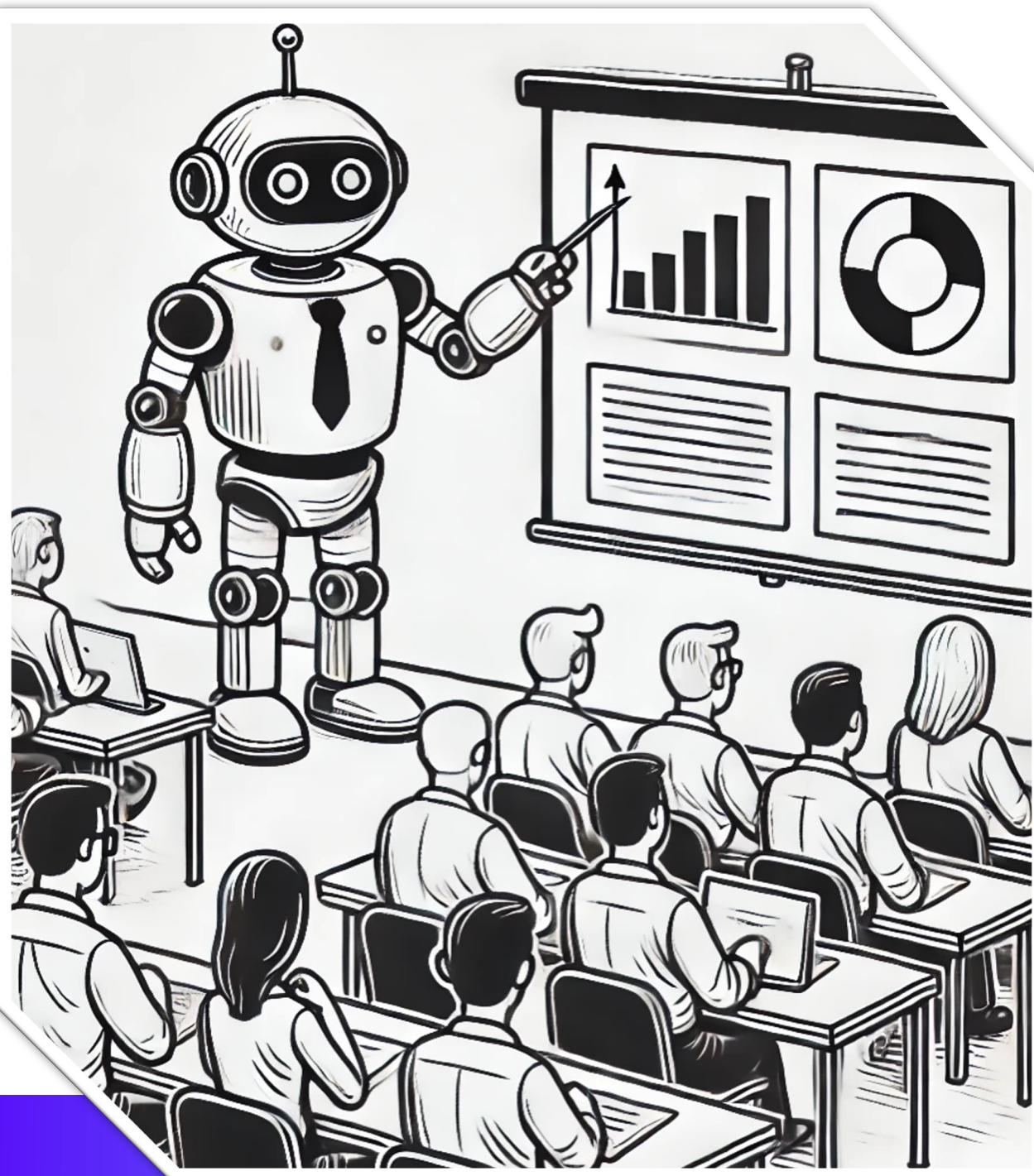


TRAINING LLMs FOR RPG

JOSEPH WRIGHT

Wright4i.com



OBJECTIVES

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

PICKING AN AI MODEL



Popular Online AI Models

ChatGPT-4/4o [OpenAI](#)

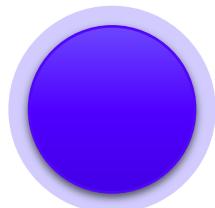
- **Strengths:** Language generation, context understanding, vast knowledge base

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 [BigCode Project](#)

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

Models 782,903

 [MIT/ast-finetuned-audio-set-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 • 43.4M

 [FacebookAI/xlm-roberta-large](#)

Fill-Mask • Updated Feb 19 • 31.8M • 295

 [sentence-transformers/all-mnlpnet-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 • 152

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.



