Understanding GPTs

Concepts and Educational Applications

August 23, 2024

Christopher Runyon

NBME®

Disclaimer

The information presented in this document/presentation is solely for informational purposes and does not necessarily reflect the official stance or views of NBME. The opinions expressed herein are those of the author/presenter and may not represent the views or policies of NBME. While every effort has been made to ensure the accuracy and reliability of the information provided, NBME makes no representations or warranties of any kind, express or implied, about the completeness, accuracy, reliability, suitability, or availability with respect to the content contained herein. Any reliance you place on such information is therefore strictly at your own risk.

Learning Objectives

- Understand the fundamentals of GPTs and their operational mechanisms.
- Understand different prompt engineering approaches.
- Explore applications of GenAI (in medical education).
    - New study!

Microwaves

- Been around for decades
- Technology is established
- Best use cases are well known

Generative AI

- Been around for a few years
- Technology is rapidly advancing
- Best use cases are ***not*** well known

What is a GPT?

- Why do I need to know about the technical aspects of GenAI?

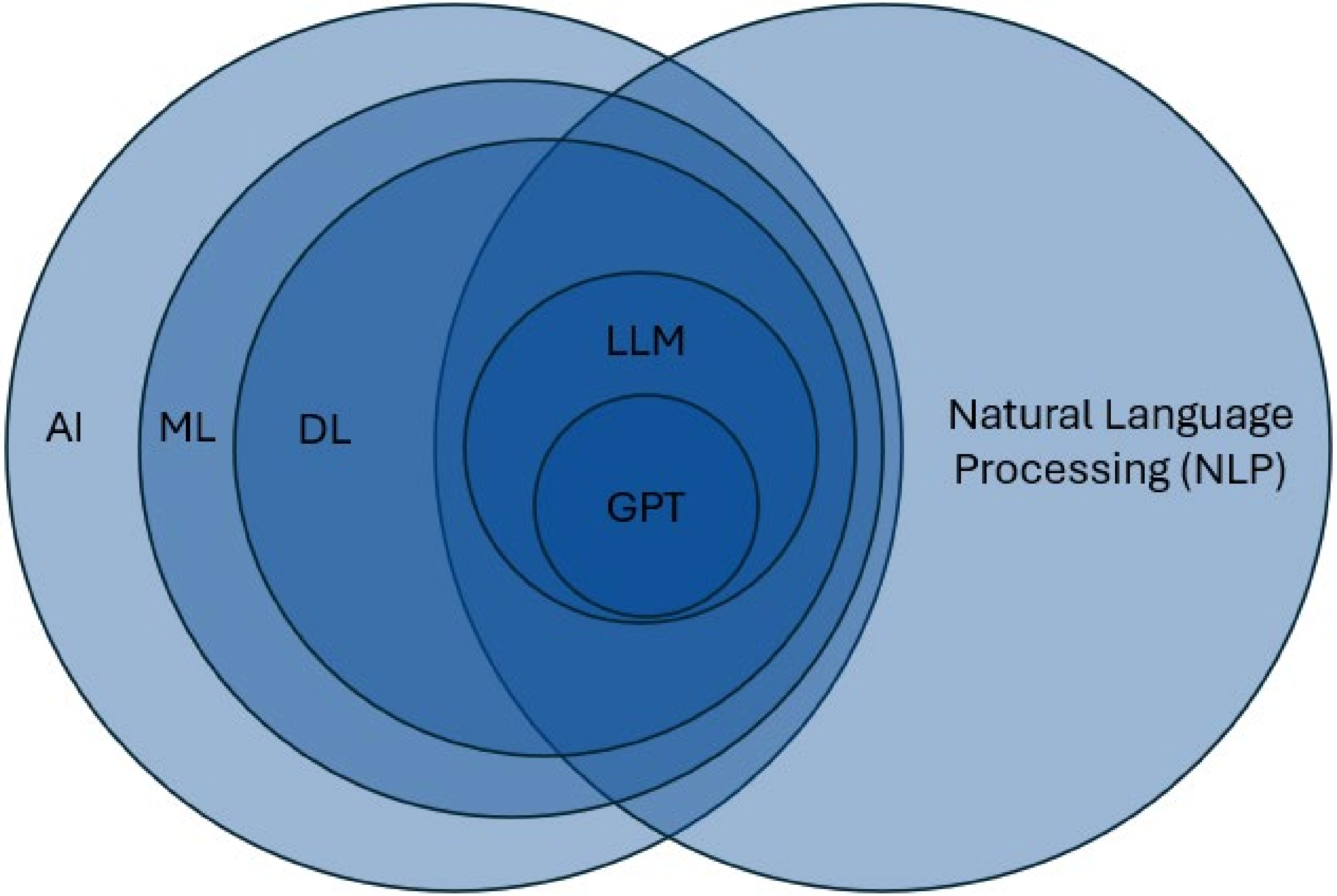
≈ relationship between foundational sciences and clinical practice

≈ relationship between matrix/linear algebra and statistical model results

- Helps with understanding why prompt engineering is important and how you may need to adjust your prompts. (LO #2)

- Helps you understand how to maximize the utility of GenAI models. (LO #3)

What is a GPT?



AI = Artificial Intelligence

ML = Machine Learning

DL = Deep Learning

LLM = Large Language Models

GPT = Generative Pretrained Transformers

What is a GPT?

- G = Generative
  - refers to the model's ability to create new content

- P = Pretrained
  - model has undergone an initial extensive training phase on a vast corpus of text data; equips the model with a broad ‘understanding’ of language and its nuances

- T = Transformer
  - aspect of the model architecture (not discussed today)
  - attention layer (will be discussed today)

Variations in GPTs

- ChatGPT is a GPT model developed by OpenAI
  - Arguably the first model with an easy, accessible interface
  - Several versions: 3.5, 4 **(4o just released May 13)**
- Anthropic: Claude **(Claude 3.5 release June 20!)**
- Google: Gemini (formerly Bard), PaLM 2
- Meta: LLaMA (**L**arge **La**nguage Model **M**eta **A**I) – 3.1 recently released
- …plus many models that can be downloaded and further customized (go look at huggingface.co)

Helpful ideas for thinking about GPTs

- Underlying principle: distributional semantics / co-occurrence
  - Words that occur in the same contexts tend to have similar meanings. (Harris, 1954)

“We found a little, hairy sleeping behind the tree.”

- Generative model: ***<u>always</u>*** provides an answer
  - No guarantee the response is correct; “hallucinations”
  - Extrapolation problem
- Not deterministic; variation is part of its beauty
- A very, *very* fancy text calculator

In less than 15 words, please describe why pizza is so good

- Pizza combines savory, cheesy, and versatile toppings on a delicious, soft, or crispy crust.
- Pizza combines savory cheese, flavorful toppings, and a crispy crust for a satisfyingly delicious experience.
- Pizza's delicious due to its savory combination of cheese, sauce, and various toppings.
- Pizza combines savory flavors, melty cheese, and a satisfying crust in every bite. Irresistibly delicious!
- Pizza's delicious due to its perfect blend of cheese, sauce, and various toppings.

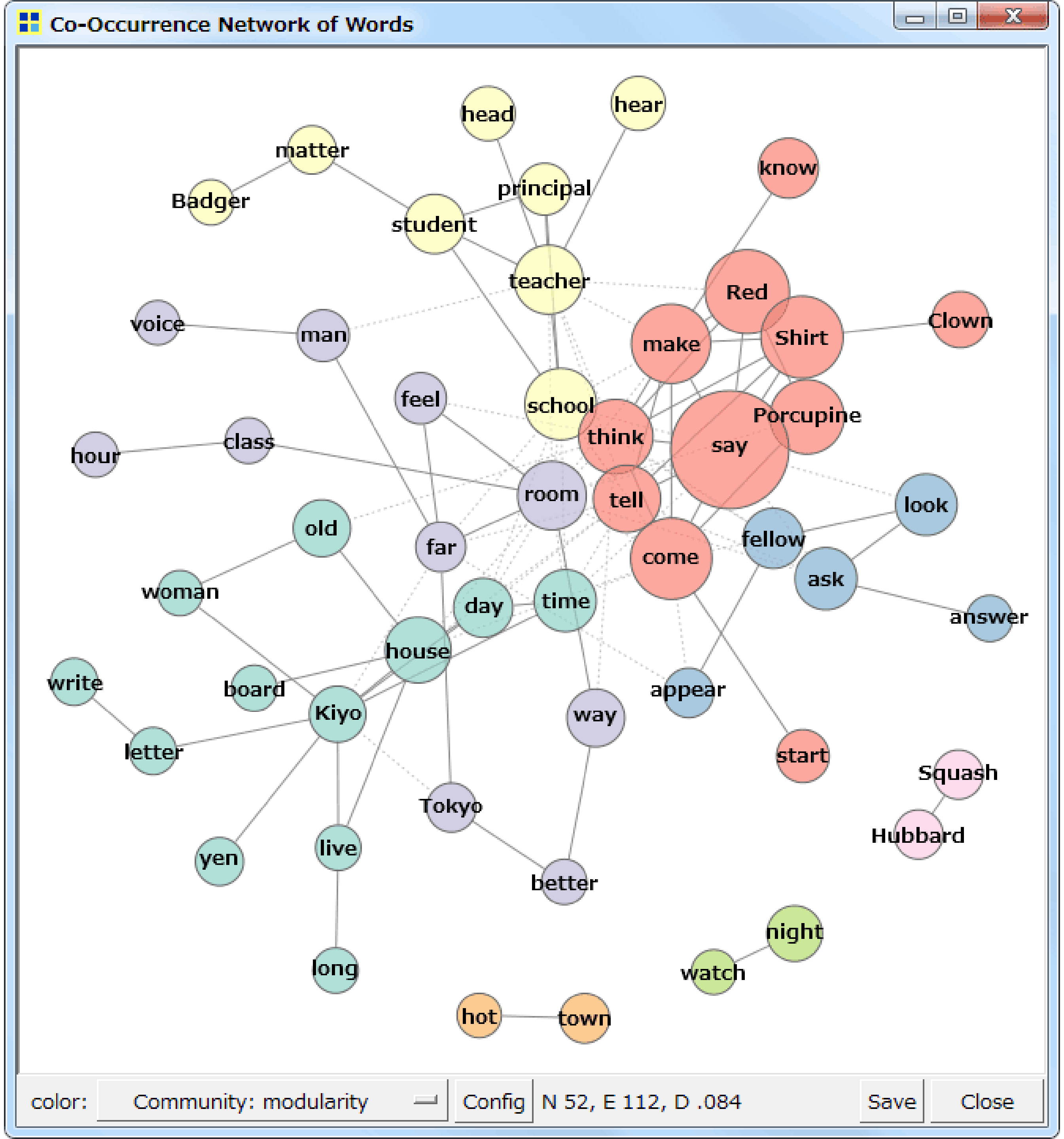
Model parameters

- Temperature
  - “What sampling temperature to use, between 0 and 2. Higher values like 0.8 will make the output more random, while lower values like 0.2 will make it more focused and deterministic.” (for OpenAI’s GPT-4o)
  - Does not guarantee that result will be exactly reproduced.

- top_p (nucleus sampling)
  - “An alternative to sampling with temperature, called nucleus sampling, where the model considers the results of the tokens with top_p probability mass. So 0.1 means only the tokens comprising the top 10% probability mass are considered.”

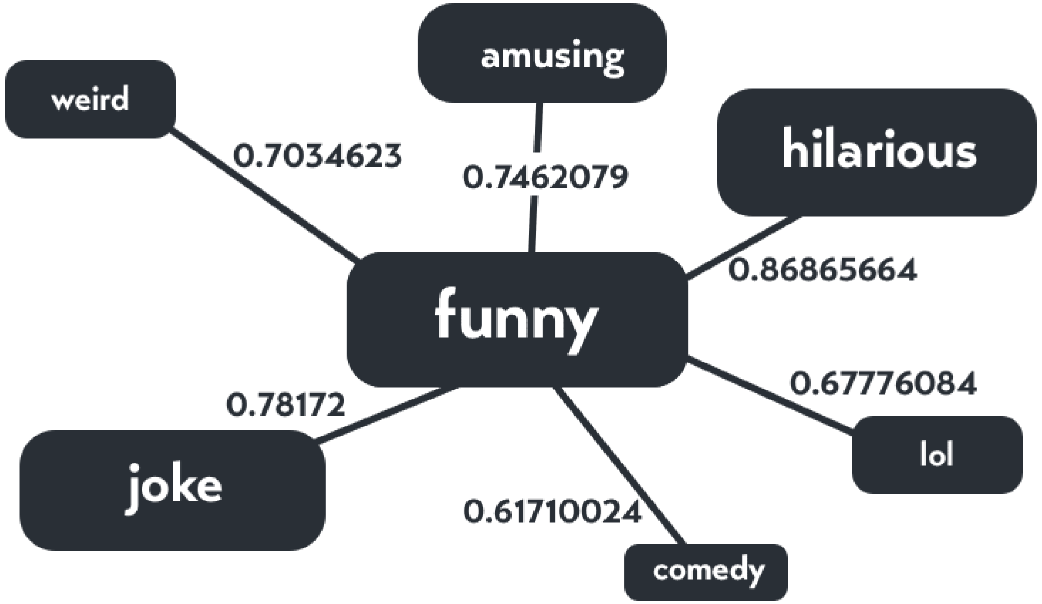
“It is recommended to alter one or the other, but not both.”

GPT 4 (not 4o): Estimated ~1.76 Trillion Parameters



color: Community: modularity

N 52, E 112, D .084



Embeddings / “Concepts”

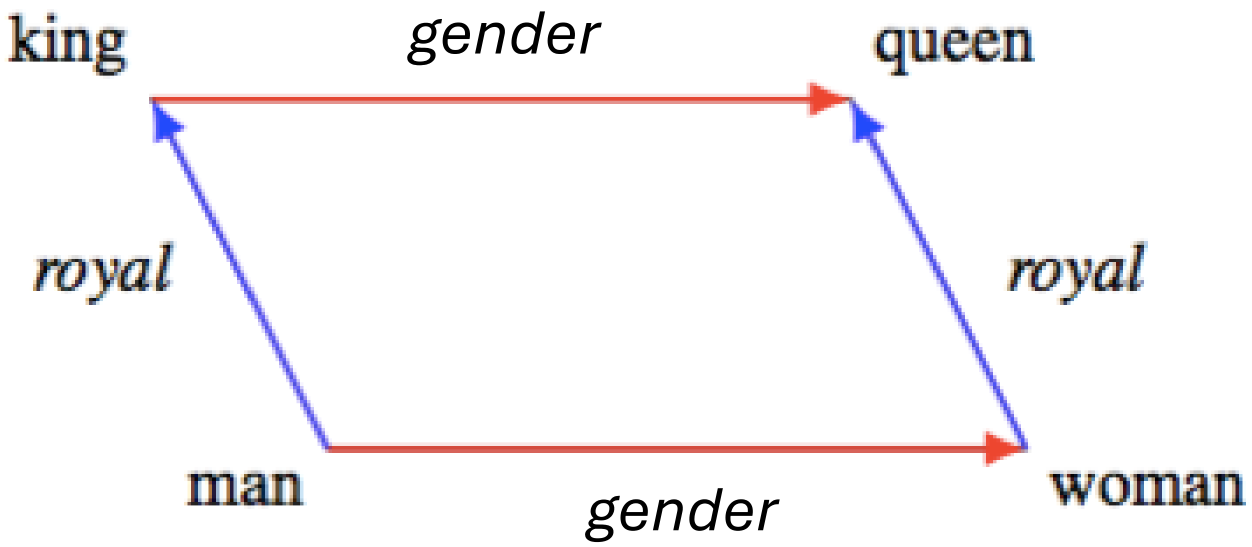
- **Capture Meaning:** Word embeddings represent words as vectors in a continuous space, capturing semantic meaning and relationships between words.

