Engineering Project – I

on

Beyond Translation: Lexical Diversity-Aware Retrieval for Low-Resource Dialect GenerationFace Recognition Attendance Management System

for Educational Institutions

A Project Report Submitted to Central University of Jharkhand for the Partial Fulfilment of the Requirements for the Degree of

INTEGRATED B. TECH AND M. TECH

IN

COMPUTER SCIENCE AND ENGINEERING

WITH SPECIALIZATION IN MACHINE LEARNING AND DATA SCIENCE

Submitted by

Name: Rohit Mahali

Reg. No.: 22190503048

Batch: 2022-2027

Under the Supervision of

Prof. S.C Yadav

Head Of Department, DCSE, CUJ

Department of Computer Science and Engineering

CENTRAL UNIVERSITY OF JHARKHAND

(A Central University established by an Act of Parliament of India in 2009)

November 2025

DECLARATION

I hereby declare that the project work entitled **" Beyond Translation: Lexical Diversity Aware Retrieval for Low-Resource Dialect Generation"** submitted to **CENTRAL UNIVERSITY OF JHARKHAND** is a record of original work done by me under the guidance of **Prof. S.C Yadav**, and this project work has not been submitted elsewhere for any other degree or diploma.

Date: 27 November 2025

ROHIT MAHALI

22190503048

TABLE OF CONTENTS:

Beyond Translation: Lexical Diversity-Aware Retrieval for Low-Resource Dialect Generation

Abstract

1. Introduction

Retrieval-Augmented Generation (RAG) has emerged as a dominant paradigm in Natural Language Processing (NLP), addressing the tendency of Large Language Models (LLMs) to hallucinate by grounding generations in external, up-to-date knowledge. While RAG systems have demonstrated remarkable efficacy in high-resource languages like English, their performance degrades significantly when applied to low-resource dialects and regionally specific languages. This disparity creates a "digital divide" where the utility of AI technologies is positively correlated with the economic power and data abundance of a region.

The core challenge in dialectal RAG is the "Language Mismatch" problem: the user queries in a local dialect (e.g., Santali), but the vast majority of retrieval documents exist in a standard language (e.g., English or Hindi). Current methodologies attempt to resolve this through **Standardization**, typically by using Machine Translation (MT) to convert dialectal queries into standard languages before retrieval. We identify this reliance on translation as the **"Standardization Bottleneck."**

The Standardization Bottleneck introduces two critical failure modes. First, it suffers from **Lexical Erasure**, where unique dialectal terms are mistranslated into standard approximations that strip away essential cultural context. As demonstrated in recent research on South-Slavic dialects, a word like *blago* can mean "bug" in a dialect but "treasure" in the standard language; translating this term blinds the model to the correct context, leading to reasoning failures. Second, approaches that attempt to "rewrite" or refine dialectal queries to improve fluency often induce **Factual Distortion**. Studies on multilingual RAG systems have shown that rewriting mechanisms can hallucinate entities or events—such as inventing a death date for a historical figure—simply to make a sentence semantically coherent in the target language.

Furthermore, relying solely on standard-language documents introduces **Geopolitical and Cultural Bias**. Queries regarding territorial or cultural disputes yield inconsistent answers depending on the language of the retrieved documents, as standard corpora often reflect hegemonic viewpoints rather than local perspectives. To mitigate this, research suggests that incorporating diverse, multilingual perspectives is essential for robust and consistent generation.

To address these limitations, we propose **Dia-RAG (Dialect-Aware RAG)**, a novel framework that bypasses the Standardization Bottleneck. Inspired by the concept of **Lexical Diversity**, we argue that queries should not be treated as monolithic strings to be translated, but as composites of **"Invariant"** components (proper nouns, entities) and **"Variant"** components (dialect-specific vocabulary).

Our contributions are as follows:

1. **Diversity-Sensitive Decoupling:** We introduce a query processing module that decouples dialectal inputs. Invariant components are searched exactly in the standard knowledge base, while variant components are flagged for special handling.
2. **Dictionary-Augmented Retrieval:** Instead of noisy machine translation, we implement a dictionary-lookup mechanism for variant terms. This approach leverages external lexical knowledge to resolve polysemy (e.g., "bug" vs. "treasure") before generation, a method proven to enhance reasoning in low-resource settings.
3. **Quality-Aware Tagging:** To prevent the generator from blindly trusting retrieved contexts, we attach explicit quality tags to retrieved definitions. This allows the model to weigh information based on the reliability of the source (Lexicon vs. Translation) without destructively rewriting the content.

