Explainability and Interpretability of Multilingual Large Language Models: A Survey

Lucas Resck[1] **and Isabelle Augenstein**[2] **and Anna Korhonen**[1]
[1]Language Technology Lab, University of Cambridge
[2]University of Copenhagen
{ler44, alk23}@cam.ac.uk, augenstein@di.ku.dk

Abstract

Multilingual large language models (MLLMs) demonstrate state-of-the-art capabilities across diverse cross-lingual and multilingual tasks. Their complex internal mechanisms, however, often lack transparency, posing significant challenges in elucidating their internal processing of multilingualism, cross-lingual transfer dynamics and handling of language-specific features. This paper addresses this critical gap by presenting a survey of current explainability and interpretability methods specifically for MLLMs. To our knowledge, it is the first comprehensive review of its kind. Existing literature is categorised according to the explainability techniques employed, the multilingual tasks addressed, the languages investigated and available resources. The survey further identifies key challenges, distils core findings and outlines promising avenues for future research within this rapidly evolving domain.

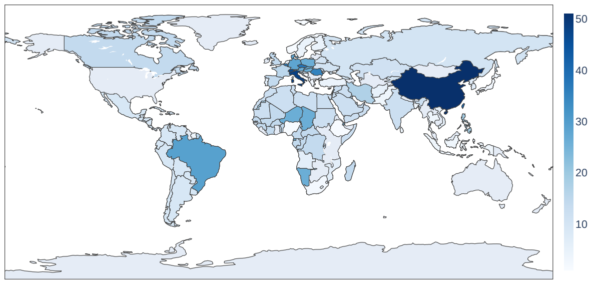

Figure 1: Global distribution of research on non-English language interpretability. Languages are mapped to countries according to their official, de facto, regional, minority or national status. For a detailed analysis of language explainability, refer to Section 4.

1 Introduction

Large language models (LLMs) have markedly advanced the field of natural language processing (NLP), attaining human-comparable, state-of-the-art performance across a multitude of tasks, including those requiring cross-lingual and multilingual capabilities (OpenAI et al., 2024). Despite their impressive capabilities, the opaque "black-box" nature of LLMs presents considerable challenges. Ensuring explainability and interpretability is crucial, and particularly pressing in the case of multilingual LLMs (MLLMs). Due to their training on linguistically and culturally diverse data, often including low-resource languages, MLLMs are particularly susceptible to generating biased or inaccurate outputs across varied linguistic contexts.

While previous surveys have explored explainability methods for various LLMs (Zhao et al., 2024a; Luo and Specia, 2024), they have not focussed on the distinct challenges of multilinguality. These challenges include elucidating how models internally process multiple languages, the dynamics of cross-lingual transfer, the handling of language-specific features (e.g., scripts, word orders, phonemes), the manifestation of language- and culture-specific biases and the scarcity of resources for most world's languages. Conversely, existing reviews of MLLMs have largely overlooked the dimension of interpretability (Qin et al., 2024; Xu et al., 2024). Although Zhu et al. (2024a) touched upon the interpretability of MLLMs, the discussion lacked comprehensive scope.

We report a survey of the state-of-the-art in explainability and interpretability of MLLMs. To the best of our knowledge, this constitutes the first survey dedicated exclusively to this intersection, synthesising research from the dual perspective of explanation methodologies and multilingual applications. As in Figure 2, we categorise existing work according to the explainability methods employed (Section 2), the specific multilingual tasks addressed (Section 3), the languages under investigation and the resources available (Section 4).

Our analysis indicates a tendency for most research to apply existing explainability methods to multilingual contexts, frequently without the req-

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 20454–20486

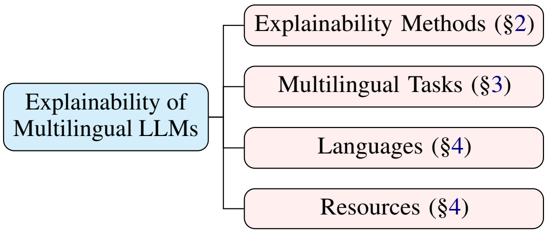

Figure 2: Structure of the survey.

uisite significant methodological innovations. The findings reveal, perhaps unsurprisingly, a considerable skew in the literature against low-resource languages. These languages are often subjected to simplistic applications of off-the-shelf explainability techniques, and impactful, cutting-edge multilingual explainability research for them remains notably scarce.

We envision promising avenues for future exploration in the development of novel multilingual explainability innovations. This includes advancing work on language-specific features and *extra-language* knowledge (e.g., cultural values, regional variations and factual knowledge), furthering the understanding of cross-lingual transfer, bridging the gap between interpretation and explanation and improving the sophistication of explainability research for low-resource languages by shifting the focus from NLP applications to core, more foundational NLP tasks.

2 Explainability Methods for Multilinguality

2.1 Probing

Probing is a common explainability technique that involves training simple classifiers on a model's internal representations to predict its capacity to encode specific properties, often within multilingual contexts (Pires et al., 2019; Vulić et al., 2020). It is widely applied to analyse how multilingual models encode linguistic information (Starace et al., 2023), assess cross-lingual transfer (Vulić et al., 2023) and detect issues such as multilingual gender bias (Steinborn et al., 2022). In the cross-lingual domain, studies probe lexical knowledge in multilingual sentence encoders (Vulić et al., 2023), the dynamics of how models acquire cross-lingual abilities (Blevins et al., 2022) and knowledge transfer from artificial languages with implications for multilingual understanding (Ri and Tsuruoka, 2022).

Linguistic Probing. A substantial body of work employs probing to investigate the linguistic knowledge within MLLMs. Syntactic understanding is a key focus with studies on multilingual linguistic acceptability (Zhang et al., 2024d), syntactic agreement in languages like French (Li et al., 2023a) and the localisation of syntactic information (Li et al., 2022). Semantic probing examines the encoding of predicate-argument structures across languages (Conia and Navigli, 2022), metaphors across languages such as Spanish, Russian and Persian (Aghazadeh et al., 2022) and verbal aspect in Russian in a layer-wise manner (Katinskaia and Yangarber, 2024). Research also explores how models represent general linguistic categories across languages and model layers (Starace et al., 2023) and specific challenges such as Chinese causative-passive homonymy (Xu and Markert, 2022). Please refer to Appendix H for a more complete description of multilingual probing works.

Takeaways. Probing is a straightforward yet potent and widely applicable methodology, frequently employed to analyse encoded linguistic information and assess cross-lingual transfer. A notable trend in recent studies is the prevalence of layer-wise probing to localise where specific information is represented. Despite its broad application, certain multilingual dimensions appear under-represented, offering fertile ground for future investigation. These include the probing of cultural and moral values across diverse languages (Pawar et al., 2024) and the examination of how models encode distinctly language-specific information, such as lexical tone in tonal languages (Shen et al., 2024).

2.2 Latent Space Analysis

Latent space analysis provides a powerful tool for understanding MLLMs, often focusing on cross-lingual representations (Chen et al., 2022; Sun et al., 2024). Wen-Yi and Mimno (2023) report that multilingual input layer embeddings show similarity between token translations, despite no explicit translation objective during training. Icard et al. (2025) analyse French writing style effects on embeddings, while Liang et al. (2020) use interpretable subspaces for multilingual gender bias removal. Wendler et al. (2024) find that models' internal representations of non-English languages become closer to English in intermediate layers. A set of work also explores language-agnostic latent spaces: Zeng et al. (2025) propose a "Lingua

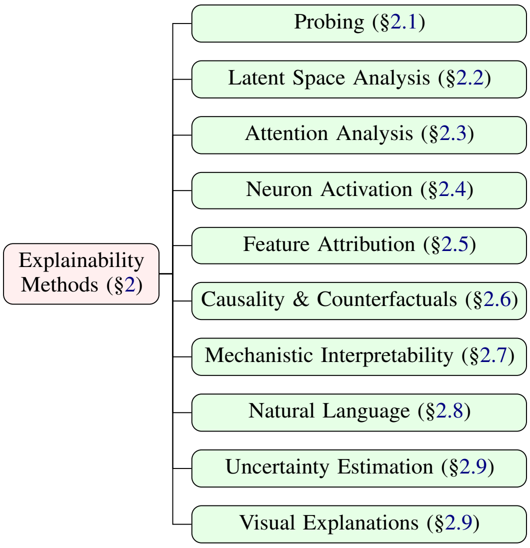

Figure 3: Overview of explainability methods.

Franca", Abdullah et al. (2024) examine language-neutral subspaces in speech translation, Utpala et al. (2024) identify language-agnostic components in code representations and Dumas et al. (2025) play with activation patching to support language-agnostic concept representations. Other analyses cover cross-lingual concept alignment (Xu et al., 2023), training stage effects on multilingual embeddings (Yan et al., 2024; Thanh et al., 2023), similarity to perceptual modalities (shape, sound, colour) (Boldsen et al., 2022) and investigating pruning (Kurz et al., 2024).

Takeaways. Multilingual LLM representations are pivotal for understanding their cross-lingual capabilities, despite their notable complexity. Research increasingly focuses on spontaneous multilingual alignment and identifying language-agnostic latent spaces. Future research may explore intrinsic language-agnostic properties, such as how linguistic information is encoded and evolves across training and layers. Future work could also focus on identifying interpretable subspaces for multilingual attributes beyond bias (e.g., cultural and regional differences) and further examining low-resource language embedding structures.

2.3 Attention Analysis

Transformers revolutionised the field of NLP particularly due to the self-attention mechanisms. While the explainability of attention has been debated (Bibal et al., 2022), it has been widely used to analyse MLLMs, for example, in multilingual bias detection (Liang et al., 2020) and linguistic tasks (Kozlova et al., 2024). For instance, Ma et al. (2021) found that pruning attention heads generally improves cross-lingual task performance, while Voita et al. (2019) noted that specialised heads are last to be pruned in Russian machine translation. Gopinath and Rodriguez (2024) found diverse attention heads in self-supervised speech models irrespective of language, with diagonal heads being key for cross-lingual phoneme classification.

Takeaways. Self-attention is considered intuitive with its visual explanation potential. Research trends cover analysing attention in diverse multilingual tasks and specialised head roles. While the debate on attention's true explanatory power for model behaviour persists, attention analysis is being used to inform model improvements (e.g., jailbreak mitigation). For instance, pruning attention heads improves cross-lingual performance (Ma et al., 2021). Future work could further explore the benefits of pruning, investigate other interventions like attention re-weighting for MLLMs and, crucially, enhance attention explanation faithfulness.

2.4 Neuron Activation

Analysing neuron activation patterns across different languages offers crucial insights into the cross-lingual capabilities and multilingualism of LLMs (Wang et al., 2024c; Zhao et al., 2024c). Studies leveraging neuron activity reveal similar activation patterns for semantically identical inputs across languages (Zeng et al., 2025) and demonstrate the development of consistent cross-lingual representations (Sun et al., 2024). In contrast, other works explore and intervene in language-specific neurons, particularly in early and final layers, to steer the output language (Tang et al., 2024; Kojima et al., 2024). Complementarily, neuron analysis supports hypotheses such as knowledge-free reasoning processes sharing similar neurons cross-lingually while knowledge is stored in a more language-specific manner (Hu et al., 2024). For instance, Mu et al. (2024) found that providing input in multiple parallel languages leads to more precise neuron activation, while other research leverages activation analysis for controlling the syntactic form of the output in machine translation (Patel et al., 2022), guiding model pruning and sparsity techniques (Liu et al., 2024d; Kurz et al., 2024) and tracing factual knowledge (Zhao et al., 2024b).

Takeaways. While some analyses of neuron activation patterns reveal consistent behaviour towards semantically identical inputs across languages, other works highlight the importance of language-specific features. Current research trends centre on leveraging such activations to inform output manipulation, model intervention and data attribution. Whether these patterns can also be utilised to mitigate multilingual bias or trace its sources remains an open question. Furthermore, future investigations may explore the potential of neuron activation analysis to *enhance low-resource language performance* by capitalising on patterns observed in high-resource languages.

2.5 Feature Attribution

Feature attribution methods aim to identify parts of the input, such as important tokens, that most influence a model's predictions in multilingual contexts. Research explores explanation faithfulness across different model types (Zhao and Aletras, 2024), analyses feature interactions in multilingual semantic similarity (Vasileiou and Eberle, 2024) and attributes key neurons for understanding cross-lingual transfer (Wang et al., 2024a). Feature attribution also aids in localising bias via token sense components in Chinese models (Sun and Hewitt, 2023), identifying syntactic information in French (Li et al., 2022) and is supported by datasets with human rationales like those for Austrian German offensive language (Pachinger et al., 2024). These techniques, including widely-used off-the-shelf methods such as LIME, are frequently applied within specific NLP applications, as discussed further in Section 3.5. For conciseness, additional works are detailed in Appendix I.

Takeaways. Widely and flexibly applied in diverse multilingual NLP, including domain-specific applications, feature attribution often utilises off-the-shelf techniques, aided by evaluation resources like human rationale datasets. However, *explanation faithfulness is a critical challenge*: current accessible token-based attributions frequently lack insightful interpretability for complex multilingual tasks and nuanced model workings. Future research must prioritise more sophisticated, genuinely interpretable and *demonstrably faithful methods*, especially techniques inherently designed for the multilingual context – such as understanding cross-lingual transfer – rather than merely adapted.

2.6 Causality and Counterfactuals

Causal and counterfactual analyses interpret MLLMs by modelling systems causally (Liu et al., 2021; Li et al., 2023c) or intervening on inputs and internal states. Such methods investigate linguistic information like Russian verbal aspect (Katinskaia and Yangarber, 2024) or French syntactic processing (Li et al., 2023a, 2022). Counterfactual input interventions are used to evaluate nationality bias in diverse languages including Maori and Basque (Barriere and Cifuentes, 2024b,a) or for German retrieval augmented generation (RAG) attribution (Roy et al., 2024). Furthermore Srinivasan et al. (2023) counterfactually probe embeddings to change language prediction and Mueller et al. (2022) intervene on neuron activations with counterfactual perturbations to study multilingual syntactic agreement.

Takeaways. Causal analysis primarily centres on input, neuron and representation interventions, often within linguistic contexts. A key benefit over other methods is its *shift from correlation to causation*, enabling more robust conclusions. A notable trend involves using counterfactuals to study bias, especially in low-resource languages. Future work could expand the analyses to other multilingual aspects, like cultural biases or cross-lingual transfer.

2.7 Mechanistic Interpretability

Mechanistic interpretability aims to uncover MLLM internal workings, frequently via circuit analysis applied to tasks like Spanish sequence continuation (Lan et al., 2024), German n-gram processing (Quirke et al., 2023) or understanding how shared circuits and language-specific components handle syntax across languages like English and Chinese (Zhang et al., 2024a). For example Ferrando and Costa-jussà (2024) studied a subject-verb agreement circuit in English and Spanish, finding a language-agnostic residual stream direction with causal effects on predictions. Ferrando and Voita (2024) introduced a more efficient circuit uncovering method, revealing that important attention heads often specialise for English versus non-English tasks. Other mechanistic approaches, using tools like causal intervention or dictionary learning, examine multilingual alignment performance (Zhang et al., 2024b), language bias and cross-lingual toxicity effects (Hinck et al., 2024; Li et al., 2024c), Arabic synthetic data effectiveness (Boughorbel et al., 2024) and internal information flow in medical LLMs (Zheng et al., 2024).

Takeaways. Circuit analysis offers a promising approach to concretely revealing the internal workings of MLLMs. Identifying circuits is, however, computationally and labour-intensive, and many studies present specific case studies over broadly generalisable findings. Future research should thus aim to uncover more general mechanisms of cross-lingual transfer and further explore the potential of language-agnostic latent spaces (Ferrando and Costa-jussà, 2024).

2.8 Natural Language Explanations

Natural language explanations (NLEs) offer interpretable free-text model insights, often generated by LLMs using Chain-of-Thought (CoT) prompting (e.g. for explainable machine translation evaluation; Lu et al., 2024) or post-hoc justifications (e.g. for Persian stance detection or Korean SMS phishing; Lee and Han, 2024; Zarharan et al., 2025). Prompting strategies are crucial for eliciting NLEs, for instance in Chinese legal judgement prediction (Jiang and Yang, 2023) and multilingual, multicultural norm discovery (Fung et al., 2022). The evaluation of NLEs for metrics like plausibility and faithfulness is also key, as explored in multilingual text classification (English, Danish, Italian) (Brandl and Eberle, 2024). For conciseness, refer to Appendix J for a more complete list of papers.

Takeaways. NLEs are easily obtainable via model prompting, requiring no additional specialised techniques, and are highly interpretable to end-users. Current research, however, predominantly focuses on multilingual applications, with less emphasis on methodological advancements. Key areas for improvement therefore include evaluating the cross-lingual consistency of NLEs and enhancing their faithfulness, particularly for post-hoc generated explanations, potentially through refined prompting strategies.