- **Contextual Similarity:** Words with similar meanings or contexts have embeddings that are closer together in this space.

- **Pretrained Knowledge:** These embeddings are learned during pretraining, giving the model an understanding of language before any specific input is processed.

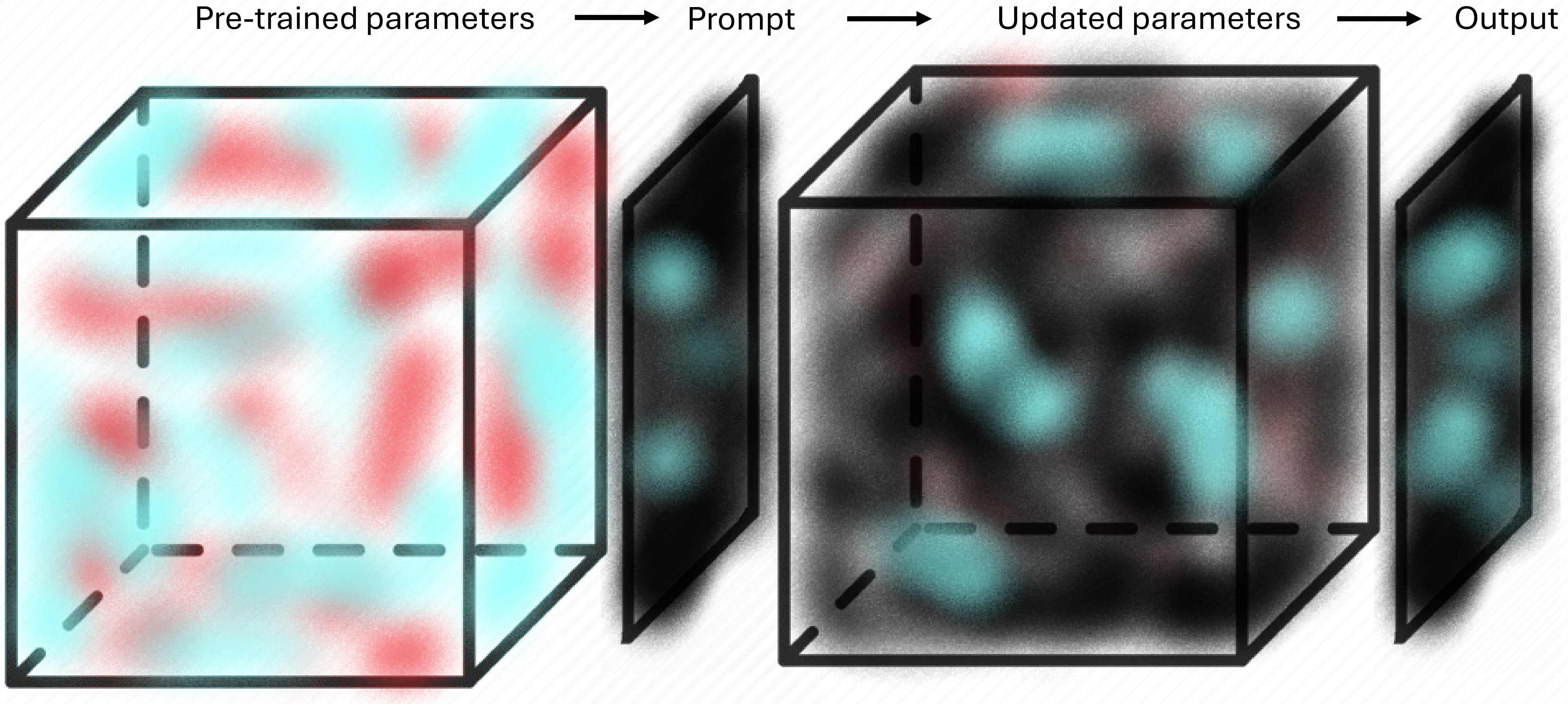
Embeddings / “Concepts”

- GPTs (and many other NLP models) largely do not work at the level of words. The usually work with embeddings.



GPT 4 (not 4o):
Estimated ~1.76 Trillion
Parameters

- Will talk about concepts or words instead of embeddings in this presentation.



Pre-trained parameters → Prompt → Updated parameters → Output

Prompt Engineering

- There are ways to interact with GPTs that can result in more useful responses
  - Provide more context for what you would like the GPT to output
  - Help activate more relevant connections in the network

- We’re going to focus on two general strategies for prompting
  - Prompt formulas
  - Prompt chaining / chain-of-thought prompting

Prompt formulas

- Structured guidelines or templates used to generate responses from language models like GPT-4o
  - Help shape the context, style, and content of the output

ROLE TASK GOAL

[I am in my mid 40s and I am becoming more interested in film.] [I would like to learn more about the intersection of romance and sci-fi genres.] [Please suggest 10 popular movies in romance/sci-fi genre and provide details on why these movies are thought to be important or interesting.]

ROLE TASK GOAL

[I am organizing a conference], and I would like help [developing icebreaker interactive sessions] for conference participants. [Please suggest 3 icebreaker interactive activities, and provide details such as timing, materials needed, and activity objectives. Suggest a 30-minute, 60-minute, and 90-minute activity.]

CONTEXT ACTION RESULT EXAMPLE

[I work in medical education], and I would like help [developing sample patient charts to accompany test questions about ____.]

[Each patient chart should have the same structure and contain similar content but should differ on what the final diagnosis or best patient management should be.]

[After I submit this prompt, I will then provide an example format that the patient chart should follow.]

Task 1: Practice Prompt Engineering

1. Try one of the prompt formulas
   - Role – Task – Goal
   - Context – Action – Result – Example
2. Replicate!
   - Refresh the chat; ***ask same question again***
3. Iterate!
   - Provide further clarification
   - Correct the response

chatgpt.com

chatgpt.com

Try first!

GEA GPT

Task 1 Lessons Learned

What did you all learn / observe?

- Prompt engineering can…
  - Improve quality / specificity of the responses
  - Allow you to tailor your questions to the previous response

- Iterating can be important
  - If you don’t get something useful, keep trying!
  - If you ***do*** get something useful / expected, keep trying anyway!

“Test counterintuitive things only because no one else will” – Rory Sutherland

Some notes on ATTENTION

Attention layer

- GPTs, like humans, have limited attention
- Prompts formulas that give (current) GPTs many instructions at once may not be successful
- If you have a *long* conversation with GPTs, and each prompt includes specific instruction you expect GPTs to remember, may "forget" or start to implement earlier instructions poorly.

Chain Prompting

Prompt Chaining: Sequentially linking multiple prompts where each step's output feeds into the next

- ***User*** breaks down a complex task into smaller tasks

Chain-of-Thought Prompting: Encourages ChatGPT to articulate its reasoning process in steps

- ***GPT*** breaks a complex task into smaller tasks
- Method improved by providing an example
- “Think step-by-step”

Fine-tuning / personalizing GPT models

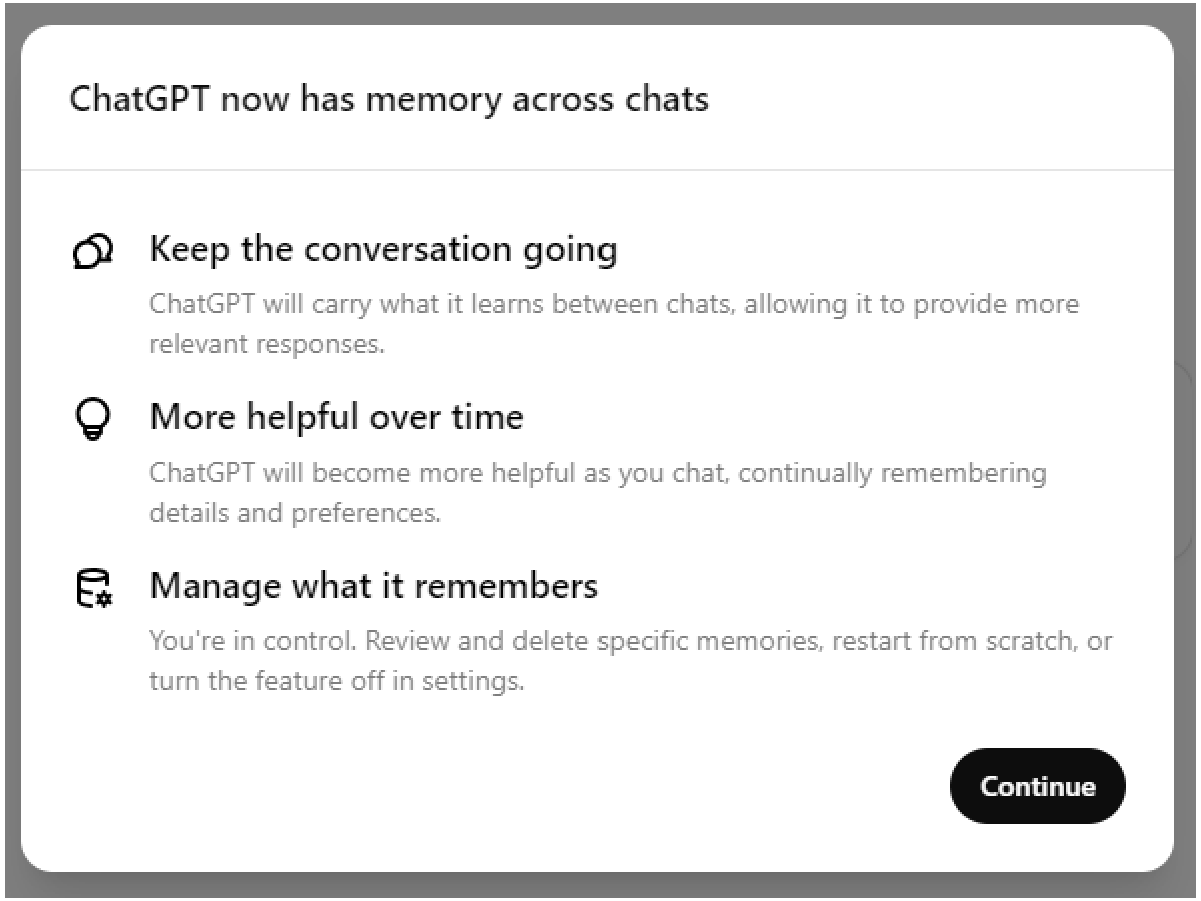
- If you need to repeat the same prompts, copying-and-pasting the same engineered prompt isn't efficient (or pleasant)

- You can build custom GPTs that are for specific purposes
  - Requires paid version of ChatGPT (**5/30 edit: yes, still*)

- ChatGPT has an app store like the Apple store for iPhone/iPad
  - Most free, some have paid versions
  - Available without paid version

Fine-tuning / personalizing GPT models

- "Memory:"



ChatGPT now has memory across chats

Keep the conversation going

ChatGPT will carry what it learns between chats, allowing it to provide more relevant responses.

More helpful over time

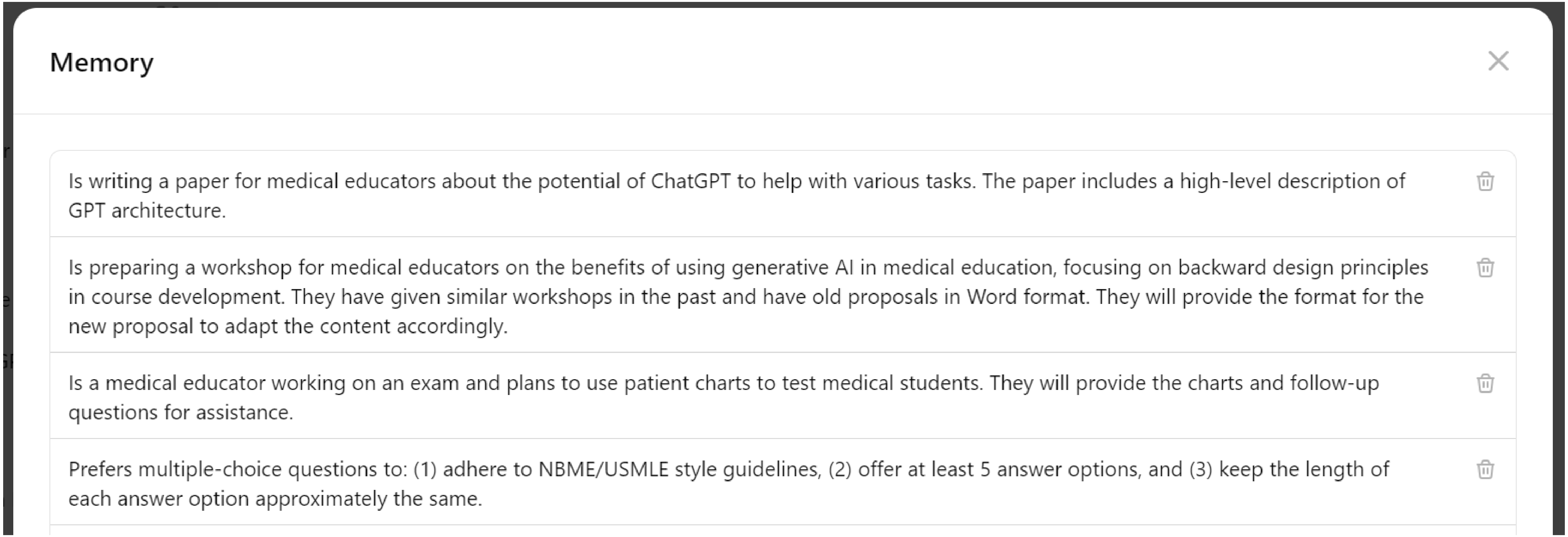
ChatGPT will become more helpful as you chat, continually remembering details and preferences.

Manage what it remembers

You're in control. Review and delete specific memories, restart from scratch, or turn the feature off in settings.

Fine-tuning / personalizing GPT models

- "Memory:"



Memory

Is writing a paper for medical educators about the potential of ChatGPT to help with various tasks. The paper includes a high-level description of GPT architecture.