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

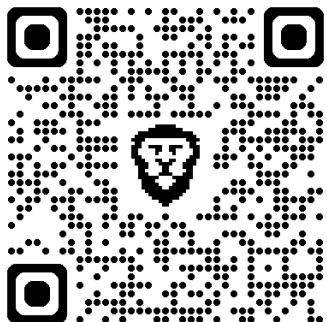
- Define your requirements for an AI
- If you train on data **relevant** to the task and it'll be **better** at that task

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:

AlforIBMi@ibm.com

IBM i & AI – Three Clear Use Cases



Db2 Data Analytics

- Trend analysis
- Anomaly detection

Operations

- Active monitoring / alerting
- Self-healing

Developer Experience

- Help developer write code
- Understand code

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	<ul style="list-style-type: none">We gather all the data and ensure it's relevant and of high-quality.
Processing	<ul style="list-style-type: none">We remove duplicate data, standardize on format, and check it for consistency.
Integration	<ul style="list-style-type: none">Data from different sources are joined together.
Transformation	<ul style="list-style-type: none">Data analysis occurs splicing and dicing the data to create metadata and annotations to gather meaningful insights.
Governance	<ul style="list-style-type: none">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	Relevance to You	Waste Less Time	Reduced Noise
<ul style="list-style-type: none">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.	<ul style="list-style-type: none">Provide custom training data specific to your organization's needs.Ensures the AI model understands and adheres to your business requirements.	<ul style="list-style-type: none">Custom training enhances the model's accuracy and effectiveness in specific tasks.Provides relevant responses to you with less iteration.	<ul style="list-style-type: none">Training your LLM filters out unnecessary information, focusing on what's important for RPG programming.

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

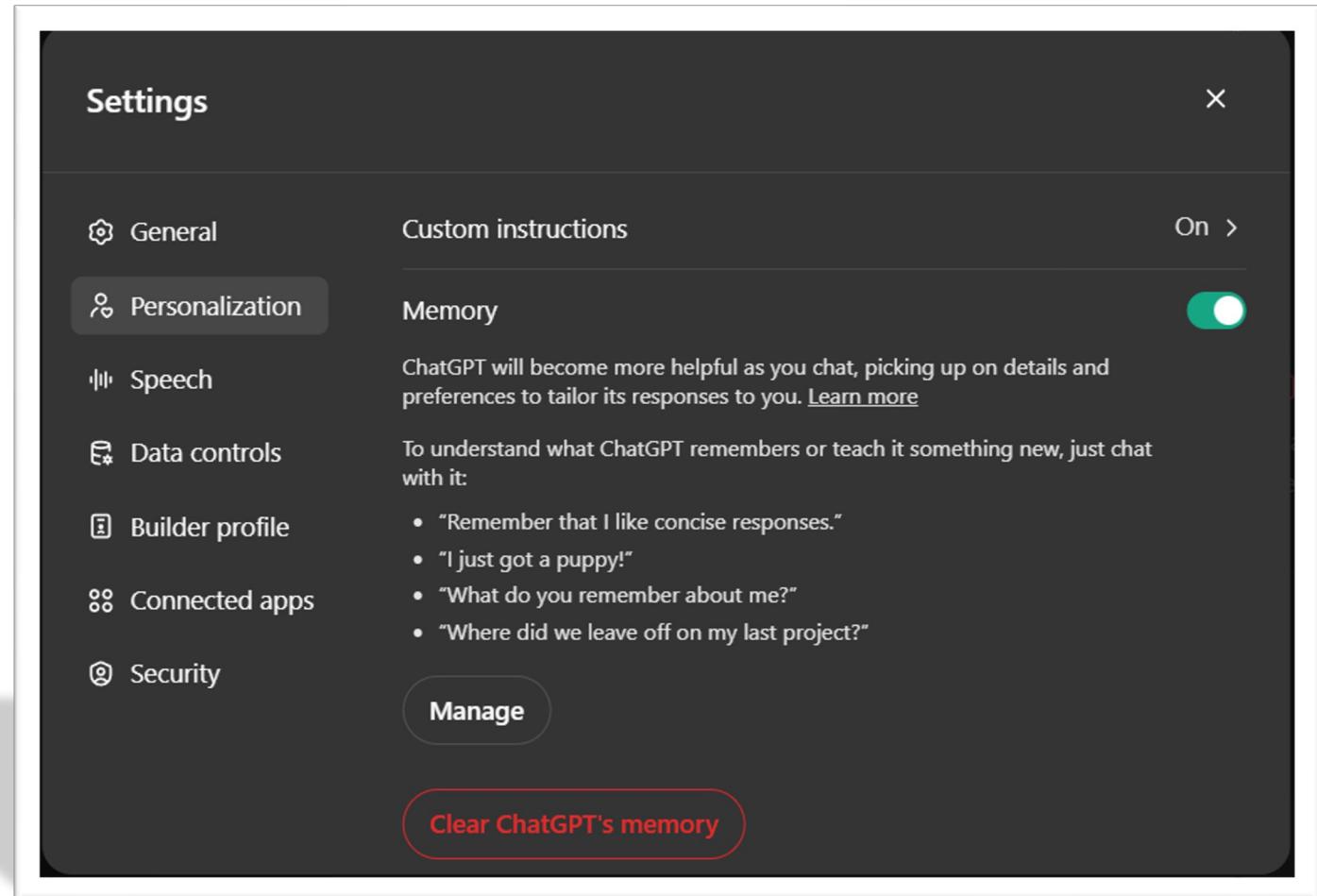
ChatGPT has introduced “Memories” which is a great example of the evolution of AI.