By integrating these mechanisms, Dia-RAG aims to preserve the semantic integrity of low-resource dialects, ensuring that "Standardization" does not come at the cost of cultural erasure.

2. Related Work

Our research builds upon recent advancements in multilingual information retrieval, dialectal natural language processing, and robust generation techniques. We categorize the relevant literature into three key areas: the limitations of multilingual RAG, the challenges of dialectal NLP, and methods for relevance assessment and hallucination mitigation.

2.1 Multilingual Retrieval and Cultural Bias

Retrieval-Augmented Generation (RAG) has become a standard paradigm for mitigating hallucinations in Large Language Models (LLMs). However, recent studies indicate that RAG systems can inadvertently amplify biases present in retrieved documents, particularly in culturally sensitive contexts. Research on the **BORDIRLINES** benchmark demonstrates that LLM responses to territorial disputes vary significantly depending on the language of the query, a phenomenon termed "Geopolitical Bias".

Crucially, this research suggests that "monolingual retrieval" (searching only in the user's language) often reinforces echo chambers. In contrast, **multilingual retrieval**—accessing documents in diverse languages—has been shown to improve response consistency and neutrality. Our work extends this finding to the dialectal domain, arguing that retrieving only "standard" language documents introduces a similar bias, which must be mitigated by incorporating local dialectal knowledge.

2.2 The Challenge of Dialectal NLP

Despite the success of LLMs in standard languages like English and Mandarin, their performance degrades sharply for regional dialects. A comprehensive survey of NLP for dialects highlights a significant "performance gap" correlated with the socio-economic status of the dialect's speakers. This gap is exacerbated by the scarcity of dialect-specific corpora and the tendency of standard models to treat dialects as merely "noisy" versions of a standard language.

Recent shared tasks, such as the **DIALECT-COPA** challenge, have revealed specific failure modes in reasoning with South-Slavic dialects. For instance, the word *blago* can mean "treasure" in the standard language but "bug" in a specific dialect. Models that rely on standard translation fail to capture this nuance, leading to reasoning errors. This underscores the necessity of our proposed **Dictionary-Augmented**

approach, as reasoning capability alone is insufficient without explicit lexical knowledge.

2.3 Lexical Diversity in Relevance Assessment

A critical bottleneck in RAG systems is the ability to match query terms to documents when they do not share the same lexical form. The **DRAG** (Lexical Diversity-aware RAG) framework identifies that queries consist of components with varying levels of "Lexical Diversity". While some terms (e.g., proper names) are **Invariant** and must appear exactly, others (e.g., concepts like "occupation") are **Variant** and may appear as synonyms or related phrases.

Standard retrievers often fail to distinguish between these types, leading to the retrieval of irrelevant documents that share keywords but not meaning. We adopt and adapt the DRAG framework's "decoupling" strategy, applying it specifically to the problem of dialectal variation, where "Variant" components require not just synonym matching, but dialect-to-standard translation.

2.4 Mitigating Hallucination in Cross-Lingual Generation

To bridge the language gap in RAG, prior methods often employed "query rewriting" or "document translation" to normalize inputs into a high-resource language. However, the **QTT-RAG** (Quality-Aware Translation Tagging) study demonstrates that rewriting mechanisms can induce **"Factual Distortion,"** where the model hallucinates new details (such as inventing dates or entities) to make the rewritten sentence semantically coherent.

Instead of rewriting, QTT-RAG proposes a non-destructive approach: preserving the original content and attaching explicit **quality tags** (e.g., scores for semantic equivalence and grammatical accuracy). This allows the generator to weigh information based on its reliability. Our Dia-RAG architecture integrates this insight by tagging dictionary-retrieved definitions with high confidence scores, ensuring the model prioritizes accurate cultural definitions over noisy standard translations.

