Explainability and Interpretability of Multilingual Large Language Models: A Survey

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

[1]Language Technology Lab, University of Cambridge
[2]University of Copenhagen
{ler44, alk23}@cam.ac.uk, augenstein@di.ku.dk

EMNLP 2025, Suzhou, China

Motivation & Research Questions

- Multilingual LLMs (MLLMs) demonstrate state-of-the-art capabilities across diverse cross-lingual and multilingual tasks.
- Their complex internal mechanisms, however, often lack transparency.
- This lack of transparency is a critical issue, especially given the linguistic and cultural diversity of their training data.

- What are the **unique challenges** for multilingual explainability?
- What is the current **literature landscape**?
- What are the most promising **future directions**?

Our Contribution: The First Comprehensive Survey

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

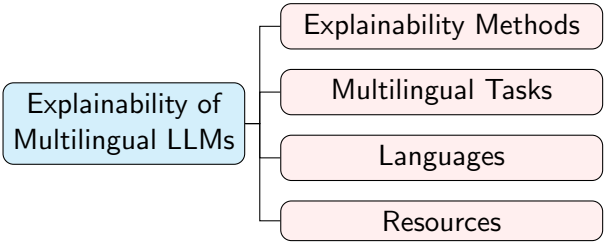
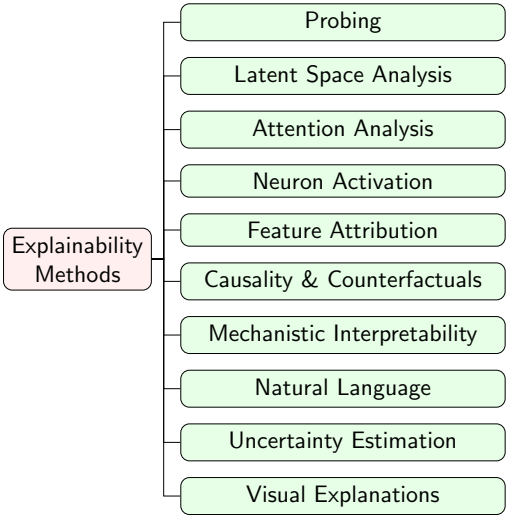


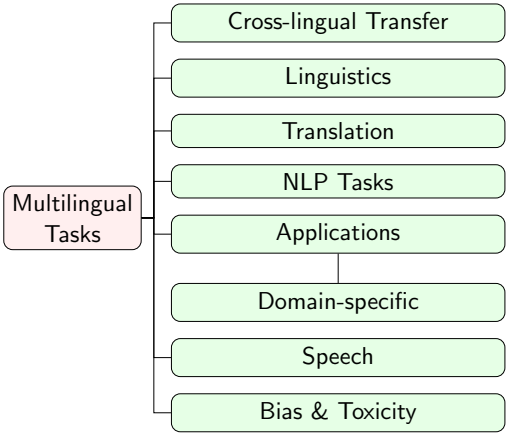
Figure: Structure of the survey.

Landscape I: Explainability Methods for Multilinguality



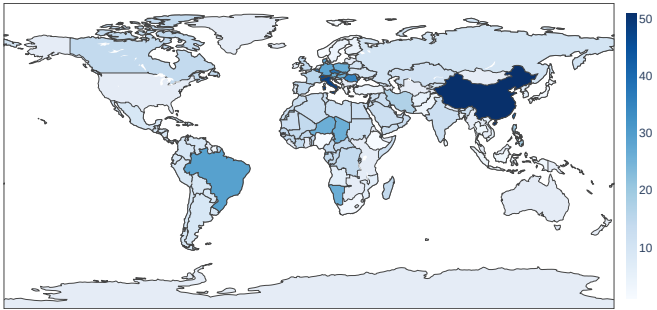
- Neuron activation analysis could help **boost low-resource performance**.
- **Explanation faithfulness** is a critical challenge, especially for feature attribution methods.
- Causal analysis allows a shift from correlation to **causation**, yielding more robust conclusions.

Landscape II: Explainability of Multilingual Tasks



- Low-resource languages are mostly studied in the context of specific applications.
- A significant gap exists in applying explainability to **core NLP tasks** (e.g., parsing, NER) for these languages.
- **Extending core NLP to low-resource contexts** warrants further research.

Landscape III: Languages



- High and mid-resource languages are frequently used as general-purpose, "non-English" test cases.
- **Low-resource languages are severely underrepresented**, with studies focusing on simple applications of existing techniques published in less impactful venues.

Landscape IV: Resources

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

Takeaways

Unique Challenges

1. Processing of **multilingualism**.
2. Dynamics of **cross-lingual transfer**.
3. Handling of **language-specific features**.
4. Language- and culture-specific **biases**.
5. Scarcity of evaluation **resources** for low-resource languages.
6. Interpreting **extra-language** features.

Core Findings

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

Future Directions

- Develop **novel explainability methods designed for multilinguality**.
- **Bridge the gap** between interpreting internal behavior and explaining final decisions.
- Interpret **extra-language features**, such as cultural values, regional variations and factual knowledge.
- **Shift the research focus** for low-resource languages from NLP applications to core, foundational NLP tasks.

Our Contribution: The First Comprehensive Survey

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

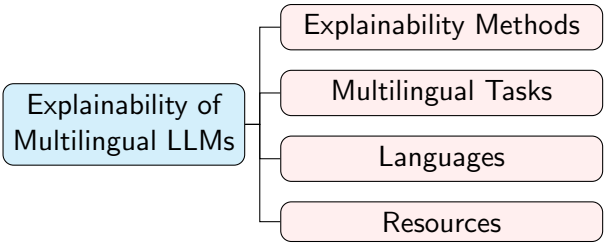


Figure: Structure of the survey.

Explainability and Interpretability of Multilingual Large Language Models: A Survey

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

[1]Language Technology Lab, University of Cambridge
[2]University of Copenhagen
{ler44, alk23}@cam.ac.uk, augenstein@di.ku.dk

EMNLP 2025, Suzhou, China