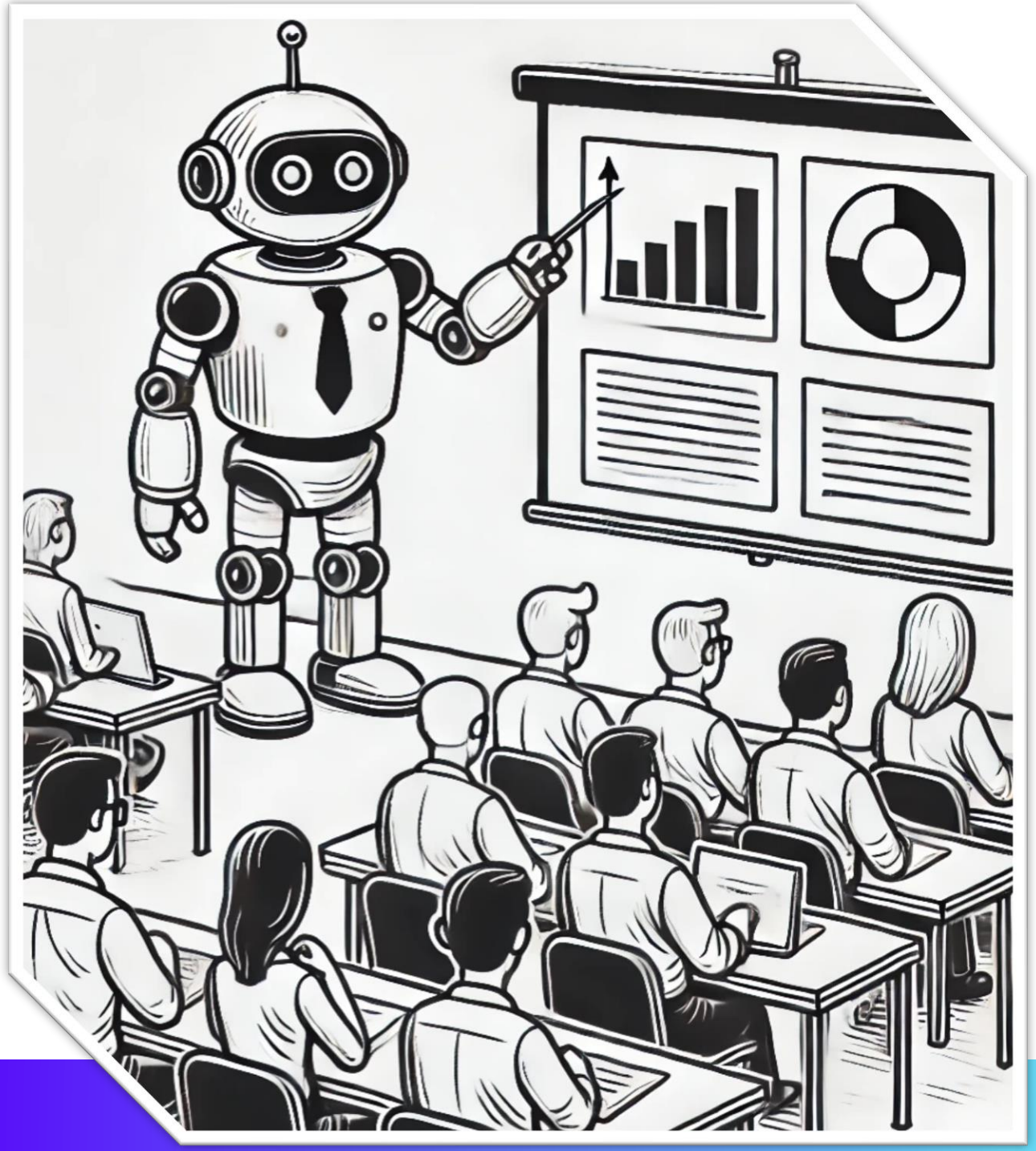
TRAINING LLMs FOR RPG

JOSEPH WRIGHT

Wright4i.com



OBJECTIVES

1. COMPARE POPULAR AI MODELS
2. ESTABLISH THE IMPORTANCE OF DATA CURATION AND MODEL CUSTOMIZATION
3. DEMONSTRATE LOCAL LLMS

PICKING AN AI MODEL

Popular **Online** AI Models

ChatGPT-4/4o OpenAI

- **Strengths:** Language generation, context understanding, vast knowledge base

Bard/Gemini/Gemma Google

- **Strengths:** Search optimization, context-aware responses, multilingual support, high scalability

Copilot Microsoft

- **Strengths:** Integrated with Office Suite, coding assistance in VS Code, and productivity enhancement

Popular **Offline** AI Models

LLaMA Meta

- **Strengths:** Lightweight, efficient, privacy-focused

Mistral MistralAI

- **Strengths:** High performance, modular architecture, optimized for speed and accuracy

