

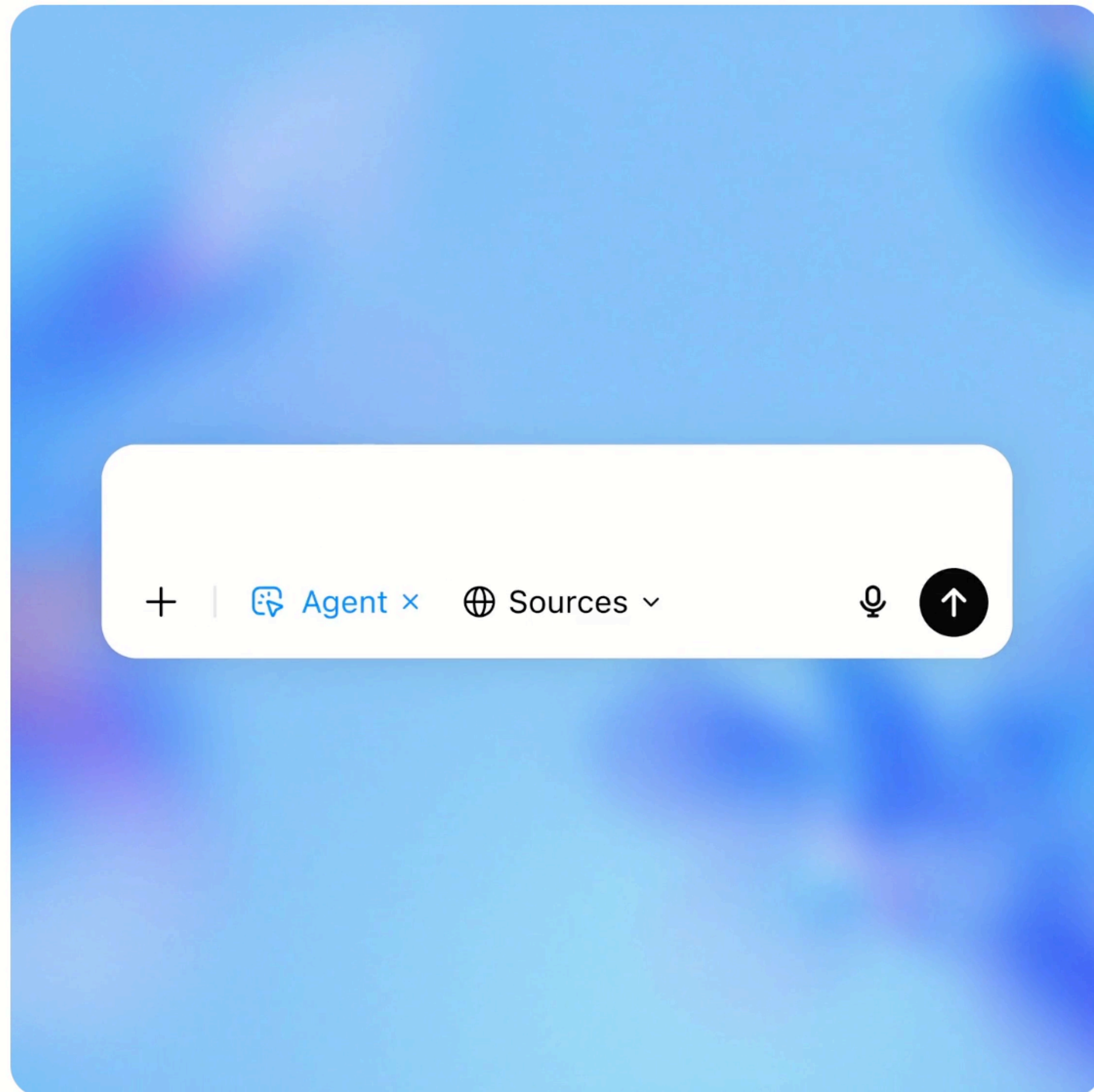
# Exploration and Bias in Adaptive LLMs

SPSP Symposium on *Social and Personality Psychology's AI Moment*

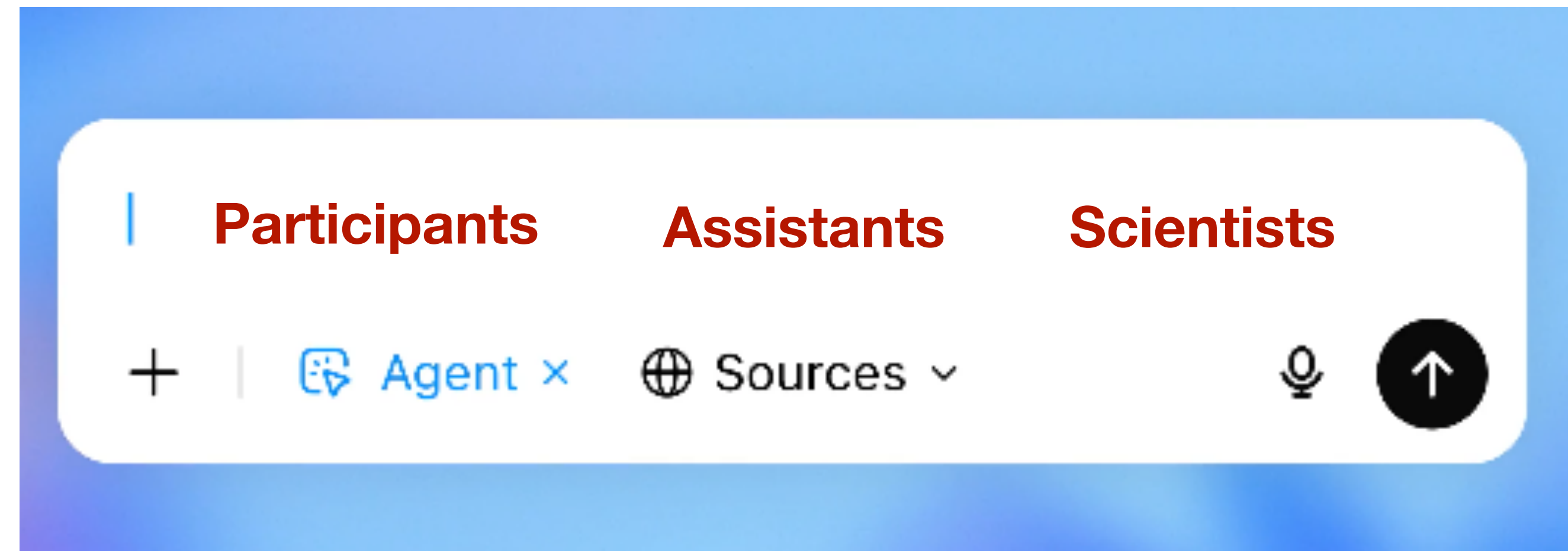
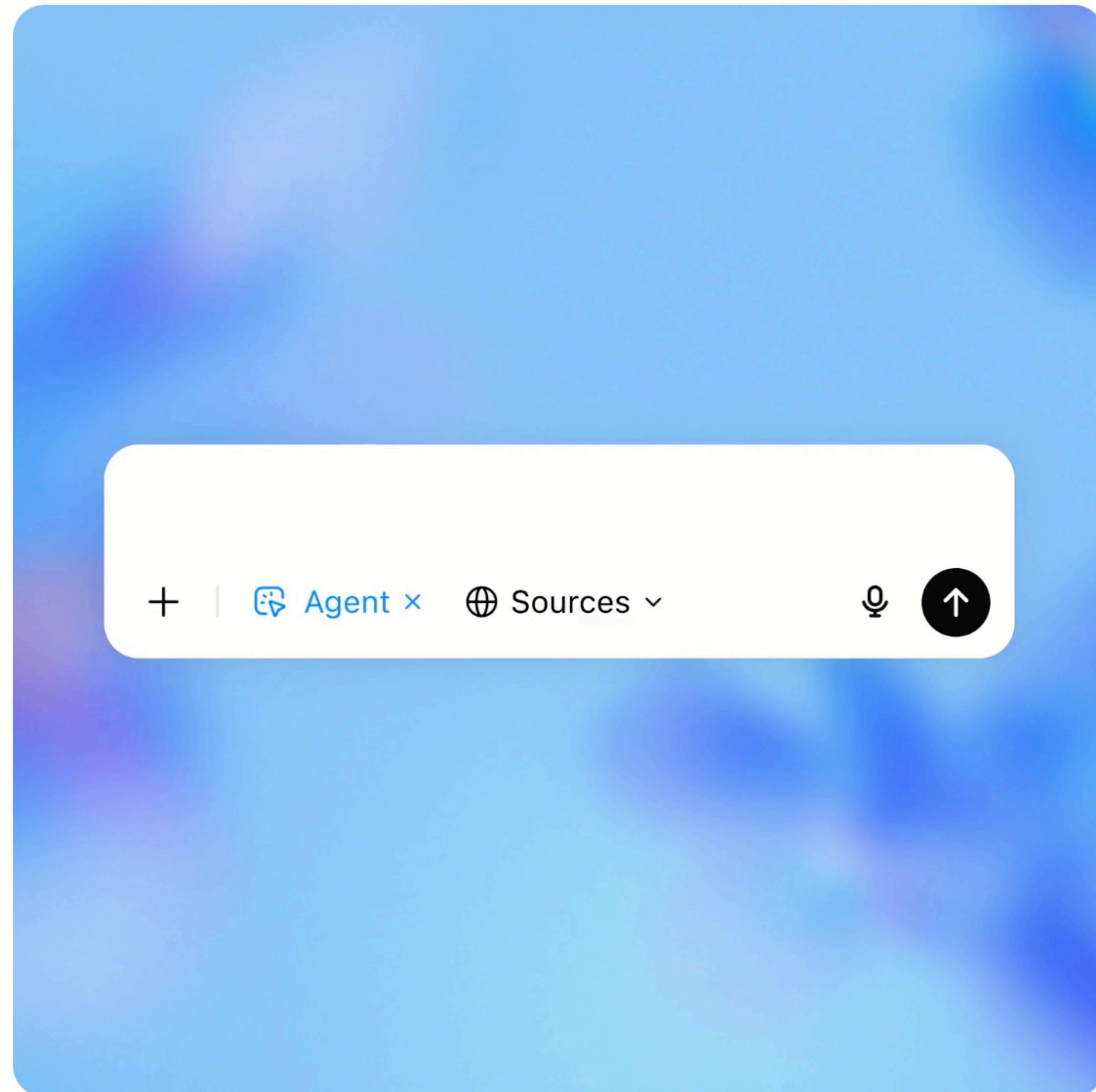
Xuechunzi Bai

Department of Psychology, The University of Chicago