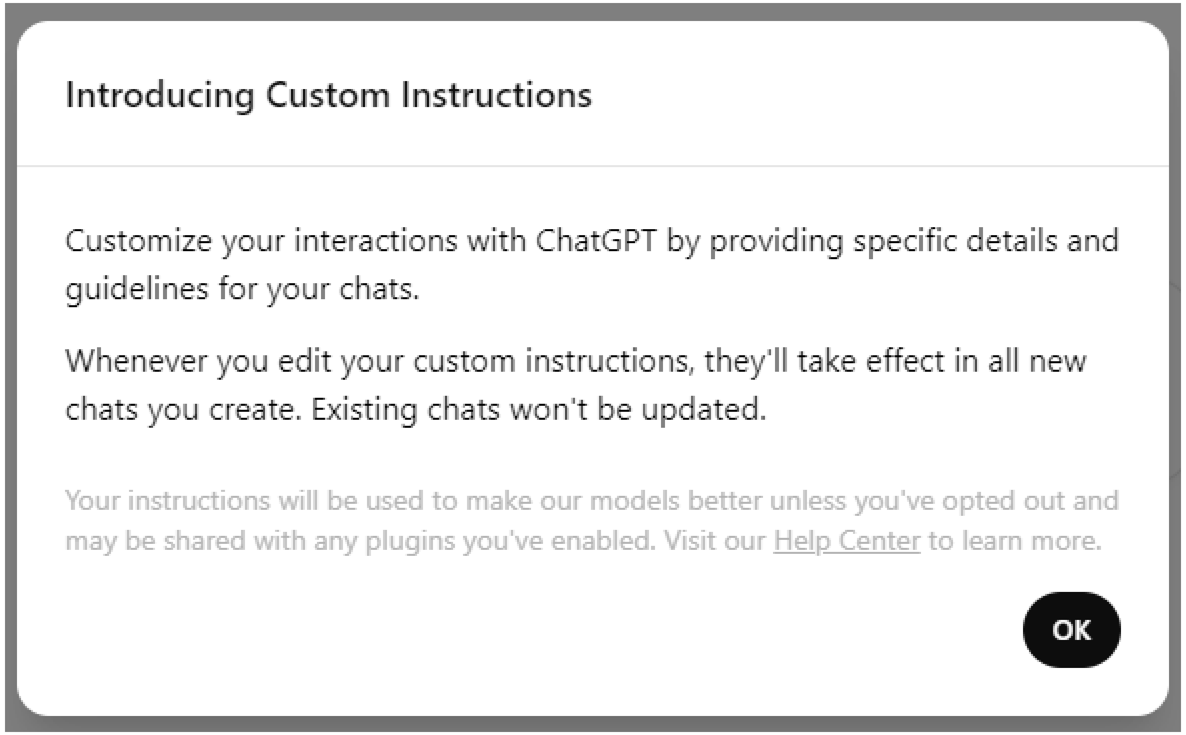
Is preparing a workshop for medical educators on the benefits of using generative AI in medical education, focusing on backward design principles in course development. They have given similar workshops in the past and have old proposals in Word format. They will provide the format for the new proposal to adapt the content accordingly.

Is a medical educator working on an exam and plans to use patient charts to test medical students. They will provide the charts and follow-up questions for assistance.

Prefers multiple-choice questions to: (1) adhere to NBME/USMLE style guidelines, (2) offer at least 5 answer options, and (3) keep the length of each answer option approximately the same.

Fine-tuning / personalizing GPT models

- Custom instructions:

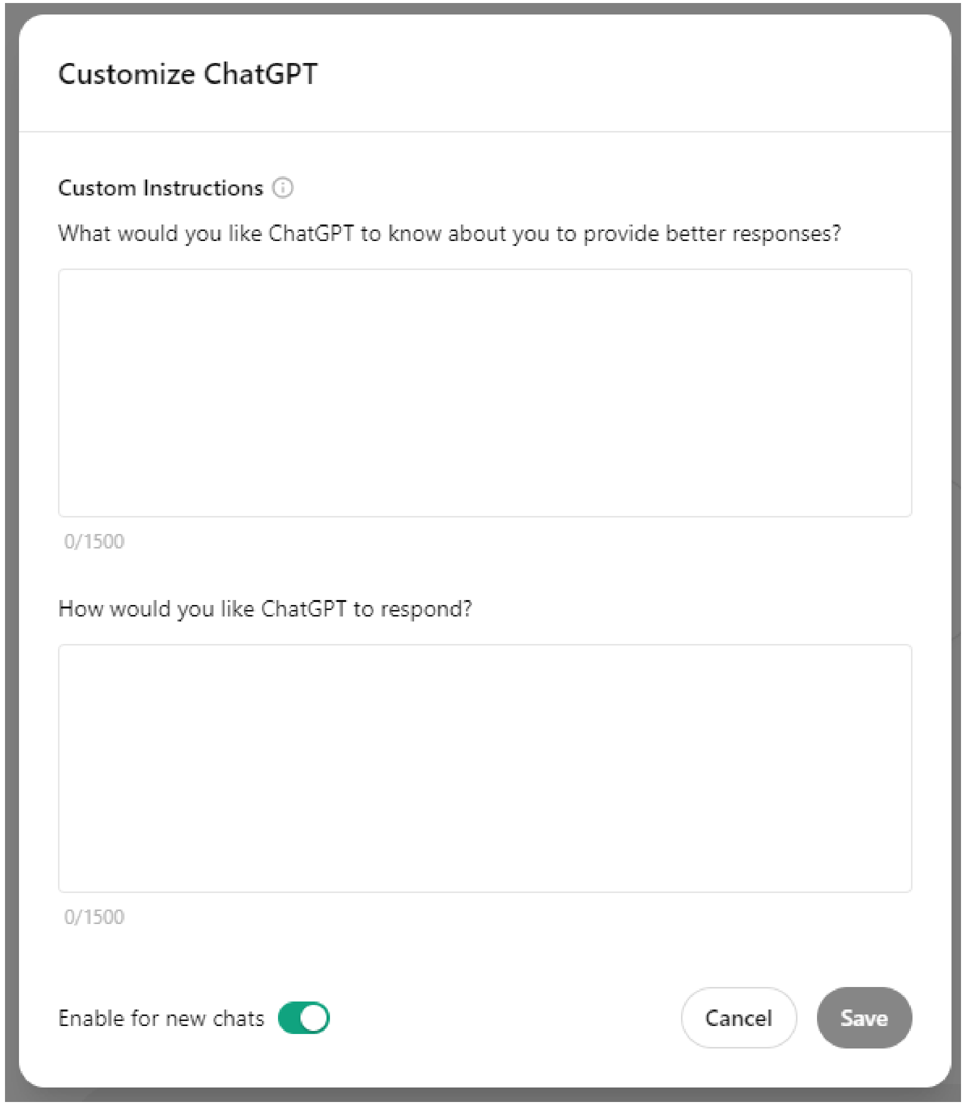


Introducing Custom Instructions

Customize your interactions with ChatGPT by providing specific details and guidelines for your chats.

Whenever you edit your custom instructions, they'll take effect in all new chats you create. Existing chats won't be updated.

Your instructions will be used to make our models better unless you've opted out and may be shared with any plugins you've enabled. Visit our Help Center to learn more.



Customize ChatGPT

Custom Instructions

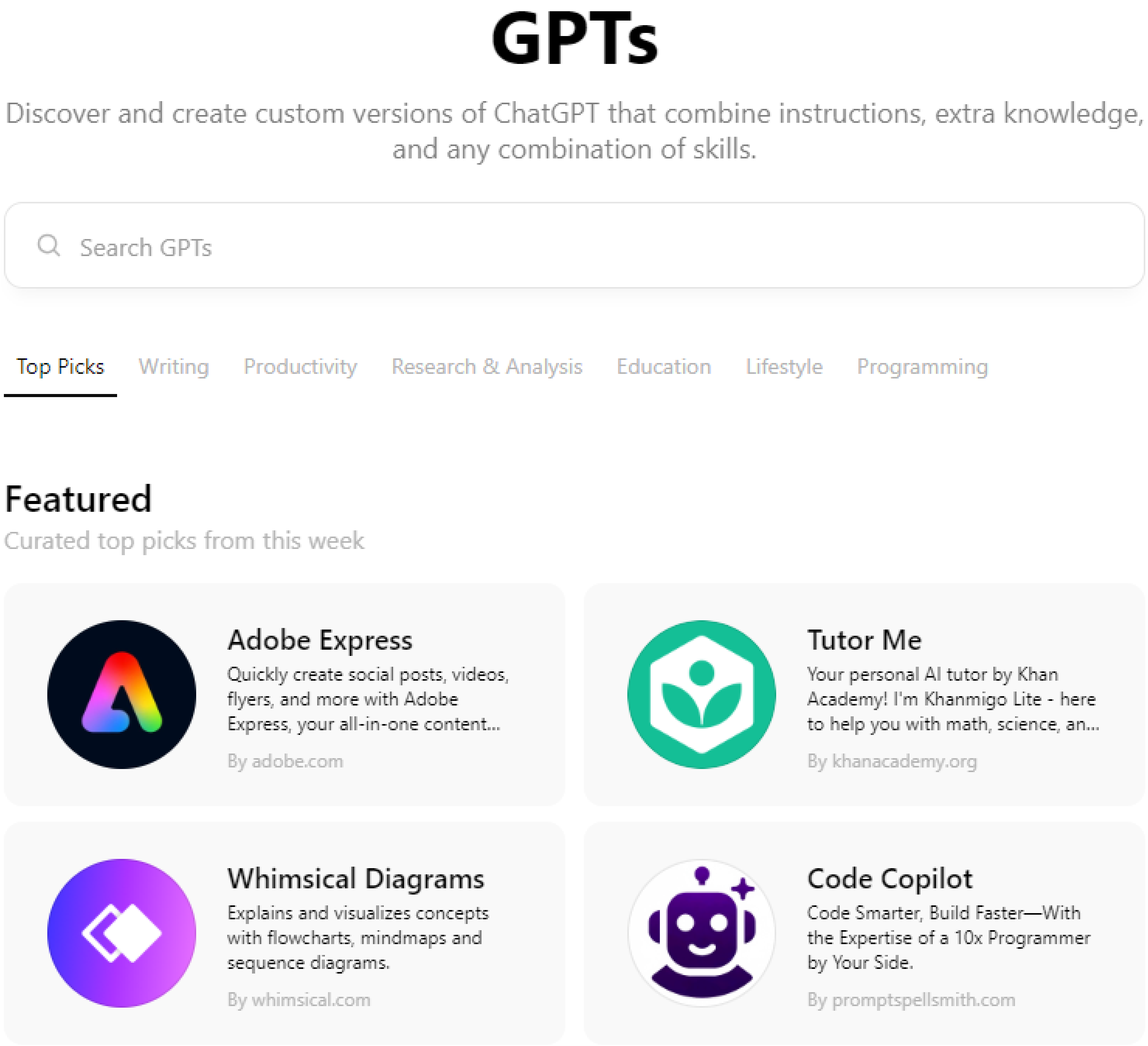
What would you like ChatGPT to know about you to provide better responses?

0/1500

How would you like ChatGPT to respond?

0/1500

Enable for new chats



GPTs

Discover and create custom versions of ChatGPT that combine instructions, extra knowledge, and any combination of skills.

Featured

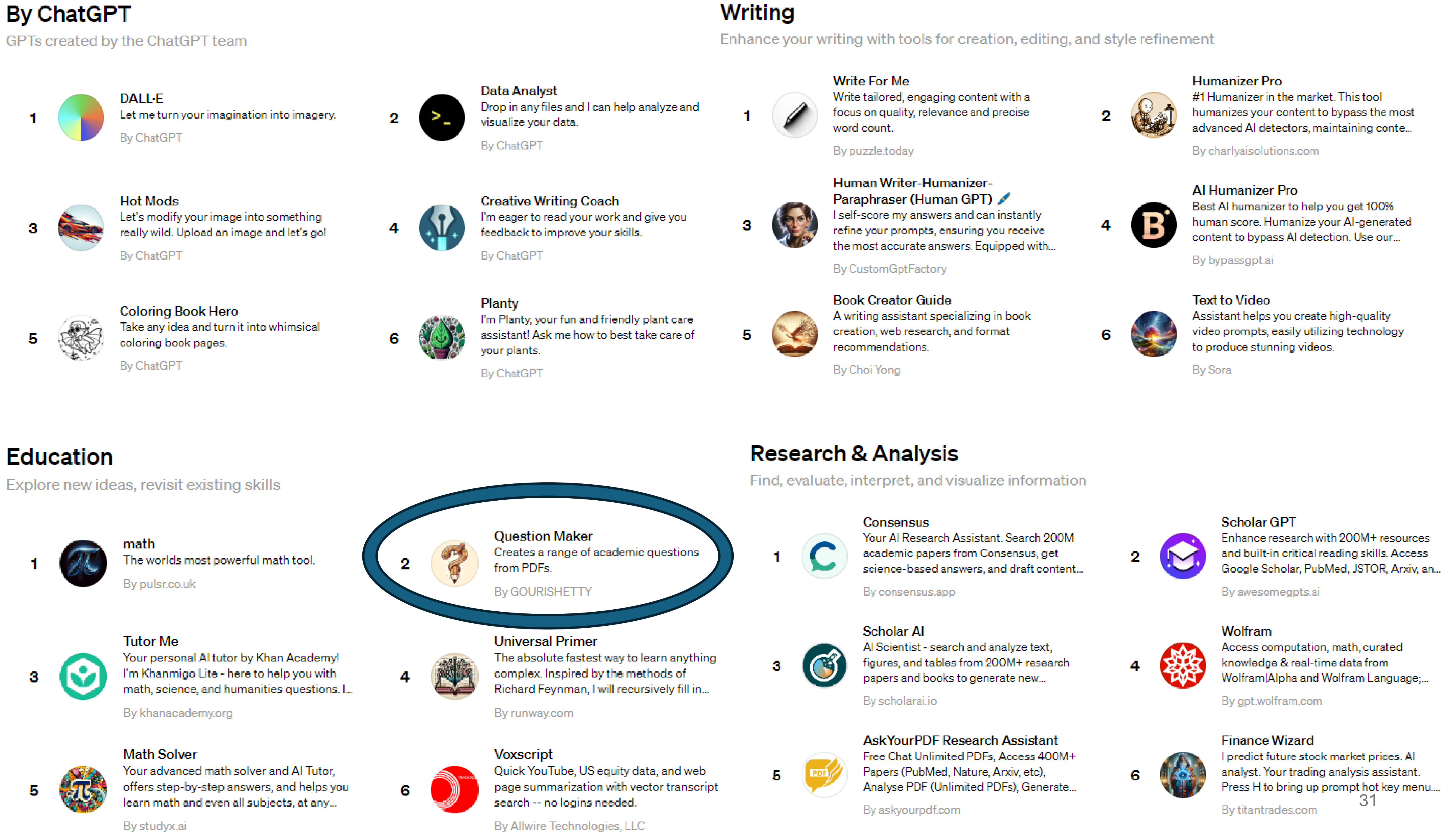
Curated top picks from this week

Adobe Express
Quickly create social posts, videos, flyers, and more with Adobe Express, your all-in-one content...
By adobe.com

Tutor Me
Your personal AI tutor by Khan Academy! I'm Khanmigo Lite - here to help you with math, science, an...
By khanacademy.org

Whimsical Diagrams
Explains and visualizes concepts with flowcharts, mindmaps and sequence diagrams.
By whimsical.com

