University of Udine

Department of Mathematical, Computer and Physical Sciences

Master's Degree in Multimedia Communication and Information Technologies

Recommender Systems – A.A. 2024/25

Lecturer: prof. Kevin Roitero

Intelligent selection of question subsets for benchmarking LLMs

Student: Angela Rossi

Registration number: 174288

PROJECT OBJECTIVES

The project aims to explore strategies to reduce the size of the benchmarks used in the evaluation of Large Language Models (LLMs), while maintaining a good approximation of the metric obtained on the entire set of questions.
The central hypothesis is that, by intelligently selecting a representative subset of questions, it is possible to reduce the computational cost of the evaluation, without significantly compromising the reliability of the result.

DATASET AND EXPERIMENTAL SETUP

The starting dataset is a subset of the MMLU benchmark containing 3 topics:

- high_school_macroeconomics
- professional_law
- professional_psychology

Each entry contains:

- The text of the question (question)
- The answers (choices)
- The index of the correct answer (answer)
- Embedding the question
- The output of the model and whether it is correct (correct)

For now the project focuses ona single LLM model, evaluated by the metric correct.

BASELINE: RANDOM SELECTION

In the first experiment I calculated the average of correct answers by randomly extracting increasingly larger groups of questions (k=1, 2, . . . , 300). For each k, the extraction was repeated 1000 times and from these the average of the averages was obtained. This operation was performed independently for each topic.

Results

The graphs show that:

- The sample mean converges rapidly to the global mean

- Already with k between 50 and 100, the estimate becomes very stable.

- The curves behave differently in different topics: noisier where the variance is higher

REPRESENTATIVE SELECTION WITH KMEANS CLUSTERING

As an alternative strategy to random selection, I used KMeans clustering to select representative question subsets based on question vector embeddings.

The idea is that by grouping the questions into k clusters, it is possible to choose the question closest to the centroid of each cluster, thus obtaining a subset that covers the content of the dataset in a broad and balanced way.

Method

For each of the three topics: professional_psychology, professional_law, and high_school_macroeconomics, I applied the KMeans algorithm testing from 1 to 300 clusters; for each k I identified the question closest to each centroid. I then went to calculate the average of the correct metric on the k questions thus chosen and finally archived the results, comparing them with those obtained through random selection.

- Across all topics, KMeans tends to stabilize close to the global mean, albeit with some fluctuations.

- The approach provides a more structured selection than random chance.

- In alcuni casi (es.professional_law), convergence is less smooth, suggesting that embedding does not always perfectly capture the domain structure.

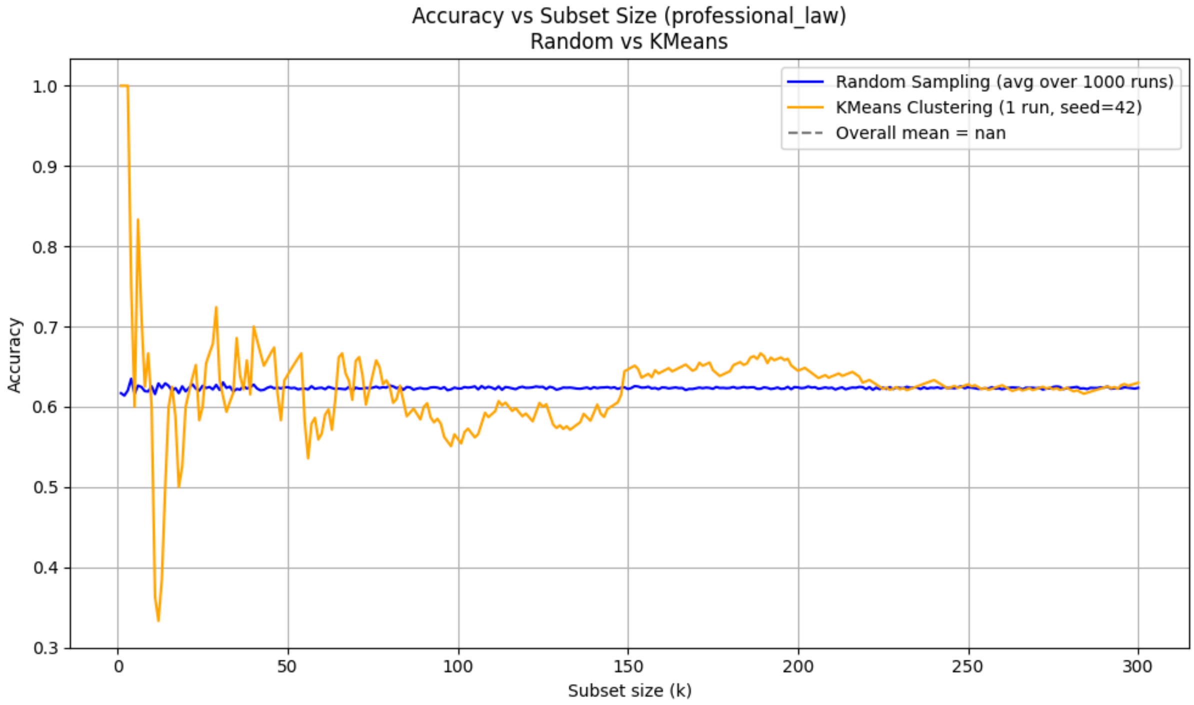
- Representative selection works best on topics with more homogeneous content (high_school_macroeconomics).