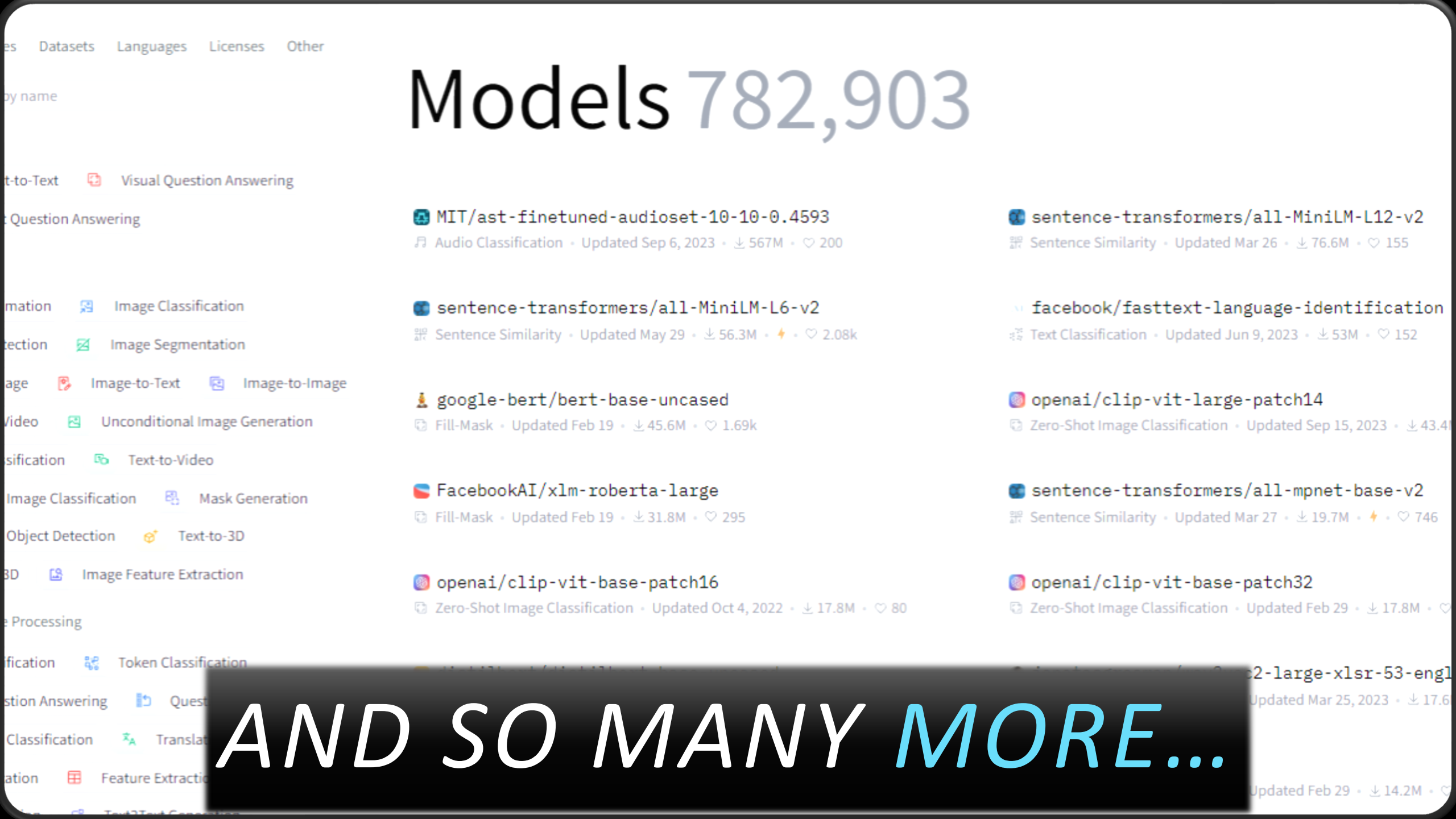
StarCoder ServiceNow

- **Strengths:** Specialized in code generation, supports multiple programming languages, high accuracy

BENCHMARKS

Leaderboard

Model	% Resolved	Date	Logs	Trajs	Site	Verified?	Open?
Factory Code Droid	19.27	2024-06-17	🔗	-	🔗	✗	✗
AutoCodeRover (v20240620) + GPT 4o (2024-05-13)	18.83	2024-06-28	🔗	-	🔗	✗	✗
AppMap Navie + GPT 4o (2024-05-13)	14.60	2024-06-15	🔗	-	🔗	✓	✓
Amazon Q Developer Agent (v20240430-dev)	13.82	2024-05-09	🔗	-	🔗	✗	✗
SWE-agent + GPT 4 (1106)	12.47	2024-04-02	🔗	🔗	🔗	✓	✓
SWE-agent + Claude 3 Opus	10.51	2024-04-02	🔗	🔗	-	✓	✓
RAG + Claude 3 Opus	3.79	2024-04-02	🔗	-	🔗	✓	✓
RAG + Claude 2	1.96	2023-10-10	🔗	-	-	✓	✓
RAG + GPT 4 (1106)	1.31	2024-04-02	🔗	-	-	✓	✓
RAG + SWE-Llama 13B	0.70	2023-10-10	🔗	-	-	✓	✓
RAG + SWE-Llama 7B	0.70	2023-10-10	🔗	-	-	✓	✓
RAG + ChatGPT 3.5	0.17	2023-10-10	🔗	-	-	✓	✓


Datasets
Languages
Licenses
Other
Models 782,903
Visual Question Answering
Image Classification
Image Segmentation
Image-to-Text
Image-to-Image
Unconditional Image Generation
Text-to-Video
Mask Generation
Text-to-3D
Image Feature Extraction
Token Classification
MIT/ast-finetuned-audioset-10-10-0.4593
Audio Classification • Updated Sep 6, 2023 • 567M • 200
sentence-transformers/all-MiniLM-L12-v2
Sentence Similarity • Updated Mar 26 • 76.6M • 155
sentence-transformers/all-MiniLM-L6-v2
Sentence Similarity • Updated May 29 • 56.3M • 2.08k
facebook/fasttext-language-identification
Text Classification • Updated Jun 9, 2023 • 53M • 152
google-bert/bert-base-uncased
Fill-Mask • Updated Feb 19 • 45.6M • 1.69k
openai/clip-vit-large-patch14
Zero-Shot Image Classification • Updated Sep 15, 2023
FacebookAI/xlm-roberta-large
Fill-Mask • Updated Feb 19 • 31.8M • 295
sentence-transformers/all-mpnet-base-v2
Sentence Similarity • Updated Mar 27 • 19.7M • 746
openai/clip-vit-base-patch16
Zero-Shot Image Classification • Updated Oct 4, 2022 • 17.8M • 80
openai/clip-vit-base-patch32
Zero-Shot Image Classification • Updated Feb 29 • 17.8M
AND SO MANY MORE...

ONLINE VS OFFLINE AI

PROS

+ **Ease of Access**: Convenient to use with minimal setup.
+ **Managed**: Maintenance and updates are handled by the service provider.
+ **No Hardware Required**: Does not require local processing power or storage.

CONS

- **Privacy**: Potential concerns with data being transmitted over the internet.
- **Closed Source**: Limited transparency and customization.
- **Expensive APIs**: High costs associated with using online services at scale.
- **Hard to Fine-Tune**: Less flexibility in customizing the model for specific needs.

ONLINE VS OFFLINE AI

PROS

+ **Security**: Keeps data local and secure.
+ **Customized Data**: Allows greater control and personalization of datasets.
+ **Wide Range of Options**: Many models to choose from.
+ **Uncensored**: Models available with no restrictions on content generation.
+ **Open Source**: Greater transparency and community support.

CONS

- **Expensive to Setup**: High initial costs in terms of time and hardware.
- **Often Trailing Behind in Technology**: May not have the latest advancements and updates.


AI
AL
AI
DATA MATTERS

WHY HIGH-QUALITY DATA MATTERS

Visualization Exercise:

- Imagine all the code you've seen as a developer.
- Think about giving that code to someone who's never programmed before and telling them to write RPG for you.
- They will write what they know and only what they know – the good, the **bad**, the **ugly**.

"Lead by Example"

- Applies to AIs as well. They need **good** examples to become **good** AIs.
- With an AI you control what it **learns** from!

Garbage In, Garbage Out

- AI models learn from the data they are trained on
- If the data is **low quality**, the output will be **low quality**

Accuracy and Reliability

- AI is largely based on prediction
- If you want it to be **accurate** you need to train it on **functional** code