2.9 Additional Methods

In addition to the explainability methods presented, we also discuss uncertainty estimation and visual explanations (Appendices B and C). Key takeaways include that most uncertainty and visualisation methods are adaptations of existing monolingual techniques to multilingual contexts, and the potential for multilingual overconfidence mitigation.

2.10 Takeaways from Explainability Methods

The application of diverse explanation methods to MLLMs reveals several key trends and challenges. Research on low-resource languages remains underrepresented, counterfactual analysis being a notable exception. Moreover, studies often apply existing techniques to multilingual tasks rather than developing dedicated multilingual explanation methodologies, apart from limited work on probing, latent space and neuron activation analyses. *The correlation versus causation debate is crucial*: causality and mechanistic methods aim for causal insights, unlike feature attribution, attention and probing, which may offer less robust correlational findings. A pressing need exists to expand beyond linguistic case studies and cross-lingual transfer to areas such as cultural values, regional variations and language bias, and to generalise mechanistic findings. The exploration of language-agnostic latent spaces, using representational and mechanistic approaches (including non-natural languages), is a promising trend. Finally, potential lies in integrating diverse explanation types (e.g., causal-mechanistic, visual-attention), guiding model improvements especially for low-resource languages and enhancing multilingual explanation faithfulness.

3 Explainability of Multilingual Tasks

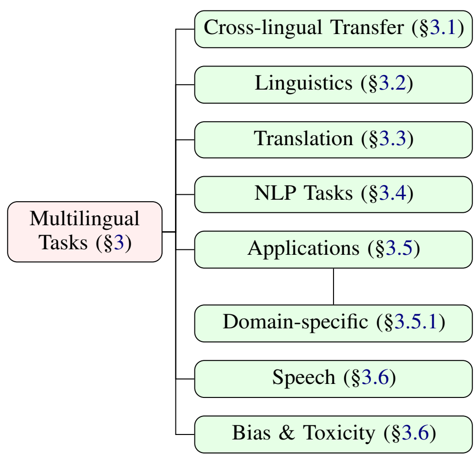

Figure 4: Overview of multilingual tasks.

3.1 Cross-lingual Transfer

Cross-lingual transfer, the process of transferring knowledge between languages, is a key multilingual task. Wang et al. (2024a) use probing and neuron attribution to demonstrate a high correlation between cross-lingual neuron overlap and transfer performance. Other work investigates explicit structural concept alignment (Xu et al., 2023) or uses mechanistic interpretability for spontaneous multilingual alignment, e.g. with unseen languages or question-translated data (Zhang et al., 2024b).

Probing techniques are widely applied to study cross-lingual transfer (Pires et al., 2019), examining aspects such as lexical knowledge in sentence encoders (Vulić et al., 2023), acquisition timing of cross-lingual abilities (Blevins et al., 2022), performance gaps between language resource levels (Li et al., 2024a) and reasoning transfer (Chen et al., 2023). Notably Ri and Tsuruoka (2022) use artificial language pre-training to probe knowledge transfer to natural languages, linking it to encoded contextual information.

Srinivasan et al. (2023) employ causal analysis to project embeddings, demonstrating a language-agnostic component allowing embeddings to be pushed towards another language. This concept of a language-neutral latent space (see also Section 2.2) is explored by other studies too (Zeng et al., 2025; Abdullah et al., 2024).

Shared multilingual properties are further explored via neuron activation similarities (Liu et al., 2024d; Wang et al., 2024c) and representation analyses (Chen et al., 2022): e.g. embedding similarity of token translations (Wen-Yi and Mimno, 2023), cross-language consistency (Sun et al., 2024), links to perceptual modalities (Boldsen et al., 2022) or how similarity impacts knowledge versus reasoning transfer (Hu et al., 2024). Transfer is also assessed with influence functions (Grosse et al., 2023) and circuit analysis (Ferrando and Voita, 2024), while uncertainty estimation (Xu et al., 2021) and attention pruning (Ma et al., 2021) aim performance.

Takeaways. Studies of cross-lingual transfer primarily analyse representation similarity, neuron activation overlap and linguistic feature probing, yielding valuable insights and enhancing model performance. A significant trend is the examination of language-neutral latent spaces theorised to facilitate inter-language knowledge sharing. Artificial languages offer a promising testbed, with recent work linking transfer efficacy to encoded contextual information. However, current findings reveal divergent neuron activation for knowledge-intensive tasks, indicating scope for refinement. Future research should further explore temporal and layer-specific transfer dynamics, alongside the transference of specific knowledge types (e.g., factual, common sense, cultural, logical, biases), particularly within low-resource contexts.

3.2 Linguistic Analysis

Linguistic analysis explainability aims to clarify how multilingual LLMs represent linguistic features. Beyond the prevalent probing methods (see Section 2.1) other approaches include mechanistic interpretability for syntactic circuits (e.g. in Spanish or for English and Chinese; Ferrando and Costa-jussà, 2024; Zhang et al., 2024a), attention analysis for Russian anaphora resolution (Kozlova et al., 2024), feature analysis for French text linguistic features (Rahman et al., 2023) and causal analysis of syntactic agreement (Mueller et al., 2022). Nonetheless the majority of linguistic studies use probing to investigate aspects such as multilingual syntax-related information (Zhang et al., 2024d) or morphosyntactic features across many languages (Serikov et al., 2022). A comprehensive discussion of linguistic probing is available in Section 2.1.

Takeaways. Building on the takeaways in Section 2.1, probing, especially layer-wise analysis, is the predominant method in linguistic analysis, with mechanistic interpretability emerging as a significant trend. Current research often addresses general cross-linguistic features, yet language-specific investigations, particularly in low-resource contexts, offer considerable potential. The exploration of "language-agnostic" features intrinsic to natural language also presents a promising avenue for future research.

3.3 Machine Translation

Machine translation (MT) is a key area of multilingual NLP where various explainability methods enhance system understanding. For instance, approaches include tracking source and target prefix attribution (Ferrando et al., 2022), developing MT methods with intrinsically linked source and target tokens for greater explainability (Stahlberg et al., 2018) and applying neuron analysis to understand how translation models process sentence structure (Patel et al., 2022).

Interpretable evaluation of MT is a relevant research focus, including human-like evaluation with NLEs (Lu et al., 2024) and more interpretable metrics correlating with human judgements (Shafayat et al., 2024). For the task of explainable quality estimation (QE) of MT, works explore interpretable multi-metric frameworks (Park and Padó, 2024), uncertainty quantification fusion (Wang et al., 2021), CoT prompting for better token alignment (Yang et al., 2023a) and token-level relevance, sometimes word-level explainers (Tao et al., 2022; Treviso et al., 2021; Kabir and Carpuat, 2021).

Specific Explainable AI (XAI) techniques also provide insights: integrated gradients are used for low-resource language MT (e.g. South Asian, African) (Islam et al., 2024; Malinga et al., 2024) and methods like SHAP and BERTViz are employed for language pairs such as Luganda-English (Kobusingye et al., 2023).

Takeaways. Various explainability methods are employed to understand MT internal mechanisms (e.g., neuron, attention analysis) and elucidate translation outputs (e.g., feature attribution). A key research gap is the synthesis of these methods for explanations faithful to model processes, such as faithful NLEs. Within the growing field of QE, a distinction is needed: some works develop inherently interpretable metrics, while others apply XAI tools for evaluation, sometimes causing confusion. For low-resource languages, studies often utilise simpler explanation techniques (e.g., SHAP), indicating a need for more profound exploration.

3.4 NLP Tasks

Explainability methods are applied to a range of core NLP tasks beyond machine translation. In NER, for instance, approaches include uncertainty quantification for cross-lingual settings (Hashimoto et al., 2024) and subword impact analysis on multilingual bias (Calix et al., 2022). For other specific tasks, Radman et al. (2023) employ feature attribution with gradients for Arabic singular-to-plural noun conversion, while Lu et al. (2022) develop interpretable first-order logic rules for multilingual short-text entity linking. Multilingual explainability also extends to mechanistic interpretability of sequence continuation in Spanish (Lan et al., 2024), probing internal representations for Chinese NLI tasks (Xu and Markert, 2022), investigating semantic text similarity through feature interactions (Vasileiou and Eberle, 2024) and localising knowledge to attribute language-agnostic neurons (Cao et al., 2024). Furthermore, established techniques like LIME and NLEs are utilised across diverse tasks, including Chinese sentence pair matching (Guo et al., 2024), Italian acceptability judgements (Buonaiuto et al., 2024), multilingual text classification (Brandl and Eberle, 2024) and Portuguese sentence similarity (Rodrigues and Marcacini, 2022).

Takeaways. Diverse explainability methods are applied to the understanding of various NLP tasks (e.g., NER, NLI) in multilingual contexts, including language-specific and cross-lingual settings. These traditional NLP tasks, however, predominantly focus on high- to mid-resource languages; applications for low-resource languages are mainly concentrated in specific domains or proper applications (see Section 3.5). Future research should explore core NLP tasks in low-resource languages to enhance their applicability and performance.

3.5 NLP Applications

Explainability methods are widely applied to multilingual NLP applications to clarify model behaviour. These include stance detection with NLEs in Persian (Zarharan et al., 2025), multilingual fact-checking using referenced explanations (Zeng et al., 2024) or natural logic justifications (Strong et al., 2024) and question answering (QA), such as Japanese multi-hop QA with derivation triples (Ishii et al., 2024) or mechanistic analyses of language bias in multimodal QA (Hinck et al., 2024).

Feature Attribution in NLP Applications. Feature attribution methods (see Section 2.5) specifically explain model predictions within NLP applications. Examples include multilingual QA using eye-tracking to compare human gaze with human-annotated rationales (Brandl et al., 2024), sparse retrieval for Chinese QA (Zhao et al., 2021) and elucidating hate speech detection models in low-resource languages like Urdu and Sindhi using LIME (Siddiqui et al., 2024), For conciseness, refer to Appendix I for a more detailed discussion.

Takeaways. Multilingual NLP applications are the subject of numerous studies, demonstrating methodological diversity in explanations and notable trends towards QA, fact-checking and the development of explainable datasets. Given the widespread use of feature attribution techniques, such research often inherits their limitations (see Section 2.5), including a reliance on off-the-shelf techniques – prevalent in low-resource language studies – and challenges in ensuring the faithfulness of explanations.

3.5.1 Domain-specific Applications

Explainability enhances various domain-specific multilingual applications. In law, NLEs support French legal QA (Louis et al., 2023) and Chinese legal judgements (Jiang and Yang, 2023). Health applications include Chinese medical NLE datasets (Li et al., 2023b) and uncertainty quantification in Korean mental health diagnosis (Kang et al., 2024), while finance sees explainable Chinese stock prediction (Wang et al., 2024b). Understanding societal aspects involves probing multilingual sociodemographic knowledge in LLMs (Lauscher et al., 2022). For conciseness, additional examples in these and other domains are detailed in Appendix I.

Takeaways. In domain-specific multilingual applications, particularly high-stakes areas like finance, law and health, NLEs and accompanying datasets are frequently employed, likely owing to their accessibility for end users; feature attribution techniques are also notably common. While visual explanations hold significant potential, this area remains largely underexplored. A key research challenge, especially for non-English contexts within these critical domains, is ensuring the faithfulness of NLEs. Furthermore, probing multilingual domain-specific knowledge constitutes a relevant and promising direction for future research.

3.6 Additional Tasks

In addition to the multilingual tasks presented, we also discuss speech processing and bias and toxicity (Appendices D and E). Key takeaways include the potential for probing of language-specific speech information and the scarcity of speech data, while there is a need for bias research in diverse multilingual areas.

3.7 Takeaways from Multilingual Tasks

Building upon prior takeaways (Section 2), multilingual tasks present distinct trends and challenges. Probing (notably layer-wise analysis) and mechanistic interpretability are prevalent in linguistic contexts and speech. Feature attribution is common for applications and domain-specific tasks, especially with low-resource languages, while NLEs see increasing adoption in multilingual domain-specific settings; visual explanations show promise but remain underexplored here. Regarding data, NLEs for applications gain prominence and speech resource scarcity persists. Low-resource languages often feature in application studies over core NLP tasks, typically addressed with high and mid-resource languages. *Extending core NLP to low-resource contexts* warrants further research.

A significant need exists to interpret "extra-language" knowledge (domain-specific, cultural, moral, factual, common sense, bias) in multilingual contexts, using probing and cross-lingual transfer; examining cultural and moral knowledge is crucial for human-aligned models. Furthermore, bridging the disparity between MLLM inner-working interpretation versus explanation of model decision is essential. Improving NLE faithfulness is a promising avenue here (Section 2.10). Probing language-specific linguistic features (e.g., dialects, accents) in speech data also holds considerable potential.

4 Languages and Resources

This section outlines our approach to categorising and analysing the languages and resources featured in the surveyed literature. Figure 1 provides an overview of the languages across the surveyed papers. A more comprehensive analysis, presented in Appendix F, categorises languages as high-mid-resource, low-resource and non-natural. Their findings complement our previous observations on the tendency to apply existing explainability methods to multilingual tasks rather than develop proper multilingual methodologies, especially for low-resource languages.

Interpretability resources are essential for the development and application of explainability methods in multilingual contexts and are summarised in Table 1 and explored in detail in Appendix G. Our categorisation divides them into three groups: evaluation resources, techniques and metrics. The discussion highlights prevalent trends, such as how evaluation resources often facilitate interpretation extraction over direct explanation assessment and the common simplistic application of techniques and lack of metrics explicitly designed for MLLMs.

5 Discussion and Future Directions

This section elaborates on the takeaways identified in the previous sections and discusses challenges, core findings, novelty of methods and future directions for multilingual explainability.

Challenges. *What are the unique challenges for multilingual explainability?* The challenges extend beyond the cross-linguality and multilinguality of the models to encompass the specific languages and their available resources, including: (i) how models internally process multilingualism; (ii) the dynamics of cross-lingual transfer; (iii) the handling of language-specific features (e.g., different scripts, word orders, phonemes and intonations);

Resources		Aid interpretation extraction	Evaluate explanations
Evaluation	Benchmarks		Attanasio et al. (2022); Park and Padó (2024)
	Datasets	Zeng et al. (2024); Barriere and Cifuentes (2024a,b); Zhang et al. (2024d)	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. (2024a); 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. Refer to Appendix G for a detailed discussion.

(iv) the manifestation of language- and culture-specific biases; and (v) the scarcity of resources (data and models) for low-resource and non-natural languages. *Extra-language* multilingual features, such as cultural knowledge, also introduce distinct explainability challenges.

Core Findings. Our survey reveals tendency to apply existing explainability methods – typically developed for English and/or a handful of high resource languages – to multilingual settings, either off-the-shelf techniques or via simple adaptation. This is particularly common in NLP applications. It may result in broad generalisability but can compromise the usefulness and, ultimately fairness, for specific target languages. Furthermore, our findings indicate, albeit not unexpectedly, a significant skew in the literature against low-resource languages which primarily involves simplistic applications of out-of-the-shelf explainability techniques (e.g., LIME, SHAP). Impactful, cutting-edge multilingual explainability research for these languages, particularly African and Asian ones, proved notably scarce.

Novelty of Methods. While our survey reveals that cutting-edge multilingual explainability research is scarce, several areas of genuine innovation are presented. One such area is the analysis of latent representations to understand cross-lingual transfer, which has revealed the emergence of both spontaneous multilingual alignment and language-agnostic latent spaces. Other novel approaches target specific linguistic contexts, such as probing language-specific features like lexical tone or using artificial languages as testbeds to analyse knowledge transfer. Notably, causal and counterfactual analysis represents a significant exception, offering robust methods for typically low-resource settings. Finally, task-specific innovation is evident in the development of inherently interpretable metrics for machine translation quality estimation.

Future Directions. First, a primary imperative is the development of multilingual explainability innovations, rather than merely adapting existing methods to non-English contexts, by advancing prior work on language-specific features, such as dialects, accents, tones (e.g., probing dialectal knowledge in MLLMs and language-specific tonal information in speech models), and furthering the understanding of cross-lingual transfer – for instance, identifying at which stage of training language-agnostic latent spaces emerge and how they affect text generation in multiple languages. Second, *bridging the gap between interpreting inner model behaviour and explaining final model decisions* is essential, for example, by enhancing the faithfulness of multilingual explanations (e.g., developing training techniques, such as prompt optimisation, to improve faithfulness, which can be evaluated with existing datasets).

Interpreting *extra-language*, external knowledge, such as cultural values, regional variations and factual knowledge (e.g., probing and localising cultural knowledge in MLLMs), also constitutes a promising future direction, particularly concerning its knowledge transfer and its role in cultural adaptation of NLP across domains and languages. Lastly, shifting the research focus to low-resource languages from NLP applications to core NLP tasks (e.g., interpreting syntax-related information in low-resource languages) represents another vital avenue for future work.

