README *for* Leveraging Large Language Models for Fuzzy String Matching in Political Science

This README file contains three sections. The first section, *General Description*, describes the contents in this supplementary folder. The second section, *Hardware and Software Requirements*, describes what hardware is used for running the scripts and the relevant packages. The third section, *Instructions*, reports on how to run the scripts.

1 General Description

The replication package contains the following files:

- README.pdf
- evaluation_set.csv: this file comes from Kaufman & Klevs (2022).
- data.csv: this file is derived from Kaufman & Klevs (2022).
- experiment_1.py: generates experiment_1_temp0.2.csv and experiment_1_temp0.2.csv.
- experiment_1_analysis.py: generates the key metrics for Experiment #1.
- figure_1.py: generates the 1st figure in the paper.
- Survey_Final_Results.csv: base results from Kaufman & Klevs (2022).
- experiment_2.py: core script for generating results in experiment_2_v{1|2}_temp{0.2|1.0}.csv.
- experiment_2_analysis: generates the key metrics for Experiment #2.
- figure_2.py: generates the 2nd figure in the paper.

2 Hardware and Software Requirements

All the scripts are run on a MacBook Pro, with a Apple M1 Pro chip, 16 GB memory. The required python packages are matplotlib, openai, and pandas.

3 Instructions

To generate the raw results with ChatGPT scores, we run the experiment_1.py and experiment_2.py.

- To generate the raw results with ChatGPT scores, we run the experiment_1.py and experiment_2.py.
- To generate the key metrics for each experiment, we run the experiment_1_analysis.py and experiment_2_analysis.py.
- To generate the metrics figure for each experiment, we run the figure_1.py and figure_2.py.

References

Kaufman, A. R., & Klevs, A. (2022). Adaptive fuzzy string matching: How to merge datasets with only one (messy) identifying field. *Political Analysis*.