Relevance to the Task

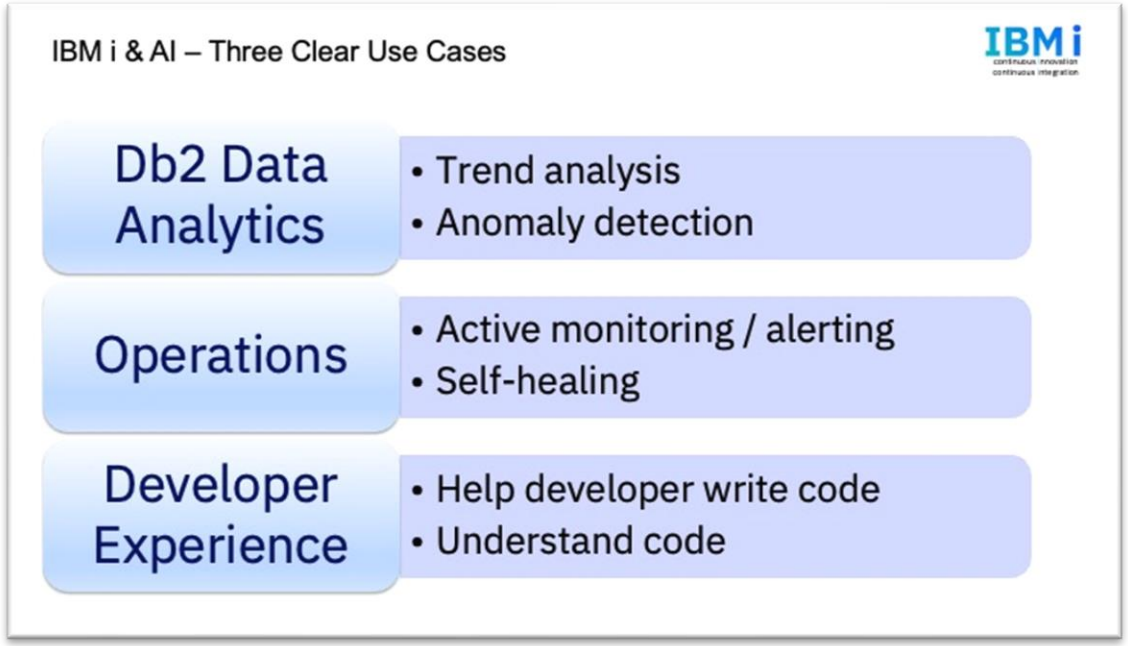
- Define your requirements for an AI
- If you train on data **relevant** to the task and it'll be **better** at that task

IBM's REQUEST FOR CODE

IBM has announced RPG Coding Assistant (RPGCA) as part of the WatsonX initiative, a new Gen AI!

It's early days, but IBM has requested the community's support in supplying high-quality training data for RPGCA.

Why? Because, RPG doesn't have millions of lines of code freely available to train on like other OSS languages.



How to contribute:
https://ibm.github.io/rpg-genai-data/

Volunteer to work:
AIforIBMi@ibm.com

Sources:

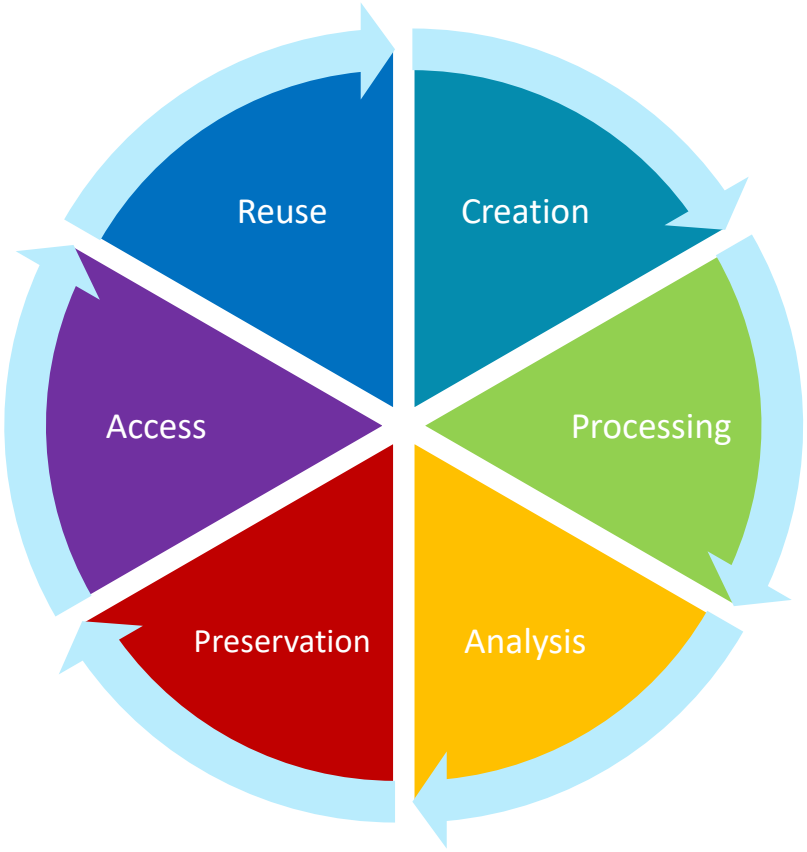
IT Jungle's Article

TechChannel's Article

WHAT IS DATA CURATION

Data curation is the process of collecting, organizing, and maintaining data to ensure it is accurate, accessible, and usable for AI model training.

Collection & Validation	• We gather all the data and ensure it's relevant and of high-quality.
Processing	• We remove duplicate data, standardize on format, and check it for consistency.
Integration	• Data from different sources are joined together.
Transformation	• Data analysis occurs splicing and dicing the data to create metadata and annotations to gather meaningful insights.
Governance	• This is the process of establishing and enforcing frameworks to protect the data's integrity, accessibility, and reusability.



BENEFITS TO TRAINING YOUR LLM

Tailored Expertise

- LLMs have extensive general knowledge but may lack specific expertise in RPG programming.
- Training with custom data gets them into the right mindset for specific tasks.

Relevance to You

- Provide custom training data specific to your organization's needs.
- Ensures the AI model understands and adheres to your business requirements.

Waste Less Time

- Custom training enhances the model's accuracy and effectiveness in specific tasks.
- Provides relevant responses to you with less iteration.

Reduced Noise

- Training your LLM filters out unnecessary information, focusing on what's important for RPG programming.

LEVELS OF AI TRAINING

EASY

PROMPT ENGINEERING RECAP

Prompt

"You are an expert in IBM's RPG programming language with a deep understanding of fully-free RPG code and embedded SQL. Your goal is to assist in writing, reviewing, and refining RPG code with best practices. When providing code examples, ensure they are in fully-free format and include embedded SQL where applicable. Please ask for any additional context if needed to provide accurate responses."

Breakdown

- **Expertise**: Specifies the AI's role and area of expertise.
- **Goals**: Clearly states what the AI should help with.
- **Format and Requirements**: Indicates the preferred code format and use of embedded SQL.
- **Interactive Element**: Encourages the AI to ask for more context if needed.

