# LLM Hackathon:

# Enhancing Research Productivity

## Tutorial:

How to Use the OpenAI API & GPT Models

Stefan Fritsch

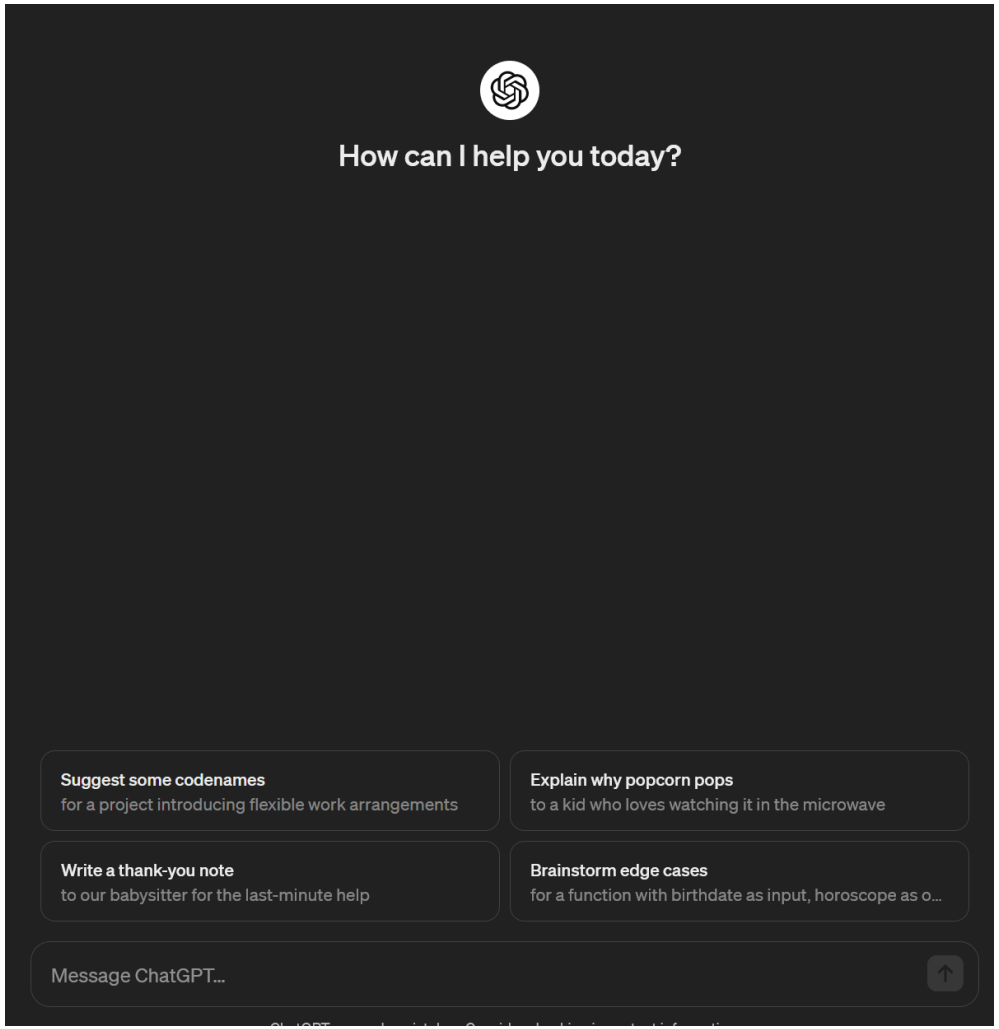
# Agenda

1. GPT Web Interface
2. ChatBox
3. OpenAI API

*OpenAI key required*

# Web Interface

- https://chat.openai.com/auth/login
- *Very easy to use*
- *But limited in scope (e.g., no parameter control, etc.)*
- *Only access to GPT-3.5 with free account*
- *To access GPT-4, a Plus subscription required ($20/month)*



