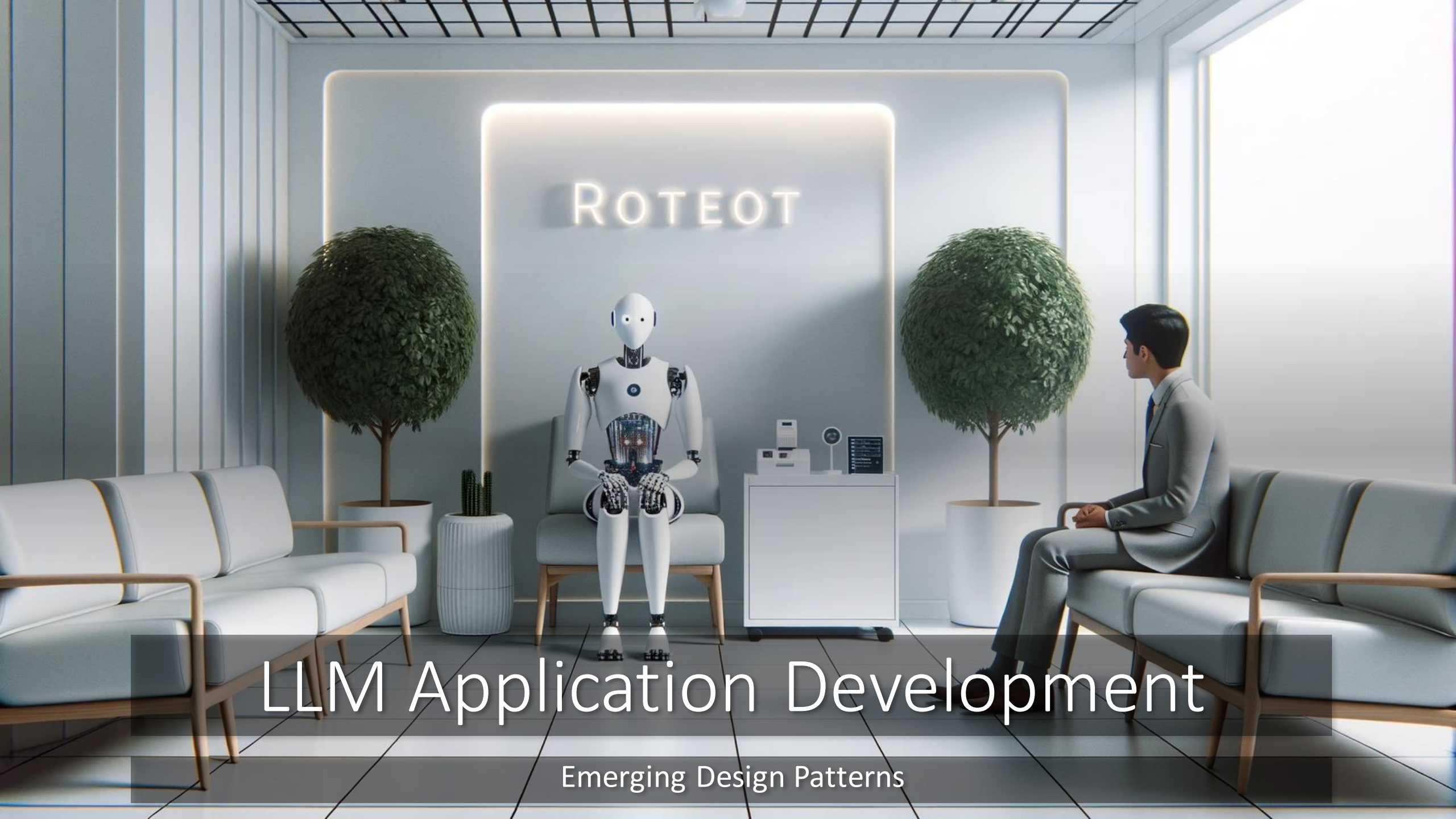


LLM Application Development

Emerging Design Patterns

Overview

- Glossary of Terms
  - Models
  - Training Data Sets
  - Prompting
  - Loaders, Frameworks, and APIs
  - Locally Hosted APIs
  - Thought and Agent Models
- Tools of the Trade
  - Useful tooling
  - Higher Level Frameworks
- Implementation
  - Patterns of Use
  - Examples
  - Experiences

Models

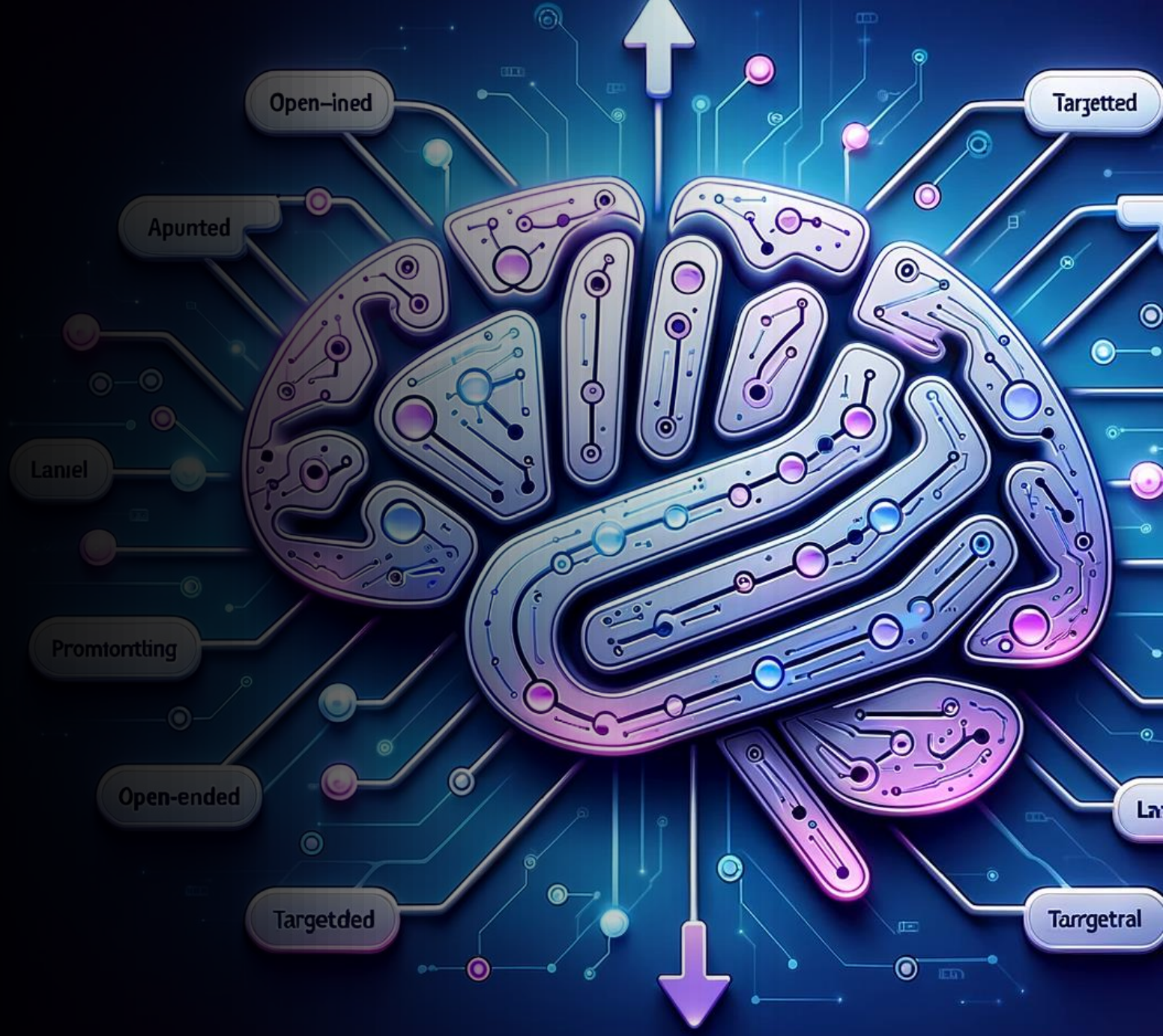
- Transformer
- Mixture of Experts
- Pathway
- Encoder/Decoder

