OBJECTIVE: Recent advancements in the few-shot prompting field have shown considerable progress in scenarios with limited labeled data. However, the effectiveness of few-shot prompting methods significantly depends on the choice of in-context demonstrations. Various techniques, such as Auto-CoT, Active-CoT, and Retrieval-CoT have demonstrated their effectiveness, emphasizing diversity (Auto-CoT), uncertainty (Active-CoT), and similarity with the test example (Retrieval-CoT). Notably, while some methods, like Random-CoT and Auto-CoT, utilize labels generated by a large language model, Active-CoT relies on human-generated labels. This thesis aims to evaluate and compare several prompting methods using closed-source and open-source models in arithmetic reasoning tasks under both labeling scenarios: labels generated by GPT-3.5-turbo and human-generated labels. It also aims to develop new prompting methods by combining diversity, uncertainty, or similarity with the test example.

RESULTS: In experiments on GSM8K using GPT-3.5-turbo, Diverse-CoT, Retrieval-CoT, and Active-CoT outperform Random-CoT by margins of 1.1%, 1.7%, and 3.5%, respectively. Similarly, in evaluations on AQUA, Diverse-CoT and Active-CoT surpass the Random approach by margins of 3.9% and 2.7%, respectively. These results are consistent when GPT-3.5-turbo labels the questions. However, when human-generated labels are used, these methods achieve performance comparable to or even lower than Random baseline.

Additionally, I propose new methods: Diverse-Active-KMeansPlusPlus-CoT, which combines diversity and uncertainty, and Diverse-Active-KMeansPlusPlus-Retrieval-CoT, integrating similarity with the test question in the Diverse-Active-KMeansPlusPlus-CoT method. These new methods outperform the Random baseline by 1.9% and 2.5%, respectively, on GSM8K when using GPT-3.5-turbo-generated labels. On AQUA, Diverse-Active-KMeansPlusPlus-Retrieval-CoT surpasses Random by 3.3%.

Moreover, Falcon-40B-Instruct achieves 37.3% accuracy on GSM8K and 18.5% on AQUA, while Falcon-7B-Instruct achieves 5.4% on GSM8K and 11.4% on AQUA. This emphasizes that larger models perform better in a few-shot setting but exhibit inferior performance compared to GPT-3.5-turbo, with margins exceeding 30%.

The code is available at https://github.com/Lori10/Master-Thesis-Few-Shot-CoT-Prompting-LLM.

ZIEL: Aktuelle Entwicklungen im Bereich des Few-Shot Prompting haben deutliche Fortschritte in Szenarien mit begrenzten gelabelten Daten gezeigt. Die Effektivität von Few-Shot Prompting-Methoden hängt jedoch wesentlich von der Wahl der kontextbezogenen Beispiele ab. Verschiedene Techniken wie Auto-CoT, Active-CoT und Retrieval-CoT haben ihre Wirksamkeit unter Beweis gestellt, indem sie die Diversität (Auto-CoT), die Unsicherheit (Active-CoT) und die Ähnlichkeit mit dem Testbeispiel (Retrieval-CoT) hervorheben. Während einige Methoden, wie z.B. Random-CoT und Auto-CoT, von einem großen Sprachmodell generierte Labels verwenden, basiert Active-CoT auf von Menschen erstellten Labels. Ziel dieser Arbeit ist es, verschiedene Prompting-Methoden zu evaluieren und zu vergleichen, die Closed-Source- und Open-Source-Modelle in arithmetischen Denkaufgaben verwenden, und zwar unter beiden Labeling-Szenarien: Labels, die von GPT-3.5-turbo generiert werden, und von Menschen erzeugte Labels. Außerdem sollen neue Prompting-Methoden entwickelt werden, die Diversität, Unsicherheit oder Ähnlichkeit mit dem Testbeispiel kombinieren.

ERGEBNISSE: In Experimenten auf GSM8K mit GPT-3.5-Turbo übertreffen Diverse-CoT, Retrieval-CoT und Active-CoT den Random-CoT-Ansatz mit einer Differenz von 1,1%, 1,7% bzw. 3,5%. In ähnlicher Weise übertreffen Diverse-CoT und Active-CoT bei den AQUA-Evaluierungen den Random-Ansatz mit einer Differenz von 3,9% bzw. 2,7%. Diese Ergebnisse sind konsistent, wenn GPT-3.5-turbo die Fragen labelt. Werden jedoch von Menschen erstellte Labels verwendet, erreichen diese Methoden eine Performance, die mit der des Random-Ansatzes vergleichbar oder sogar niedriger ist.

Zusätzlich schlage ich neue Methoden vor: Diverse-Active-KMeansPlusPlus-CoT, das Diversität und Unsicherheit kombiniert, und Diverse-Active-KMeansPlusPlus-Retrieval-CoT, das die Ähnlichkeit mit der Testfrage in die Diverse-Active-KMeansPlusPlus-CoT-Methode integriert. Diese neuen Methoden übertreffen die Random-Baseline auf GSM8K um 1,9% bzw. 2,5%, wenn sie mit GPT-3.5-Turbo-generierten Labels eingesetzt werden. Auf AQUA übertrifft Diverse-Active-KMeansPlusPlus-Retrieval-CoT die Random-Methode um 3,3%.

Außerdem erreicht Falcon-40B-Instruct eine Genauigkeit von 37,3% bei GSM8K und 18,5% bei AQUA, während Falcon-7B-Instruct 5,4% bei GSM8K und 11,4% bei AQUA erreicht. Dies betont, dass größere Modelle in einer Few-Shot Situation besser abschneiden, aber im Vergleich zu GPT-3.5-turbo eine schlechtere Performance zeigen, mit Abweichungen von über 30%.

Der Code ist verfügbar unter https://github.com/Lori10/Master-Thesis-Few-Shot-CoT-Prompting-LLM.