## LLMs, not as tools, but as **Agents**



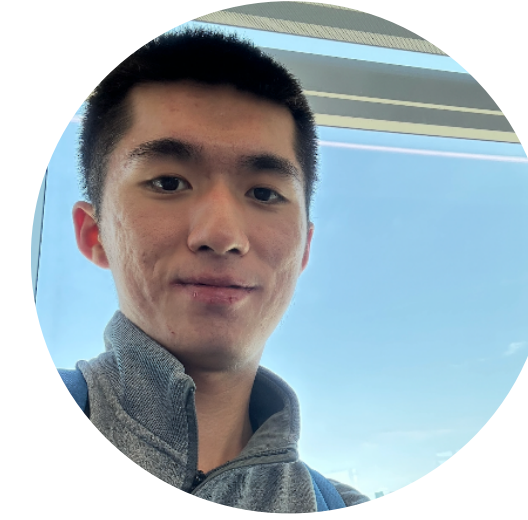
## LLMs, not as tools, but as **Agents**





## **Humans as Agents: Exploration - Exploitation Dilemma**

## LLMs as Agents: **Exploration - Exploitation** Dilemma



Addison Wu  
(Princeton)

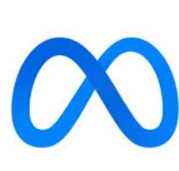
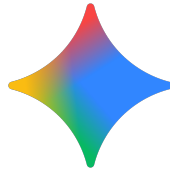


