Rafeed Mohammad Sultan

Natural Language Processing / Multilingual LLM Evaluation, Safety Alignment, and Agentic Retrieval

Sydney, NSW, Australia | +61 424 379 583 | sultanrafeed@gmail.com

Google Scholar | ORCID 0009-0008-4291-1570 | ResearchGate | github.com/sultanrafeed | linkedin.com/in/rafeed-sultan | Portfolio

Research Interests

The field now grades language models with language models, a substitution adopted faster than it has been validated. It holds up worst where the stakes are highest: outside English, on items where human annotators genuinely disagree, and in retrieval pipelines that should abstain rather than answer. I work on evaluation and safety in those settings, and as a native Bangla speaker with published bilingual Bangla/English work, I build the annotated resources these questions cannot be answered without.

1. **Judge reliability outside English.** LLM-as-judge is now the default evaluation paradigm, yet cross-lingual judgment consistency is poor in low-resource languages, and neither multilingual training data nor model scale reliably repairs it. I want to measure rather than assume judge reliability in Bangla and comparable Global-South languages: treating human preference as distributional rather than a single consensus point, testing whether rubric redesign, reference grounding or judge ensembling actually recovers agreement, and asking whether reliability degrades specifically on culturally situated items. Builds directly on my LLM-as-judge pipeline over human-annotated ethics data [2] and bilingual Bangla/English retrieval [1].
2. **Whether safety alignment transfers across languages.** Attack success and defence robustness vary sharply by language, and safety tuning done in English does not reliably carry into low-resource, non-Latin-script settings. I am interested in building refusal and red-team evaluation resources for Bangla with human-written rationales, then testing whether alignment can be transferred by parameter-space methods (safety-vector merging, layer-wise swapping, self-distillation) rather than full retraining. Extends the rationale-annotated dataset design and reason-supervised fine-tuning of [2], using the distillation and model-souping techniques of [4] as tools rather than as an end.
3. **Faithfulness and abstention in agentic retrieval.** Faithfulness and attribution remain unsolved even in RL-trained agentic search systems, and out-of-context or unanswerable queries remain a recognised failure mode, one I documented in my own work. I want component-level diagnosis of where in a retrieval-and-reasoning pipeline faithfulness breaks, run on local open models rather than proprietary ones, with out-of-context handling treated as a primary metric rather than an error bucket. Direct sequel to [1] and to the production RAG and MCP tool-use systems I built in industry.

Candidature sought: Master's by Research in the above areas,open to an earlier Master's-to-PhD transfer where a supervisor and their faculty support it. Currently in Australia on a subclass 500 student visa.

Publications: Peer-Reviewed (Accepted / Published)

All four items below have completed peer review and are published at the venue named; the arXiv identifiers link author preprint versions of the same accepted papers, not separate submissions. I have no items currently under review or preprint-only. Equal first-author contribution on [1]; second author on [2] and [3]. Three CORE-ranked conferences and one Q1 journal.

[1] M. R. Kabir*, **R. M. Sultan***, F. Rahman, M. R. Amin, S. Momen, N. Mohammed, S. Rahman. **LegalRAG: A Hybrid RAG System for Multilingual Legal Information Retrieval**. *International Joint Conference on Neural Networks (IJCNN)*, 2025. **[Published]** arXiv:2504.16121
**Equal contribution.* Added a relevance-check and query-refinement stage over vanilla RAG for bilingual Bangla/English legal documents (cosine similarity 0.76 to 0.82, human evaluation 3.41 to 3.70 of 5), and identified out-of-context and unanswerable queries as the failure mode no configuration tested resolved, the diagnosis that motivates research direction 3.

[2] M. R. Kabir, **R. M. Sultan**, I. H. Asif, J. Ibn Ahad, F. Rahman, M. R. Amin, N. Mohammed, S. Rahman. **Beyond Labels: Aligning Large Language Models with Human-like Reasoning**. *International Conference on Pattern Recognition (ICPR)*, 2024. **[Published]** arXiv:2408.11879 Code & Data
Designed the annotation protocol for DFAR, a 5,000-statement ethics dataset carrying human-written rationales alongside labels, extending ETHICS, annotated by twelve screened annotators; proposed the Misalignment Rate, a metric separating a correct label from correct reasoning. Reason-supervised QLoRA fine-tuning took Llama-2 7B from 36.4% to 89.4% accuracy and cut Misalignment Rate from 52.0% to 9.4%; cross-dataset validation on ETHOS.

