# Generative Artificial Intelligence for Code Generation (1)

COMP4431 Lab 7

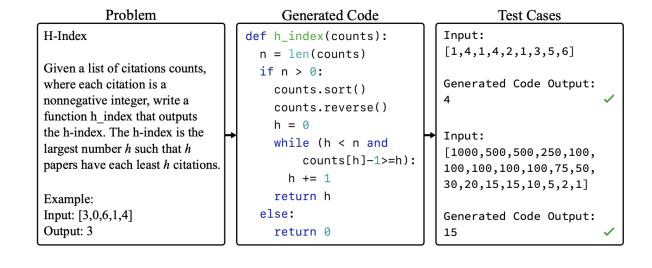
### Lab 07 Agenda

- Introduction to Generative Artificial Intelligence for Code Generation
  - Concepts
  - Leaderboard
  - Access
- Basic Techniques
  - Tab and Chat
  - Models
  - Context Management
  - Prompt Engineering
- Case Study

### Concepts

- Code Generation
  - Code Generation is an important field focused on predicting explicit code or program structures from multimodal data sources, including:
    - Incomplete code
    - Programs in other programming languages
    - Natural language descriptions
    - Execution examples
  - Code Generation tools can assist the development of automatic programming tools to improve programming productivity.

Ganesh, K. Srivatsa, et al. "A Survey of using Large Language Models for Generating Infrastructure as Code." Proceedings of the 20th International Conference on Natural Language Processing. 2023.

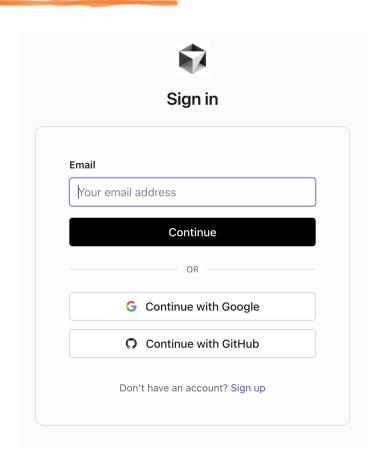


### Leaderboard

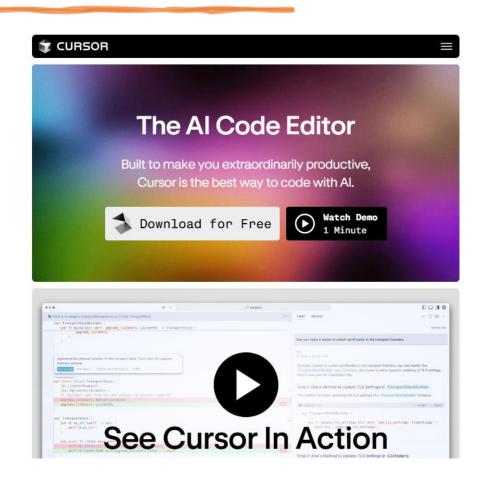
- Recent development in Code Generation
  - Having been trained on a large amount of text for predicting the next word, LLMs have shown remarkable performance in tasks related to code generation and validation.
- Leaderboard
  - BigCodeBench, a benchmark that challenges LLMs to invoke multiple function calls as tools from 139 libraries and 7 domains for 1,140 fine-grained tasks.
  - Pass@K evaluates the first k
    generated code's functional
    accuracy. K=1 evaluates if the first
    generated code can pass all the
    provided test cases.

| #  | Model                                      | Pass@1 |
|----|--|--------|
| 1  | ŏ <u>GPT-40-2024-05-13</u> +               | 56.1   |
| 2  | <u>ŏ DeepSeek-V2-Chat (2024-06-28)</u> ;+  | 54.1   |
| 3  | <u>ŏDeepSeek-Coder-V2-Instruct</u> ;       | 54     |
| 4  | GPT-4-Turbo-2024-04-09;+                   | 53.2   |
| 5  | Claude-3.5-Sonnet-20240620;+               | 52.7   |
| 6  | GPT-4o-mini-2024-07-18;+                   | 51.8   |
| 7  | GPT-4-0613;+                               | 51.6   |
| 8  | Claude-3-0pus-20240229;+                   | 51.5   |
| 9  | Qwen2.5-72B-Instruct'+                     | 50.8   |
| 10 | Gemini-1.5-Pro-API-0514;→                  | 50.6   |
| 10 | Hermes-2-Theta-Llama-3-70B <sup>1</sup> ;+ | 50.6   |

