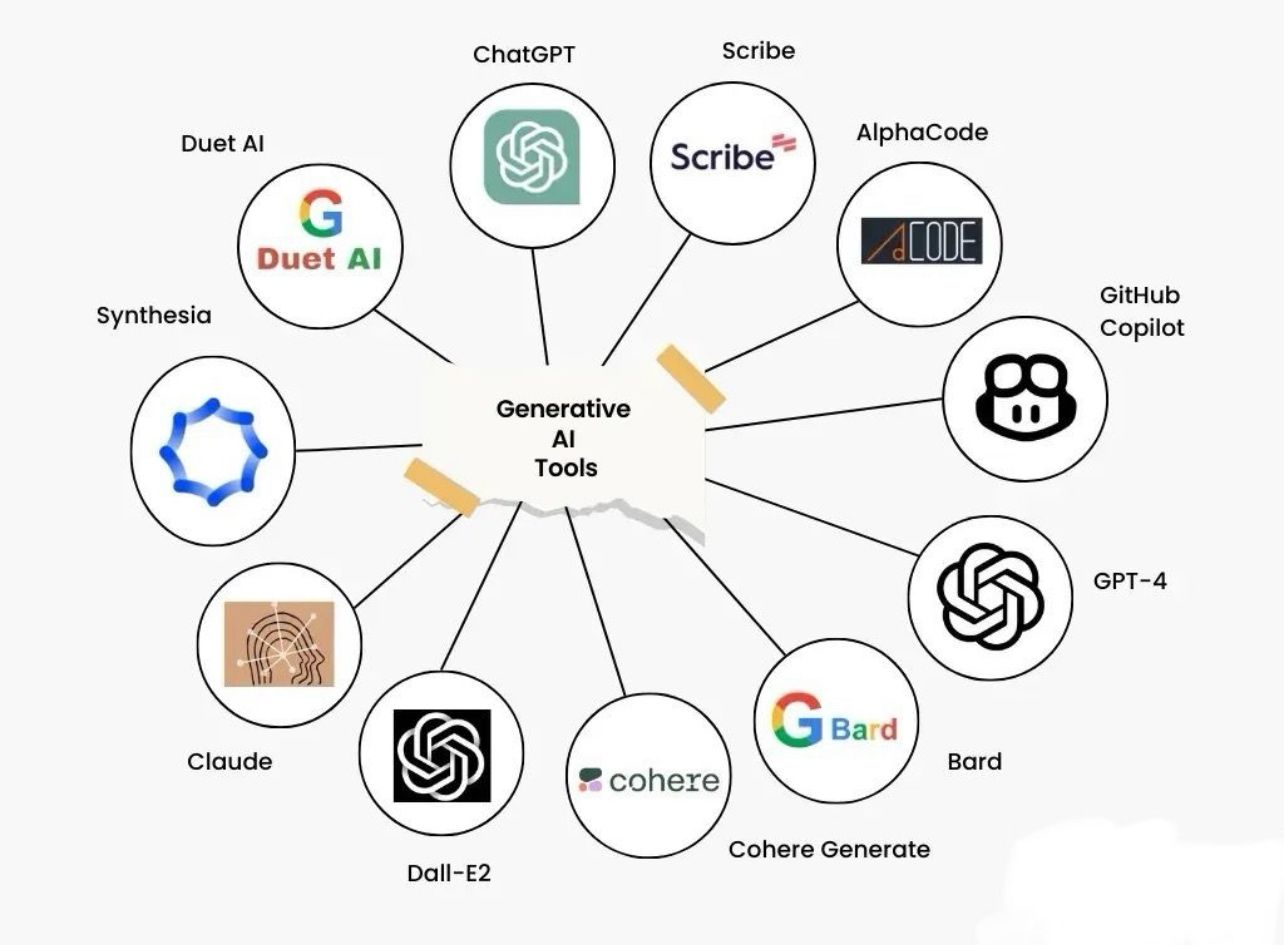
**Matthew Yeseta**

**Gen AI Architect**

**Here is how one must lead as a Gen AI Architect**

**End to End Gen Architecture Engineering**

Gen AI LLM, in leadership that transcends technical prowess, focusing on people, accountability, and strategic alignment. It requires balancing innovation with fiscal responsibility, ensuring that projects meet both executive expectations and customer needs.



