Journal of Applied Big Data Analytics, Decision-Making, and Predictive Modelling Systems

Hallucination in Low-Resource Languages: Amplified Risks and Mitigation Strategies for Multilingual LLMs

Mostafa Abdelrahman[1]

[1]*Fayoum University, Department of Computer Science, El-Gamaa, Fayoum, Egypt.*

Abstract

Hallucinations in low-resource languages present challenges for multilingual language models in domains where factual accuracy and linguistic nuance are paramount. Large Language Models (LLMs) often rely on extensive corpora for training, yet many dialects and underrepresented languages lack substantial textual resources. This scarcity can amplify hallucination, where models generate statements devoid of factual grounding, leading to misinformation and undermining user trust. Recent advancements in neural architectures provide partial solutions through language transfer and specialized fine-tuning, but limitations persist when data exhibits inconsistent spellings, code-switching, and limited availability of authoritative references. The resulting outputs may include fabricated entities, incorrect translations, or contextually discordant facts. These hallucinations pose risks in settings such as healthcare, legal documentation, and governmental communication. Empirical findings indicate that increasing the size of training data and leveraging cross-lingual transfer techniques can mitigate certain errors, though no single strategy fully eradicates hallucination. The surge in real-world deployments of LLMs amplifies ethical concerns over content authenticity, fairness in resource allocation, and long-term user reliance on automated systems. Future research directions highlight the importance of balanced corpora, robust evaluation metrics, and lexicon-based validation strategies to enhance reliability in low-resource contexts. Methods for systematic error analysis, domain adaptation, and targeted oversight offer promising steps toward higher-fidelity multilingual generation.

1. Introduction

Hallucination phenomena arise from complex interactions between model architectures, training data distributions, and linguistic intricacies. Multilingual language models trained on massive corpora often display proficiency in high-resource languages while showing marked deficiencies when dealing with lower-resource variants. These deficiencies include inaccurate translations, invented references, and unwarranted contextual leaps. Training pipelines tend to optimize for broad-coverage performance, emphasizing average accuracy across many languages instead of tailoring strategies for limited-resource contexts. The resulting outputs can become unreliable, with higher hallucination rates manifesting in narrative texts and informational passages.

Multilingual LLMs aim to unify multiple languages within a single model, leveraging shared subword embeddings and cross-lingual transfer. Such techniques yield robust generalization across some linguistic families but can falter when subword tokenization fails to capture underrepresented vocabulary. Rare morphological structures, regional idioms, and sparse lexical forms escape the model's learned patterns. Fine-tuning on small or noisy datasets introduces further risk of amplifying existing biases, sometimes embedding misinformation or linguistic artifacts into the model's generative processes.

Organizations deploying automated systems that target a global audience encounter ethical and practical implications when hallucination rates vary by linguistic group. Users of minority languages may receive subpar informational content, possibly misaligned with cultural norms or factual records. This discrepancy raises concerns over unequal access to reliable machine-generated texts and perpetuates broader digital inequalities. Regulatory bodies and public institutions rely on textual data to drive decisions, compounding the risks posed by model-generated inaccuracies.

Studies on hallucination in major languages typically examine factual corrections, retrieval-augmented generation, or specialized fine-tuning. However, minimal attention has been paid to the data scarcity challenges afflicting underrepresented languages, where the corpus volume might be insufficient for effective knowledge transfer. Models can appear fluent while conveying misleading content that remains undiscovered due to a lack of reference materials or expertise among system evaluators. Such overlooked hallucination can propagate myths, rumors, and incomplete cultural narratives.

Tokenization lies at the core of many modeling challenges. Subword-based systems excel in splitting text for languages with large corpora, yet they struggle to adapt to languages featuring unique alphabets, complex diacritics, or frequent borrowing from multiple linguistic sources. Intricate morphological rules can result in multiple valid tokenizations, creating inconsistencies during training. With few domain-specific corpora, data augmentation efforts risk introducing artificial patterns that fail to generalize. The underlying embeddings then form tenuous semantic associations, raising the incidence of fabricated claims during generation.

Data curation in low-resource scenarios complicates the matter further. Collection strategies sometimes merge user-generated content from social media, partially curated texts from local publications, and transcriptions of oral interviews. The resulting data often exhibits code-switching, unconventional spelling variants, or informal registers of language. Automatic cleaning pipelines might inadvertently discard relevant linguistic markers that are crucial for meaning, leading to impoverished model understanding. Efforts to filter offensive or nonsensical content reduce noise at the expense of trimming an already small dataset, prompting trade-offs between

representativeness and quality.