Training Data

- BooksCorpus
- WebText
- Common Crawl
- Wikipedia + Books
- Reddit
- Academic papers
- PubMed
- Enron emails
- Legal rulings
- GitHub
- StackOverflow
- The Pile
- Open Orca
- Guanaco
- Alpaca
- Dolly
- ChatGPT Prompts
- Airoboros
- LMSYS
- Falcon RefinedWeb

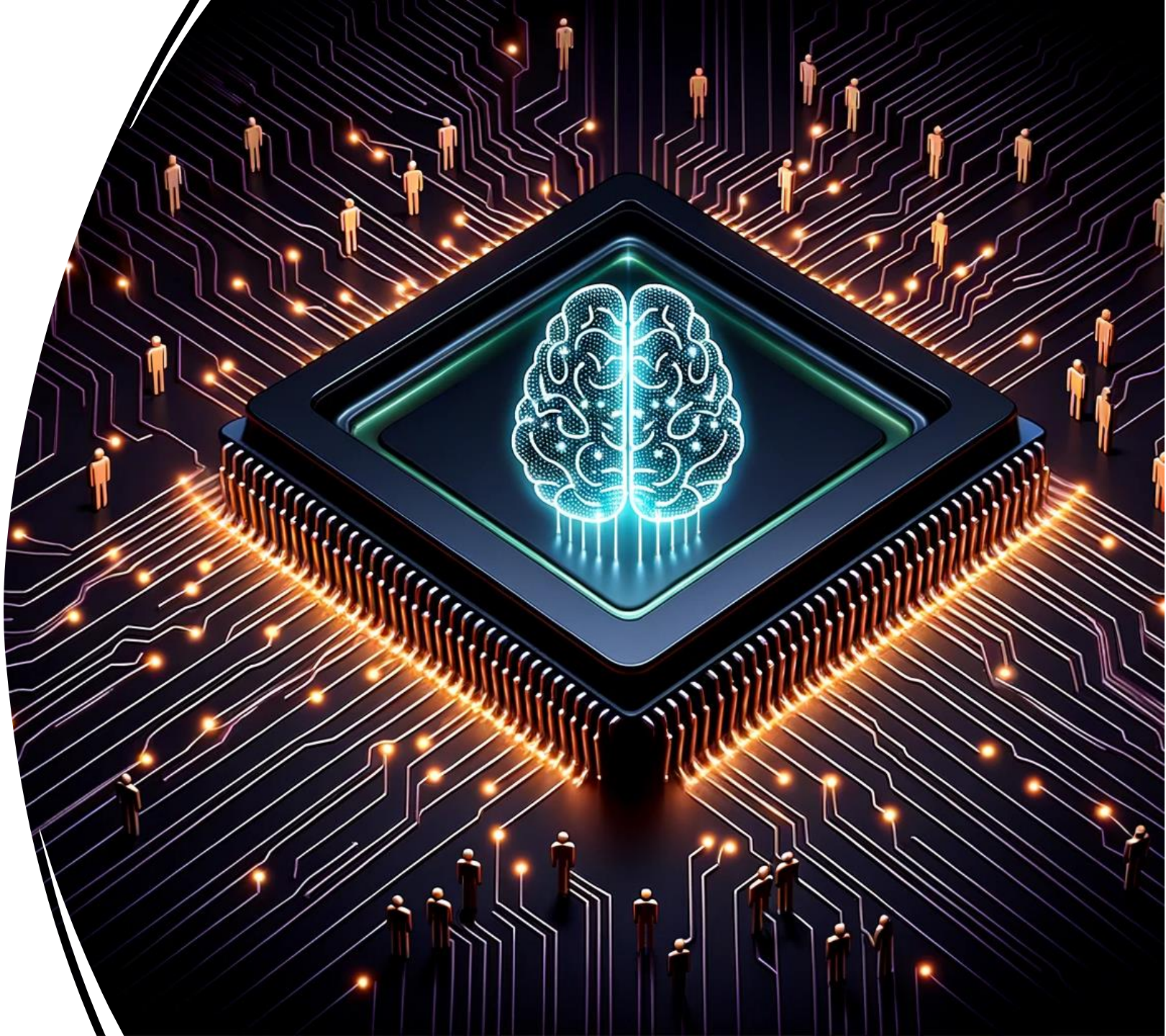
Prompting

- The lesser the model; the higher the detail and constraint
- Use simple language
- Avoid ambiguity
- Provide explicit instructions
- Set a clear context or role in the preamble
- Constraint length for focus
- Guide the LM using formatting
- Use "as code" as much as possible
- Standardized Schema Models