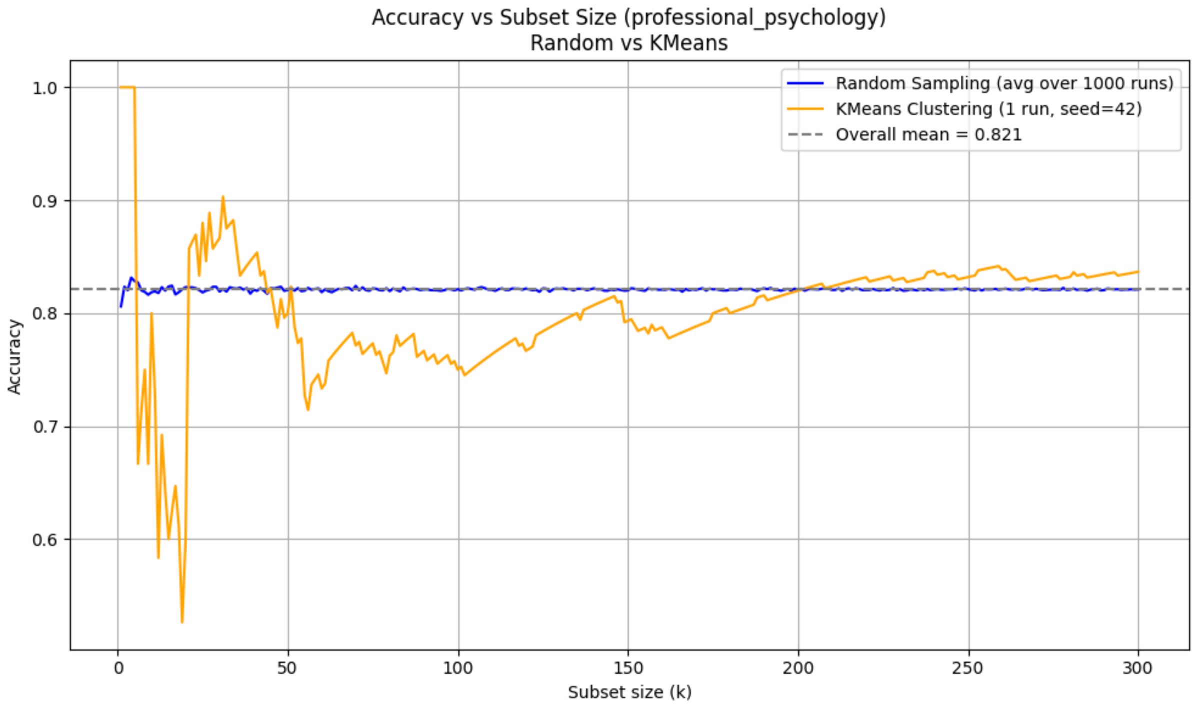
COMPARISON BETWEEN RANDOM AND KMEANS

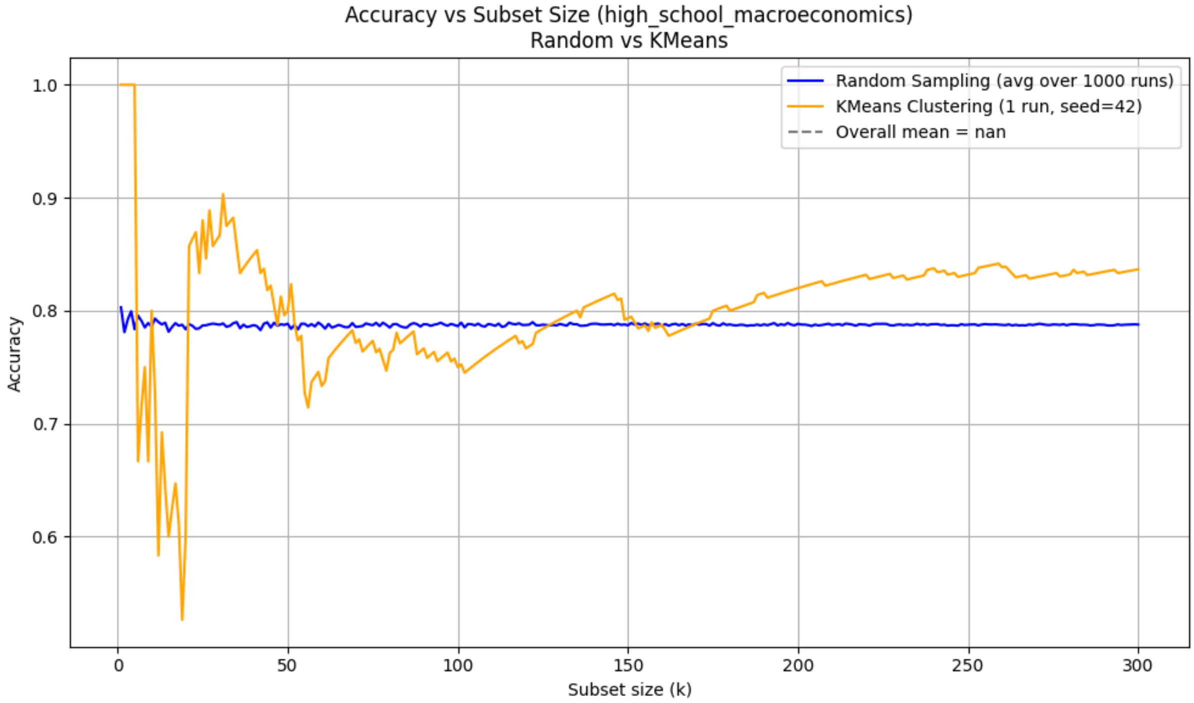
To compare the two strategies, I plotted the average number of correct responses by subset size, both for random selection, where the response set is scrolled and randomly selected 1000 times, and for representative selection via KMeans for all subsets, in a single run for size k.

The results show that random selection is able to converge to the global mean with sufficient stability, but with higher fluctuations, especially under low-k conditions.

KMeans provides a smoother curve, although in specific cases, such as professional_law in Figure 6B, convergence is nonlinear for some choices of k; this suggests limitations in the ability of embeddings to provide a semantically averaged representation. In general, KMeans is better at covering the semantic variety present in and of questions, especially in less diverse contexts.







CONCLUSIONS

- Random selection is effective, but it has high variability at low k values.
- KMeans offers a more stable strategy and semantically informed.
- Representativeness is more easily achieved in more uniform domains.

- Benchmark reduction strategies are practicable, and represent a valid compromise between accuracy and efficiency computational.

FUTURE DEVELOPMENTS

1. To evaluate the effect of subset selection onsorting between different models (in TREC style).

2. Explore other selection techniques, such as diversity-aware sampling of methods learnable.