Ethical dimensions cannot be ignored when exploring hallucination in underrepresented languages. Disinformation campaigns and propaganda may exploit the fact that authoritative corpora are scarce or non-existent. Automated systems could inadvertently generate distorted narratives that shape public perception of historical events or communal knowledge. Community-driven initiatives to compile open corpora have emerged, aiming to mitigate the resource gap, but adoption is inconsistent. The reliance on volunteer contributors and sporadic validation means that systematic errors can propagate, fueling hallucination tendencies in LLM outputs.

Emerging frameworks for measuring hallucination offer partial solutions by identifying divergences between model output and a reference knowledge base, yet these methods are typically validated on high-resource settings. There remains a pressing need to adapt and refine such frameworks for languages lacking robust ontologies or widely accepted lexical databases. This gap underscores the necessity of analyzing the root causes of hallucinations, adopting context-sensitive mitigation strategies, and incorporating domain-specific knowledge in model training [1], [2].

A closer examination of hallucination in low-resource contexts reveals specific risk factors, including limited corpora, inconsistent orthographies, and insufficient representation of domain-specific lexicons. Transliteration pipelines and cross-lingual adapters show promise in bridging gaps, though they may introduce additional artifacts. Practical applications demand rigorous testing protocols that account for sociolinguistic variations, domain constraints, and real-world use cases. The following sections discuss the nature of hallucination, patterns of amplification in low-resource scenarios, and technical strategies aimed at reducing hallucination rates through improved data augmentation, cross-lingual knowledge transfer, and architecture-specific mitigation techniques.

2. The Nature of Hallucination Phenomena in LLMs

Hallucination stems from a mismatch between internal model representations and real-world referential content. Neural language models generate text by conditioning on learned distributions, approximating the probability of sequences that appear coherent in training data [3]. In high-resource contexts, the abundance of text, documentation, and curated corpora allows the model to cross-reference factual statements more effectively. In low-resource contexts, the probability distributions capture incomplete or skewed representations of reality, leading to fabricated details that superficially match the model's internal syntax and lexical patterns [4].

At the core of hallucination lies a shortfall in grounding mechanisms, where models construct coherent sentences without ensuring alignment with external facts. Traditional language modeling objectives minimize perplexity by predicting tokens that follow a given context. This optimization does not necessarily enforce veracity. Incorporating external knowledge sources, such as knowledge graphs or curated ontologies, can mitigate hallucination by linking text generation to factual evidence. However, these resources are frequently unavailable or underdeveloped for many underrepresented languages, creating structural vulnerabilities.

Prominent forms of hallucination include invented entities, distorted historical references, and mismatch of named entities to their proper contexts. Words might be strung together coherently but fail to reflect reality. When users engage with such content in a language with limited digital presence, fact-checking is challenging. Many indigenous languages lack established dictionaries or electronic archives, so standard knowledge alignment procedures become impractical. As a result, model outputs carry a higher risk of containing factually baseless statements, which may be accepted at face value by readers if alternative references cannot be easily located.

Subtle forms of hallucination manifest as partial truths interwoven with incorrect details. Such mixtures can be more problematic than outright fabrications since they appear plausible. Models exploit statistical cues gleaned from training data, generating texts that mimic the style and structure of well-resourced languages. When training sets contain parallel sentences or bilingual dictionaries, cross-lingual embeddings might conflate contexts and produce erroneous translations that sound credible due to consistent grammar. A phrase might reference an accurate event but assign it to an incorrect place or date, rendering the entire statement misleading.

Ambiguity in morphological structures also leads to unintentional hallucination. Certain languages exhibit agglutinative or polysynthetic morphology, where multiple grammatical elements attach to a single word root. If the training data is sparse, the model's segmented embeddings may fail to capture the compositional aspects of meaning. Generation then becomes guesswork regarding possible affix combinations, leading to nonsensical expressions or plausible-sounding terms that do not exist in actual usage. This phenomenon highlights how linguistic complexity interacts with data scarcity to amplify error rates.

Data contamination occurs when scraped texts conflate multiple dialects or incorporate erroneous translations from volunteer-driven platforms. If the model learns from such flawed corpora, it normalizes incorrect forms and references, effectively baking hallucination into the parameter space. Post-processing steps, such as grammar checks or dictionary lookups, are less feasible for languages without well-established computational tools. The absence of robust morphological analyzers, part-of-speech taggers, or named entity recognizers in these languages leaves the hallucination unchecked.

