Welcome

We will begin promptly at 11 AM ET.

If you are unable to hear the speakers, please let us know in the chat box. You may enter your questions in the Q&A, we will address them at the end of the presentation. You can find a copy of the slide deck and recording of this webinar: www.fdpinstitute.org/webinars
Financial Data Professional Institute

FDP Institute provides world class training and education to financial professionals to meet the accelerating needs of digital transformation in the industry.
Introductions

Today’s Topic:

Leveraging Large Language Models (LLMs) in Finance

Avi Patel, CAIA, FDP
Head of Marketing & Data Science, Fulton Bank

Raul Salles de Padua
Founder & Principal Data Scientist, Quod Analytics

Dr. Hossein Kazemi, CFA
Senior Advisor, CAIA Association & FDP Institute
Leveraging Large Language Models (LLMs) in Finance

Avi Patel, FDP, CAIA
Raul Salles de Padua
Contact info & resources

Avi Patel:
Email: avipatel68@gmail.com
LinkedIn: https://www.linkedin.com/in/avipatel/
YouTube: https://www.youtube.com/@AviPatel68
Github: https://github.com/mktaop

Raul Salles de Padua:
Email: raul.padua@iese.net
LinkedIn: https://www.linkedin.com/in/raulpadua/
YouTube: https://www.youtube.com/@sallespadua
Github: https://github.com/raul-padua
HF Spaces: https://huggingface.co/raul-padua
Agenda

1. Build up to Large Language Models (LLMs)
2. LLMs use cases and lifecycle
3. Prompt Engineering & Configuration Parameters
4. MLOps in GenAI Prompt Engineered Systems
Agenda

1. Build up to Large Language Models (LLMs)
2. High-level overview in the Transformers architecture
3. LLMs use cases and lifecycle
4. Prompt Engineering & Configuration Parameters
Real world application with financial documents

The Busy Banker's AI Wingman

Empowering Bankers, One Query at a Time.

Upload your Financial Report in a PDF file

Drag and drop file here
Limit 200MB per file • PDF

NASDAQ_AAPL_2022.pdf 0.7MB

Input your prompt on the financial document here

What were the company's Revenues last fiscal year?

The company's Revenues last fiscal year were $394,328 million.

https://www.youtube.com/watch?v=Q5wAmPT7Ixs
Build up to LLMs - Vector Space Models

From static embeddings to contextual embeddings ...on large corpora... but fail to draw out the domain-specific relationship/semantics

Couple of results from word embeddings experimentation from my paper
Build up to LLMs - RNN Architecture

For many years, language processing was done with recurrent neural networks. Predicting each token is based on the most recent word plus the hidden state of the past. Two key problems with this:

- Not parallelizable
- Not good at long range word associations since they weigh recent words more highly
Build up to LLMs - Transformers: Attention is all you need

**Encoder**

- Encodes inputs ("prompts") with sequences contextual understanding
- E.g. Bert: MLM, PNS

**Decoder**

- Takes in new tokens and generates new tokens
- E.g. GPT-3: Predict next token, recursive
Build up to LLMs—Transformers: Attention is all you need. Learn context for each combination of words.
Build up to LLMs - Size Matters

GPT-3 by far was the largest model, in 2020

Since then, just an explosion in scaling

Source: https://towardsdatascience.com/gpt-3-the-new-mighty-language-model-from-openai-a74ff35346f
Source: https://thelowdown.momentum.asia/the-emergence-of-large-language-models-lms/
That’s how we got to LLMs, scale and architecture. An emergent property of these LLMs is “in-context” learning, which was not expected. They can complete tasks that they were not explicitly trained on!

In-Context Learning

Zero-shot Learning

- Translate English to French
- Red =>

One-shot Learning

- Translate English to French
- Blue => Bleue
- Red =>

Few-shot Learning

- Translate English to French
- Blue => Bleue
- Yellow => Jaune
- Red =>

Allows us to skip the time and $ required for fine-tuning. *Using the same LLM, we can perform multiple tasks.*
Agenda

1. Build up to Large Language Models (LLMs)
2. LLMs use cases and lifecycle
3. Prompt Engineering & Configuration Parameters
4. MLOps in GenAI Prompt Engineered Systems
Leverage LLMs to perform tasks using your domain knowledge

Text Summarization

Question Answering

Machine Translation

EDA - Model Build w/ Prompts
Disrupting for better, also into AI development?

**Supervised Learning**

- \(~1-2\) months: Get labeled data
- \(3\) months: Train a model on that data
- \(~2-3\) months: Deploy & Call model in apps

**Prompt Engineering**

- \(~1-2\) days: Define prompt
- < 1 day: Call model
There are mainly three ways to leverage adoption of LLMs' Gen AI capabilities:

<table>
<thead>
<tr>
<th>Approach</th>
<th>How to...</th>
<th>Effort/ Cost</th>
</tr>
</thead>
</table>
| **1. Prompt Engineering** | - Commercial/ Open Source API calls  
- Instruct LLM through prompts to pass role and context  
- In-Context prompts: zero to few-shot | Low                         |

| **2. Fine-tuning** | - Instruction-led fine-tuning of pre-trained LLM with additional examples related to the task | Costly                      |

| **3. Build your own** | - Create and Train a new model from scratch | Very Costly                  |
… as they are setting stones within LLMOps lifecycle

**Scope & Constraints**
- Define application & Use case

**Model Selection**
- Choose model
  - pretrained
  - build your own pretrained model

**Tailor your model**
- Prompt Engineering
- Fine-tuning
- Align with human feedback

**Deploy into apps integration**
- Deploy for inference
- Build LLM application
- Monitor
Agenda

1. Build up to Large Language Models (LLMs)
2. LLMs use cases and lifecycle
3. Prompt Engineering & LLM's instruction led finetuning
4. MLOps in GenAI Prompt Engineered Systems
What is prompt engineering?

