

Register Always Matters: Analysis of LLM Pretraining Data Through the Lens of Language Variation

Amanda Myntti, Erik Henriksson, Veronika Laippala, Sampo Pyysalo

TurkuNLP, University of Turku



We evaluate the effects of the **linguistic register (or genre)** on LLM performance by training identical generative models with register-filtered datasets. We then evaluate and compare these models using well-known benchmarks, revealing how each register impacts model capabilities.

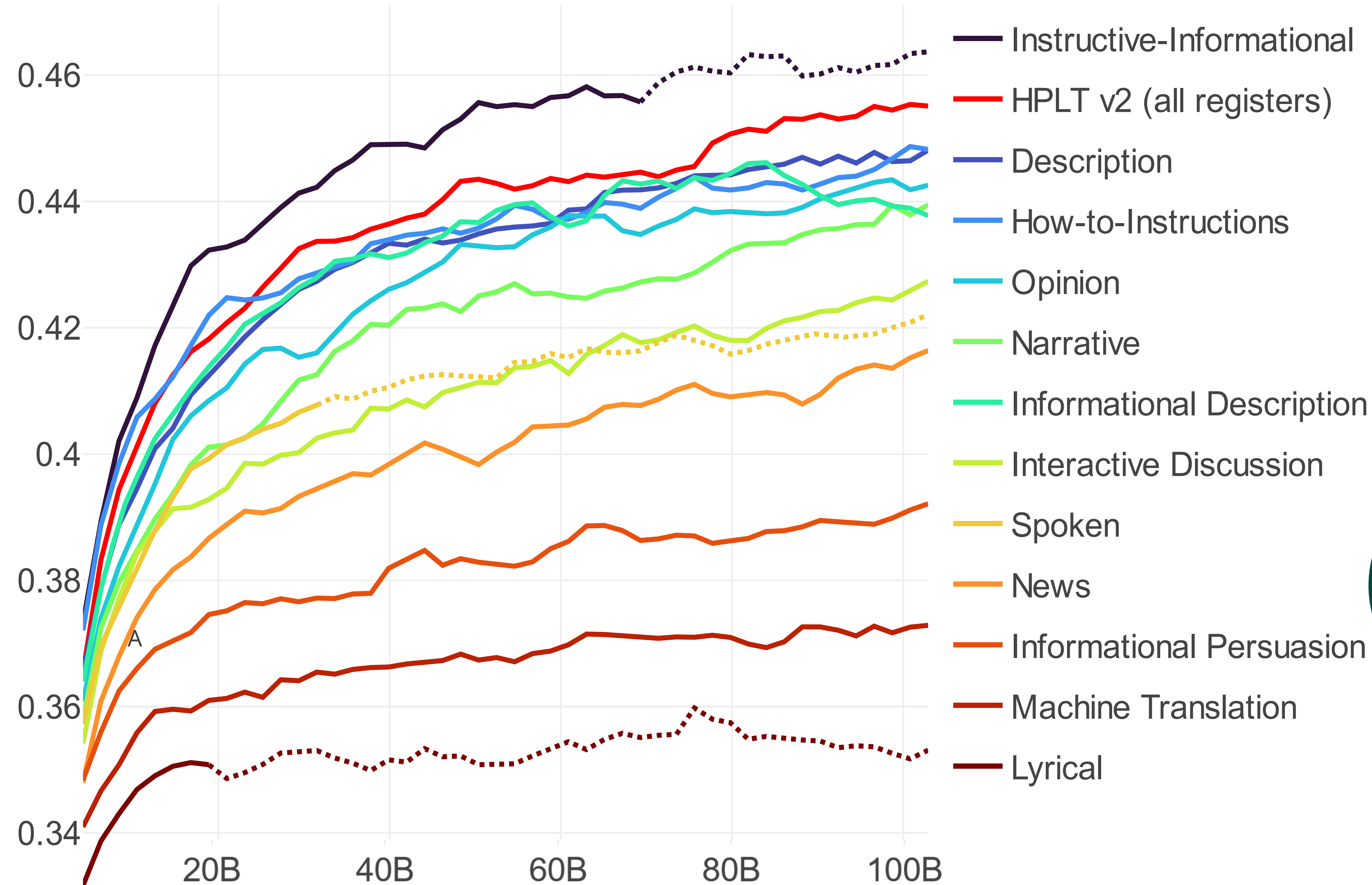
REGISTERS

Situationally characterised text varieties, including categories such as *news*, *reviews* and *song lyrics*