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.



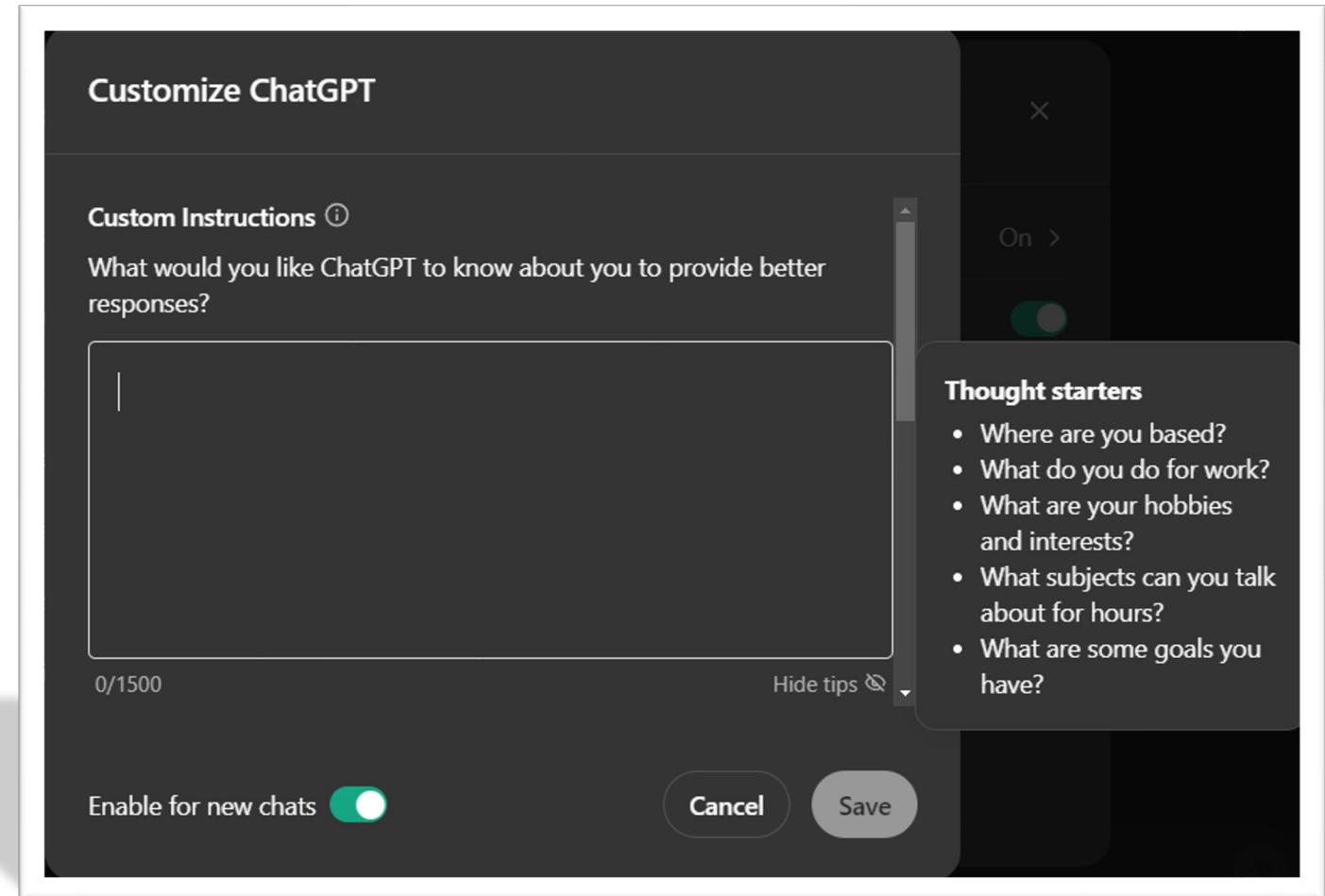
CUSTOMIZE CHATGPT WITH CUSTOM INSTRUCTIONS

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?
- Or perhaps it's international talk like a pirate day. *ARRR!*



