Rebuilding Trust in Historical Records: Leveraging Deep Learning and Generative AI for Enhanced Data Accuracy and Comprehensive Dataset Collection

In the context of the rapidly evolving digital landscape, the integrity and reliability of historical records have emerged as critical concerns for researchers in the humanities and social sciences. Traditional methodologies often fall short in addressing the complexities and nuances inherent in historical documents, leading to transcription errors and misinterpretations that can significantly skew data analysis. This study introduces innovative approaches that leverage the capabilities of deep learning and generative artificial intelligence (AI) to enhance the accuracy and trustworthiness of historical records.

My research, conducted in partnership with the Minnesota Population Center, focuses on the monumental task of scrutinizing over 15 million pages of U.S. Census records. The objective was to rectify approximately 1-2 million instances of "extra persons" erroneously recorded due to transcription errors or OCR errors. These inaccuracies, arising from common misinterpretations such as confusing "No One" with "Noah" or "Vacant" with "Vincent," underscore deeper challenges related to trust and verification within the historical record. have developed a tool that achieved 95% accuracy on a large validation dataset, successfully identifying and correcting these inaccuracies.

I also implemented innovative semantic segmentation models to enhance the precision in localizing tasks within Optical Character Recognition (OCR) processes. These models dissect documents into semantically meaningful segments, enabling more effective identification of inconsistencies. A key aspect of my methodology is the use of transfer learning, which leverages pre-trained models to bring a rich understanding of visual patterns to historical document analysis. This accelerates the learning process and enhances model performance. Additionally, I explored advanced scaling techniques to efficiently manage large image datasets, maximizing GPU utilization to significantly reduce processing times and achieve unprecedented scalability in analysis.

A key aspect of my research involved using a Large Language Model (LLM) to address the complex challenge of name matching within a dataset, which included variations in spelling. This task was tested on a specially designed dataset of thousands of artificially generated individual names, tailored to replicate difficult scenarios such as matching full names with their abbreviations and distinguishing individuals with identical first and last names but different middle names. The LLM's performance was outstanding, achieving a 93.57% accuracy rate. This highlighted its capability to accurately identify suffixes,

prefixes, and adapt to spelling variations, demonstrating AI's significant potential to improve historical data analysis precision.

An example of the LLM's effectiveness was its precise matching of "Ms. Bethany Heather Maria Gordon L.L.B." with “Ms. Bethany Heather Maria Gordon L.L.B.” from a list of candidates. This showcased the model's ability to discern and accurately match complex name components, reflecting AI's nuanced approach to data interpretation.

This work underscores the transformative impact of Large Language Models and deep learning technologies in advancing social sciences and humanities research. By applying these technologies to historical document analysis, my study enhances the reliability and accuracy of historical records, contributing to discussions on identity, kinship, and community in the digital age. This innovative effort has wide implications for scholars in the social sciences, setting new benchmarks for integrating advanced technology into historical research.