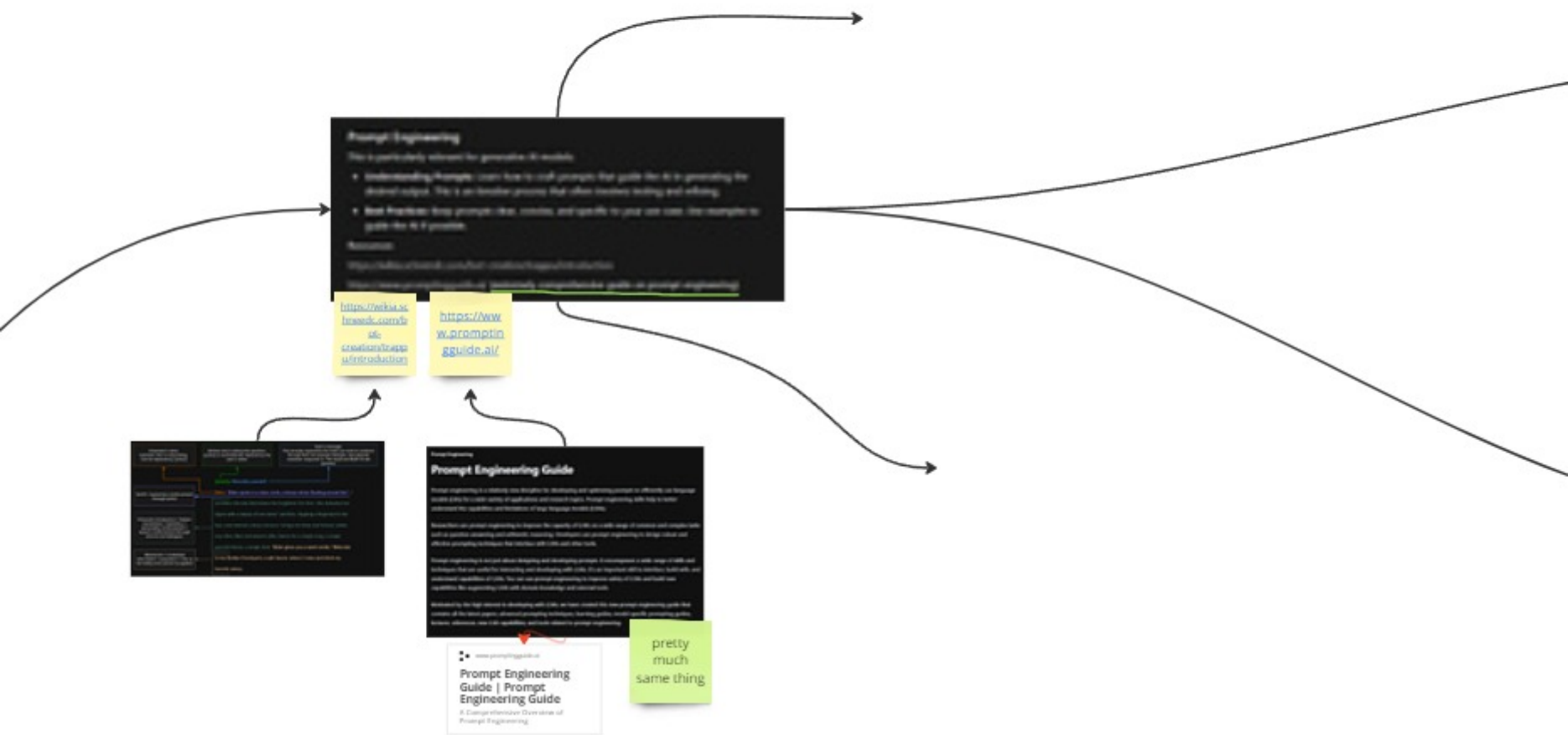


## Choosing the AI model



# Managing prompts/Prompting



# Tuning AI

**Fine-Tuning Model with Datasets**

When you train a model, you're essentially teaching it to perform a task. In this case, you're training a model to generate text based on the input it receives. The model is trained on a dataset of text, and it learns to predict the next word in the sequence. This process is called fine-tuning, and it's a key part of training a model to perform a specific task.

Intentional fine-tuning further, such as, is obtaining more specific knowledge or context (e.g. about  
language, images, time, location, etc.)