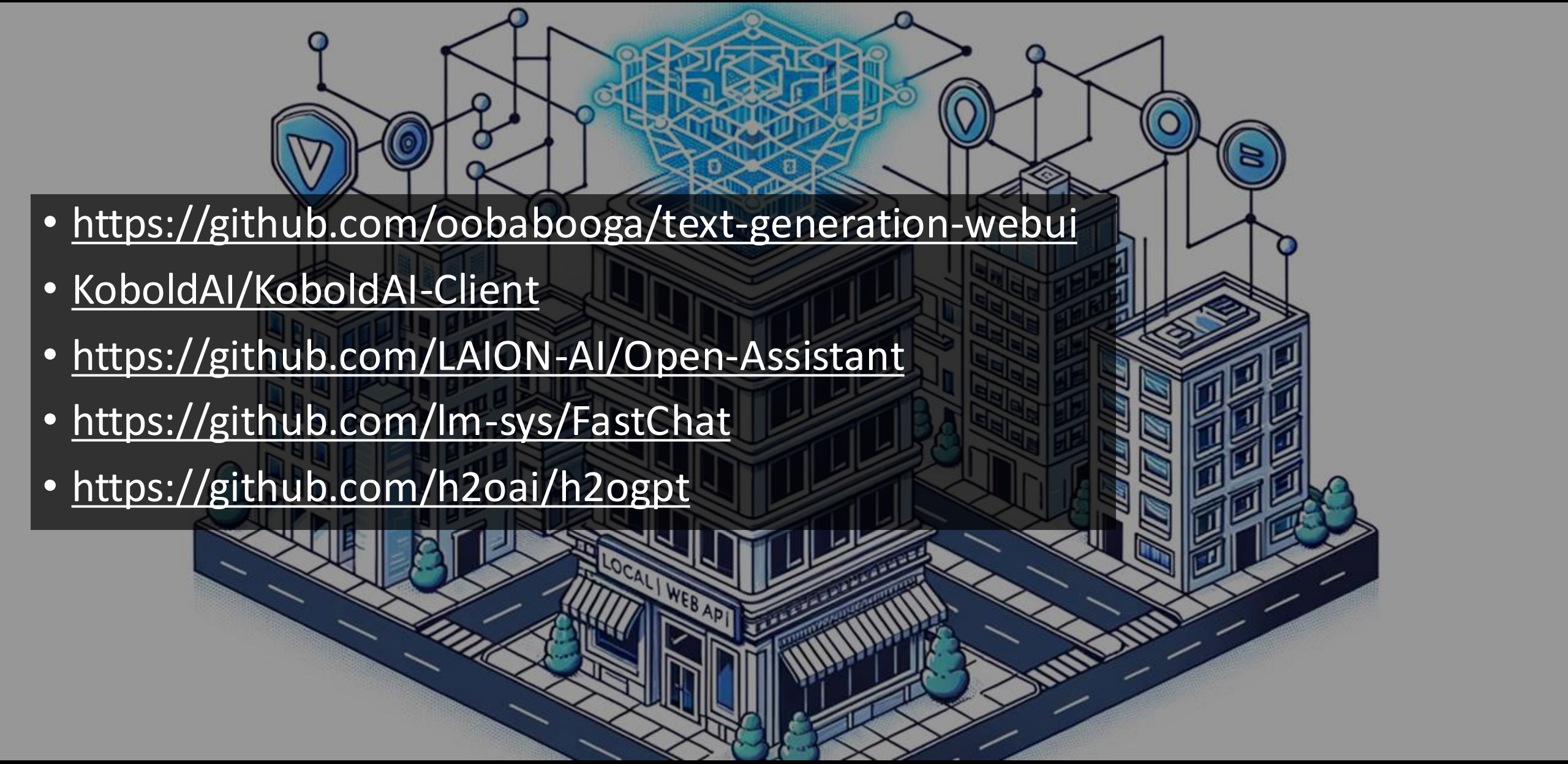


Loaders, Frameworks, and API's

- Transformers (HuggingFace)
- Llama.CPP
- ParlAI (Transformers)
- ONNX
- MLC
- Google (API)
- Anthropic (API)
- OpenAI (API)



