An Empirical Investigation into the Utility of Large Language Models in Assisting Human Data Categorization

Chris Soria* Claude Fischer †

Draft Version: February 6, 2024

Abstract

This research evaluates Large Language Models (LLMs) in multi-class classification tasks against human categorization as a benchmark and provides Python tools and templates for survey researchers to classify open-ended responses. We release a comprehensive dataset of open-ended survey responses (with personal information redacted), sociologist-validated categorizations, and all human and LLM categorizations, serving as benchmarks for methodological enhancement and further research.

*Department of Demography, University of California, Berkeley ChrisSoria@berkeley.edu.
†Department of Sociology, University of California, Berkeley Fischer1@berkeley.edu.

1 Introduction

PLACEHOLDER

2 LLM Data Categorization

2.1 Background

2.2 Training

Prior research has explored various prompting strategies to enhance the accuracy and output quality of Large Language Models (LLMs) for reasoning tasks. Fine-tuning (FT) the LLM, updating its weights for a particular task, is effective (Ziegler et al., 2020), especially for tasks like surveys repeated over time. While fine-tuning improves task-specific results, it demands large, task-specific datasets and can reduce the model's general applicability (McCoy, Pavlick and Linzen, 2019). Our current focus is on assessing the model's performance across a hypothetical broad range of open-ended survey response categorization tasks without tailoring it to any specific one. We aim to explore fine-tuning in future work.

In contrast, so called Zero-Shot (ZS), One-Shot (OS) and Few-Shot techniques do not alter the model in anyway and instead focus on crafting the ideal string object to feed the model to achieve a desired output. While ZS provides a set of instructions to the model, it does not include any demonstration of how the outcome is desired. OS and ZS techniques, on the other hand, provide the model with demonstrations of the task without altering the model's weights. For these approaches, a small amount of task-specific instruction is necessary, where we provide K examples of the "proper" categorization of a survey responses. This approach frames the models as unsupervised multitask learners (Radford et al., 2019).

Below are the three approaches on demonstrating a correct response we will test:

1. Few-Shot (FS): $K=N$
2. One-Shot (OS): $K=1$
3. Zero-Shot (ZS): $K=0$

2.3 Instructions

Significant research has focused on optimizing instructions to enhance model performance. Wang Ling's team proposes simplifying algebraic word problems into smaller steps for better accuracy (Ling et al., 2017), while Jason Wei and colleagues demonstrate that the "Chain-of-Thought" (CoT) framework notably enhances LLM performance on reasoning tasks (Wei et al., 2023). In evaluating this framework, researchers discovered that while it raises the likelihood of obtaining a correct answer, it does not guarantee consistency. Consequently, researchers suggest further dividing the CoT process into two main phases: one for generating the answer and another for organizing it into a specified structured format. This augmentation of CoT, known as Structured CoT (SCoT), significantly improves the correctness of the format of a desired output (Li et al., 2023).

Notwithstanding, LLMs frequently generate responses detached from reality, known as "hallucinations." Researchers suggest enhancing model accuracy by incorporating a self-critique phase before finalizing the response, a method that can increase correctness by up to 20 percent for reasoning tasks (Press et al., 2023; Madaan et al., 2023). Building on this, Shehzaad Dhuliawala et al. discovered that employing an independent prompt for the model to verify its output, a technique named Chain-of-Verification (CoVe), further improves accuracy for complex tasks that require long-form responses and lists as the output Dhuliawala et al. (2023). This CoV framework, especially when the task involves listing categories within survey responses, shows promise in reducing hallucinations and ensuring the comprehensiveness and accuracy of the categories generated.

Below are the three instructional approaches we will test:

1. Chain-of-Thought (CoT): *Requiring the task to be done in steps*
2. Structured Chain-of-Thought (SCoT): *Asking output to be formatted in s separate step*
3. Chain-of-Verification (CoVe): *Requiring the model to verify its work*

2.4 Structuring the String

Despite limited research, effective prompting is recognized for its structured, clear, and concise presentation of instructions and labeled objects. AI practitioners generally recommend enclosing user input intended for model interaction within delimiters to clearly separate it from instructions. It's also advised to offer an alternative output option, in this case, if the data does not match any provided categories. Lastly, some have suggested that providing the model with an "inner monologue," or a persona, might improve performance. This study investigates how these adjustments to a prompt enhance output quality for survey categorization tasks, aiming to create a tailored template for multi-label classification tasks.

Below are the two string structures we will test:

1. Clearly labeled instructions and delimited survey input and user-provided categories
2. Concise instructions without clear delimitation

3 Human Data Categorization

4 Methods

4.1 Data

We selected three open-ended questions from the UCNets survey, focusing on US internal mobility (Questions 1 and 2) and emergency family support (Question 3). Question 3 consists of two parts: a rating of confidence in family support on a 1-4 scale, followed by an open-ended explanation of their rating.