What were the company's Net Revenues?

What were the company's Net Revenues?

$247 billions
Zero-shot prompt engineering

Q: What were the company's Net Revenues?
A: 

Role: you are a financial advisor...

Q: What were the company's Net Revenues?
A: $247 billions
One-shot prompt engineering

Prompt

Q: What were the company's Net Revenues?
A: $247 billions

Q: What were the company's Net Profits?
A:

Completion

Role: you are a financial advisor...

Q: What were the company's Net Revenues?
A: $247 billions

Q: What were the company's Net Profits?
A: $7.89 billions
Few-shot prompt engineering

Q: What were the company's Net Revenues?
A: $247 billions
...
Q: What were Net Revenues in 2021?
A:

Role: you are a financial advisor...

Q: What were the company's Net Revenues?
A: $247 billions
...
Q: What were Net Revenues in 2021?
A: $230 billions

around 5-6 examples
Fine-tuning a pretrained model with instruction

Instruction-led finetuning: prompt-completion pairs with specific instructions to the LLM (1,000+ examples)

Role: you are a financial advisor...
Text: What were the company's Net Revenues?
Label: $247 billions

Role: you are a financial advisor...
Text: What were Net Revenues in 2021?
Label: $230 billions

Role: you are a financial advisor...
Text: What were the company's Net Profits?
Label: $7.89 billions
Sample prompts in instruction-led fine-tuning

Classification / Sentiment Analysis

"Given the following news article:
{{news_article}}
predict the associated sentiment (positive, neutral, and negative)"

Agent: Question Answering

""""You are an excellent document analyst specialized in the financial markets and industry.\nYou are great at performing the task queried in the prompt in a straightforward and easy to understand manner.\nYou answer the prompts always in English, translating from the document if it's not written in English language.\nIf the query is out of scope from the document answer back with {{out_scope}}.\nHere is the prompted query:
{{prompt}}""""
LLMs' finetuning process

Prompt: you are a financial advisor…
Text: What were the company's Net Revenues?
Prediction: $247 billions
Label: $247 billions
How to bring this to life in your organization?

- Sell the art of the possible to generate use-case(s) from the business line folks
  - Start with an use-case with lowest amount of risk to the firm, think POC
- Engage key stakeholders for buy-in, definitely:
  - Legal, Compliance, Information Security, Technology (IT), Risk Management
- Highlight the challenges that do exist, such as:
  - Latency, Cost, Context Window, Building Complex Chains, Model & Alignment Choice
Agenda

1. Build up to Large Language Models (LLMs)
2. LLMs use cases and lifecycle
3. Prompt Engineering & LLM's instruction led finetuning
4. High-level MLOps in GenAI Prompt Engineered Systems
"...start with the customer experience and work backwards with the technology".

Steve Jobs
... as they are setting stones within LLMOps lifecycle

Scope & Constraints

- Define application & Use case

Model Selection

- Choose model
  - pretrained
  - build your own pretrained model

Tailor your model

- Prompt Engineering
  - Fine-tuning
  - Align with human feedback

Deploy into apps integration

- Deploy for inference
- Build LLM application
- Monitor
Prompt Engineered system example

Cloud Infrastructure / On-premise (following security requirements)

Frontend
- django
- Streamlit

Backend
- LangChain
- llamaindex
- Caching
- Prompt templates
- Data Retrieval

API calls

Select model based on your criteria

Cost Mgmt
Config. Mgmt
Monitoring

OpenAI
LLaMA

How you validate the users' prompts?

What additional data to add context to your model do you need to include in the prompts?
Prompt Engineered system example

Cloud Infrastructure/ On-premise (following security requirements)

- **Frontend**
  - django
  - Streamlit

- **Backend**
  - prompt engine
  - llama-index
  - Caching
  - Data Retrieval

- API calls

- Consider where to store templates and user prompts outside your application

- Monitor/ Manage costs
  - Consider where to store templates and user prompts outside your application

- Ensure to stay within token constraints

- **Performance & Cost**

- **OpenAI**
  - LLaMA

- **Cost Mgmt**
- **Config. Mgmt**
- **Monitoring**
Q&A Example - Using domain-specific knowledge

Source: Sascha Heyer
Key takeaways

- Unmeasurable "low hanging fruits" prompt engineering LLMs
- You may fine-tune LLMs to specialize on your task, but keep in mind of the trade-offs
- Same MLOps principles apply when LLMs serve applications in production
Questions
Please join us for our upcoming webinar:

Register Here: https://bit.ly/3YsZIVi
Additional useful resources

- Original GPT-3 Paper: Language models are few-shot learners
- Attention is all you need – Transformer architecture
- Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings
- Word-to-Vector Paper
- ChatGPT is not all you need. A State of the Art Review of large Generative AI models.
- On the Opportunities and Risks of Foundation Models
- Holistic Evaluation of Language Models
- Text to powerpoint, yes you can
State of the art GenAI Models- Landscape from Jan 2023

Source: https://arxiv.org/abs/2301.04655