Limitations

This work is presented as a survey rather than a systematic literature review; consequently, our methodological choices reflect this specific scope. For instance, the keyword selection for paper retrieval was tailored to provide a representative overview, which differs from the exhaustive coverage characteristic of a systematic review. Furthermore, inherent limitations and occasional inconsistencies within large-scale paper repositories, such as Semantic Scholar, may mean some relevant publications were not identified. The assessment of paper relevance, while informed by the authors' domain expertise, naturally incorporates a degree of subjectivity inherent in a survey format. Similarly, the categorisation of papers, particularly within nuanced or overlapping areas like "probing of representations" and "latent space analysis", involves an element of interpretative judgement. These considerations are consistent with our objective to map the representative terrain of the field.

Acknowledgements

This work was supported by the UK Research and Innovation (UKRI) Frontier Research Grant EP/Y031350/1 EQUATE.

This research was co-funded by the European Union (ERC, ExplainYourself, 101077481), and supported by the Pioneer Centre for AI, DNRF grant number P1. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

Lucas Resck gratefully acknowledges funding from the Cambridge Commonwealth, European and International Trust through a PhD scholarship.

AI tools were employed to assist with specific tasks, including coding, text refinement and information summarisation, enhancing overall workflow efficiency. The authors meticulously reviewed all AI-assisted outputs and bear full responsibility for the final content of this manuscript.

References

Maged Abdelaty and Shaimaa Y. Lazem. 2024. Investigating the Robustness of Arabic Offensive Language Transformer-Based Classifiers to Adversarial Attacks. *Internet, Multimedia Systems and Applications*.

Badr M. Abdullah, Mohammed Maqsood Shaik, and Dietrich Klakow. 2024. Wave to Interlingua: Analyzing Representations of Multilingual Speech Transformers for Spoken Language Translation. *Interspeech*.

Miguel Abreu-Cardenas, Saúl Calderón-Ramírez, and M. Solis. 2023. Uncertainty Estimation for Complex Text Detection in Spanish. *2023 IEEE 5th International Conference on BioInspired Processing (BIP)*.

Ehsan Aghazadeh, Mohsen Fayyaz, and Yadollah Yaghoobzadeh. 2022. Metaphors in pre-trained language models: Probing and generalization across datasets and languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2037–2050, Dublin, Ireland. Association for Computational Linguistics.

Manex Agirrezabal, Sidsel Boldsen, and Nora Hollenstein. 2023. The hidden folk: Linguistic properties encoded in multilingual contextual character representations. In *Proceedings of the Workshop on Computation and Written Language (CAWL 2023)*, pages 6–13, Toronto, Canada. Association for Computational Linguistics.

Saud Althabiti, M. Alsalka, and Eric Atwell. 2024. TA'KEED the First Generative Fact-Checking System for Arabic Claims. *Artificial Intelligence and Applications*.

Tejaswini Ananthanarayana, Nikunj Kotecha, Priyanshu Srivastava, Lipisha Chaudhary, Nicholas Wilkins, and Ifeoma Nwogu. 2021. Dynamic Cross-Feature Fusion for American Sign Language Translation. *IEEE International Conference on Automatic Face & Gesture Recognition*.

Akhilesh Aravapalli, Mounika Marreddy, S. Oota, Radhika Mamidi, and Manish Gupta. 2024. IndicSentEval: How Effectively do Multilingual Transformer Models encode Linguistic Properties for Indic Languages? *arXiv.org*.

Giuseppe Attanasio, Debora Nozza, Eliana Pastor, and Dirk Hovy. 2022. Benchmarking post-hoc interpretability approaches for transformer-based misogyny detection. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pages 100–112, Dublin, Ireland. Association for Computational Linguistics.

Nadiah A. Baghdadi, Yousry M Abdulazeem, Hanaa ZainEldin, T. A. Farrag, Mansourah Aljohani, Amer Malki, Mahmoud Badawy, and Mostafa A. Elhosseini. 2024. Toward Robust Arabic Sign Language Recognition via Vision Transformers and Local Interpretable Model-agnostic Explanations Integration. *Journal of Disability Research*.

Nicolas Ballier, Léa Burin, Behnoosh Namdarzadeh, Sara B Ng, Richard Wright, and Jean-Baptiste Yunès. 2024. Probing Whisper Predictions for French, English and Persian Transcriptions. *International Conference on Natural Language and Speech Processing*.

Valentin Barriere and Sebastian Cifuentes. 2024a. Are text classifiers xenophobic? a country-oriented bias detection method with least confounding variables. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1511–1518, Torino, Italia. ELRA and ICCL.

Valentin Barriere and Sebastian Cifuentes. 2024b. A study of nationality bias in names and perplexity using off-the-shelf affect-related tweet classifiers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 569–579, Miami, Florida, USA. Association for Computational Linguistics.

Hadas Ben-Atya, Naama Gavrielov, Zvi Badash, Gili Focht, Ruth Cytter-Kuint, Talar Hagopian, Dan Turner, and Moti Freiman. 2025. Agent-Based Uncertainty Awareness Improves Automated Radiology Report Labeling with an Open-Source Large Language Model. *arXiv preprint*.

Matteo Berta, Salvatore Greco, Giuseppe Tipaldo, and Tania Cerquitelli. 2024. Decoding Narratives: Towards a Classification Analysis for Stereotypical Patterns in Italian News Headlines. *BigData Congress [Services Society]*.

P. D. Bianco, Oscar Stanchi, F. Quiroga, Franco Ronchetti, and Enzo Ferrante. 2024. SignAttention: On the Interpretability of Transformer Models for Sign Language Translation. *arXiv.org*.

Adrien Bibal, Rémi Cardon, David Alfter, Rodrigo Wilkens, Xiaoou Wang, Thomas François, and Patrick Watrin. 2022. Is Attention Explanation? An Introduction to the Debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3889–3900, Dublin, Ireland. Association for Computational Linguistics.

Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. 2016. Layer-Wise Relevance Propagation for Neural Networks with Local Renormalization Layers. In *Artificial Neural Networks and Machine Learning – ICANN 2016*, volume 9887 of *International Conference on Artificial Neural Networks*, pages 63–71, Cham. Springer International Publishing.

Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2022. Analyzing the mono- and cross-lingual pretraining dynamics of multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3575–3590, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jérémie Bogaert, Emmanuel Jean, Cyril de Bodt, and François-Xavier Standaert. 2023. Fine-tuning is not (always) overfitting artifacts. *The European Symposium on Artificial Neural Networks*.

Jérémie Bogaert, M. Marneffe, Antonin Descampe, Louis Escouflaire, Cédrick Fairon, and François-Xavier Standaert. 2024. Explanation sensitivity to the randomness of large language models: the case of journalistic text classification. *arXiv.org*.

Sidsel Boldsen, Manex Agirrezabal, and Nora Hollenstein. 2022. Interpreting character embeddings with perceptual representations: The case of shape, sound, and color. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6819–6836, Dublin, Ireland. Association for Computational Linguistics.

Sabri Boughorbel, Md Rizwan Parvez, and Majd Hawasly. 2024. Improving language models trained on translated data with continual pre-training and dictionary learning analysis. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 73–88, Bangkok, Thailand. Association for Computational Linguistics.

Stephanie Brandl and Oliver Eberle. 2024. Comparing zero-shot self-explanations with human rationales in multilingual text classification. *arXiv.org*.

Stephanie Brandl, Oliver Eberle, Tiago Ribeiro, Anders Søgaard, and Nora Hollenstein. 2024. Evaluating webcam-based gaze data as an alternative for human rationale annotations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6544–6556, Torino, Italia. ELRA and ICCL.

Giuseppe Buonaiuto, Raffaele Guarasci, Aniello Minutolo, G. Pietro, and M. Esposito. 2024. Quantum Transfer Learning for Acceptability Judgements. *Quantum Machine Intelligence*.

R. A. Calix, Jj Ben-Joseph, Nina Lopatina, Ryan Ashley, Mona Gogia, George P. Sieniawski, and Andrea L. Brennen. 2022. Saisiyat Is Where It Is At! Insights Into Backdoors And Debiasing Of Cross Lingual Transformers For Named Entity Recognition. *2022 IEEE International Conference on Big Data (Big Data)*.

Pengfei Cao, Yuheng Chen, Zhuoran Jin, Yubo Chen, Kang Liu, and Jun Zhao. 2024. One Mind, Many Tongues: A Deep Dive into Language-Agnostic Knowledge Neurons in Large Language Models. *arXiv.org*.

Tommaso Caselli, Irene Dini, and Felice Dell'Orletta. 2022. How about time? probing a multilingual language model for temporal relations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3197–3209, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Beiduo Chen, Wu Guo, Quan Liu, and Kun Tao. 2022. Feature Aggregation in Zero-Shot Cross-Lingual Transfer Using Multilingual BERT. *International Conference on Pattern Recognition*.

Nuo Chen, Ning Wu, Shining Liang, Ming Gong, Linjun Shou, Dongmei Zhang, and Jia Li. 2023. Is Bigger and Deeper Always Better? Probing LLaMA Across Scales and Layers. *arXiv.org*.

Yanfang Chen, Ding Chen, Shichao Song, Simin Niu, Hanyu Wang, Zeyun Tang, Feiyu Xiong, and Zhiyu Li. 2024a. HRDE: Retrieval-Augmented Large Language Models for Chinese Health Rumor Detection and Explainability. *arXiv.org*.

Yongjian Chen and M. Farrús. 2022. Neural Detection of Cross-lingual Syntactic Knowledge. *IberSPEECH Conference*.

Yuyan Chen, Yichen Yuan, Panjun Liu, Dayiheng Liu, Qinghao Guan, Mengfei Guo, Haiming Peng, Bang Liu, Zhixu Li, and Yanghua Xiao. 2024b. Talk Funny! A Large-Scale Humor Response Dataset with Chain-of-Humor Interpretation. *AAAI Conference on Artificial Intelligence*.

Long Cheng, Qihao Shao, Christine Zhao, Sheng Bi, and Gina-Anne Levow. 2024. TEII: Think, explain, interact and iterate with large language models to solve cross-lingual emotion detection. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 495–504, Bangkok, Thailand. Association for Computational Linguistics.

Ahmadul Karim Chowdhury, Saidur Rahman Sujon, Md. Shirajus Salekin Shafi, Tasin Ahmmad, Sifat Ahmed, Khan Md Hasib, and Faisal Muhammad Shah. 2024. Harnessing Large Language Models Over Transformer Models for Detecting Bengali Depressive Social Media Text: A Comprehensive Study. *Natural Language Processing Journal*.

Simone Conia and Roberto Navigli. 2022. Probing for predicate argument structures in pretrained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4622–4632, Dublin, Ireland. Association for Computational Linguistics.

Yiming Cui, Ting Liu, Wanxiang Che, Zhigang Chen, and Shijin Wang. 2022. Teaching Machines to Read, Answer and Explain. *IEEE/ACM Transactions on Audio Speech and Language Processing*.

Yiming Cui, Wei-Nan Zhang, Wanxiang Che, Ting Liu, Zhigang Chen, and Shijin Wang. 2021. Multilingual Multi-Aspect Explainability Analyses on Machine Reading Comprehension Models. *arXiv preprint*.

Thao Anh Dang, Limor Raviv, and Lukas Galke. 2024. Tokenization and Morphology in Multilingual Language Models: A Comparative Analysis of mT5 and ByT5. *arXiv.org*.

Ant'on de la Fuente and Dan Jurafsky. 2024. A layer-wise analysis of Mandarin and English suprasegmentals in SSL speech models. *Interspeech*.

Loic De Langhe, Orphee De Clercq, and Veronique Hoste. 2023. What does BERT actually learn about event coreference? probing structural information in a fine-tuned Dutch language model. In *Proceedings of the Fourth Workshop on Insights from Negative Results in NLP*, pages 103–108, Dubrovnik, Croatia. Association for Computational Linguistics.

Darshan Deshpande, Selvan Sunitha Ravi, Sky CH-Wang, B. Mielczarek, Anand Kannappan, and Rebecca Qian. 2024. GLIDER: Grading LLM Interactions and Decisions using Explainable Ranking. *arXiv.org*.

Clément Dumas, Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2025. Separating Tongue from Thought: Activation Patching Reveals Language-Agnostic Concept Representations in Transformers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31822–31841, Vienna, Austria. Association for Computational Linguistics.

Sebastian-Vasile Echim, Razvan-Alexandru Smadu, and Dumitru-Clementin Cercel. 2024. Benchmarking Adversarial Robustness in Speech Emotion Recognition: Insights into Low-Resource Romanian and German Languages. *European Conference on Artificial Intelligence*.

Bojan Evkoski and Senja Pollak. 2023. XAI in Computational Linguistics: Understanding Political Leanings in the Slovenian Parliament. *arXiv preprint*.

Jiahai Feng, Stuart Russell, and Jacob Steinhardt. 2024. Monitoring Latent World States in Language Models with Propositional Probes. *arXiv.org*.

Javier Ferrando and Marta R. Costa-jussà. 2024. On the similarity of circuits across languages: a case study on the subject-verb agreement task. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10115–10125, Miami, Florida, USA. Association for Computational Linguistics.

Javier Ferrando, Gerard I. Gállego, Belen Alastruey, Carlos Escolano, and Marta R. Costa-jussà. 2022. Towards opening the black box of neural machine translation: Source and target interpretations of the transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8756–8769, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Javier Ferrando and Elena Voita. 2024. Information flow routes: Automatically interpreting language models at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17432–17445, Miami, Florida, USA. Association for Computational Linguistics.

Y. Fung, Tuhin Chakraborty, Hao Guo, Owen Rambow, S. Muresan, and Heng Ji. 2022. NormSAGE: Multi-Lingual Multi-Cultural Norm Discovery from Conversations On-the-Fly. *Conference on Empirical Methods in Natural Language Processing*.

Lance Calvin Lim Gamboa and Mark Lee. 2024. A Novel Interpretability Metric for Explaining Bias in Language Models: Applications on Multilingual Models from Southeast Asia. *arXiv.org*.

Rouzbeh Ghasemi and S. Momtazi. 2023. How a Deep Contextualized Representation and Attention Mechanism Justifies Explainable Cross-Lingual Sentiment Analysis. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*

Sai Gopinath and Joselyn Rodriguez. 2024. Probing self-attention in self-supervised speech models for cross-linguistic differences. *arXiv.org*.

R. Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, Evan Hubinger, Kamil.e Lukovsiut.e, Karina Nguyen, Nicholas Joseph, Sam McCandlish, Jared Kaplan, and Sam Bowman. 2023. Studying Large Language Model Generalization with Influence Functions. *arXiv.org*.

Muzhe Guo, Muhao Guo, Juntao Su, Junyu Chen, Jiaqian Yu, Jiaqi Wang, Hongfei Du, Parmanand Sahu, Ashwin Assysh Sharma, and Fang Jin. 2024. Bayesian Iterative Prediction and Lexical-based Interpretation for Disturbed Chinese Sentence Pair Matching. *The Web Conference*.

Wataru Hashimoto, Hidetaka Kamigaito, and Taro Watanabe. 2024. Are data augmentation methods in named entity recognition applicable for uncertainty estimation? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18852–18867, Miami, Florida, USA. Association for Computational Linguistics.

Ehtesham Hashmi, Muhammad Mudassar Yamin, Shariq Imran, Sule YAYILGAN YILDIRIM, and Mohib Ullah. 2024a. Enhancing Misogyny Detection in Bilingual Texts Using FastText and Explainable AI. *2024 International Conference on Engineering & Computing Technologies (ICECT)*.

Ehtesham Hashmi, Sule YAYILGAN YILDIRIM, Ibrahim A. Hameed, M. Yamin, Mohib Ullah, and Mohamed Abomhara. 2024b. Enhancing Multilingual Hate Speech Detection: From Language-Specific Insights to Cross-Linguistic Integration. *IEEE Access*.

Linyang He, Ercong Nie, Helmut Schmid, Hinrich Schutze, N. Mesgarani, and Jonathan R. Brennan. 2024. Large Language Models as Neurolinguistic Subjects: Identifying Internal Representations for Form and Meaning. *arXiv.org*.

Musashi Hinck, Carolin Holtermann, Matthew Lyle Olson, Florian Schneider, Sungduk Yu, Anahita Bhiwandiwalla, Anne Lauscher, Shao-Yen Tseng, and Vasudev Lal. 2024. Why do LLaVA vision-language models reply to images in English? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13402–13421, Miami, Florida, USA. Association for Computational Linguistics.

MD. Mithun Hossain, Md Shakil Hossain, Md Shakhawat Hossain, M. F. Mridha, Mejdl S. Safran, and Sultan Alfarhood. 2024a. TransNet: Deep Attentional Hybrid Transformer for Arabic Posts Classification. *IEEE Access*.