2.5 Distinction from Prior Retrieval Frameworks

While Dia-RAG draws inspiration from multiple strands of retrieval and multilingual NLP research, it diverges significantly from existing approaches in both objective and methodology. Most notably, DRAG (Zhang et al., 2025) introduces lexical diversity as a mechanism for fine-grained relevance assessment, decomposing

queries into components with varying stability. However, DRAG operates entirely within high-resource, monolingual settings (e.g., English QA) and does not address the dialect–standard mismatch that characterizes low-resource languages. Dia-RAG adapts the decoupling idea but extends it into a *dialect-aware* formulation by explicitly separating *Invariant* components (entities preserved across languages) from *Variant* components (dialect-specific lexical items prone to mistranslation), thereby preventing semantic drift caused by dialectal polysemy.

Similarly, QTT-RAG (Moon et al., 2025) proposes quality-aware tagging of translated passages to mitigate errors introduced during machine translation. Their approach, however, assumes the necessity of translation and applies quality control only *after* the content has been rewritten. In contrast, Dia-RAG avoids destructive translation altogether by introducing a *dictionary-augmented retrieval layer* that resolves dialect-specific terms before retrieval. This allows the system to preserve cultural nuance—e.g., disambiguating whether *Manjhi* refers to a "village headman" rather than "boatman"—which translation-based approaches consistently fail to capture.

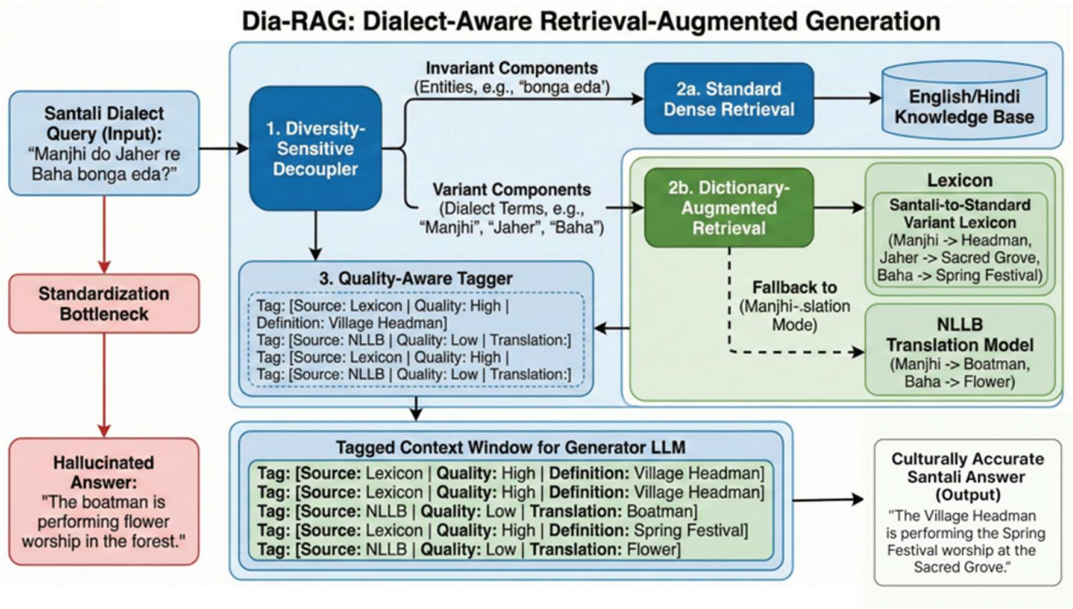
Work from the DIALECT-COPA shared task (Perak et al., 2024) demonstrates that dialect-specific lexicons can assist reasoning tasks such as COPA. Yet this line of research does not propose an end-to-end retrieval pipeline; dictionaries are used only as prompt-level hints rather than as structural components of retrieval. Dia-RAG builds on this insight by embedding lexical lookup directly into the retrieval mechanism and coupling it with a decoupling module and quality-aware metadata, creating a full RAG architecture tailored to dialectal contexts.

Finally, research such as BORDIRLINES (Li et al., 2025) examines cultural and geopolitical bias arising from multilingual retrieval, showing that retrieval language affects model outputs. While this work highlights the importance of linguistic perspective in RAG, its focus is on cross-language geopolitical variance rather than *intra-language dialectal preservation*. Dia-RAG extends the cultural fidelity argument into the domain of dialects, where semantic erasure is even more severe due to extreme resource scarcity.