Evaluation protocols for hallucination often rely on reference-based metrics or human assessment. Many reference-based methods (e.g., BLEU, ROUGE) measure lexical overlap rather than factual correctness, leading to incomplete detection of invented content. Human reviewers, if not fluent in underrepresented languages, might overlook subtle inaccuracies. Others might lack authoritative sources to verify statements in their native tongue. In scenarios where a language's textual tradition is primarily oral, the challenge of validating model outputs grows further, as oral narratives are not systematically archived or standardized.

Addressing hallucination in LLMs trained on low-resource languages involves acknowledging these multifaceted aspects: data sparsity, linguistic complexity, contamination from noisy sources, and limited validation tools. Methods proposed in high-resource settings require adaptation to languages with different grammatical structures and cultural contexts. Col-

Factor	Description	Impact	Mitigation Strategy
Training Data Distribution	Skewed towards high-resource languages	Increased errors in low-resource texts	Balanced data curation
Subword Tokenization	Fails for rare morphological structures	Loss of linguistic nuances	Adaptive tokenization methods [5]
Noise in Fine-Tuning	Small datasets amplify biases [6]	Embeds misinformation	Robust validation protocols
Cross-Lingual Transfer	Works well for some families but not all	Unstable generalization	Language-specific adaptation layers
Data Augmentation	Risk of artificial patterns	Poor generalization in real-world texts	Context-aware augmentation

Table 1. Factors contributing to hallucination in multilingual language models.

Challenge	Description	Consequence	Potential Solution
Sparse Training Corpora	Insufficient text data for many languages	Poor model fluency	Data collection initiatives
Orthographic Variability	Inconsistent spelling	Tokenization instability	Standardization efforts
Code-Switching	Mixed-language usage in texts	Confused language modeling	Specialized handling techniques
Lack of Lexical Resources	No comprehensive dictionaries or corpora	Difficult model alignment	Crowdsourced lexicon development
Ethical Considerations	Risk of misinformation	Reinforcement of biases	Transparent evaluation metrics

Table 2. Challenges in modeling low-resource languages and potential solutions.

laborative data curation with local communities could improve corpus quality, though such efforts demand sustained engagement and resource allocation. Techniques like leveraging cross-lingual transfer from semantically closer languages offer partial relief, yet do not wholly eliminate the inherent data limitations.

A deeper understanding of how language models internalize linguistic features is crucial for developing robust hallucination detection and prevention mechanisms. Hybrid approaches that combine rule-based components with neural backbones might provide interpretability benefits. Symbolic knowledge representations, although less flexible, can help ground neural generation by enforcing constraints on permissible factually relevant outputs. Structured alignment with domain-specific terminology and integration of context-aware retrieval might further bolster factual correctness. Continued exploration of these strategies can illuminate paths toward reducing the prevalence of hallucination in low-resource language contexts.

3. Amplification of Hallucination in Low-Resource Scenarios

Sparse data availability exacerbates the tendency of LLMs to generate content that deviates from factual accuracy. Training procedures generally involve sampling from text corpora that are heavily skewed toward popular languages. This imbalance leads to well-tuned representations for languages with significant digital footprints while leaving minimal capacity for underrepresented languages. Subword vocabulary selection typically favors frequent tokens drawn from high-resource corpora, producing suboptimal segmentations in underrepresented languages and undermining semantic fidelity.

Code-switching behavior adds another layer of complexity. Speakers of low-resource languages often incorporate terms from major lingua francas, resulting in heterogeneous textual data. Language models trained on mixed-code texts may attempt to generate content by applying translation rules or morphological transformations that do not align with actual usage. This phenomenon can contribute to hallucinated terms that sound superficially correct but deviate from real linguistic norms. Official documents in minority languages often rely on loanwords when specialized terminology is unavailable, introducing further irregularities that the model might misinterpret or exaggerate.

Orthographic inconsistencies, common in languages lacking standardized scripts, lead to fragmentation of the training corpus. Variations in diacritic use, transliteration schemes, or region-specific spelling conventions fragment the data, preventing robust pattern extraction. The model may interpret different spellings of the same word as unrelated tokens, diluting any aggregated representation. Consequently, the chance of generating invented forms rises, since the model extrapolates from incomplete morphological and lexical patterns. Unsupervised subword methods can inadvertently capture partial tokens that do not map cleanly to actual morphological units, sowing the seeds of nonsense outputs.

Limited domain coverage further intensifies hallucination risks. High-resource languages benefit from data spanning news articles, scientific literature, social media content, and various specialized fields. In contrast, low-resource languages might only have substantial data in a single genre, such as religious texts or folklore. Domain-specific corpora imbue the model with a narrow worldview, prompting it to fill knowledge gaps through unwarranted extrapolation. For instance, if the only available texts are historical narratives, the model may hallucinate contemporary facts that mirror archaic language or conflate historical events with modern ones.