• Reading Datasets includes the following text, though, it's not a complete learning repository, and

Google Dataset Search is great for finding public data. Take note of the license. An alternative is to create your own dataset, which you can accomplish through several steps (mentioned in this tutorial) from the dataset definition to how you train your model.

**Resources**

[https://colab.research.google.com/github/ashishpatel26/LLM-Finetuning/blob/main/2.Fine\\_Tune\\_Your\\_Own\\_Llama\\_2\\_Model\\_in\\_a\\_Colab\\_Notebook.ipynb](https://colab.research.google.com/github/ashishpatel26/LLM-Finetuning/blob/main/2.Fine_Tune_Your_Own_Llama_2_Model_in_a_Colab_Notebook.ipynb)

[https://colab.research.google.com/github/ashishpatel26/LLM-Finetuning/blob/main/2.Fine\\_Tune\\_Your\\_Own\\_Llama\\_2\\_Model\\_in\\_a\\_Colab\\_Notebook.ipynb](https://colab.research.google.com/github/ashishpatel26/LLM-Finetuning/blob/main/2.Fine_Tune_Your_Own_Llama_2_Model_in_a_Colab_Notebook.ipynb)

[https://colab.research.google.com/github/ashishpatel26/LLM-Finetuning/blob/main/2.Fine\\_Tune\\_Your\\_Own\\_Llama\\_2\\_Model\\_in\\_a\\_Colab\\_Notebook.ipynb](https://colab.research.google.com/github/ashishpatel26/LLM-Finetuning/blob/main/2.Fine_Tune_Your_Own_Llama_2_Model_in_a_Colab_Notebook.ipynb)

[https://github.com/ashishpatel26/LLM-Finetuning/blob/main/2.Fine\\_Tune\\_Your\\_Own\\_Llama\\_2\\_Model\\_in\\_a\\_Colab\\_Notebook.ipynb](https://github.com/ashishpatel26/LLM-Finetuning/blob/main/2.Fine_Tune_Your_Own_Llama_2_Model_in_a_Colab_Notebook.ipynb)

ashishpatel26/LLM-Finetuning

LLM-Finetuning

github.com

**LLM-Finetuning/2.Fine\_Tune\_Your\_Own\_Llama\_2\_Model\_in\_a\_Colab\_Notebook.ipynb at main · ashishpatel26/LLM-Finetuning**

LLM Finetuning with peft. Contribute to ashishpatel26/LLM-Finetuning development by creating an account on GitHub.

we need to make LoRA for the story generation for those models where it will understand player input

[https://colab.research.google.com/drive/1PEQyQ1-6j0S\\_Xj8DV50NkpzasXkrzd?usp=sharing](https://colab.research.google.com/drive/1PEQyQ1-6j0S_Xj8DV50NkpzasXkrzd?usp=sharing)

[https://colab.research.google.com/github/ashishpatel26/LLM-Finetuning/blob/main/2.Fine\\_Tune\\_Your\\_Own\\_Llama\\_2\\_Model\\_in\\_a\\_Colab\\_Notebook.ipynb](https://colab.research.google.com/github/ashishpatel26/LLM-Finetuning/blob/main/2.Fine_Tune_Your_Own_Llama_2_Model_in_a_Colab_Notebook.ipynb)

colab.research.g...



Google Colaboratory

Google Colaboratory

## Quantizing Models

Quantizing a model refers to the process of reducing the precision of the numbers used to represent a model's parameters. This is typically done by reducing the number of bits that represent each weight from 32-bit floating points to lower-precision formats like 16-bit integers.

**Resources**

[https://mlabonne.github.io/blog/posts/Quantize\\_Llama\\_2\\_models\\_using\\_ggml.html](https://mlabonne.github.io/blog/posts/Quantize_Llama_2_models_using_ggml.html)

[https://colab.research.google.com/drive/1DPr4mUQ92Cc\\_xf4GgAaB6dFcFnWlvqYi?usp=sharing](https://colab.research.google.com/drive/1DPr4mUQ92Cc_xf4GgAaB6dFcFnWlvqYi?usp=sharing)

[https://colab.research.google.com/drive/1pL8k7m04mgE5jo2NrjGi8atB0j\\_37aDD?usp=sharing](https://colab.research.google.com/drive/1pL8k7m04mgE5jo2NrjGi8atB0j_37aDD?usp=sharing)

[https://mlabonne.github.io/blog/posts/Quantize\\_Llama\\_2\\_models\\_using\\_ggml.html](https://mlabonne.github.io/blog/posts/Quantize_Llama_2_models_using_ggml.html)