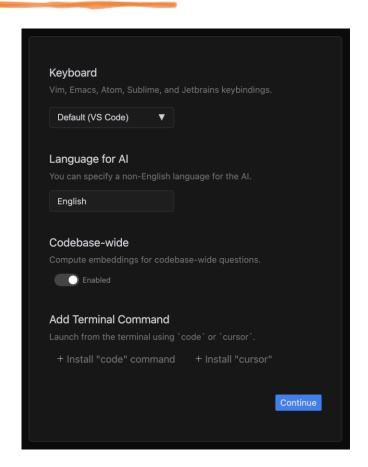
- Step-by-step guide to access to Cursor
  - Navigate to the official Cursor AI website https://www.cursor.com/
  - Create an account
    - Hobby (free)
      - 14 days Pro trial; 50 slow premium model uses; 200 cursorsmall uses; 2000 completions uses
    - Pro (\$20.00/month)
      - 500 fast premium model uses per month; Unlimited slow premium model uses; Unlimited cursor-small uses; Unlimited completions completions; 10 Claude Opus uses per month
  - Download the appropriate version for your operating system
  - Run the installer and follow the on-screen instructions
  - Launch Cursor AI and complete any additional setup prompts



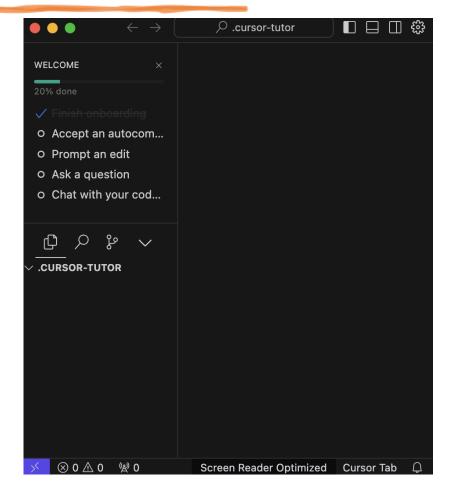
- Step-by-step guide to access to Cursor
  - Navigate to the official Cursor AI website https://www.cursor.com/
  - Create an account
    - Hobby (free)
      - 14 days Pro trial; 50 slow premium model uses; 200 cursor-small uses; 2000 completions uses
    - Pro (\$20.00/month)
      - 500 fast premium model uses per month; Unlimited slow premium model uses; Unlimited cursorsmall uses; Unlimited completions completions; 10 Claude Opus uses per month
  - Download the appropriate version for your operating system
  - Run the installer and follow the on-screen instructions
  - Launch Cursor AI and complete any additional setup prompts



- Step-by-step guide to access to Cursor
  - Navigate to the official Cursor AI website https://www.cursor.com/
  - Create an account.
    - Hobby (free)
      - 14 days Pro trial; 50 slow premium model uses; 200 cursorsmall uses; 2000 completions uses
    - Pro (\$20.00/month)
      - 500 fast premium model uses per month; Unlimited slow premium model uses; Unlimited cursor-small uses; Unlimited completions completions; 10 Claude Opus uses per month
  - Download the appropriate version for your operating system
  - Run the installer and follow the on-screen instructions
  - Launch Cursor AI and complete any additional setup prompts



- Step-by-step guide to access to Cursor
  - Navigate to the official Cursor AI website https://www.cursor.com/
  - Create an account
    - Hobby (free)
      - 14 days Pro trial; 50 slow premium model uses; 200 cursor-small uses; 2000 completions uses
    - Pro (\$20.00/month)
      - 500 fast premium model uses per month; Unlimited slow premium model uses; Unlimited cursorsmall uses; Unlimited completions completions; 10 Claude Opus uses per month
  - Download the appropriate version for your operating system
  - Run the installer and follow the on-screen instructions
  - Launch Cursor AI and complete any additional setup prompts



### Lab 07 Agenda

- Introduction to Generative Artificial Intelligence for Code Generation
  - Concepts
  - Leaderboard
  - Access
- Basic Techniques
  - Tab and Chat
  - Models
  - Context Management
  - Prompt Engineering
- Case Study

#### Cursor Tab

- Cursor Tab can suggest edits around your cursor position, including:
  - Insertions of additional code
  - Suggestions based on your recent changes
  - Recommendations based on linter errors

#### Inline Generation

 Cursor Chat allows you to ask questions or solve problems within your codebase using advanced language models, all within your editor. When no code is selected and you press Ctrl/Cmd K, Cursor will generate new code based on your input in the prompt bar.

#### Inline Edits

• For in-place edits, simply select the code you want to modify and enter your instructions in the prompt bar.

#### Cursor terminal

• In the built-in terminal, pressing Ctrl/\(\mathbb{H}\) K opens a prompt bar at the bottom of the terminal.

```
two_sum.py
     class Solution(object):
         def twoSum(self, nums, target):
             :type nums: List[int]
             :type target: int
             :rtype: List[int]
14
             for i in range(len(nums)):
                     if nums[i] + nums[j] == target:
```

#### Cursor Tab

Cursor Tab can suggest edits around your cursor position.