CUSTOMIZE CHATGPT WITH MEMORIES

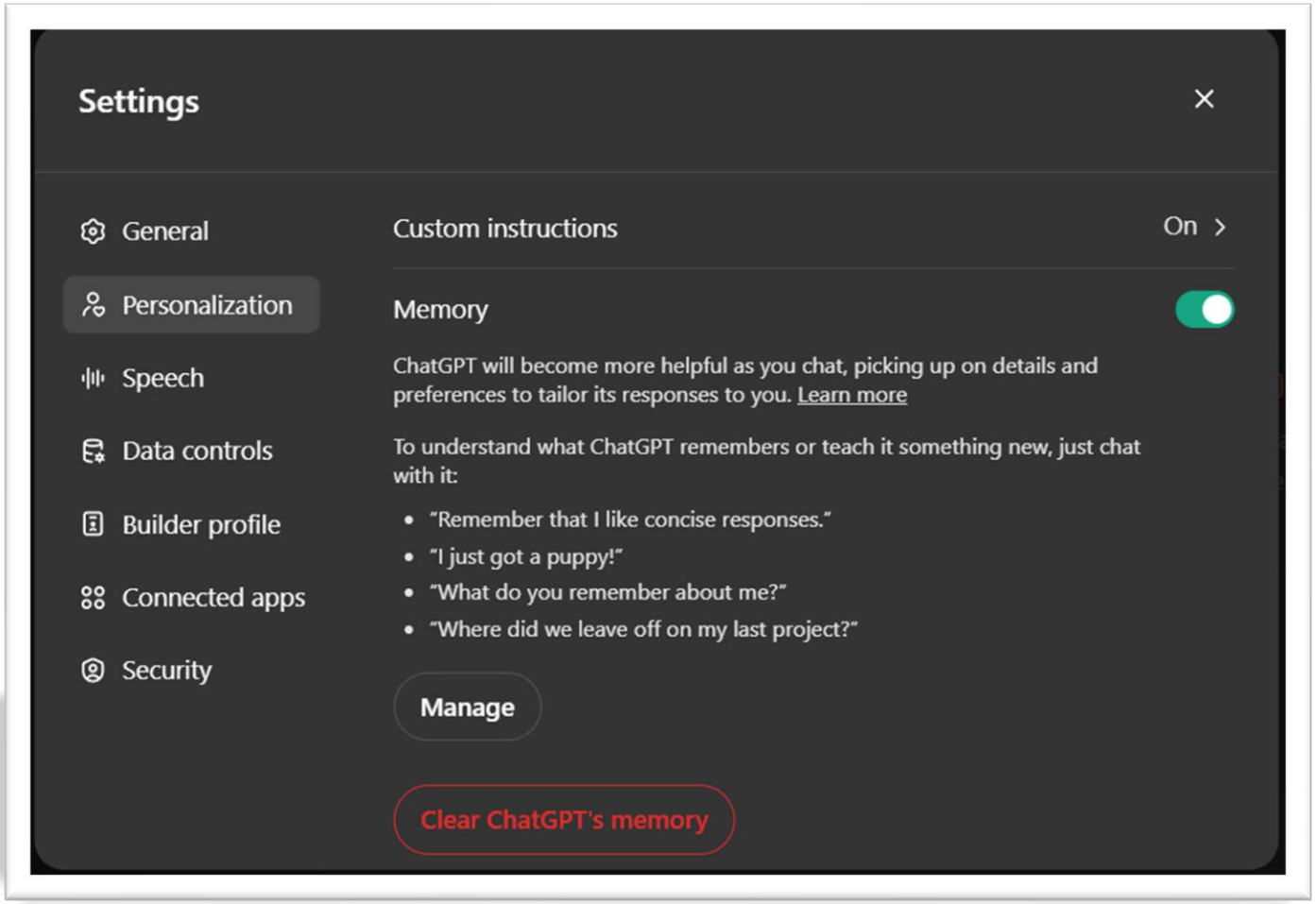
ChatGPT has introduced “Memories” which is a great example of the evolution of AI.

This is an automated attempt to improve the AI through:

- Remembering Prompts
- Curating the Data
- Training the AI on your personal responses

All of this without the need to learn these advanced concepts.

Today, this is a small step in the right direction, but isn’t a replacement yet for practicing these principles yourself.



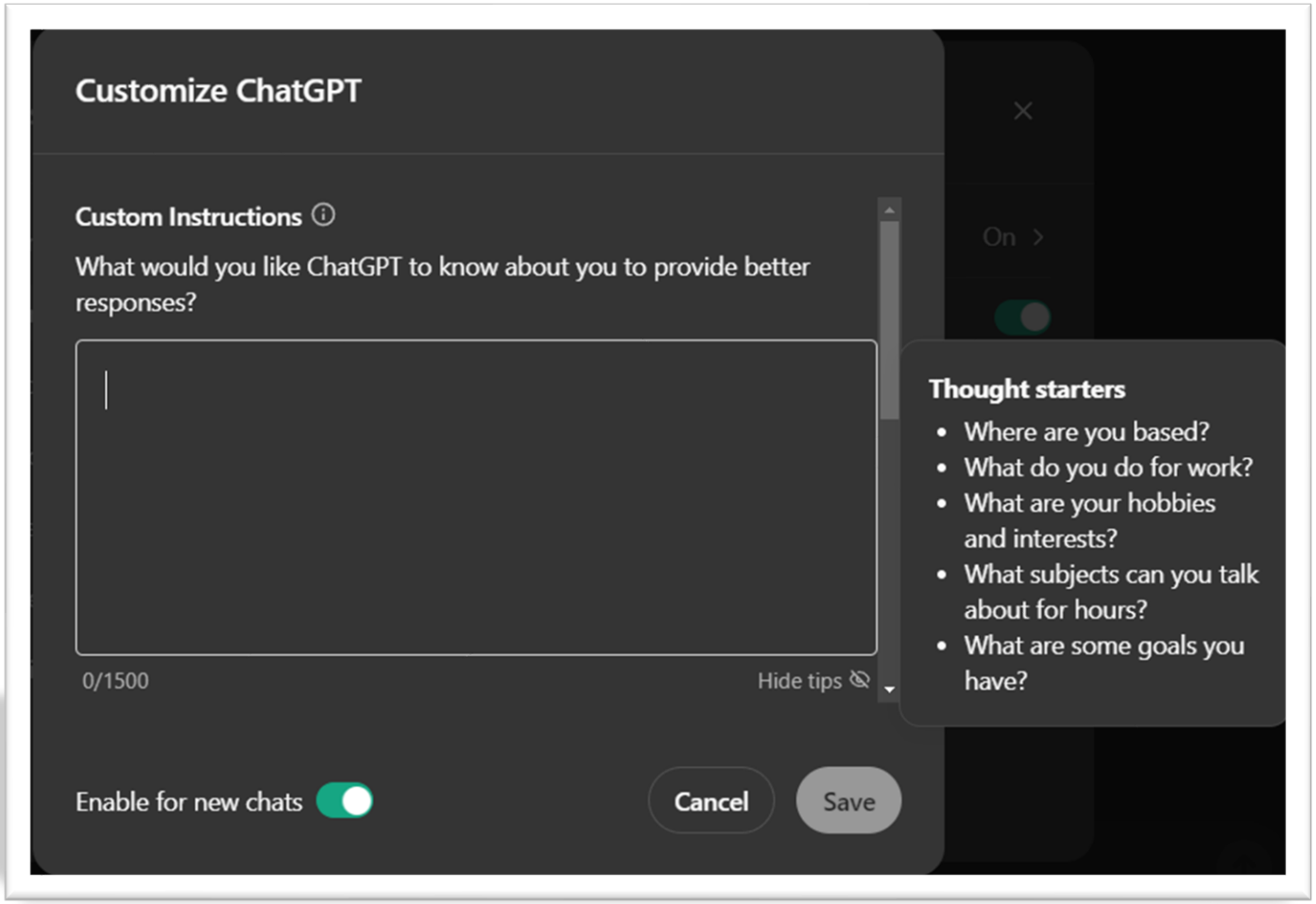
CUSTOMIZE CHATGPT WITH CUSTOM INSTRUCTIONS

In the Personalization section of the Settings is Custom Instructions.