MD. Mithun Hossain, Md Shakil Hossain, Mejdl S. Safran, Sultan Alfarhood, Meshal Alfarhood, and M. F. Mridha. 2024b. A Hybrid Attention-Based Transformer Model for Arabic News Classification Using Text Embedding and Deep Learning. *IEEE Access*.

Peng Hu, Sizhe Liu, Changjiang Gao, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian Huang. 2024. Large Language Models Are Cross-Lingual Knowledge-Free Reasoners. *arXiv.org*.

Zhihong Huang, Longyue Wang, Siyou Liu, and Derek F. Wong. 2023. How Does Pretraining Improve Discourse-Aware Translation? *Interspeech*.

Benjamin Icard, Evangelia Zve, Lila Sainero, Alice Breton, and Jean-Gabriel Ganascia. 2025. Embedding style beyond topics: Analyzing dispersion effects across different language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3511–3522, Abu Dhabi, UAE. Association for Computational Linguistics.

Verena Irrgang, Veronika Solopova, Steffen Zeiler, Robert M. Nickel, and D. Kolossa. 2024. Features and Detectability of German Texts Generated with Large Language Models. *Conference on Natural Language Processing*.

Ai Ishii, Naoya Inoue, Hisami Suzuki, and Satoshi Sekine. 2024. JEMHopQA: Dataset for Japanese explainable multi-hop question answering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9515–9525, Torino, Italia. ELRA and ICCL.

Saiful Islam, Md Jabed Hosen, Fowzia Rahman Taznin, Naznin Sultana, Md. Injamul Haque, and Shakil Rana. 2024. An Efficient Framework for Transliteration Sentence Identification of Low Resource Languages Using Hybrid BERT-BiGRU. *International Conference on Computing Communication and Networking Technologies*.

Yalong Jia, Zhenghui Ou, and Yang Yang. 2022. SPDB innovation lab at SemEval-2022 task 10: A novel end-to-end structured sentiment analysis model based on the ERNIE-M. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1401–1405, Seattle, United States. Association for Computational Linguistics.

Cong Jiang and Xiaolei Yang. 2023. Legal Syllogism Prompting: Teaching Large Language Models for Legal Judgment Prediction. *International Conference on Artificial Intelligence and Law*.

Xinke Jiang, Ruizhe Zhang, Yongxin Xu, Rihong Qiu, Yue Fang, Zhiyuan Wang, Jinyi Tang, Hongxin Ding, Xu Chu, Junfeng Zhao, and Yasha Wang. 2023. HyKGE: A Hypothesis Knowledge Graph Enhanced Framework for Accurate and Reliable Medical LLMs Responses. *arXiv preprint*.

Rasmus Jørgensen, Fiammetta Caccavale, Christian Igel, and Anders Søgaard. 2022. Are multilingual sentiment models equally right for the right reasons? In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 131–141, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tasnim Kabir and Marine Carpuat. 2021. The UMD submission to the explainable MT quality estimation shared task: Combining explanation models with sequence labeling. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 230–237, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Migyeong Kang, Goun Choi, Hyolim Jeon, Ji Hyun An, Daejin Choi, and Jinyoung Han. 2024. CURE: Context- and uncertainty-aware mental disorder detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17924–17940, Miami, Florida, USA. Association for Computational Linguistics.

Md. Rezaul Karim, Sumon Dey, and Bharathi Raja Chakravarthi. 2020. DeepHateExplainer: Explainable Hate Speech Detection in Under-resourced Bengali Language. *International Conference on Data Science and Advanced Analytics*.

Anisia Katinskaia and Roman Yangarber. 2024. Probing the category of verbal aspect in transformer language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3347–3366, Mexico City, Mexico. Association for Computational Linguistics.

Ayush Kaushal and Kyle Mahowald. 2022. What do tokens know about their characters and how do they know it? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2487–2507, Seattle, United States. Association for Computational Linguistics.

Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David Graham, Fangzhou Hu, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu, Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler Murray, Chris Newell, Smita Rao, Shaurya Rohatgi, Paul Sayre, Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, Amber Tanaka, Alex D. Wade, Linda Wagner, Lucy Lu Wang, Chris Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Madeleine Van Zuylen, and Daniel S. Weld. 2025. The Semantic Scholar Open Data Platform. *arXiv preprint*. ArXiv:2301.10140 [cs].

Belinda Marion Kobusingye, A. Dorothy, J. Nakatumba-Nabende, and Ggaliwango Marvin. 2023. Explainable Machine Translation for Intelligent E-Learning of Social Studies. *2023 7th International Conference on Trends in Electronics and Informatics (ICOEI)*.

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. 2024. On the Multilingual Ability of Decoder-based Pre-trained Language Models: Finding and Controlling Language-Specific Neurons. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6919–6971, Mexico City, Mexico. Association for Computational Linguistics.

Anastasia Kozlova, Albina Akhmetgareeva, Aigul Khanova, Semen Kudriavtsev, and Alena Fenogenova. 2024. Transformer attention vs human attention in anaphora resolution. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 109–122, Bangkok, Thailand. Association for Computational Linguistics.

Lea Krause, Wondimagegnhue Tufa, Selene Baez Santamaria, Angel Daza, Urja Khurana, and Piek Vossen. 2023. Confidently wrong: Exploring the calibration and expression of (un)certainty of large language models in a multilingual setting. In *Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023)*, pages 1–9, Prague, Czech Republic. Association for Computational Linguistics.

Simon Kurz, Jian-Jia Chen, Lucie Flek, and Zhixue Zhao. 2024. Investigating Language-Specific Calibration For Pruning Multilingual Large Language Models. *arXiv preprint*.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. Okapi: Instruction-tuned Large Language Models in Multiple Languages with Reinforcement Learning from Human Feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327, Singapore. Association for Computational Linguistics.

Michael Lan, Philip Torr, and Fazl Barez. 2024. Towards interpretable sequence continuation: Analyzing shared circuits in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12576–12601, Miami, Florida, USA. Association for Computational Linguistics.

Anne Lauscher, Federico Bianchi, Samuel R. Bowman, and Dirk Hovy. 2022. SocioProbe: What, when, and where language models learn about sociodemographics. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7901–7918, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

You-Qian Lee, Ching-Tai Chen, Chien-Chang Chen, Chung-Hong Lee, Pei-Tsz Chen, Chi-Shin Wu, and Hong-Jie Dai. 2023. Unlocking the Secrets Behind Advanced Artificial Intelligence Language Models in Deidentifying Chinese-English Mixed Clinical Text: Development and Validation Study. *Journal of Medical Internet Research*.

Yunseung Lee and Daehee Han. 2024. KorSmishing explainer: A Korean-centric LLM-based framework for smishing detection and explanation generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 642–656, Miami, Florida, US. Association for Computational Linguistics.

Bingzhi Li, Guillaume Wisniewski, and Benoit Crabbé. 2022. How distributed are distributed representations? an observation on the locality of syntactic information in verb agreement tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 501–507, Dublin, Ireland. Association for Computational Linguistics.

Bingzhi Li, Guillaume Wisniewski, and Benoît Crabbé. 2023a. Assessing the capacity of transformer to abstract syntactic representations: A contrastive analysis based on long-distance agreement. *Transactions of the Association for Computational Linguistics*, 11:18–33.

Daoyang Li, Mingyu Jin, Qingcheng Zeng, Haiyan Zhao, and Mengnan Du. 2024a. Exploring Multilingual Probing in Large Language Models: A Cross-Language Analysis. *arXiv.org*.

Dongfang Li, Jindi Yu, Baotian Hu, Zhenran Xu, and Min Zhang. 2023b. ExplainCPE: A free-text explanation benchmark of Chinese pharmacist examination. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1922–1940, Singapore. Association for Computational Linguistics.

Jiahong Li, Yiyuan Chen, Yichi Wang, Yiqiang Ye, Min Sun, Haopan Ren, Weibin Cheng, and Haodi Zhang. 2023c. Interpretable Pulmonary Disease Diagnosis with Graph Neural Network and Counterfactual Explanations. *2023 2nd International Conference on Sensing, Measurement, Communication and Internet of Things Technologies (SMC-IoT)*.

Jie Li, Yi Liu, Chongyang Liu, Ling Shi, Xiaoning Ren, Yaowen Zheng, Yang Liu, and Yinxing Xue. 2024b. A Cross-Language Investigation into Jailbreak Attacks in Large Language Models. *arXiv.org*.

Xiaochen Li, Zheng Xin Yong, and Stephen Bach. 2024c. Preference tuning for toxicity mitigation generalizes across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13422–13440, Miami, Florida, USA. Association for Computational Linguistics.

Sheng Liang, Philipp Dufter, and Hinrich Schütze. 2020. Monolingual and multilingual reduction of gender bias in contextualized representations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5082–5093, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yu-Ting Lin, Yuan-Xiang Deng, Chu-Lin Tsai, Chien-Hua Huang, and Lijuan Fu. 2024. Interpretable Deep Learning System for Identifying Critical Patients Through the Prediction of Triage Level, Hospitalization, and Length of Stay: Prospective Study. *JMIR Medical Informatics*.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2024a. Culturally Aware and Adapted NLP: A Taxonomy and a Survey of the State of the Art. *arXiv preprint*. ArXiv:2406.03930.

Fan Liu, Yue Feng, Zhao Xu, Lixin Su, Xinyu Ma, Dawei Yin, and Hao Liu. 2024b. JAILJUDGE: A Comprehensive Jailbreak Judge Benchmark with Multi-Agent Enhanced Explanation Evaluation Framework. *arXiv.org*.

Qi Liu, Matt Kusner, and Phil Blunsom. 2021. Counterfactual data augmentation for neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 187–197, Online. Association for Computational Linguistics.

Shuo Liu, J. Keung, Zhen Yang, Fang Liu, Qilin Zhou, and Yihan Liao. 2024c. Delving into Parameter-Efficient Fine-Tuning in Code Change Learning: An Empirical Study. *IEEE International Conference on Software Analysis, Evolution, and Reengineering*.

Weize Liu, Yinlong Xu, Hongxia Xu, Jintai Chen, Xuming Hu, and Jian Wu. 2024d. Unraveling Babel: Exploring multilingual activation patterns of LLMs and their applications. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11855–11881, Miami, Florida, USA. Association for Computational Linguistics.

Antoine Louis, G. van Dijck, and Gerasimos Spanakis. 2023. Interpretable Long-Form Legal Question Answering with Retrieval-Augmented Large Language Models. *AAAI Conference on Artificial Intelligence*.

Qingyu Lu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. 2024. Error analysis prompting enables human-like translation evaluation in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8801–8816, Bangkok, Thailand. Association for Computational Linguistics.

Qiuhao Lu, Sairam Gurajada, Prithviraj Sen, Lucian Popa, Dejing Dou, and Thien Nguyen. 2022. Cross-lingual short-text entity linking: Generating features for neuro-symbolic methods. In *Proceedings of the Fourth Workshop on Data Science with Human-in-the-Loop (Language Advances)*, pages 8–14, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, volume 30. Curran Associates, Inc.

Haoyan Luo and Lucia Specia. 2024. From Understanding to Utilization: A Survey on Explainability for Large Language Models. *arXiv preprint*. ArXiv:2401.12874 [cs].

Weicheng Ma, Kai Zhang, Renze Lou, Lili Wang, and Soroush Vosoughi. 2021. Contributions of transformer attention heads in multi- and cross-lingual tasks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1956–1966, Online. Association for Computational Linguistics.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605.

K. Mabokela, Mpho Primus, and Turgay Çelik. 2024. Explainable Pre-Trained Language Models for Sentiment Analysis in Low-Resourced Languages. *Big Data and Cognitive Computing*.

VÍıtor Machado, C. Bom, Kary Ocaña, R. Terra, and Miriam Chaves. 2022. Using Deep Learning Transformer Networks to Identify Symptoms Associated with COVID-19 on Twitter. *Notas técnicas*.

Daniil Maksymenko and Oleskii Turuta. 2024. Interpretable Conversation Routing via the Latent Embeddings Approach. *Computation*.

Melusi Malinga, Isaac Lupanda, Mike Wa Nkongolo, and Jacobus Philippus van Deventer. 2024. A Multilingual Sentiment Lexicon for Low-Resource Language Translation using Large Languages Models and Explainable AI. *arXiv.org*.

Mamta Mamta, Zishan Ahmad, and Asif Ekbal. 2023. Elevating code-mixed text handling through auditory information of words. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15918–15932, Singapore. Association for Computational Linguistics.

D. Mareček, H. Çelikkanat, Miikka Silfverberg, Vinit Ravishankar, and J. Tiedemann. 2020. Are Multilingual Neural Machine Translation Models Better at Capturing Linguistic Features? *Prague Bulletin of Mathematical Linguistics*.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29):861.

Tim Metzler, P. Plöger, and Jörn Hees. 2024. Computer-Assisted Short Answer Grading Using Large Language Models and Rubrics. *Jahrestagung der Gesellschaft für Informatik*.

Alessio Miaschi, Chiara Alzetta, D. Brunato, F. Dell'Orletta, and Giulia Venturi. 2023a. Testing the Effectiveness of the Diagnostic Probing Paradigm on Italian Treebanks. *Inf.*

Alessio Miaschi, D. Brunato, F. Dell'Orletta, and Giulia Venturi. 2023b. On Robustness and Sensitivity of a Neural Language Model: A Case Study on Italian L1 Learner Errors. *IEEE/ACM Transactions on Audio Speech and Language Processing*.

Alessio Miaschi, Gabriele Sarti, D. Brunato, F. Dell'Orletta, and Giulia Venturi. 2022. Probing Linguistic Knowledge in Italian Neural Language Models across Language Varieties. *Italian Journal of Computational Linguistics*.

Anamaria-Monica Migea, Vlad-Andrei Negru, Sebastian-Antonio Toma, C. Lemnaru, and R. Potolea. 2024. Cook Smarter Not Harder: Enhancing Learning Capacity in Smart Ovens with Supplementary Data. *International Conference on Computational Photography*.

Vladislav Mikhailov, Oleg Serikov, and Ekaterina Artemova. 2021a. Morph call: Probing morphosyntactic content of multilingual transformers. In *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*, pages 97–121, Online. Association for Computational Linguistics.

Vladislav Mikhailov, Ekaterina Taktasheva, Elina Sigdel, and Ekaterina Artemova. 2021b. RuSentEval: Linguistic source, encoder force! In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 43–65, Kiyv, Ukraine. Association for Computational Linguistics.

Hosein Mohebbi, Grzegorz Chrupała, Willem Zuidema, and Afra Alishahi. 2023. Homophone disambiguation reveals patterns of context mixing in speech transformers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8249–8260, Singapore. Association for Computational Linguistics.

Yongyu Mu, Peinan Feng, Zhiquan Cao, Yuzhang Wu, Bei Li, Chenglong Wang, Tong Xiao, Kai Song, Tongran Liu, Chunliang Zhang, and Jingbo Zhu. 2024. Revealing the Parallel Multilingual Learning within Large Language Models. *arXiv preprint*.

Aaron Mueller, Yu Xia, and Tal Linzen. 2022. Causal analysis of syntactic agreement neurons in multilingual language models. In *Proceedings of the*

26th Conference on Computational Natural Language Learning (CoNLL), pages 95–109, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

A. Munthuli, P. Pooprasert, N. Klangpornkun, P. Phienphanich, C. Onsuwan, Kankamol Jaisin, Keerati Pattanaseri, Juthawadee Lortrakul, and C. Tantibundhit. 2023. Classification and analysis of text transcription from Thai depression assessment tasks among patients with depression. *PLoS ONE*.

Ahmad Mustafa, Saja Nakhleh, R. Irsheidat, and R. Alruosan. 2024. Interpreting Arabic Transformer Models: A Study on XAI Interpretability for Quranic Semantic Search Models. *Jordanian Journal of Computers and Information Technology*.

Nikolaos Mylonas, Nikolaos Stylianou, T. Tsikrika, S. Vrochidis, and I. Kompatsiaris. 2024. A Multi-Task Text Classification Pipeline with Natural Language Explanations: A User-Centric Evaluation in Sentiment Analysis and Offensive Language Identification in Greek Tweets. *arXiv.org*.

Aleksandra Mysiak and Jacek Cyranka. 2023. Is German secretly a Slavic language? what BERT probing can tell us about language groups. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 86–93, Dubrovnik, Croatia. Association for Computational Linguistics.

Inez Okulska and Emilia Wiśnios. 2023. Towards Harmful Erotic Content Detection through Coreference-Driven Contextual Analysis. *CRAC*.

