NLP4Biocuration: Getting going with AI

AgBioData Community Workshop May 2025
Tanya Berardini

Members

- Adam Wright (Co-chair), OICR
- Andrew Olson, CSHL
- Alyssa Proia, TAIR
- Bob Cottingham, ORNL
- Carson Andorf, USDA-ARS
- Doreen Ware, CSHL
- Edwin Ong Jun Kiat, Queen's University of Belfast
- Irene Cobo, ICIFOR-INIA, CSIC
- Jacqueline Campbell, Soybase, USDA-ARS
- James Koltes, Iowa State University
- Jodi Callwood, Iowa State University
- Kapeel Chougule, CSHL
- Laurel Cooper, Oregon State University
- Marcela Tello-Ruiz, CSHL
- Parul Gupta, Oregon State University
- Pierre Larmande, IRD
- Qi Li, Iowa State University
- Rex Nelson, Soybase, USDA-ARS
- Sook Jung, Washington State University
- Srikanth Kumar Karaikal, Cornell University
- Sudhansu Dash, NCGR
- Sushma Naithani, Oregon State University
- Taner Sen, USDA-ARS/UC Berkeley
- Tanya Berardini (Chair), TAIR
- Trish Whetzel, University of North Carolina at Chapel Hill
- Zhiliang Hu, Iowa State University

Goals

- Summarize existing NLP models, tools, and curated training sets, and identify their limitations with AgBioData-curated content.

- Define use cases for applying NLP in biocuration for AgBioData databases (e.g., key research questions).

- Identify common entities curated across AgBioData databases for NLP-driven extraction/curation.

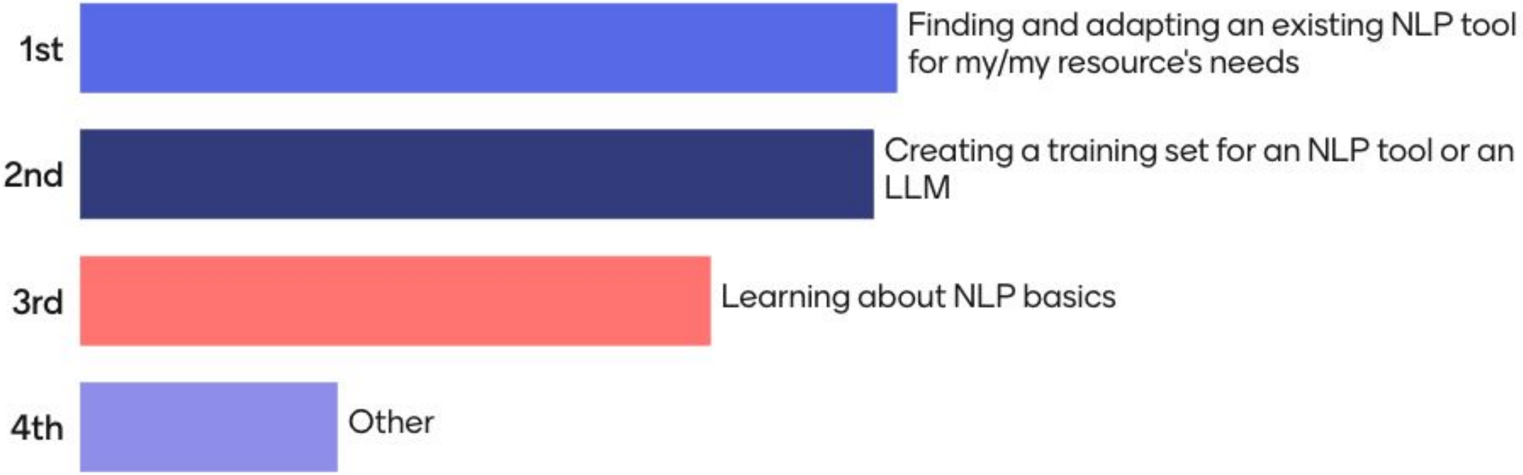
- Recommend strategies and next steps to address these limitations and advance NLP for biocuration to the consortium.