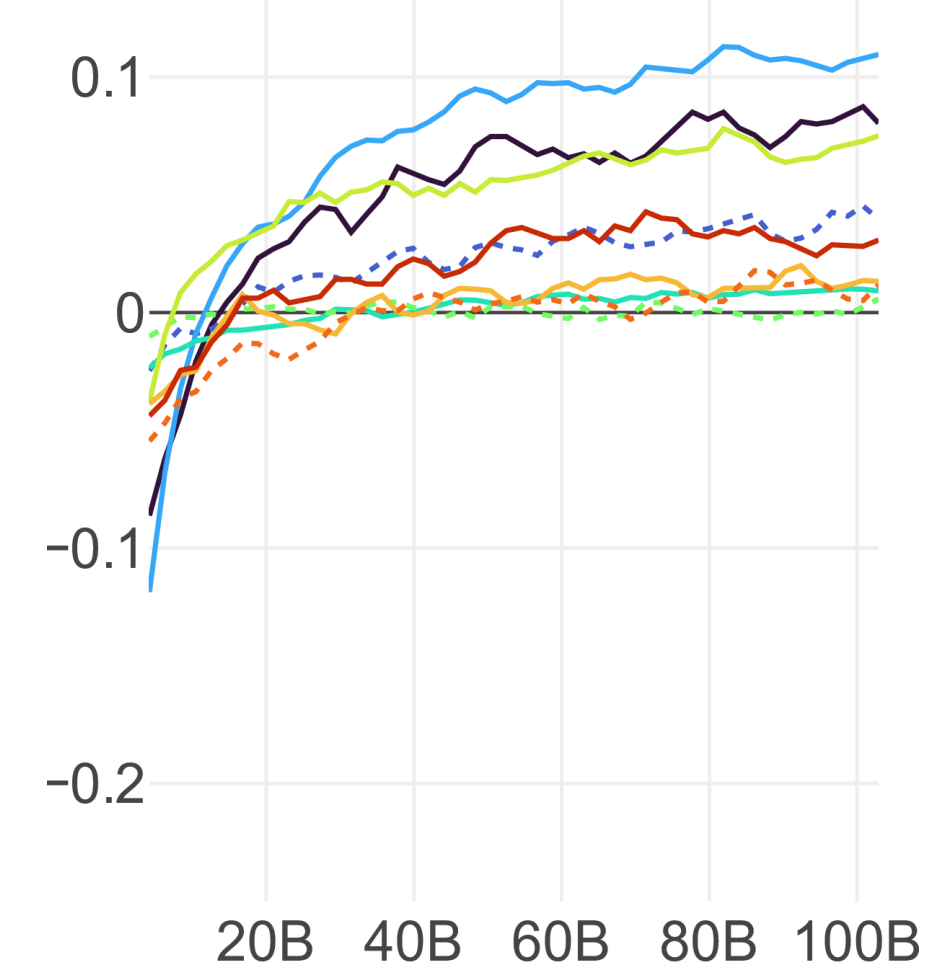
DATA

Register classified HPLT v2 [1]