Hilário Oliveira, Rafael Ferreira Mello, Bruno Alexandre Barreiros Rosa, Mladen Raković, Péricles B. C. Miranda, T. Cordeiro, Seiji Isotani, I. Bittencourt, and D. Gašević. 2023. Towards explainable prediction of essay cohesion in Portuguese and English. *International Conference on Learning Analytics and Knowledge*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. GPT-4 Technical Report. *arXiv preprint*. ArXiv:2303.08774 [cs].

Asli Umay Ozturk, R. Çekinel, and Pinar Senkul. 2024. Make Satire Boring Again: Reducing Stylistic Bias of Satirical Corpus by Utilizing Generative LLMs. *arXiv.org*.

Pia Pachinger, Janis Goldzycher, Anna Planitzer, Wojciech Kusa, Allan Hanbury, and Julia Neidhardt. 2024. AustroTox: A dataset for target-based Austrian German offensive language detection. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11990–12001, Bangkok, Thailand. Association for Computational Linguistics.

Dojun Park and Sebastian Padó. 2024. Multi-dimensional machine translation evaluation: Model evaluation and resource for Korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11723–11744, Torino, Italia. ELRA and ICCL.

D.E. Pashchenko, E. Razova, A. Kotelnikova, S. Vychegzhanin, and E. V. Kotelnikov. 2022. Interpretation of Language Models Attention Matrices in Texts Sentiment Analysis. *2022 VIII International Conference on Information Technology and Nanotechnology (ITNT)*.

Gal Patel, Leshem Choshen, and Omri Abend. 2022. On neurons invariant to sentence structural changes in neural machine translation. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 194–212, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Rajaswa Patil, Jasleen Dhillon, Siddhant Mahurkar, Saumitra Kulkarni, Manav Malhotra, and Veeky Baths. 2021. Vyākarana: A colorless green benchmark for syntactic evaluation in indic languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 153–165, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2024. Survey of Cultural Awareness in Language Models: Text and Beyond. *arXiv preprint*. ArXiv:2411.00860.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Luca Putelli, A. Gerevini, A. Lavelli, T. Mehmood, and I. Serina. 2022. On the Behaviour of BERT's Attention for the Classification of Medical Reports. *XAI.it@AI*IA*.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2024. Multilingual Large Language Model: A Survey of Resources, Taxonomy and Frontiers. *arXiv preprint*. ArXiv:2404.04925 [cs].

Lucia Quirke, Lovis Heindrich, Wes Gurnee, and Neel Nanda. 2023. Training Dynamics of Contextual N-Grams in Language Models. *arXiv.org*.

Azzam Radman, Mohammed Atros, and Rehab Duwairi. 2023. Neural Arabic singular-to-plural conversion using a pretrained Character-BERT and a fused transformer. *Natural Language Engineering*.

Rashedur M. Rahman, Gw'enol'e Lecorv'e, and Nicolas B'echet. 2023. Age Recommendation from Texts and Sentences for Children. *arXiv.org*.

Mirco Ramo, G. Silvestre, and F. Balado. 2023. Small, Multilingual, Explainable Transformers for Online Handwriting Decoding. *Irish Signals and Systems Conference*.

Juncang Rao and Yanshan He. 2024. Research on Automated Scoring Method for HSK Essays with Hybrid Features. *International Conference on Asian Language Processing*.

Manikandan Ravikiran and Bharathi Raja Chakravarthi. 2022. Zero-shot code-mixed offensive span identification through rationale extraction. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 240–247, Dublin, Ireland. Association for Computational Linguistics.

Ryokan Ri and Yoshimasa Tsuruoka. 2022. Pretraining with artificial language: Studying transferable knowledge in language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7302–7315, Dublin, Ireland. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144, New York, NY, USA. Association for Computing Machinery.

Phillip Richter-Pechanski, Philipp Wiesenbach, Dominic M. Schwab, Christina Kiriakou, Nicolas Geis, Christoph Dieterich, and Anette Frank. 2024. Clinical information extraction for Low-resource languages with Few-shot learning using Pre-trained language models and Prompting. *Natural Language Processing*.

M. T. Rietberg, Van Bach Nguyen, J. Geerdink, Onno Vijlbrief, and Christin Seifert. 2023. Accurate and Reliable Classification of Unstructured Reports on Their Diagnostic Goal Using BERT Models. *Diagnostics*.

Sandy Ritchie, Daan van Esch, Uche Okonkwo, Shikhar Vashishth, and Emily Drummond. 2024. LinguaMeta: Unified Metadata for Thousands of Languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10530–10538, Torino, Italia. ELRA and ICCL.

Ana Carolina Rodrigues and R. Marcacini. 2022. Sentence Similarity Recognition in Portuguese from Multiple Embedding Models. *International Conference on Machine Learning and Applications*.

Samuel Rönnqvist, Jenna Kanerva, Tapio Salakoski, and Filip Ginter. 2019. Is multilingual BERT fluent in language generation? In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 29–36, Turku, Finland. Linköping University Electronic Press.

Rishiraj Saha Roy, Joel Schlotthauer, Chris Hinze, Andreas Foltyn, Luzian Hahn, and Fabian Kuech. 2024. Evidence Contextualization and Counterfactual Attribution for Conversational QA over Heterogeneous Data with RAG Systems. *arXiv.org*.

Abdulwahab Sahyoun and Shady Shehata. 2023. AraDiaWER: An explainable metric for dialectical Arabic ASR. In *Proceedings of the Second Workshop on NLP Applications to Field Linguistics*, pages 64–73, Dubrovnik, Croatia. Association for Computational Linguistics.

Oleg Serikov, Vitaly Protasov, Ekaterina Voloshina, Viktoria Knyazkova, and Tatiana Shavrina. 2022. Universal and independent: Multilingual probing framework for exhaustive model interpretation and evaluation. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 441–456, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Sheikh Shafayat, Dongkeun Yoon, Woori Jang, Jiwoo Choi, Alice Oh, and Seohyon Jung. 2024. A 2-step Framework for Automated Literary Translation Evaluation: Its Promises and Pitfalls. *arXiv.org*.

Naomi Shapiro, Amandalynne Paullada, and Shane Steinert-Threlkeld. 2021. A multilabel approach to morphosyntactic probing. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4486–4524, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gaofei Shen, Michaela Watkins, Afra Alishahi, Arianna Bisazza, and Grzegorz Chrupała. 2024. Encoding of lexical tone in self-supervised models of spoken language. *arXiv preprint*.

Ya-Ming Shen, Lijie Wang, Ying Chen, Xinyan Xiao, Jing Liu, and Hua Wu. 2022. An Interpretability Evaluation Benchmark for Pre-trained Language Models. *arXiv.org*.

J. A. Siddiqui, S. Yuhaniz, Ghulam Mujtaba, Safdar Ali Soomro, and Zafar Ali Mahar. 2024. Fine-Grained Multilingual Hate Speech Detection Using Explainable AI and Transformers. *IEEE Access*.

Anant Singh and Akshat Gupta. 2023. Decoding Emotions: A comprehensive Multilingual Study of Speech Models for Speech Emotion Recognition. *arXiv.org*.

Tooba Sohail, Atiqa Aiman, Ehtesham Hashmi, A. Imran, Sher Muhammad Daudpota, and Sule YAYILGAN YILDIRIM. 2024. Hate Speech Detection in Code-Mixed Datasets Using Pretrained Embeddings and Transformers. *International Conference on Frontiers of Information Technology*.

Guijin Son, Dongkeun Yoon, Juyoung Suk, Javier Aula-Blasco, Mano Aslan, Vu Trong Kim, Shayekh Bin Islam, Jaume Prats-Cristià, Lucía Tormo-Bañuelos, and Seungone Kim. 2024. MM-Eval: A Multilingual Meta-Evaluation Benchmark for LLM-as-a-Judge and Reward Models. *arXiv preprint*. ArXiv:2410.17578.

Anirudh Srinivasan, Venkata S Govindarajan, and Kyle Mahowald. 2023. Counterfactually Probing Language Identity in Multilingual Models. *MRL*.

Felix Stahlberg, Danielle Saunders, and Bill Byrne. 2018. An operation sequence model for explainable neural machine translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 175–186, Brussels, Belgium. Association for Computational Linguistics.

Giulio Starace, Konstantinos Papakostas, Rochelle Choenni, Apostolos Panagiotopoulos, Matteo Rosati, Alina Leidinger, and Ekaterina Shutova. 2023. Probing LLMs for joint encoding of linguistic categories. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7158–7179, Singapore. Association for Computational Linguistics.

Victor Steinborn, Philipp Dufter, Haris Jabbar, and Hinrich Schuetze. 2022. An information-theoretic approach and dataset for probing gender stereotypes in multilingual masked language models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 921–932, Seattle, United States. Association for Computational Linguistics.

Marek Strong, Rami Aly, and Andreas Vlachos. 2024. Zero-shot fact verification via natural logic and large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17021–17035, Miami, Florida, USA. Association for Computational Linguistics.

Hao Sun and John Hewitt. 2023. Character-level Chinese backpack language models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 106–119, Singapore. Association for Computational Linguistics.

Haoran Sun, Renren Jin, Shaoyang Xu, Leiyu Pan, Supryadi, Menglong Cui, Jiangcun Du, Yikun Lei, Lei Yang, Ling Shi, Juesi Xiao, Shaolin Zhu, and Deyi Xiong. 2024. FuxiTranyu: A multilingual large language model trained with balanced data. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1499–1522, Miami, Florida, US. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328, Sydney, Australia. PMLR. ISSN: 2640-3498.

Kerdkiat Suvirat, Detphop Tanasanchonnakul, Kanakorn Horsiritham, C. Kongkamol, Thammasin Ingviya, and Sitthichok Chaichulee. 2022. Automated Diagnosis Code Assignment of Thai Free-text Clinical Notes. *2022 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*.

Andrea Tagarelli and Andrea Simeri. 2021. Unsupervised Law Article Mining based on Deep Pre-Trained Language Representation Models with Application to the Italian Civil Code. *arXiv preprint*.

Ekaterina Taktasheva, Vladislav Mikhailov, and Ekaterina Artemova. 2021. Shaking syntactic trees on the sesame street: Multilingual probing with controllable perturbations. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 191–210, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-Specific Neurons: The Key to Multilingual Capabilities in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715, Bangkok, Thailand. Association for Computational Linguistics.

Shimin Tao, Su Chang, Ma Miaomiao, Hao Yang, Xiang Geng, Shujian Huang, Min Zhang, Jiaxin Guo, Minghan Wang, and Yinglu Li. 2022. CrossQE: HW-TSC 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 646–652, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Kushal Tatariya, Heather Lent, Johannes Bjerva, and Miryam de Lhoneux. 2024. Sociolinguistically informed interpretability: A case study on Hinglish emotion classification. In *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 66–74, St. Julian's, Malta. Association for Computational Linguistics.

Nguyen Ha Thanh, Vu Tran, Phuong Minh Nguyen, Le-Minh Nguyen, and Ken Satoh. 2023. How Fine Tuning Affects Contextual Embeddings: A Negative Result Explanation. *International Conference on Agents and Artificial Intelligence*.

Julia Thomas, Antonia Lucht, Jacob Segler, Richard Wundrack, M. Miché, R. Lieb, Lars Kuchinke, and Gunther Meinlschmidt. 2024. An Explainable Artificial Intelligence Text Classifier for Suicidality Prediction in Youth Crisis Text Line Users: Development and Validation Study. *JMIR Public Health and Surveillance*.

Baojie Tian, Liangjun Zang, Jizhong Han, and Songlin Hu. 2024. Capture Long-Range Dependency with Meta-Path Transformer for De-Anonymization of Q&A Sites. *International Conference on Computer Supported Cooperative Work in Design*.

Isidora Chara Tourni and Derry Tanti Wijaya. 2023. Relevance-guided Neural Machine Translation. *arXiv.org*.

Marcos Treviso, Nuno M. Guerreiro, Ricardo Rei, and André F. T. Martins. 2021. IST-unbabel 2021 submission for the explainable quality estimation shared task. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 133–145, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dennis Ulmer. 2024. On Uncertainty In Natural Language Processing. *arXiv.org*.

Saiteja Utpala, Alex Gu, and Pin-Yu Chen. 2024. Language agnostic code embeddings. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 678–691, Mexico City, Mexico. Association for Computational Linguistics.

Alexandros Vasileiou and Oliver Eberle. 2024. Explaining text similarity in transformer models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7859–7873, Mexico City, Mexico. Association for Computational Linguistics.

Jesse Vig. 2019. A Multiscale Visualization of Attention in the Transformer Model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy. Association for Computational Linguistics.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Davorin Vukadin, A. S. Kurdija, G. Delač, and M. Šilić. 2021. Information Extraction From Free-Form CV Documents in Multiple Languages. *IEEE Access*.

Ivan Vulić, Goran Glavaš, Fangyu Liu, Nigel Collier, Edoardo Maria Ponti, and Anna Korhonen. 2023. Probing cross-lingual lexical knowledge from multilingual sentence encoders. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2089–2105, Dubrovnik, Croatia. Association for Computational Linguistics.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Lennart Wachowiak and Dagmar Gromann. 2022. Systematic analysis of image schemas in natural language through explainable multilingual neural language processing. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5571–5581, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Hetong Wang, Pasquale Minervini, and Edoardo Ponti. 2024a. Probing the emergence of cross-lingual alignment during LLM training. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12159–12173, Bangkok, Thailand. Association for Computational Linguistics.

Ke Wang, Yangbin Shi, Jiayi Wang, Yuqi Zhang, Yu Zhao, and Xiaolin Zheng. 2021. Beyond glass-box features: Uncertainty quantification enhanced quality estimation for neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4687–4698, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Meiyun Wang, Kiyoshi Izumi, and Hiroki Sakaji. 2024b. LLMFactor: Extracting profitable factors through prompts for explainable stock movement prediction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3120–3131, Bangkok, Thailand. Association for Computational Linguistics.

Weixuan Wang, B. Haddow, Wei Peng, and Alexandra Birch. 2024c. Sharing Matters: Analysing Neurons Across Languages and Tasks in LLMs. *arXiv.org*.

Y. Wang, Jonas Pfeiffer, Nicolas Carion, Yann LeCun, and Aishwarya Kamath. 2023. Adapting Grounded Visual Question Answering Models to Low Resource Languages. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Yuqi Wang, Zimu Wang, Nijia Han, Wei Wang, Qi Chen, Haiyang Zhang, Yushan Pan, and Anh Nguyen. 2024d. Knowledge distillation from monolingual to multilingual models for intelligent and interpretable multilingual emotion detection. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 470–475, Bangkok, Thailand. Association for Computational Linguistics.

A. Wawer and Justyna Sarzyńska-Wawer. 2022. Detecting Deceptive Utterances Using Deep Pre-Trained Neural Networks. *Applied Sciences*.

Sabine Wehnert, Christian Scheel, Simona Szakács-Behling, Maret Nieländer, Patrick Mielke, and Ernesto William De Luca. 2021. HOTTER: Hierarchical optimal topic transport with explanatory context representations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4856–4866, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Andrea W Wen-Yi and David Mimno. 2023. Hyperpolyglot LLMs: Cross-lingual interpretability in token embeddings. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1124–1131, Singapore. Association for Computational Linguistics.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do Llamas Work in English? On the Latent Language of Multilingual Transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Tim Z. Xiao, Aidan N. Gomez, and Y. Gal. 2020. Wat zei je? Detecting Out-of-Distribution Translations with Variational Transformers. *arXiv.org*.

Chen-Wei Xie, Jianmin Wu, Yun Zheng, Pan Pan, and Xiansheng Hua. 2022. Token Embeddings Alignment for Cross-Modal Retrieval. *ACM Multimedia*.

Feng Xiong, Jun Wang, Geng Tu, and Ruifeng Xu. 2024. HITSZ-HLT at WASSA-2024 shared task 2: Language-agnostic multi-task learning for explainability of cross-lingual emotion detection. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 476–482, Bangkok, Thailand. Association for Computational Linguistics.

Liyan Xu, Xuchao Zhang, Xujiang Zhao, Haifeng Chen, Feng Chen, and Jinho D. Choi. 2021. Boosting cross-lingual transfer via self-learning with uncertainty estimation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6716–6723, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ningyu Xu, Qi Zhang, Jingting Ye, Menghan Zhang, and Xuanjing Huang. 2023. Are structural concepts universal in transformer language models? towards interpretable cross-lingual generalization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13951–13976, Singapore. Association for Computational Linguistics.

Shanshan Xu and Katja Markert. 2022. The Chinese causative-passive homonymy disambiguation: an adversarial dataset for NLI and a probing task. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4316–4323, Marseille, France. European Language Resources Association.

