My research interest lies in aligning computational models with low-resource languages. This serves two comprehensive goals: (i) building robust models and (ii) ensuring they benefit all language communities. Surprisingly, while current models excel at complex reasoning in rich-resourced languages, they often fail on basic tasks in under-resourced ones. Closing this gap is a fascinating goal toward improving model interpretability and understanding human language universality. It also has the potential to influence other modalities and advance perceptual grounding. The following research questions form the cornerstone of the goals I aim to pursue in my PhD:

(RQ: 1) How can we integrate practical linguistic differences into models' representational learning?

During my bachelor's third year, I was excited about the potential of byte-pair-encoding (BPE) to address **the out-of-vocabulary** (OOV) issues in low-resourced morphologically-rich Bangla language processing. In one case, this means BPE to generate as many highly productive subwords (replacing the need for labour-intensive phoneme and morpheme annotation), which thereafter maps to unique vector representations depending on the tasks (conventional word2vec fails on this). For the Bangla language, which has flexible **word order** and significant **orthographic variation**, i.e., a single word may appear in different spellings while conveying the same meaning, we noticed that multi-lingual BERT's encoding mechanism fails to grasp the context (Islam et al. 2021). In another project, we showed that monolingual training of this model is not sufficient to overcome these challenges (Islam et al. 2022). Interestingly, many global languages have these traits—flexible word order in Russian and Turkish, orthographic variation in Hausa, and both in Hindi and Arabic—suggesting a broader vulnerability in current LLMs. These challenges raise the question of whether the root problem lies in the representational layer of current models. Moving forward, I aim to explore how injecting practical linguistic intuitions—such as phonological cues available through standard phonological dictionaries—can improve **representational learning** for these languages.

(RQ: 2) What does the internal multilingual region tell us about how it deals with different languages?

A relevant question to *(RQ: 1)* is: Do these LLMs understand the gold prompt for these languages? In one case, there is human guidance that the models seem to contextualize well for many NLP tasks. On the other hand, these models are under-trained [from *(RQ: 1)*]. With this in mind, we set up a **cross-lingual** evaluation (Islam et al. 2025) on a **multi-hop reasoning** non-binary answer QA dataset on a small parameterized model (Llama-3.2-1B). Intuitively, if the models perform competitively well across languages, they are theoretically multi-lingual. Surprisingly, although we notice premise-supportive results, the model put substantially lower importance across English sentences, compared to its translated Bangla counterpart. This asymmetry implies that morphological richness may shape how the model distributes attention across languages. Building on this observation, I am interested in **probing** the model's internal multilingual regions to understand how they encode linguistic differences and what this reveals about the model's cross-language understanding.

(RQ: 3) Can we incorporate linguistic structural reward policies in post-training?

The results of *(RQ: 2)* motivated a natural question. How well do LRMs perform in low-resource languages when evaluated through generation? Given that **reward policies** in current **post-training** pipelines heavily shape stylistic and behavioural preferences, I was particularly interested in whether they improve the performance of these undertrained models. In my recent

project, we evaluated Qwen3 by comparing its non-thinking (LLM) and thinking (LRM) variants across parameter scales on high-, mid-, and low-resource languages using the MGSM8K dataset [*preprint to appear soon*]. Surprisingly, LRM consistently outperformed LLM across all languages, with performance gains increasing only modestly as the parameter scale grew to $8B$. Importantly, the stylistic reward cues did not produce substantial performance changes at larger scales. Motivated by this, I now aim to investigate whether linguistic structural reward policies can more effectively guide learning, especially at small parameter scales across various resourced languages.

(RQ: 4) How can we align with other modalities to amplify under-resourced performance?

In my recent course projects, I explored how additional modalities can support textual tasks within a multi-modal framework. During ancient Hebrew OCR fine-tuning, we observed that **vLM**s converged faster than other architectural pre-trained models (Westerdijk et al. 2025). In another project on **audio-text** fallacy detection, Qwen3 adapted its thinking block when supplied with paralinguistic cues (Zhou et al. 2025). These findings suggest that different modalities provide complementary cues, although each requires task-specific adjustments. For instance, adopting more interpretable vLM architectures to reduce reliance on monolingual decoders and to support more efficient visual encoding, as well as improving the use of internal prompting mechanisms to ensure **faithfulness** when injecting non-textual data into LLMs. Building on these insights, I aim to understand how principled cross-modal alignment strategies can systematically boost performance in under-resourced settings.

References

[1] Khondoker Ittehadul Islam, Sudipta Kar, Md Saiful Islam, and Mohammad Ruhul Amin. "SentNoB: A dataset for analysing sentiment on noisy Bangla texts". In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. 2021, pp. 3265–3271.

[2] Khondoker Ittehadul Islam and Gabriele Sarti. "Reveal-Bangla: A Dataset for Cross-Lingual Multi-Step Reasoning Evaluation". In: *arXiv preprint arXiv:2508.08933* (2025).

[3] Khondoker Ittehadul Islam, Tanvir Yuvraz, Md Saiful Islam, and Enamul Hassan. "Emonoba: A dataset for analyzing fine-grained emotions on noisy bangla texts". In: *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. 2022, pp. 128–134.

[4] Hylke Westerdijk, Ben Blankenborg, and Khondoker Ittehadul Islam. "Improving OCR for Historical Texts of Multiple Languages". In: *arXiv preprint arXiv:2508.10356* (2025).

[5] Hongxu Zhou, Hylke Westerdijk, and Khondoker Ittehadul Islam. "Joint Effects of Argumentation Theory, Audio Modality and Data Enrichment on LLM-Based Fallacy Classification". In: *arXiv preprint arXiv:2509.11127* (2025).