Aligning Variables across Cohort Studies: A Comparative Study of Large Language Models and Fuzzy Match Approaches

Zexu Li, MS[1], Suraj P. Prabhu, MS[2], Zachary T. Popp, MPH[1], Ting Fang Alvin Ang, MD, MPH[1,3,4], Rhoda Au, PHD[1,3,4,5], Jinying Chen, PHD[6,7]

1.Department of Anatomy and Neurobiology, Neurology and Medicine, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, USA; 2.Department of Bioinformatics, Boston University school of Engineering and Graduate School of Art & Science, Boston, MA, USA ; 3. Framingham Heart Study, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, USA; 4. Slone Epidemiology Center, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, USA; 5. Department of Epidemiology, Boston University School of Public Health, Boston, MA, USA; 6. Department of Medicine/Section of Preventive Medicine and Epidemiology, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, USA; 7. Data Science Core, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, USA

Introduction

➢Merging data collected by multiple studies (i.e., data harmonization, meta-analysis) is a common strategy to increase sample size and statistical power of data analysis. However, even studies using similar research protocols may used different variable naming conventions and coding schemes.

➢Objective: to develop and validate Natural Language Processing (NLP) methods that align variables from different studies to support data harmonization.

❑ Are NLP methods applicable to the variable alignment task?

❑ Which NLP methods have the best performance on variable alignment?

Data and Sample

➢Source of evaluation data: data variables from European and Japan GERAS cohort studies.

❑ Similar protocol was used to collect data across these 2 cohorts, but data variables were coded and named differently.

	Japan Cohort (324 Variables)	EU Cohort (928 Variables)
Variable Label	ADTTERM:AD Treatment Name	SDYTRTTERM: Study Treatment Dictionary Term
Data Sheet	ADTR: All AD medication as recorded	SDYTRT: Study Treatment
Variable Definition	Donepezil, Galantamine, Memantine, Rivastigmine, Yokukansan (Chinese herbal medicine)…	Approved AD treatment Donepezil Galantamine Investigational product Memantine Rivastigmine…

