wonderlic.

SIOP ML Challenge 2024 Results Presentation

Wonderlic ML

The Team

Wonderlic: Unlock People Potential With Predictive Assessments

Guglielmo Menchetti
Senior Machine Learning Engineer

Lea Cleary
Machine Learning Engineer

Annie Brinza
Manager of Data Science and Engineering

Our Approach

What we tried

- ✓ Prompt Engineering and ICL
- ✓ Prompt Tuning
- ✓ SetFit

Why we tried them

- ✓ Prompt Engineering and ICL
  - Easy to implement and to use as baseline
  - ✓ Often produces great results
- ✓ Prompt Tuning
  - ✓ Fine-Tuning on additional datasets
  - ✓ Limited amount of time and resources
- ✓ SetFit
  - ✓ Small amount of data
  - ✓ Quick training on (relatively) small LLMs

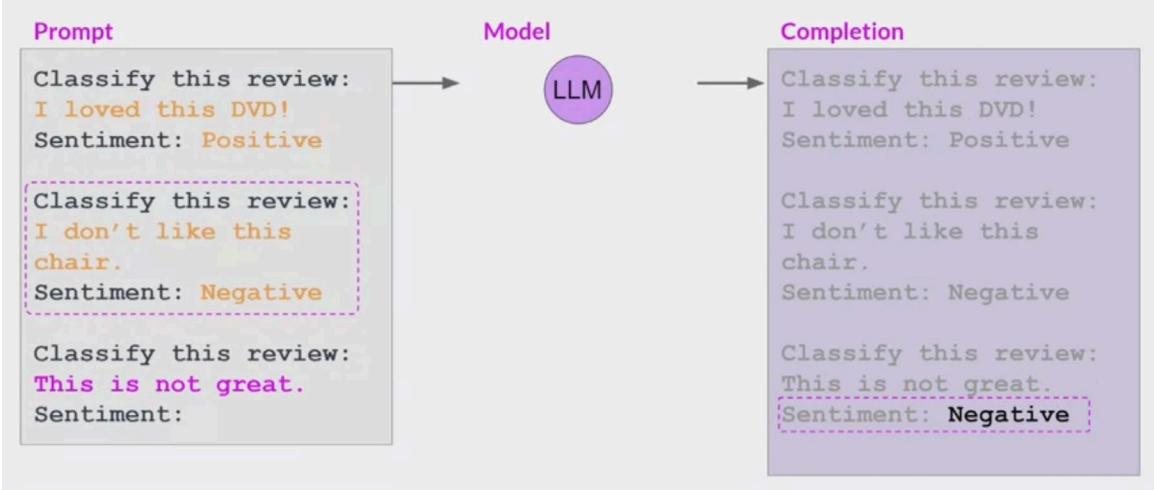
Prompt Engineering and ICL- Overview

Prompt Engineering

- ✓ Adapt LLMs knowledge and capabilities for a particular task
- ✓ Design **input prompts** to elicit the desired response from the LLM
- ✓ **Iterative refinement**
  - ✓ Adjust the prompts based on experimentation
  - ✓ **Example** : trying different phrasing or structure

In Context Learning

- ✓ **Include examples** of the task in the input prompt
- ✓ Additional step: **dataset preparation**



Source: Coursera - Generative AI with Large Language Models

Prompt Engineering and ICL – ML Competition

Methods we tried

- ✓ Zero-shot and Few-Shot
- ✓ Few-shot limited by the model's max input tokens
- ✓ Clarity, Fairness and Empathy
  - ✓ Used the training data as context examples
  - ✓ Each model required a different prompt
- ✓ Interview
  - ✓ Interview responses from the same person as context
  - ✓ Personality data

Models we tried

- ✓ GPT4 Turbo
- ✓ Flan T5
- ✓ Mistral 7B

What we learned

Prompt Engineering and ICL – Interview

Adding personality in interview

What we learned

Prompt Engineering and ICL– GPT Prompts

System Message all but Interview : "You are an Industrial -Organizational Psychologist."
System Message Interview : "You are a job candidate and you are responding to behavioral questions during an interview."

Empathy

User: "Job candidates were asked to provide empathetic responses to a difficult workplace situation. Classify whether empathy was demonstrated or not. Only respond with 0 (no empathy) or 1 (empathy).
[TEXT 1]"

Assistant: " [LABEL 1]" …

User: "[TEXT N] "

Fairness

User: "Respondents compared two organizational policies and voted on which was fairest. Identify which policy received the majority vote as the fairer option. Answer only with 'first' or 'second', even in uncertainty. 'Both' as a response is not valid.