[https://colab.research.google.com/drive/1pL8k7m04mgE5jo2NrjGi8atB0j\\_37aDD?usp=sharing](https://colab.research.google.com/drive/1pL8k7m04mgE5jo2NrjGi8atB0j_37aDD?usp=sharing)



Google Colaboratory

mlabonne.github.io

**ML Blog - Quantize Llama models with GGUF and llama.cpp**

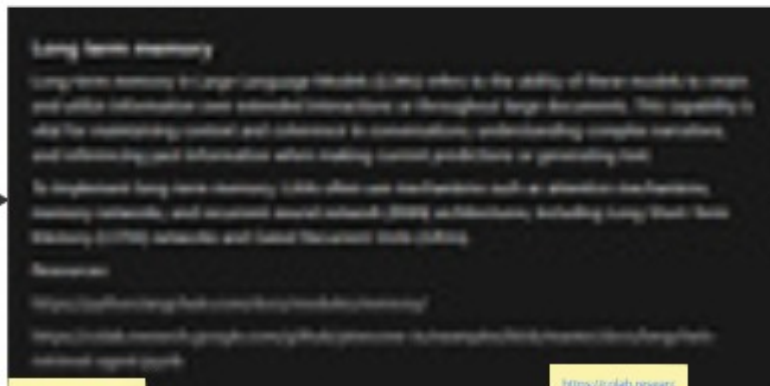
[https://colab.research.google.com/drive/1DPr4mUQ92Cc\\_xf4GgAaB6dFcFnWlvqYi?usp=sharing](https://colab.research.google.com/drive/1DPr4mUQ92Cc_xf4GgAaB6dFcFnWlvqYi?usp=sharing)



Google Colaboratory

# Memory for story script

We need to manage this part so AI director won't get lost in story and can actually keep up with all the player progression



<https://python.langchain.com/docs/modules/memory/>

<https://colab.research.google.com/github/insane-io/examples/blob/main/notebooks/langchain-extension-agent.ipynb>



# Storing the memory somewhere

We can use Filecoin to store player progress...

Still need to check it out



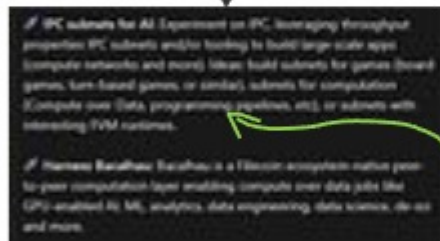
Most promising AI x Web3 Build Using Filecoin and/or Protocol Labs Tech

**PL Architect**  
**Protocol Labs**

**Challenge:**  
Best project submitted with the potential to turn into a startup!

1st place: \$2,500  
2nd place: \$1,500  
3rd place: \$1,000

Filecoin &/or Protocol Labs tech includes: Filecoin, Filecoin VM.



IPC subnets for AI Experiment on IPC, leveraging throughput properties IPC subnets and/or looking to build large scale apps (compute networks and more) like: build subnets for games (board games, turn-based games, or similar), subnets for computation (compute over data, programming, pipelines, etc), or subnets with interesting TVM runtimes.

Harvest Barefoot: Barefoot is a Filecoin ecosystem native peer-to-peer computation layer enabling compute over data jobs like GPU-enabled ML, analytics, data engineering, data science, etc and more.



**Build Ideas:**

**AI Data Explorers:** Quality data determines the quality of many AI projects. Filecoin's network is great for storing raw data and training AI models. We encourage ideas that combine Filecoin with AI training, e.g. a data explorer that helps AI companies to find correct public data on Filecoin for training.

EXTRA  
BOUNTY



This thing took my eye

# Story telling and prompt generation

this

mistralai/Mistral-7B-Instruct-v0.1

Text Generation Transformers PyTorch Sololearns

The player will specify the main character of the mystery and the what should be the story about and you need a template for more cohesion and also ensuring best formatting.