- Identify a tool that would be useful for the most number of resources represented in the WG, at least one use case, develop and test a workflow that can be applied across multiple groups.

Join at menti.com | use code 8884 6240

Mentimeter

Please rank your goals in joining this working group.

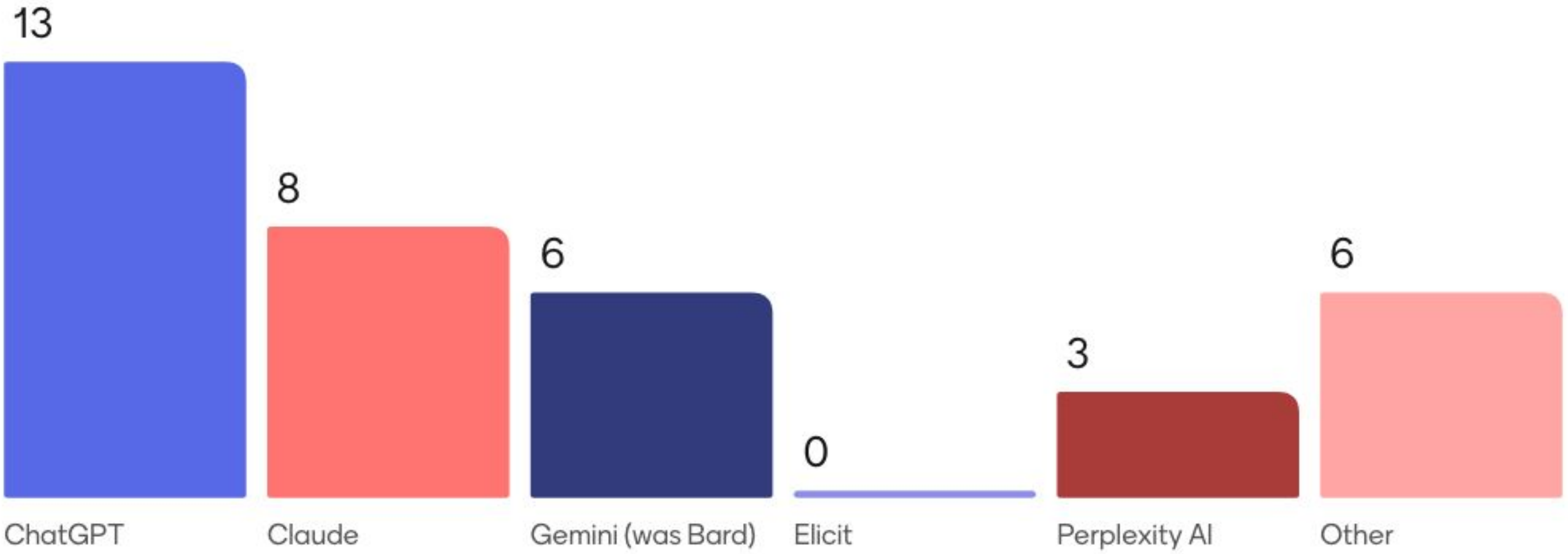


14 / 0

Join at menti.com | use code 8884 6240

Mentimeter

Which of the following LLMs/LLM-based tools have you used?



14 / 0

Join at menti.com | use code 8884 6240

Mentimeter

How have you used LLM tools (current use cases)?