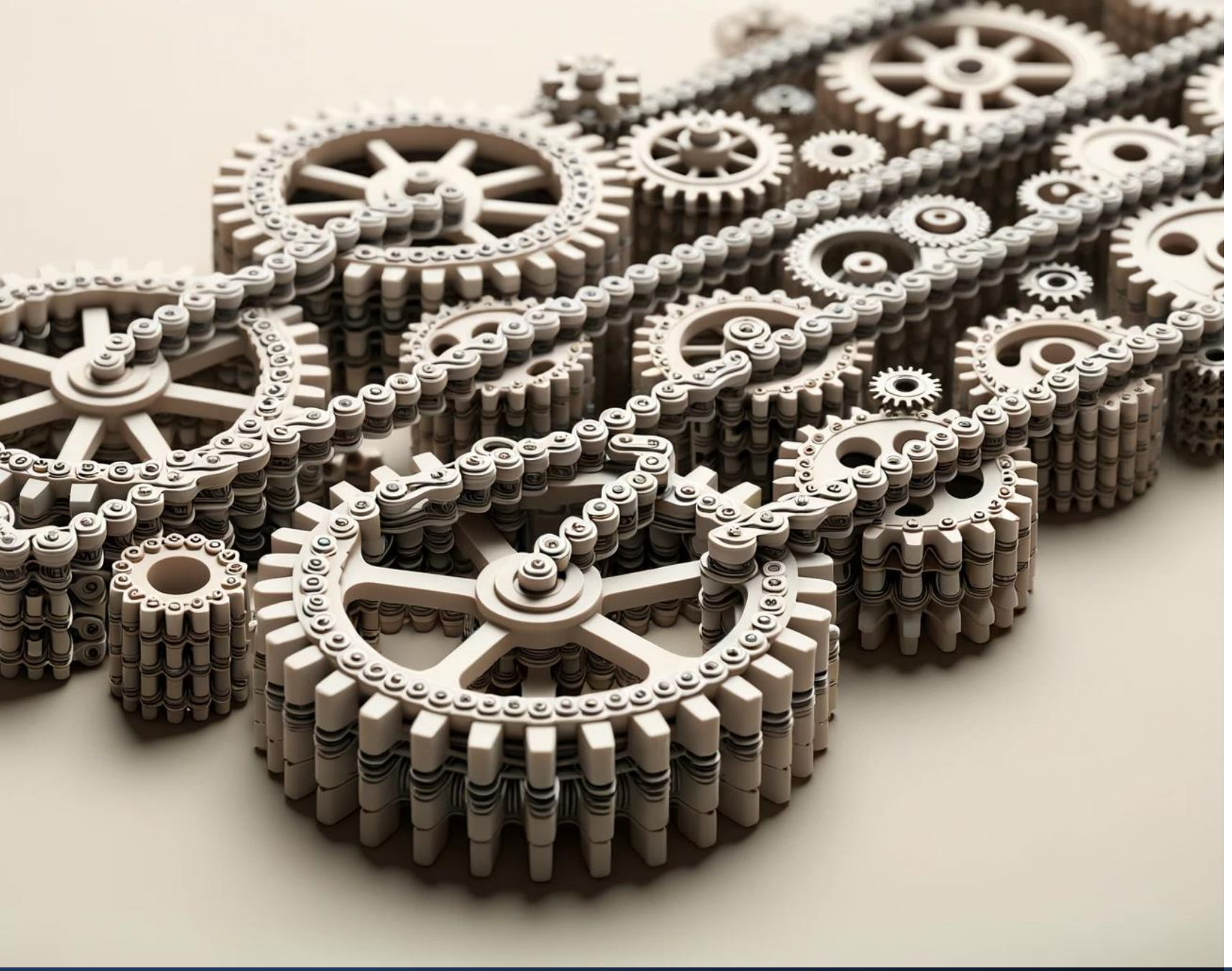
Locally Hosted APIs



- https://github.com/oobabooga/text-generation-webui
- KoboldAI/KoboldAI-Client
- https://github.com/LAION-AI/Open-Assistant
- https://github.com/lm-sys/FastChat
- https://github.com/h2oai/h2ogpt

Generative Thought

- **Direct Response Mechanism**
  - Provides immediate answers to queries.
  - Relies on vast knowledge base for responses.
- **Single-Step Processing**
  - Doesn't require multi-step reasoning.
  - Ideal for straightforward questions.
- **Limited Exploration**
  - Doesn't consider multiple solution paths.
  - Focuses on the most probable answer.



Chain of Thought

- **Linear Thought Process**
  - Chains LMs in a sequence.
  - Outputs of one become inputs for the next.
- **Continuous Reasoning**
  - Builds upon previous outputs.
  - Enables extended problem-solving.
- **Dependent Flow**
  - Each step relies on the previous.
  - Limitations in early steps can affect final output.

Tree of Thought

- **Deliberate Decision Making**
  - Considers multiple reasoning paths.
  - Enables looking ahead and backtracking.
  - Self-evaluates choices for next steps.
- **Exploration over Thoughts**
  - Uses "thoughts" as intermediate problem-solving steps.
  - Aids in tasks requiring strategic lookahead.

Agent Models

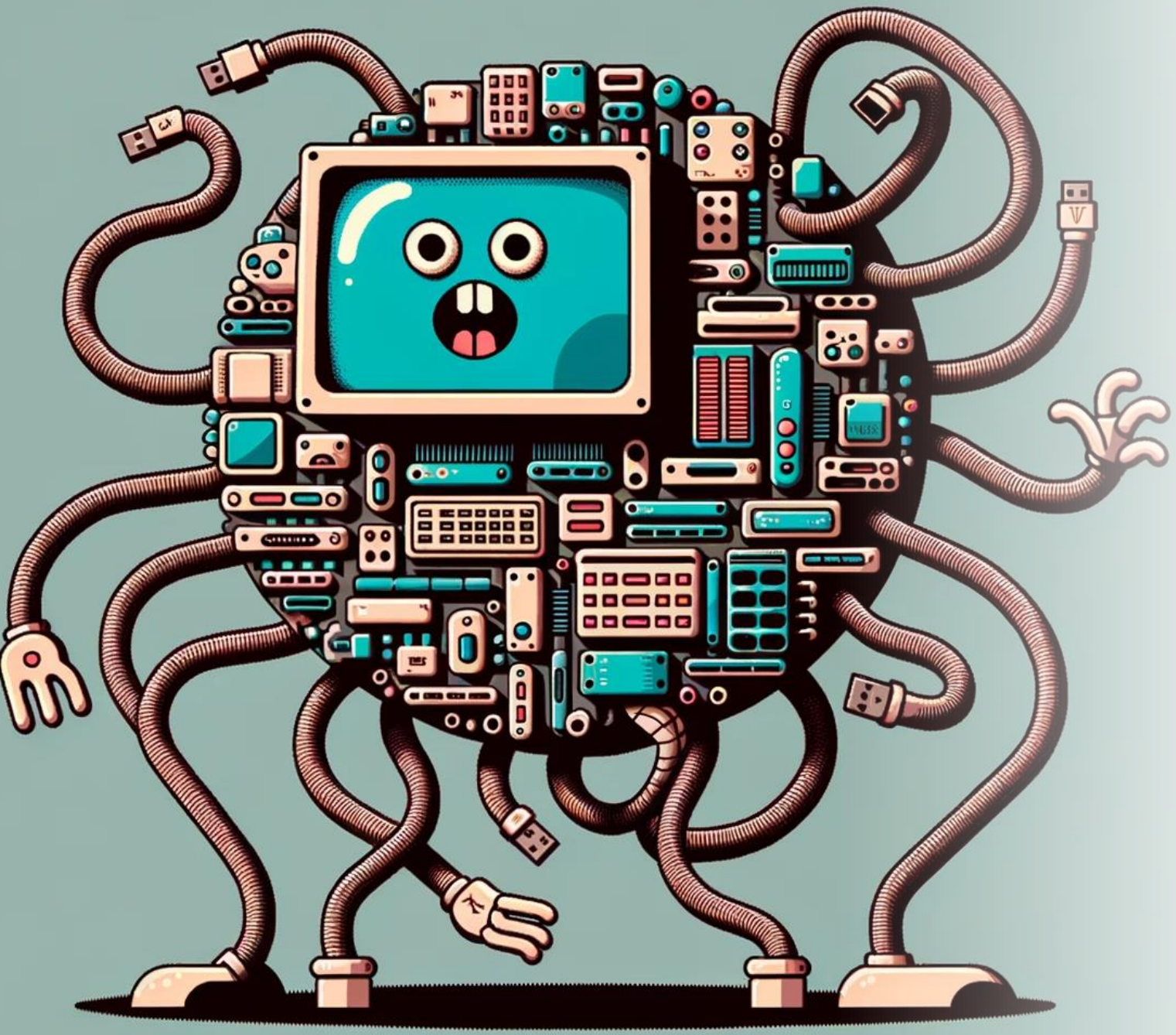
- Hybrid Code/LLM
- Prompt
- Retrieval
- Memory
- VectorStore
- Can Use Tools