#### Inline Generation

- Cursor Chat allows you to ask questions or solve problems within your codebase using advanced language models, all within your editor. When no code is selected and you press Ctrl/Cmd K, Cursor will generate new codes based on your input in the prompt bar.
  - Write test cases to test the twoSum function.

#### Inline Edits

• For in-place edits, simply select the code you want to modify and enter your instructions in the prompt bar.

#### Cursor terminal

• In the built-in terminal, pressing Ctrl/\mathbb{H} K opens a prompt bar at the bottom of the terminal.

```
Write test cases to test the twoSum function
  第← Accept #図 Reject Follow-up instructions... 公策K
def test twoSum():
    solution = Solution()
   assert solution.twoSum([2, 7, 11, 15], 9) == [0, 1]
   assert solution.twoSum([3, 2, 4], 6) == [1, 2]
   assert solution.twoSum([3, 3], 6) == [0, 1]
   assert solution.twoSum([1, 2, 3, 4], 10) == []
   assert solution.twoSum([-1, -2, -3, -4, -5], -8) == [2, 4]
   print("All test cases passed!")
test twoSum()
```

#### Cursor Tab

• Cursor Tab can suggest edits around your cursor position.

#### Inline Generation

 Cursor Chat allows you to ask questions or solve problems within your codebase using advanced language models, all within your editor. When no code is selected and you press Ctrl/Cmd K, Cursor will generate new codes based on your input in the prompt bar.

#### Inline Edits

- For in-place edits, simply select the code you want to modify and enter your instructions in the prompt bar.
  - Use a hash table. We can iterate through the array once, and for each element, check if the target minus the current element exists in the hash table. If it does, we have found a valid pair of numbers. If not, we add the current element to the hash table.

#### Cursor terminal

• In the built-in terminal, pressing  $Ctrl/\Re$  K opens a prompt bar at the bottom of the terminal.

```
two_sum.py
       indices of the two numbers such that they add up to target.
     \vee class Solution(object):
             Use a hash table. We can iterate through the array once, and for each
             element, check if the target minus the current element exists in the hash
             table. If it does, we have found a valid pair of numbers. If not, we add the
             current element to the hash table.
             #← Accept #図 Reject Follow-up instructions... 公業K
           def twoSum(self, nums, target):
               :type nums: List[int]
                :type target: int
               :rtype: List[int]
                for i in range(len(nums)):
                    for j in range(i + 1, len(nums)):
                        if nums[i] + nums[j] == target:
                            return [i, j]
                num dict = {}
                for i, num in enumerate(nums):
                    complement = target - num
                    if complement in num_dict:
                        return [num_dict[complement], i]
                    num dict[num] = i
               return []
```

#### Cursor Tab

• Cursor Tab can suggest edits around your cursor position.

#### Inline Generation

 Cursor Chat allows you to ask questions or solve problems within your codebase using advanced language models, all within your editor. When no code is selected and you press Ctrl/Cmd K, Cursor will generate new code based on your input in the prompt bar.

#### Inline Edits

• For in-place edits, simply select the code you want to modify and enter your instructions in the prompt bar.

#### Cursor terminal

- In the built-in terminal, pressing Ctrl/\(\mathbb{H}\) K opens a prompt bar at the bottom of the terminal. You can:
  - Accept the command by pressing Esc
  - Execute the command immediately using Ctrl/ $\mathbb{H}$  + Enter
- Run the code and log to two\_sum.log

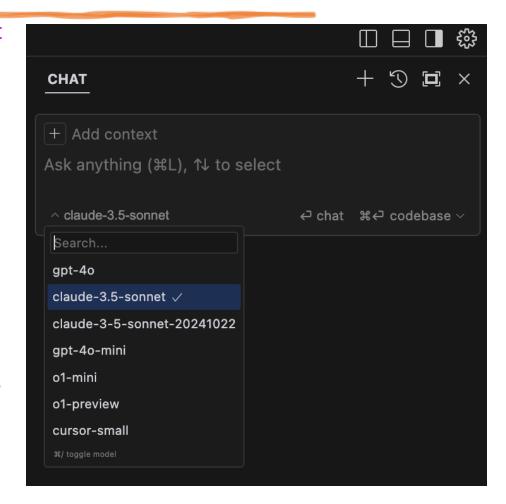
```
# Test case 3: Same number used
       twice
            assert solution.twoSum([3, 3], 6) ==
        [0, 1]
TERMINAL
 (base)
         geovanni@Mac-mini-2
 (base) geovanni@Mac-mini-2
                               ~/Work/COMP4431
python two sum.py > two sum.log 2>&1
Run the code and log to two_sum.log
          \C \to quick

← Submit

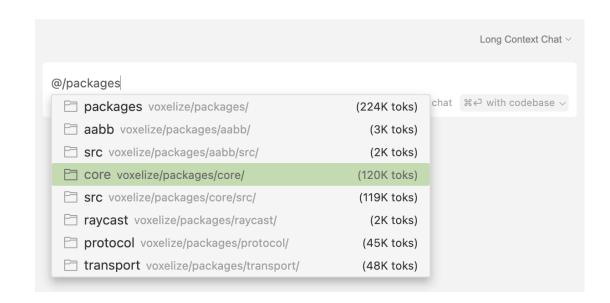
                        ∨ claude-3.5-s...
                                      業K to toggle focus
```