Here is how I bring my people and tech leadership accomplishments (again) to be hired and lead for the Generative AI Architect Leadershikp’

**Chat GPT Generative AI My Leadership Practice That I Offer**

On Gen AI my leadership in projects is outlined from my experience and best practices. Here I offer my Gen AI LLM Chat GPT strategy on leading the team needs to navigate decision with the sponsors and stakeholders.

For this leadership to deliver Gen AI solutions the key is to focus on what benefit that business shall be delivered. How will this Gen AI improve customer engagements, customer loyalty and meet business targets, goals, KPI for Gen AI solutions in the business. As lead by example Generative AI Architect it is my responsibility to manage risk of Gen AI LLM limitations such as hallucination. My leadership will work with both business and teams to ensure that Gen AI LLM models are not bias. Also manage ethical Gen AI be delivered and any machine learning models that must be unbiased and inclusive for GPT Chats for customer responses. Key importance is to manage test case generation, manage fine tuning LLM Chat optimization. Look for approach to minimize vulnerability behavior and work with the team on best practices on defect analysis.

**Prompt Engineering for a Strategic Approach**

At its foundation, prompt engineering involves the deliberate crafting of inputs to guide a large language model (LLM) like OpenAI's GPT-4.x to generate precise and contextually relevant outputs. Given that the model's responses are constrained by the knowledge it has been trained on, effective prompt engineering acts as a crucial mechanism to elicit the desired information. Prompt engineering is the process of framing and refining questions or prompts to extract more accurate and meaningful results, akin to formulating the right query to get the most reliable response.

**Developing a Plan for Gen AI Prompt Engineering**

When developing prompt engineering for GPT-4.x or similar models, a team should adopt a systematic and strategic approach. Here’s a step-by-step outline for planning and executing prompt engineering:

**Objective Definition**

Begin by defining core objectives of the AI domain for the project implementation. The team needs to have a clear understanding of the desired planned outcomes from specific domain LLMs. We need to work with the sponsor to define use cases for 1) summation, 2) content generation, 3) sentiment analysis, 4) question / response. Well-defined objectives will ensure that prompt engineering aligns with the business goals and user requirements.

**Prompt Question Contextual Definition**

The team should understand the limitations and capabilities of GPT-4.x. We need to recognize that the GPT-4.x model is not able to access all information beyond its pre-training (see RAG for such solutions). It is essential to focus on prompts that leverage the knowledge encoded within the model and avoid scenarios where external, real-time data is required. This guides the types of questions and inputs the team should prioritize.

**Iterative Design and Testing**

Designing prompts is best done in an iterative process so we have some time for this effort. First find a use case domain and begin by experimenting on various question / response and observe how slight adjustments impact the model's responses. Plan for testing and refinement, which will lead to an optimal set of prompts that generate the most relevant and accurate responses. Prompt engineering are setup as zero-shot or few-shot learning responses to be applied to LLM.

**LLM and prompt engineering Evaluation, Metrics**

Together with the sponsor, team needs to establish metrics by evaluating the quality of responses generated by the model. This could include accuracy, relevance, coherence, and usability, depending on the application. The team should conduct A/B testing or user feedback sessions to measure the effectiveness of different prompts and fine-tune them accordingly.

**Collaboration and Cross-Functional Input**

Collaboration across departments, such as data science, product design, and subject matter experts, is critical. A cross-functional approach ensures that the prompts are not only technically sound but also aligned with real-world applications and user needs.

**Strategic Generative AI Thinking for Projects Delivery**

**Define Objective Alignment**

Work with the sponsor to define core business domain objectives that will lead to a clearer understanding how GenAI and LLM can be applied to specific domains in Chat and specific use cases.

**Partner with Sponsor for Requirement Analysis**

We will need to conduct thorough analysis of data that is availability and for which domain topics this applies to. We may also need computational resources for repository sizing or cloud sizing; hence we need to assess data size, data type, structure or unstructured, real time or batch. Also ask if there are any regulatory requirement compliance and how that will affect our design.