molecular bioogy chat app

reading about llm-nlp

entity recognition

answering random question

coding and development

coding check

data annotation

general questions

experiment with pdfs

training dataset

questions on topcis

coding

generating sqls

editing tests

ontology analyses

draft travel itinerary

pdf extraction

lliterature search

latest research updates

experimentation with pdfs

brainstorming

design experiment

text generation

for text mining

code dev

composing draft statement

ask bioinfo questions

review github issues

13 / 0

Join at menti.com | use code 8884 6240

Mentimeter

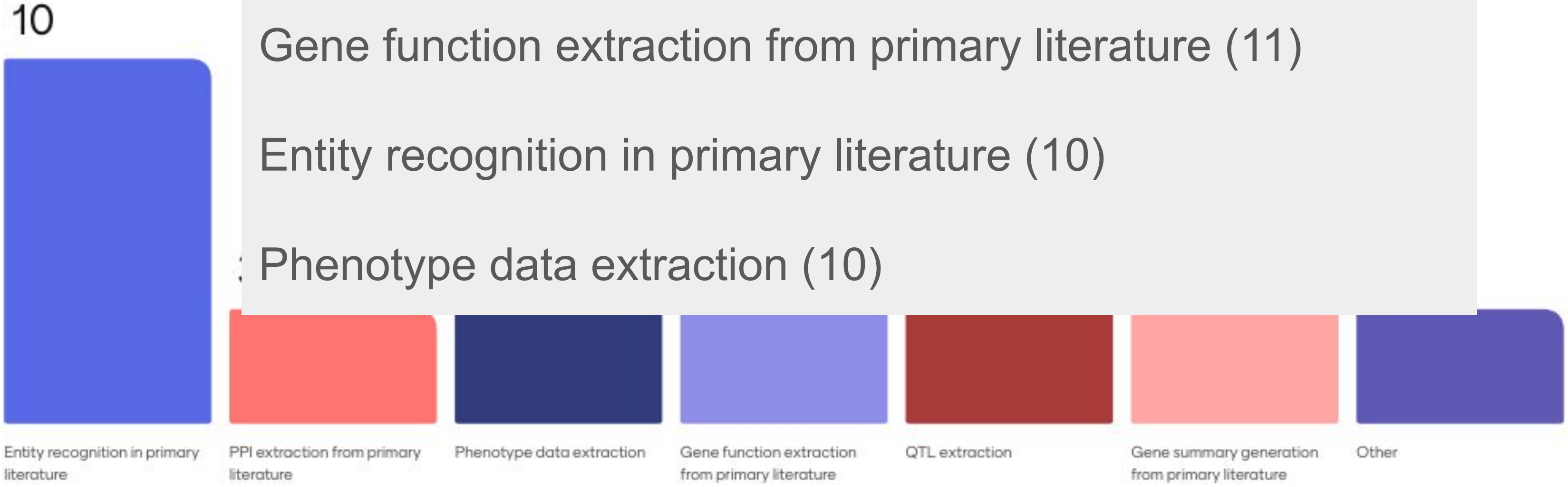
What type of tasks would you like to apply NLP to (future use cases)?

Gene summary generation from primary literature (12)

Gene function extraction from primary literature (11)

Entity recognition in primary literature (10)

Phenotype data extraction (10)

