Welo Data Research Reveals Multilingual Gaps in LLM Safety Guardrails, Exposing *Global Risk*

(Doha, Qatar) - Welo Data, a division of Welocalize and the leader in high-quality multilingual AI training data, has released new research showing that safety protections in leading large language models (LLMs) do not reliably transfer across languages. The study evaluated 10 leading models using 210,000 model-prompt pairs across 79 languages and safety categories such as *Hate and Discrimination*, *Self-harm and Suicide*, *Violence and Threats*, and *Misinformation and Disinformation.* Unlike most existing safety evaluations, which are conducted primarily in English, this study measured cross-lingual safety gaps, exposing multilingual safety as a major challenge. The results reveal a consistent pattern: safety alignment is strongest in English but deteriorates across models in low-resource, or digitally under-represented, languages.

Even in areas where guardrails are effective in English, like *Hate and Discrimination* and *Self-harm and Suicide*, the same prompts in low-resource languages can “jailbreak” those protections. A safe response in English often becomes an unsafe response in another language, exposing how safety reduces as linguistic representation declines. This erosion exposes a global security gap: even strong English guardrails can be bypassed simply by translating harmful prompts into another language using readily available translation tools. It also creates a safety blind spot for populations that primarily interact with LLMs in non-English languages, leaving them with weaker protection against harmful content. For global developers and organizations deploying LLMs across regions, this is not just a technical limitation; it’s also a safety liability.

"Our research shows that LLM policy enforcement is still strongest in English, where most training data, policy design, and red-teaming are concentrated,” said Dr. Fernando Migone, VP of Research and Innovation at Welo Data and Web Summit Qatar presenter. “The challenge we see in the data is that once harmful prompts are translated into low-resource languages, responses that were safe in English often become unsafe, effectively turning translation into a jailbreak vector. This breakdown points to gaps in multilingual safety data and evaluation, which is why multilingual safety must be treated as core infrastructure for responsible global deployment rather than an optional localization step."

Key observations from the study include:

- **Models vary in their safety behavior in English.** Safety performance in English varies by model, indicating that alignment mechanisms operate differently across harm domains. In our sample, the categories of *Hate and Discrimination* and *Self-harm and Suicide* are the most consistently protected across models.
- **Prompting in low resource languages can increase the likelihood of a harmful response by 4-5X.** Even the categories with the strongest guardrails in English degrade sharply in low-resource languages. For instance, unsafe response rates in the *Hate and Discrimination* category climbed from below 10% in English to 40–50% in several low-resource languages.
- **Models vary in their multilingual safety.** Some models maintain relatively consistent safety performance from English to low-resource languages, while

others show sharper degradation. These differences suggest that safety mechanisms do not generalize equally well across models.
- **The gap is systematic, not incidental.** Safety degradation correlates with language family. Languages from the Niger-Congo and Nilo-Saharan families exhibit the greatest increases in unsafe completions.

Dr. Migone will present at Web Summit Qatar. To view the full research report, visit welodata.ai.

About Welo Data
Welo Data, a division of Welocalize, stands at the forefront of the AI training data industry, delivering exceptional data quality and security. Supported by a global network of over 500,000 AI training professionals and domain experts, along with cutting-edge technological infrastructure, Welo Data fulfills the growing demand for dependable training data across diverse AI applications. Its service offerings span a variety of critical areas, including data annotation and labeling, large language model (LLM) enhancement, data collection and generation, and relevance and intent assessment. Welo Data's technical expertise ensures that datasets are not only accurate but also culturally aligned, tackling significant AI development challenges like minimizing model bias and improving inclusivity. Its NIMO (Network Identity Management and Operations) framework guarantees the highest level of accuracy and quality in AI training data by leveraging advanced workforce assurance methods. *welodata.ai*

About Welocalize, Inc.
Welocalize is a leading language service provider. It bridges language and AI to power global success in complex and regulated environments. Welocalize's expertise in translation, localization, and AI training data ensures precise, scalable, and compliant solutions for businesses worldwide. Whether high-quality multilingual content for regulated industries or structured datasets to train AI models, Welocalize helps its clients communicate, innovate, and accelerate their businesses globally in over 300 languages. Its investment in AI innovation has produced patented, award-winning products that deliver multilingual content with unmatched efficiency, speed, scale, and accuracy. These solutions are delivered within a framework of data security, compliance, and safety, underpinned by its 7 ISO certifications. Welocalize is headquartered in New York with offices all over the globe. *welocalize.com*