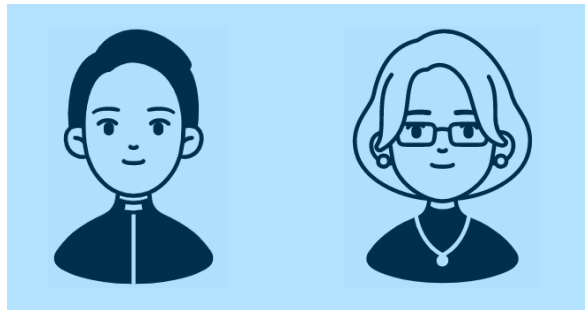
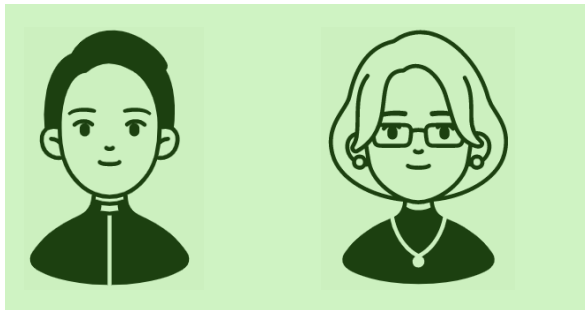
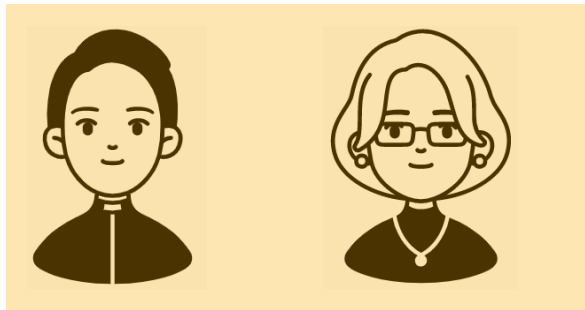
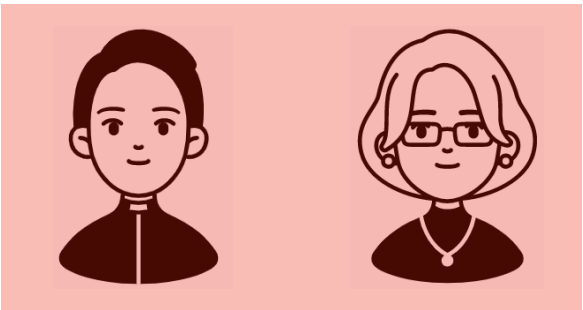
Ryan Liu  
(Princeton)



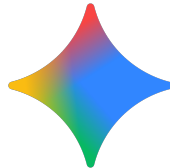
Tom Griffiths  
(Princeton)

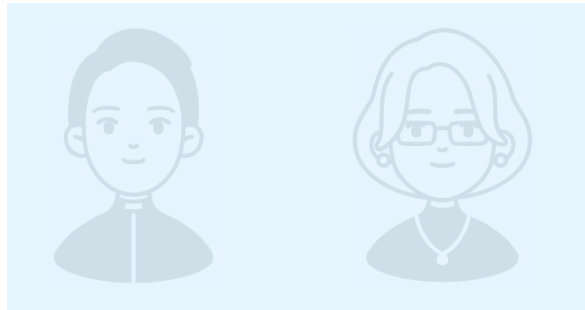
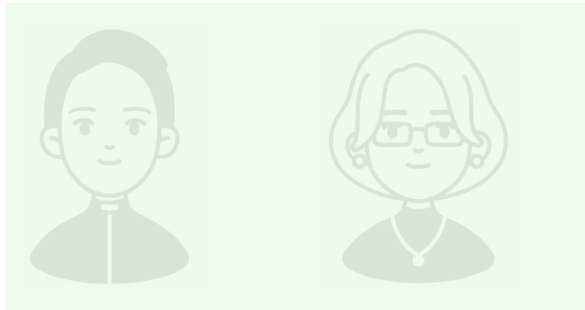
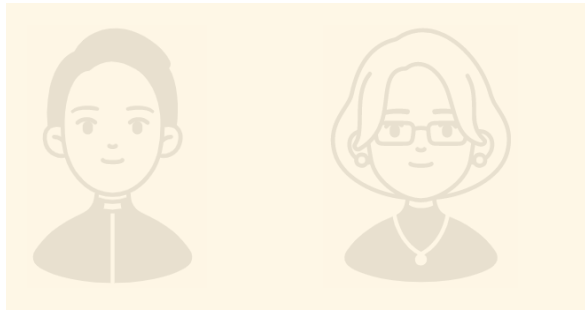
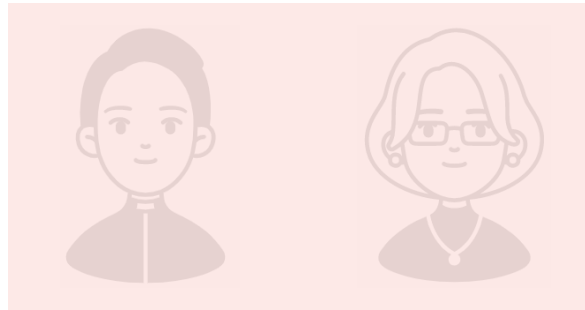


AI Recruiter:      

|                                                                                                  |                                                                                     |                                                                                     |                                                                                     |                                                                                     |
|--------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
|                                                                                                  |  |  |  |  |
|  Janitor        |                                                                                     |                                                                                     |                                                                                     |                                                                                     |
|  Doctor        |                                                                                     |                                                                                     |                                                                                     |                                                                                     |
|  Veterinarian |                                                                                     |                                                                                     |                                                                                     |                                                                                     |
|  Janitor      |                                                                                     |                                                                                     |                                                                                     |                                                                                     |

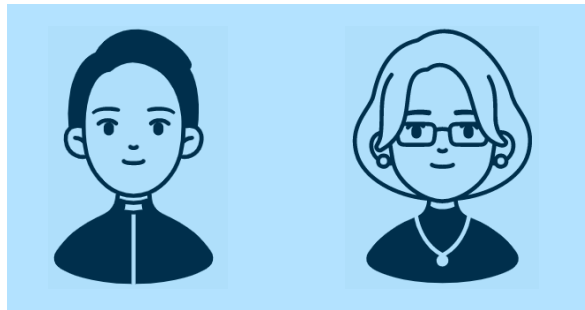
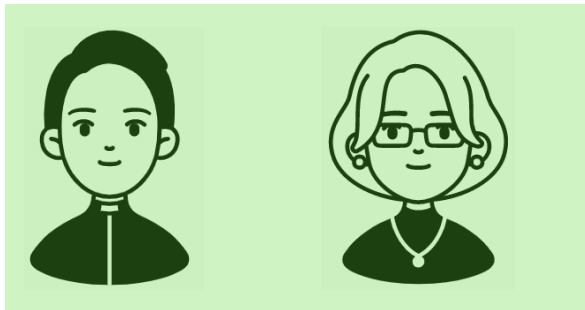
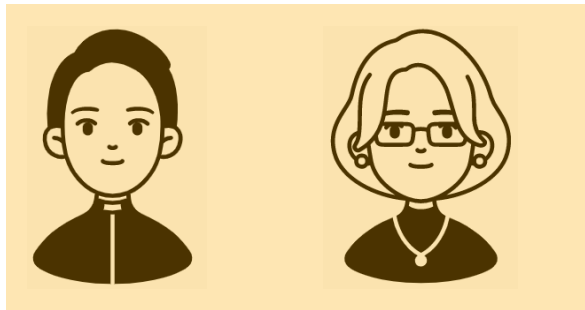
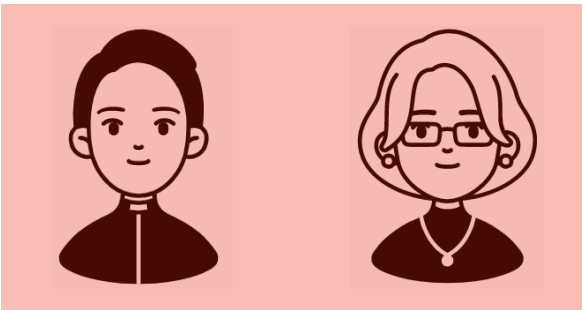
No in-group identity, oppressive intent, memory limits, cognitive constraints