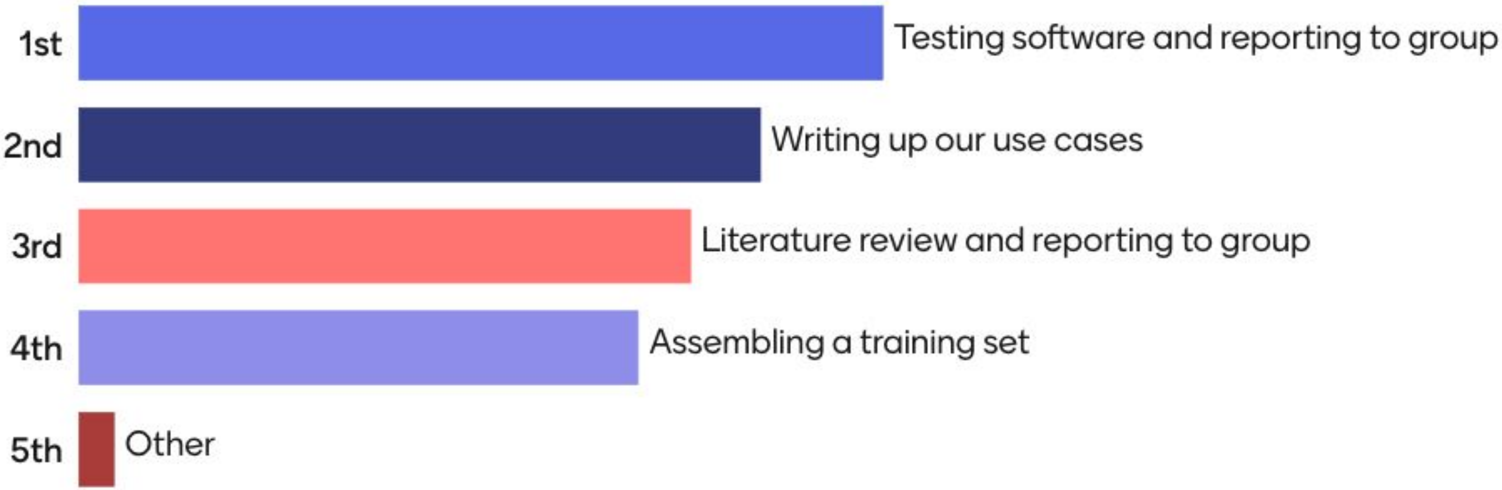


13 / 0

Join at menti.com | use code 8884 6240

Mentimeter

How would you feel comfortable contributing to the group?



13 / 0

Tool Reviews

- **Adam**: React-to-Me Chatbot

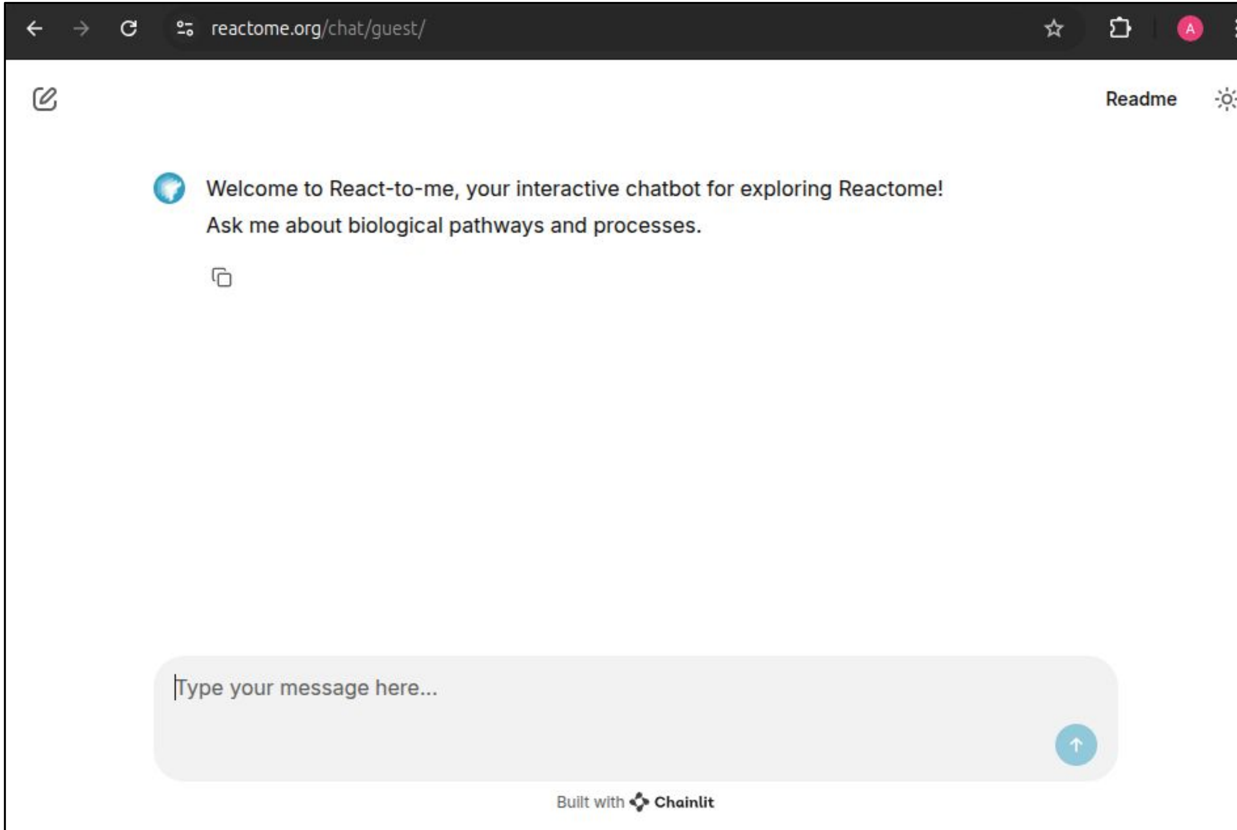
- **Qi**: Recognizing Animal Traits in PubMed Abstracts
- **Edwin**: OntoGPT in AgBio
- **Kapeel**: ChatGPT - Custom GPT Model for Gene Information