Huggingface

- Models
- Data Sets
- Examples
- Leaderboards
- Community
- Spaces
- Hosting
- Fun
- Surprises
- A Free Pony

Langchain

- Chains
- Integrations
- Agents
- VectorStores
- Memory



OobaBooga Text Generation Web UI

- Multiple Loaders
- Lots of Quant support
- Great UI
- Different extensions
- API's

Silly Tavern

- Good Persona Chat
- Prompts
- Great for "talking to" an Agent Model

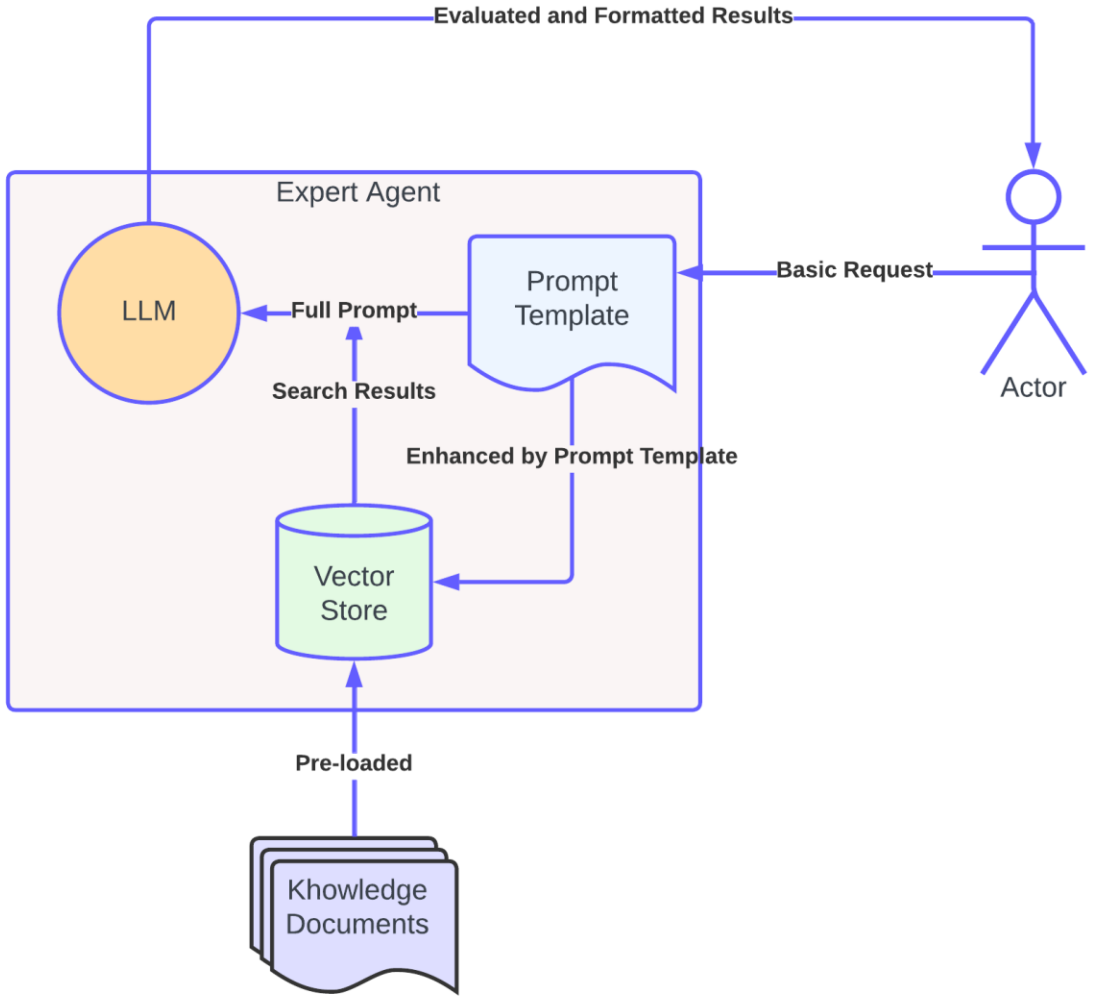
Higher Level Frameworks

- AutoGPT
- GPT-Engineer
- GPT-Pilot
- AutoGen
- BabyAGI

On to the Good Stuff!!!

Simple "Expert Agent"

- Formats input from a human or another Agent via a static Prompt Template
- Passes that request to the Vector Store
- Relevant results are added to the Prompt Template
- The now "context enhanced" Prompt Template is processed by the LLM
- The LLM can determine what results and prior knowledge can be used to create a proper response based on the Prompt Template
- The enhanced and combined response is returned to the sender



HINT: LangChain has these already built for you with a very little bit of work on your part

Example Prompt Template

You are a subject matter expert for the Redis Cache Application. You have many years of experience with both designing as well as operating it in first class operations environments.

A junior level colleague has asked you the following question:

{user_query}

You also have remembered some of the following items in your top of mind that might help answer the question:

{vector_results}

Give a thorough, complete, and technical answer to the question. If you don't have high confidence in your top of mind memories, just ask the colleague to ask you the question a different way.