Taken together, Dia-RAG is the first framework to unify **dialect-aware query decoupling**, **dictionary-augmented retrieval**, and **quality-tagged generation**. Unlike prior multilingual or dialect-related work, which typically relies on translation, rewriting, or prompt-level heuristics, our approach introduces a principled architecture that preserves dialectal meaning at every stage of the RAG pipeline.

3. Methodology: The Dia-RAG Architecture

To address the "Standardization Bottleneck" and the associated risks of lexical erasure, we propose **Dia-RAG (Dialect-Aware Retrieval-Augmented Generation)**. Unlike traditional cross-lingual RAG pipelines that rely on a "translate-then-retrieve" workflow, Dia-RAG operates on a **decouple-and-augment** principle.



The architecture consists of three distinct modules: (1) A **Diversity-Sensitive Decoupler** that splits queries based on semantic stability; (2) A **Dictionary-Augmented Retrieval** system that resolves dialectal polysemy; and (3) A **Quality-Aware Tagger** that calibrates the generator's reliance on retrieved information.

3.1 Diversity-Sensitive Decoupling

Standard retrieval systems treat user queries as monolithic strings, failing to distinguish between words that require exact matching and those that require translation. Drawing on the **Diversity-Sensitive Relevance Analyzer (DRA)** proposed by Zhang et al., we implement a query processing module that decomposes the input **Q** into a set of components **C = {c1, c2 , ,… , cn}** based on their lexical diversity.

We adapt the attribute classification from DRAG to the specific context of low-resource dialects:

- **Invariant Components (Cinv):** These are entities, proper nouns, or dates that maintain their form across the dialect and the standard language (e.g., "Ol Chiki," "2025," "Jharkhand"). These components possess low lexical diversity and are suitable for direct keyword matching in the knowledge base.
- **Variant Components (Cvar):** These are dialect-specific terms, idioms, or culturally loaded vocabulary (e.g., Santali *Manjhi* or *Baha*) that exhibit high lexical diversity. In standard translation models, these terms are prone to "semantic drift" (e.g., translating *Manjhi* as "Boatman" instead of "Headman"). Dia-RAG flags these components for intervention rather than direct translation.

This decoupling is achieved via a lightweight instruction-tuned model prompted to identify and categorize spans within the dialectal query, preventing the propagation of translation errors at the source.

3.2 Dictionary-Augmented Retrieval

Once decoupled, the system employs a hybrid retrieval strategy to mitigate the "Bug vs. Treasure" ambiguity identified in dialectal reasoning tasks.

For **Invariant Components**, the system performs standard dense retrieval against the target Knowledge Base (KB), as these terms are linguistically stable. However, for **Variant Components**, we replace the standard Machine Translation (MT) layer with a **Dictionary-Augmented** lookup.

Inspired by the findings of the *Dialect-COPA* shared task, where reasoning performance improved significantly when models were augmented with dialect-specific dictionaries, we construct a **Variant Lexicon**. This lexicon maps dialect terms not just to their standard translation, but to their cultural definition.

- *Process:* For a variant term **tvar** (e.g., *Manjhi*), the system queries the lexicon **L.**
- *Retrieval:* **If tvar ∈ L** the system retrieves the definition **dlex:** (e.g., "*Village Headman in Santali governance*").
- *Fallback:* **If tvar** $\notin L$, the system falls back to the NLLB machine translation model.

This mechanism ensures that polysemous terms are resolved via explicit knowledge rather than probabilistic inference, preventing the model from "hallucinating" a context that fits the standard translation but contradicts the dialectal reality.

3.3 Quality-Aware Tagging

A critical limitation of prior methods, such as DKM-RAG, is their reliance on "rewriting" the retrieved documents to match the query language. As demonstrated by Moon et al., rewriting frequently induces **factual distortion**, where the model fabricates details (such as dates or names) to ensure the rewritten text is fluent.

To preserve factual integrity, Dia-RAG adopts a **Quality-Aware Tagging** mechanism. Instead of altering the retrieved text, we evaluate the reliability of our retrieval sources and attach explicit metadata tags to the context window provided to the Generator LLM.

