Explainability and Interpretability of Multilingual LLMs: A Survey

Lucas Resck [1] Isabelle Augenstein [2] Anna Korhonen [1]

[1]Language Technology Lab, University of Cambridge [2]University of Copenhagen

Motivation

- Multilingual LLMs (MLLMs) demonstrate state-of-the-art capabilities across diverse cross-lingual and multilingual tasks.
- Their complex internal mechanisms, however, often lack transparency.

★ *What are the **unique challenges** for multilingual explainability?*
★ *What is the **literature landscape**?*
★ *What are **future directions**?*

Contributions

- We **survey** the current explainability and interpretability methods for MLLMs.
- 220+ papers.
- To our knowledge, it is the **first comprehensive review** of its kind.

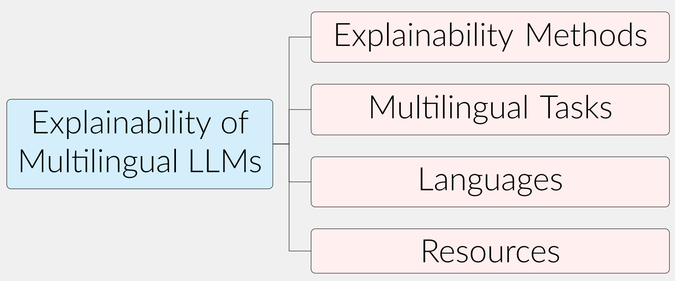

Figure 1. Structure of the survey.

Explainability Methods for Multilinguality

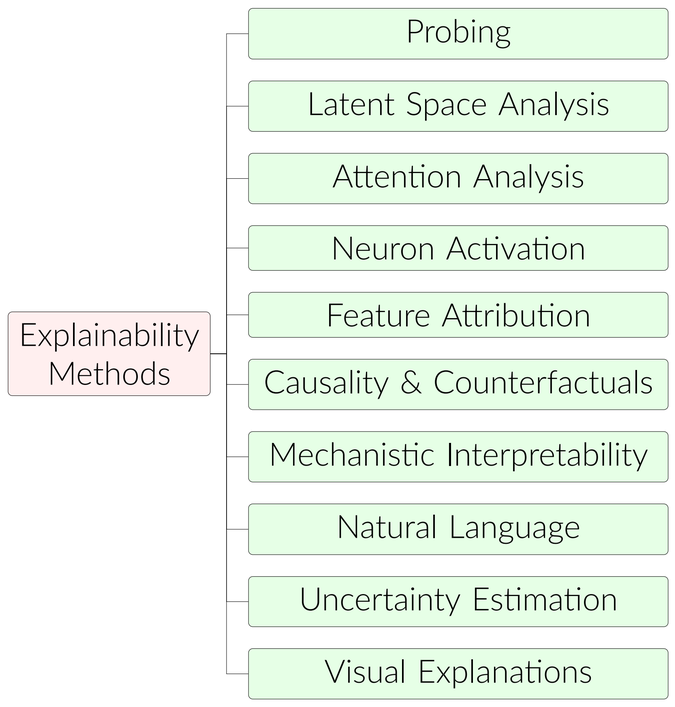

Figure 2. Overview of explainability methods.

- Future work may explore neuron activation analysis to ***enhance low-resource language performance*** by capitalising on high-resource language patterns.
- ***Explanation faithfulness*** *is a critical challenge* for feature attribution methods.
- Causal analysis ***shifts from correlation to causation***, with more robust conclusions.

Explainability of Multilingual Tasks

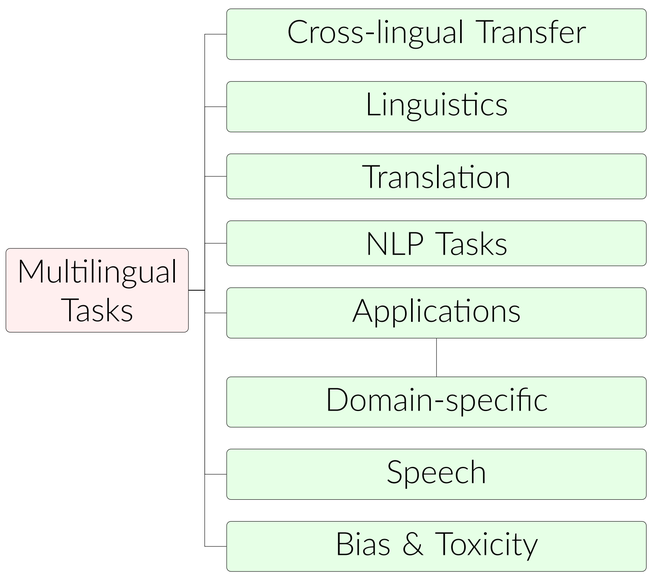

Figure 3. Overview of multilingual tasks.