- **Jacqueline**: ChatPDF
- **Tanya**: Elicit
- **Parul:** Gemini
- **Sudhansu:** Gemini, ChatPDF, MS CoPilot
- **Sri:** Google Notebook LLM

Reviews

React-to-Me Reactome Chatbot

Able to chat with Pathway information. Currently working on extending to include other databases including UniProt and Alliance of Genome Resources.



reactome.org/chat/guest/

Welcome to React-to-me, your interactive chatbot for exploring Reactome!
Ask me about biological pathways and processes.

Type your message here...

Tool Reviews

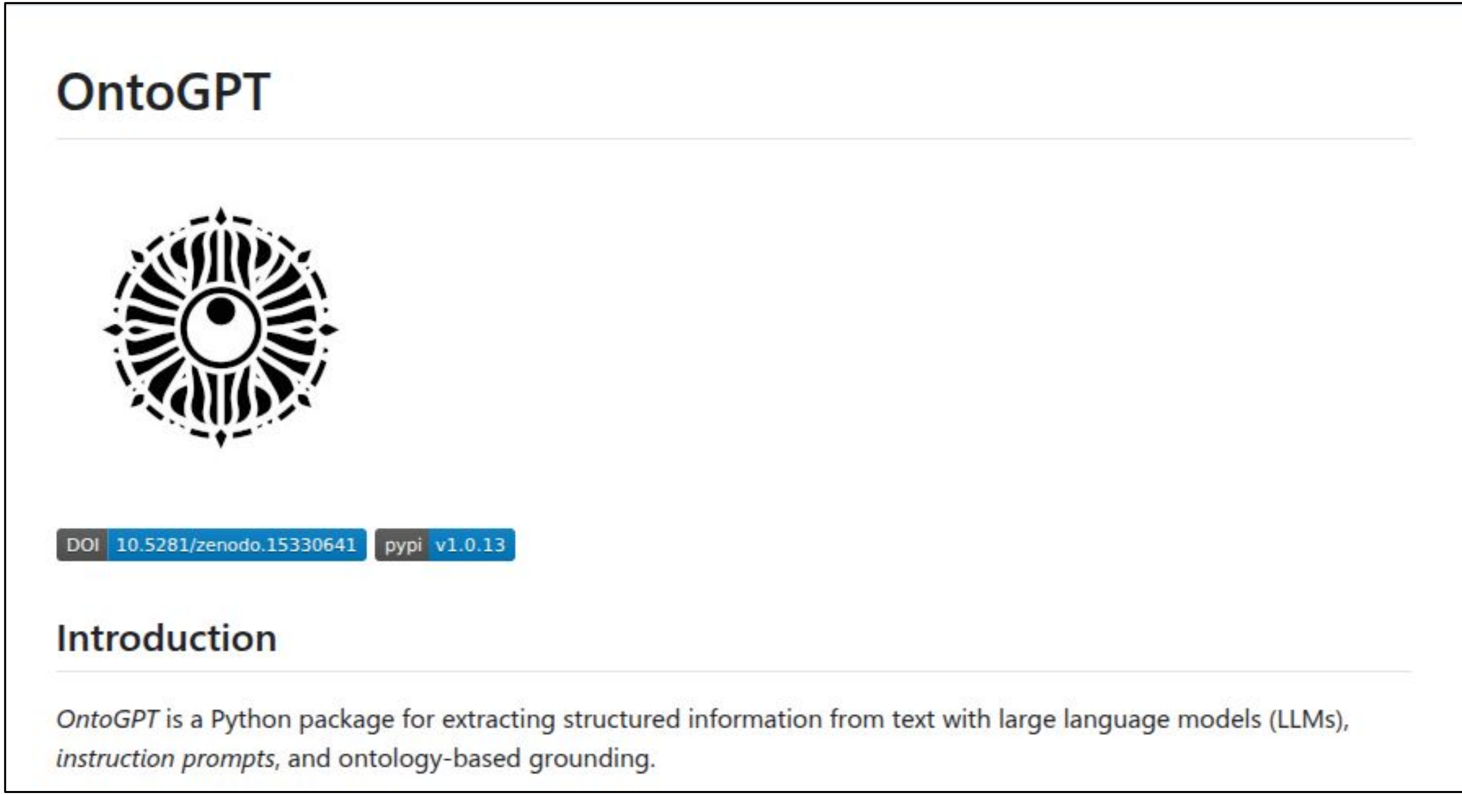
- **Adam**: React-to-Me Chatbot

