Building Apps with GenAI in Production

Code Explorer and Work Journal – AI Coach

About you

- How many of you are individual contributors?
  - i.e. Software Engineer, Senior SE, Lead SE etc.
- How many of you are managers?
  - i.e. Team Lead, Engineering Manager, Senior EM etc.
- How many of you work in startups?
- How many of you worked on a side project using GenAI?
  - Can anyone of you share what you built?
- How many of you have a GenAI use case in production?
  - Can anyone of you share details of the use case?

About me

- Bachelors in Computer Engineering from National University of Singapore (2011 - 2015)
- 10 years in Software Engineering Industry in Singapore
- Headed Engineering (like CTO) for an acquired startup in public company listed in NYSE.
  - Managed a distributed software team of 15 (including contractors) across 5 countries.

LLM Basics

Under the hood

Supervised Learning

- Computer learns input and output mapping through labelled training data
- Steps
  - Get Labeled data
  - Train AI model on data
  - Deploy and call model
- Main tool used to train LLM

Neural Network

- Enormous, complex mathematical model of language
- Stores information
  - which words are commonly used with each other
  - which order they appear
  - high level capture what these words mean in context
- Mathematical representation of language (i.e. model) is used to generate new text

LLM Training

- Mathematical model has billions of individual parameters or numerical weights
- Before training model outputs random text
- During training, LLM shown incomplete pieces of text from training data
- LLM tries to predict which words come next
- Based on accuracy of predictions it will update internal parameters
- Model learns to produce factual information and linguistic styles

LLM Fundamentals

- Use neural network
- Repeatedly predict next word (actually token)
- Original phrase is called prompt
- Each of the final completed phrases is called completion
- Generates a probability distribution across every token
- Token choices made earlier will impact token selections later

Usual LLM Question

- Running the same prompt multiple times usually leads to different completions?
  - Randomly choses a token from probability distribution
  - Chooses one direction and takes it to completion (i.e. autoregressive meaning self-influencing)

LLM Types

- Base LLM
    - Predict next word based on text training data
- Instruction Tuned LLM
    - Tries to follow instructions
    - Trained first on base LLM and then on top with instructions
    - Refined using RLHF (Reinforcement Learning with Human Feedback)
- Most applications practical use – instruction tuned LLM

Prompt

Revolutionize AI app development

Prompt message roles

- System
  - Specifies overall tone of what you want LLM to do
- User
  - Specific instruction that you want to carry out
- Assistant
  - Message that was sent by LLM previously

Guidelines

- Give clear and specific instructions
  - Clear prompt is not equal to short prompt
- Give model time to think
- There is probably no perfect prompt to start with
- New Early Application – develop with one example
- Existing Mature Application – refine prompts with batch of examples

Tactics

- Use delimiters to separate distinct parts
- Helps avoid prompt injection
- Examples
  - triple backtics - ` ` `
  - Quotes – “”
  - triple Hyphen - ---
  - Tags - <hello>

Tactics

- Ask for structured output – HTML, JSON
- Check whether conditions are satisfied
  - Check assumptions required to do task
- Few-shot prompting
  - Give successful example of completing task
  - Then ask model to perform the task
- Specify steps required to complete a task
- Instruct model to work out own solution before rushing to conclusion

Usage

- Use at-most 5 sentences / 70 words / 350 characters
- Can use “extract” instead of ”summarize”
- Understand sentiment
  - Positive/negative, identify emotions
- Spelling or grammar corrections
- Transform inputs to other languages
- Transform formats – html to json

Temperature

- Allows variety or randomness in model response
- For tasks that require predictability
  - temperature=0
  - Same output every time
- For tasks that require variety
  - temperature=0.7
  - Different output every time

Code Explorer

Generate Summary of Pull Requests

Problem

- Created many Pull Requests (PRs) over a period of time
- Don’t remember details of the code written
- Not enough time to view all the diffs in all the PRs
- Has anyone faced this problem?
  - If yes, when did you face it?
  - Did you try to solve it?
  - If yes, what did you do?

Demo

- https://code.dayatwork.info
- Vibe coded almost entire app
- Code is open source
  - Frontend
  - Backend

Work Journal – AI Coach

Diary for Work Details

Problem

- Do you remember the details of work done last month?
- Have you struggled to recall about your past work in discussion with your team or manager?
- In an interview, have you faced the problem of not remembering details of the work?

Demo

iOS app

Android app

Questions?

FAQ: If the app is free, how will you make money?