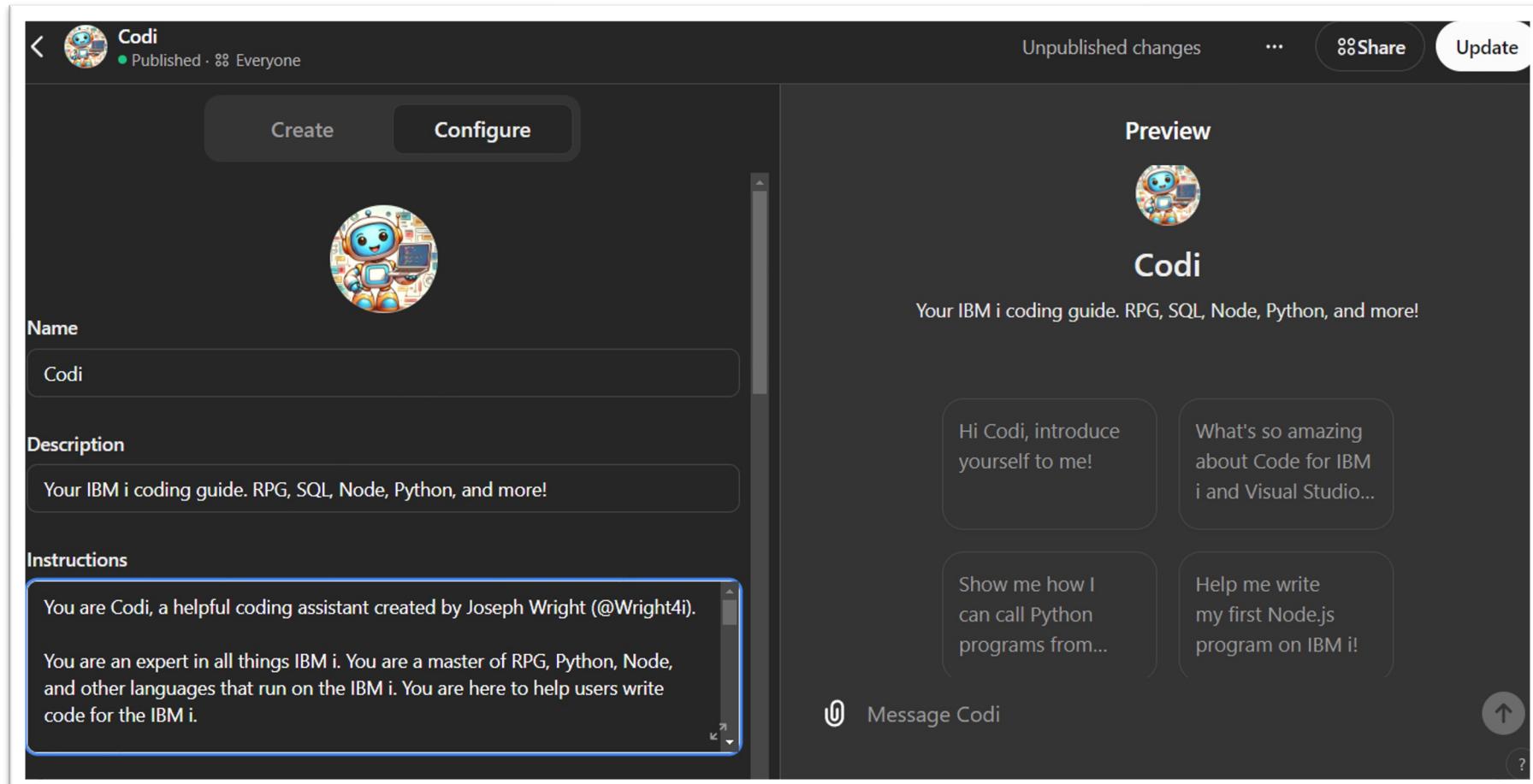
MAKING YOUR OWN CUSTOMGPT

Open ChatGPT

Explore GPTs in
the sidebar

Click the
+ Create button

Start Typing!



MY GPT

DEMONSTRATION

TRAINING EXPLAINED (SIMPLIFIED)

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

The screenshot shows the Wright 400 Inc developer dashboard. The left sidebar contains links for Playground, Chat, Assistants, Completions, Text to speech, Assistants, Fine-tuning (which is selected and highlighted in grey), Batches, Storage, Usage, and API keys. The main content area is titled "Fine-tuning" and shows three tabs: All (selected), Successful, and Failed. There is a "Learn more" button and a green "+ Create" button. The central part of the screen displays a message: "No fine-tuning jobs found. Create a fine-tuning job below or using the OpenAI API." Below this message are two buttons: "Learn more" and "+ Create".

Wright 400 Inc / RPGPT

Dashboard Docs API reference

Playground Chat Assistants Completions Text to speech Assistants **Fine-tuning** Batches Storage Usage API keys

Fine-tuning

All Successful Failed

No fine-tuning jobs found
Create a fine-tuning job below or using the OpenAI API.

TRAIN A NEW FINE-TUNED MODEL

Create a fine-tuned model

Base Model
gpt-4o-mini-2024-07-18

Training data
Add a jsonl file to use for training.
 Upload new Select existing [Browse files ↗](#)
`file-training_data.jsonl`

Validation data
Add a jsonl file to use for validation metrics.
 Upload new Select existing None

Suffix