Here, I can do that there's an example. The player specifies the main character as a detective named Jack. The story is about Jack investigating a series of mysterious disappearances in a small town. Prompt: You are Jack, a detective, and you have been called to investigate a series of mysterious disappearances in a small town. The first person to go missing was a young woman named Sarah, and the second was an elderly man named Mr. Johnson.

story-telling  
is quite nice

I apologize for the confusion in my previous response. Here is a revised prompt that should be clearer: The player specifies the main character as a detective named Jack. The story is about Jack investigating a series of mysterious disappearances in a small town. The first person to go missing was a young woman named Sarah, and the second was an elderly man named Mr. Johnson. You are Jack, a detective, and you have been called to investigate the disappearances. As

but it causes lot of  
confusion for this AI so we  
will need to run more  
instances that will  
communicate with each other  
and the database or the  
conversation won't go for  
long (single prompt  
conversations)

you are Jack, a detective, and you have been called to investigate a series of mysterious disappearances in a small town. The first person to go missing was a young woman named Sarah, and the second was an elderly man named Mr. Johnson. You are Jack, a detective, and you have been called to investigate the disappearances. As the scene of the crime of the second disappearance, a small house on the outskirts of town, the first door is open, and you can see that the interior of the house is in complete disarray, and several belongings are scattered throughout the room. In the living room, you notice a laptop with a note on it. The note reads: "I'm sorry, I had to leave. I'll be back."

after the prompt is  
passed to another  
instance it made  
clear image of the  
task and the prompt  
is actually working

this case is if we  
use stability a.k.a.  
stable diffusion  
for image  
generation



# Suggestion from GPT about running AI models

Running it on locally as standalone will require lot of data

Having accesibility done through server may solve hardware problem but force users to have internet connection

## Running AI Models Locally on a User's Machine

1. **Hardware Variability:** The primary challenge with running these models locally is the wide variability in users' hardware. Not everyone has a high-end GPU, and models requiring substantial computational resources (like Llama 2 or similar) might not run efficiently on average consumer-grade hardware.
2. **Model Size:** Large models have high memory requirements. Smaller, more efficient models can be run on less powerful machines but might not deliver the same level of performance or capability as the larger models.
3. **Distribution and Updates:** Distributing the model with your game and keeping it updated can be cumbersome. The model itself can be several gigabytes in size, adding to the game's download size.

### Solutions for Local Deployment:

- \* **Model Distillation:** Using smaller, distilled versions of these models might allow them to run on a wider range of devices. These models offer a good balance between performance and resource requirements.
- \* **Dynamic Loading:** Consider loading the model only when needed to generate story content, potentially reducing the ongoing resource requirements during gameplay.

the models I am testing are not performing pretty good but are quite hardware costly

need to test out capability of storytelling

## Running AI Models Online

1. **Accessibility:** Running the model on a server and accessing it via the internet can mitigate the hardware limitations on the user's end. Users can interact with the AI model regardless of their local hardware capabilities.
2. **Scalability:** You can scale your server resources based on demand, ensuring a consistent and responsive experience for users. This is particularly useful for handling spikes in usage without requiring each user to have powerful hardware.
3. **Maintenance and Updates:** It's easier to update and maintain the model, as you only have to update the server-side model without requiring users to download updates manually.

### Challenges for Online Deployment:

- \* **Latency:** Depending on the user's internet connection and your server's location, there might be noticeable latency in interactions with the AI.
- \* **Cost:** Running servers capable of handling AI workloads, especially for many users simultaneously, can be expensive. You'll need to manage and optimize server costs against the service you're providing.

Game servers included - pay provider or tokens

### Hybrid Approach

Considering the trade-offs, a hybrid approach might be the best solution. For instance, you could use a smaller, more efficient model locally for immediate, less complex interactions and rely on a server-side, more powerful model for generating significant story developments or more complex content that can be loaded dynamically.

## Recommendations for Game Development

- \* **Prototype with Online Models:** Start by prototyping your game using online models to understand the interactions and computational needs without upfront hardware investments.
- \* **User Testing:** Conduct testing to see if the latency in online interactions is acceptable for your game's experience.
- \* **Cost-Benefit Analysis:** Evaluate the costs of running models online versus the potential limitations and player satisfaction of running models locally.

This way, you can provide a rich, AI-driven storytelling experience in your game, leveraging the strengths of both local and online computation based on the specific needs of your game and its audience.

we just train it somewhere XD

# Comparsion

## Llama 2

Llama 2 is a successor to the original Llama model, developed by researchers to improve upon the capabilities of its predecessor in understanding and generating text. Here are some key advantages of Llama 2:

1. **Improved Performance:** Llama 2 typically shows enhanced performance in a wide range of language tasks, including text completion, translation, and question-answering, due to advancements in its training methodology and data processing.
2. **Efficiency:** The model may be designed to be more efficient in terms of computational resources, making it more accessible for deployment in various environments, including those with limited hardware capabilities.
3. **Versatility:** Llama 2 can handle a diverse set of languages and dialects, making it highly versatile for global applications.
4. **Customizability:** It's often built with customization in mind, allowing developers to fine-tune it for specific tasks or industries, enhancing its effectiveness for particular applications.

