Large Language Models Enable Few-Shot Clustering

Moncef Bouhabel , Salah Eddine Mohamed OUIKENE , Mouloud MAKHLOUF, Yagoub BAKIR

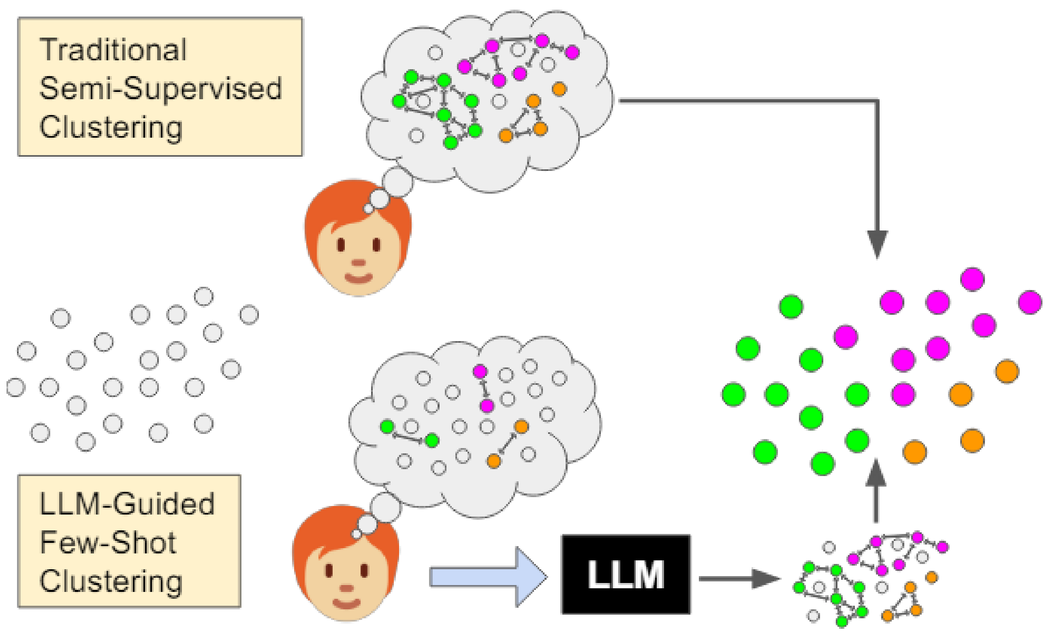
1. Overview

Few-Shot Clustering introduces a novel approach to text clustering by incorporating Large Language Models (LLMs). The Code and paper is publicly available on **GitHub.**

Github repo link :
https://github.com/Salaheddine-ouikene/LLM-Clustering

2. Abstarct

Unlike traditional unsupervised clustering, semi-supervised clustering allows users to provide meaningful structure to the data, which helps the clustering algorithm to match the user's intent. Existing approaches to semisupervised clustering require a significant amount of feedback from an expert to improve the clusters. In this paper, we ask whether a large language model can amplify an expert's guidance to enable query-efficient, fewshot semi-supervised text clustering. We show that LLMs are surprisingly effective at improving clustering.
We explore three stages where LLMs can be incorporated into clustering: **before clustering** (improving input features), **during clustering** (by providing constraints to the clusterer), and **after clustering** (using LLMs post-correction). We find incorporating LLMs in the first two stages can routinely provide significant improvements in cluster quality, and that LLMs enable a user to make trade-offs between cost and accuracy to produce desired clusters. We release our code and LLM prompts for the public to use.



3. Model Selection

We use the following models:

1. (OpenAI) `gpt-3.5-turbo-0613`
2. (OpenAI) `gpt-4-0613`
3. (MetaAI) `llama-2-7b-chat-hf`
4. (MistralAI) `Mistral-7b`

4.Outcomes of Utilizing key phrase expansion technique with GPT-3.5

Seed		42		5		25		100	
Metric		NMI	acc	NMI	acc	NMI	acc	NMI	acc
250 Articles	Simple	0.369	0.496	0.349	0.65	0.393	0.544	0.481	0.712
	GPT-3.5	0.881	0.956	0.656	0.704	0.668	0.776	0.886	0.96
500 Articles	Simple	0.659	0.806	0.639	0.8	0.35	0.558	0.35	0.558
	GPT-3.5	0.697	0.684	0.554	0.67	0.737	0.866	0.737	0.866
1000 Articles	Simple	0.71	0.57	0.61	0.605	0.63	0.658	0.59	0.70
	GPT-3.5	0.789	0.919	0.804	0.927	0.715	0.844	0.70	0.67
2250 Articles	Simple	0.815	0.926	0.637	0.733	0.668	0.704	0.697	0.805
	GPT-3.5	0.769	0.901	0.824	0.938	0.625	0.715	0.772	0.902

5.Outcomes of Utilizing Open-Source Language Models (Llama2 , Mistral)

Seed		42	
Metric		NMI	acc
50 Articles	Simple	0.436	0.6
	Llama2	0.583	0.7
100 Articles	Simple	0.609	0.77
	Llama2	0.818	0.86

Seed		42	
Metric		NMI	acc
50 Articles	Simple	0.436	0.6
	Mistral	0.605	0.64

6. Pros and Cons

Advantages :

1. Higher Accuracy and NMI
2. Ease of integration
3. Reduction of human supervision costs

Challenges :

1. Computational Overhead
2. Cost of LLMs

7. Conclusions

LLM clustering presents a promising approach for handling complex data analysis tasks, offering notable benefits such as higher accuracy and normalized mutual information (NMI), ease of integration into existing systems, and a reduction in costs associated with human supervision. These attributes make it particularly valuable in environments where precision and efficiency are paramount. However, the approach is not without its drawbacks.
The main challenges include the computational overhead required to manage these sophisticated models and the inherent costs associated with procuring and maintaining LLMs. Therefore, while LLM clustering is an effective strategy for enhancing data processing capabilities, it necessitates careful consideration of its computational and financial implications. Organizations should weigh these factors to determine the practicality and sustainability of implementing LLM clustering solutions in their specific contexts.

8. References

- Wagstaff, K., Cardie, C., Rogers, S., Schroedl, S. (2001, June).Constrained K-Means Clustering with Background Knowledge.In ICML (Vol. 1, pp. 577–584).
- Basu, S., Banerjee, A., Mooney, R. J. (2004, April).Active Semi-Supervision for Pairwise Constrained Clustering. In Proceedings of the 2004 SIAM international conference on data mining (pp. 333–344). Society for Industrial and Applied Mathematics.