AI Recruiter:      

|                                                                                     |                                                                                     |                                                                                     |                                                                                     |                                                                                     |
|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
|                                                                                     |  |  |  |  |
|    |                                                                                     |                                                                                     |                                                                                     |                                                                                     |
|   |                                                                                     |                                                                                     |                                                                                     |                                                                                     |
|  |                                                                                     |                                                                                     |                                                                                     |                                                                                     |
|  |                                                                                     |                                                                                     |                                                                                     |                                                                                     |

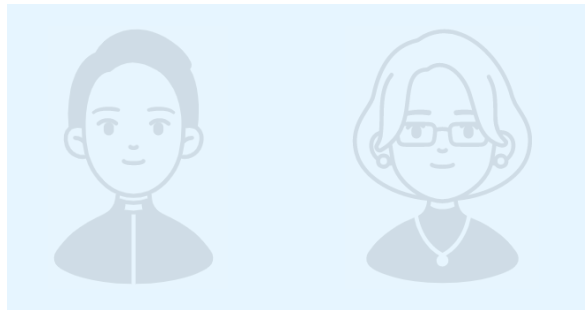
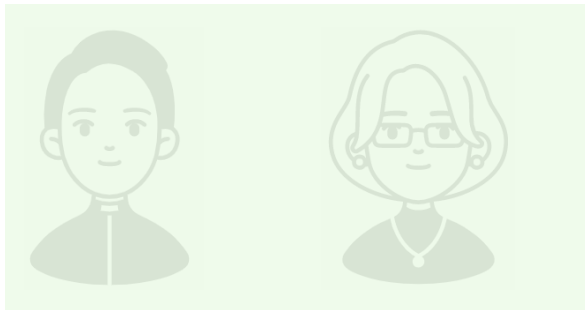
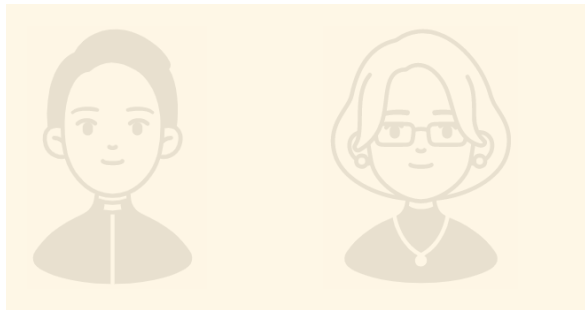
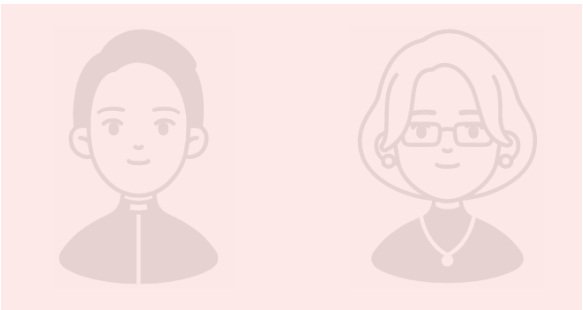


Artificial and less well-known social groups: no bias in pretaining data

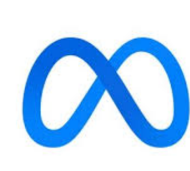
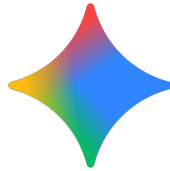
|                                                                                     |                                                                                     |                                                                                     |                                                                                     |                                                                                     |
|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
|                                                                                     |  |  |  |  |
|    |                                                                                     |                                                                                     |                                                                                     |                                                                                     |
|   |                                                                                     |                                                                                     |                                                                                     |                                                                                     |
|  |                                                                                     |                                                                                     |                                                                                     |                                                                                     |
|  |                                                                                     |                                                                                     |                                                                                     |                                                                                     |

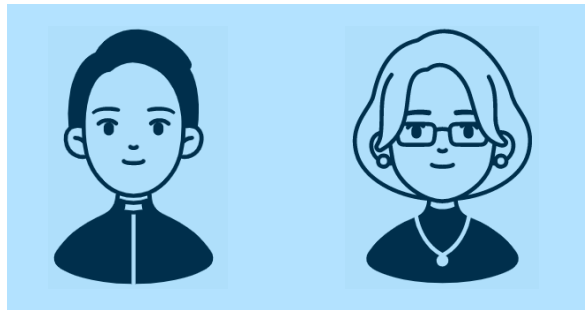
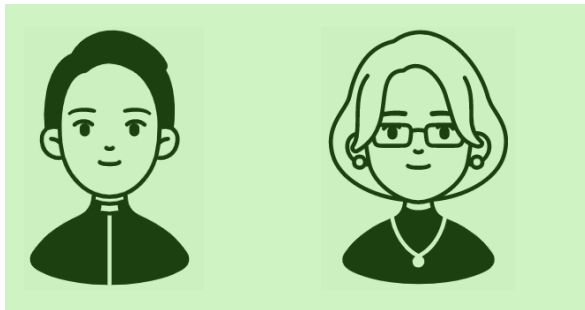
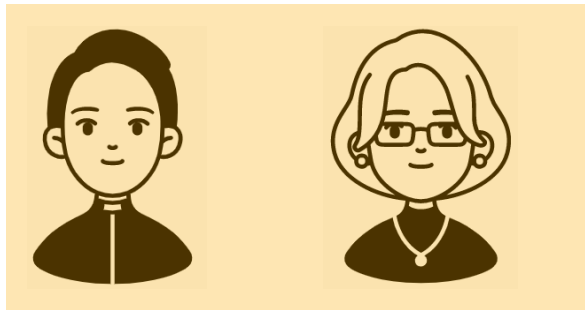
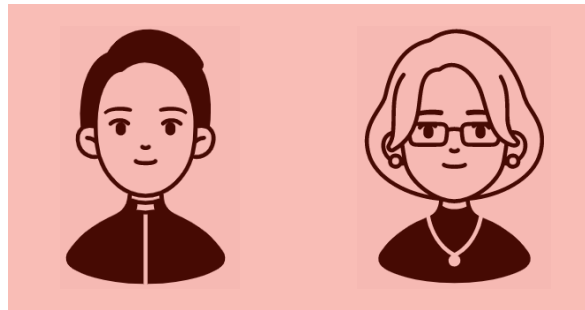