**LLM Framework**

LLM offers model selection and for LLM we need to know if there are any available pre-trained models that can base our initial exploratory analysis. Ask if any former fine-tuned for specific business domains have received solutions. OR determine if we need to define custom LLM for domain-specific knowledge (again we need to know which domain business we are to work on) Once we have a plan for data then we can plan for how we data repository for data data, or a plan how we shall upload custom curated data to a private data space on HuggingFaces. Also needed is to begin making a decision and authentication for OpenAI for access login to either our LLM with database back-end, or HuggingFaces authentication. Someone needs to take the lead on this since it may require some time to get authenticated. Next plan if the LLM can be supported by LangChain which can connect to our LLM or do we need to connect to RAG for external data sources.

**LangChain Utility**

LangChain is good at build data pipelines that permit dynamic interaction with LLM or with RAG for external data sources or with HuggingFaces for our curated customer data (questions/responses). LangChain supports memory, prompts, and user interaction loops.

**For AI and Gen AI Engineering here is an execution plan**

**Data Preparation and Curation**

Organize domain-specific data collection we can as a team utilized to improve the data responses generated by GPT-4.x. In order for a successful prompt engineer effort, all training datasets must contain diverse, high-quality samples. Data for the LLM should be made to closely resemble the type of interaction and outcomes we expect to use the final production system.

**Cleaning in Data Preprocessing**

Key before any LLM data or curated data is to clean and pre-process our data. Here we need to focus on removing inappropriate tones or biases. These efforts will align the LLM training set to represents the necessary professional conversation that is wrought in a normal discussion. Preprocessing focus can involve word tokenization, text semantics, word vectorization, and language transformation for the data into compatible formats for our LLM and repository.

**Data Augmentation**

Data augmentation is vital as we may need multiple source connections for the chat question-response data providers. LangChain will be used to connect with either RAG retrieval repository or other real-time data retrieval systems. This strategy with LangChain permits the project to be available for dynamic prompts conversational sessions and easier data adjustments and testing from sources.

**Prompt Engineering and Fine-Tuning**

The strategy for prompt engineering and fine tuning is to establish a baseline of several prompt templates, each of which reflects the business objectives. Build a zero-shot prompt templates to be defined and tested (fine tune) for when we need to classify a sentence as positive, negative, or neutral, then we use a zero-shot chat template. Build a few-shot prompt temples when we need a language translation for small tasks to understand the context, then we use a few-short chat template. Build a Chain-of-thought prompt template that can be utilized to generate intermediate steps, reasoning before delivering a final answer. LangChain will be used to chain these templates and prompts and orchestrate the chat effectively.

**Focus Iterative Fine-Tuning**

Prompt tuning involves fine-tuning the input prompts to improve the model's response to specific task, We will have initial prompts and tuned prompts. Prompt tuning adjusts the specificity or scope of the prompt question-response conversation model to closely match user needs. To be effective on this we need to engage in continuous refinement of prompt templates based on real-world interactions, feedback loops. And implement fine tuning with reward modeling and PEFT and LoRA in order to test and to optimize the quality of LLM outputs based on user feedback and specific business success metrics.

**Cross-functional Coordination**

For leading AI Engineers Data Scientists, lead with responsible for fine-tuning the LLMs, developing API, and ensuring that LangChain pipelines are effectively integrated with external data sources and workflows.

Lead team how to handle the preprocessing and curation of data, creating the training datasets and performing model validation and quality checks.

Ensure seamless deployment on cloud environments, handling the infrastructure scaling, system monitoring, and API performance tuning.

Product Owners must be aligned for AI development with business goals, defining the success metrics and ensuring that the LLM solutions meet end-user needs.

**Technical Leadership and Knowledge Transfer**

As a senior leader, it's essential to foster a learning environment where team members stay updated with advancements in LLM technologies. Organize regular knowledge-sharing sessions to explore the latest developments in transformers, fine-tuning techniques, and LangChain utilization for the business goals and targets to increase customer engagements, trust, loyalty.