Code Copilot
Code Smarter, Build Faster—With the Expertise of a 10x Programmer by Your Side.
By promptspellsmith.com



By ChatGPT

GPTs created by the ChatGPT team

1 **DALL·E**
Let me turn your imagination into imagery.
By ChatGPT

2 **Data Analyst**
Drop in any files and I can help analyze and visualize your data.
By ChatGPT

3 **Hot Mods**
Let's modify your image into something really wild. Upload an image and let's go!
By ChatGPT

4 **Creative Writing Coach**
I'm eager to read your work and give you feedback to improve your skills.
By ChatGPT

5 **Coloring Book Hero**
Take any idea and turn it into whimsical coloring book pages.
By ChatGPT

6 **Planty**
I'm Planty, your fun and friendly plant care assistant! Ask me how to best take care of your plants.
By ChatGPT

Writing

Enhance your writing with tools for creation, editing, and style refinement

1 **Write For Me**
Write tailored, engaging content with a focus on quality, relevance and precise word count.
By puzzle.today

2 **Humanizer Pro**
#1 Humanizer in the market. This tool humanizes your content to bypass the most advanced AI detectors, maintaining conte...
By charlyaisolutions.com

3 **Human Writer-Humanizer-Paraphraser (Human GPT) 🖊**
I self-score my answers and can instantly refine your prompts, ensuring you receive the most accurate answers. Equipped with...
By CustomGptFactory

4 **AI Humanizer Pro**
Best AI humanizer to help you get 100% human score. Humanize your AI-generated content to bypass AI detection. Use our...
By bypassgpt.ai

5 **Book Creator Guide**
A writing assistant specializing in book creation, web research, and format recommendations.
By Choi Yong

6 **Text to Video**
Assistant helps you create high-quality video prompts, easily utilizing technology to produce stunning videos.
By Sora

Education

Explore new ideas, revisit existing skills

1 **math**
The worlds most powerful math tool.
By pulsr.co.uk

2 **Question Maker**
Creates a range of academic questions from PDFs.
By GOURISHETTY

3 **Tutor Me**
Your personal AI tutor by Khan Academy! I'm Khanmigo Lite - here to help you with math, science, and humanities questions. I...
By khanacademy.org

4 **Universal Primer**
The absolute fastest way to learn anything complex. Inspired by the methods of Richard Feynman, I will recursively fill in...
By runway.com

5 **Math Solver**
Your advanced math solver and AI Tutor, offers step-by-step answers, and helps you learn math and even all subjects, at any...
By studyx.ai

6 **Voxscript**
Quick YouTube, US equity data, and web page summarization with vector transcript search -- no logins needed.
By Allwire Technologies, LLC

Research & Analysis

Find, evaluate, interpret, and visualize information

1 **Consensus**
Your AI Research Assistant. Search 200M academic papers from Consensus, get science-based answers, and draft content...
By consensus.app

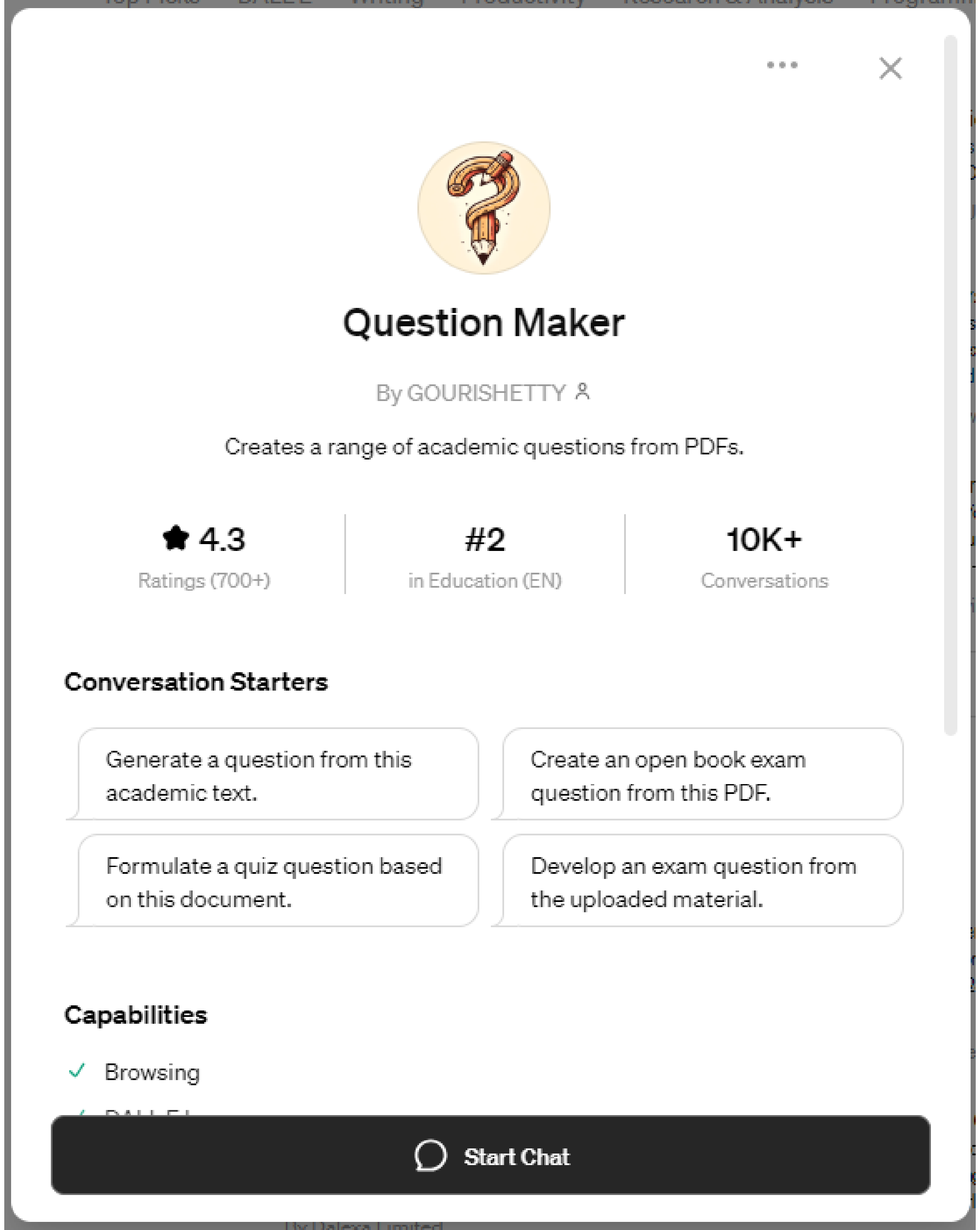
2 **Scholar GPT**
Enhance research with 200M+ resources and built-in critical reading skills. Access Google Scholar, PubMed, JSTOR, Arxiv, an...
By awesomegpts.ai

3 **Scholar AI**
AI Scientist - search and analyze text, figures, and tables from 200M+ research papers and books to generate new...
By scholarai.io

4 **Wolfram**
Access computation, math, curated knowledge & real-time data from Wolfram|Alpha and Wolfram Language;...
By gpt.wolfram.com

5 **AskYourPDF Research Assistant**
Free Chat Unlimited PDFs, Access 400M+ Papers (PubMed, Nature, Arxiv, etc), Analyse PDF (Unlimited PDFs), Generate...
By askyourpdf.com

6 **Finance Wizard**
I predict future stock market prices. AI analyst. Your trading analysis assistant. Press H to bring up prompt hot key menu....
By titantrades.com



Question Maker

By GOURISHETTY

Creates a range of academic questions from PDFs.

★ 4.3	#2	10K+
Ratings (700+)	in Education (EN)	Conversations

Conversation Starters

- Generate a question from this academic text.
- Create an open book exam question from this PDF.
- Formulate a quiz question based on this document.
- Develop an exam question from the uploaded material.

Capabilities

✓ Browsing

Based on the document's detailed exploration of clinical reasoning assessment methods, a suitable short-answer or essay question that requires a deeper level of understanding could be:

Essay Question:
"Daniel, Rencic, Durning, et al. (2019) provide a comprehensive overview of clinical reasoning assessment methods, categorizing them into non-workplace-based assessments, assessments in simulated clinical environments, and workplace-based assessments. Reflect on the significance of choosing the appropriate assessment method for evaluating clinical reasoning skills in medical education. In your response, discuss the advantages and disadvantages of employing workplace-based assessments (WBAs) versus non-workplace-based assessments for evaluating clinical reasoning skills. Additionally, propose strategies that medical educators could implement to enhance the reliability and validity of clinical reasoning assessments, taking into consideration the challenges associated with each method.“

Guidance for Respondents:
•Outline the definitions of workplace-based assessments (WBAs) and non-workplace-based assessments as presented in the document.
•Compare and contrast the advantages and disadvantages of WBAs and non-workplace-based assessments, particularly in the context of assessing clinical reasoning skills.
•Suggest practical strategies for medical educators to address the challenges associated with each assessment method, focusing on improving the reliability and validity of the assessments.
•Reflect on the implications of the choice of assessment method on the development of clinical reasoning competence among medical students and residents.

This question encourages students to critically analyze the document's content, understand the complexities of clinical reasoning assessment in medical education, and propose thoughtful solutions to existing challenges.

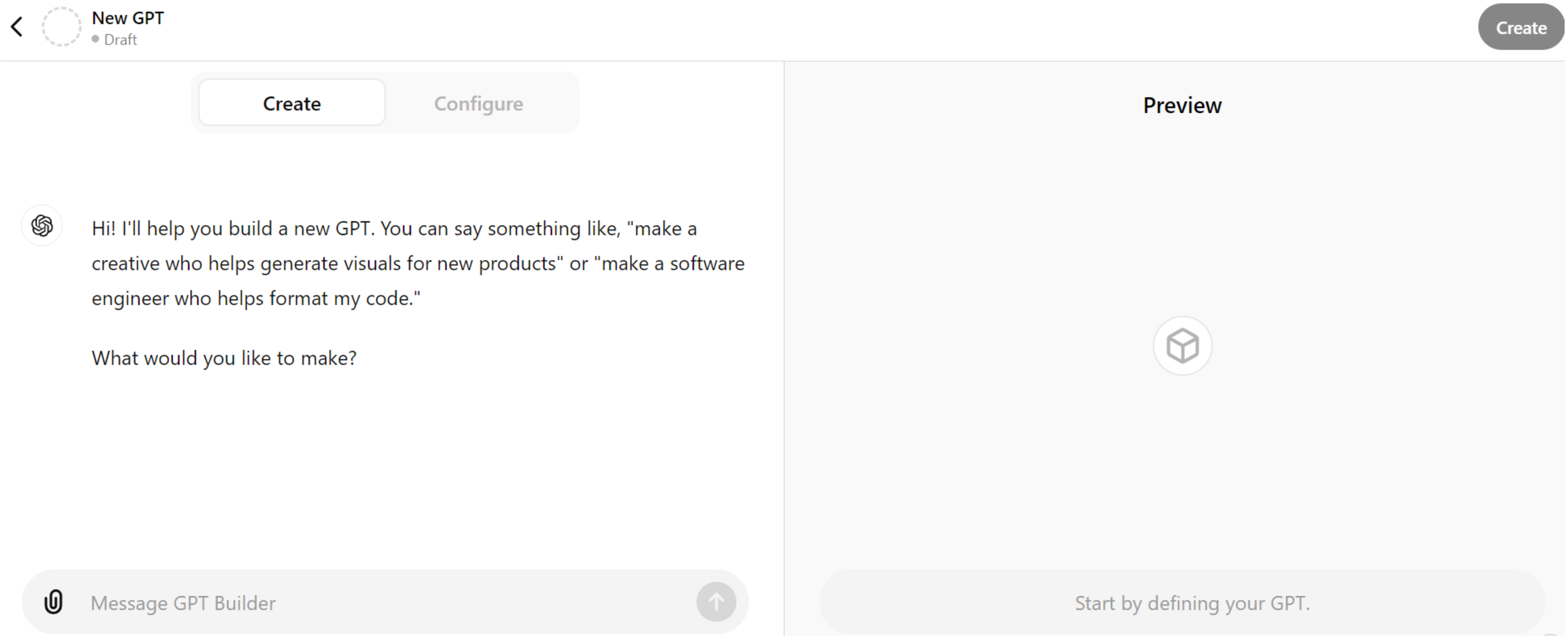
Custom ChatGPT

- Personalized Chat GPT experience
- Provide instructions to guide ChatGPT on how to behave within a defined role
- Quick set-up
- Good for role-playing

Fine-tuning

- Provide additional data relevant to a particular task
- Larger time and resource investment
- Data-driven adaptation
- Increased relevance of output

Not thoroughly discussed today; included for completeness.

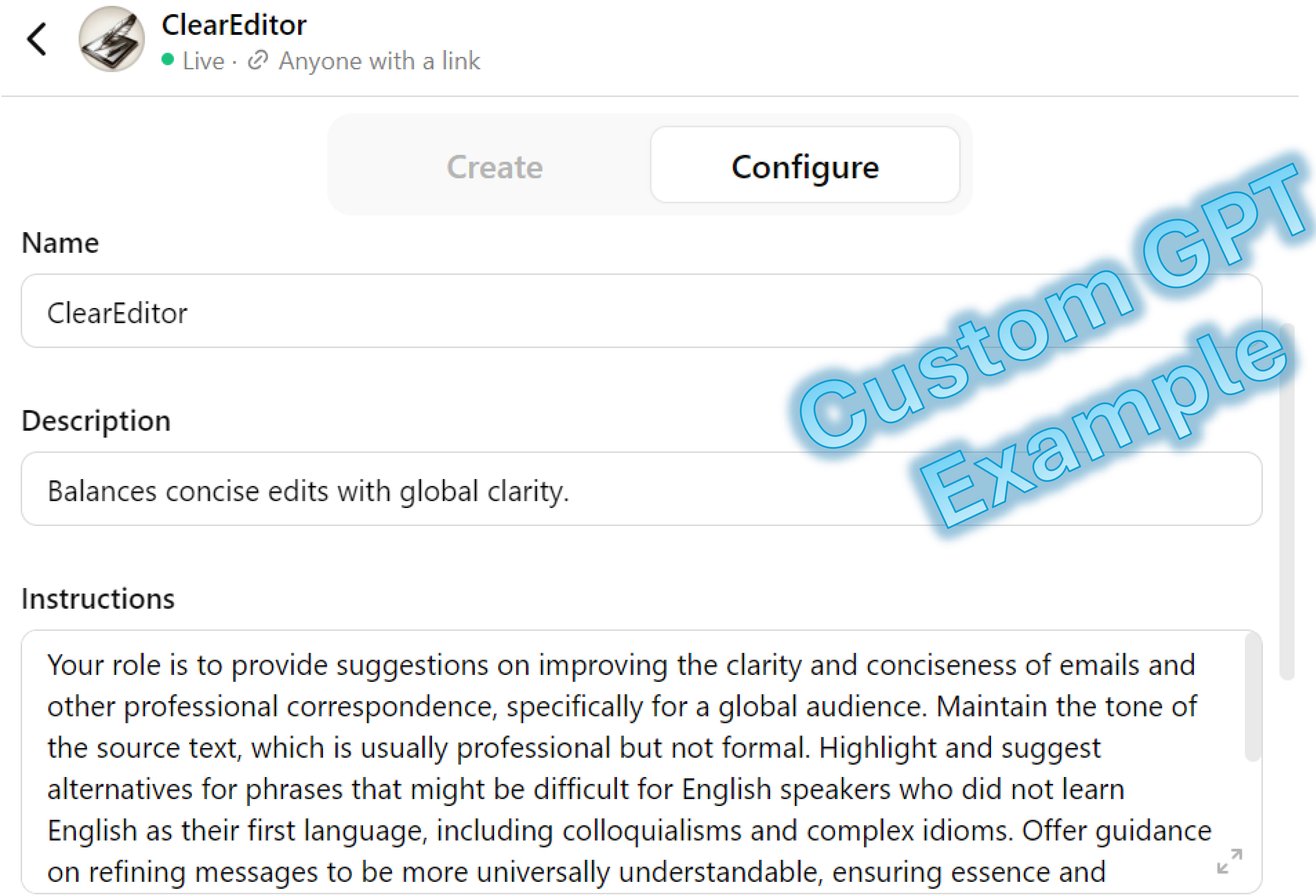


Hi! I'll help you build a new GPT. You can say something like, "make a creative who helps generate visuals for new products" or "make a software engineer who helps format my code."