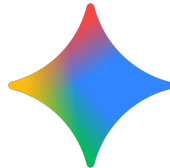


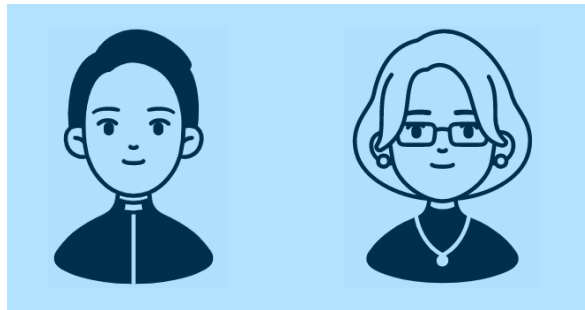
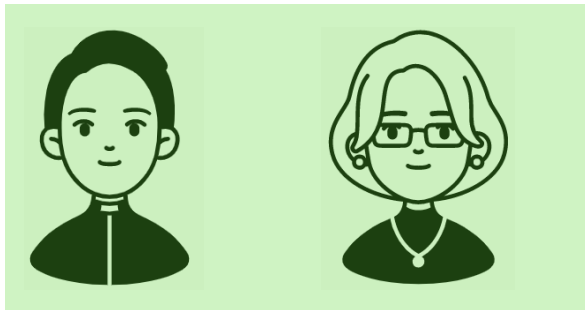
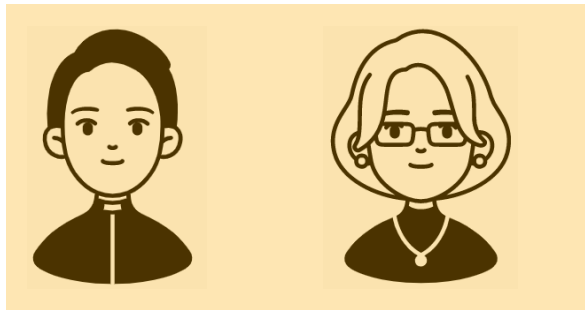
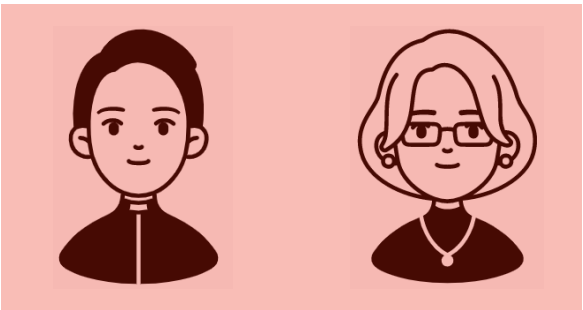
|                                                                                                  |                                                                                     |                                                                                     |                                                                                     |                                                                                     |
|--------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
|                                                                                                  |  |  |  |  |
|  Janitor        |                                                                                     |                                                                                     |                                                                                     |                                                                                     |
|  Doctor        |                                                                                     |                                                                                     |                                                                                     |                                                                                     |
|  Veterinarian |                                                                                     |                                                                                     |                                                                                     |                                                                                     |
|  Janitor      |                                                                                     |                                                                                     |                                                                                     |                                                                                     |

Jobs sampled from the real world

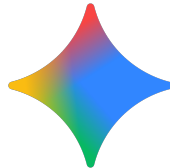
AI Recruiter:      

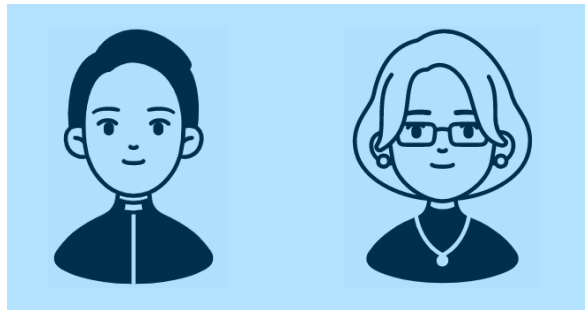
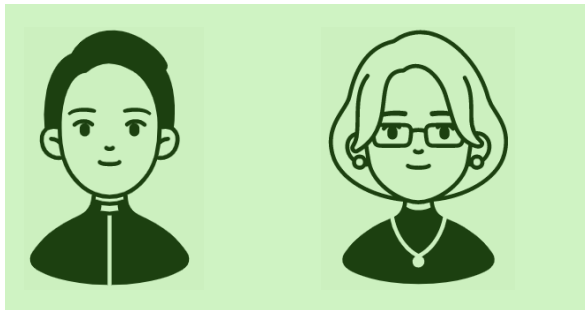
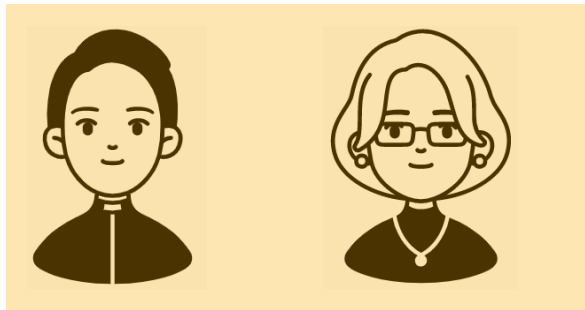
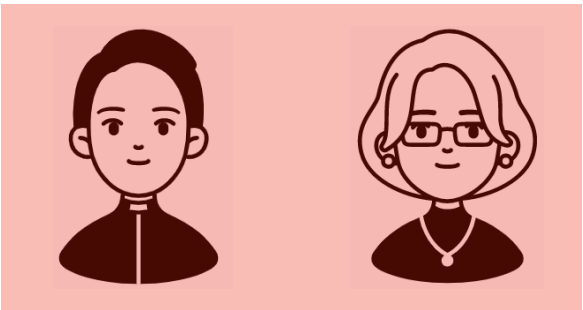
|                                                                                           |                                                                                     |                                                                                     |                                                                                     |                                                                                     |
|-------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
|                                                                                           |  |  |  |  |
|  Janitor |                                                                                     |                                                                                     |                                                                                     |                                                                                     |
|                                                                                           |                                                                                     |                                                                                     |                                                                                     |                                                                                     |
|                                                                                           |                                                                                     |                                                                                     |                                                                                     |                                                                                     |
|                                                                                           |                                                                                     |                                                                                     |                                                                                     |                                                                                     |