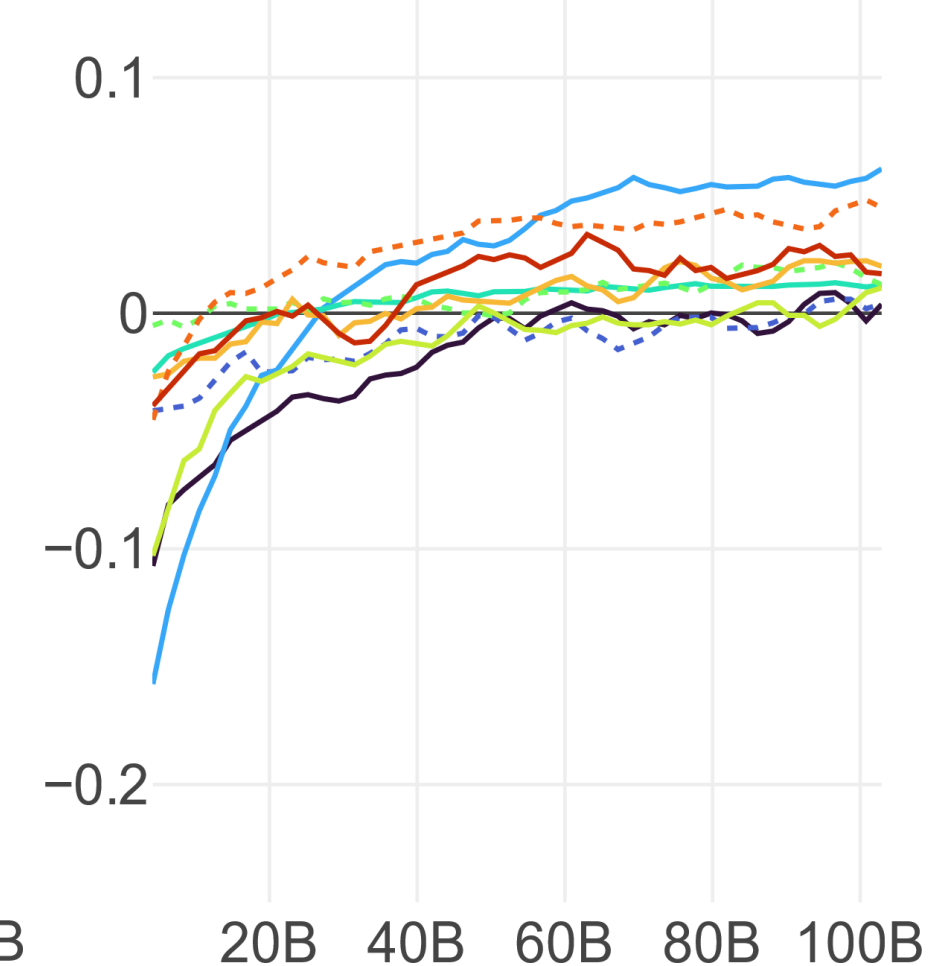
Our results show that although the ranking of registers somewhat aligns with intuitions about their content, like informational and instructive texts yielding powerful models, surprisingly, *Opinion* yields the 4th best performing model, and the *News* register performs among the worst. Combinations of selected registers further improve performance.