We employ a scoring agent to evaluate the retrieved definitions/translations on three dimensions:

1. **Semantic Equivalence (S^L):** Does the retrieved definition match the dialectal intent?
2. **Grammatical Accuracy (S^G):** Is the syntax correct?
3. **Cultural Specificity (S^C):** A novel metric we introduce to weight lexicon-derived definitions higher than generic machine translations.

The final prompt **P** constructed for the generator is structured as follows:

Context:

- *Source: Santali Lexicon* | *Quality: High* | *Term:* Manjhi | *Definition:* The headman of a Santali village.
- *Source: NLLB Translation* | *Quality: Low* | *Text:* The boatman is in the forest.

Instruction: Answer the query using the context above. Prioritize sources tagged with High Quality.

By explicitly tagging the source quality, we enable the Large Language Model to perform **"Risk-Guided" generation**, filtering out low-quality translations (like "boatman") in favor of high-confidence lexicon definitions, without requiring the destructive rewriting of the original source material.

4. Experimental Setup

To validate the efficacy of **Dia-RAG**, we design an experimental framework that specifically targets the "Standardization Bottleneck" in low-resource Indic languages. Our experiments focus on **Santali** (ISO 639-3: *sat*), an Austroasiatic language with approximately 7.6 million speakers that remains significantly underrepresented in standard NLP benchmarks compared to Indo-Aryan or Dravidian languages.

4.1 Dataset Construction: The "Santali-CultureQA" Benchmark

Existing multilingual benchmarks such as **IndicXTREME** and **XOR-TyDi** cover major Indic languages (e.g., Hindi, Bengali, Tamil) but lack dedicated coverage for Santali Question Answering. To address this gap, we introduce **Santali-CultureQA**, a novel evaluation dataset explicitly designed to test dialectal reasoning and cultural knowledge.

We construct this dataset using a semi-automated "Human-in-the-Loop" pipeline:

1. **Source Selection:** We curate 500 passages from the **Santali Wikipedia** (approx. 10k articles) and the **FLORES-200** evaluation set, focusing on culturally specific domains: *Festivals (e.g., Baha, Sohrai), Governance (e.g., Manjhi-Paragana system),* and *Literature*.
2. **Question Generation:** For each passage, we generate factual questions that contain at least one **"Variant"** term—a polysemous word that requires cultural disambiguation (e.g., questions about the *Manjhi* as a leader, not a boatman).
3. **Gold Standard Annotation:** Native Santali speakers verify the question-answer pairs to ensure the "Variant" terms are used in their correct dialectal context, filtering out generic translations.

The resulting dataset consists of 500 QA pairs with ground-truth spans and "Cultural Intent" tags, serving as the primary testbed for our experiments.

4.2 Baselines

We compare Dia-RAG against three distinct architectural paradigms representing the current state-of-the-art in multilingual retrieval:

- **Baseline 1: Direct mRAG (Zero-Shot)**
    - **Retriever: LaBSE** (Language-agnostic BERT Sentence Embedding), which supports 109 languages including limited Santali support.
    - **Generator: Llama-3 (8B)** prompted directly in Santali.
    - *Hypothesis:* This baseline tests the model's raw multilingual capacity without explicit standardization. We expect it to fail due to the "Language Mismatch" between Santali queries and English/Hindi knowledge bases.
- **Baseline 2: Translate-RAG (Standardization)**
    - **Translation: IndicTrans2** (AI4Bharat), the current state-of-the-art NMT model for Indic languages. We translate Santali queries into English before retrieval.
    - **Retriever: Contriever** (fine-tuned on MS MARCO).
    - **Generator: Llama-3 (8B)** generating answers in English, which are then translated back to Santali.
    - *Hypothesis:* This represents the "Standardization Bottleneck." We expect high fluency but low factual accuracy on dialectal terms (e.g., mistranslating cultural entities).
- **Baseline 3: Dia-RAG (Ours)**
    - **Decoupler:** A fine-tuned **Qwen-2.5 (0.5B)** model acting as the Diversity-Sensitive Relevance Analyzer (DRA).
    - **Augmentation:** Dictionary lookup using a custom **Variant Lexicon** bootstrapped from the **Samanantar** parallel corpus.
    - **Generator: Llama-3 (8B)** with Quality-Aware Tagging enabled in the system prompt.