First=[FIRST TEXT 1] – Second=[SECOND TEXT 1]"

Assistant: "[LABEL 1]" …

User: " First=[FIRST TEXT N] – Second=[SECOND TEXT N]"

Clarity

User: "Respondents rated the clarity of personality test items using a 7 -point scale from 1 = extremely unclear to 7 = extremely clear. Your task is to rate the clarity of the following item from an industrial -organizational psychology perspective. Respond with a rational number between 1 and 7.
[ITEM TEXT 1]"

Assistant: " [LABEL 1]" …

User: "[ITEM TEXT N] "

Interview

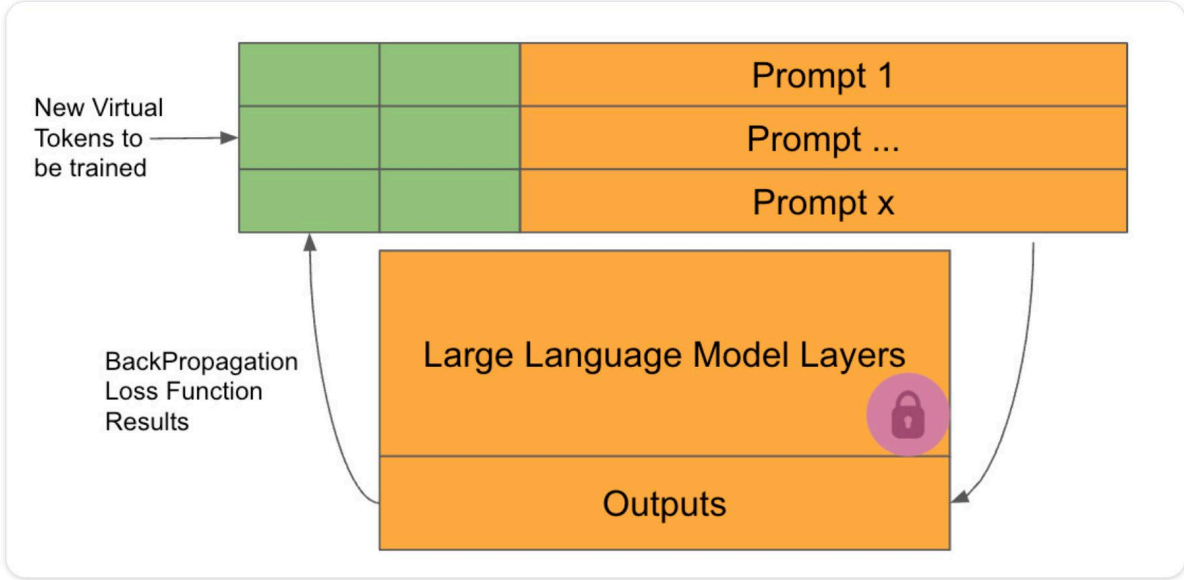
User: " You scored moderate on Openness which means that …
You will be provided with questions and answers you already answered in between ### markers. You will need to answer the last question in a consistent style, using less than five sentences.

Question: [QUESTION 1] – Response: [RESPONSE 1] ###...
Question: [QUESTION N] – Response:"

Prompt Tuning - Overview

What is it

- ✓ **Soft Prompting** method
  - ✓ Adapt the model to a **downstream task**
  - ✓ Add a small amount of **learnable parameters** to the input embeddings
  - ✓ Keep the pretrained **LLM parameter's frozen**
- ✓ **Sequence Classification**
  - ✓ Cast the problem as a **next token generation**
  - ✓ Prompts are **added to the input** as a series of tokens
  - ✓ Train the model for a **Causal LM task**



Source: Hugging Face - https:// huggingface.co /learn/cookbook/ prompt_tuning_peft

Pros

- ✓ Few parameters to train
- ✓ Fast adaptation to specialized tasks
- ✓ Efficient resource utilization

Cons

- ✓ Lack of interpretability of the tuned prompts

Prompt Tuning - ML Competition

What we tried

- ✓ Model trained on QA and Entailment **external datasets**
  - ✓ Zero-shot on **Empathy** and **Fairness**
  - ✓ Didn't work as expected
- ✓ Model trained on **SIOP data** as a QA and Entailment problems
  - ✓ **Entailment** : Empathy and Fairness (+ additional data), unified datasets and separate
  - ✓ **QA:** All tasks separate
- ✓ **Models**
  - ✓ BLOOMZ (560M, 1.1B, 7B)
  - ✓ GPT2 Medium
  - ✓ LLaMa2 7B