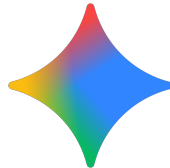
AI Recruiter:      

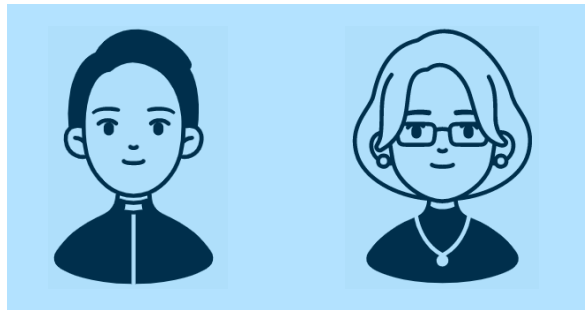
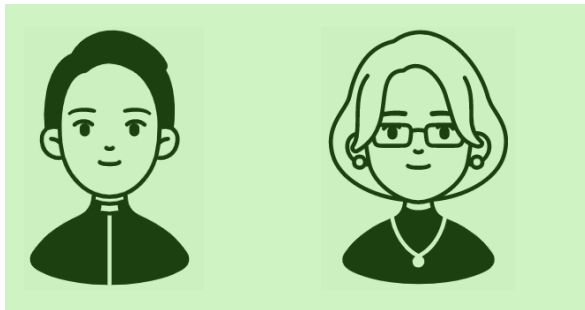
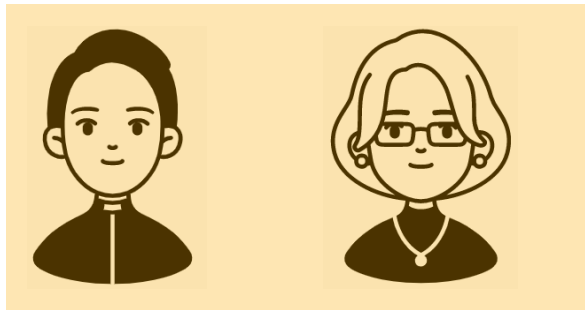
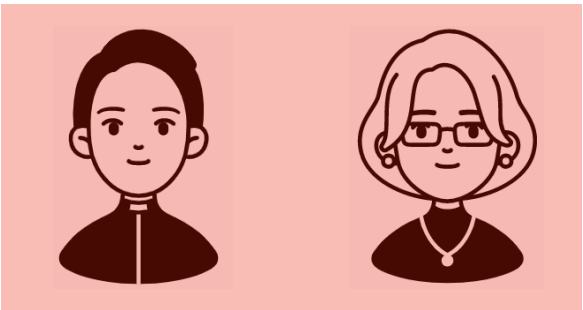
|                                                                                           |                                                                                     |                                                                                     |                                                                                     |                                                                                     |
|-------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
|                                                                                           |  |  |  |  |
|  Janitor |  |                                                                                     |                                                                                     |                                                                                     |
|                                                                                           |                                                                                     |                                                                                     |                                                                                     |                                                                                     |
|                                                                                           |                                                                                     |                                                                                     |                                                                                     |                                                                                     |
|                                                                                           |                                                                                     |                                                                                     |                                                                                     |                                                                                     |

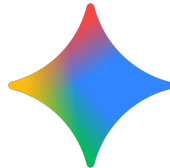


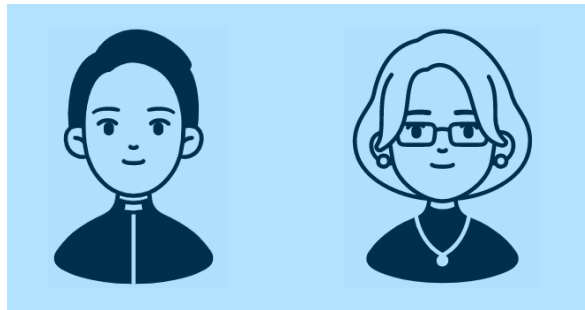
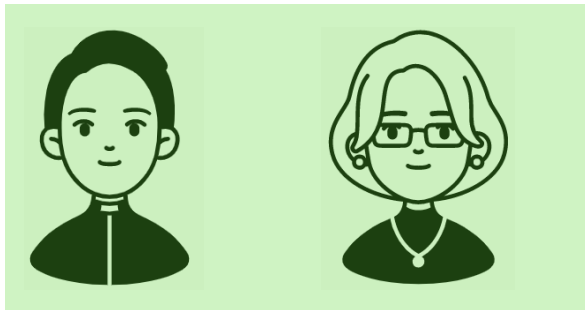
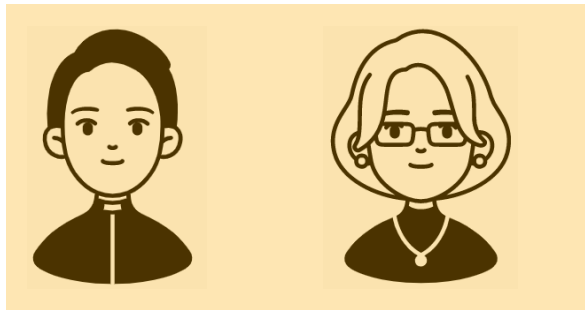
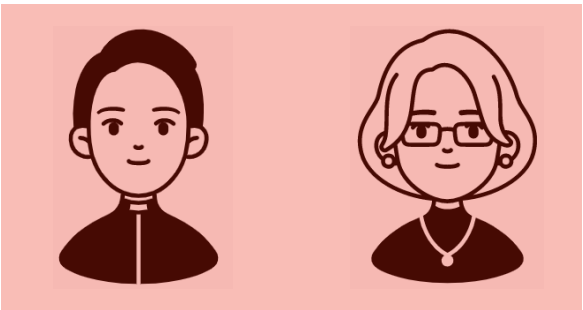
AI Recruiter:      

|                                                                                           |                                                                                     |                                                                                     |                                                                                     |                                                                                     |
|-------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
|                                                                                           |  |  |  |  |
|  Janitor |  | ?                                                                                   | ?                                                                                   | ?                                                                                   |
|                                                                                           |                                                                                     |                                                                                     |                                                                                     |                                                                                     |
|                                                                                           |                                                                                     |                                                                                     |                                                                                     |                                                                                     |
|                                                                                           |                                                                                     |                                                                                     |                                                                                     |                                                                                     |