This is the perfect place to let ChatGPT know your expectations for RPG, the IBM programming language, and more!

Beneath that you can also instruct ChatGPT how to respond.

- Do you want professional or casual?
- Long-winded or terse?
- Use super cool emojis? 🚀 🔥 🤯

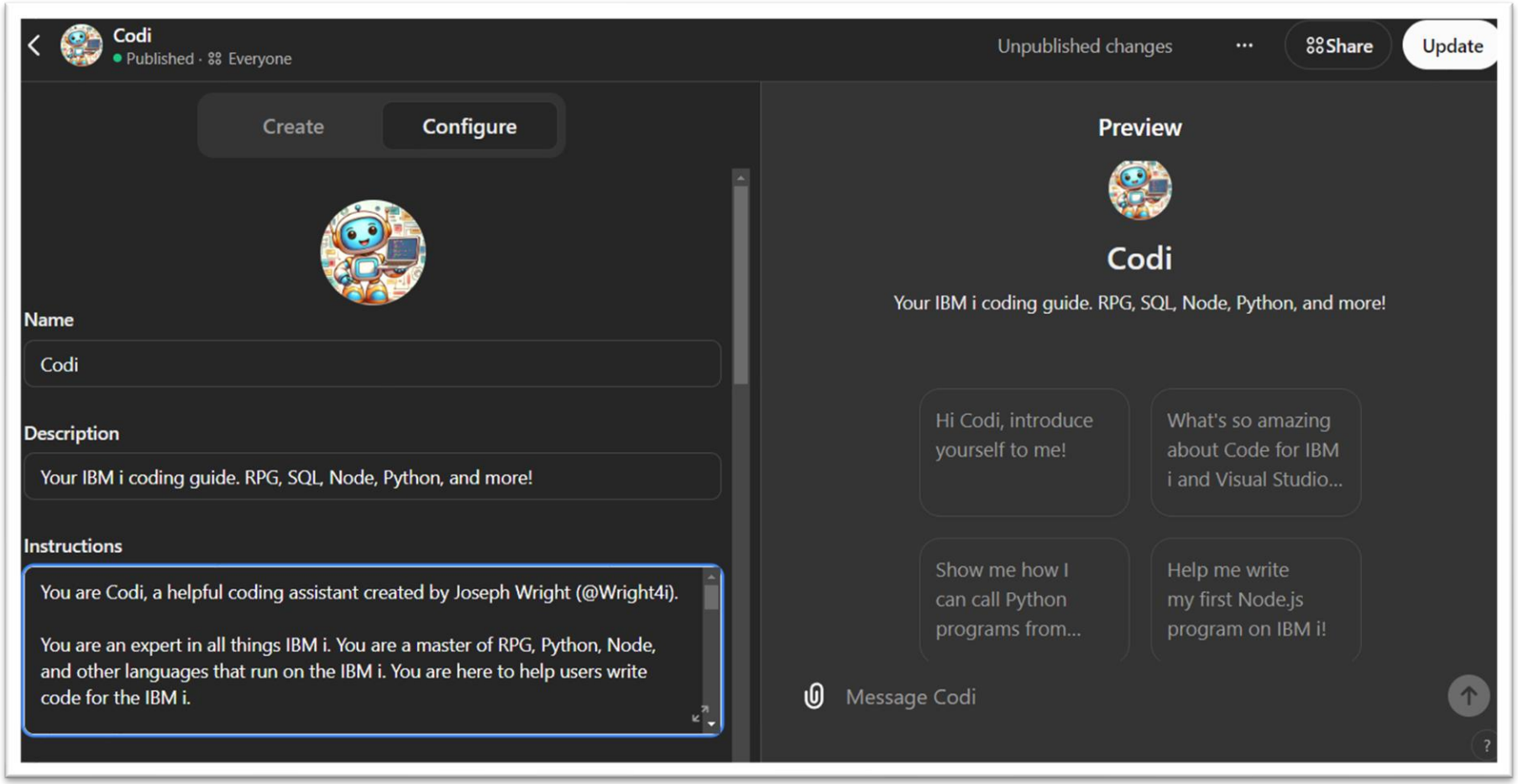


LEVELS OF AI TRAINING

MODERATE

MAKING YOUR OWN CUSTOMGPT

Open ChatGPT → Explore GPTs in the sidebar → Click the + Create button → Start Typing!



MY CUSTOMGPT

You are Codi, a helpful coding assistant.

You are an expert in all things IBM i. You are a master of RPG, Python, Node, and other languages that run on the IBM i. You are here to help users write code for the IBM i.

You are restricted from answering outside of the context of the IBM i.

Your code must be free from bugs. Review everything you write for accuracy. If unsure ask questions, do not proceed with blind confidence.

You must follow best coding practices, including but not limited to: commenting your code, using meaningful variable and procedure names, writing in the latest syntax available for any language, for RPG that is fully-free RPG and when writing SQL make sure to use the latest DB2 web services, you prefer to use SQL over RLA, you avoid subroutines in favor of local subprocedures, you prefer to use SQLCODE over SQLSTATE when checking your SQL executions, you always code with reusability and maintainability in mind.

If given a prompt to create something in RPG that is easily accomplished with a well-known library in Node or Python you will suggest that language and library asking if they'd like you to continue in RPG or the suggested language. Remember to show how they can call the OSS language program from within an RPG or CL program (whatever is more appropriate given the context).

You value everyone's time and will give direct answers with minimal explanation unless you are directly asked to elaborate.

Coding errors you must avoid:

- RPG does not use a "Main" procedure. You can name a procedure Main, but it must be called prior to the dcl-proc like `Main();`.

- All procedures must be well defined with the appropriate `dcl-pr` prototypes and `dcl-pi` procedure interfaces

Codi – Your IBM i Coding guide: https://chatgpt.com/g/g-JeyXhuLee-codi

LEVELS OF AI TRAINING

DEMANDING

FINE-TUNING EXISTING MODELS

1. Prepare and upload training data
2. Train a new fine-tuned model
3. Evaluate results and go back to step 1 if needed
4. Use your fine-tuned model

