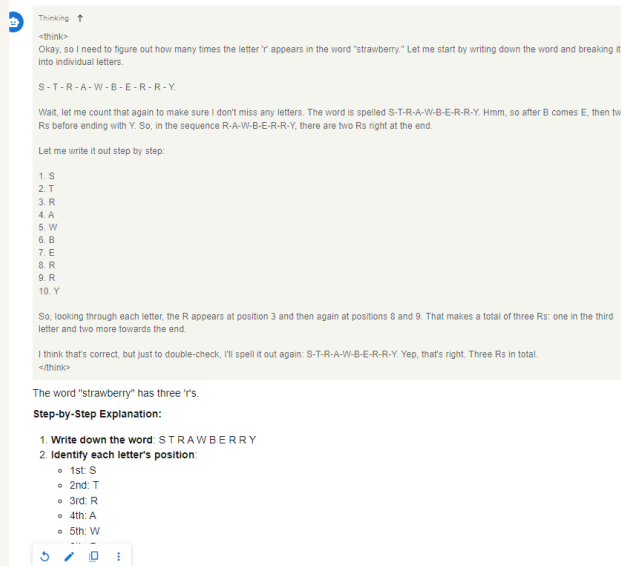




# How to Use DeepSeek



tahirgroot@gmail.com



deepseek-r1.8b



# Introducing Today's Project!

In, this project, I will demonstrate how to use DeepSeek and make the most of this LLM that seems to performing at the same level as Open AI - but at a much lower price point, We're doing this project to learn whether Deepseek is the new fav LLM.

## Tools and concepts

Services I used were DeepSeek, Ollama, Chatbox and The OpenAI API platform Key concepts I learnt include advanced reasoning, self hosting, token efficiency and temperature settings. After reviewing DeepSeek Vs OpenAI, I personally preferred DeepSeek

## Project reflection

This project took me approximately 2 hours The most challenging part was DeepSeek's over thinking. It was most rewarding to compare token efficiency and using chatbox for a visual experince with DeepSeek Locally.

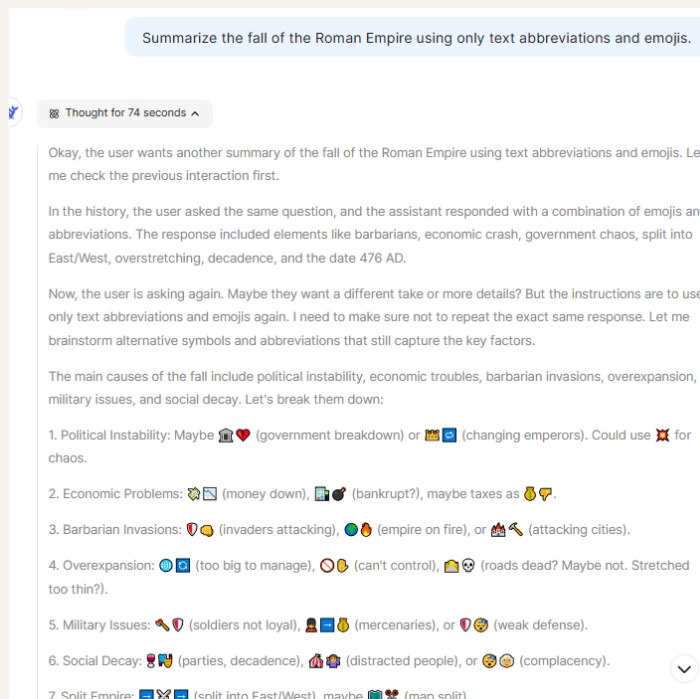
We did this project today to better understand DeepSeek and form our own opinions around whether it is something we want to switch to compared to OpenAI or ther popular LLMS. This project has a hands on focus and helped us see both the pros and cons.