### Models

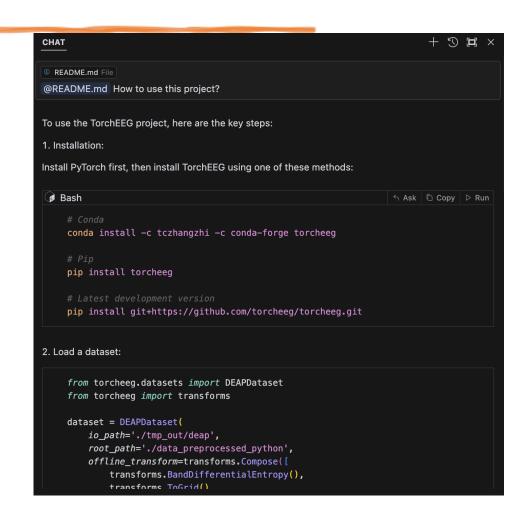
- Underneath the AI input box, you will see a dropdown that allows you to select the model you want to use.
- By default, Cursor has these models ready to use:
  - gpt-4o
  - gpt-4
  - claude 3.5 Sonnet
  - cursor-small
- In long context chat, the model selection is limited to the models that support long context:
  - gpt-4o-128k
  - gemini-1.5-flash-500k
  - claude-3-haiku-200k
  - claude-3-sonnet-200k
  - claude-3-5-sonnet-200k
- You can also set your own Azure API key so that you will be using Azure OpenAI models at your own cost.
  - https://docs.cursor.com/advanced/api-keys



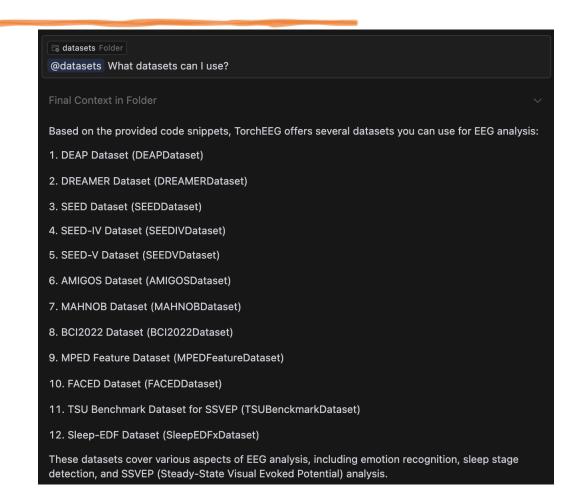
- For language models to give good answers, they need to know specific things that are relevant to your codebase — context.
  - Cursor has several built-in features to provide context in chat, such as:
    - automatically including context across your entire codebase
    - searching the web
    - indexing documentation, and user-specified references to code blocks
  - You can use @ symbols by typing @. A popup menu will appear with a list of suggestions, and it will automatically filter to only show the most relevant suggestions based on your input.
    - @Files
    - @Folders
    - @Code
    - @Docs
    - @Web



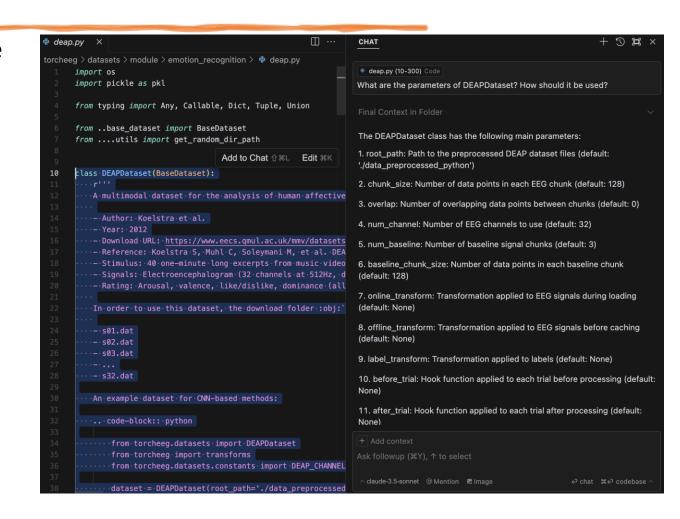
- Cursor has several built-in features to provide context.
   You can use @ symbols by typing @.
  - @Files
    - you can reference entire files by using @Files.
  - @Folders
    - You can reference entire folders by using @ Folders.
  - @Code