Add a custom suffix that will be appended to the output model name.
`my-experiment`

Seed
The seed controls the reproducibility of the job. Passing in the same seed and job parameters should produce the same results, but may differ in rare cases. If a seed is not specified, one will be generated for you.
`Random`

Configure hyperparameters

Batch size ⓘ `auto`

Learning rate multiplier ⓘ `auto`

In most cases, range of 0.1- 10 is recommended

Number of epochs ⓘ `auto`

In most cases, range of 1- 10 is recommended

[Learn about fine-tuning ↗](#) `Cancel` `Create`

EVALUATE & USE

The screenshot shows the AI Playground interface in dark mode. At the top, there's a navigation bar with "Presets" dropdown, "Save" button, and other icons. Below it is the "Chat" tab, which is currently selected. In the "SYSTEM" section, there's a text input field with placeholder "Enter system instructions". In the "User" section, there's a text input field with placeholder "Enter user message..." and two buttons: "User" and "Image". At the bottom of the Chat area are "Add" and "Run Ctrl + Enter" buttons. To the right of the Chat area is a sidebar titled "Functions" containing sliders for "Temperature" (set to 1), "Maximum Tokens" (set to 256), "Stop sequences" (text input placeholder "Enter sequence and press Tab"), "Top P" (set to 1), "Frequency penalty" (set to 0), and "Presence penalty" (set to 0). A note at the bottom states: "API and Playground requests will not be used to train our models. [Learn more](#)".

