Yogesh Haribhau Kulkarni• You

AI Coach | PhD in Geometric Modeling | Google Developer E...

2w • Edited • 🌐

•••

Wish to test different LLMs yourself? 😃

LangChain makes it possible. 👍 Be it an OpenAI's GPT (Generative Pretrained Transformer) or any other LLM (Large Language Model) available at Hugging Face, you can programmatically run them over same set of prompts and compare the results. ⚖️

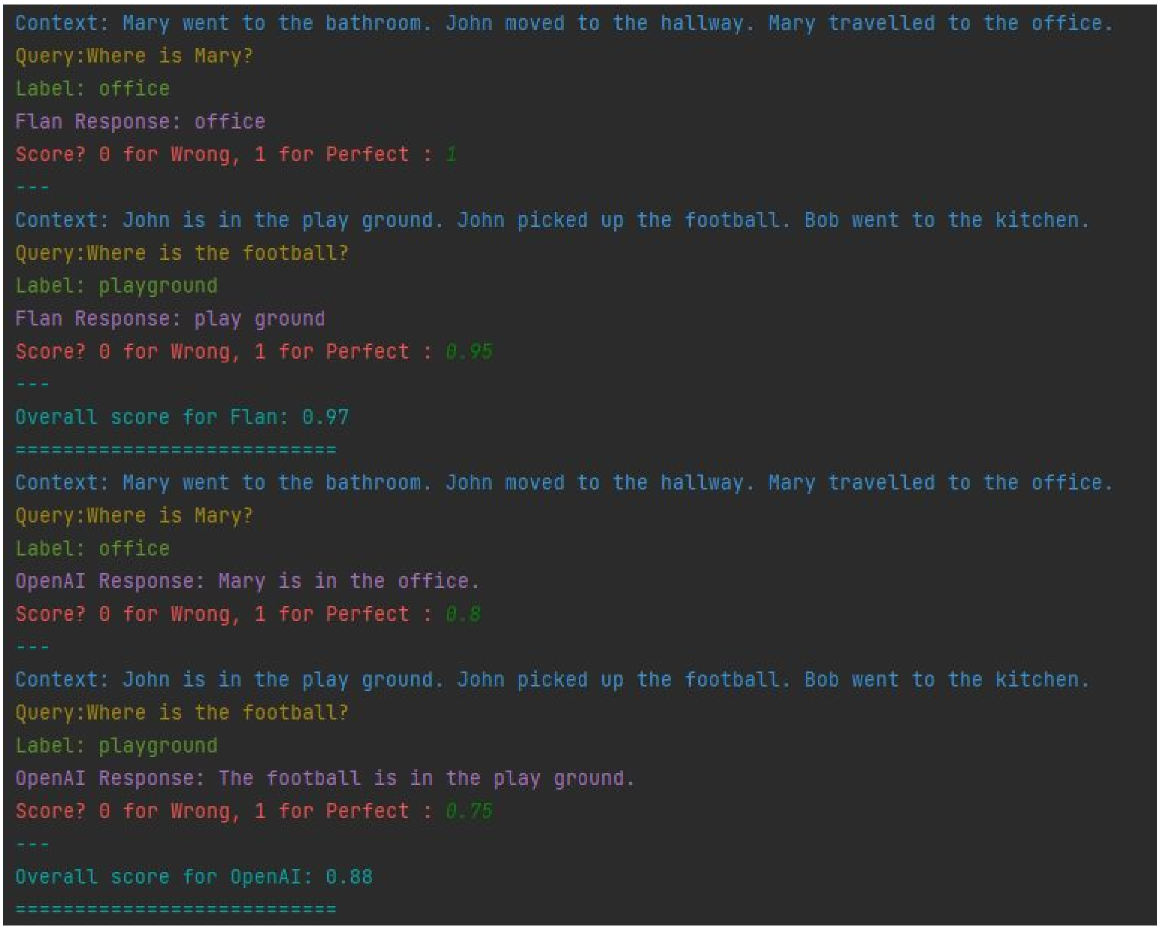
Written a toy-test-bed to do gauge the efficacy of these LLMs and open-sourced the program at my GitHub repo: yogeshhk/Sarvadnya/src/bAbi_tasks

There I run two models, OpenAI's GPT-3 and Google's Flan-T5-small, on a few bAbi tasks prompts (from Facebook AI), and let user (i.e. you) score the results. 💯

At the end, average score per model is given. Simple, crude but gives some way to compare. 💪

Of course, this can be expanded further in various ways, i.e. UI, auto-compare, more models, different prompts sets, etc., but you get the idea. 💡

#chatgpt #gpt #ai #openai #google #t5 #flan #nlp #langchain #facebook #babitasks #github #language



```
Context: Mary went to the bathroom. John moved to the hallway. Mary travelled to the office.
Query:Where is Mary?
Label: office
Flan Response: office
Score? 0 for Wrong, 1 for Perfect : 1
---
Context: John is in the play ground. John picked up the football. Bob went to the kitchen.
Query:Where is the football?
Label: playground
Flan Response: play ground
Score? 0 for Wrong, 1 for Perfect : 0.95
---
Overall score for Flan: 0.97
===========================
Context: Mary went to the bathroom. John moved to the hallway. Mary travelled to the office.
Query:Where is Mary?
Label: office
OpenAI Response: Mary is in the office.
Score? 0 for Wrong, 1 for Perfect : 0.8
---
Context: John is in the play ground. John picked up the football. Bob went to the kitchen.
Query:Where is the football?
Label: playground
OpenAI Response: The football is in the play ground.
Score? 0 for Wrong, 1 for Perfect : 0.75
---
Overall score for OpenAI: 0.88
===========================
```

51 Yogesh Sajanikar and 50 others 5 comments 2 reposts