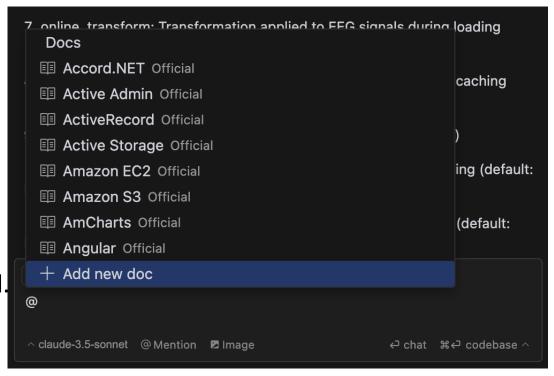
- Cursor has several built-in features to provide context. You can use @ symbols by typing @.
  - @Files
    - you can reference entire files by using @Files.
  - @Folders
    - You can reference entire folders by using @ Folders.
  - @Code
    - You can select the code you want to reference, and click on Add to Chat (Ctrl/\mathbb{H} Shift L)



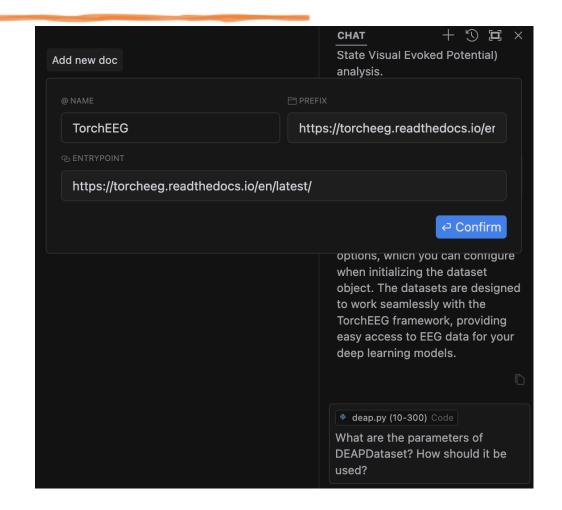
- Cursor has several built-in features to provide context. You can use @ symbols by typing @.
  - @Files
    - you can reference entire files by using @Files.
  - @Folders
    - You can reference entire folders by using @ Folders.
  - @Code
    - You can select the code you want to reference, and click on Add to Chat (Ctrl/\mathbb{H} Shift L)



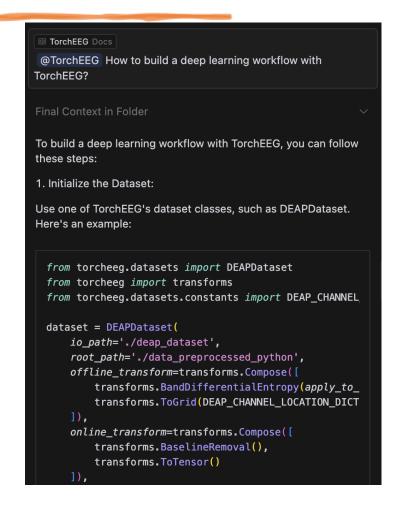
- Cursor has several built-in features to provide context. You can use @symbols by typing @.
  - Cursor comes with a set of third-party docs crawled, indexed, and ready to be used as context.
  - If you want to crawl and index custom docs that are not already provided, you can do so by @Docs > Add new doc.
  - The modal will appear after you've pasted in the URL of your desired doc.
  - You can access them by using the @Docs symbol.



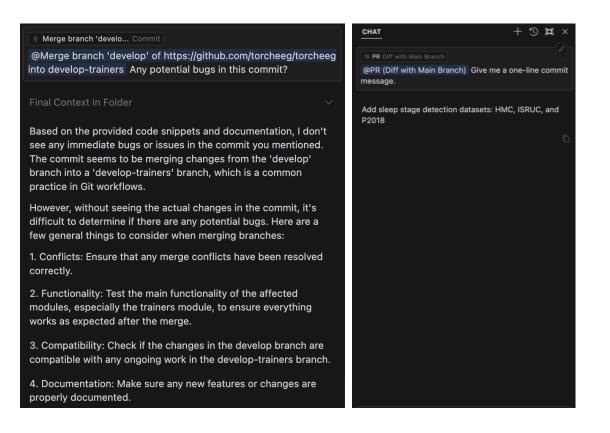
- Cursor has several built-in features to provide context.
   You can use @symbols by typing @.
  - Cursor comes with a set of third-party docs crawled, indexed, and ready to be used as context.
  - If you want to crawl and index custom docs that are not already provided, you can do so by @Docs > Add new doc.
  - The modal will appear after you've pasted in the URL of your desired doc.
  - You can access them by using the @Docs symbol.