- *Extending core NLP to low-resource contexts* warrants further research.

Languages

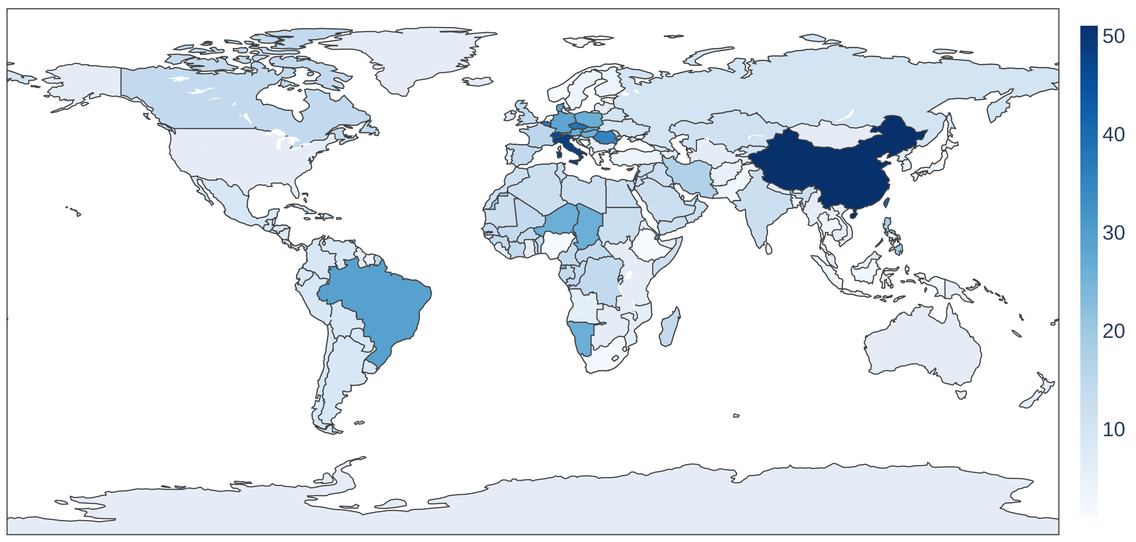

Figure 4. Global distribution of research on non-English language interpretability. Languages are mapped to countries according to their official, de facto, regional, minority or national status.

- High-mid languages are mostly used as case studies and multilingual test datasets for any task that is "non-English".
- **Low-resource languages are underrepresented in the literature**, with a focus on simple applications of existing techniques and less impactful venues.

Resources

Resources		Aid interpretation extraction	Evaluate explanations
Evaluation	Benchmarks		Attanasio et al. (2022); Park and Padó (2024)
	Datasets	Zeng et al. (2024); Barriere and Cifuentes (2024b,a); Zhang et al. (2024)	Jørgensen et al. (2022)
	Human evaluation	Serikov et al. (2022)	Brandl et al. (2024); Kozlova et al. (2024); Zarharan et al. (2025)
Explanation techniques	Feature attribution	Jørgensen et al. (2022); Mamta et al. (2023); Tourni and Wijaya (2023); Vasileiou and Eberle (2024); Guo et al. (2024)	Zhao and Aletras (2024)
	Uncertainty	Kang et al. (2024); Cao et al. (2024)	
	Visualisation	Tagarelli and Simeri (2021); Lin et al. (2024)	
	Others	Wang et al. (2024); Grosse et al. (2023)	

Table 1. A summary of resources used for multilingual explainability and interpretability and how they are adopted. Works are selected based on recency and representativeness.

Takeaways

Challenges

What are the unique challenges for multilingual explainability?

1. How models internally process **multilingualism**;
2. The dynamics of **cross-lingual transfer**;
3. The handling of **language-specific** features (e.g., different scripts);
4. The manifestation of language- and culture-specific **biases**;
5. The **scarcity of resources** for low-resource and non-natural languages;
6. ***Extra-language*** multilingual features, such as cultural knowledge.

Core Findings

- Tendency to **apply existing explainability methods** to multilingual settings.
- A significant **skew in the literature** against low-resource languages.

Future Directions

- Development of **multilingual explainability innovations**, rather than merely adapting existing methods to non-English contexts.
- ***Bridging the gap*** *between interpreting inner model behaviour and explaining final model decisions*.
- Interpreting ***extra-language***, external knowledge, such as cultural values, regional variations and factual knowledge.
- **Shifting the research focus** to low-resource languages from NLP applications to core NLP tasks.