**Introduction**

Large Language Models (LLMs) have emerged as potent tools for data categorization, providing a cost-effective and efficient alternative to traditional manual categorization by human coders, which often incurs higher costs, demands more time, and requires extensive coordination to maintain accuracy and consistency. Despite the promise, the application of LLMs in this realm is still in its infancy, necessitating further research to ascertain effective practices. This study endeavors to evaluate the efficacy of LLMs in data categorization tasks, identify methods to enhance the data categorization process for those preferring human expertise, and explore various prompting techniques to bridge the operational gap between machine-driven and human-centric data categorization approaches. Through this inquiry, we aim to facilitate a more accessible, efficient, and cost-effective data categorization framework.

**Methods**

We employed four human coders to manually categorize data. The correlation among human coders' outputs was computed to establish a baseline for "acceptable" variation. Subsequently, a series of tests were conducted employing different prompting techniques to evaluate the LLM's ability to approximate the acceptable variation. Through comparative analysis, we aim to ascertain the efficacy of LLMs in assisting or potentially substituting human effort in data categorization tasks. Additionally, we devised a technique that utilizes the LLM to augment, but not replace, the human coder by employing it as a tool for reaching a consensus on the "correct" categorization.

Utilizing the UC Berkeley Social Networks study, we analyzed open-ended text responses to the question, "Why did you move?" Subsequently, six categories of action were defined and the human coders were tasked with categorizing the responses accordingly. Parallelly, GPT-3.5 was employed through API calls to categorize the responses, utilizing various prompting techniques to ascertain the most fitting categories for each response. In a separate exercise, we utilized the LLM as a third coder alongside the human coders to achieve a "majority vote" on the best categorization for each response.

**Results**

Various prompting techniques were employed, with the most effective being a two-step process where the model summarized the response to extract major reasons for the move before categorizing, yielding a 0.77 correlation to a "gold standard" human coding. Superior results were observed with the "Chain of Thought" method, where the summary from one instance was fed to a separate instance for categorization, achieving a 0.79 correlation. Additionally, employing the LLM as a third coder to resolve disagreements between human coders reduced the manual work of verifying categorization by up to 40%, showcasing a substantial enhancement in the categorization workflow efficiency.

**Discussion**

While the attained correlation of 0.79 with the "Chain of Thought" method indicates promise, there's further exploration required to align closer with the 0.87 correlation achieved by human coders against a "gold standard." Interestingly, the model outperformed human coders in certain areas, unveiling potential for this technique. Utilizing LLMs as auxiliary to human coders significantly reduced the data categorization workload by up to 40%, highlighting the efficiency and cost-effectiveness that can be introduced to the data categorization process, while still aiming for a high degree of accuracy and correlation with established human coding standards.

Table 1: Comparing Correlations Between Human Coders and Various Prompting Techniques

A table with numbers and lines

Description automatically generated

Figure 1: Diagram on How Employing LLM’s as a Third Coder Can Improve a Workflow

A diagram of a diagram

Description automatically generated