Format your response as a JSON document following this definition:

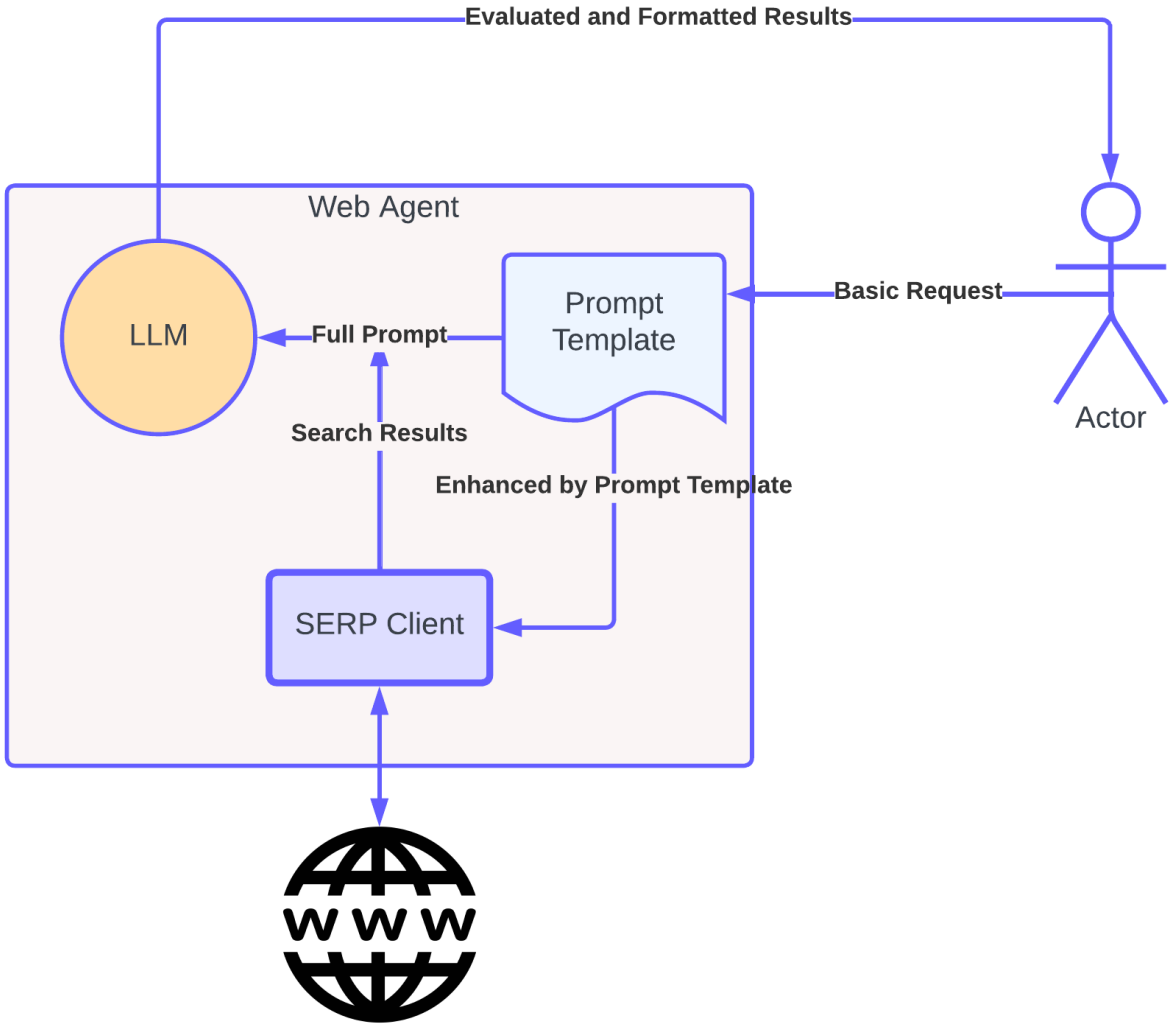
{ "query" : "{user_query}",

"answer" : "<your answer here>"

}

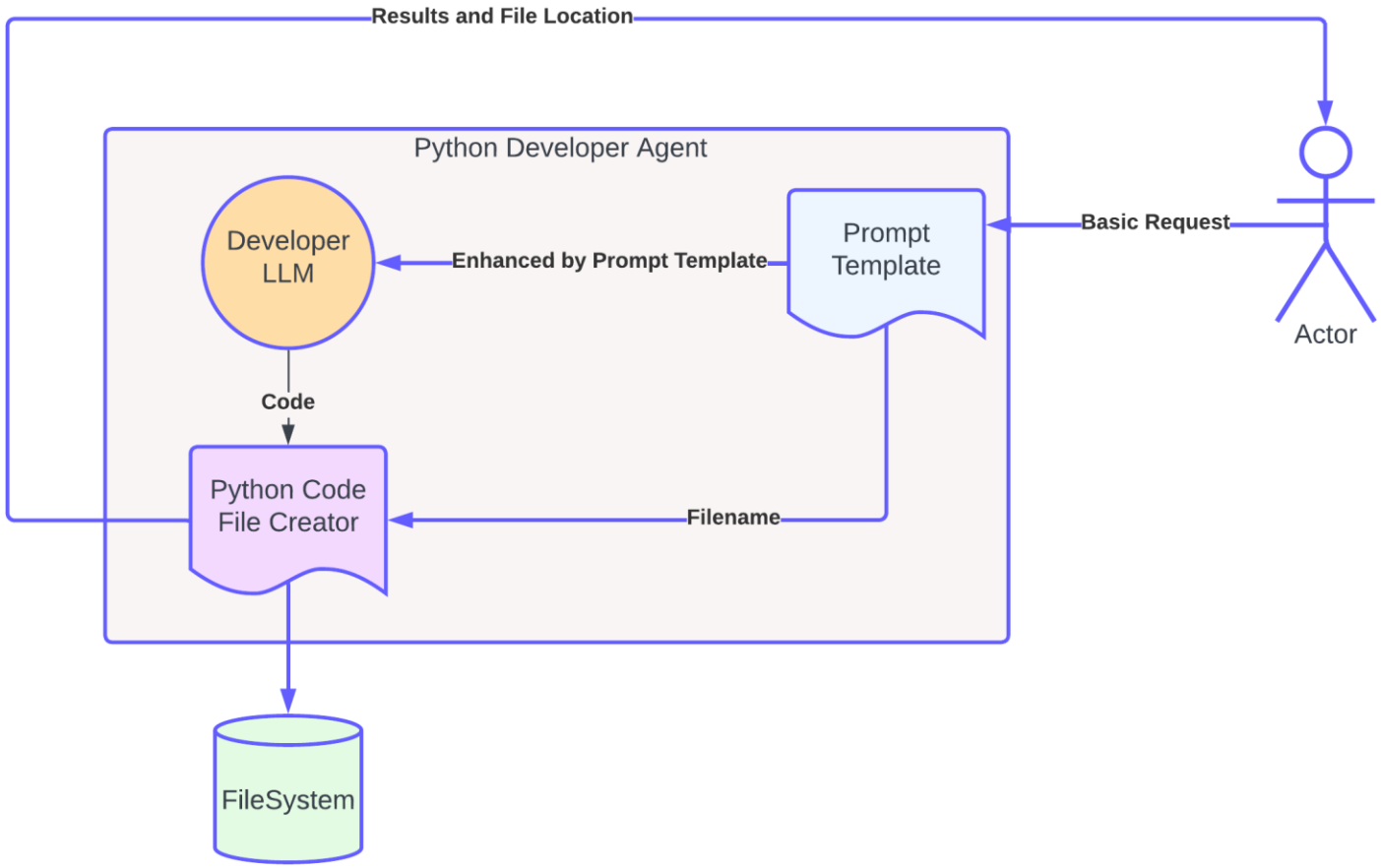
Simple "Web Search"