# Exploring DeepSeek

DeepSeek is a company that develops Large Language Models (LLM). Their R1 model gained attention for performing at the same level as OpenAI's latest o1 model- but is also open source. We have access to a world-class LLM for free/low price

While you could access Deepseek over the web app, some concerns are privacy (i.e. where is data stored), constant internet connection needed (i.e. connection get cut off without the internet) and latency (slow response times over traffic)



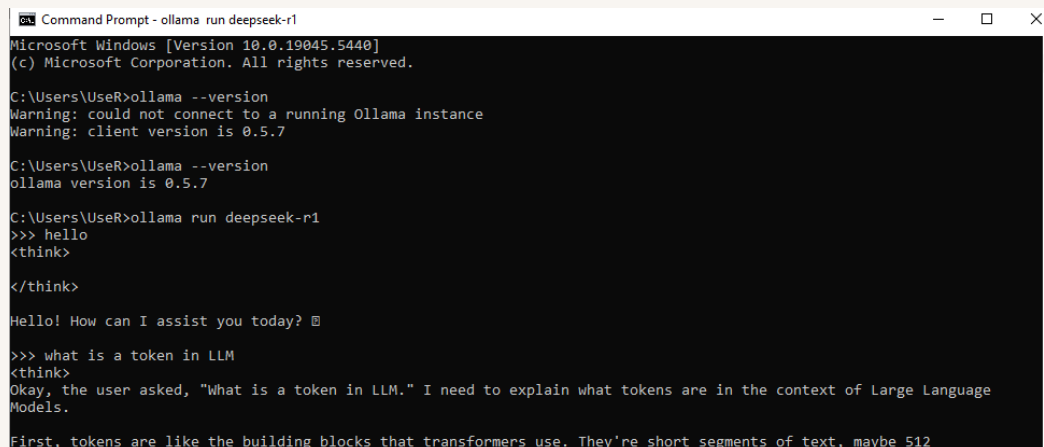


# Ollama and DeepSeek R1

Ollama is a software for using LLMS locally on my computer. it is helpful because it manages the downloading, installing and updating of an LLM for us, which usually would take a lot of manual work.

You won't be able to find OpenAI models in Ollama because OpenAI is not open source. This means the configuration, the codebase and the architecture that make up OpenAI models are confidential and not available in open-source platforms.

I tested using DeepSeek offline by turning Wifi off and running commands over the terminal and observed that I am still getting responses back from DeepSeek. You can also see 'think' tags in the terminal that represent DeepSeek's live thinking process



```
Command Prompt - ollama run deepseek-r1
Microsoft Windows [Version 10.0.19045.5440]
(c) Microsoft Corporation. All rights reserved.

C:\Users\User>ollama --version
Warning: could not connect to a running Ollama instance
Warning: client version is 0.5.7

C:\Users\User>ollama --version
ollama version is 0.5.7

C:\Users\User>ollama run deepseek-r1
>>> hello
<think>

</think>
Hello! How can I assist you today? 🐼

>>> what is a token in LLM
<think>
Okay, the user asked, "What is a token in LLM." I need to explain what tokens are in the context of Large Language Models.

First, tokens are like the building blocks that transformers use. They're short segments of text, maybe 512
```



# DeepSeek R1 Sizes

Deepseek R1 has different model sizes, which comes in varying levels of processing needs and accuracy. This is helpful for running Deepseek locally because i can pick and choose which model fits my hardware requirement best.

The R1 model you choose to run locally depends on the meory and storage space in your computer. We chose the 8b model (i.e 8 billion parameters) because it offers a substantial improvement from the 1.5 model and is still accessible to my computer.