Socioeconomic and political factors also contribute to data scarcity. Minority communities may face limited internet access, reducing the volume of digital text available for model training. Government policies might prioritize official languages, leaving underrepresented languages absent from official documents and large-scale digitization efforts. Commercial incentives drive content creation in widely spoken languages, overlooking smaller linguistic communities. This lack of institutional support constrains the diversity of textual resources and fosters an environment where hallucination thrives.

Annotation quality exerts substantial influence on model outputs. Machine translation systems for low-resource languages often rely on crowdsourced data with varying quality levels. Inconsistent or erroneous annotations can embed systematic biases and factual inaccuracies into training corpora. Reinforcement learning steps, if guided by suboptimal reward signals, may further reinforce hallucinations. Reviewers who are not fully fluent or trained in the language might miss subtle factual errors, thereby introducing a reinforcing loop where the model is validated on incomplete or incorrect references.

Cross-lingual transfer techniques attempt to alleviate data scarcity by mapping representations from high-resource languages to low-resource ones. These methods rely on shared subword vocabularies, bilingual dictionaries, or parallel texts. While beneficial for bootstrapping coverage, such transfers risk injecting the target language with biases and structural assumptions from the source language. If the source and target languages belong to different families, morphological and syntactic mismatches can lead to hallucinated constructs absent from either language's norms. Even genetically related languages may differ sufficiently in orthography or phonology to trigger systematic generation errors.

Domain adaptation strategies involving fine-tuning on specialized corpora can inadvertently amplify hallucination if the specialized corpora are small or of questionable quality. Overfitting to limited domain data can produce illusions of factual knowledge. The model internalizes domain-specific terms without a balanced sense of their real-world frequency or context. Rare words seen only in one domain become associated with other contexts, triggering out-of-domain hallucinations. Reliance on any single data source during fine-tuning risks overshadowing general language competence, culminating in a model that generates domain-distorted text.

Human-machine feedback loops pose additional concerns. Interactive systems where users correct or refine model outputs can aid iterative improvement. Yet in low-resource settings, there may be few expert reviewers available, and the process of repeated correction is time-consuming and requires compensation. Underqualified annotators might introduce further uncertainties, rationalizing erroneous outputs instead of discarding them. This dynamic perpetuates a cycle in which ambiguous or fabricated content gains acceptance and eventually becomes re-ingested into future training iterations.

Collective reinforcement of hallucination occurs when multiple systems trained on overlapping data corroborate fabricated statements. If one system invents a claim, subsequent language models may adopt it as accepted knowledge when that claim appears in re-scraped or user-generated content. This cyclical contamination is challenging to detect or reverse, especially in minor languages lacking robust fact-checking frameworks. Even if discovered, retraction is complicated by the absence of an official correction mechanism, risking permanent entrenchment of spurious information in the digital record.

Researchers and practitioners must grapple with these amplifying factors to devise targeted solutions. Bolstering data collection and curation processes, promoting standardized orthographies, and encouraging community-driven lexical resources can all help reduce hallucination prevalence. Frameworks that integrate external validation signals—ranging from knowledge bases to specialized domain experts—offer additional safeguards. Preprocessing pipelines that unify spelling and address code-switching through consistent transliteration can minimize model confusion. The next section explores how data augmentation, cross-lingual transfer, and advanced modeling approaches can mitigate hallucination by leveraging structured methods that refine and expand the training corpus while preserving linguistic integrity.

4. Data Augmentation and Cross-Lingual Transfer Approaches

Data augmentation strategies for low-resource languages serve as a critical avenue to counterbalance sparse textual resources. These strategies manipulate existing data to simulate new training instances, thereby increasing corpus size and diversity. Methods such as back-translation utilize machine translation pipelines, translating text from the low-resource language into a high-resource language and back. Although this approach can create varied sentence structures, inaccuracies in the translation process can introduce artificial errors or reinforce existing ones. Selecting reliable pivot languages and robust translation models is essential for maintaining fidelity to the original semantics.

Language modeling pipelines can also benefit from paraphrasing techniques. Generative models trained in high-resource languages paraphrase low-resource sentences, which are then translated back into the original language. This creates multiple expressions of similar content, expanding the pool of training samples. Ensuring that morphological and cultural subtleties remain intact is crucial. If automated paraphrasing loses essential elements of meaning or style, the augmented data risks amplifying hallucination by normalizing inaccuracies.