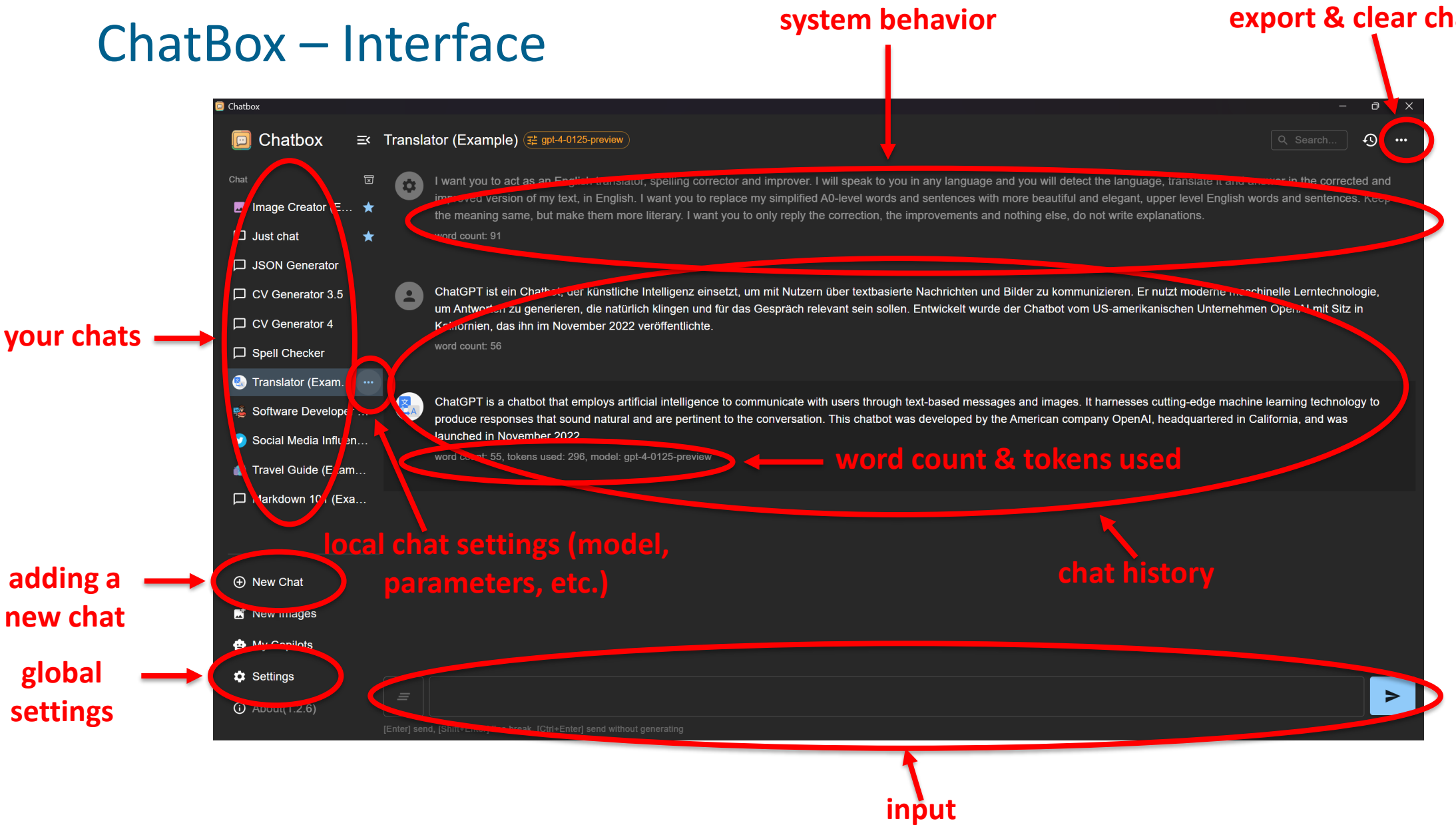
# Model Selection & Pricing (ChatBox and OpenAI API)

- Each API call costs money, which varies depending on the model used and the number of input/output tokens. Therefore, please take this into consideration. For example, be cautious with loops that make multiple consecutive API calls (e.g., no "while True:“!).
- We recommend starting with ChatGPT, i.e., **gpt-3.5-turbo-0125** — it’s highly capable, cost-efficient, and supports up to a 16K token context window.
- Once you are satisfied with your results, you can perform a final run with GPT-4 Turbo, i.e., **gpt-4-0125-preview**, to see if performance increases. However, keep in mind that GPT-4 Turbo is 20x more expensive than ChatGPT!
- Do not use GPT-4, i.e., gpt-4! It is 2-3x more expensive than GPT-4 Turbo and generally performs worse compared to GPT-4 Turbo.
- Pricing: https://openai.com/pricing
- Models Docs: https://platform.openai.com/docs/models