1. "Why did you move?"
2. "After this last move, what steps, if any, did you take in order to make new friends?"
3. "If you had a serious problem, like a life-threatening illness or possibly losing your home, do you feel that you have some relatives that you can rely on to help (scale)?

What is it about your relatives or your connection to them that makes you say that (open-ended)?"

In order to stress test the model's capabilities, we progressively increased the difficulty of the categorization task.

1. Five simple categories that require no logical reasoning to sort.
2. Nine categories, two of which require nuanced differentiation and one that requires simple conditional logic
3. Twenty categories, two many of which require nuanced differentiation and two that require slightly more difficult conditional logic

4.2 Controls

4.3 Evaluation

To evaluate the models against human performance, we'll use the Jaccard Similarity index, also known as the Jaccard coefficient, to compare categorized data columns. We measure the similarity between each column categorized by humans and the LLM using the Jaccard coefficient (1), then average these coefficients to generate a single, comparable score (eq. 2.). A score closer to 1 indicates greater similarity between human and LLM categorizations, while a score of 0 signifies the opposite.

$$J_{\text{Qk},C_i} = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

$$S_{\text{Qk}}(A, B) = \frac{\sum_{i=1}^{n} J_{\text{Qk},C_i}(A, B)}{n} \tag{2}$$

In line with our binary multi-class classification, where each variable is a set of indices where the variable is 1 (or true) if the category is present in the response, then:

- $|A \cap B|$ is the count of indices where both vectors are 1.
- $|A \cup B|$ is the count of indices where either vector is 1 (or both).

In this context:

- Qk corresponds to the extracted coefficient for question k.
- A corresponds to the set of attributes, choices, or decisions made by a human.
- B corresponds to the set of attributes, choices, or decisions predicted or generated by an LLM.

To compare the categorization of a single response, we assign a unique categorization ID to each response's categorization (3). We compare these IDs between human and LLM categorizations, generating a 1 for a match and a 0 for any difference (4). For inspecting all differences between human and LLM categorization, we subset dataframes where concordance flag = 0.

$$id = [a_1, a_2, \ldots, a_n] \tag{3}$$

$$\begin{cases} 1, & \text{if } id_h = id_{llm} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

5 Results

5.1 Question 1: Simplest Task

	Average_Jaccard
Claude	0.806572
Rebeca	0.769453
bad	0.675987
good	0.718006
cot	0.749850
cove	0.687312
oneshot	0.736034
fewshot	0.734055

Figure 1: Categorization of "Why did you move?"

References

Dhuliawala, Shehzaad, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz and Jason Weston. 2023. "Chain-of-Verification Reduces Hallucination in Large Language Models.". arXiv:2309.11495 [cs].
URL: *http://arxiv.org/abs/2309.11495*

Li, Jia, Ge Li, Yongmin Li and Zhi Jin. 2023. "Structured Chain-of-Thought Prompting for Code Generation.". Publisher: arXiv Version Number: 3.
URL: *https://arxiv.org/abs/2305.06599*

Ling, Wang, Dani Yogatama, Chris Dyer and Phil Blunsom. 2017. Program Induction by Rationale Generation: Learning to Solve and Explain Algebraic Word Problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ed. Regina Barzilay and Min-Yen Kan. Vancouver, Canada: Association for Computational Linguistics pp. 158–167.
URL: *https://aclanthology.org/P17-1015*

Madaan, Aman, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh and Peter Clark. 2023. "Self-Refine: Iterative Refinement with Self-Feedback.". arXiv:2303.17651 [cs].
URL: *http://arxiv.org/abs/2303.17651*

McCoy, R. Thomas, Ellie Pavlick and Tal Linzen. 2019. "Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference.". arXiv:1902.01007 [cs].
URL: *http://arxiv.org/abs/1902.01007*

Press, Ofir, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith and Mike Lewis. 2023. "Measuring and Narrowing the Compositionality Gap in Language Models.". arXiv:2210.03350 [cs].
URL: *http://arxiv.org/abs/2210.03350*

Radford, Alec, Jeff Wu, Rewon Child, D. Luan, Dario Amodei and I. Sutskever. 2019. Language Models are Unsupervised Multitask Learners.
URL: *https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe*

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le and Denny Zhou. 2023. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.". arXiv:2201.11903 [cs].
URL: *http://arxiv.org/abs/2201.11903*

Ziegler, Daniel M., Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano and Geoffrey Irving. 2020. "Fine-Tuning Language Models from Human Preferences.". arXiv:1909.08593 [cs, stat].
URL: *http://arxiv.org/abs/1909.08593*