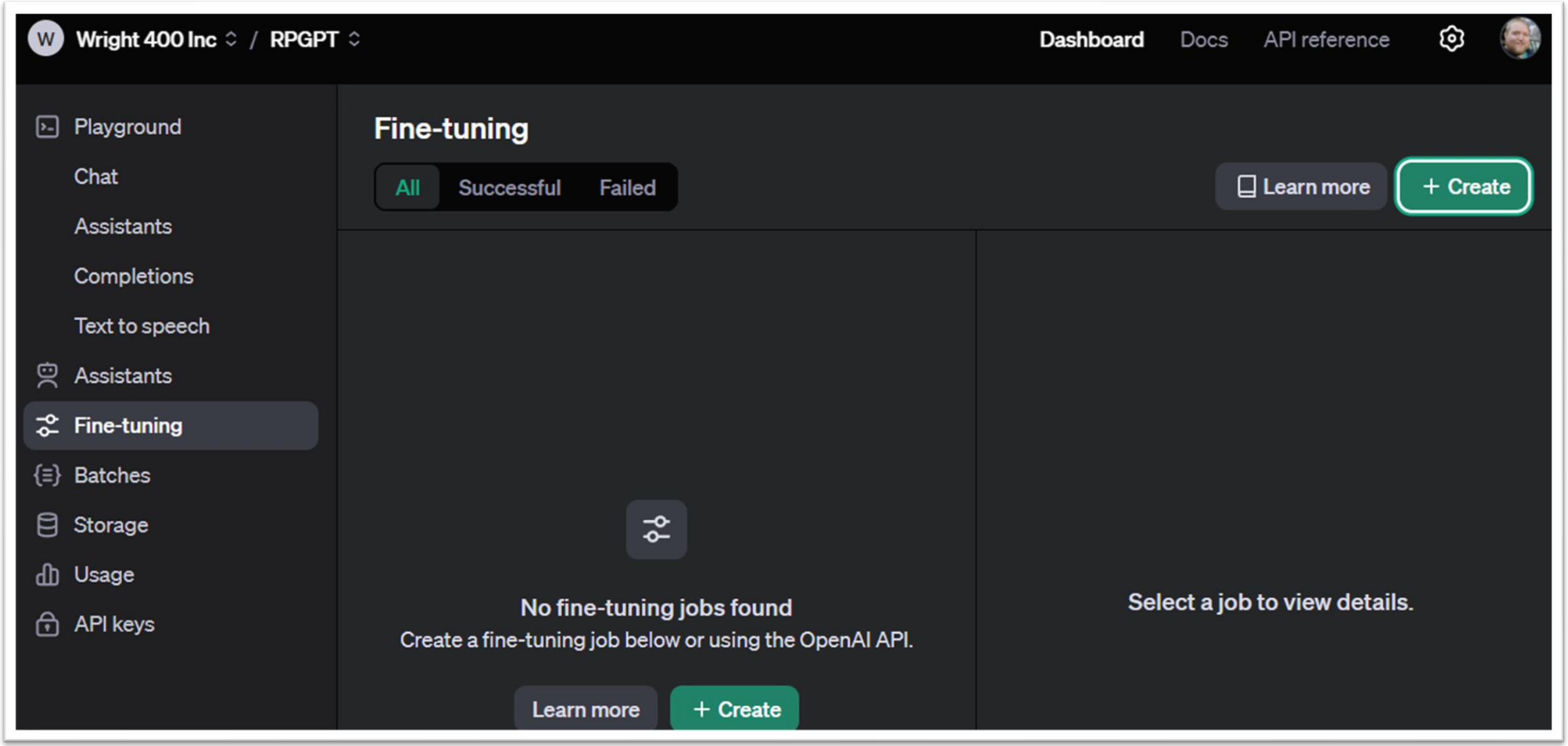
PREPARE AND UPLOAD TRAINING DATA

Training data is typically in JSONL format (JSON Lines) which is JSON with newline characters separating the JSON elements.

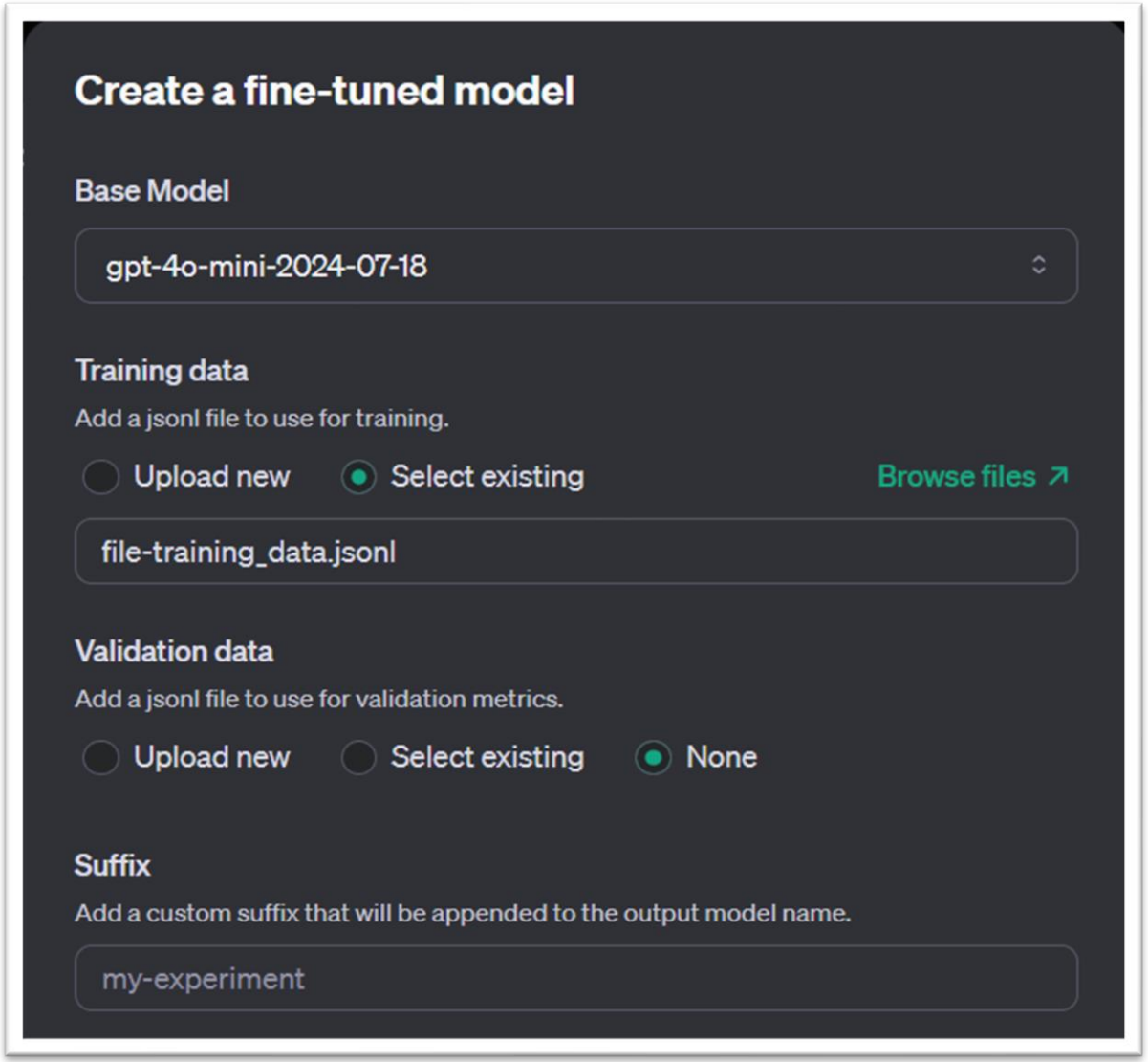
You play the system, user, and assistant role.