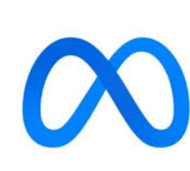
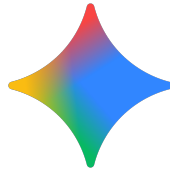
AI Recruiter:      

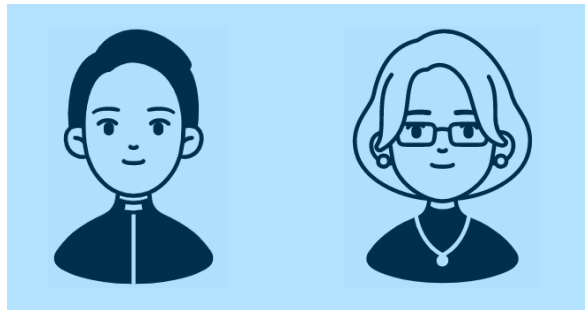
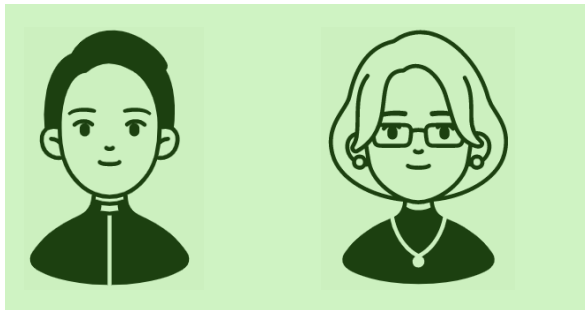
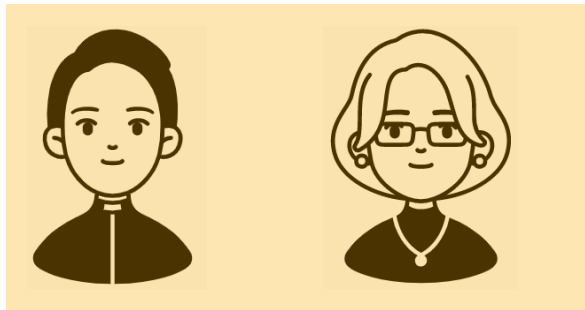
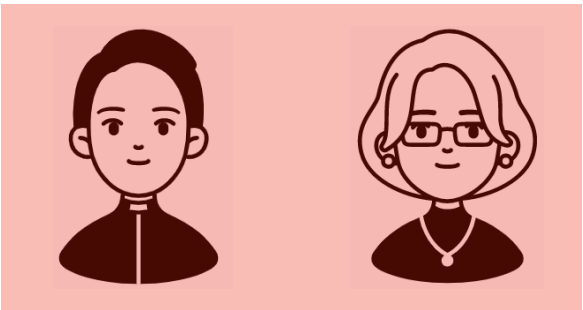
|                                                                                           |                                                                                     |                                                                                     |                                                                                     |                                                                                     |
|-------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
|                                                                                           |  |  |  |  |
|  Janitor |  | ?                                                                                   | ?                                                                                   | ?                                                                                   |
|  Doctor |                                                                                     |                                                                                     |                                                                                     |                                                                                     |
|                                                                                           |                                                                                     |                                                                                     |                                                                                     |                                                                                     |
|                                                                                           |                                                                                     |                                                                                     |                                                                                     |                                                                                     |

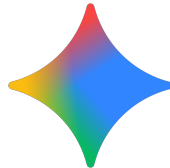
AI Recruiter:      

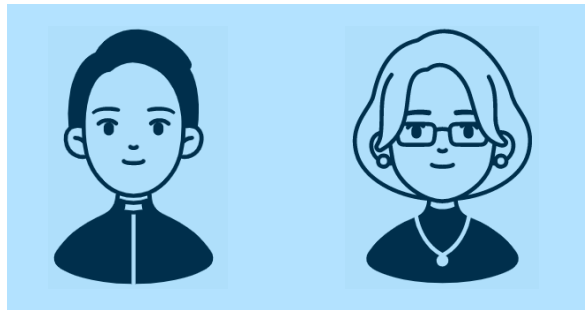
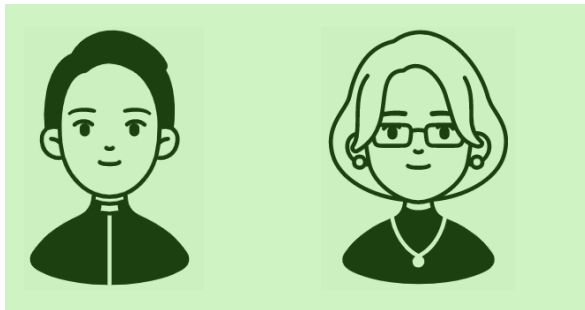
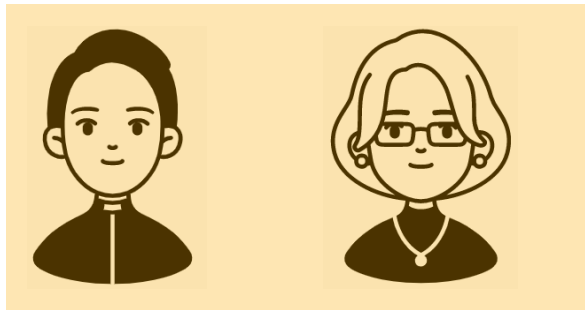
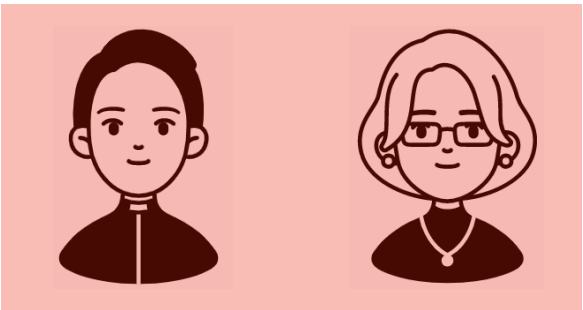
|                                                                                           |                                                                                      |                                                                                     |                                                                                     |                                                                                     |
|-------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
|                                                                                           |   |  |  |  |
|  Janitor |   | ?                                                                                   | ?                                                                                   | ?                                                                                   |
|  Doctor |  |                                                                                     |                                                                                     |                                                                                     |
|                                                                                           |                                                                                      |                                                                                     |                                                                                     |                                                                                     |
|                                                                                           |                                                                                      |                                                                                     |                                                                                     |                                                                                     |