Noisy channel approaches augment data by introducing controlled noise, simulating real-world typing errors, code-switching, or domain-specific lexical usage. These approaches expose the model to plausible variations in spelling or grammar, building resilience against unpredictable inputs. However, poorly configured noise levels can lead to confusion be-

Technique	Description	Benefit	Risk
Back-Translation	Translate to a high-resource language and back	Increases diversity	Introduces translation errors
Paraphrasing	Generate alternative sentence structures	Expands linguistic variation	Risk of semantic drift
Noisy Channel	Simulate real-world noise (typos, code-switching)	Improves robustness	Can confuse model with random noise
Syntactic Transformations	Reorder sentence components while preserving meaning	Enhances syntactic flexibility	May create ungrammatical outputs [7], [8]
Domain-Specific Expansion	Generate structured text from expert templates	Ensures domain relevance	Limits generalization

Table 3. Comparison of data augmentation techniques and their effects.

Strategy	Description	Advantage	Challenge
Multilingual Pretraining	Train on multiple languages simultaneously	Enables cross-lingual knowledge sharing	Poor generalization for distant languages [8]–[10]
Fine-Tuning on Low-Resource Data	Adapt a pretrained model with specific data	Improves specialization	Risk of overfitting on small datasets [11]
Unsupervised Domain Adaptation	Align representations using monolingual corpora	Enhances language-specific embeddings	Requires large high-quality corpora
Parallel Corpus Learning	Train with bilingual or multilingual aligned texts	Strengthens cross-lingual coherence	Parallel data is scarce in low-resource languages
Transliteration Pipelines	Convert scripts into a shared representation	Facilitates knowledge transfer	Can introduce orthographic inconsistencies

Table 4. Overview of cross-lingual transfer strategies and their trade-offs.

Approach	Integration Method	Benefit	Risk
Semi-Supervised Learning	Combine unlabeled monolingual data with annotated samples	Enhances generalization	Requires careful validation of labeled data
Multitask Learning	Incorporate auxiliary tasks (NER, disambiguation)	Strengthens factual consistency	Labeled data is resource-intensive
Iterative Knowledge Distillation	Use teacher-student models for pseudo-labeling	Transfers structured knowledge	Errors from teacher propagate to student
Hybrid Augmentation Pipelines	Mix multiple augmentation strategies adaptively	Balances data diversity	Needs fine-tuned orchestration to avoid bias
Community-Based Evaluation	Involve native speakers for validation	Improves cultural accuracy	Resource-intensive and slow scaling

Table 5. Hybrid strategies for reducing hallucination in low-resource language models.

tween genuine linguistic artifacts and random mistakes. This confusion might worsen hallucination if the model becomes excessively tolerant of spurious sequences. Calibration of noise injection requires a detailed understanding of the language's orthographic rules and morphological structure.

Syntactic transformations offer another augmentation av-

enue. Certain frameworks parse sentences into syntactic trees, apply transformations such as subject-object inversion or clause reordering, and then reconstruct sentences. Such transformations generate fresh training instances that preserve semantic content while varying structural patterns. This technique can bolster the model's grasp of syntax. Yet for languages with complex morphology or free word order, naive transformations may distort meaning or create ungrammatical constructs. Customized syntactic rules, guided by linguistic expertise, become essential for reliable augmentation.

Cross-lingual transfer leverages alignments between the target language and a better-resourced sister language, assuming some degree of lexical or typological proximity. Shared subword vocabularies facilitate knowledge transfer, although the usefulness of this approach diminishes for distant language pairs. Training a bilingual or multilingual model, then fine-tuning it on limited data from the low-resource language, can hasten model convergence and improve performance. The risk of hallucination persists if the model learns misleading patterns from the higher-resource language that do not apply in the target language. Careful curation of parallel corpora or bilingual dictionaries helps alleviate these issues, though truly parallel data sets are rarely abundant in low-resource contexts.

Unsupervised domain adaptation methods aim to refine cross-lingual representations using monolingual texts from the target language. During self-supervised training, the model learns to predict masked tokens, capturing the statistical regularities of the language's syntax and semantics. When integrated with cross-lingual objectives, these representations become more aligned, allowing partial transfer of linguistic knowledge. The final performance hinges on the quantity and quality of monolingual data for the low-resource language. If domain coverage remains narrow, the resulting improvements in language modeling may not reduce hallucination in broader contexts.

Multitask learning paradigms can incorporate additional objectives, such as named entity recognition or morphological disambiguation, to enrich the model's internal representations. Such tasks anchor the generative model to factual and structural components of language. Including tasks that require explicit identification of factual statements or constraints on entity relationships can provide stronger guardrails against hallucination. However, preparing labeled data for these tasks remains challenging in languages with limited resources. Often, it demands experts who can label complex morphological forms or domain-specific terminologies.