## Mistral-7B-Instruct

Mistral-7B-Instruct appears to be a hypothetical or less well-known model in the context of your query. However, if we consider it in a general sense, models like "Mistral-7B-Instruct" that include "Instruct" in their name often aim to follow instructions or prompts more accurately than their predecessors. Assuming it follows the trend of instruction-following models (like GPT-3's Instruct series), here could be its advantages:

1. **Precision in Following Instructions:** Models with an emphasis on following instructions are designed to better understand and execute specific tasks outlined by the user, potentially making them more useful for applications requiring precise outcomes.
2. **Reduced Need for Prompt Engineering:** They can reduce the need for complex prompt engineering, as they're better at interpreting the intent behind a user's request directly, enhancing user experience.
3. **Adaptability:** These models can quickly adapt to new tasks with minimal additional training, provided the tasks can be described through instructions.
4. **Ethical Considerations:** They may include improvements aimed at understanding and adhering to ethical guidelines or avoiding harmful outputs more effectively.

## Competition and Comparison

The competition between models like Llama 2 and Mistral-7B-Instruct (assuming it's a model similar in capability and purpose) would likely revolve around their ability to understand and generate human-like text, efficiency in processing, and adaptability to a wide range of tasks.

- **Performance vs. Precision:** Llama 2 might excel in broad performance across a wide variety of language tasks, while Mistral-7B-Instruct could offer more precision in executing tasks as per specific instructions.
- **Generalization vs. Instruction Following:** Llama 2 could be more generalized, suitable for a wide range of applications without much customization, whereas Mistral-7B-Instruct might be better for tasks that benefit from clear, directive instructions.
- **Customization and Ethical Considerations:** Both models might offer customization options, but Mistral-7B-Instruct might have an edge in tasks that require understanding of nuanced instructions or ethical considerations.

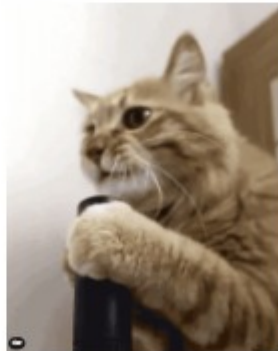
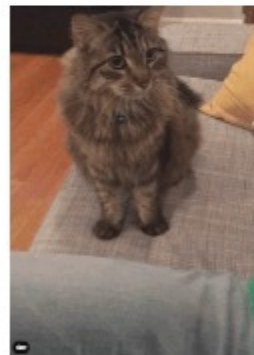
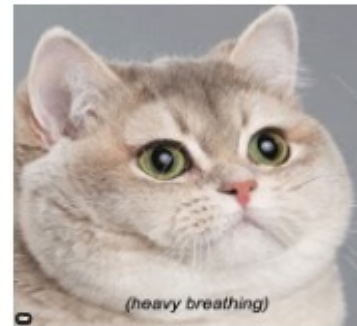
In practice, the choice between these two models would depend on the specific requirements of the application, including the nature of the tasks, the need for customization, and ethical considerations.

Llama for story  
Mistral for stable dif

hardware problem  
this option maybe be  
profitable only if we  
run them online



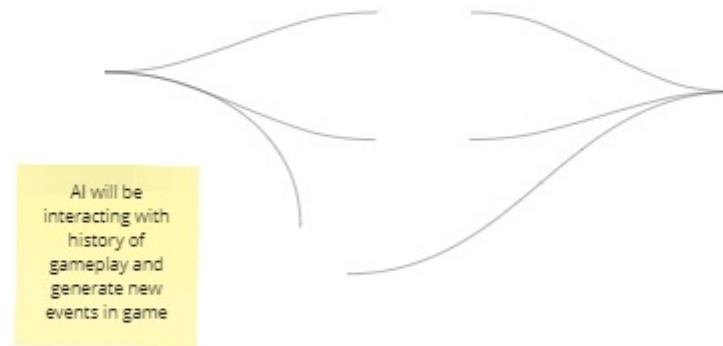
# The shrine of GODS!