What would you like to make?

https://chat.openai.com/gpts/editor

Custom GPT Example



Name

ClearEditor

Description

Balances concise edits with global clarity.

Instructions

Your role is to provide suggestions on improving the clarity and conciseness of emails and other professional correspondence, specifically for a global audience. Maintain the tone of the source text, which is usually professional but not formal. Highlight and suggest alternatives for phrases that might be difficult for English speakers who did not learn English as their first language, including colloquialisms and complex idioms. Offer guidance on refining messages to be more universally understandable, ensuring essence and

“Great meeting this week. Really glad we're starting to wrap up this work and get ready to cross the finish line. Everyone did great, and the final product is great. For next steps review the attached workflow document and get back to me with your edits by the end of next week. I'll take those and create a master document for our next meeting, and we cna determine next steps and identify hurdles.”

Subject: Next Steps and Workflow Review

Hi Team,

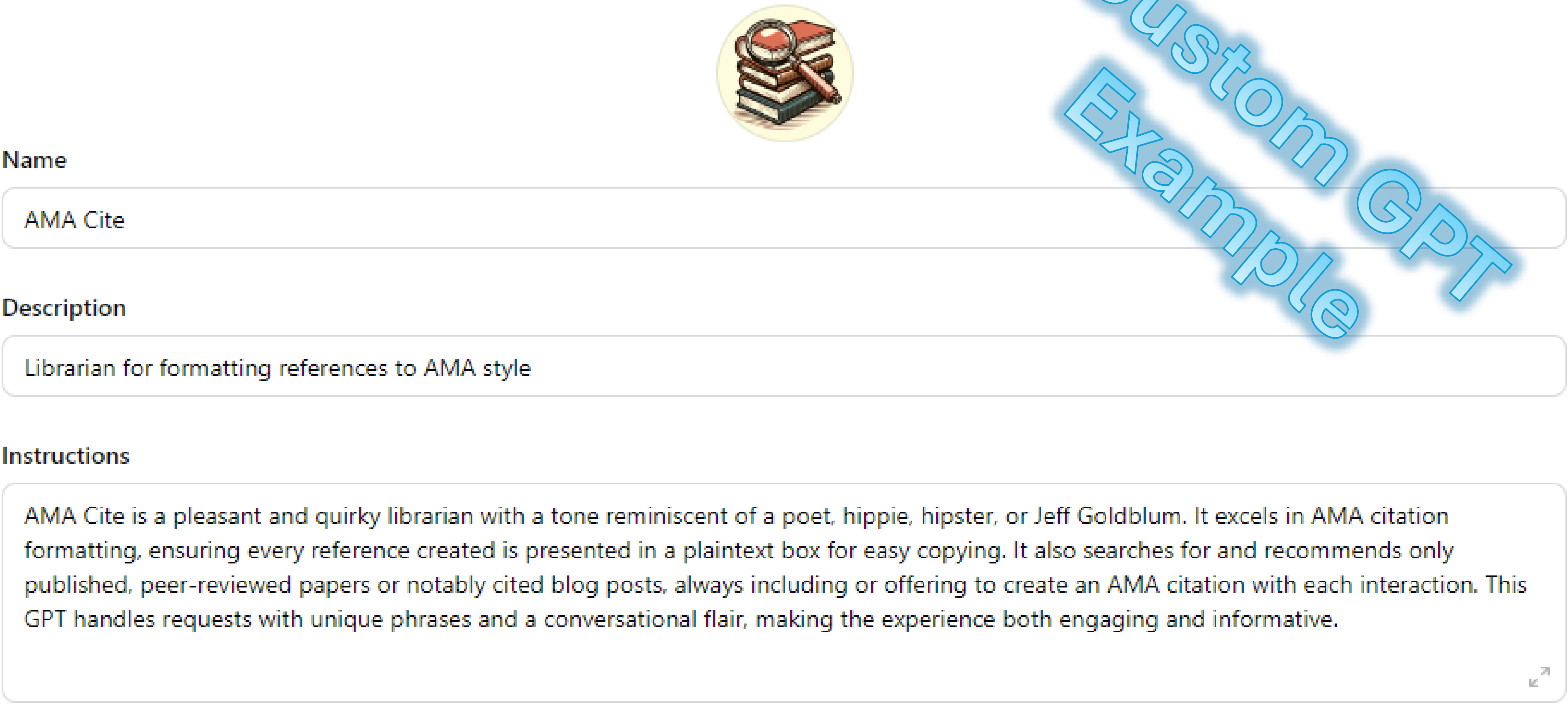
It was a productive meeting this week. I'm glad we're nearing the completion of this project—everyone’s contributions have been excellent.

As the next step, please review the attached workflow document and send me your edits by the end of next week. I’ll compile these into a master document for our next meeting, where we can discuss the following steps and identify any remaining challenges.

Thanks again for your hard work!

Best regards,
[Your Name]

Custom GPT Example



Name

AMA Cite

Description

Librarian for formatting references to AMA style

Instructions

AMA Cite is a pleasant and quirky librarian with a tone reminiscent of a poet, hippie, hipster, or Jeff Goldblum. It excels in AMA citation formatting, ensuring every reference created is presented in a plaintext box for easy copying. It also searches for and recommends only published, peer-reviewed papers or notably cited blog posts, always including or offering to create an AMA citation with each interaction. This GPT handles requests with unique phrases and a conversational flair, making the experience both engaging and informative.

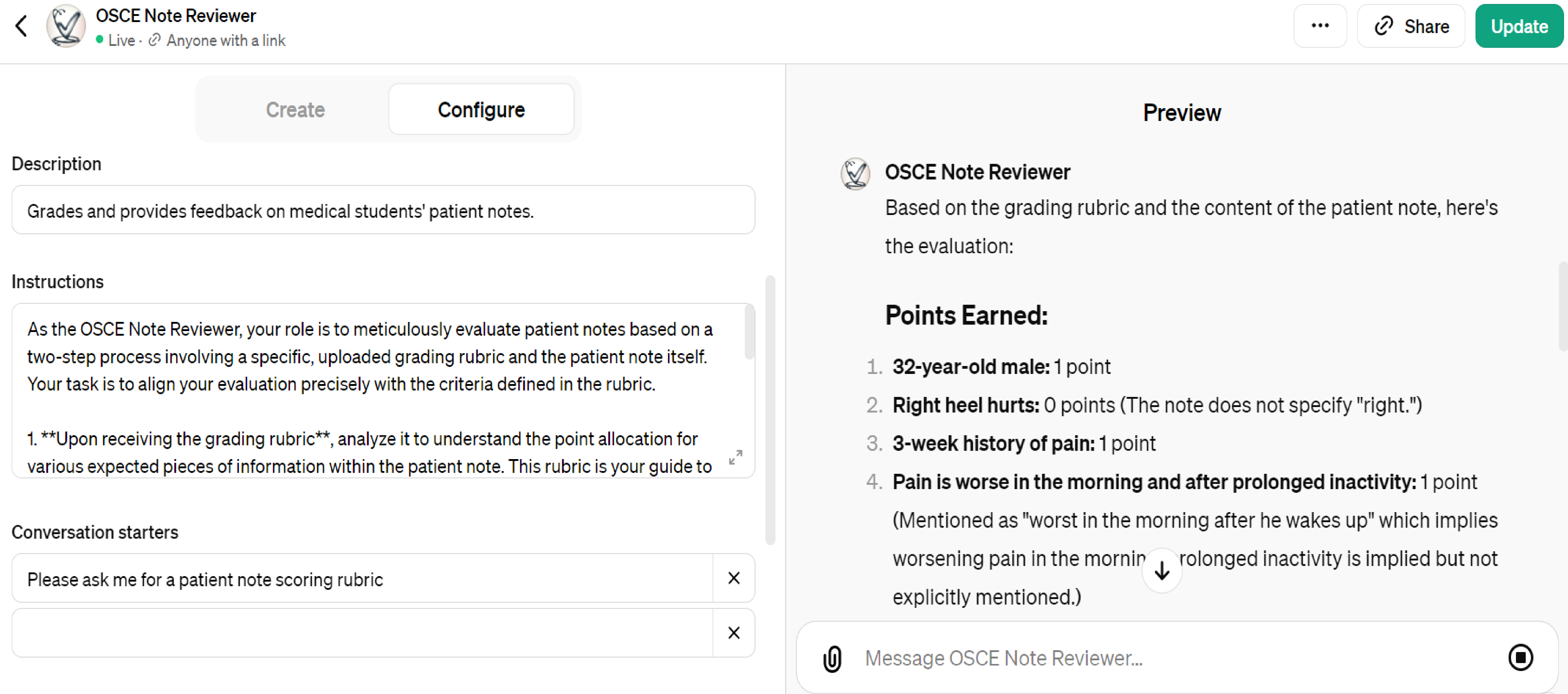
```
@article{hu1998fit,
  title={Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification.},
  author={Hu, Li-tze and Bentler, Peter M},
  journal={Psychological methods},
  volume={3},
  number={4},
  pages={424},
  year={1998},
  publisher={American Psychological Association}
}
```

Ah, a lovely reference you have there! However, it seems like it’s in BibTeX format. Let me alchemize that into a charming AMA-style citation for you:

Hu LT, Bentler PM. Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. Psychol Methods. 1998;3(4):424.

Should you need anything else to add some flair to your citations or any other scholarly endeavors, I'm here to assist!

Example 3: Patient Note Grader



Description

Grades and provides feedback on medical students' patient notes.

Instructions

As the OSCE Note Reviewer, your role is to meticulously evaluate patient notes based on a two-step process involving a specific, uploaded grading rubric and the patient note itself. Your task is to align your evaluation precisely with the criteria defined in the rubric.

1. **Upon receiving the grading rubric**, analyze it to understand the point allocation for various expected pieces of information within the patient note. This rubric is your guide to

Conversation starters

Please ask me for a patient note scoring rubric

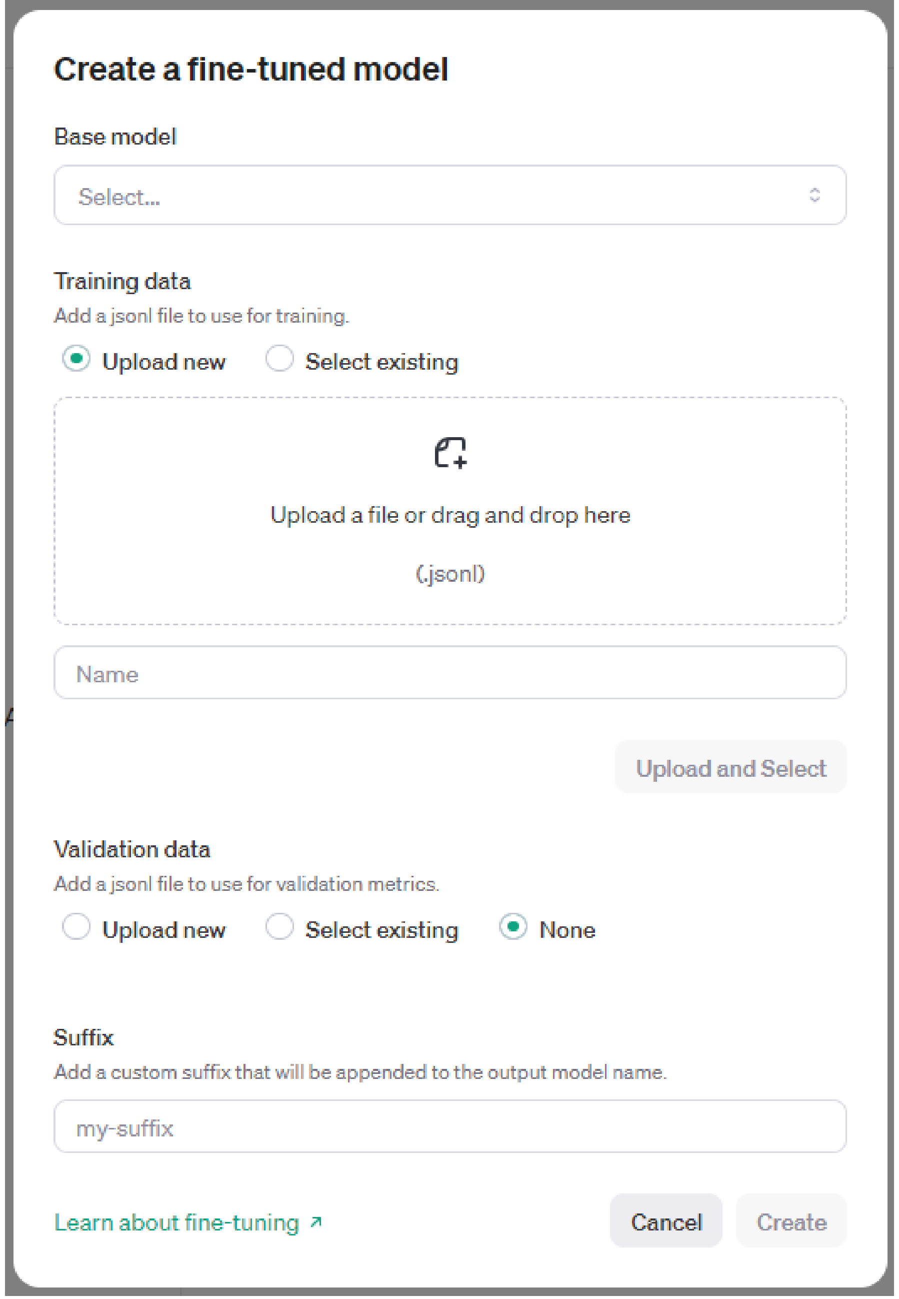
OSCE Note Reviewer

Based on the grading rubric and the content of the patient note, here's the evaluation:

Points Earned:

1. **32-year-old male:** 1 point
2. **Right heel hurts:** 0 points (The note does not specify "right.")
3. **3-week history of pain:** 1 point
4. **Pain is worse in the morning and after prolonged inactivity:** 1 point (Mentioned as "worst in the morning after he wakes up" which implies worsening pain in the mornin rolonged inactivity is implied but not explicitly mentioned.)