[3] J. Ibn Ahad, **R. M. Sultan**, A. Kaikobad, F. Rahman, M. R. Amin, N. Mohammed, S. Rahman. **Empowering Meta-Analysis: Leveraging Large Language Models for Scientific Synthesis**. *IEEE International Conference on Big Data (IEEE BigData)*, 2024, pp. 1615–1624. **[Published]** arXiv:2411.10878 Code

Built MAD (625 meta-articles, 6,344 support abstracts) and introduced Inverse Cosine Distance, a loss for long-context fine-tuning; human-evaluation protocol over 13 annotators returned 87.6% relevance, with irrelevant output falling from 4.56% to 1.9%.

[4] M. R. Kabir, R. H. Borshon, M. K. Wasi, **R. M. Sultan**, A. Hossain, R. Khan. **Efficient Skin Cancer Detection Using Lightweight Model Souping and Ensembling Knowledge Distillation for Memory-Constrained Devices**. *Intelligence-Based Medicine* (Elsevier, Q1), vol. 10, 2024, art. 100176. **[Published]** DOI:10.1016/j.ibmed.2024.100176
Multi-teacher distillation into a 0.35M-parameter student, 82× smaller than DenseNet161, reaching 88.0% on HAM10000 at 1.05M parameters via model souping and ensembling. *Included as evidence of compression and model-merging technique, which I apply to cross-lingual safety transfer (direction 2), not as a separate research programme.*

Research Experience

Independent Researcher: Cross-Model LLM-as-Judge Evaluation
2024 – Present

github.com/sultanrafeed — Independent

- Built an LLM-as-judge pipeline in which four small open judges (Phi-3 Mini 3.8B, StableLM Zephyr 3B, Gemma 2B, GPT-2) scored 500 statement–reason pairs on a six-point rubric across six response systems (Llama-2 7B and Mistral 7B, each pre-trained, label-tuned, and reason-tuned), benchmarked against three-evaluator majority-vote human labels rather than against a single reference model.
- Judge scores recovered the human ranking of systems at Spearman $\rho = -0.60$ against Misalignment Rate (negative by construction; higher score, lower misalignment), but agreement was highly judge-dependent, ranging from $\rho = -0.71$ (Phi-3 Mini, Gemma 2B) to $\rho = -0.09$ (Zephyr 3B): consistent judge *rankings* did not imply consistent judge *reliability*.
- Identified a rank inversion in which the worst system by human evaluation (Llama-2 7B pre-trained, 55.0% Misalignment Rate) placed near the top by judge score, and a degenerate judge (GPT-2, mean 1.21–2.73 against 2.83–3.56 for Phi-3 Mini), evidence that judge capability, not judge count, bounds panel reliability.

Undergraduate Researcher
2023 – 2024

Apurba-NSU R&D Lab, North South University — Dhaka, Bangladesh

- Contributed to four peer-reviewed papers under Dr. Shafin Rahman and Dr. Nabeel Mohammed: dataset curation and annotation protocol design, QLoRA fine-tuning pipelines, custom loss implementation, retrieval architecture, baseline benchmarking, human-evaluation design and manuscript preparation.
- Designed and ran human-evaluation protocols at scale (twelve screened annotators writing rationales for DFAR and three-evaluator majority-vote assessment of generated reasons in [2], thirteen-annotator relevance assessment for [3]), and proposed a metric (Misalignment Rate) that separates label correctness from reasoning correctness.
- Ran the full research cycle end to end on [1] and [2]: research question, experimental pipeline, baseline comparison, human evaluation, and write-up for peer review.