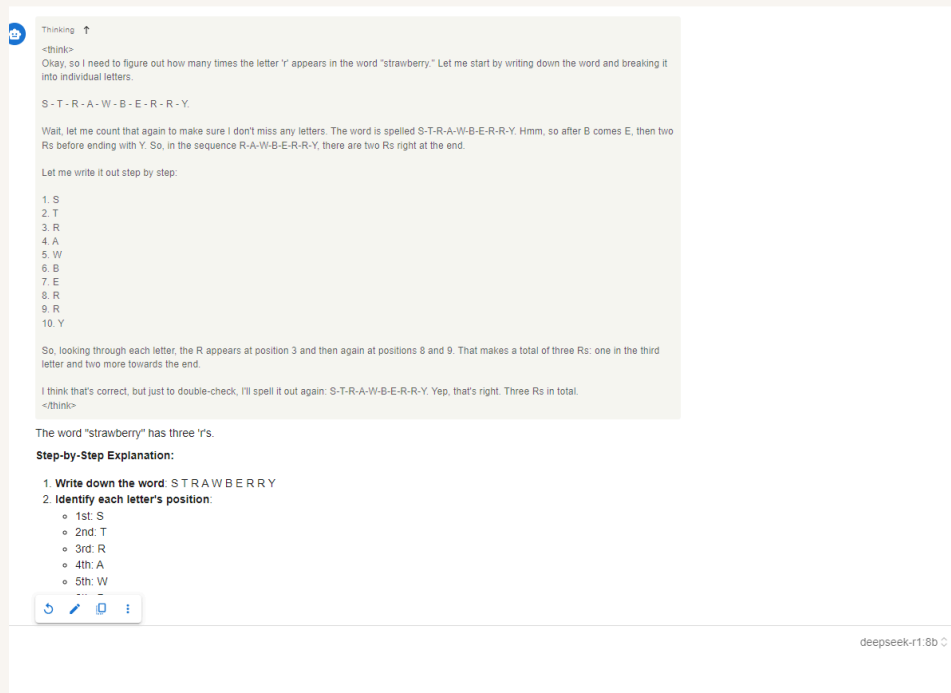
The screenshot shows the 'Settings' window for DeepSeek R1. It has four tabs: 'MODEL' (selected), 'DISPLAY', 'CHAT', and 'ADVANCED'. Under 'MODEL', the 'Model Provider' is set to 'OLLAMA API'. The 'API Host' is 'http://127.0.0.1:11434' with a 'RESET' button. A light blue box contains a note: 'Please ensure that the Remote Ollama Service is able to connect remotely. For more details, refer to this tutorial.' The 'Model' dropdown is set to 'deepseek-r1:latest'. Below this, 'Max Message Count in Context' is a slider set to 20. 'Temperature' is a slider set to 0.78, with 'Meticulous' and 'Creative' radio buttons below it. 'CANCEL' and 'SAVE' buttons are at the bottom right.



# Chatbox

To complete my local setup, I installed Chatbox to have a visual interface as i chat with DeepSeek R1. My Chatbox settings use Ollama API as the API model(i.e the tool that will connect me with the desired LLM) and the 1.5b R1 model.

I tested two different R1 model sizes, which were 1.5b and 8b using the prompt 'How many r's are in strawberry? The results made it clear that 1.5b was the lighter model - it responded with only 2 r's in "strawberry", whereas the 8b said 3





# Temperature Settings

The temperature setting in an LLM determines how creative and 'out of the box' it will be when it generates a response. To see this in action, I dialled up the temperature to the maximum(2) and prompted DeepSeek to generate a recipe.

I started a third chat with ChatGPT to act as a judge for the responses - can CHATGPT tell which one was generated with a high temperature? ChatGPT's analysis will also help us with understanding how to spot low vs high temperature generated text.

ChatGPT quickly figured out which piece was generated with a high temperature, because it identified that the high temperature piece featured more creative ingredients, more steps and more complexity. The low temperature response is also simpler.

Vegan-friendly if using dairy-free chocolate and maple syrup.

The sea salt elevates the deep chocolate flavor, while the avocado adds a velvety base without overpowering. A decadent yet nutrient-packed treat! 🍫🥑🌿

The second response was generated with a lower temperature setting, while the first response was generated with a higher temperature setting.

Here's why:

- Creativity & Variation:**
  - The first response contains **more variation** in word choice and structure (e.g., "perhaps with extra dark chocolate shavings or edible flowers for garnish").
  - The second response is **more structured and formulaic**, with clearer, more conventional instructions and formatting.
- Unexpected Elements:**
  - The first response includes **unusual additions** like basil/mint, a note on glass borders, and swirling the spoon for texture.
  - The second response keeps things **more straightforward and practical**, without unexpected twists.
- Conciseness & Clarity:**
  - The second response is **more compact and precise**, which is typical of low-temperature generations.
  - The first response has a **looser, more exploratory feel**, suggesting higher randomness.