Semi-supervised approaches combine the benefits of unlabeled monolingual text with smaller sets of annotated data. The model learns general patterns from large unlabeled corpora, refining them using the annotated subset for tasks such as question answering or summarization. If the annotated data includes factual consistency checks or reference alignment, it can guide the model away from hallucinated content. The inherent limitation is that the annotated subset must be carefully validated for accuracy, as any noise can pollute the entire training process.

Iterative knowledge distillation can also reduce hallucination. A teacher model trained in a high-resource language setting can generate pseudo-labels or translations in the low-resource language, and a student model is fine-tuned on these outputs. This method hinges on the assumption that the teacher model's knowledge is accurate and adaptable across languages. If the teacher model itself exhibits hallucination or if the language gap is too large, the student model may inherit errors. Validation of the teacher's outputs by native speakers or domain experts can mitigate the risk, but resource constraints often limit such oversight.

Domain-specific data augmentation involves targeted collection and synthesis of text relevant to specialized areas like healthcare, law, or education. Techniques such as rule-based template generation, where domain experts define canonical sentence structures, can produce controlled expansions of training data. By populating these templates with plausible entities and scenarios, the model learns robust associations grounded in realistic contexts. This method ensures coverage of critical topics that might be missing in generic corpora. Yet it demands considerable domain expertise and systematic evaluation to confirm correctness. Overreliance on template-based data may lead to overly rigid text generation habits, limiting the model's ability to adapt to novel contexts within the same domain.

The success of any data augmentation or cross-lingual transfer strategy depends on the ability to preserve factual integrity. If the process inadvertently introduces inconsistencies or artificially inflates data without proper quality checks, the risk of hallucination heightens. Feedback loops that combine automated validation with expert reviews can maintain a cycle of continuous improvement. As new data is generated, it should be assessed for factual consistency before reintegration into the training corpus. Community-based initiatives, where local speakers offer feedback, represent a cost-effective path toward iterative refinement. Such involvement fosters a sense of ownership and cultural accuracy, ensuring that the models respect linguistic norms and communal knowledge systems.

Comprehensive data augmentation pipelines demand a balance between quantity, quality, and authenticity. Overly aggressive transformations or unregulated cross-lingual injections risk overshadowing the true characteristics of the target language, creating superficial illusions of data richness. On the other hand, limited augmentation or conservative transfer methods may fail to alleviate the scarcity problem, leaving the model prone to hallucination. Intelligent orchestration of multiple strategies—back-translation, paraphrasing, syntactic transformations, code-switching simulation, and domain-specific expansions—can diversify the training distribution sufficiently to reduce hallucination occurrences.

5. Model Architectures and Mitigation Techniques

Transformer-based architectures dominate multilingual LLM research due to their capacity for handling large training corpora and modeling long-range dependencies. Pre-training on multilingual corpora allows a single model to learn cross-lingual representations, but these representations can become entangled when language-specific tokens are infrequent. Researchers have introduced modular layers or language-specific adapters that slot into a shared backbone, aiming to retain generalization benefits while isolating language-specific peculiarities. Adapters trained on low-resource languages can capture morphological and syntactic patterns without overwriting parameters learned from high-resource languages, though

they may still inherit some cross-lingual confusion.

Retrieval-augmented generation (RAG) integrates external knowledge bases to ground model outputs. At generation time, the model retrieves relevant documents or structured data, merging them with the internal decoder states. This method reduces hallucination risk by forcing the model to justify statements using external evidence. For low-resource languages, building and maintaining a knowledge base is nontrivial. Digitized texts, encyclopedic entries, or parallel corpora are sparse, making retrieval less reliable. However, targeted curation of domain-relevant knowledge repositories can yield substantial gains in factual accuracy, provided that coverage aligns with user queries.

Factual calibration techniques focus on adjusting the model's confidence in its generated outputs. Temperature scaling, for instance, can dampen overconfidence in uncertain contexts, curtailing the production of definitive yet inaccurate statements. By lowering the temperature, the model diversifies its output and is less prone to hallucinating unwavering claims. This approach does not inherently improve factual grounding but can reduce misleading certainty. Calibration must be tuned per language or domain to reflect realistic uncertainty levels, an endeavor complicated by limited test data in low-resource contexts.