Research Projects

Predictive Modelling and Explainable AI Analysis of White Spot Disease in Farmed Shrimps. Built a disease-prevalence classifier reaching 72.67% accuracy and used SHAP to attribute feature impact for non-technical stakeholders. Code

Optimizer Performance Evaluation. Benchmarked a custom VGG-based CNN across optimisers and batch sizes to characterise convergence and generalisation trade-offs. Code

Efficiency Comparison of Programming Languages in Algorithm Development. Measured runtime, memory and CPU utilisation across languages implementing algorithms of varying complexity, informing language choice for performance-critical systems. Code

Applied Research & Engineering Experience

Software Engineer (promoted from Associate Software Engineer, Apr 2026)
Jan 2025 – Jun 2026

SELISE Digital Platforms — Dhaka, Bangladesh

- Designed and ran model evaluation protocols for production AI features, assessing LLM outputs for accuracy, robustness and instruction adherence prior to release. Direct continuity with the evaluation methodology in [1] and [3].
- Built a production RAG ingestion and retrieval pipeline for multimodal (PDF/image) context on Azure PostgreSQL, and a voice agent combining guardrailed prompting with refusal handling and a Model Context Protocol tool server.
- Led development of an agentic orchestration framework and expanded MCP server tooling for structured LLM tool-use; authored the architecture documentation for model orchestration and tool-calling.

Intern Software Developer
Oct 2024 – Jan 2025

SELISE Digital Platforms — Dhaka, Bangladesh

- Built finance and procurement automation tooling, improving cycle time and data quality for operational reporting.

Teaching Experience

Private Tutor, self-employed, Dhaka (2023 to 2025). One-on-one tuition in Physics, Chemistry, Biology and Mathematics to A Level; individualised lesson design and exam-technique coaching.

Teacher's Assistant, Study Town, Lalmatia, Dhaka (Nov 2019 to Jun 2020). Marked A Level Biology and Chemistry against published criteria; small-group tutoring with written feedback.

Education

Master of Artificial Intelligence — *Jul 2026 – Jun 2028 (expected)*
University of Technology Sydney (UTS) — Sydney, Australia

BSc in Computer Science & Engineering — *2020 – 2024*
North South University | CGPA 3.38/4.00 (scale 4.00) — Dhaka, Bangladesh

Edexcel A Levels & O Levels — *2017 – 2019*
Cephalon International School / Academia — Dhaka, Bangladesh

Technical Skills

Evaluation: human-evaluation protocol design (five- and thirteen-annotator studies), inter-annotator agreement analysis, LLM-as-judge methodology, metric design, experimental design, baseline benchmarking, ablation studies, rubric design, rank-agreement and calibration analysis

Multilingual & low-resource: Bangla/English dataset curation and annotation design, bilingual retrieval and evaluation, multilingual embedding models (bge-m3); Bangla (native) as an annotation and evaluation research instrument

Alignment & Training: parameter-efficient fine-tuning (LoRA/QLoRA), instruction tuning, reason-supervised fine-tuning, custom loss design, knowledge distillation (multi-teacher), model souping and merging

Retrieval & Agents: RAG (dense and sparse retrieval, query refinement, relevance filtering), Model Context Protocol tool-use, LangChain, FAISS, ChromaDB

Stack: PyTorch, Hugging Face (Transformers, PEFT, TRL/SFTTrainer, Datasets), scikit-learn, SHAP, pandas, NumPy; Python, TypeScript, Java, C++, SQL, LaTeX; PostgreSQL, MongoDB, Azure, Git; multi-GPU training under quantised configurations

Certifications

IBM via Coursera *(2024)*: Databases & SQL for Data Science (Honors), Python for Data Science, AI & Development, Data Science Methodology, Tools for Data Science, Introduction to Software Engineering

Google Data Analytics via Coursera *(2024)*: Foundations: Data, Data, Everywhere, Ask Questions to Make Data-Driven Decisions, Prepare Data for Exploration

Academic Service, Community & Languages

Open source: public research code for [2] and [3] and four independent project repositories: github.com/sultanrafeed

Volunteer: JAAGO Foundation, Dhaka (2016), community outreach and fundraising for youth education; raised over BDT 20,000.

Languages: English (IELTS Academic 7.0, CEFR C1), Bangla (native, see Technical Skills; underpins research directions 1 and 2).

Referees

Dr. Shafin Rahman, Associate Professor, Dept. of Electrical & Computer Engineering, North South University *(senior author, [1][2][3])*. shafin.rahman@northsouth.edu

Dr. Nabeel Mohammed, Associate Professor, Dept. of Electrical & Computer Engineering, North South University *(co-author, [1][2][3])*. nabeel.mohammed@northsouth.edu

Dr. Riasat Khan, Dept. of Electrical & Computer Engineering, North South University *(corresponding author, [4])*. riasat.khan@northsouth.edu

Work rights: currently in Australia on a subclass 500 student visa with associated work rights.