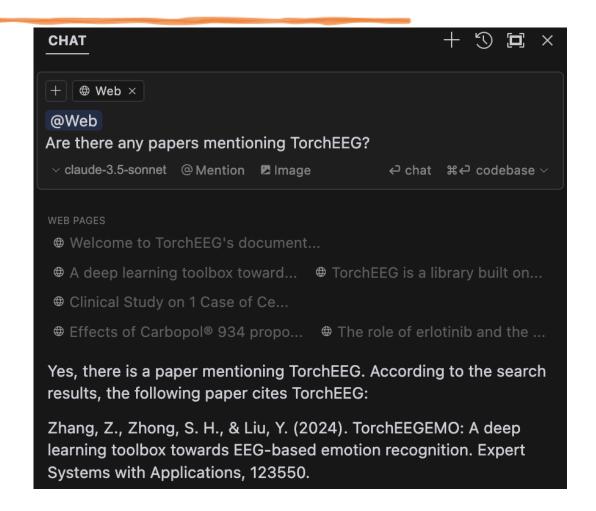
- Cursor has several built-in features to provide context. You can use @symbols by typing @.
  - Cursor comes with a set of third-party docs crawled, indexed, and ready to be used as context.
  - If you want to crawl and index custom docs that are not already provided, you can do so by @Docs > Add new doc.
  - The modal will appear after you've pasted in the URL of your desired doc.
  - You can access them by using the @Docs symbol.



- Cursor has several built-in features to provide context. You can use @symbols by typing @.
  - Cursor comes with a set of third-party docs crawled, indexed, and ready to be used as context.
  - If you want to crawl and index custom docs that are not already provided, you can do so by @Docs > Add new doc.
  - The modal will appear after you've pasted in the URL of your desired doc.
  - You can access them by using the @Docs symbol.



- Cursor has several built in features to provide context. You can use @ symbols by typing @.
  - With @Web, Cursor constructs a search query based on the query and the context you've provided, and searches the web to find relevant information as additional context.
  - This is particulary useful for finding the most up-to-date information.



### Prompt Engineering

- Start general, then get specific.
  - When writing a prompt for LLM coder, first give LLM coder a broad description of the goal or scenario. Then list any specific requirements.
    - Write a function that tells me if a number is prime
    - The function should take an integer and return true if the integer is prime
    - The function should error if the input is not a positive integer
- Use examples to help LLM coder understand what you want.
  - You can provide example input data, example outputs, and example implementations.
    - Write a function that finds all dates in a string and returns them in an array.
    - Example:
    - findDates("I have a dentist appointment on 11/14/2023 and book club on 12-1-23")
    - Returns: ["11/14/2023", "12-1-23"]
- Break the process down into smaller tasks, and ask LLM coder to accomplish them one by one.

```
Write a function that tells me if a number is prime.
The function should take an integer and return true if the integer is prime.
The function should error if the input is not a positive integer.
Certainly! I'll write a Python function that determines if a number is prime and
meets the requirements you specified. Here's the implementation:
  def is_prime(n):
      Check if a number is prime.
      n (int): The number to check.
      Returns:
      bool: True if the number is prime, False otherwise.
      Raises:
      ValueError: If the input is not a positive integer.
      if not isinstance(n, int) or n <= 0:</pre>
           raise ValueError("Input must be a positive integer.")
      if n == 1:
           return False # 1 is not considered prime
      if n == 2:
```

### Prompt Engineering

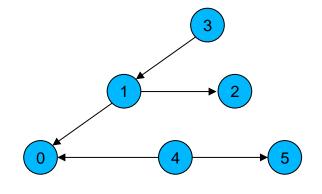
- Start general, then get specific.
  - When writing a prompt for LLM coder, first give LLM coder a broad description of the goal or scenario. Then list any specific requirements.
    - Write a function that tells me if a number is prime
    - The function should take an integer and return true if the integer is prime
    - The function should error if the input is not a positive integer
- Use examples to help LLM coder understand what you want.
  - You can provide example input data, example outputs, and example implementations.
    - Write a function that finds all dates in a string and returns them in an array.
    - Example:
    - findDates("I have a dentist appointment on 11/14/2023 and book club on 12-1-23")
    - Returns: ["11/14/2023", "12-1-23"]
- Break the process down into smaller tasks, and ask LLM coder to accomplish them one by one.

