**Proposal: Enhancing Climate Co-Benefits Tagging Framework in Fragile, Conflict, and Violence-Affected Contexts Using NLP and LLM Techniques**

**Project Title**

**"Refining Climate Co-Benefits Classification with NLP and LLMs for Fragile Contexts"**

**Context and Challenge**

Fragile, Conflict, and Violence-affected (FCV) regions face unique climate adaptation and mitigation challenges, intertwined with socio-economic and security vulnerabilities. The current World Bank framework for tagging climate co-benefits does not adequately account for the complexities of FCV contexts, limiting the identification of critical adaptation and mitigation measures.

This project addresses the need for an enhanced tagging framework that integrates FCV-specific considerations, leveraging advanced Natural Language Processing (NLP) and Large Language Models (LLMs). The ultimate goal is to improve the World Bank's ability to allocate resources effectively, measure the impact of climate actions, and support climate resilience in FCV regions.

**Beneficiaries**

1. **Development Organizations and Policymakers**: Enhanced tagging enables better decision-making, resource allocation, and impact measurement.
2. **Communities in FCV Regions**: Improved project alignment with their unique vulnerabilities and resilience needs.
3. **University Students and Data Scientists**: Gain real-world experience applying NLP and LLMs to solve global challenges.
4. **Global Development Community**: A replicable framework for addressing climate resilience challenges in fragile settings.

**Goals and Objectives**

1. **Framework Development**: Develop an FCV-sensitive tagging framework for climate co-benefits.
2. **Model Innovation**: Implement and compare traditional NLP models and state-of-the-art LLMs for classification and tagging.
3. **Resource Allocation Support**: Enable data-driven prioritization of climate resilience investments.
4. **Scalable and Reproducible Solution**: Build a methodology that can be adapted to future projects.
5. **Research Contribution**: Produce a research paper to document methodology, insights, and comparative results.

**Data Details**

* **Primary Data**: World Bank Project Appraisal Documents (PADs), including project descriptions, component details, and expected outcomes.
* **Supplemental Data**: Geographic and socio-economic data, such as conflict intensity and climate risk exposure, to enhance tagging accuracy.
* **Data Size**: Approximately 1-5 GB, primarily textual.

**Proposed Methodology**

**Phase 1: Data Collection and Preprocessing**

* **Data Compilation**: Extract relevant sections from PADs, focusing on project descriptions, components, and expected outcomes.
* **Preprocessing**: Clean data, tokenize text, remove stop words, and handle domain-specific terminology.
* **Tagging Framework**: Develop a lexicon of FCV-sensitive terms to guide initial tagging and annotation.

**Phase 2: Model Development and Fine-Tuning**

1. **Traditional NLP Models**
   * Train and evaluate models like **Support Vector Machines (SVM)**, **Random Forests**, and **Naive Bayes** with feature extraction using TF-IDF and Bag-of-Words.
   * Use topic modeling (e.g., **Latent Dirichlet Allocation**) to identify underlying themes in the text.
2. **Advanced LLMs**
   * Fine-tune BERT-based models (e.g., BERT, RoBERTa, or DistilBERT) for FCV-specific classification.
   * Leverage state-of-the-art LLMs, such as **GPT-4** or **Google Gemini**, for zero-shot or few-shot classification of climate co-benefits.
3. **Named Entity Recognition (NER)**
   * Develop an NER pipeline using pretrained models to identify and categorize FCV-specific terms like “conflict-sensitive infrastructure” and “natural disaster risk reduction.”
   * Incorporate a custom lexicon of FCV-sensitive terms to enhance NER performance.
4. **Hybrid Approach**
   * Combine predictions from traditional models and LLMs using ensemble techniques to improve classification accuracy.

**Phase 3: Validation and Refinement**

* Validate model outputs with subject-matter experts in FCV and climate resilience.
* Incorporate feedback and additional training data to improve tagging accuracy.
* Evaluate performance using metrics such as precision, recall, F1 score, and accuracy.

**Phase 4: Deployment and Reporting**

* Deploy the final model on a pilot set of PADs to classify climate co-benefits.
* Generate a detailed methodology report and provide recommendations for integrating the model into the World Bank’s framework.

**Phase 5: Knowledge Sharing and Research Dissemination**

* Host a workshop with stakeholders to share findings and lessons learned.
* Draft a research paper highlighting the comparative performance of traditional NLP methods and LLMs in addressing climate co-benefits in FCV contexts.

**Deliverables**

1. **Enhanced Tagging Framework**: A refined classification system incorporating FCV-specific considerations.
2. **NER Pipeline**: A robust system to identify FCV-relevant terms and entities in PADs.
3. **Comparative Model Analysis**: Insights into the efficacy of traditional NLP models versus LLMs.
4. **Pilot Deployment**: A model deployed on a subset of PADs, demonstrating real-world applicability.
5. **Research Paper**: A publication documenting the methodology, results, and implications for development practice.

**Expected Outcomes**

1. **Improved Accuracy**: A classification framework tailored to FCV contexts, ensuring accurate tagging of climate co-benefits.
2. **Resource Optimization**: Enhanced capacity for data-driven resource allocation and strategy development in fragile regions.
3. **Knowledge Generation**: Contributions to the field of development analytics through innovative applications of NLP and LLMs.
4. **Scalability**: A replicable tagging methodology adaptable to other organizations and contexts.