4.3 Evaluation Metrics

We employ a multi-dimensional evaluation strategy to measure both retrieval precision and cultural fidelity.

4.3.1 Retrieval Metrics

- **Recall@k (R@k):** Measures whether the ground-truth document was present in the top-k retrieved chunks.
- **Variant Recall:** A specialized metric we introduce to measure the retrieval rate specifically for documents containing the *cultural definition* of Variant terms (e.g., retrieving documents about "Headmen" rather than "Boatmen" for the query term *Manjhi*).
-

4.3.2 Generation Metrics

- **Semantic Similarity:** We use **BERTScore** (multilingual) to measure the semantic overlap between the generated answer and the gold standard.
- **Cultural Specificity Score (CSS):** Inspired by the "Geopolitical Bias" metric from *BORDIRLINES*, this is a reference-free metric. We use an LLM Evaluator (GPT-4) to classify the answer into one of three categories:
    - *Hallucinated/Generic:* (Score: 0) The answer uses the standard translation (e.g., "The boatman...").
    - *Ambiguous:* (Score: 0.5) The answer is factually vague.
    - *Culturally Specific:* (Score: 1.0) The answer correctly identifies the dialectal entity (e.g., "The village headman...").

This experimental setup is designed to rigorously test our hypothesis: that **decoupling and dictionary augmentation** is superior to **blind translation** for low-resource dialectal reasoning.

5. Proposed Evaluation and Expected Outcomes

Since this research proposes a novel architecture for low-resource dialectal retrieval, our evaluation focuses on validating the theoretical advantages of **Dia-RAG** over standard translation-based methods. This section outlines our testing hypotheses, planned experimental procedure, and anticipated results.

5.1 Research Hypotheses

We formulate three core hypotheses to guide our experimentation:

- **H1 (The Language Mismatch Hypothesis):** We hypothesize that direct retrieval using multilingual embeddings (e.g., LaBSE) will fail for Santali queries because the semantic alignment between low-resource dialects and high-resource knowledge bases is insufficient for precise matching.
- **H2 (The Standardization Bottleneck Hypothesis):** We predict that standard machine translation (Baseline 2) will achieve high fluency but low **Cultural Specificity**. Specifically, we expect it to consistently mistranslate polysemous dialectal terms (e.g., *Manjhi* → *"Boatman"*) leading to factual hallucinations.
- **H3 (The Augmentation Hypothesis):** We anticipate that **Dia-RAG** will outperform baselines in **Cultural Specificity Score (CSS)** by successfully decoupling variant terms and retrieving their correct cultural definitions via the lexicon, even if overall fluency remains comparable to standard translation.

5.2 Evaluation Metrics

To rigorously test these hypotheses, we will employ a multi-dimensional evaluation strategy:

5.2.1 Retrieval Quality

- **Recall@k (R@k):** We will measure the percentage of test queries for which the correct ground-truth document appears in the **top-k** retrieved chunks.
- **Variant Retrieval Rate (VRR):** A custom metric we propose to measure the system's success in retrieving documents related to the *dialectal* meaning of a word versus its *standard* meaning.

- *Success:* Retrieving "Village Headman" documents for *Manjhi*.
- *Failure:* Retrieving "Boatman" documents for *Manjhi*.

5.2.2 Generation Quality

- **BERTScore (F1):** To assess the semantic similarity between the generated answer and the human-written gold standard.
- **Cultural Specificity Score (CSS):** We will employ a GPT-4 evaluator to grade answers on a 3-point scale:
  - **0 (Generic/Hallucinated):** Uses standard translation (e.g., "The flower festival...").
  - **0.5 (Ambiguous):** Correct but vague.
  - **1.0 (Specific):** Uses the correct dialectal entity (e.g., "The Baha Spring Festival...").