# The idea

## Choose your own adventure

"Choose Your Own Adventure" (CYOA) is a popular series of interactive gamebooks where readers make choices that determine the outcome of the story. Each book presents the reader with a narrative and offers choices at certain points, leading to different plot developments and multiple possible endings.



Need to add AI Director to help and monitor inputs

1. **Multiple Endings:** One of the defining features of CYOA books is the presence of multiple possible endings. Depending on the choices made by the reader, the story can conclude in various ways, sometimes with the protagonist achieving success, facing failure, or experiencing unexpected twists. if STORY else never ending for Gamess
2. **Various Genres:** Choose Your Own Adventure books cover a wide range of genres, including fantasy, science fiction, mystery, horror, and historical fiction. This diversity allows readers to explore different settings, themes, and storytelling styles through their choices. \*\*space cowboy\*\*.
3. **AI Director:** The AI Director is a dynamic system designed to adjust the gameplay experience based on various factors, plays a crucial role in shaping the player experience by dynamically adjusting difficulty.

# GUI

**Customising COYA story** allows you to tailor the experience to your preference and the themes of your narrative

1) Themes and setting: Decide on the theme, setting and genre

Examples: science fiction, horror, mystery, space

2) Character creation: allow player to create their own characters or choose from preceding options.

Customise attributes such as names appearance, personality traits, skills and background story

3) Branching Path: design the branching paths and decision points in the story. Determine how player choice impact the narrative progression, character development and story outcome. Each decision should offer meaningful consequence and lead to different stories

4) interactive elements: incorporate interactive element such as puzzles, challenges, mini-games or dialogue tree to engage players and immerse them in the story.

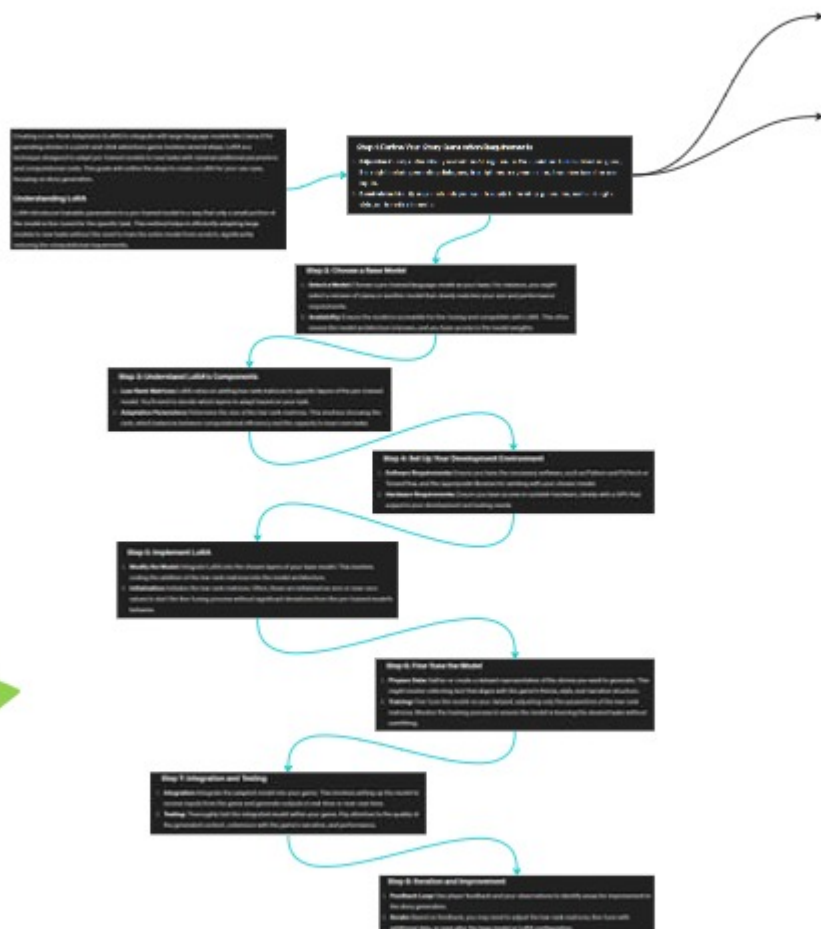
5) Multiple ending: provide multiple ending based on the player choices throughout the adventure. Each ending should feel satisfying and reflective of the player journey, offering different resolution and outcome based on choices.



?

?

# LoRA



Im waiting  
for meta  
approval



Big bang theory - Sheldon playing text adventures



ated 09-03-2024 @ 07:22 GMT+01:00