- **Qi**: Recognizing Animal Traits in PubMed Abstracts
- **Edwin**: OntoGPT in AgBio
- **Kapeel**: ChatGPT - Custom GPT Model for Gene Information

- **Jacqueline**: ChatPDF
- **Tanya**: Elicit
- **Parul:** Gemini
- **Sudhansu:** Gemini, ChatPDF, MS CoPilot
- **Sri:** Google Notebook LLM

Reviews

OntoGPT - Allows users to pull ontology based information from text



OntoGPT

Introduction

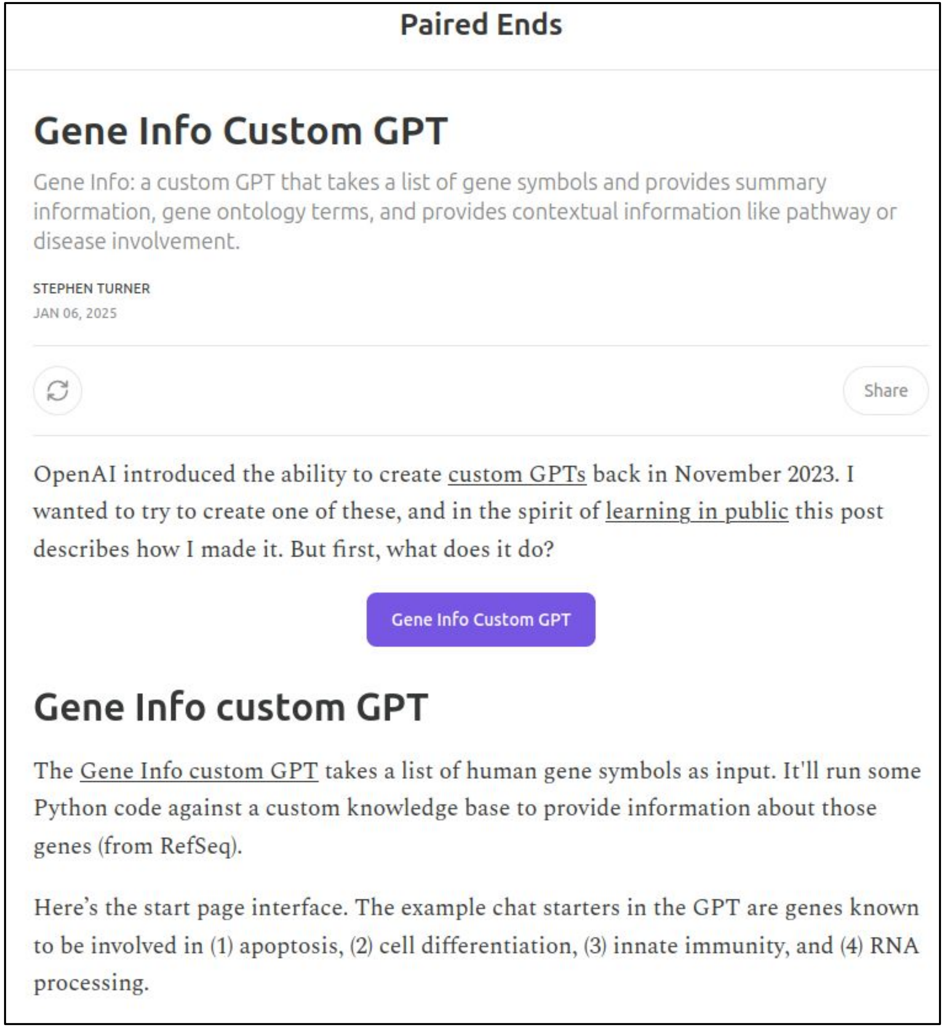
OntoGPT is a Python package for extracting structured information from text with large language models (LLMs), *instruction prompts*, and ontology-based grounding.