Methodology

Figure 2. Large Language Model for variable alignment

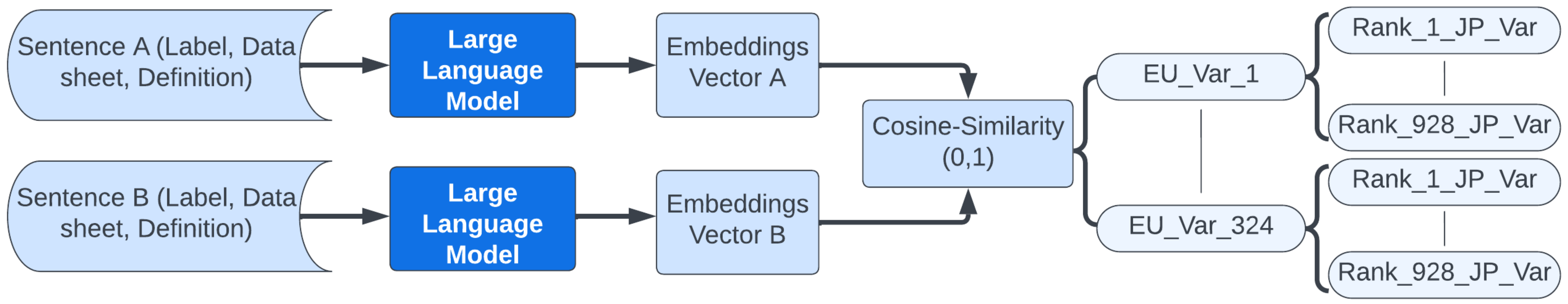


Figure 3. Fuzzy Match method for variable alignment

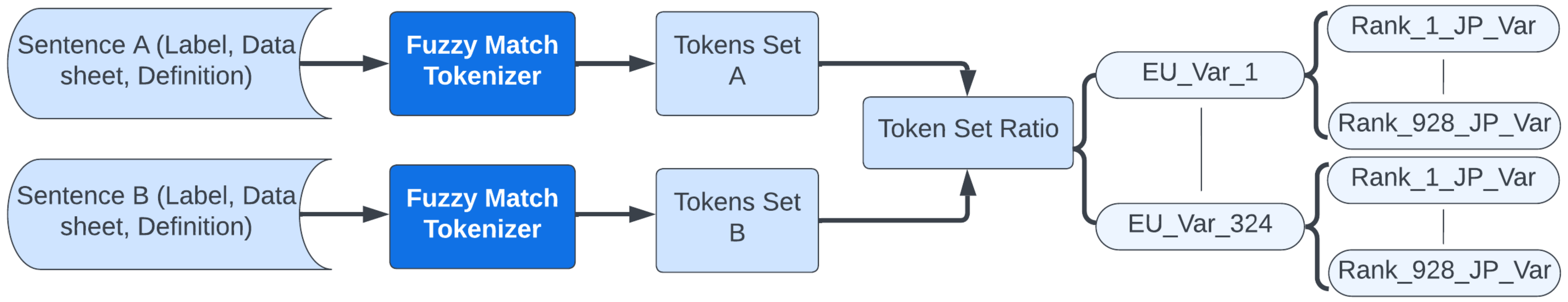


Table 1. NLP Methods Detail Descriptions

Models	Training Data	Description	Complexity
E5-Large-V2	270M sentences pairs	Large language model (LLM), extension of Bidirectional Encoder Representations from Transformers (BERT)	CPU times: total: 27min 28s Wall time: 4min 17s
All-MiniLM-L12-V2	1B sentences pairs	LLM, extension of BERT	CPU times: total: 2min 40s Wall time: 53.1 s
All-Mpnet-base-v2	1B sentences pairs	LLM, extension of BERT	CPU times: total: 6min 37s Wall time: 1min 29s
Token-set-ratio	None	Fuzzy Match Based on Tokens	Total time: 13.1s

➢Evaluation metrics: Hit Ratio, Mean Reciprocal Rank, Mean Average Precision

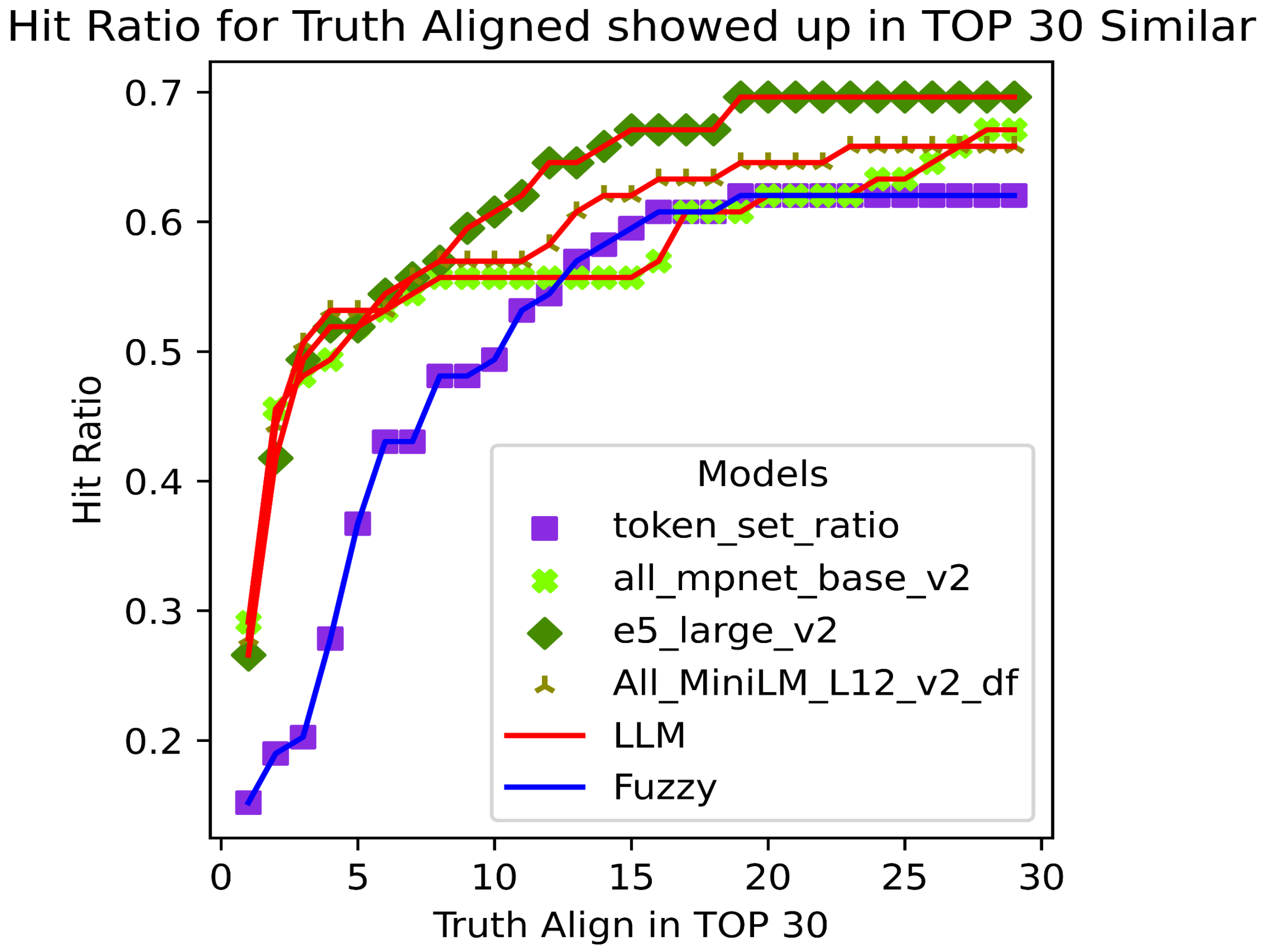
❑ Hit Ratio: Proportion of correct alignments (between source and target variables) in the top-n target variables ranked by the NLP algorithms.

