Opdracht 2 - LLM Zephyr 3B + prompt engineering + experiment opzet

Johan Weiland

April 10, 2024

Abstract

Your abstract.

1 Interesante links

- stability.ai - https://stability.ai/news/stablelm-zephyr-3b-stability-llm
- Voorbeeld notebook - https://www.kaggle.com/code/prbo123/stablelm-zephyr-3b
- Voorbeeld notebook2 - https://colab.research.google.com/drive/..
- What actually happens when we run some text in a pipeline created with the Hugging Face Transformers library - https://www.youtube.com/watch?v=1pedAIvTWXk
- interactie voorbeeld tussen een llm en Pettingzoo - https://pettingzoo.farama.org/tutorials/langchain/langchain/
- Blackjack op gym - https://github.com/langchain-ai/langchain/blob/master/cookbook/gymnasium$_{a}gent_{s}imulation$

2 literatuuronderzoek

2.1 Stable LM Zephyr 3B

Stable LM Zephyr 3B is Large Language Model (LLM) with 3 billion parameters. This is 60% smaller than the Zephyr 7B models, allowing accurate, and responsive output on a variety of devices without requiring high-end hardware [Isl23].

2.2 Promp Engineering

With the recent popularity of llm like ChatGpt, a lot of research is done in prompt engineering [SSS+24][LYF+23]. For this experiment various prompt engineering techniques where used: Zero-shot, one-shot, few-shot[BMR+20] and Chain-of-Thought [WWS+22].

2.2.1 Zero-shot prompting

In zero-shot learning, a model is tasked with making predictions or generating outputs for tasks it has never seen during training. The model relies on its understanding of the underlying concepts and relationships learned from the training data to perform these new tasks. Essentially, it's like asking someone a question they've never heard before and expecting them to give a reasonable answer based on their general knowledge.

```
Example:
In which direction the cart of a cart pole should be moved if the pole direction is 1.5 degrees
off to the left?
```

2.2.2 One-shot prompting

One-shot learning involves training a model given only a single example or piece of data per class or task.

```
Example:
When the pole tilts to the right the cart should be moved to the right.
In which direction the cart should be moved if the pole direction is 1.5 degrees off to the left?
```

2.2.3 Few-shot prompting

In few-shot learning, the model is provided with a small number of examples (typically fewer than what's required for traditional supervised learning) for each class or task during training. The model learns to generalize from this limited data and perform well on new instances of the same classes or tasks [BMR+20].

```
Example:
When the pole tilts to the right the cart should be moved to the right.
When the pole tilts to the left the cart should be moved to the left.
In which direction the cart should be moved if the pole direction is 1.5 degrees off to the left?
```

2.2.4 Chain-of-Thought prompting

In addition, we use Chain-of-Thought (CoT) Prompting. LLMs often stumble in the face of complex reasoning, limiting their potential. Aiming to bridge this gap, [WWS+22] introduced Chain-of-Thought (CoT) prompting as a technique to prompt LLMs in a way that facilitates coherent and step-by-step reasoning processes.

```
Example:
Q: In which direction the cart of a cart pole should be moved if
the pole direction is 5 degrees off to the left
A: Left
Q: In which direction the cart of a cart pole should be moved if
the pole direction is 5 degrees off to the right
A: Right
Q: In which direction the cart of a cart pole should be moved if
the pole direction is 1.5 degrees off to the left?
```

3 Experiment setup

In this experiment 2 single agent RL environments will be used: PoleCart, Blackjack and one multi agent RL enviroment named TicTacToe. The goal of this assigment is to use two LLM models (Phi en Zephyr 3B) to play the reinforcement learning envirfonments. These will be compared to a trained DQn and ActorCritic model. For baseline purposes a Random agent is also used.

The first part of the experiment different types of prompts are experimented with in the single agent enviroments. Next the traditonal RL models DQN and actor critc are trained. Every model runs for a 100 episodes in every single agent environment so a comparison of the rewards can be made.

In the second part of the experiment the different models will compete in a multi agent Tic-Tac-Toe environment.

3.1 Gym environments cartPole

1. 100 episodes
2. 4 Prompt methods: Zero-shot, One-shot, Few-shot, Chain-of-Though
3. Parameters:
   (a) 4x temperature 0, 0.3, 0.6, 0.9.

(b) more?

A grid search will result in 4*4=16 experiments of 100 runs per LLM. The best parameters from the grid search are used for comparison with the other models (Phi, DQN, Actor-Critic, and a Random Action baseline).

3.2 Gym environments BlackJack

Zelfde setup als cartPole, prompts volgen nog.

3.3 Gym environments TicTacToe

A 2 agent environment. 5 agents will be competing Phi, DQN and Actor-Critic this means 20 agent combinations. Each combination will play 100 games.

3.4 Openstaande vragen

Gebruiken we bij DQN ook binaire waarden bij cardepole. dus of 1 =rechts en 0 = links, of continious -1 t/m 1? Wel zo

4 Data analysis setup

statistische vergelijking van de resultaten om te kijken of er significant verschil zit in de agents t-test?

References

[BMR+20] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[Isl23] Anel Islamovic. Introducing stable lm zephyr 3b: A new addition to stable lm, bringing powerful llm assistants to edge devices, Dec 2023.

[LYF+23] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

[SSS+24] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*, 2024.

[WWS+22] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.