

Prompting as Panacea? A Case Study of In-Context Learning Performance for Qualitative Coding of Classroom Dialog

Ananya Ganesh, Chelsea Chandler, Sidney D'Mello, Martha Palmer, Katharina von der Wense
University of Colorado Boulder
Boulder, CO, 80309
Ananya.Ganesh@colorado.edu

ABSTRACT

One of the areas where Large Language Models (LLMs) show promise is for automated qualitative coding, typically framed as a text classification task in natural language processing (NLP). Their demonstrated ability to leverage *in-context learning* to operate well even in data-scarce settings poses the question of whether collecting and annotating large-scale data for training qualitative coding models is still beneficial. In this paper, we empirically investigate the performance of LLMs designed for use in prompting-based in-context learning settings, and draw a comparison to models trained with task-specific annotated data, specifically for tasks involving qualitative coding of classroom dialog. Compared to other domains where NLP benchmarking studies are typically situated, classroom dialog is much more natural and therefore variable and complex. Moreover, tasks in this domain are nuanced, theoretically grounded and require a deep understanding of the conversational context. We provide a comprehensive evaluation across five datasets, including tasks such as talk move prediction and collaborative problem solving skill identification. Our findings show that task-specific fine-tuning strongly outperforms in-context learning, underscoring the ongoing need for high-quality annotated training datasets.

Keywords

large language models, natural language processing, qualitative coding, classroom dialog

1. INTRODUCTION

In recent years, the proliferation of Natural Language Processing (NLP), Artificial Intelligence (AI), and Large Language Models (LLMs) has revolutionized various facets of educational technology, from the development of conversational tutors to automated grading systems, significantly impacting student learning experiences and instructional methodologies [4, 33, 20]. LLMs, particularly those that can be used

off-the-shelf with minimal to no training such as the GPT [22] and LLaMa [31] models are slowly being adopted as the de facto models for text generation tasks such as generating hints [25], providing feedback to students [18], or assisting teachers [33]. More recently, LLMs have gained attention as an alternative to "traditional" NLP models for automated qualitative coding tasks [34]. However, their feasibility for coding tasks, especially for challenging constructs commonly found in the educational domain, has yet to be systematically investigated.

Automated qualitative coding problems are typically formalized as classification tasks in NLP. That is, given text (such as student conversation or writing), the task is to predict the most likely label from a pre-defined set of classes (such as if an utterance is a question). Classification models have until recently been developed through the pre-training/fine-tuning paradigm: pre-trained language models such as BERT [7], trained on large web-based corpora are *fine-tuned*, or undergo task-specific training on human-annotated data. While these models leverage the rich representations of language learned in the pre-training stage to understand meaning, they require several examples of each class to learn to distinguish between them, ultimately necessitating datasets with thousands of examples. Acquiring such large datasets incurs substantial costs, as trained experts must invest time and effort into the annotation process. On the other hand, the ability of LLMs to learn to solve complex tasks using very few or no examples ostensibly provides a way to bypass the expensive data collection process. LLMs achieve this through the mechanism of in-context learning, a process that involves interacting with LLMs through natural language instructions, without any training.

This ability has been demonstrated on NLP benchmarks that challenge the language understanding abilities of models [28]. Some of these benchmarks even test for domain knowledge and reasoning on topics such as physics or medicine [21, 12]. Despite such rigorous testing, applying these models to qualitative coding, particularly in education, should still be treated with caution for the following reasons: 1) benchmark tasks tend to be well-defined and sometimes shallow, in contrast to the nuanced, theoretically-motivated frameworks that drive qualitative coding, 2) benchmark tasks may be highly similar to tasks that some LLMs are explicitly trained on (e.g., question answering); and 3) benchmark datasets, particularly if openly available online, may have

A. Ganesh, C. Chandler, S. D'Mello, M. Palmer, and K. Kann. Prompting as panacea? a case study of in-context learning performance for qualitative coding of classroom dialog. In B. Paaßen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 835-843, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

© 2024 Copyright is held by the author(s). This work is distributed under the Creative Commons Attribution NonCommercial NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license.
<https://doi.org/10.5281/zenodo.12729966>