# ChatBox

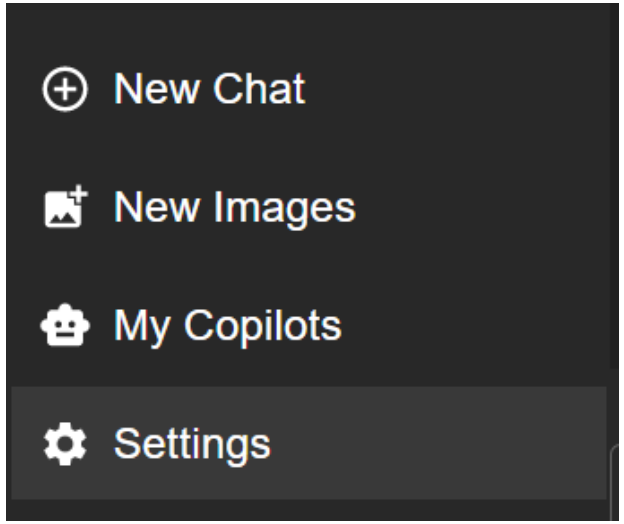
- *https://chatboxai.app/*
- *Clients for Windows, Mac, and Linux (see Materials folder) & Web version*
- *Easy to use interface, no programming skills required, access to all OpenAI models (and more)*