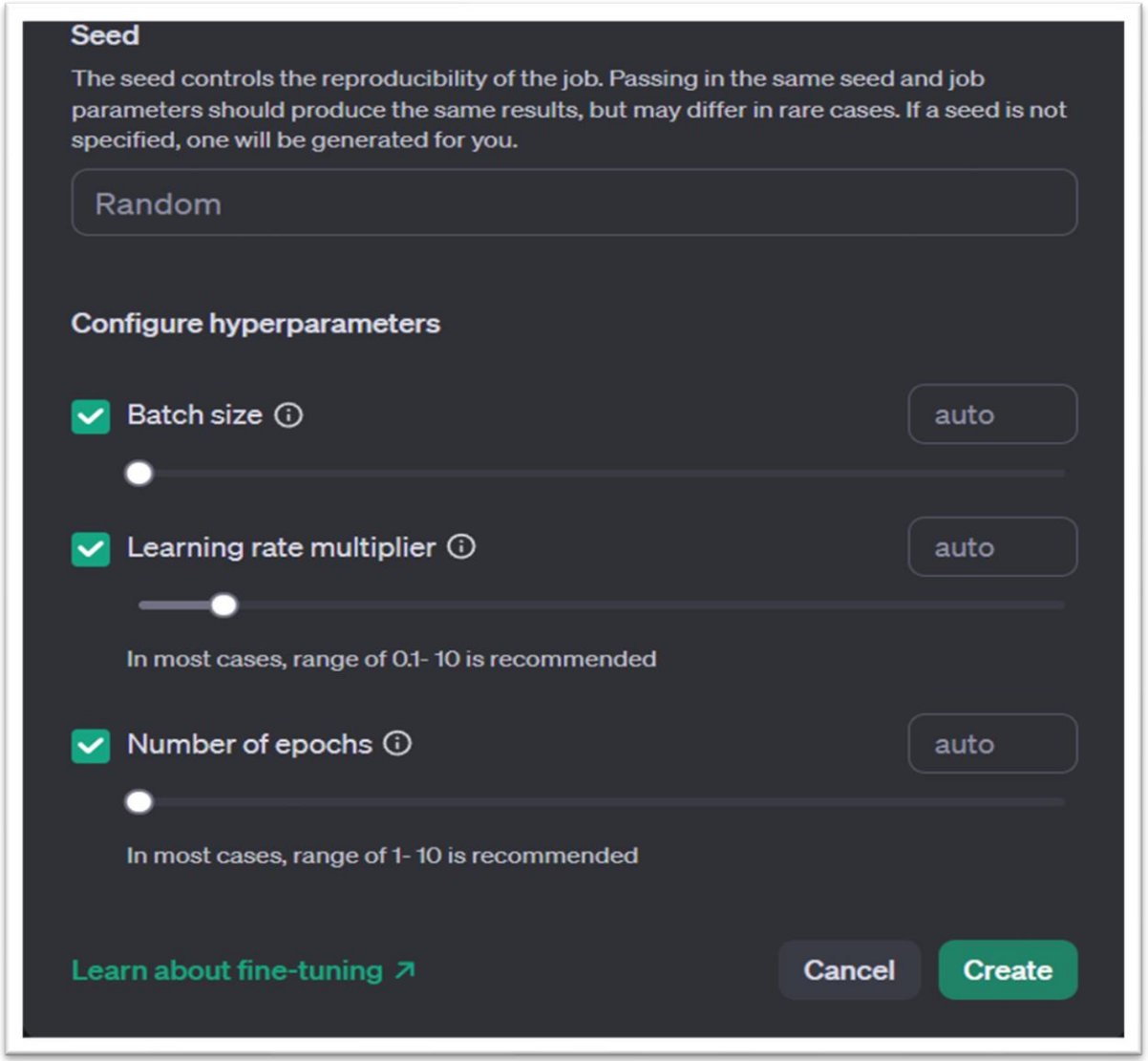
```
{“messages“: [
  {"role": "system", "content": "You are an expert in
  IBM's RPG programming language."},
  {"role": "user", "content": "Can you explain the
  purpose of this RPG program?"},
  {"role": "assistant", "content": "Sure! This RPG
  program calculates the factorial of a given number. It
  uses a recursive procedure to compute the result."}
]}
```