Regularization strategies such as label smoothing or dropout can help models avoid overfitting to small corpora. Overfitting can accentuate hallucination if the model clings to memorized but incorrect training patterns. Careful hyperparameter tuning is necessary to strike a balance between memorization of relevant linguistic forms and generalization to unseen contexts. Non-adaptive regularization can blunt the model's capacity to capture fine-grained nuances in morphological-rich languages. Dynamic regularization schedules, which adjust dropout or learning rates according to the model's perceived difficulty of the training samples, might yield more stable outcomes.

Hybrid architectures incorporate symbolic reasoning modules or rule-based constraints alongside neural components. The symbolic module enforces consistency checks or domain rules, while the neural module handles fluent generation and broad coverage. This synergy provides interpretability and a safeguard against blatant factual errors. Symbolic constraints can validate named entities, numeric consistency, or known relationships among concepts. However, implementing such modules for lesser-known languages is difficult, as it requires codifying grammar rules or compiling structured knowledge. When symbolic resources are incomplete, the system risks degenerating to purely neural generation, reintroducing hallucination potential.

Active learning frameworks solicit human feedback where the model exhibits high uncertainty or domain specificity. Presenting uncertain outputs to native speakers or domain experts and incorporating corrections into the training loop iteratively refines the model's performance. This process can target hallucination hotspots, such as specialized terms or cultural references, but the method's success hinges on the availability of qualified annotators. Low-resource language communities often lack extensive pools of linguistic experts, and the cost of such interventions can be prohibitive. Crowd-based solutions might help, but verifying crowd-sourced corrections poses another challenge in resource-scarce environments.

Prompt engineering has emerged as a technique to steer model behavior without retraining. Designing prompts that include contextual clues or references to external knowledge sources can reduce the model's reliance on incomplete internal representations. By explicitly providing relevant facts or constraints in the prompt, the model is less likely to hallucinate contradictory information. In low-resource scenarios, prompt templates can incorporate bilingual or lexical cues that anchor generation in known expressions. This approach remains limited by the model's underlying capacity to handle the language, and prompt design can be time-intensive.

Evaluation metrics for measuring hallucination must evolve to address the unique challenges of low-resource languages. Automated benchmarks often rely on reference translations or fact-based question answering. In languages lacking extensive parallel corpora, alternative methods involve human evaluation of factual correctness, linguistic fluency, and cultural appropriateness. Criteria for correctness may vary across dialects or communities, complicating consensus. Adversarial testing can expose the model to queries or prompts designed to trigger potential hallucinations, uncovering failure modes not apparent in standard benchmarks. Integrating these insights into training fosters a cycle of iterative refinement.

Ensemble methods can reduce hallucination by combining outputs from multiple models that vary in architecture, data selection, or training objectives. If the models exhibit complementary strengths, voting or confidence-based selection of final outputs can filter out the most egregious hallucinations. The approach hinges on diversity among model predictions and an effective aggregation mechanism. Training multiple models for low-resource languages demands computational resources, which might be scarce for community-driven or academic projects. Suboptimal ensembles can degrade performance if each component model shares similar biases or data limitations [1], [12].

Mitigation efforts must remain vigilant against cultural biases and colonial patterns embedded in data collection. Automated systems often reflect the viewpoints of dominant groups, risking erasure or distortion of minority narratives. Hallucination compounds these issues by fabricating statements about cultural practices or historical events that lack documented sources. Rigorous community involvement and transparency in dataset construction can align the model's outputs with authentic linguistic and cultural realities. Ultimately, an integrated approach combining data augmentation, cross-lingual knowledge transfer, architectural adaptations, and robust evaluation protocols offers the most promise for mitigating hallucination in low-resource contexts [13]. Such a multifaceted strategy requires collaboration between technologists, linguists, community members, and policymakers to ensure inclusive and accurate multilingual systems.

6. Conclusion

Multilingual language models exhibit heightened vulnerability to hallucination in low-resource contexts, posing risks for domains where factual accuracy and cultural authenticity are indispensable. Training methods designed for high-resource languages often underperform when data remains sparse or nonstandard, leading to outputs that misrepresent reality. Code-switching, orthographic diversity, limited domain coverage, and scarce lexicographic resources contribute to the model's

inability to ground text in verifiable facts. These challenges are reinforced by socioeconomic factors and institutional biases that prioritize global languages over minority ones, perpetuating cycles of limited data availability and suboptimal model performance [14], [15].

Architectural interventions that incorporate adapters, retrieval-based grounding, or symbolic constraints can curtail some forms of hallucination, but none serve as a universal solution. Data augmentation tactics, including back-translation, paraphrasing, syntactic transformations, and cross-lingual transfer, alleviate data scarcity to varying degrees. Their effectiveness hinges on stringent validation to ensure that artificially expanded corpora do not introduce new inaccuracies. Techniques such as factual calibration, active learning, and domain adaptation offer additional layers of refinement, though all depend on carefully curated resources and competent human oversight.