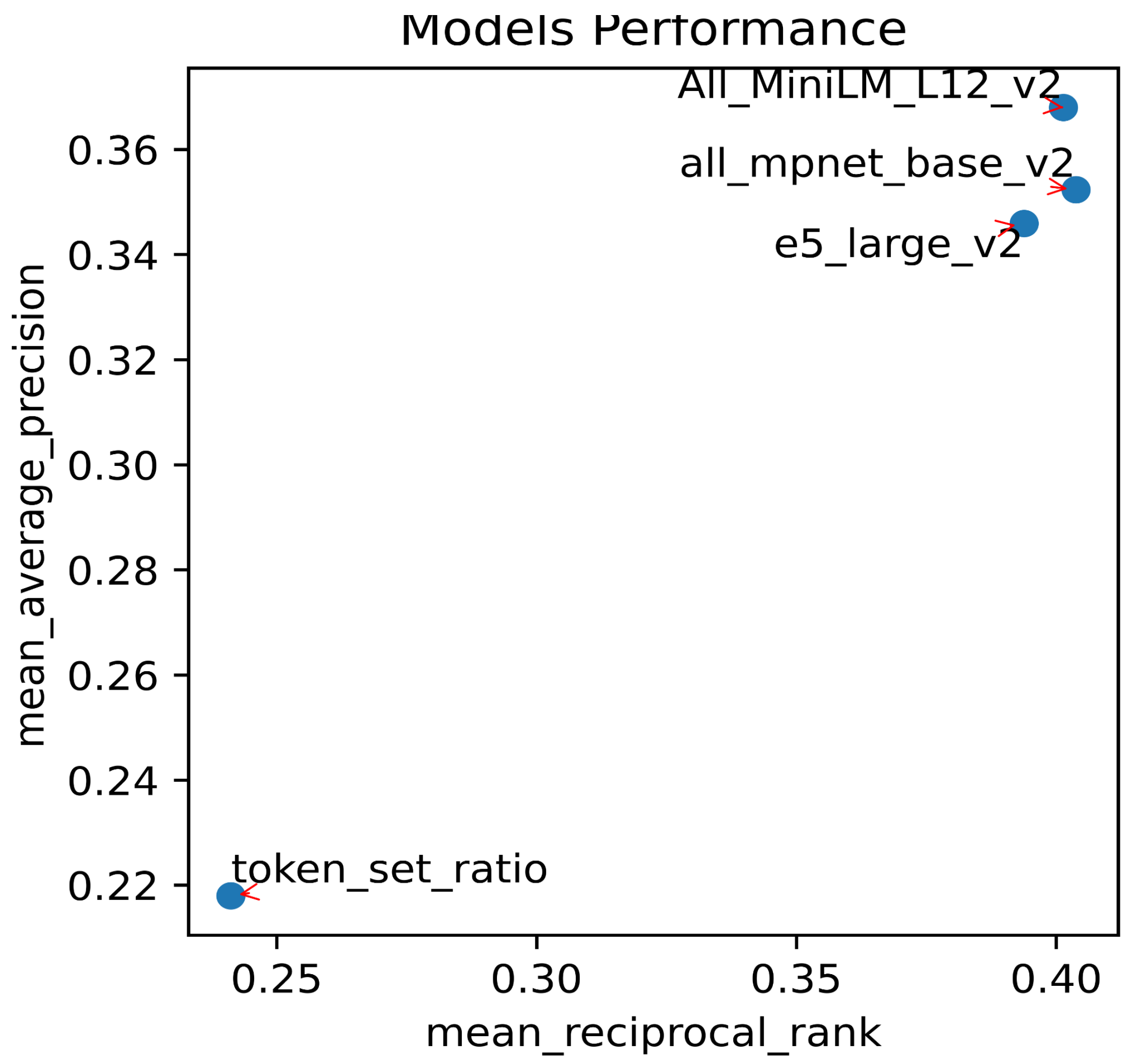
❑ Mean Reciprocal Rank: Mean value of reciprocal rank (one divided by the rank of first appeared correctly aligned target variable).

❑ Mean Average Precision: Mean value of average precision (consider ranks of all correctly aligned target variables).

❑ Truth Map/evaluation set: 160 pairs of source (EU) and target (Japan) variables that were manually identified and validated by three co-authors (ZL, SP, ZTP).

Results





Conclusion

➢ NLP methods showed adequate results for variable alignment tasks.

➢ LLMs outperformed fuzzy match for aligning variables.

➢ Among the LLMs, the E5 model has the best performance and MiniLM model has the lowest running time.

Future Work

➢ Improve LLM models using task-specific training data.

➢ Validate approach with other datasets.

Acknowledgements: This study is supported by Alzheimer’s Disease Data Initiative (ADDI). JC was also supported by a 2023 pilot award from the Framingham Heart Study Brain Aging Program, which is funded by the National Institute on Aging (U19-AG068753), which also provides support to RA, AA. The American Heart Association (20SFRN35360180) provides funding support to RA, ZL, ZP. The funders had no decisional role in study design, data collection and analysis, interpretation of data, or preparation of the abstract.