Sihao Xu, Wei Zhang, and Fan Zhang. 2020. Multi-Granular BERT: An Interpretable Model Applicable to Internet-of-Thing devices. *2020 IEEE International Conference on Energy Internet (ICEI)*.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin XU, Yuqi Ye, and Hanwen Gu. 2024. A Survey on Multilingual Large Language Models: Corpora, Alignment, and Bias. ArXiv:2404.00929 [cs].

Yuzi Yan, J. Li, Yipin Zhang, and Dong Yan. 2024. Exploring the LLM Journey from Cognition to Expression with Linear Representations. *International Conference on Machine Learning*.

Hao Yang, Min Zhang, Shimin Tao, Minghan Wang, Daimeng Wei, and Yanfei Jiang. 2023a. Knowledge-Prompted Estimator: A Novel Approach to Explainable Machine Translation Assessment. *International Conference on Advanced Communication Technology*.

Xianjun Yang, Wei Cheng, Linda Petzold, William Yang Wang, and Haifeng Chen. 2023b. DNA-GPT: Divergent N-Gram Analysis for Training-Free Detection of GPT-Generated Text. *International Conference on Learning Representations*.

Cuicui Ye, Jing Yang, and Yan Mao. 2024a. FDHFUI: Fusing Deep Representation and Hand-Crafted Features for User Identification. *IEEE transactions on consumer electronics*.

Jingheng Ye, Shang Qin, Yinghui Li, Xuxin Cheng, Libo Qin, Hai-Tao Zheng, Peng Xing, Zishan Xu, Guo Cheng, and Zhao Wei. 2024b. EXCGEC: A Benchmark of Edit-wise Explainable Chinese Grammatical Error Correction. *arXiv.org*.

Zi yu Chen, Fei Xiao, Xiao kang Wang, Wen hui Hou, Rui Huang, and Jian qiang Wang. 2023. An interpretable diagnostic approach for lung cancer: Combining maximal clique and improved BERT. *Expert Syst. J. Knowl. Eng.*

Majid Zarharan, Maryam Hashemi, Malika Behroozrazegh, Sauleh Eetemadi, Mohammad Taher Pilehvar, and Jennifer Foster. 2025. FarExStance: Explainable stance detection for Farsi. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10125–10147, Abu Dhabi, UAE. Association for Computational Linguistics.

Hongchuan Zeng, Senyu Han, Lu Chen, and Kai Yu. 2025. Converging to a lingua franca: Evolution of linguistic regions and semantics alignment in multilingual large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10602–10617, Abu Dhabi, UAE. Association for Computational Linguistics.

Yirong Zeng, Xiao Ding, Yi Zhao, Xiangyu Li, Jie Zhang, Chao Yao, Ting Liu, and Bing Qin. 2024. RU22Fact: Optimizing evidence for multilingual explainable fact-checking on Russia-Ukraine conflict. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14215–14226, Torino, Italia. ELRA and ICCL.

Wei Zhai, Nan Bai, Qing Zhao, Jianqiang Li, Fan Wang, Hongzhi Qi, Meng Jiang, Xiaoqin Wang, Bing Xiang Yang, and Guanghui Fu. 2024. MentalGLM Series: Explainable Large Language Models for Mental Health Analysis on Chinese Social Media. *arXiv.org*.

Ruochen Zhang, Qinan Yu, Matianyu Zang, Carsten Eickhoff, and Ellie Pavlick. 2024a. The Same But Different: Structural Similarities and Differences in Multilingual Language Modeling. *arXiv.org*.

Shimao Zhang, Changjiang Gao, Wenhao Zhu, Jiajun Chen, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian Huang. 2024b. Getting more from less: Large language models are good spontaneous multilingual learners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8037–8051, Miami, Florida, USA. Association for Computational Linguistics.

Xiaocheng Zhang, Xi Wang, Yifei Lu, Zhuangzhuang Ye, Jianing Wang, Mengjiao Bao, Peng Yan, and Xiaohong Su. 2024c. Augmenting the Veracity and Explanations of Complex Fact Checking via Iterative Self-Revision with LLMs. *arXiv.org*.

Ziyin Zhang, Yikang Liu, Weifang Huang, Junyu Mao, Rui Wang, and Hai Hu. 2024d. MELA: Multilingual evaluation of linguistic acceptability. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2658–2674, Bangkok, Thailand. Association for Computational Linguistics.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024a. Explainability for Large Language Models: A Survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):20:1–20:38.

Shuai Zhao, Qing Li, Yuer Yang, Jinming Wen, and Weiqing Luo. 2023. From Softmax to Nucleusmax: A Novel Sparse Language Model for Chinese Radiology Report Summarization. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*

Tiancheng Zhao, Xiaopeng Lu, and Kyusong Lee. 2021. SPARTA: Efficient open-domain question answering via sparse transformer matching retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 565–575, Online. Association for Computational Linguistics.

Xin Zhao, Naoki Yoshinaga, and Daisuke Oba. 2024b. Tracing the roots of facts in multilingual language models: Independent, shared, and transferred knowledge. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2088–2102, St. Julian's, Malta. Association for Computational Linguistics.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024c. How do Large Language Models Handle Multilingualism? In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, volume 37, pages 15296–15319, Vancouver, Canada. Curran Associates, Inc.

Zhixue Zhao and Nikolaos Aletras. 2024. Comparing explanation faithfulness between multilingual and monolingual fine-tuned language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3226–3244, Mexico City, Mexico. Association for Computational Linguistics.

Guorui Zheng, Xidong Wang, Juhao Liang, Nuo Chen, Yuping Zheng, and Benyou Wang. 2024. Efficiently Democratizing Medical LLMs for 50 Languages via a Mixture of Language Family Experts. *arXiv.org*.

Jianyu Zheng and Ying Liu. 2023. What does Chinese BERT learn about syntactic knowledge? *PeerJ Computer Science*.

Shaolin Zhu, Supryadi, Shaoyang Xu, Haoran Sun, Leiyu Pan, Menglong Cui, Jiangcun Du, Renren Jin, António Branco, and Deyi Xiong. 2024a. Multilingual Large Language Models: A Systematic Survey. *arXiv preprint*. ArXiv:2411.11072 [cs].

Wenhao Zhu, Sizhe Liu, Shujian Huang, Shuaijie She, Chris Wendler, and Jiajun Chen. 2024b. Multilingual contrastive decoding via language-agnostic layers skipping. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8775–8782, Miami, Florida, USA. Association for Computational Linguistics.

Judit Ács, Endre Hamerlik, Roy Schwartz, Noah A. Smith, and András Kornai. 2023. Morphosyntactic probing of multilingual BERT models. *Natural Language Engineering*.

Judit Ács, D'aniel L'evai, D. Nemeskey, and András Kornai. 2021. Evaluating Contextualized Language Models for Hungarian. *arXiv.org*.

A Methodology

This section outlines the methodology employed to identify, review and categorise the literature relevant to this survey on the explainability and interpretability of multilingual LLMs. Our approach is designed as a literature survey, aiming to capture key trends, techniques and challenges in the field, rather than an exhaustive systematic review covering every publication. The methodology is inspired by recent work and good practices in the field (Liu et al., 2024a).

Paper Selection. We initiated the process by defining a set of search keywords targeting the three core concepts of our survey: "explainability" (and its synonyms like "interpretability"), "large language models" and "multilinguality". To ensure broad linguistic coverage within the multilinguality aspect, our keyword list included terms for every language reported by LinguaMeta (Ritchie et al., 2024) as having over one million speakers. This keyword set, detailed in Appendix K, was intentionally curated to be representative rather than exhaustive, aligning with our survey's objective as previously stated. Using these keywords, we performed searches querying the titles and abstracts of publications indexed in the ACL Anthology, arXiv and Semantic Scholar repositories via their respective APIs (Kinney et al., 2025). The search had a knowledge cut-off date of February 2025, adhering to the ACL Rolling Review's 3-month recency policy[1] relevant to our target submission timeline. This initial search yielded 721 candidate papers.

Paper Screening. The retrieved papers underwent a multi-stage selection process. First, we performed deduplication based on DOIs and titles, prioritising versions from the ACL Anthology where duplicates existed across sources. Second, the unique papers were subjected to a manual relevance screening conducted by the authors. To be included at this stage, a paper needed to substantially address the intersection of explainability/interpretability techniques with large language models in a multilingual context. This screening phase resulted in a shortlist of 250 relevant papers (98 from ACL Anthology). During this initial review, we also extracted pertinent keywords and annotations from each paper to aid subsequent analysis and categorisation.

[1] https://aclrollingreview.org/reviewerguidelines

Paper Categorisation. Based on the additional keywords identified during the initial screening, we developed the categorisation that structures this survey. The shortlisted papers then underwent a second review phase. The primary goals of this phase were to assign each paper to one or more relevant categories concerning explainability methods (Section 2), multilingual tasks (Section 3), languages (Appendix F) and resources (Appendix G), while also confirming their continued relevance to the survey's scope. Specific categories (e.g., feature attribution and NLP applications) have an expressive number of papers, therefore choosing the works to discuss in the main text versus the appendix leveraged paper's impact and venue. When paper categories are fuzzy (e.g., probing & latent representations, or attention & visualisation), we opted for the most foundational category. Finally, the distinction between high-mid- and low-resource languages was made based on the data ratio in Common Crawl[2][3], as in previous work (Lai et al., 2023; Son et al., 2024).

This structured process of search, screening and categorisation yielded the 226 papers forming the core basis for this survey. A small number of additional relevant papers were also included beyond this search process, either because they were published after our cut-off date or were identified through other means (e.g., organic discovery, recommendations from peers).

B Uncertainty Estimation

Uncertainty estimation in LLMs is vital for explainability due to their potential overconfidence. Uncertainty is studied in cross-lingual settings (Hashimoto et al., 2024; Xu et al., 2021), multilingual analysis (Ulmer, 2024), machine translation (MT) quality estimation and out-of-distribution detection (Wang et al., 2021; Xiao et al., 2020) and specific domains such as medicine (Kang et al., 2024; Ben-Atya et al., 2025). Krause et al. (2023) showed models can exhibit exaggerated cross-lingual confidence, while Ben-Atya et al. (2025) improved Hebrew radiology labelling by filtering uncertain samples. Methods include specific metrics adapted for cross-lingual transfer (e.g. LEU, LOU, EVI; Xu et al., 2021), data augmentation (Hashimoto et al., 2024), Monte Carlo dropout and ensembles (Abreu-Cardenas et al., 2023), Spectral-normalized Neural Gaussian Process (SNGP) (Kang et al., 2024) and using model-intrinsic signals in MT (Wang et al., 2021; Xiao et al., 2020).

Takeaways. Uncertainty estimation methods are applied across a diverse range of multilingual tasks. Nevertheless, many metrics are adaptations from monolingual to multilingual contexts, and recent research indicates that models demonstrate cross-lingual overconfidence (Krause et al., 2023). Key open questions include how this overconfidence affects low-resource languages and the development of effective mitigation strategies.

C Visual Explanations

Visual explanations significantly aid the interpretability of multilingual LLMs particularly through dimensionality reduction methods like t-SNE (Maaten and Hinton, 2008) and UMAP (McInnes et al., 2018). Embedding visualisation has proven valuable for analysing multilingual conversational routing (Maksymenko and Turuta, 2024), understanding speech emotion recognition in German and Romanian (Echim et al., 2024) and assessing embedding quality within Dialectal Arabic automatic speech recognition systems (Sahyoun and Shehata, 2023). This approach also extends to inspecting embeddings in French (Bogaert et al., 2023) and legal Italian (Tagarelli and Simeri, 2021) LLMs and for multilingual information extraction from curricula vitae (Vukadin et al., 2021). Furthermore dedicated visualisation tools, such as BERTViz (Vig, 2019), are employed to inspect attention mechanisms for tasks including information extraction (Vukadin et al., 2021), multilingual handwriting recognition (Ramo et al., 2023) and within legal domain models (Tagarelli and Simeri, 2021).

Takeaways. Embedding and attention visualisation are the predominant methods for visual explanations, increasingly applied across diverse modalities. Most studies, however, apply these to multilingual tasks rather than designing visualisations to explore multilinguality itself – e.g., visualising cross-lingual embedding spaces. Significant potential lies in the visual inspection of multilingual embeddings, such as to analyse resource levels and language-agnostic latent spaces, develop interactive visualisation tools and integrate these visual approaches with other interpretability methods.

[2] Over 0.1% as high-mid, otherwise low.

[3] http://commoncrawl.org/

D Speech Processing

Understanding model behaviour in complex multilingual speech processing tasks necessitates diverse explainability techniques. For instance Mohebbi et al. (2023) probe Transformers for French speech homophony information, while other studies perform layer-wise probing of suprasegmentals and lexical tone in Mandarin (de la Fuente and Jurafsky, 2024) or investigate lexical tone encoding in Mandarin and Vietnamese spoken language models (Shen et al., 2024). Further probing analyses examine aspects like architectural bias for French in Whisper's multilingual transcription (Ballier et al., 2024) and identify key layers for multilingual speech emotion recognition (Singh and Gupta, 2023). Beyond probing, research explores self-attention mechanisms in cross-lingual self-supervised speech models (Gopinath and Rodriguez, 2024), the nature of latent spaces in multilingual speech translation (Abdullah et al., 2024) and applies various explainability methods, for instance to speech emotion recognition in German and Romanian (Echim et al., 2024) and to the study of speech systems for Egyptian Arabic (Sahyoun and Shehata, 2023).

Takeaways. While probing, particularly layer-wise analysis, is a prevalent method for investigating linguistic information in speech models, there is also a growing utilisation of other XAI techniques. Speech data offer a unique source of multilingual information, such as spoken variations (e.g., dialects, accents) typically absent in textual data; this characteristic presents promising avenues for future research. Such data also facilitate more language-specific analyses, encompassing spoken-language-specific features (e.g., lexical tone, suprasegmentals), with potential for employing mechanistic approaches and XAI methods tailored to the distinct properties of speech. Nevertheless, the scarcity of high-quality spoken data, especially for low-resource languages, is evident in current literature and constitutes a significant research gap.

E Bias and Toxicity

Explainability methods are crucial for addressing bias and toxicity in multilingual LLMs. Counterfactual analysis, for instance, reveals nationality biases in various languages including Maori and Basque by perturbing inputs (Barriere and Cifuentes, 2024b,a). Mechanistic interpretability also explains how toxicity mitigation via preference tuning in English can generalise cross-lingually (Li et al., 2024c). Further explainability efforts include probing for gender stereotypes in multiple languages (Steinborn et al., 2022) and architectural biases for French (Ballier et al., 2024); analysing interpretable representations or character senses to address gender bias in languages like Chinese (Sun and Hewitt, 2023; Liang et al., 2020); developing bias attribution metrics, for instance within Southeast Asian LLMs (Gamboa and Lee, 2024); and examining subword impacts on NER bias (Calix et al., 2022).

Takeaways. A diverse range of explanation methods are utilised to investigate biases within multilingual LLMs, including those related to gender and nationality, thereby informing mitigation strategies. Nevertheless, further research is crucial in three principal areas: cross-language bias (e.g., towards English and other high-resource languages), intra-language bias (e.g., between regions and dialects), and extra-language bias (e.g., towards cultural and moral values). The last of these is especially pertinent for preference-tuned models, where an understanding of such biases can shape mitigation efforts and enhance fairness across diverse cultural contexts.

F Explainability of Languages

This section analyses the surveyed languages, categorised as high-mid-resource, low-resource and non-natural languages. Due to the volume of papers, discussion is limited to representative works.

F.1 High-mid-resource

Within high and medium resource languages, studies frequently involve Chinese (30% of such works), German (16%) and French (10%). These often serve as testbeds for diverse explainability techniques, such as explainable stock movement prediction in Chinese (Wang et al., 2024b), comparing human and model attention multilingually (Brandl et al., 2024) or probing cross-lingual generalisation (Aghazadeh et al., 2022), commonly prioritising method or task over language-specific insights.

Some research, however, targets language-specific features, like the encoding of tonal information in Mandarin (Shen et al., 2024; de la Fuente and Jurafsky, 2024). Other works use languages as case studies, for example, analysing French writing style effects on embeddings (Icard et al., 2025) or the mechanistic interpretability of Spanish numbers (Lan et al., 2024; Ferrando and Costa-jussà, 2024).

New datasets with interpretability features are also developed, e.g. for Persian stance detection or Japanese QA (Zarharan et al., 2025; Ishii et al., 2024). Domain-specific analyses with explainability are also prevalent, spanning French legal applications (Louis et al., 2023), Korean mental health (Kang et al., 2024) and diverse Chinese financial or medical contexts (Wang et al., 2024b; Li et al., 2023b; Chen et al., 2024b).

Key Takeaway. High-mid languages are mostly used as case studies and multilingual test datasets for any task that is "non-English". More research is needed to explore the specificities of languages regarding explainability.

