropic), and Cursor AI as a programming assistant.

GitHub was also used for version control, source code management, and collaboration among team members during the development of the TriloByte project.

AI tools were employed as development aids; however, all methodological decisions, experimental configurations, result analyses, and final conclusions were reviewed and validated by the project team. No findings were included in the final report without undergoing the evaluation and validation procedures described in the methodology. Additionally, Gemini and Claude models were part of the object of study of this research and were evaluated using the proposed framework to analyze their ability to generate semantically aligned definitions for Bolivian Quechua lexical entries. Their outputs were assessed through a combination of embedding-based similarity metrics and human evaluation.