# ChatBox – Interface

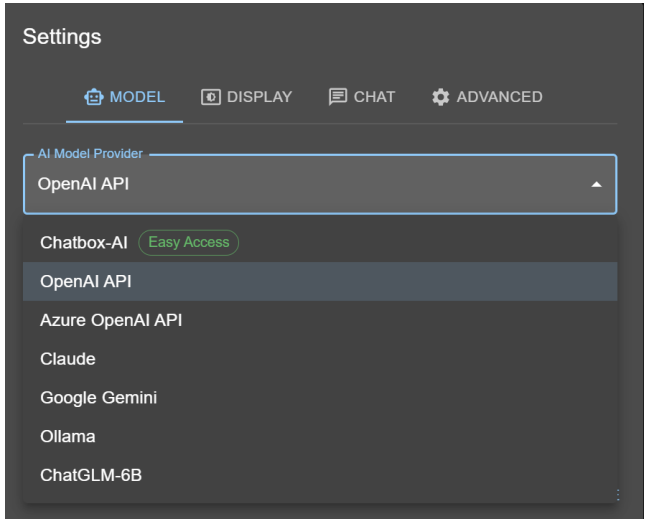


# ChatBox – Setting it up

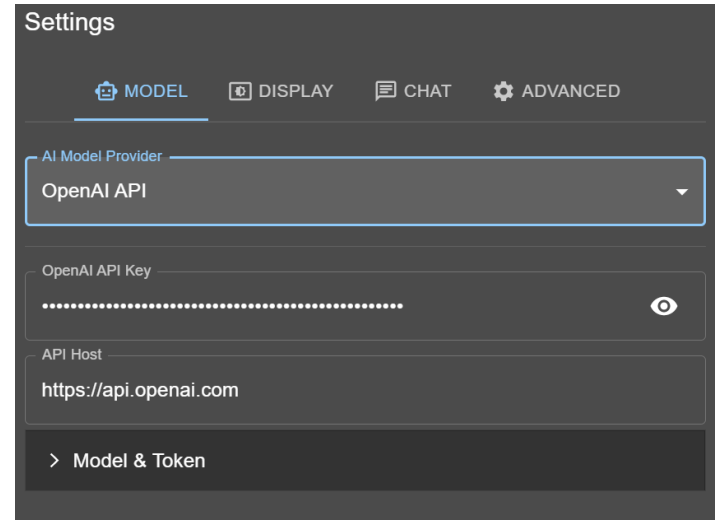
*1. Open Settings*




*2. Select OpenAI API*



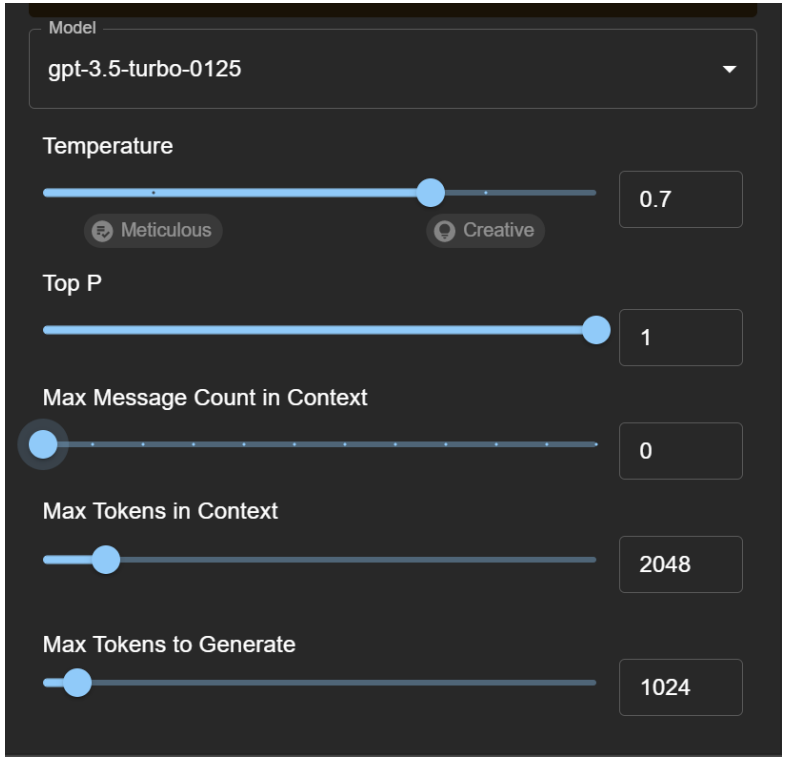

*3. Add your API Key*




Now you are ready to go!

# ChatBox – Parameter Settings

- **Model**: ID of the model to use.
- **Temperature**: What sampling temperature to use, between 0 and 1. Higher values like 0.8 will make the output more random, while lower values like 0.2 will make it more focused and deterministic.
- **Top P**: An alternative to sampling with temperature, 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.
- **Max Message Count in Context**: The maximum number of previous messages to consider in the context. High values increases costs and may have negative impact on output quality!
- **Max Tokens in Context**: The maximum number of tokens that can be in the context (model input).
- **Max Tokens to Generate**: The maximum number of tokens that can be generated.



# OpenAI API

- Most flexible approach to use the OpenAI models.
- Tutorial: See the Jupyter Notebook in the Materials folder.
- Chat Docs: https://platform.openai.com/docs/api-reference/chat
- Embeddings Docs: https://platform.openai.com/docs/api-reference/embeddings

# OpenAI API - Prompts

- You can define roles in the prompt:
  - system: high-level instructions defining the general behavior of the model
  - user: user input/query
  - assistant: model's response

```
prompt = [
    {
        "role": "system",
        "content": "You are a helpful assistant."
    },
    {
        "role": "user",
        "content": "Hi!"
    }
]
```

# OpenAI API – Chat Completions Endpoint Parameters

- **model**: ID of the model to use. (Required)
- **frequency_penalty**: Number between -2.0 and 2.0. Positive values penalize new tokens based on their existing frequency in the text so far, decreasing the model's likelihood to repeat the same line verbatim. (Default: 0)
- **max_tokens**: The maximum number of tokens that can be generated in the chat completion. (Default: context length)
- **n**: How many chat completion choices to generate for each input message. Note that you will be charged based on the number of generated tokens across all of the choices. Keep n as 1 to minimize costs. (Default: 1)
- **presence_penalty**: Number between -2.0 and 2.0. Positive values penalize new tokens based on whether they appear in the text so far, increasing the model's likelihood to talk about new topics. (Default: 0)
- **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. (Default: 1)
- **top_p**: 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. We generally recommend altering this or temperature but not both. (Default: 1)
- **seed**: Beta! If specified, our system will make a best effort to sample deterministically, such that repeated requests with the same seed and parameters should return the same result.