Reviews

ChatGPT Pro

Make your own Custom GPT

List of genes and symbols



Paired Ends

Gene Info Custom GPT

Gene Info: a custom GPT that takes a list of gene symbols and provides summary information, gene ontology terms, and provides contextual information like pathway or disease involvement.

STEPHEN TURNER
JAN 06, 2025

OpenAI introduced the ability to create custom GPTs back in November 2023. I wanted to try to create one of these, and in the spirit of learning in public this post describes how I made it. But first, what does it do?

Gene Info custom GPT

The Gene Info custom GPT takes a list of human gene symbols as input. It'll run some Python code against a custom knowledge base to provide information about those genes (from RefSeq).

Here's the start page interface. The example chat starters in the GPT are genes known to be involved in (1) apoptosis, (2) cell differentiation, (3) innate immunity, and (4) RNA processing.

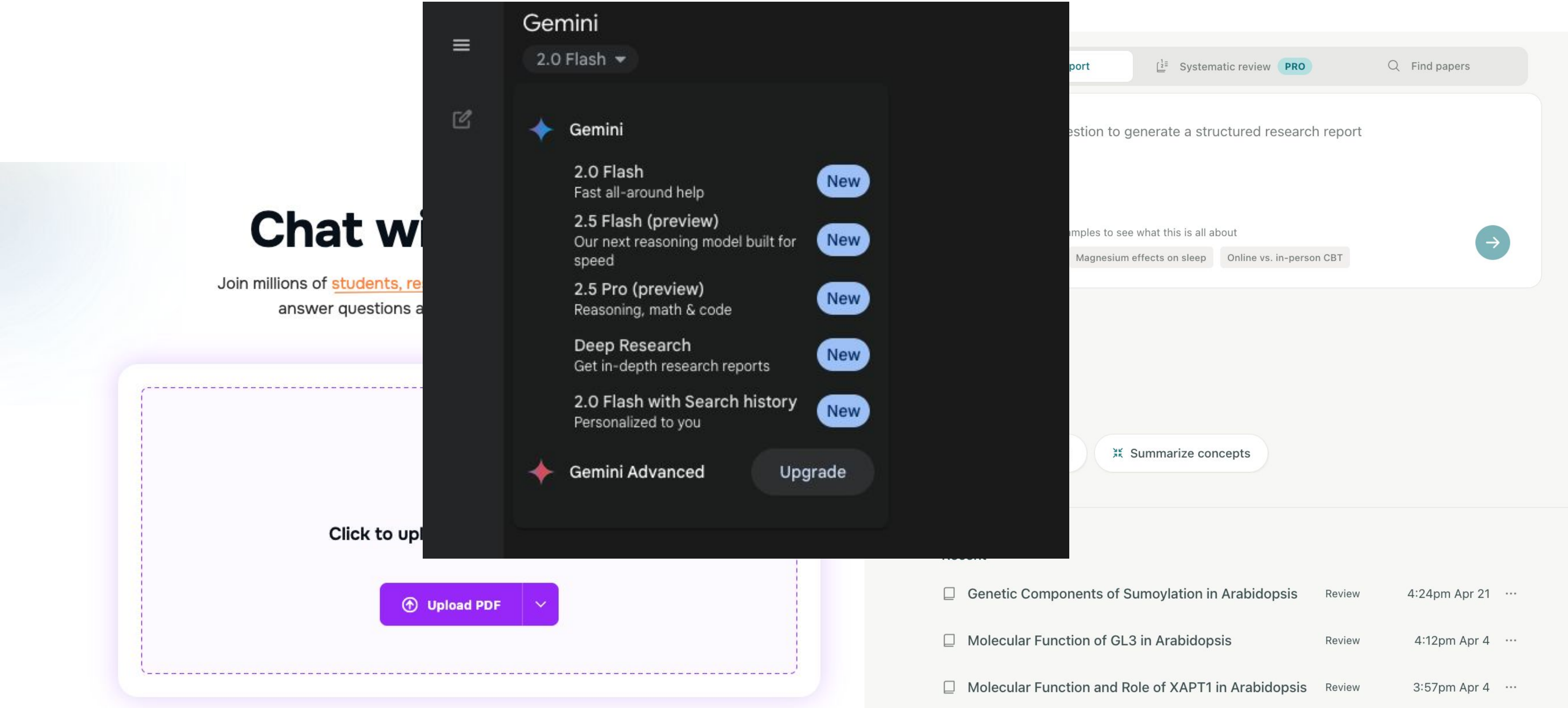
Tool Reviews

- **Adam**: React-to-Me Chatbot

- **Qi**: Recognizing Animal Traits in PubMed Abstracts
- **Edwin**: OntoGPT in AgBio
- **Kapeel**: ChatGPT - Custom GPT Model for Gene Information

- **Jacqueline**: ChatPDF
- **Tanya**: Elicit
- **Parul:** Gemini
- **Sudhansu:** Gemini, ChatPDF, MS CoPilot