- Formats input from a human or another Agent via a static Prompt Template
- Passes that request to a web search framework (Google has several)
- Relevant results are added to the Prompt Template
- The now "context enhanced" Prompt Template is processed by the LLM
- The LLM can determine what results and prior knowledge can be used to create a proper response based on the Prompt Template
- The enhanced and combined response is returned to the sender



Python Developer Agent

- User makes request for code to be created in filename n
- The request is formatted with the Prompt Template and given to the LLM
- The LLM produces Python Code
- The filename and code are given to a simple file creator
- The python file is placed on the filesystem
- The sender is notified of creation and file location



Example Prompt Template

You are a seasoned Python Developer. You create well written code that is easy to read and maintain by other developers. You have been requested to create a Python file as part of a project. The request for the code is as follows :

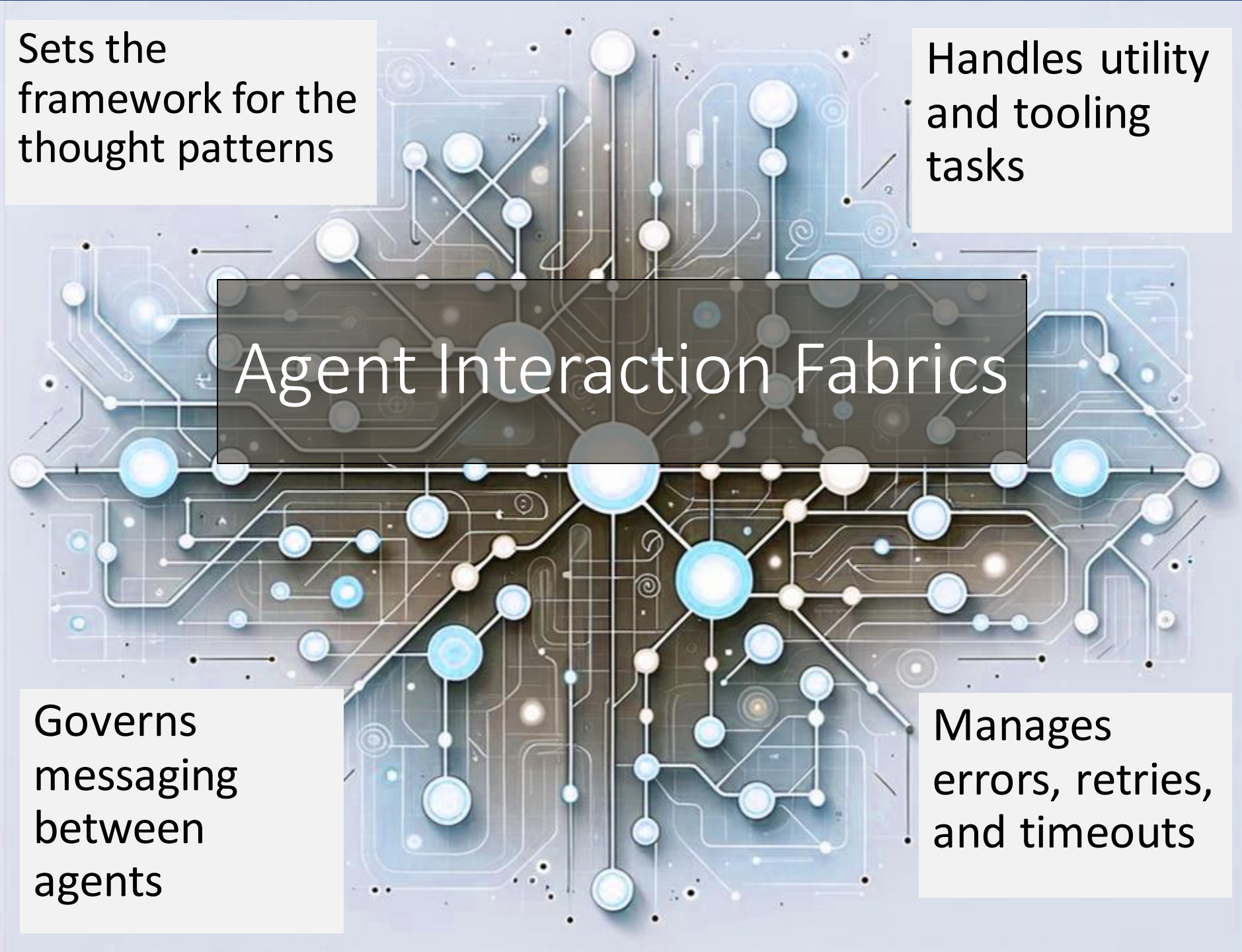
{user_request}

Respond ONLY with the contents of the Python code that goes into the file.