Research efforts continue to explore ensemble approaches, specialized metrics, and interpretability frameworks to detect and reduce hallucination. Low-resource languages demand bespoke strategies that respect linguistic complexity, sociocultural context, and community needs. Collaborative initiatives between researchers, local communities, and policymakers have the potential to reshape the data ecosystem, fostering more comprehensive repositories and culturally aligned benchmarks. Sustained investment in language documentation, knowledge base construction, and domain-specific annotation will prove essential for future model robustness.

Systematic and transparent evaluation remains pivotal for understanding progress in mitigating hallucination. Metrics should capture factual correctness, contextual relevance, and cultural nuance, rather than focusing narrowly on token overlap or perplexity. Community-led error analysis can reveal linguistic phenomena that automated checks overlook, spurring more targeted and inclusive model refinements. Continued development of these methods can strengthen user trust in multilingual systems and extend the reach of natural language processing to underserved linguistic communities. By combining technical innovation with ethical responsibility, the field can move closer to equitable and reliable language technologies for all.

References

[1] G. Geigle, A. Jain, R. Timofte, and G. Glavaš, "Mblip: Efficient bootstrapping of multilingual vision-llms," *arXiv preprint arXiv:2307.06930*, 2023.

[2] M. Berbatova and Y. Salambashev, "Evaluating hallucinations in large language models for bulgarian language," in *Proceedings of the 8th Student Research Workshop associated with the International Conference Recent Advances in Natural Language Processing*, 2023, pp. 55–63.

[3] R. Mehta, A. Hoblitzell, J. O'Keefe, H. Jang, and V. Varma, "Metacheckgpt–a multi-task hallucination detection using llm uncertainty and meta-models," *arXiv preprint arXiv:2404.06948*, 2024.

[4] H. Ye, T. Liu, A. Zhang, W. Hua, and W. Jia, "Cognitive mirage: A review of hallucinations in large language models," *arXiv preprint arXiv:2309.06794*, 2023.

[5] F. Yuan, S. Yuan, Z. Wu, and L. Li, "How multilingual is multilingual llm?" *arXiv preprint arXiv:2311.09071*, 2023.

[6] S. V. Bhaskaran, "Tracing coarse-grained and fine-grained data lineage in data lakes: Automated capture, modeling, storage, and visualization," *International Journal of Applied Machine Learning and Computational Intelligence*, vol. 11, no. 12, pp. 56–77, 2021.

[7] Y. Qiu, Y. Ziser, A. Korhonen, E. M. Ponti, and S. B. Cohen, "Detecting and mitigating hallucinations in multilingual summarisation," *arXiv preprint arXiv:2305.13632*, 2023.

[8] Z. Ji, T. Yu, Y. Xu, N. Lee, E. Ishii, and P. Fung, "Towards mitigating llm hallucination via self reflection," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 1827–1843.

[9] R. Mehta, A. Hoblitzell, J. O'keefe, H. Jang, and V. Varma, "Halu-nlp at semeval-2024 task 6: Metacheckgpt-a multi-task hallucination detection using llm uncertainty and meta-models," in *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, 2024, pp. 342–348.

[10] N. M. Guerreiro, D. M. Alves, J. Waldendorf, *et al.*, "Hallucinations in large multilingual translation models," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1500–1517, 2023.

[11] S. V. Bhaskaran, "A comparative analysis of batch, real-time, stream processing, and lambda architecture for modern analytics workloads," *Applied Research in Artificial Intelligence and Cloud Computing*, vol. 2, no. 1, pp. 57–70, 2019.

[12] Z. Cao, Y. Yang, and H. Zhao, "Autohall: Automated hallucination dataset generation for large language models," *arXiv preprint arXiv:2310.00259*, 2023.

[13] S. V. Bhaskaran, "Integrating data quality services (dqs) in big data ecosystems: Challenges, best practices, and opportunities for decision-making," *Journal of Applied Big Data Analytics, Decision-Making, and Predictive Modelling Systems*, vol. 4, no. 11, pp. 1–12, 2020.

[14] Q. Cheng, T. Sun, W. Zhang, *et al.*, "Evaluating hallucinations in chinese large language models," *arXiv preprint arXiv:2310.03368*, 2023.

[15] A. Bruno, P. L. Mazzeo, A. Chetouani, M. Tliba, and M. A. Kerkouri, "Insights into classifying and mitigating llms' hallucinations," *arXiv preprint arXiv:2311.08117*, 2023.