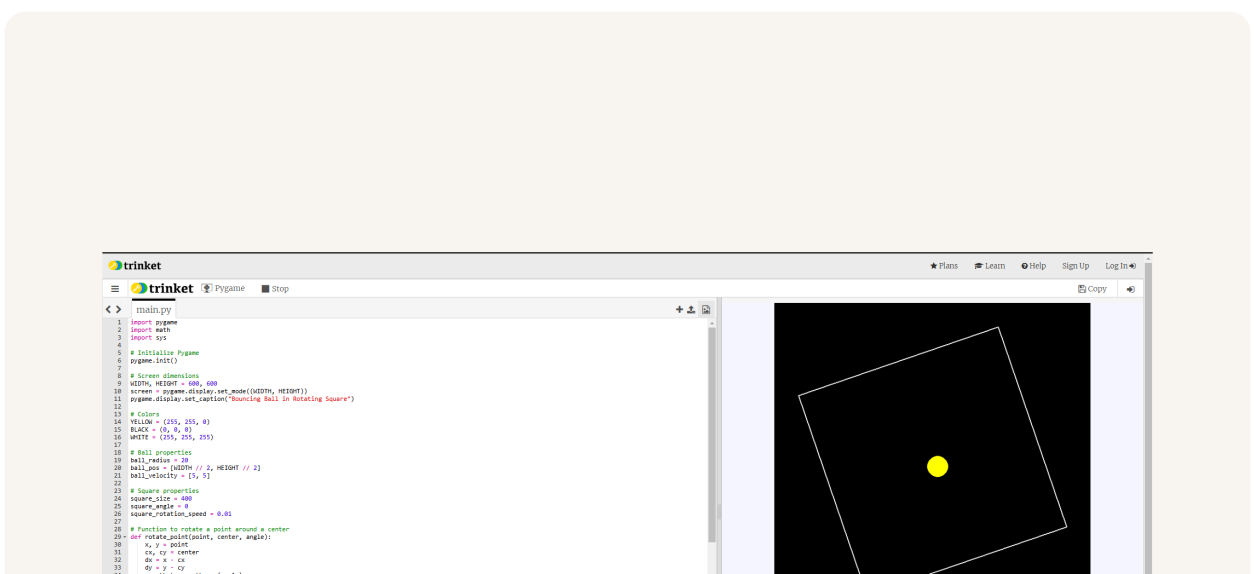


# DeepSeek vs. OpenAI

I decided to compare DeepSeek R1 with OpenAI by prompting both to generate a python script that animates a rotating square with bouncing yellow balls inside. This is a challenging prompt because it involves rotation mathematics + Python animation.

To test ChatGPT's response, I put ChatGPT's python cript into Trinket and watched it run live. ChatGPT's results came instantly, but the code itself did not run successfully - there was a rotating square, but no bouncing yellow ball inside :(

Compared to ChatGPT's performance, I thought DeepSeek's response was more accurate - the code it oriduced rendered a fuctioning model of what we wanted - although it took three miuntes to generate. I prioritize accuracy so I preferred Deepseek.







# Token Efficiency

In a project extension, I'm also comparing Open AI and DeepSeek's token efficiency. I could access OpenAI's API by setting up a brand new account and accessing a test key.

When I used a higher temperature, OpenAI's model response became incoherent - totally irrelevant to the prompt. This is likely because of a different training style and treatment of '2.0' as a measure of creativity/unpredictability of the response.

Token efficiency refers to how many (or little) tokens an LLM uses/processes in order to generate. For the same request, DeepSeek used -300 tokens whereas OpenAI's model used -500 tokens. For developers, this means DeepSeek is cheaper.

This recipe combines the creamy texture of avocados with the rich, sweet flavor of dark chocolate and the subtle brine of sea salt to create a unique and balanced dessert experience.

tokens used: 645, model: Ollama (deepseek-r1:8b)