What we learned

Prompt Tuning - Data formats

Input data format

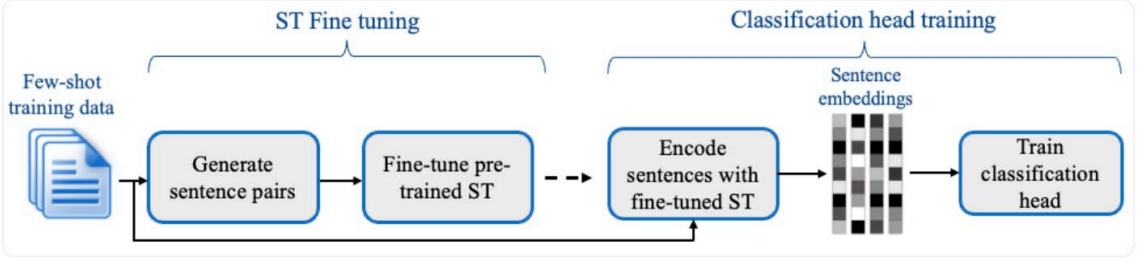
- ✓ Entailment - Hypothesis
  - ✓ Empathy
    ”The premise demonstrates an empathetic response to a difficult workplace situation.”
  - ✓ Fairness
    “The first option was voted by the majority of respondents as the fairer organizational policy.”
- ✓ Question Answering - Question
  - ✓ Empathy
    “Does the passage demonstrate an empathetic response to a difficult workplace situation?”
  - ✓ Fairness
    “Was the first option voted by the majority of respondents as the fairer organizational policy?”
  - ✓ Clarity
    “What is the average clarity rating on a scale of 1 -7 of the personality item?”
  - ✓ Interview
    “Job candidates responded to four common interview questions. You will be given the text of three question and response pairs as reference. Your task is to generate a likely text response for the last question based on the previous responses.”

SetFit: Efficient Fine-Tuning of Sentence Transformers

What is it

- ✓ Fine-Tuning of **Sentence Transformer** model
- ✓ **Few-shot learning**
  - ✓ Small amount of data
  - ✓ Leveraging prior knowledge
- ✓ **Contrastive Learning**
  - ✓ Learn differences between similar/dissimilar pairs
  - ✓ Create pairs of similar or dissimilar items

How it works



Source: Efficient Few-Shot Learning Without Prompts - https://arxiv.org/pdf/2209.11055.pdf

Pros

- ✓ Only few datapoints needed
- ✓ Save time and resources

Cons

- ✓ Results are comparable but not the same as training on a big dataset

SetFit – ML Competition

What we tried

- ✓ Only tried on **Empathy**
- ✓ **Created sentence pairs**
  - ✓ Pairs: 420
  - ✓ Validation: 8 samples
- ✓ **Hyperparameter tuning**
  - ✓ Number of epochs
  - ✓ Learning rate
- ✓ **Models**
  - ✓ MPNET
  - ✓ BART-Large

What we learned

Final Models

Empathy

- ✓ SetFit Fine-Tuning
- ✓ BART-Large
- ✓ Hyperparameter tuning
- ✓ **Score: 0.1225**

Fairness

- ✓ In-Context Learning
- ✓ GPT4 – Full-Shot
- ✓ Temperature: 0
- ✓ **Score: 0.19828**

Interview

- ✓ In-Context Learning
- ✓ Personality with buckets and interpretation
- ✓ GPT4
- ✓ Temperature: 1
- ✓ **Score: 0.11539**

Clarity

- ✓ Prompt-Tuning
- ✓ Ensemble of models
- ✓ BLOOMZ 1B + GPT2 Medium
- ✓ **Score: 0.19346**

Learnings/Takeaways

- ✓ Great chance to **explore possible uses of LLMs**
- ✓ **Prompt Engineering** and **ICL** are easy to setup and use as first models
- ✓ **GPT4** was the easiest one to setup
  - ✓ Most reliable outputs
  - ✓ Worth paying for the API access
- ✓ **Prompt -Tuning** allows to save on resources
  - ✓ Still requires time to train the models
  - ✓ Easy to setup using the **PEFT Library** (Huggingface)
- ✓ **SetFit** easy solution when few datapoints are available
  - ✓ Easy to setup using the **SetFit Library** (Huggingface)

Thank you for your attention

Appendix A

Prompt Tuning using external data

- ✓ Check if prompt -tuned models could be transferred to **Fairness** and **Empathy** tasks
- Prompt Tuned **LLaMa2-7B** with **100 tokens**
- Two datasets
  - **BoolQ – Question Answering**
    - Around 9k examples
    - Passage-Question pairs with “yes”/”no” labels

is confectionary sugar the same as powdered sugar	true	Powdered sugar, also called confectioners' sugar, icing sugar, and icing cake, is a finely ground sugar…
is elder scrolls online the same as skyrim	false	As with other games in The Elder Scrolls series, the game is set on the continent of Tamriel. The events of…

  - **RTE**
    - Around 2.5k examples
    - Premise-Hypothesis pairs with “entailment”/”not entailment” labels

No Weapons of Mass Destruction Found in Iraq Yet.	Weapons of Mass Destruction Found in Iraq.	1	0	not entailment
A place of sorrow, after Pope John Paul II died, became a place of…	Pope Benedict XVI is the new leader of the Roman Catholic Church.	0	1	entailment