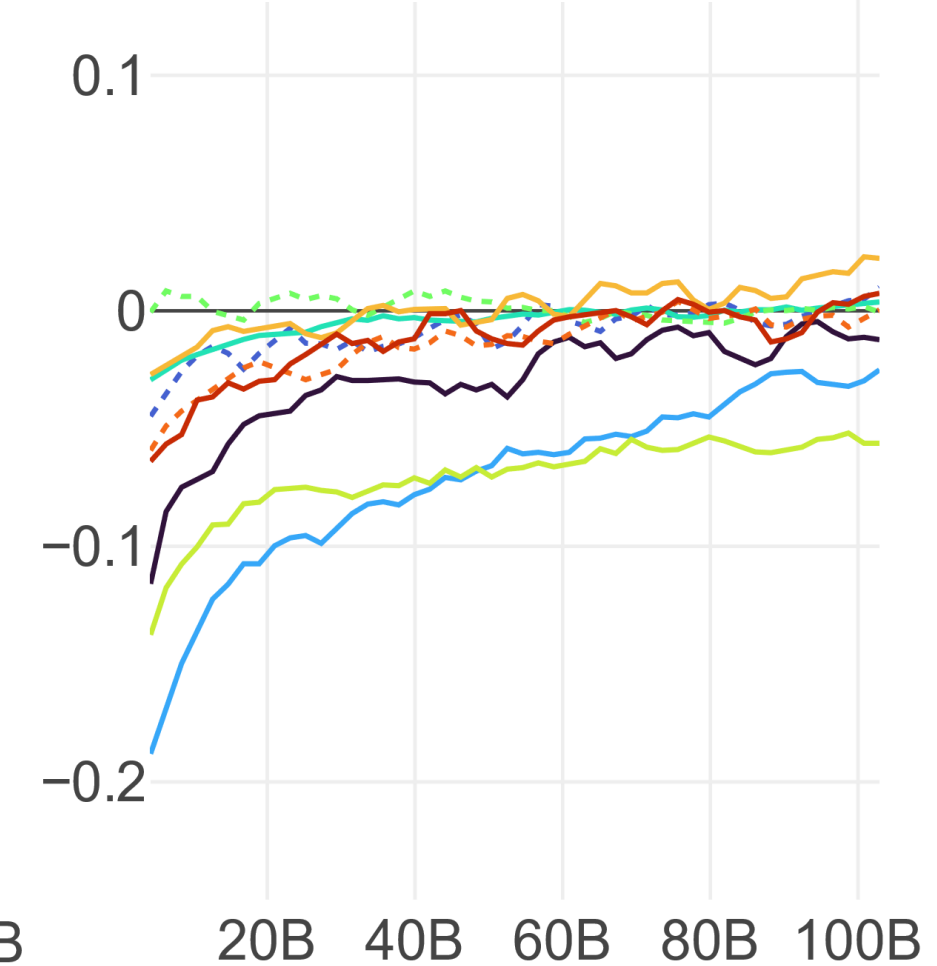
Instructive-Inform.