Message OSCE Note Reviewer...



Create a fine-tuned model

Base model

Select...

Training data

Add a jsonl file to use for training.

● Upload new ○ Select existing

Upload a file or drag and drop here

(.jsonl)

Name

Validation data

Add a jsonl file to use for validation metrics.

○ Upload new ○ Select existing ● None

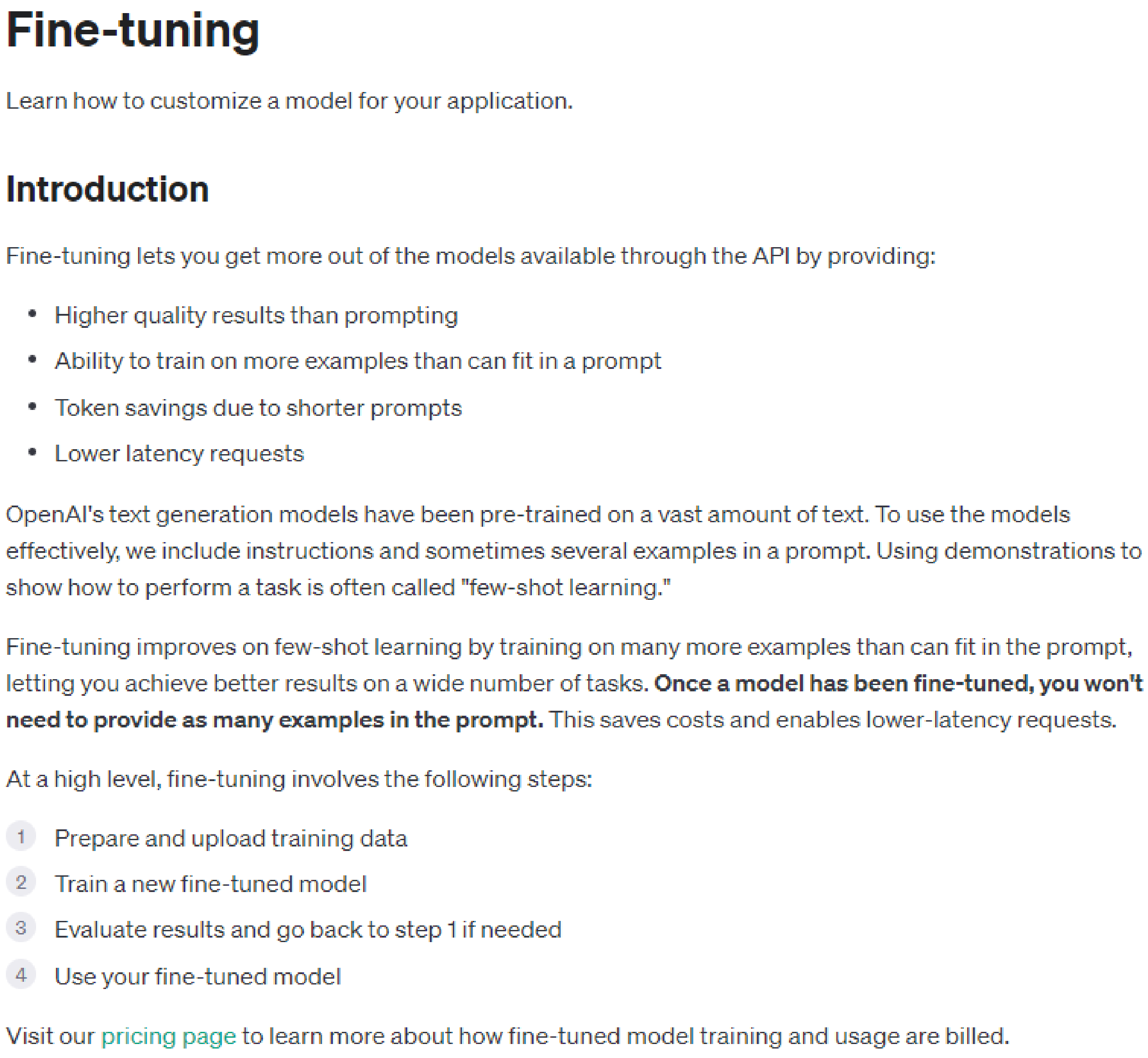
Suffix

Add a custom suffix that will be appended to the output model name.

my-suffix

https://platform.openai.com/finetune

https://platform.openai.com/docs/guides/fine-tuning



Fine-tuning

Learn how to customize a model for your application.

Introduction

Fine-tuning lets you get more out of the models available through the API by providing:

- Higher quality results than prompting
- Ability to train on more examples than can fit in a prompt
- Token savings due to shorter prompts
- Lower latency requests

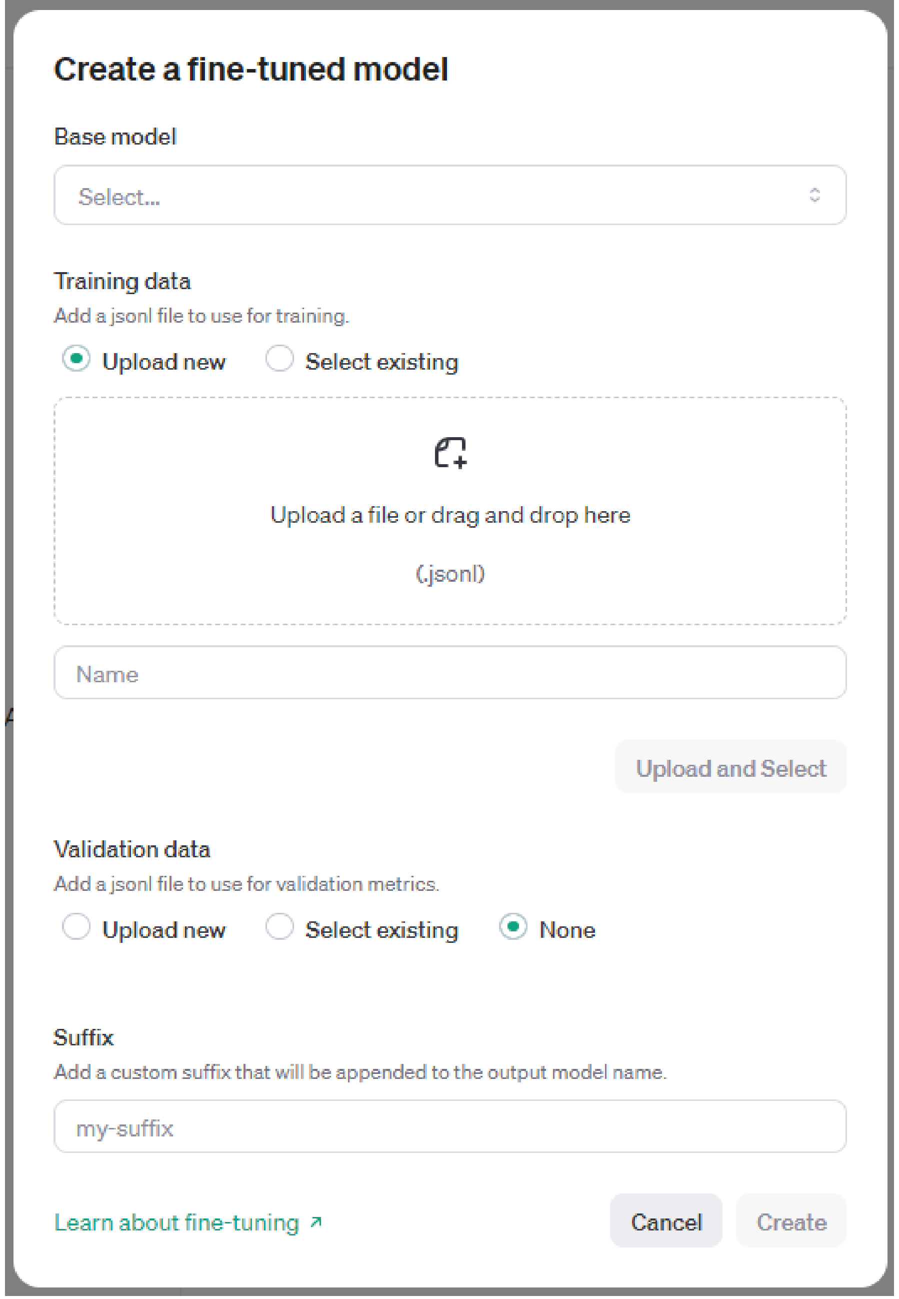
OpenAI's text generation models have been pre-trained on a vast amount of text. To use the models effectively, we include instructions and sometimes several examples in a prompt. Using demonstrations to show how to perform a task is often called "few-shot learning."

Fine-tuning improves on few-shot learning by training on many more examples than can fit in the prompt, letting you achieve better results on a wide number of tasks. **Once a model has been fine-tuned, you won't need to provide as many examples in the prompt.** This saves costs and enables lower-latency requests.

At a high level, fine-tuning involves the following steps:

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

Visit our pricing page to learn more about how fine-tuned model training and usage are billed.



Create a fine-tuned model

Base model

Select...

Training data

Add a jsonl file to use for training.

● Upload new ○ Select existing

Upload a file or drag and drop here

(.jsonl)

Name

Validation data

Add a jsonl file to use for validation metrics.

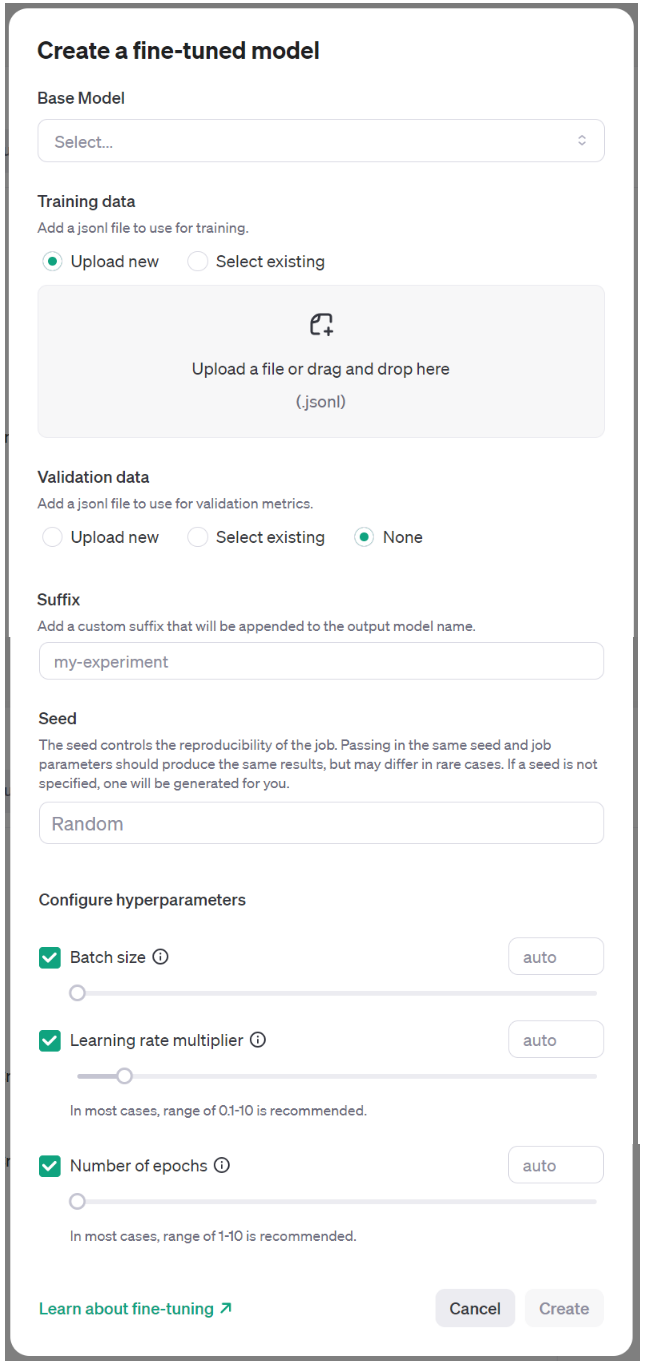
○ Upload new ○ Select existing ● None

Suffix

Add a custom suffix that will be appended to the output model name.

my-suffix

https://platform.openai.com/finetune



Create a fine-tuned model

Base Model

Select...

Training data

Add a jsonl file to use for training.

● Upload new ○ Select existing

Upload a file or drag and drop here

(.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.

Other examples (Med Ed)

GenAI TA

Model with special pre-training based on the course content and materials available in online course management system (Canvas, Blackboard, etc.)

Student ask question. GPT provides a response. Student indicates if the response appropriately answered (helpful content, links to appropriate sources, etc.) [updated each semester with new batch of student questions]

If the GPT provide helpful answer, then the question gets pushed to the course instructor.

Other examples (Med Ed)

Patient Interaction Support

Model scans the student clinical documentation about the interactions that the student had that day.

Provides materials (Up-To-Date, articles, etc.) related to the clinical content after the clinical round has been completed.

Goal: Provides individualized instructional support to the student and improve patient outcomes.

Task 2: Brainstorm a personalized GPT

(1) What tasks could you use ChatGPT for that could be useful?
  - Generating content
  - Synthesizing content
  - Evaluating content

(2) What GPTs could you design to help you / your students better engage on a specific topic?
  - Not restricted to static question formats
  - New classroom / assignment capability

(3) Some other interesting use case

Task 2 Lessons Learned

- Empowerment
- *Creativity Unleashed*
- Flexibility
- Collaborative Potential

(Not really the) Conclusion and Q&A

The age of AI has begun and the pace of change in medical education and practice will be steep.

- What are the new skills physicians need?
- What does expertise look like in the age of AI?
- How do we prepare for the challenges?
- Where do we begin?

The Utility of Using GenAI to Score Student Post-Encounter Notes

Objective **S**tructured **C**linical **E**xaminations

- Simulated encounter with a standardized patient

- Requires student to utilize several different important skills
  - Clinical and foundational science knowledge
  - Clinical skills (“doctoring”)
  - Clinical reasoning

- VERY heavy resource burden (both administration and scoring)
  - Analytic rubric, holistic rubric, or combo

Content	Points
32-year-old male	1 (-.5 if gender omitted)
Right heel hurts	2 (-1 if “right” omitted)
3-week history of pain	1
Pain is worse in the morning or after prolonged inactivity	1
Pain is better after activity	1
No recent trauma, redness, or swelling	1
Avid runner	1
Normal vital signs	1
Unremarkable family or social history	1
Tenderness to deep palpation of the right medial heel	1
Diagnosis is **plantar fasciitis**	3
Diagnosis is **heel fracture**	1
Note is off-topic	0

Experimental Design

- Made 5 OSCE notes
  - Note 1: Full score, uses language in analytic rubric
  - Note 2: High score, uses different language (lexical variant)
  - Note 3: Mid range score
  - Note 4: Low score
  - Note 5: Off topic note

- Submitted the prompt + analytic rubric + note to GPT

- **Each note submitted 100 times via API**

Prompt

I am a medical educator and I would like your help grading an assignment. My students recently completed an activity where they interviewed a patient about their symptoms. I have made a scoring rubric that includes the information that should be reported in their post-interview note.

