



RESEARCH PROGRAM

OTED AI Research Initiative

Program charter and initial research protocol

Measuring the reliability and safety of large language models in Dari and Pashto

Initial project | Multilingual model evaluation

Organization for Talent and Entrepreneurship Development (OTED)

oted.org | oted.us.org | info@oted.org

Program charter

OTED has established the OTED AI Research Initiative as an internal research program. The program studies how artificial intelligence systems perform in languages and settings that receive limited attention in mainstream evaluation. Its first research line focuses on Dari and Pashto, two languages used widely by communities connected to OTED's education and development work.

Research focus. The initiative's first project measures factual accuracy, translation quality, citation reliability, and response consistency in Dari and Pashto. It also tests whether retrieval and structured verification reduce common errors.

Institutional relationship

The initiative operates within OTED and uses OTED's administrative, financial, and organizational controls. OTED is a nonprofit, non-governmental organization established in 2013. Its public programs include education, training, entrepreneurship, health, relief, and community development. The research initiative extends the education and talent-development parts of that mission through systematic computer science research.

Purpose

- Conduct reproducible research on multilingual AI reliability, safety, and usefulness.
- Develop evaluation datasets and open research materials for Dari and Pashto where licensing and participant protections allow.
- Separate research claims from program advocacy, fundraising, and operational decisions.

Initial scope

The program will begin with model evaluation, benchmark construction, error analysis, and mitigation experiments. It will not train a foundation model during the initial project. Work involving students, beneficiaries, medical guidance, or identifiable personal data falls outside the first phase and requires a separate protocol.

Governance and responsibility

Role	Responsibility
Research Lead / PI-equivalent	Sets the research questions, approves the protocol, supervises experiments, reviews results, and accepts responsibility for scientific integrity.
Research contributors	Prepare benchmark items, run documented experiments, conduct language review, and maintain code and data records.
OTED administration	Manages accounts, purchasing, access, records, and compliance with OTED policies.
Independent reviewers	Assess Dari and Pashto outputs under written rubrics. Reviewers will not score outputs they authored.

Initial research project

Project title

Measuring the reliability and safety of large language models in Dari and Pashto

Study rationale

Dari and Pashto appear in multilingual resources such as FLORES-200, but broad model evaluations still give far more attention to English and other high-resource languages. A model may produce fluent text while making factual, translation, or citation errors. OTED will measure those errors under a documented protocol and test whether simple interventions reduce them.

Primary research question

How reliably do current general-purpose large language models answer factual questions, interpret short passages, and translate text in Dari and Pashto compared with matched English tasks?

Secondary questions

- Which error types are most common in each language and task?
- Do retrieval-provided evidence and a structured verification step improve factual accuracy?
- Are model confidence signals and refusals associated with actual correctness?

Testable hypotheses

- H1: Accuracy will be lower in Dari and Pashto than in matched English items.
- H2: Retrieval-provided evidence will reduce unsupported factual claims.
- H3: A structured verification step will improve citation validity and response consistency, though the effect may differ by model and task.

Planned contribution

The project will produce a documented evaluation set, an experiment runner, an anonymized results table, an error taxonomy, and a research report or preprint. The study is comparative and empirical. It will report negative or inconclusive findings alongside positive results.

Operational definition of reliability

Reliability means that a response is factually supported, responsive to the prompt, written in the requested language, and stable enough for repeated evaluation. Safety in this protocol concerns unsupported claims, fabricated citations, inappropriate certainty, and failure to refuse when the prompt lacks enough evidence. The project will not treat fluency alone as evidence of correctness.

Methods

Benchmark design

The pilot will contain 180 items: 60 factual question-answering items, 60 reading-comprehension items, and 60 translation items. Each task will have matched Dari, Pashto, and English material where a valid comparison is possible. After the pilot and quality review, the full benchmark target is 600 items.

Task	Material	Primary outcome
Factual question answering	Public, non-sensitive facts with dated source records and answer keys.	Supported correctness and unsupported-claim rate.
Reading comprehension	Short passages followed by questions answerable only from the passage.	Answer accuracy and evidence use.
Translation	Licensed parallel text, including FLORES-200 Dari and Southern Pashto material where redistribution terms permit.	Adequacy, fluency, meaning preservation, chrF++, and error type.

Systems and conditions

At least three multilingual language models available at the time of testing will be evaluated. Every run will record provider, model name, model version where available, access date, system instructions, temperature, token limit, and prompt. The same item will be tested under three conditions:

- Direct response: the model receives the task without external evidence.
- Evidence provided: the model receives a short, source-controlled context.
- Evidence plus verification: the model answers, checks each factual claim against the supplied context, and revises unsupported statements.