F.2 Low-resource

Low-resource languages (e.g. Tamil and Basque) are underrepresented, constituting only 8% of reviewed papers. Analytical work includes probing robustness in Indic languages (Aravapalli et al., 2024), phonetics in Nordic languages like Faroese (Agirrezabal et al., 2023) and morphosyntax in languages such as Marathi and Yoruba (Shapiro et al., 2021), alongside counterfactual bias detection in Maori and Basque (Barriere and Cifuentes, 2024b). Li et al. (2024a) also note a probing performance gap for these languages compared to high resource ones, with the latter exhibiting greater representational similarity among themselves.

Much research applies established feature attribution methods (e.g. LIME and SHAP) to NLP tasks like hate speech detection in Roman Urdu and Sindhi (Hashmi et al., 2024b; Sohail et al., 2024; Siddiqui et al., 2024) or sentiment analysis and machine translation for African languages (Mabokela et al., 2024; Malinga et al., 2024; Kobusingye et al., 2023). Such work often involves straightforward applications of existing techniques, particularly for African and Indic languages, and typically appears in less impactful venues. Barriere and Cifuentes (2024b) on Basque offers a notable exception in methodology and venue.

Key Takeaway. Low-resource languages are underrepresented in the literature, with a focus on simple applications of existing techniques and less impactful venues. Research needs to be more methodologically advanced.

F.3 Non-Natural Languages

Non-natural languages, including sign and programming languages, also feature in explainability research. For programming languages, Utpala et al. (2024) analyse code embeddings, identifying language-agnostic and language-specific components. Liu et al. (2024c) employ probing to evaluate fine-tuning strategies for code comprehension. In the realm of artificial languages, Ri and Tsuruoka (2022) design a language mimicking natural linguistic structures – pre-training and subsequent probing reveal that successful transfer to natural languages correlates with encoded contextual information. Finally, explainability in sign language processing is explored using feature attribution (LIME) for Arabic Sign Language (Baghdadi et al., 2024), attention analysis for Greek Sign Language (Bianco et al., 2024) and attention feature visualisation for American Sign Language (Ananthanarayana et al., 2021).

Key Takeaway. Artificial languages are a promising avenue for interpreting cross-lingual transfer, due to their potential to mimic natural languages and facilitate probing of desired features.

G Resources for Explainability

Interpretability resources are crucial for the development and application of explainability methods, spanning evaluation, techniques and metrics. Due to the large number of papers, we focus on the most relevant resources in each category.

G.1 Evaluation

Evaluation resources encompass benchmarks, datasets and human studies. Datasets provide support for NLEs in multilingual applications (e.g. fact-checking or domain-specific uses; Zeng et al., 2024; Louis et al., 2023; Li et al., 2023b), multilingual human rationales (Jørgensen et al., 2022; Pachinger et al., 2024), counterfactuals for bias detection (Barriere and Cifuentes, 2024a,b) and multilingual probing resources (Zhang et al., 2024d; Steinborn et al., 2022). Attanasio et al. (2022) propose a benchmark for hate speech interpretability approaches in English and Italian, while Park and Padó (2024) target interpretable MT quality estimation. Importantly, datasets often aid interpretation extraction rather than evaluating explanations directly (Attanasio et al., 2022).

Human evaluation explores novel data sources like webcam gaze for multilingual QA, comparable to human rationales (Brandl et al., 2024), and contrasts human with neural attention for Russian anaphora resolution (Kozlova et al., 2024). NLE quality is often human-judged across languages (e.g. Persian, Korean, Chinese and Greek; Zarharan et al., 2025; Lee and Han, 2024; Ye et al., 2024b; Mylonas et al., 2024). GUI-based systems also support multilingual linguistic probing (Serikov et al., 2022).

Key Takeaway. Evaluation resources are diverse, but most focus on enabling interpretation extraction rather than evaluating the actual explanation.

G.2 Explainability Techniques

Standard explainability techniques are widely used to interpret multilingual models. Feature attribution methods like LIME (Ribeiro et al., 2016), SHAP (Lundberg and Lee, 2017) and integrated gradients (Sundararajan et al., 2017) are prevalent in NLP and domain-specific applications (Section 3.5), for instance, in multilingual sentiment analysis (Jørgensen et al., 2022) or code-mixed text handling (Mamta et al., 2023). Their faithfulness is compared across multilingual and monolingual models in Zhao and Aletras's work (2024). Layer-wise relevance propagation (Binder et al., 2016) guides machine translation of low-resource languages (Tourni and Wijaya, 2023) and explains text similarity (Vasileiou and Eberle, 2024), while LIME has seen task-specific adaptations (Guo et al., 2024; Rodrigues and Marcacini, 2022).

Other approaches include uncertainty quantification techniques, for instance using SNGP for mental disorder detection in Korean (Kang et al., 2024) or specific methods for multilingual knowledge neuron localisation (Cao et al., 2024). Intrinsic probing techniques identify linguistic neurons (Wang et al., 2024a) and influence functions study generalisation (Grosse et al., 2023). Visualisation tools like t-SNE (Maaten and Hinton, 2008), UMAP (McInnes et al., 2018) and BERTViz (Vig, 2019) are also employed, for example, in analysing Italian legal models (Tagarelli and Simeri, 2021) or Chinese medical systems (Lin et al., 2024).

Key Takeaway. The frequent application of techniques in NLP applications shows their popularity and effectiveness, but also points out the simplicity of the analyses.

G.3 Metrics

Metrics are an important resource for evaluating explanations or measuring other properties in an interpretable way. Key explanation metrics include faithfulness (how explanations reflect model behaviour), applied when comparing multilingual models (Zhao and Aletras, 2024) or for Bengali hate speech (Karim et al., 2020), and plausibility (human understandability), used for multilingual sentiment analysis (Jørgensen et al., 2022). Others are sufficiency, compactness and consistency (Shen et al., 2022), automatic NLE metrics for Korean or Chinese (Lee and Han, 2024; Ye et al., 2024b) and sensitivity (Bogaert et al., 2024).

Beyond direct explanation evaluation, uncertainty quantification metrics are used for cross-lingual transfer and MT quality estimation (Xu et al., 2021; Wang et al., 2021). Specific metrics assess multilingual gender bias via probing (Steinborn et al., 2022) or token-level bias contributions (Gamboa and Lee, 2024). Sparsity, indicative of interpretability in Chinese QA (Zhao et al., 2021), is also measured. Furthermore, some task-specific metrics are designed for enhanced interpretability, such as for MT quality estimation (Park and Padó, 2024) or Arabic ASR (Sahyoun and Shehata, 2023).

Key Takeaway. While diverse metrics assess explanations or interpretable properties, they are often not explicitly designed for multilingual models.

H Additional Probing Papers

Additional probing studies further illuminate the capabilities of multilingual models and complement the overview in Section 2.1.

Applications extend across diverse multilingual areas including speech processing (Mohebbi et al., 2023), the encoding of multilingual sociodemographic knowledge across layers (Lauscher et al., 2022) and multilingual temporal relations (Caselli et al., 2022). The technique is also adapted for non-natural languages such as code understanding in multilingual scenarios (Liu et al., 2024c).

Morphosyntactic knowledge is extensively analysed, for instance, by probing for Universal Dependency features across many languages (Serikov et al., 2022) or using multilabel approaches for diverse languages (Shapiro et al., 2021). Furthermore, phonological information in character embeddings with cross-lingual analysis (Boldsen et al., 2022), character-level encoding across various alphabets (Kaushal and Mahowald, 2022) and lexical knowledge across diverse languages (Vulić et al., 2020) are frequently probed.

Investigations into speech and phonology include probing lexical tone encoding in Mandarin and Vietnamese (Shen et al., 2024), Whisper's ASR representations for French, English and Persian (Ballier et al., 2024), multilingual speech emotion recognition (Singh and Gupta, 2023), Mandarin and English suprasegmentals in speech models (de la Fuente and Jurafsky, 2024) and phonetic encoding in character-based models for Nordic languages (Agirrezabal et al., 2023).

Numerous works probe syntactic and morphological knowledge. For instance, studies examine Chinese BERT's syntactic knowledge (Zheng and Liu, 2023), leverage multilingual morphological datasets (Ács et al., 2023) and assess BERT's handling of Italian learner errors alongside general linguistic knowledge (Miaschi et al., 2023b,a). Other research probes coreferential relationships in Dutch BERT (De Langhe et al., 2023), cross-lingual syntax in English and Mandarin (Chen and Farrús, 2022) and morphology in Hungarian models (Ács et al., 2021). Further investigations cover morphosyntactic content across Indo-European languages (Mikhailov et al., 2021a), sensitivity to word order perturbations in English, Swedish and Russian (Taktasheva et al., 2021), syntactic evaluation using benchmarks for Indic languages (Patil et al., 2021), mBERT's syntactic capabilities (Rönnqvist et al., 2019) and the link between tokenisation strategies and morphology in models like mT5 and ByT5 (Dang et al., 2024).

Broader linguistic understanding and cross-lingual phenomena are also common targets such as the robustness of models for Indic languages under perturbation (Aravapalli et al., 2024), form versus meaning representation in Chinese and German from a neurolinguistic perspective (He et al., 2024), performance disparities across high- and low-resource languages in multilingual probing (Li et al., 2024a), how probing results reflect linguistic classifications (Mysiak and Cyranka, 2023), diverse linguistic properties in Russian (Mikhailov et al., 2021b) and linguistic feature capture in multilingual neural machine translation (Mareček et al., 2020).

Finally, probing extends to specific applications and domains like logical propositions in English and Spanish contexts (Feng et al., 2024), assessing discourse relation knowledge for Chinese-English translation (Huang et al., 2023), evaluating Llama's multilingual abilities (Chen et al., 2023) and analysing linguistic knowledge of LLMs in Italian (Miaschi et al., 2022).

I Additional Feature Attribution and NLP Applications Papers

There is a substantial body of work on feature attribution and NLP applications, including domain-specific ones, within multilingual contexts. Many studies also represent an intersection between these areas. This section expands upon subsections 2.5 and 3.5 by cataloguing additional relevant papers.

I.1 Feature Attribution

Further studies on feature attribution offer diverse insights into model interpretability. General evaluations and benchmarks are crucial; for instance, Brandl and Eberle (2024) compare NLEs with Layer-wise Relevance Propagation (LRP) for multilingual text classification in English, Danish and Italian, while Shen et al. (2022) propose benchmarks with token-level rationales for English and Chinese LLMs using methods like attention and Integrated Gradients (IG). Bayesian methods with LIME adaptation have been developed for disturbed Chinese sentence pair matching (Guo et al., 2024). The faithfulness of feature importance explanations across monolingual and multilingual models also remains a key research area, explored by Zhao and Aletras (2024).

Machine translation (MT) and quality estimation (QE) are common application areas for feature attribution. In MT, for example, these methods contribute to interpretable quality estimation for English-Korean (Park and Padó, 2024), enable tracking of source and target token contributions in multilingual MT (Ferrando et al., 2022) and support the development of self-explanatory MT for language pairs like Japanese-English (Stahlberg et al., 2018). IG has been employed to explain transliteration models for low-resource Indian languages such as Tamil (Islam et al., 2024). Explainable QE benefits from methods generating token-level scores from XLM-R (Tao et al., 2022) and from ensemble approaches across various language pairs including Estonian-English and Russian-German (Kabir and Carpuat, 2021; Treviso et al., 2021). Relevance-guided training has also been explored for neural MT involving French, Gujarati and Kazakh, particularly in low-resource settings (Tourni and Wijaya, 2023).

Understanding specific model behaviours, such as bias or knowledge encoding, is another significant focus. Metrics for token-level bias attribution are proposed for multilingual Southeast Asian LLMs (Gamboa and Lee, 2024). Subword impact analysis helps explain cross-lingual Named Entity Recognition (NER) and bias for languages like Saisiyat (Calix et al., 2022). Methods like MATRICE, using IG, quantify uncertainty in localising language-agnostic knowledge neurons in Chinese and other languages (Cao et al., 2024). Influence functions have been scaled to study LLM generalisation, including cross-lingual aspects (Grosse et al., 2023).

A variety of specific feature attribution techniques are applied broadly. SHAP aids in interpreting quantum transfer learning for Italian acceptability judgements (Buonaiuto et al., 2024) and LRP is used to study the effects of fine-tuning French CamemBERT (Bogaert et al., 2023). Gradient-based attribution helps analyse Arabic singular-to-plural conversion models (Radman et al., 2023) and SHAP combined with BERTViz explains Luganda-English MT (Kobusingye et al., 2023). Linguistic feature analysis provides explainability for age recommendation systems based on French texts (Rahman et al., 2023). Perturbation analysis and Shapley values assist in locating disambiguating information for multilingual morphosyntactic probing across numerous languages (Ács et al., 2023). LIME extensions are developed for Portuguese sentence similarity from meta-embeddings (Rodrigues and Marcacini, 2022) and multilingual features are incorporated into interpretable first-order logic frameworks for entity linking (Lu et al., 2022). Visual explanation methods like Grad-CAM++ alongside t-SNE are applied to speech emotion recognition in German and Romanian (Echim et al., 2024).

I.2 NLP Applications

Explainability research in general NLP applications continues to expand. For instance, multilingual jailbreak benchmarks are being developed that include NLEs (Liu et al., 2024b) and analyses of neuron activation investigate parallel multilingual learning within LLMs by translating input to multiple languages (Mu et al., 2024). Fact-checking in Chinese has been augmented with NLEs generated via iterative self-revision (Zhang et al., 2024c) and benchmarks for explainable Chinese grammatical error correction are also being created (Ye et al., 2024b).

Research also explores improving multilingual reasoning via interpretability-inspired contrastive decoding (Zhu et al., 2024b), tracing sources of multilingual factual knowledge through neuron activation and data attribution (Zhao et al., 2024b) and understanding internal representations of bilingual models (Yan et al., 2024). Other studies analyse French writing style effects in embeddings (Icard et al., 2025), use NLEs for multilingual norm discovery (Fung et al., 2022) and probe multilingual temporal relations (Caselli et al., 2022).

Attention visualisation is a common technique, used for analysing multilingual jailbreak patterns to inform mitigation strategies (Li et al., 2024b) and for interpreting Transformer models in the context of Greek Sign Language translation (Bianco et al., 2024). Explainable systems are also being built for Arabic fact-checking with NLE generation (Althabiti et al., 2024). Interpretable conversation routing using latent embeddings is being applied to multilingual datasets (Maksymenko and Turuta, 2024) and language-specific calibration for pruning multilingual LLMs for monolingual applications is studied via latent subspaces and neuron activation patterns (Kurz et al., 2024).

Cross-lingual emotion detection tasks benefit from NLEs and agentic workflows (Cheng et al., 2024). Uncertainty estimation methods are applied to tasks like complex text detection in Spanish (Abreu-Cardenas et al., 2023) and for multilingual question answering across diverse languages including Amharic (Krause et al., 2023). Visualisation techniques offer insights into multilingual Transformer models for applications like online handwriting decoding (Ramo et al., 2023). Interpretable structured sentiment analysis is explored using multilingual models such as ERNIE-M (Jia et al., 2022). Attention matrices are used to interpret Russian sentiment analysis models (Pashchenko et al., 2022) and broader explainability analyses are conducted for multilingual machine reading comprehension models (Cui et al., 2021).

I.3 Domain-Specific Applications

In various domain-specific contexts, explainability is proving crucial. Medical applications are prominent, with uncertainty estimation enhancing Hebrew radiology report labelling through agent-based models (Ben-Atya et al., 2025) and mechanistic interpretability guiding the development of efficient medical LLMs for up to 50 languages by analysing internal information flow (Zheng et al., 2024). Retrieval-augmented LLMs aid Chinese health rumour detection by providing NLEs (Chen et al., 2024a) and Chinese medical LLM responses are improved with explainable knowledge graphs (Jiang et al., 2023). Explainable models are also used for mental health analysis on Chinese social media, supported by new datasets (Zhai et al., 2024). Furthermore, attention visualisation helps interpret the deidentification of Chinese-English mixed clinical text (Lee et al., 2023) and counterfactual explanations support pulmonary disease diagnosis in Chinese (Li et al., 2023c). Attention patterns have also been analysed in BERT for Italian medical report classification (Putelli et al., 2022).

Specific content applications include datasets with rationales for Austrian German offensive language in news comments (Pachinger et al., 2024), explainable Korean SMS phishing detection (Lee and Han, 2024) and Chinese humor response datasets with "chain-of-humor" annotations (Chen et al., 2024b). Educational tools offer explainable German document retrieval (Wehnert et al., 2021) and, for programming languages, code analysis identifies language-specific and -agnostic embedding components (Utpala et al., 2024).