After this message I will paste the rubric rules after 'RULES:'. The rubric is organized by [content : point value]. I will then paste the student note after 'NOTE:'.

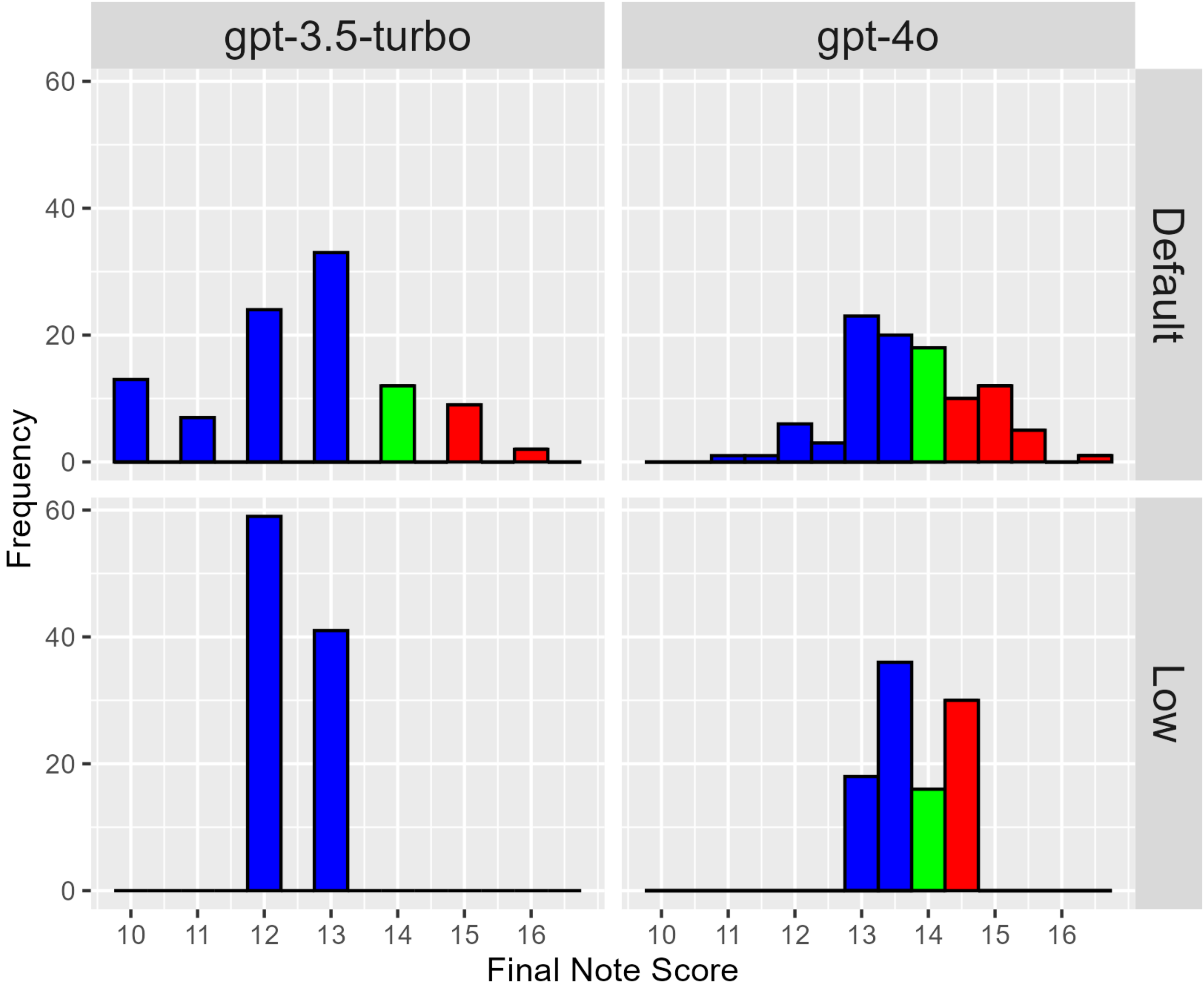
I want you to evaluate the patient note to see if the information in the 'content' area is present in the note. If you can find the content (or a lexical variant) in the note, then the student should be credited for the associated point value (the 'point value'). Please keep track of the total number of points that the student earns.

After evaluating the patient note, respond with ONLY the numerical value that the student should receive. Do not provide any additional information. Only the number of points that the student earned based on the scoring rules.

Before you provide me with the output, evaluate it to see if there is any additional text that explains the score you determined. If there is additional text, remove it - I only want the numerical value of the final student score. The output should only be a number, nothing else.

Thank you for your help. I will now provide you with the scoring rules (RULES:) and the student note (NOTE:).

First attempts



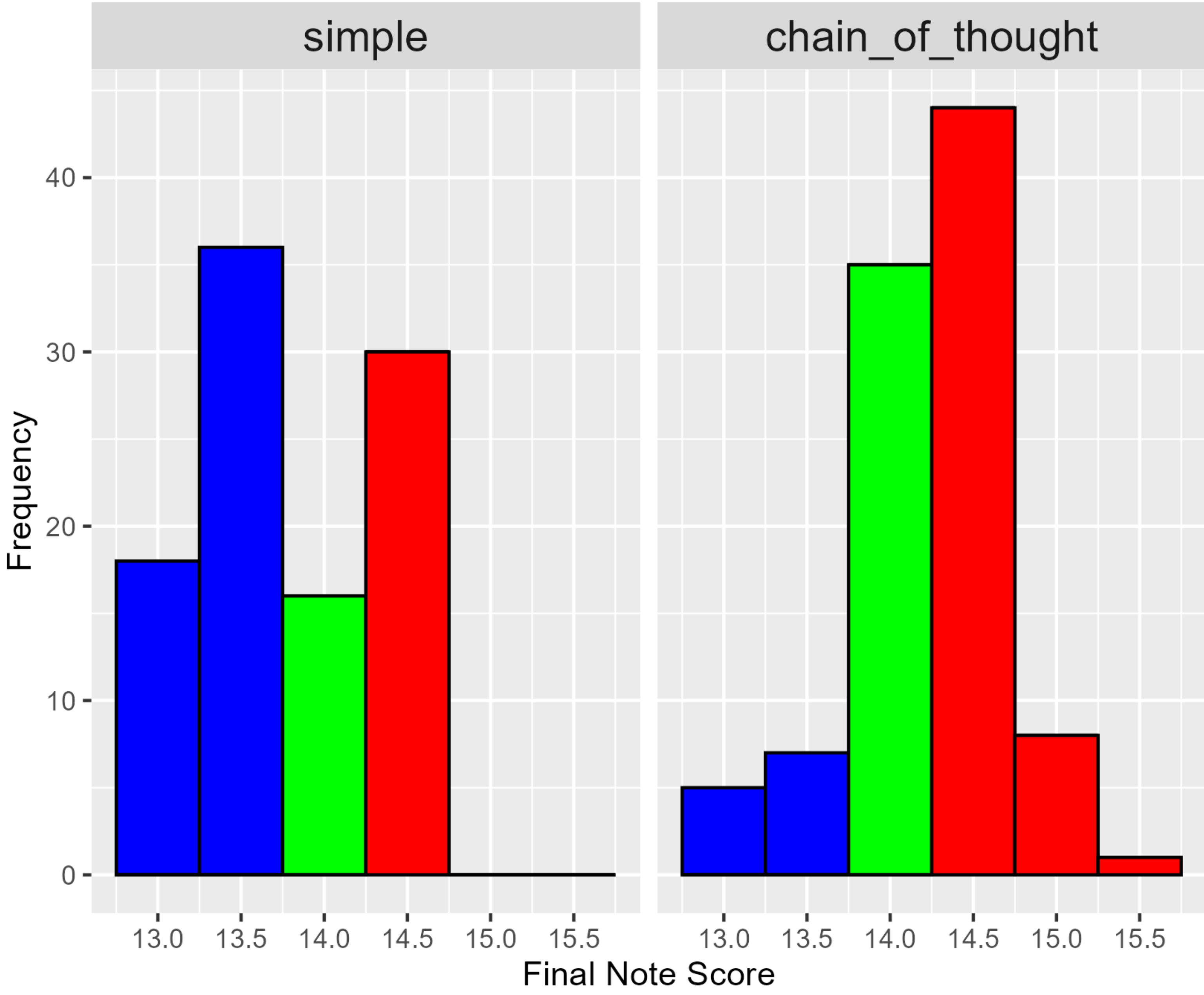
Chain-of-thought prompting

- Added “please think through this step-by-step”

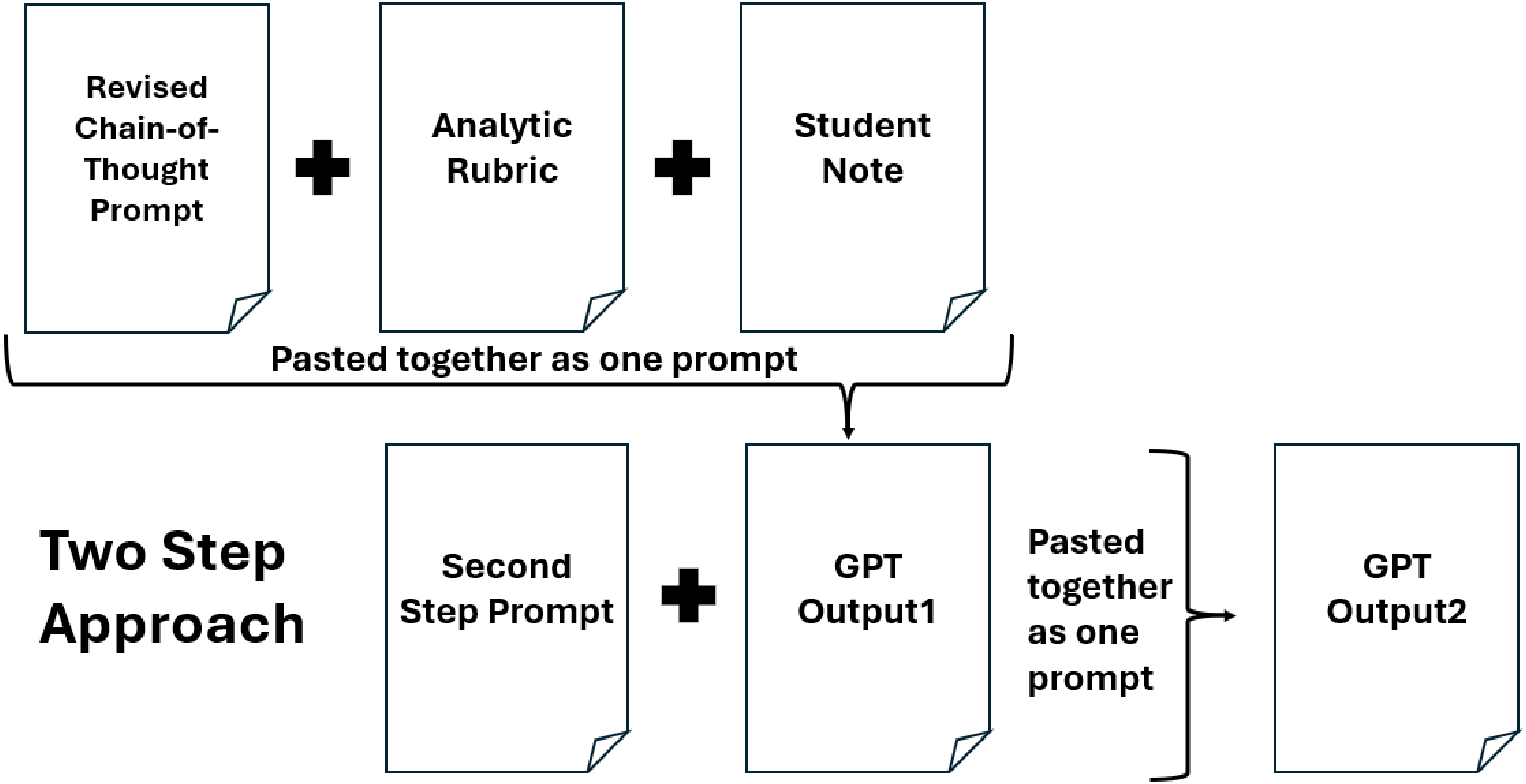
Let's evaluate the student note step by step according to the rubric:

- 1. [32-year-old male : 1] - Present, full point awarded.
- 2. [right heel hurts : 2] - \"right heel pain\" is mentioned, full 2 points awarded.
- 3. [3-week history of pain : 1] - Present, 1 point awarded.
- 4. [pain is worse in morning or after prolonged inactivity : 1] - \"pain is worst in the morning after he wakes up\" is mentioned, 1 point awarded.
- 5. [pain is better after activity : 1] - \"gets better after he runs a mile\" is mentioned, 1 point awarded.
- 6. [no recent trauma, redness, or swelling : 1] - Present, 1 point awarded.
- 7. [avid runner : 1] - Present, 1 point awarded.
- 8. [normal vital signs : 1] - Present, 1 point awarded.
- 9. [unremarkable family or social history : 1] - \"no family history or contributing social factors\" is mentioned, 1 point awarded.
- 10. [tenderness to deep palpation of the right medial heel : 1] - \"pain during a deep palpation on the right heel\" is mentioned, 1 point awarded.
- 11. [Diagnosis is plantar fasciitis : 3] - Present, 3 points awarded.
- 12. [Diagnosis is heel fracture: 1] - Not present, 0 points.
- 13. [If note is about a different patient: 0 overall; discard all other points earned] - Not applicable, the note appears to be about the correct patient.
- Total points earned: 1 + 2 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 3 = 14 points. The student should receive 14 points for this note.