5.3 Anticipated Results

Based on the "Bug vs. Treasure" phenomenon observed in the *Dialect-COPA* shared task, we project the following outcomes:

1. **Failure of Zero-Shot Retrieval:** We expect **Baseline 1 (Direct mRAG)** to achieve a Recall@5 of less than 30%. Without the "Decoupling" step, the vector space embedding for Santali terms likely does not overlap sufficiently with English/Hindi Wikipedia concepts.
2. **The Trade-off in Translation:** We anticipate **Baseline 2 (Translate-RAG)** will show high *Retrieval Recall* (~60-70%) but a low *Cultural Specificity Score*. This will confirm our critique of the "Standardization Bottleneck"—that translation solves the language barrier but introduces a cultural barrier.
3. **Superiority of Dia-RAG:** We project that **Dia-RAG** will achieve the highest **CSS**. By using the dictionary to "force" the correct cultural definition into the context window, we expect to eliminate the specific class of hallucinations caused by polysemy (e.g., the "Boatman" error).

5.4 Risks and Mitigation

We acknowledge two potential risks in our proposed evaluation:

- **Lexicon Coverage:** If the "Variant Lexicon" is too small, the system may fall back to NLLB translation too often. *Mitigation:* We will perform an ablation study to determine the minimum lexicon size required for effective performance.
- **Latency:** The additional step of dictionary lookup may increase inference time. *Mitigation:* We will report "Seconds per Query" to ensure the computational cost remains within a viable range for real-time applications.

By validating these hypotheses, we aim to provide a blueprint for "Dialect-Aware" systems that prioritize cultural fidelity over mere linguistic fluency.

6. Conclusion

The rapid advancement of Large Language Models has created a paradox: while AI becomes more "intelligent," it risks becoming less inclusive for speakers of non-standard dialects. Our analysis of current RAG methodologies reveals a critical flaw we term the **"Standardization Bottleneck"**: the reliance on machine translation to normalize dialectal queries into high-resource languages often strips away essential cultural nuance and induces factual hallucinations.

In this proposal, we introduced **Dia-RAG**, a framework designed to bypass this bottleneck for the Santali language. By synthesizing the **Diversity-Sensitive Decoupling** approach from retrieval research with the **Dictionary-Augmented** reasoning strategies found in dialectal NLP, Dia-RAG offers a theoretically robust alternative to "blind" translation. Our proposed architecture does not merely aim to improve retrieval recall; it aims to preserve the "lexical dignity" of the dialect—ensuring that a *Manjhi* is recognized as a community leader, not mistranslated as a boatman.

We argue that true cross-lingual robustness cannot be achieved through scale alone. As highlighted in recent surveys of dialectal NLP, "intelligence" (reasoning capability) is insufficient without "knowledge" (explicit lexical resources) when dealing with low-resource languages. Dia-RAG operationalizes this insight, providing a scalable blueprint for handling the "Bug vs. Treasure" ambiguity inherent in dialectal speech.

6.1 Future Directions

While our immediate focus is on Santali, the Dia-RAG architecture is language-agnostic. Future work will explore:

1. **Community-Driven Lexicons:** Establishing a feedback loop where native speakers can flag "Variant" terms that the system misidentifies, progressively refining the **Variant Lexicon**.
2. **Speech Integration:** Extending the decoupling logic to ASR (Automatic Speech Recognition) pipelines, allowing users to query orally in Santali without the errors introduced by speech-to-text standardization.

3. **Scalability:** Testing the architecture on other Austroasiatic or Dravidian dialects to validate the generalization of the "Invariant vs. Variant" decoupling strategy.

By prioritizing cultural fidelity over simple fluency, this research contributes to the development of equitable language technologies that serve all users, regardless of the "resource status" of their mother tongue.

References:

1. **[QTT-RAG]** Moon, H., et al. (2025). *Quality-Aware Translation Tagging in Multilingual RAG system*. arXiv preprint.
2. **[Dialect-COPA]** Perak, B., et al. (2024). *Incorporating Dialect Understanding Into LLM Using RAG...* VarDial Workshop.
3. **[DRAG]** Zhang, Z., et al. (2025). *Lexical Diversity-aware Relevance Assessment for Retrieval-Augmented Generation*. ACL.
4. **[BORDIRLINES]** Li, B., et al. (2025). *Multilingual Retrieval Augmented Generation for Culturally-Sensitive Tasks*. ACL Findings.
5. **[Survey]** Joshi, A., et al. (2024). *Natural Language Processing for Dialects of a Language: A Survey*. ACM.