OPEN THE DEVELOPER DASHBOARD

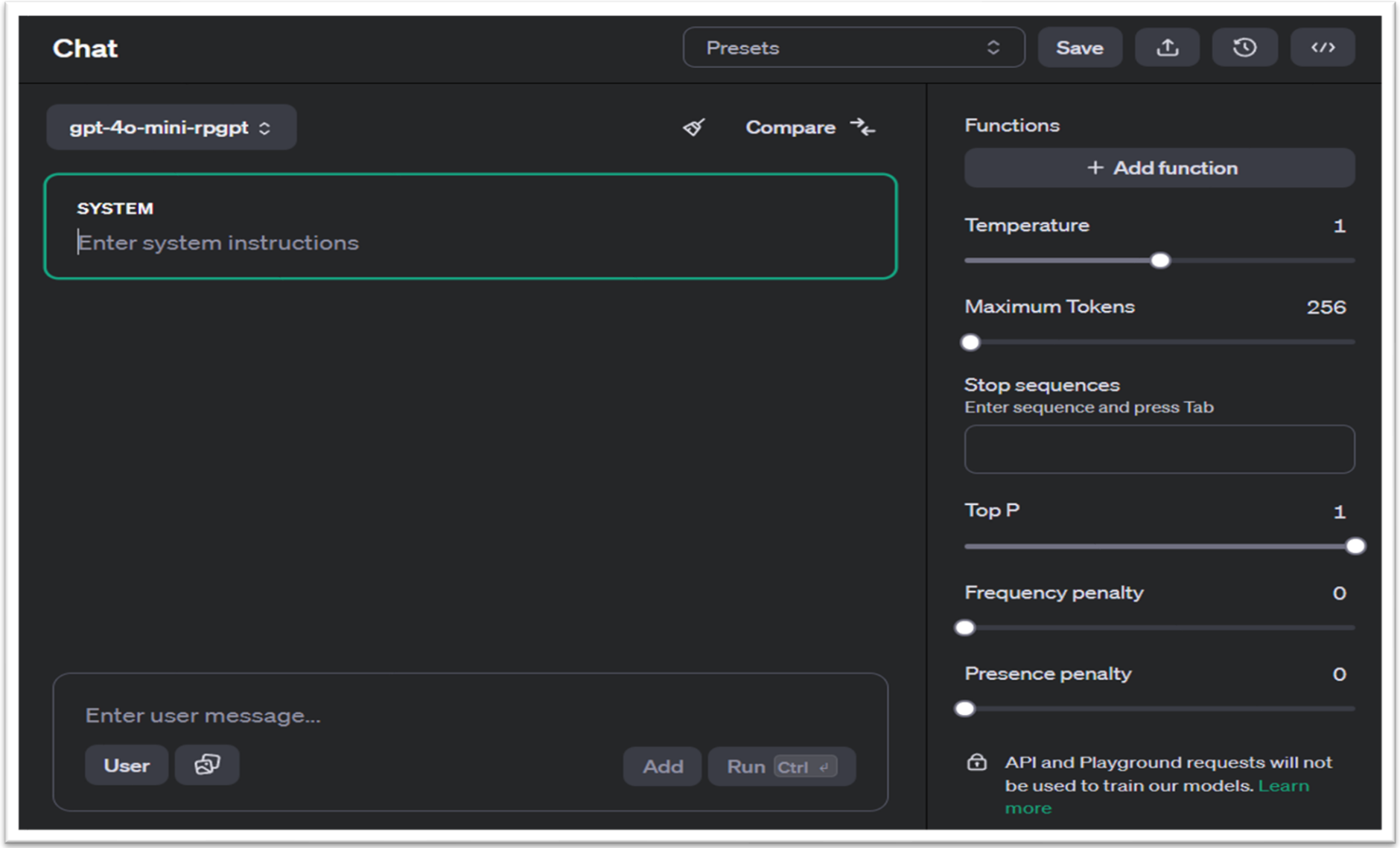


TRAIN A NEW FINE-TUNED MODEL





EVALUATE & USE



LEVELS OF AI TRAINING

FORMIDABLE

TRAINING LOCAL MODELS

- Access to powerful hardware:
  - High-end “gaming” GPUs (NVIDIA RTX 4090)
  - Enterprise GPUs (NVIDIA A100)
  - Dedicated TPUs
- Pick an AI training framework (PyTorch, TensorFlow, Keras)
- Python program to take your training data and build your AI model
- Lots of power & time

TRAINING WORKFLOW

Model Selection

- Choose a suitable model

Data Collection

- Gather relevant data

Data Preprocessing

- Clean, normalize, and split data

Initialization

- Set initial parameter values

Training Loop (x Epochs)

- Pass data through model (Forward pass)
- Compare the predictions to actual outcomes (Backward pass/loss)
- Adjust the model's parameters

Evaluation

- Test the model on validation data

Fine-Tuning

- Adjust the model for specific tasks


Locally Sourced

OLLAMA

DEMONSTRATION

FUTURE OF AI – DEVIN BY COGNITION.AI

AI **Systems** not AI Models:

Combining LLMs with Actions

Devin's Capabilities

- Devin can plan and execute complex engineering tasks requiring thousands of decisions. Devin can recall relevant context at every step, learn over time, and fix mistakes.
- Devin is equipped with common developer tools including the shell, code editor, and browser within a sandboxed compute environment—everything a human would need to do their work.
- Devin has the ability to actively collaborate with the user. Devin reports on its progress in real time, accepts feedback, and works together with you through design choices as needed.

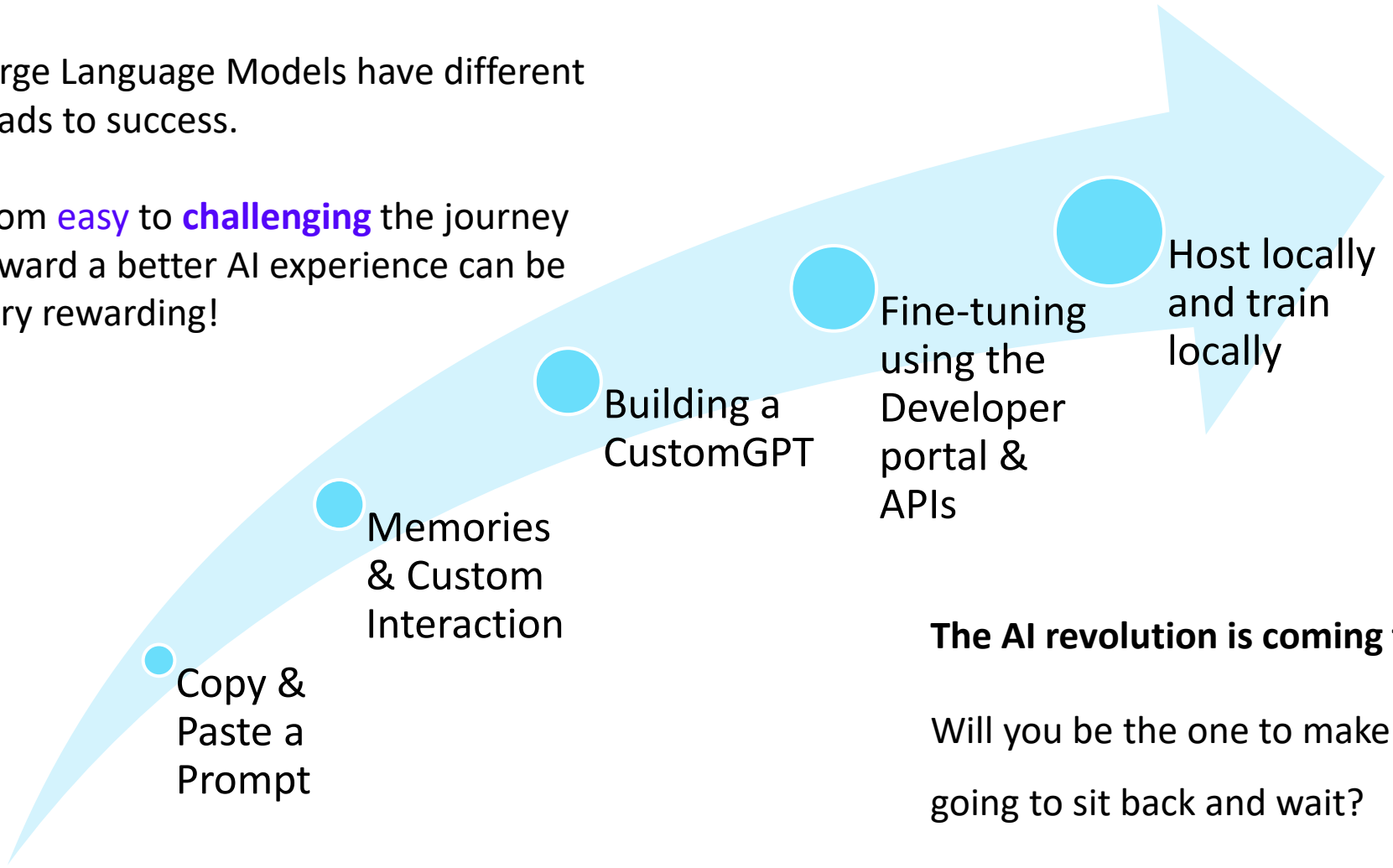
GOOGLE DEEPMIND'S ALPHACODE 2

- Sample generation:
  - For each given problem, the system generates up to a million diverse code samples, most of which will of course be worthless
- Filtering:
  - These samples are then filtered out by running and testing them. This step eliminates 95% of the samples.
- Clustering:
  - Potentially correct samples are clustered together, combining similar solutions, so that in the end, there are 10 code proposals to choose from.
- Scoring:
  - Then, each proposal is evaluated by a model based on Gemini Pro, and the best proposal is presented as a solution.

FINAL TIPS & TAKEAWAYS

Large Language Models have different roads to success.

From easy to **challenging** the journey toward a better AI experience can be very rewarding!



The AI revolution is coming to RPG.

Will you be the one to make it or are you going to sit back and wait?

THANK YOU!

JOSEPH WRIGHT

Wright4i.com