Text classification pipelines with NLE generation are tested on Greek tweets for sentiment analysis and offensive language identification (Mylonas et al., 2024). LLMs assist educators in grading student answers in German using rubrics as explanations (Metzler et al., 2024). Fine-tuning effects on contextual embeddings are analysed for legal Transformers (Thanh et al., 2023). Sparse language models aim to improve the interpretability of Chinese radiology report summarisation (Zhao et al., 2023). Multilingual CV information extraction uses attention and representation visualisation (Vukadin et al., 2021) and Italian legal BERT models (LamBERTa) are analysed using BERTViz and embedding visualisation (Tagarelli and Simeri, 2021).

For non-natural languages and specialised tasks, probing explains PEFT efficacy in cross-lingual code change learning (Liu et al., 2024c). Interpretable multi-granular BERT, converting character-level to word-level, is applied to Chinese IoT text classification improving self-attention interpretability (Xu et al., 2020).

I.4 Feature Attribution in NLP Applications

Many studies directly apply feature attribution techniques to a wide array of general NLP applications, enhancing their transparency. For instance, GLIDER serves as an LLM-as-judge evaluator offering multilingual reasoning and explainable span highlighting (Deshpande et al., 2024). LIME is frequently used, for example, to understand Transformer predictions for hate speech detection in Roman Urdu (Sohail et al., 2024). Interlanguage error features are designed to improve interpretability in the automated scoring of Chinese HSK essays (Rao and He, 2024).

Other applications are AI-generated text detection in German explained via text regeneration differences (Yang et al., 2023b), understanding code-mixed data handling via SHAP for auditory features (Mamta et al., 2023), assessing LIME and SHAP plausibility for multilingual sentiment analysis (Jørgensen et al., 2022), employing Integrated Gradients for sentiment analysis in various African low-resource contexts (Malinga et al., 2024) and using LIME for German image schema prediction from text (Wachowiak and Gromann, 2022).

The application of LIME extends to broad multilingual hate speech detection efforts covering languages such as Chinese, Spanish, Urdu, Portuguese, Indonesian, German and Italian (Hashmi et al., 2024b). Both LIME and SHAP are employed for interpreting Arabic semantic search models in the context of Quranic text (Mustafa et al., 2024). Attention-based attribution methods are utilised to explain de-anonymization processes in bilingual (Chinese-English) QA sites that use GNNs and Transformers (Tian et al., 2024). LIME also helps interpret English and Italian Transformers for misogyny detection tasks (Hashmi et al., 2024a) and explains Vision Transformers for Arabic sign language recognition (Baghdadi et al., 2024).

Knowledge distillation techniques aim to improve the identification of emotion-trigger words in multilingual models like XLM-R and E5, thus enhancing interpretability (Wang et al., 2024d). LIME is further used to study sociolinguistic biases in Hinglish (Hindi-English code-mixed) emotion classification (Tatariya et al., 2024). For user identification in Chinese, hand-crafted features are combined with mBERT to improve interpretability (Ye et al., 2024a). Comparative studies, for example between LIME and SHAP, assess methods for AI-generated text detection in German (Irrgang et al., 2024). LIME and SHAP are also used to explain Afrocentric and mainstream LLMs in sentiment analysis for low-resource South African languages (Mabokela et al., 2024) and LIME helps generate adversarial examples for Arabic offensive language detection systems (Abdelaty and Lazem, 2024).

Counterfactual attribution methods explain RAG systems for conversational QA over heterogeneous data, including German content (Roy et al., 2024). Coreference-driven feature attribution aids in the detection of harmful erotic content in Polish texts (Okulska and Wiśnios, 2023). LSTM models with attention mechanisms provide explanations for cross-lingual sentiment analysis with Transformers involving Persian (Ghasemi and Momtazi, 2023). Interpretable bounding boxes are provided for key phrases in multilingual Visual Question Answering (VQA) tasks involving languages such as Bengali, Portuguese and Indonesian (Wang et al., 2023). SHAP examines explainability in automated essay cohesion prediction for Portuguese and English (Oliveira et al., 2023). Unsupervised self-explainable frameworks using recursive dynamic gating can provide text explanations for machine reading comprehension in English and Chinese (Cui et al., 2022). IG interpret Transformer predictions for lie detection in Polish (Wawer and Sarzyńska-Wawer, 2022). Token embedding alignment coupled with visualisation techniques explains cross-modal retrieval in Chinese (Xie et al., 2022). Post-hoc token attribution methods like Gradient, IG, SHAP and Sampling-and-Occlusion have been benchmarked for misogyny detection tasks in English and Italian (Attanasio et al., 2022). LIME and IG have also been adapted for zero-shot offensive span identification in code-mixed Tamil (Ravikiran and Chakravarthi, 2022). Attention feature visualisation explains feature contributions in American Sign Language translation (Ananthanarayana et al., 2021). Finally, sensitivity analysis and LRP are employed to explain hate speech detection in the Bengali language (Karim et al., 2020).

I.5 Feature Attribution in Domain-Specific Applications

Feature attribution is also pervasively used to explain models in various domain-specific multilingual applications. In the context of social media analysis for public interest, language-agnostic multi-task learning frameworks identify binary trigger words for emotion detection in tweets across multiple languages (Xiong et al., 2024). LIME explains hybrid Transformer models designed for classifying asthma-related Arabic social media posts (Hossain et al., 2024a) and is also applied to models for general Arabic news classification (Hossain et al., 2024b).

Medical and health-related NLP frequently employs feature attribution. SHAP helps investigate suicidality prediction from German crisis helpline texts (Thomas et al., 2024) and is used in studies on explainable satirical news detection in Turkish (Ozturk et al., 2024). LIME aids in the analysis of models for detecting depression in Bengali social media text (Chowdhury et al., 2024). For clinical information extraction, SHAP validates data quality and model selection for German texts (Richter-Pechanski et al., 2024). Input perturbation techniques interpret a Chinese BERT-based medical triage system (Lin et al., 2024). SHAP also helps decode patterns in Italian political news headlines (Berta et al., 2024). An improved BERT model using attention mechanisms explains lung cancer diagnosis from Chinese electronic medical records (yu Chen et al., 2023). LIME is used to interpret XLM-R models for depression classification based on Thai speech transcriptions (Munthuli et al., 2023) and methods like LIME, SHAP and IG are used to compare Dutch medical report classifiers with domain expert explanations (Rietberg et al., 2023). LIME also interprets models predicting COVID-19 symptoms from Brazilian Portuguese tweets (Machado et al., 2022) and tools like `transformers-interpret` highlight relevant words for medical International Classification of Diseases (ICD) code assignment from Thai patient records using mBERT (Suvirat et al., 2022).

In other specialised domains, LIME, SHAP and DeepLIFT explain Transformer models for multilingual cooking recipe classification, with a focus on low-resource languages (Migea et al., 2024). LRP is used to study the sensitivity of explanations to random seeds in French journalistic text classification (Bogaert et al., 2024). SHAP also identifies important keywords for predicting political leanings from Slovenian parliamentary transcriptions (Evkoski and Pollak, 2023).

J Additional NLE Papers

Further research into Natural Language Explanations (NLEs) spans various applications and languages, expanding the insights from Section 2.8.

Advancing NLEs relies on specialised datasets:

for Persian stance detection with extractive explanations (Zarharan et al., 2025); for Chinese applications like humor responses with "chain-of-humor" (Chen et al., 2024b), medical explanations (Li et al., 2023b) and stock prediction using NL "factors" (Wang et al., 2024b); for multilingual fact-checking (e.g. Russia-Ukraine conflict; Zeng et al., 2024); and for French legal question-answering with rationales rooted in legal provisions (Louis et al., 2023).

Studies include the use of LLMs to assist educators with rubrics for grading student answers, primarily in English and German contexts (Metzler et al., 2024), and the development of a jailbreak benchmark featuring multilingual samples and explanations (Liu et al., 2024b). In machine translation, NLEs contribute to interpretable metrics for evaluating literary translations into Korean (Shafayat et al., 2024) and enhancing quality estimation through knowledge-prompted CoT (Yang et al., 2023a). The domain of mental health benefits from NLEs in analysing Chinese social media content, supported by new datasets and model explanations (Zhai et al., 2024).

Fact-checking systems increasingly incorporate NLEs, for example in an Arabic system that generates justifications (Althabiti et al., 2024) and a framework for complex Chinese fact-checking using iterative self-revision with LLMs to produce explanations (Zhang et al., 2024c). NLEs are also integrated into tools for grammatical error correction, with benchmarks for Chinese that include edit-wise explanations (Ye et al., 2024b) and for text classification in Greek where NLEs are evaluated via user studies (Mylonas et al., 2024). Other applications include retrieval-augmented LLMs for Chinese health rumour detection providing referenced answers (Chen et al., 2024a), NLEs in cross-lingual emotion detection tasks (Cheng et al., 2024) and frameworks enhancing medical LLM responses in Chinese with hypothesis knowledge graphs to improve explainability (Jiang et al., 2023).

K Search Keywords

The following list contains the keywords used to search for papers in the repositories. For details on the search methodology, please refer to Appendix A.

Explainability Keywords. "explainability", "explainable", "interpretability", "interpretable", "feature importance", "feature attribution", "counterfactual", "probing", "neuron activity", "neuron activation", "mechanistic", "circuit", "representation engineering", "uncertainty".

LLM Keywords. "language model", "llm", "transformer".

Multilinguality Keywords. "multilingual", "multilinguality", "multilingualism", "cross-lingual", "cross-linguality", "mandarin", "chinese", "hindi", "spanish", "arabic", "urdu", "bengali", "portuguese", "french", "punjabi", "swahili", "indonesian", "russian", "japanese", "western panjabi", "telugu", "lahnda", "marathi", "german", "javanese", "vietnamese", "wu chinese", "persian", "caribbean javanese", "tamil", "yue chinese", "egyptian arabic", "turkish", "korean", "filipino", "italian", "jinyu chinese", "gujarati", "thai", "pashto", "kannada", "nigerian pidgin", "min nan chinese", "odia (oriya)", "oromo", "malayalam", "xiang chinese", "sindhi", "polish", "fulah", "sudanese arabic", "algerian arabic", "amharic", "burmese", "odia", "malay", "bhojpuri", "sundanese", "hakka chinese", "moroccan arabic", "azerbaijani", "ukrainian", "hausa", "yoruba", "northern uzbek", "igbo", "saraiki", "uzbek", "cebuano", "awadhi", "antankarana malagasy", "saidi arabic", "dutch", "south azerbaijani", "malagasy", "gan chinese", "north azerbaijani", "bagirmi fulfulde", "marwari", "romanian", "nepali", "maithili", "rajasthani", "serbo-croatian", "northeastern thai", "assamese", "madurese", "mesopotamian arabic", "rangpuri", "sinhala", "magahi", "haryanvi", "zhuang", "nepali", "khmer", "chhattisgarhi", "southern pashto", "nigerian fulfulde", "zulu", "kazakh", "deccan", "chichewa", "sanaani arabic", "swedish", "greek", "iranian persian", "shona", "ta'izzi-adeni arabic", "hungarian", "kurmanji kurdish", "low german", "sorani kurdish", "tunisian arabic", "hijazi arabic", "wolof", "norwegian bokmål", "tigrinya", "ilocano", "czech", "nande", "xhosa", "north mesopotamian arabic", "kinyarwanda", "luba-lulua", "kanuri", "dhundari", "dari", "belarusian", "min dong chinese", "umbundu", "somali", "hiligaynon", "kikuyu", "congo swahili", "bambara", "haitian creole", "tajik", "hebrew", "catalan", "quechua", "sichuan yi", "bavarian", "mossi", "kimbundu", "sylheti", "kongo", "minangkabau", "serbian", "standard moroccan tamazight", "hmong", "uyghur", "rundi", "albanian", "kanauji", "santali", "afrikaans", "eastern maninkakan", "northern pinghua", "southern pinghua", "varhadi-nagpuri", "bulgarian", "northern thai", "central pashto", "mongolian", "sesotho", "krio", "swiss german", "mewati", "balochi", "tswana", "luyia", "guarani", "luganda", "libyan arabic", "betawi", "danish", "southern thai", "norwegian", "bemba", "kashmiri", "kituba", "malvi", "northeastern dinka", "sepedi", "finnish", "halh mongolian", "tok pisin", "sukuma", "hadrami arabic", "koongo", "sicilian", "ghanaian pidgin english", "slovak", "konkani", "balinese", "mainfränkisch", "paraguayan guaraní", "croatian", "huizhou chinese", "eastern oromo", "buginese", "tichurong", "mazanderani", "southern uzbek", "dinka", "konkani", "kamba", "bukit malay", "kalenjin", "gheg albanian", "banjar", "northern hindko", "borana-arsi-guji oromo", "turkmen", "makhuwa", "merwari", "zarma", "gilaki", "bosnian", "southern balochi", "sidamo", "achinese", "shekhawati", "pulaar", "chuanqiandian cluster miao", "garhwali", "shan", "lombard", "lambadi", "galician", "bangala", "central atlas tamazight", "lingala", "hmong daw", "peripheral mongolian", "georgian", "pattani malay", "kabyle", "bikol", "sankaran maninka", "gondi", "waray", "central kanuri", "omani arabic", "bundeli", "musi", "kenyi", "tachelhit", "southern kurdish", "ibibio", "hunsrik", "sabah malay", "godwari", "armenian", "zaza", "efik", "pular", "hassaniyya", "tonga", "brahui", "baoulé", "kumaoni", "sango", "maay", "kyrgyz", "aymara", "tibetan", "eastern egyptian bedawi arabic", "south bolivian quechua", "northern gondi", "tagwana senoufo", "nyankole", "jamaican creole english", "dogri", "segeju", "kedah malay", "gusii", "sasak", "pu-xian chinese", "bouyei", "dyula", "batak toba", "west albay bikol", "beja", "pampanga", "kurukh", "central bikol", "tsonga", "bini", "pahari-potwari", "sadri", "konkani", "waddar", "luba-katanga", "bagri", "chiga", "lithuanian", "soga", "chadian arabic", "dogri", "mobwa karen", "min bei chinese", "hazaragi", "swati", "meru", "kangri", "mandinka", "tulu", "southern betsimisaraka malagasy", "cameroon pidgin", "occitan", "lomwe", "chuka", "tatar", "upper saxon", "yongbei zhuang", "esperanto", "wagdi", "khandesi", "powari", "shahmirzadi", "makasar", "makassar malay", "ci gbe", "bodo", "giryama", "nyamwezi", "kipsigis", "ahirani", "defi gbe", "wolaytta", "fanti", "tumbuka", "mende", "lampung api", "slovenian", "bashkir", "northern luri", "chuvash", "eastern balochi", "tosk albanian", "amdo tibetan", "kalanga", "lugbara", "timne", "north ndebele", "central aymara", "tarifit", "nimadi", "serer", "alur", "mandeali", "teso", "dimli", "southern ma'di", "central-eastern niger fulfulde", "scots", "western maninkakan", "malawi sena", "lango", "tsimihety malagasy", "acoli", "central malay", "igala", "bhili", "lampung nyo", "pangasinan", "dombe", "sonha", "makhuwa-shirima", "qashqa'i", "liberian english", "meiteilon (manipuri)", "eastern yiddish", "surgujia", "northern dong", "maasina fulfulde", "afar", "thur", "eastern apurímac quechua", "southern dong", "takwane", "abron", "makonde", "cusco quechua", "s'gaw karen", "gujari", "tai dam", "tamashek", "western armenian", "gogo", "makhuwa-meetto", "ngandyera", "mbalanhu", "nyakyusa-ngonde", "ndonde hamba", "bukusu", "norwegian nynorsk", "machinga", "susu", "anaang", "sena", "khams tibetan", "macedonian", "tachawit", "avaric", "northern betsimisaraka malagasy", "venda", "maguindanaon", "haya", "mewari", "bulu", "masaaba", "western balochi", "marma", "sakalava malagasy", "bhilali", "napo lowland quechua", "eastern hongshuihe zhuang", "tswa", "surjapuri", "mundari", "southern pastaza quechua", "tena lowland quichua", "morisyen", "bakhtiari", "gurani", "soninke", "northern qiandong miao", "estonian", "vlaams", "northern khmer", "batak simalungun", "salasaca highland quichua", "calderón highland quichua", "tausug", "rejang", "vasavi", "k'iche'", "batak dairi", "cebaara senoufo", "anyin", "irish", "tesaka malagasy", "hadothi", "tigre", "muong", "dagaari dioula", "latvian", "gamo", "batak mandailing", "zande", "khasi", "northern dagara", "gorontalo", "sardinian", "talysh", "jambi malay", "izon", "lozi", "pwo eastern karen", "bena", "southern luri", "najdi arabic", "farefare", "newari", "rakhine", "shambala", "trinidadian creole english", "songe", "campidanese sardinian", "berom", "basque", "southern dagaare", "ngbaka", "ebira", "kabiyè", "ronga", "chuwabu", "mahasu pahari", "guibian zhuang", "nupe-nupe-tako".