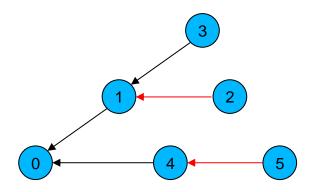
```
Write a function that finds all dates in a string and returns them in an
array.
Example:
findDates("I have a dentist appointment on 11/14/2023 and book club on
12-1-23")
Returns: ["11/14/2023", "12-1-23"]
Here's a Python function that finds all dates in a string and returns them
in an array:
  import re
  def findDates(text):
      # Regular expression pattern to match various date for
      date_pattern = r'\b(\d{1,2}[-/]\d{1,2}[-/]\d{2,4}]\d{4}
      # Find all matches in the text
      dates = re.findall(date_pattern, text)
       return dates
```

### Lab 07 Agenda

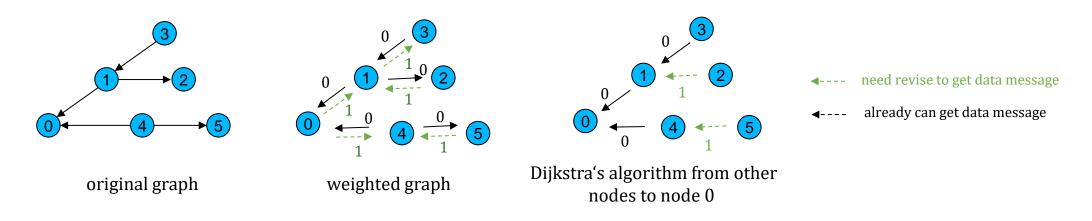
- Introduction to Generative Artificial Intelligence for Code Generation
  - Concepts
  - Leaderboard
  - Access
- Basic Techniques
  - Tab and Chat
  - Models
  - Context Management
  - Prompt Engineering
- Case Study

- The Information Center has set up a network to allow data uploads between servers. This network can be represented by a set of permitted directed transmissions, where a permitted directed transmission from server  $a_i$  to server  $b_i$  can be formulated as  $[a_i, b_i]$ . It's given that  $0 \le a_i, b_i < n, a_i \ne b_i$ . There are  $n \ (n \ge 2)$  servers in the network with with server 0 being the central server. The number of directed transmissions is less than n.
- The Information Center plans to update this network. The objective is to ensure the central server (0) can receive data uploads from every other server in the network. Your task is to reorient some of these permitted transmissions so that the central server can receive data from other servers. Return the minimum number of transmission changes required.
- If it's not feasible to enable the central server to receive data from other servers by adjusting the directions of existing permitted transmissions, then return -1.
- Example:
  - Input: transmissions = [[1, 0], [3, 1], [1, 2], [4,0], [4,5]], num\_servers=6
  - Expected output: 2
  - Transmission changes are demonstrated in red lines to ensure the central server can receive data from all the other servers.





- Break the process down into smaller tasks, and ask LLM coder to accomplish them one by one.
  - Initialize a directed graph
  - Add nodes for each server
  - Add directed edges based on transmissions. Assign a weight of 0 to all edges, indicating indicating no change
    is needed, for the edge[0] to receive from edge[1]. Assign reverse edges with a weight of 1, indicating a
    change is needed, to allow the edge[0] to receive from the edge[1]
  - Determine the shortest paths from other servers to node 0
  - Collect edges in the shortest paths and their associated weights (indicating change cost)



- Create a template to define function inputs and outputs. Follow Google Python Style Guide for input/output descriptions (https://google.github.io/styleguide/pyguide.html)
  - Args: List each parameter by name.
  - Returns: Describe the semantics of the return value, including any type information that the type annotation does not provide.
  - Raises: List all exceptions that are relevant to the interface followed by a description.
- Use Python type hints for input/output data types:
  - Type annotations provide hints to tools but are not runtime enforced.
  - Use typing module for complex type annotation.s (https://docs.python.org/3/library/typing.html)
- Never forget to write case studies.

```
transmissions = [[1, 0], [3, 1], [1, 2], [4,0], [4,5]]
num_servers = 6
answer = 2
result = solve(transmissions, num_servers)
assert result == answer, f"Test case 1: expected {answer}, got {result}"
```