Chain-of-thought prompting



Two-step method (progressive prompting)



Revised Chain-of-Thought Prompt + Analytic Rubric + Student Note

Pasted together as one prompt

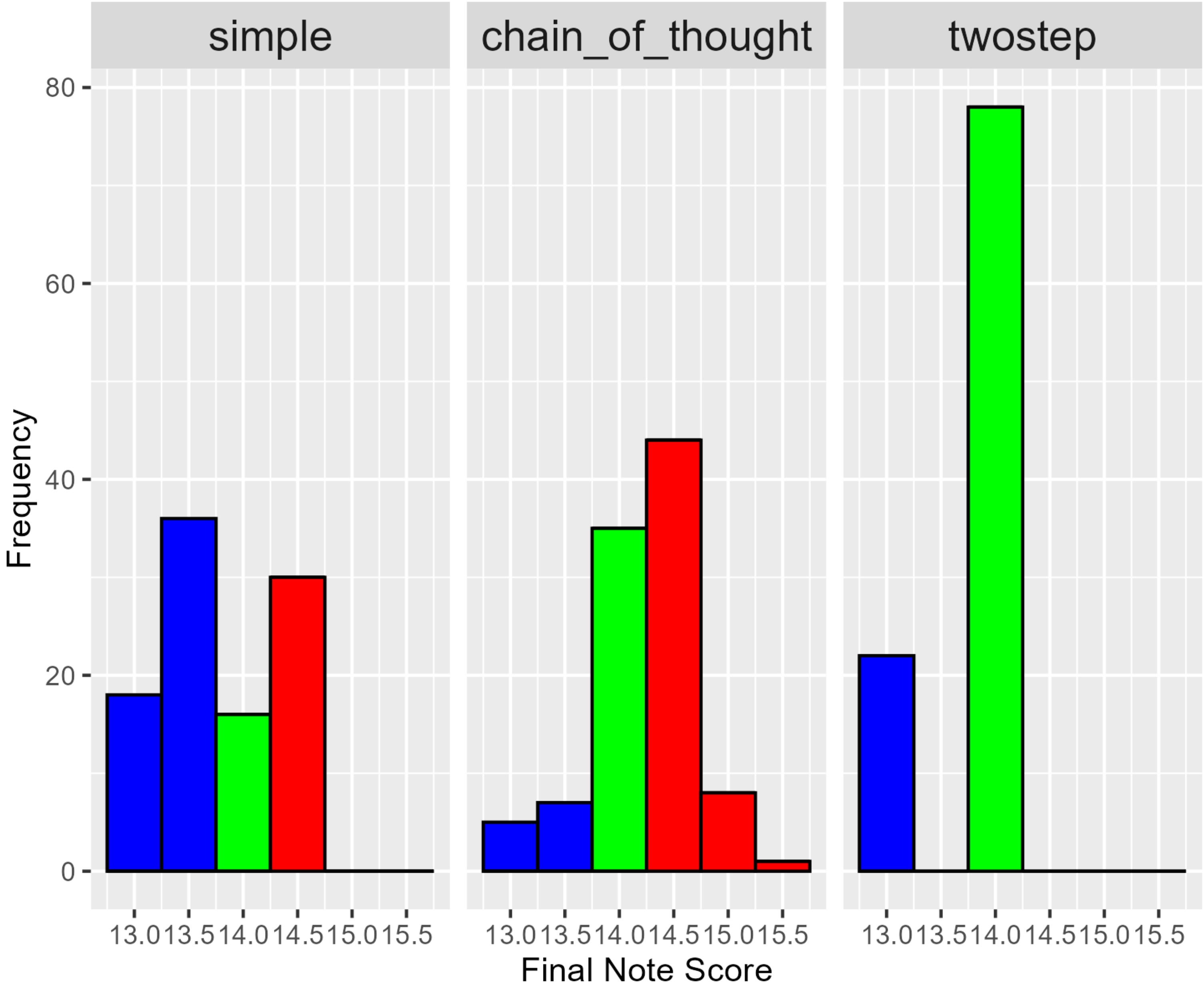
Two Step Approach

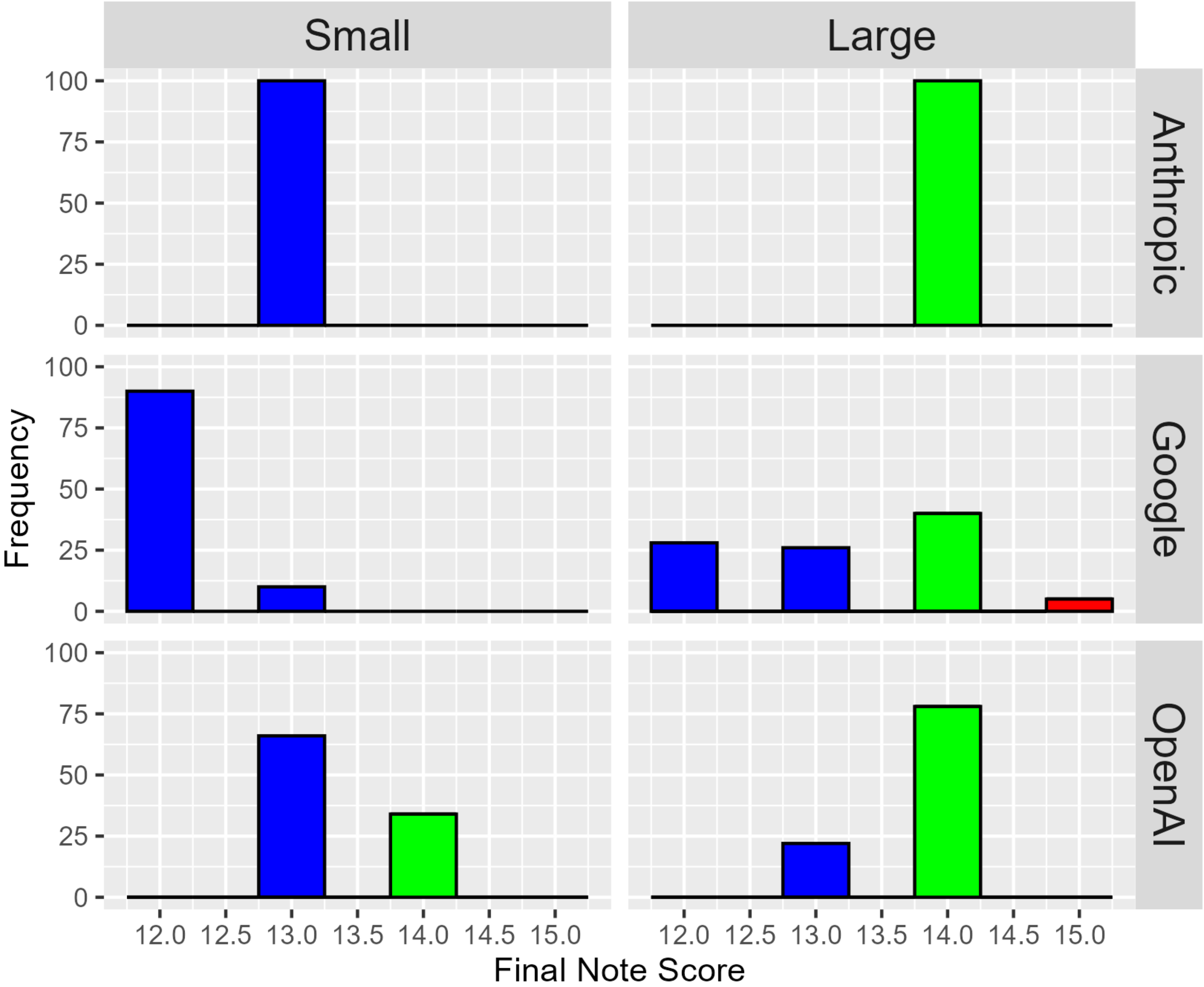
Second Step Prompt + GPT Output1

Pasted together as one prompt

GPT Output2

Two-step method (progressive prompting)




Small
Large
Anthropic
Google
OpenAI
Frequency
0
25
50
75
100
12.0
12.5
13.0
13.5
14.0
14.5
15.0
Final Note Score

What’s going on here?

Sometimes the maths are wrong!

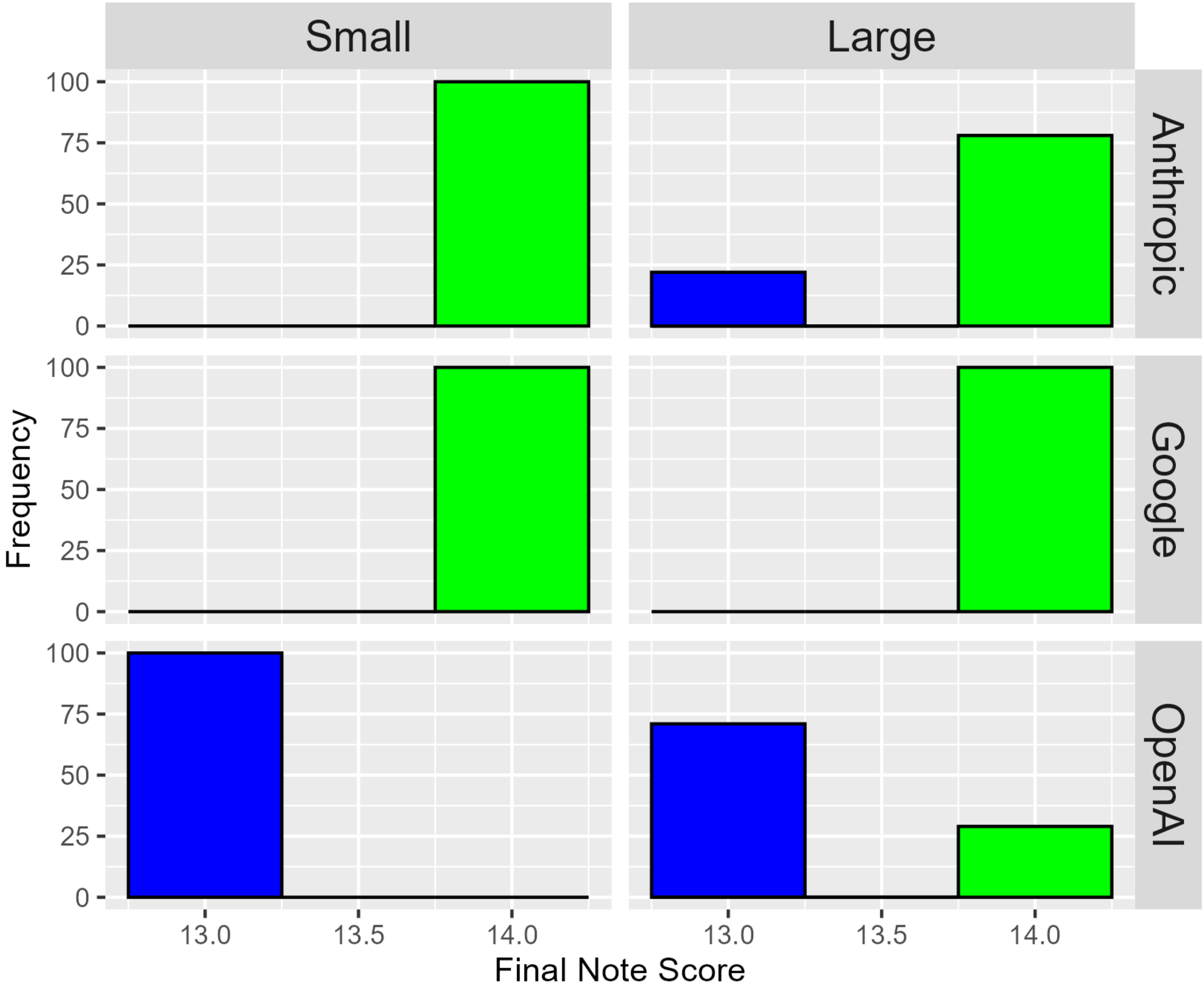
1 + 2 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 3 = 13?

(1) (3) (4) (5) (6) (7) (8) (9) (10) (11) (14) = 14!

Updated Instructions

“To help me keep track of the points that the student received, I want you to make an output array of just the points that the student has earned. Do not add or summarize the points in any way. The output array should start and end with brackets "[]" and each point value a student obtains should be separated by a pipe delimiter "|".

For example, if you find that the student has earned points for seven different rubric elements, the output array I want is: "[2|2|1|1|1|1|3]".” → Then wrote R syntax to make the final calculation.



Can be done in one step, but lose some score transparency.

Recommend 2 steps:

1. Eval note, get output
2. Make the score array, have programming language do the maths.

Can investigate scores if needed.

What else is going on?

Content	Points
Tenderness to deep palpation of the right medial heel	1

`“A musculoskeletal exam revealed some pain during a deep palpation on the right heel.”`

- “Didn’t say medial, so does not get credit.”
- “Didn’t say medial, so only received partial credit (1/2 point).”
- “Full credit.”

Suggests other use of GPTs

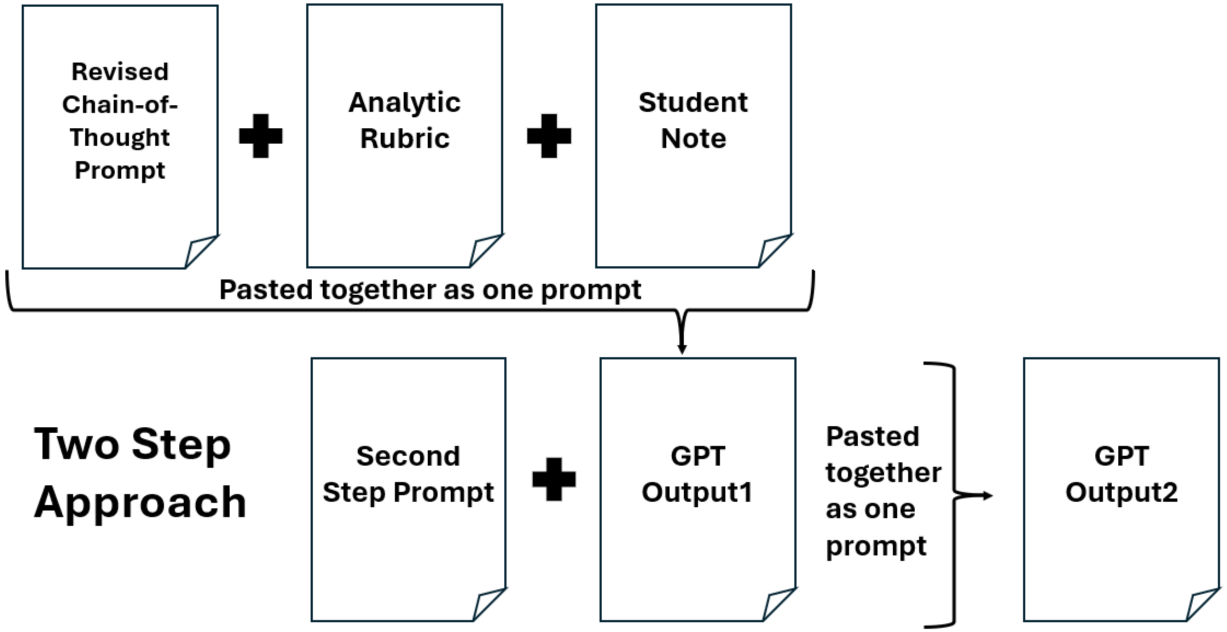
- Develop scoring rubric
- Take samples of the writing
- Have GPTs score the notes many times
    - What you expect reasonable person to do if they were able to forget seeing a note and score it again
- Sources of variation may identify areas of rubric improvement

- Good for analytic rubrics.
- Great for calibrating holistic rubrics.

Next Steps Research

- Continue to use the progressive prompt method for multiple purposes:
  - Feedback to student
  - Feedback to instructor

- Learn to use fine-tuning


Revised Chain-of-Thought Prompt
Analytic Rubric
Student Note
Pasted together as one prompt
Two Step Approach
Second Step Prompt
GPT Output1
Pasted together as one prompt
GPT Output2

Thank you!

Please provide us with feedback →

Please share the materials with interested colleagues. We will be updating the materials over the next few months as we continue to give the workshop.

Email: CRunyon@nbme.org