Opinion



News

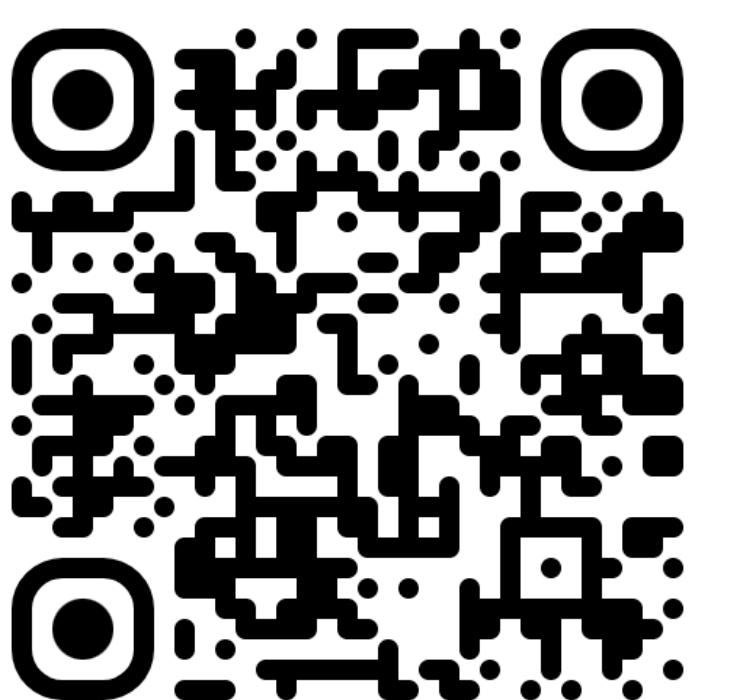


MODELS
Llama 1.71B,
trained up to
100B tokens

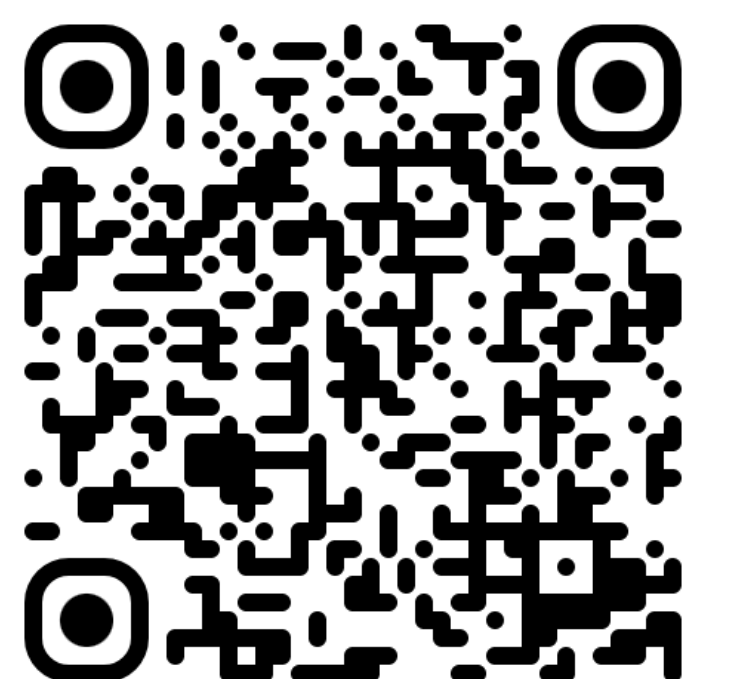
BENCHMARKS

HellaSwag [2],
WinoGrande [3], PIQA
[4], SIQA [5],
OpenBookQA [6], ARC
Easy & Challenge [7],
CommonsenseQA [8],
and MMLU [9].

We additionally
evaluate
combination of
registers! See our
full paper



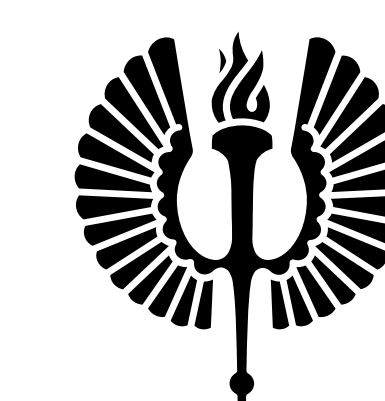
Full paper



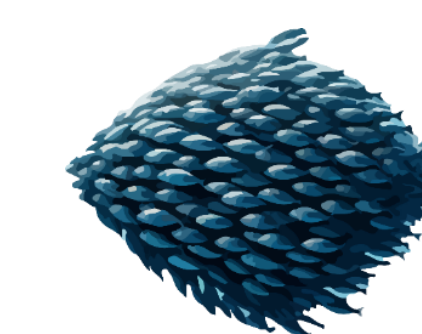
This poster

[1] Burchell et al. (2025) An Expanded Massive Multilingual Dataset for High-Performance Language Technologies (HPLT)
[2] Zellers et al. (2025) HellaSwag: Can a Machine Really Finish Your Sentence?
[3] Sakaguchi et al. (2021) WinoGrande: an adversarial winograd schema challenge at scale
[4] Bisk et al. (2019) PIQA: Reasoning about Physical Commonsense in Natural Language
[5] Sap et al. (2019) Social IQa: Commonsense Reasoning about Social Interactions
[6] Mihaylov et al. (2018) Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering
[7] Clark et al. (2018) Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge
[8] Talmor et al. (2019) CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge
[9] Hendrycks et al. (2020) Measuring Massive Multitask Language Understanding

Methodology inspired by Penedo et al. (2024) The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale



**UNIVERSITY
OF TURKU**



AI-DOC
Finnish Doctoral Program Network
in Artificial Intelligence



High Performance
Language Technologies