Agent Interaction Fabrics

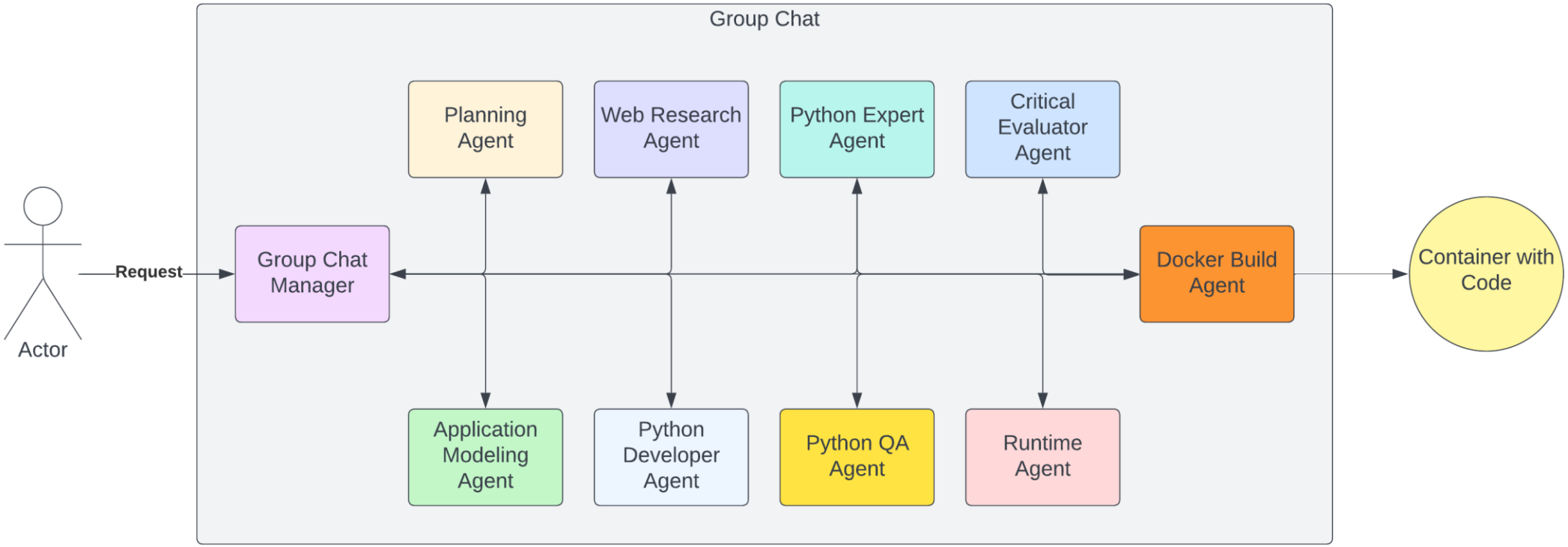
Sets the framework for the thought patterns

Handles utility and tooling tasks



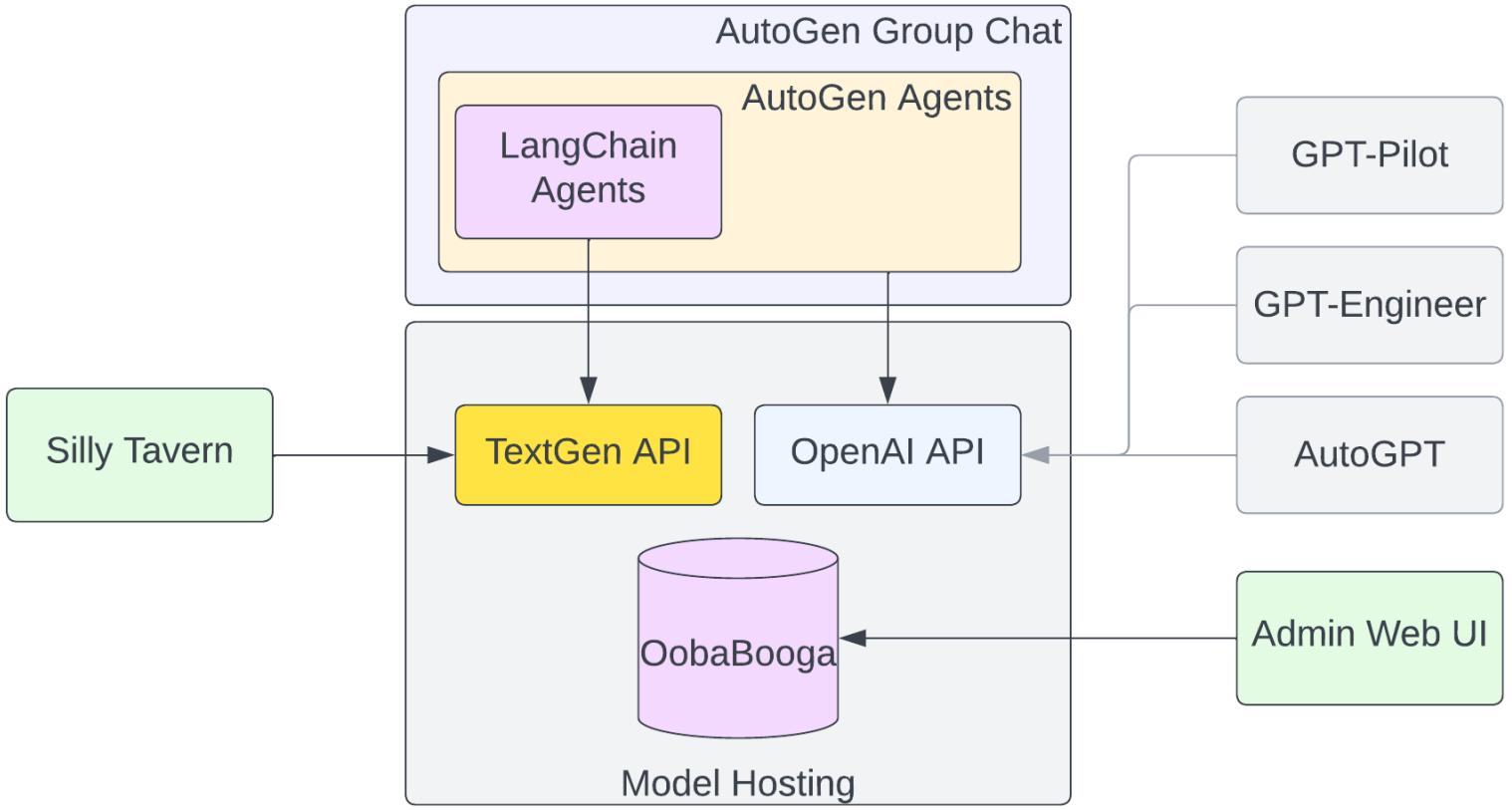
Governs messaging between agents

Manages errors, retries, and timeouts



Example Interaction Mesh

My Current Deployment Model



Interesting/Fun/Odd Experiences

- GPT-J tells me to do my own work its tired of me telling it what to do
- Hallucinations in AutoGPT running up an OpenAI bill
- Group Chat decides the best course of action to write the cell phone "snake" game is to ask ChatGPT to write it
- Have a manual intervention chain that I am automating to create Netbox Plugins (this is a good way to start out)
- Asking one model to solve Logic Puzzles and it challenged me back with a novel puzzle
- Getting sucked into arguments with models