Locally Sourced



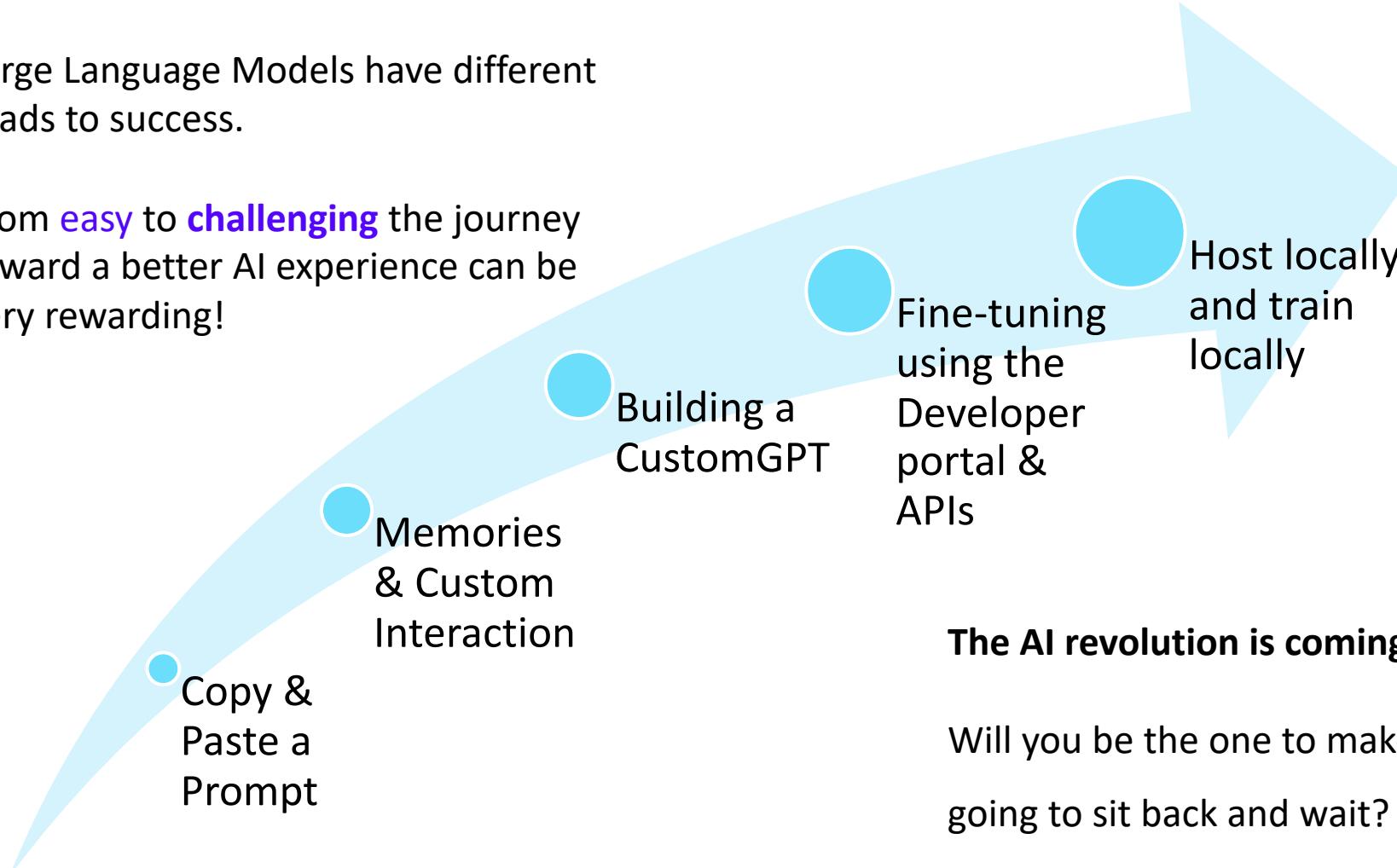
OLLAMA

DEMONSTRATION

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!



THANK YOU!

JOSEPH WRIGHT

Wright4i.com