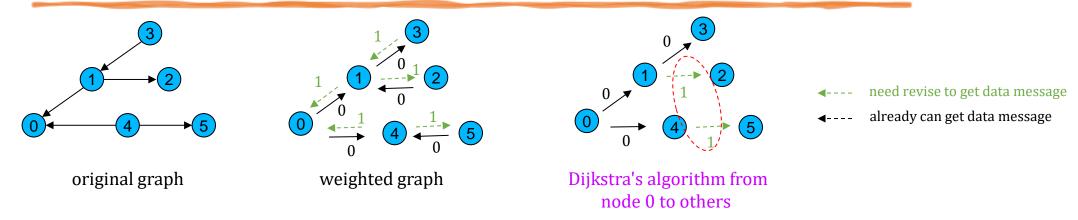
```
import networkx as nx
from typing import List
def solve(transmissions: List[List[int]], num_servers: int) -> int:
    Your solution to the problem goes in this function.
       transmissions (List[List[int]]): The permitted transmission between servers,
       e.g., [[1, 0], [3, 1], [1, 2], [4,0], [4,5]]
       num_servers (int): The number of servers in the network, e.g., 6
        int: the minimum number of transmission changes required, e.g., 2
    return -2
transmissions = [[1, 0], [3, 1], [1, 2], [4, 0], [4, 5]]
answer = 2
result = solve(transmissions, num_servers)
assert result == answer, f"Test case 1: expected {answer}, got {result}"
print('Passed test case 1...')
transmissions = [[1, 0], [3, 1], [2, 3]]
num_servers = 4
answer = 0
result = solve(transmissions, num_servers)
assert result == answer, f"Test case 2: expected {answer}, got {result}"
print('Passed test case 2...')
transmissions = [[1, 0], [0, 2], [3, 4]]
num servers = 5
result = solve(transmissions, num_servers)
assert result == answer, f"Test case 3: expected {answer}, got {result}"
print('Passed test case 3...')
```

- Inline Edits on the template
  - Initialize a directed graph
  - Add nodes for each server
  - Add directed edges based on transmissions, Assign a
    weight of 0 to all edges, indicating indicating no change
    is needed, for the edge[0] to receive from edge[1].
    Assign reverse edges with a weight of 1, indicating a
    change is needed, to allow the edge[0] to receive from
    the edge[1].
  - Determine the shortest paths from other servers to node 0
  - Collect edges in the shortest paths and their associated weights (indicating change cost)
- Inline Edits on the generated codes
  - Use a set to record the edges that need to be flipped to avoid an edge being recorded multiple times

```
the 'edge[1]' Determine the shortest paths from other servers to node 0 Collect
edges in the shortest paths and their associated weights (indicating change cost)
#← Accept #® Reject Follow-up instructions... 公業K
 G = nx.DiGraph()
 for i in range(num_servers):
     G.add node(i)
 for edge in transmissions:
     G.add_edge(edge[0], edge[1], weight=0)
     G.add_edge(edge[1], edge[0], weight=1)
 total changes = 0
 unreachable = False
 for node in range(1, num_servers):
          path = nx.shortest_path(G, node, 0, weight='weight')
          path_cost = sum(G[path[i]][path[i+1]]['weight'] for i in
          total_changes += path_cost
      except nx.NetworkXNoPath:
          unreachable = True
          break
 return -1 if unreachable else total_changes
```

- Inline Edits on the template
  - Initialize a directed graph
  - Add nodes for each server
  - Add directed edges based on transmissions. Assign a
    weight of 0 to all edges, indicating indicating no change
    is needed, for the edge[0] to receive from edge[1].
    Assign reverse edges with a weight of 1, indicating a
    change is needed, to allow the edge[0] to receive from
    the edge[1].
  - Determine the shortest paths from other servers to node 0
  - Collect edges in the shortest paths and their associated weights (indicating change cost)
- Inline Edits on the generated codes
  - Use a set to record the edges that need to be flipped to avoid an edge being recorded multiple times

```
Use a set to record the edges that need to be flipped to avoid an edge being
     recorded multiple times
                          Follow-up instructions... 公眾K
      %← Accept % Reject
   total_changes = 0
   edges_to_flip = set()
   unreachable = False
   for node in range(1, num_servers):
       try:
           path = nx.shortest_path(G, node, 0, weight='weight')
           path_cost = sum(G[path[i]][path[i+1]]['weight'] for i in
ange(len(path)-1))
           total changes += path cost
           for i in range(len(path)-1):
               if G[path[i]][path[i+1]]['weight'] == 1:
                    edges_to_flip.add((path[i], path[i+1]))
       except nx.NetworkXNoPath:
           unreachable = True
           break
   return -1 if unreachable else total_changes
  return -1 if unreachable else len(edges_to_flip)
```



- [Extensive Reading] The algorithm can be further optimized into a faster version:
  - Initialize a directed graph
  - Add nodes for each server
  - Assign a weight of 1 to all edges, indicating a change is needed, to allow the edge[0] to receive from the
    edge[1]
  - Add reverse edges with a weight of 0, indicating no change is needed, for the `edge[0]` to receive from edge[1]
  - Determine the shortest paths from node 0 to other servers
  - Collect edges in the shortest paths and their associated weights (indicating change cost)

## Question Time!

Or See you next week!