Each stochastic condition will be run three times. Item order will be randomized. Prompts will keep meaning and task difficulty as close as possible across languages; a native-language reviewer will check every final item before testing.

Data management

The project will use public or appropriately licensed text and researcher-authored evaluation items. The team will keep a provenance record for every item. Raw model outputs, scoring decisions, code versions, and exclusions will be stored in access-controlled project folders with dated version history. The released dataset will omit copyrighted passages that cannot be redistributed.

Evaluation and analysis

Measure	Definition	Reporting
Correctness	Two reviewers score 0 (incorrect), 1 (partly correct), or 2 (fully correct) against the answer key.	Mean score and share fully correct, by model, language, task, and condition.
Unsupported claims	Claims that lack support in the reference source or supplied evidence.	Claims per response and responses containing at least one unsupported claim.
Citation validity	The cited source exists and supports the associated statement.	Valid citations divided by all citations produced.
Translation quality	Human adequacy and fluency scores, plus chrF++ where reference translations exist.	Scores by language direction, model, and condition.
Consistency	Agreement in conclusion across repeated runs of the same item.	Pairwise agreement and response-level variance.
Reviewer agreement	Agreement between independent human ratings.	Cohen's kappa or weighted kappa, with adjudication counts.

Analysis plan

Primary comparisons will use paired differences between languages and experimental conditions on matched items. The report will include bootstrap confidence intervals and effect sizes, not only significance tests. The team will analyze disagreements and publish the error taxonomy. Exclusion rules and metric definitions will be fixed before the full benchmark run.

Eight-month work plan

Month	Planned work	Evidence produced
1	Protocol, literature review, source inventory, and scoring rubrics.	Protocol v1, bibliography, item log.
2-3	Build and review the 180-item pilot; implement experiment logging.	Pilot dataset, code repository, review records.
4	Run pilot experiments and revise items, prompts, and rubrics.	Dated runs, pilot analysis, protocol amendment.
5-6	Complete the 600-item benchmark and full model evaluation.	Locked dataset, outputs, rating records.
7	Statistical analysis, error review, and replication checks.	Analysis notebook and results tables.
8	Prepare the report, dataset documentation, and public release.	Preprint or report, repository, release notes.

Research integrity and dissemination

Safeguards

- The initial phase will not recruit students, children, program beneficiaries, or other human research participants.
- The benchmark will exclude personal data and prompts that request individualized medical, legal, financial, or security advice.
- Adult language reviewers will receive written instructions, clear attribution rules, and appropriate confidentiality terms.
- Any later study that collects data about people or measures learner outcomes will require a separate ethics determination before recruitment.
- Researchers will document model access dates and versions because hosted systems can change during a study.

Transparency and release

OTED plans to publish the final report after internal quality review. The public research package will include the protocol, benchmark documentation, scoring rubric, code, aggregate results, and a model card-style limitations statement. Dataset files will be released only when OTED has the right to redistribute them. The report will distinguish measured findings from interpretation and will identify unresolved limitations.

Use of Claude

Claude will support literature synthesis, code development, data-quality checks, experiment documentation, error analysis, and manuscript preparation. Where Claude is itself evaluated, the team will preserve the exact prompts, settings, access date, and raw outputs used in the analysis. Human reviewers will make the final scoring decisions.

Project stage

Work began in September 2026 with the research question, pilot design, evaluation measures, and eight-month schedule. The next phase will build the pilot set, onboard language reviewers, and run controlled pilot experiments. Findings from the pilot will guide the full benchmark.

References

1. NLLB Team et al. (2022). No Language Left Behind: Scaling Human-Centered Machine Translation. arXiv:2207.04672.
2. Liang, P. et al. (2023). Holistic Evaluation of Language Models. Transactions on Machine Learning Research; arXiv:2211.09110.
3. National Institute of Standards and Technology (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1.
4. Sindhuja, A., Kanojia, D., Orasan, C., and Qian, S. (2025). When LLMs Struggle: Reference-less Translation Evaluation for Low-resource Languages. Proceedings of LoResLM 2025, 437-459.
5. Bokaei, Z., Webber, B., and Magdy, W. (2026). Benchmarking Offensive Language Detection in Persian and Pashto. Proceedings of SilkRoadNLP 2026, 13-23.

ORGANIZATION	RESEARCH PROGRAM
Organization for Talent and Entrepreneurship Development (OTED) oted.org oted.us.org info@oted.org	OTED AI Research Initiative Internal computer science research program Research Lead / PI-equivalent governance