- **Sri:** Google Notebook LLM

Online chat with a paper/set of papers



Chat wi

Join millions of students, re

answer questions a

Click to upl

Upload PDF

Gemini

2.0 Flash

Gemini

2.0 Flash
Fast all-around help
New

2.5 Flash (preview)
Our next reasoning model built for speed
New

2.5 Pro (preview)
Reasoning, math & code
New

Deep Research
Get in-depth research reports
New

2.0 Flash with Search history
Personalized to you
New

Gemini Advanced
Upgrade

port

Systematic review PRO

Find papers

estion to generate a structured research report

mples to see what this is all about

Magnesium effects on sleep

Online vs. in-person CBT

Summarize concepts

Genetic Components of Sumoylation in Arabidopsis	Review	4:24pm Apr 21
Molecular Function of GL3 in Arabidopsis	Review	4:12pm Apr 4
Molecular Function and Role of XAPT1 in Arabidopsis	Review	3:57pm Apr 4

Next steps

- Short term:
  - Continue evaluating tools
  - Start a live document of recommendations (What do I want to do? What can I use?)

- Longer term:
  - Identify a tool/pipeline that can be adapted/developed and shared across projects

You can still join.

Slack channel: #nlp4biocuration

Meeting schedule: Every other week, Wednesdays, 8 am Pacific / 10am Central /11 am Eastern / 4pm Central European Time

Next meeting: May 21, 2025