

Opportunity Title: Optimizing Sub Word Tokenization Techniques for Low-Resource Languages and/or Non-Latin Scripts to Enhance LLM Performance

Opportunity Reference Code: ICPD-2025-28

Organization Office of the Director of National Intelligence (ODNI)

Reference Code ICPD-2025-28

How to Apply **Create and release your Profile on Zintellect** – Postdoctoral applicants must create an account and complete a profile in the on-line application system. **Please note: your resume/CV may not exceed 3 pages.**

Complete your application – Enter the rest of the information required for the IC Postdoc Program Research Opportunity. The application itself contains detailed instructions for each one of these components: availability, citizenship, transcripts, dissertation abstract, publication and presentation plan, and information about your Research Advisor co-applicant.

Additional information about the IC Postdoctoral Research Fellowship Program is available on the program website located at: <https://orise.orau.gov/icpostdoc/index.html>.

If you have questions, send an email to ICPostdoc@orau.org. Please include the reference code for this opportunity in your email.

Application Deadline 2/28/2025 6:00:00 PM Eastern Time Zone

Description **Research Topic Description, including Problem Statement:**

Improving tokenization in large language models (LLM's) is a crucial area of research as it directly impacts the model's ability to understand and generate text. It has been shown that poor tokenization impacts both the performance of models and their training efficiency (Mehdi et al, 2024). English-centric tokenization approaches can lead to a 62% loss in training efficiency with equivalent degradation in inference and generative capabilities.

This research program will seek to find more appropriate methods for tokenizing multilingual data with a focus on low resource languages and languages with non-Latin orthographies.

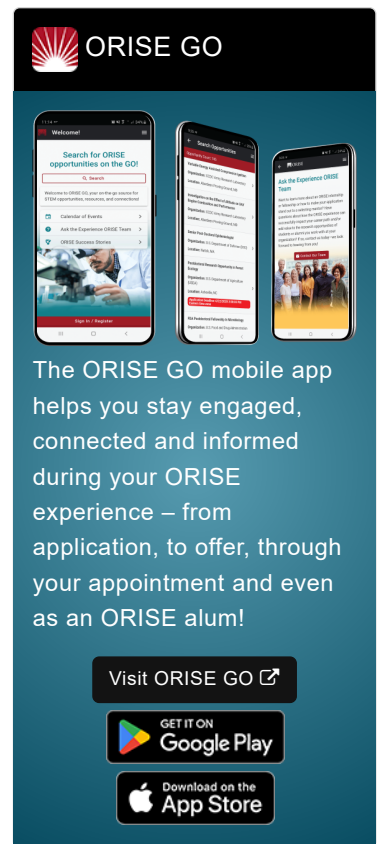
Example Approaches:

There are several ways that this problem could be approached, including but not limited to;

- investigating the effectiveness of novel sub-word tokenization techniques on target languages
- exploration of tokenization techniques that balance the needs of multiple languages, focusing on the maintenance of performance and reduction of tokenization errors
- investigate tokenization approaches for languages with a range of different morphological complexities
- a study of the relationship between tokenization methods and model interpretability.


Relevance to the Intelligence Community:


The development of LLM's is largely considered to be out of reach of small companies, academia and the government sector due to the large amount




ORISE GO

The ORISE GO mobile app helps you stay engaged, connected and informed during your ORISE experience – from application, to offer, through your appointment and even as an ORISE alum!

Visit ORISE GO 

GET IT ON
 Google Play

Download on the
 App Store

Opportunity Title: Optimizing Sub Word Tokenization Techniques for Low-Resource Languages and/or Non-Latin Scripts to Enhance LLM Performance

Opportunity Reference Code: ICPD-2025-28

of compute required to train them. Additionally, they do not perform well in the multi-lingual context particularly when low-resource languages and non-Latin scripts form part of the training data. The intelligence community will benefit from the increased performance and reduction in training costs. Gains in interpretability will also provide significant improvements in our ability to govern these types of models.

References:

- Ali, M., Fromm, M., Thellmann, K., Rutmann, R., Lübbering, M., Leveling, J., Klug, K., Ebert, J., Doll, N., Buschhoff, J.S. and Jain, C., 2023. Tokenizer Choice For LLM Training: Negligible or Crucial?. arXiv preprint arXiv:2310.08754.

Key Words: LLM, multilingual tokenization, low resource languages, orthography.

Qualifications **Postdoc Eligibility**

- U.S. citizens only
- Ph.D. in a relevant field must be completed before beginning the appointment and within five years of the appointment start date
- Proposal must be associated with an accredited U.S. university, college, or U.S. government laboratory
- Eligible candidates may only receive one award from the IC Postdoctoral Research Fellowship Program

Research Advisor Eligibility

- Must be an employee of an accredited U.S. university, college or U.S. government laboratory
- Are not required to be U.S. citizens

Point of Contact [Keri Tarwater](#)

Eligibility Requirements

- **Citizenship:** U.S. Citizen Only
- **Degree:** Doctoral Degree.
- **Discipline(s):**
 - **Chemistry and Materials Sciences** ([12](#))
 - **Communications and Graphics Design** ([3](#))
 - **Computer, Information, and Data Sciences** ([17](#))
 - **Earth and Geosciences** ([21](#))
 - **Engineering** ([27](#))
 - **Environmental and Marine Sciences** ([14](#))
 - **Life Health and Medical Sciences** ([45](#))
 - **Mathematics and Statistics** ([11](#))
 - **Other Non-Science & Engineering** ([2](#))
 - **Physics** ([16](#))
 - **Science & Engineering-related** ([1](#))
 - **Social and Behavioral Sciences** ([30](#))