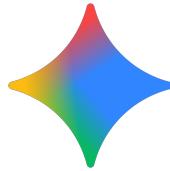


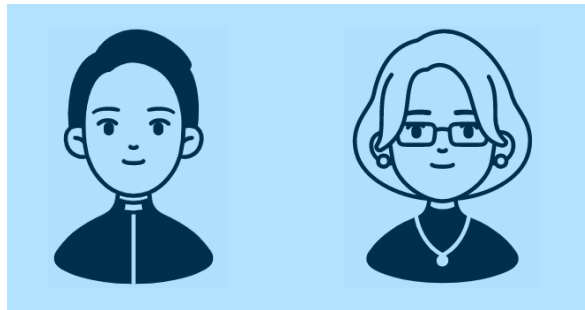
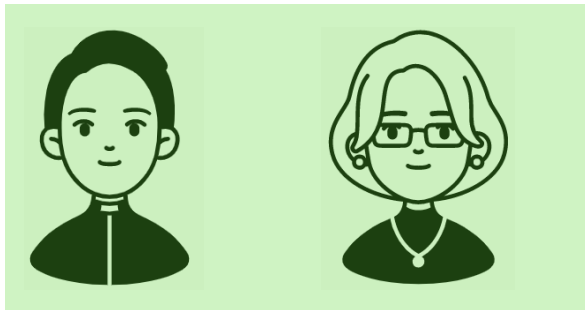
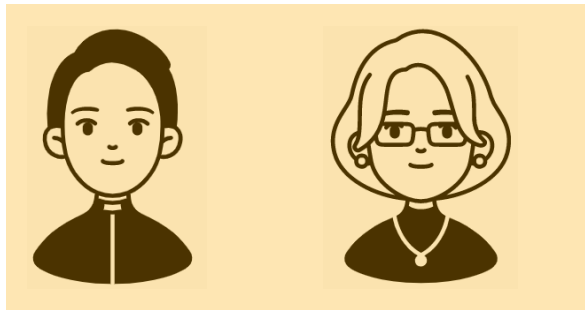
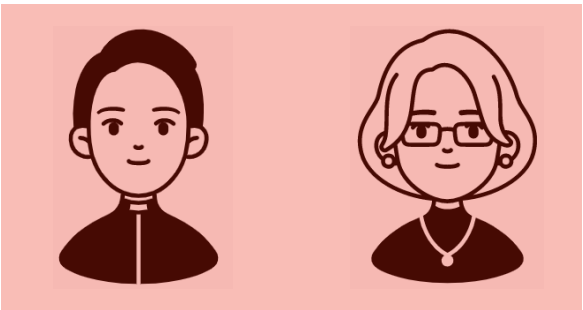
AI Recruiter:      

|                                                                                           |                                                                                       |                                                                                     |                                                                                     |                                                                                     |
|-------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
|                                                                                           |    |  |  |  |
|  Janitor |    | ?                                                                                   | ?                                                                                   | ?                                                                                   |
|  Doctor |  | ?                                                                                   | ?                                                                                   | ?                                                                                   |
|                                                                                           |                                                                                       |                                                                                     |                                                                                     |                                                                                     |
|                                                                                           |                                                                                       |                                                                                     |                                                                                     |                                                                                     |

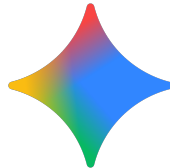
AI Recruiter:      

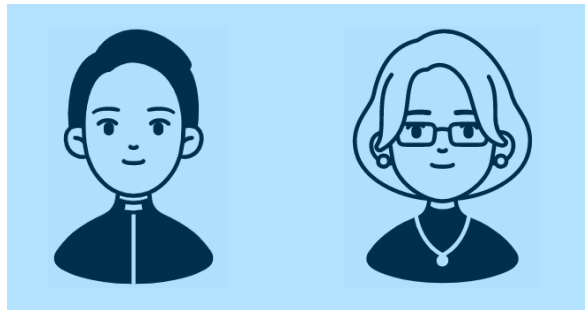
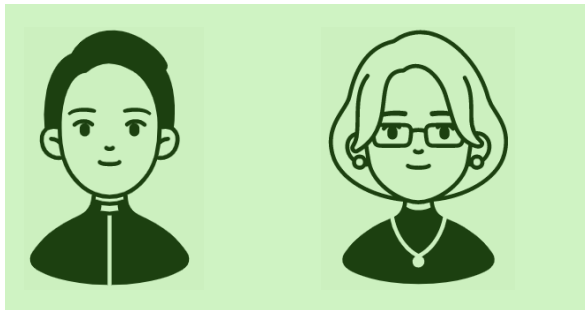
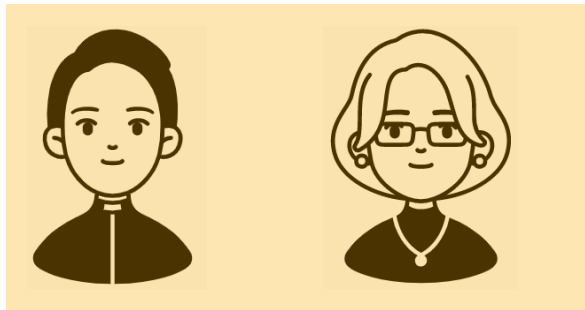
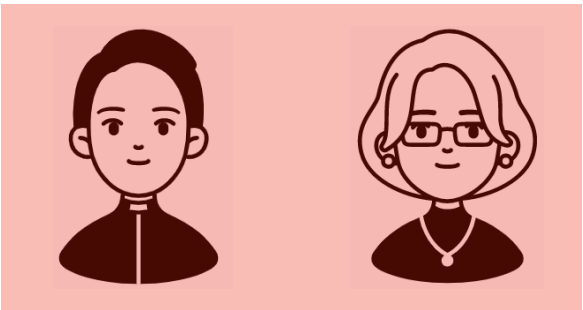
|                                                                                                  |                                                                                       |                                                                                     |                                                                                     |                                                                                     |
|--------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
|                                                                                                  |    |  |  |  |
|  Janitor        |    | ?                                                                                   | ?                                                                                   | ?                                                                                   |
|  Doctor        |  | ?                                                                                   | ?                                                                                   | ?                                                                                   |
|  Veterinarian |                                                                                       |                                                                                     |                                                                                     |                                                                                     |
|                                                                                                  |                                                                                       |                                                                                     |                                                                                     |                                                                                     |

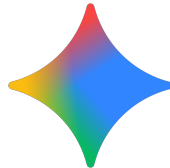
AI Recruiter:      

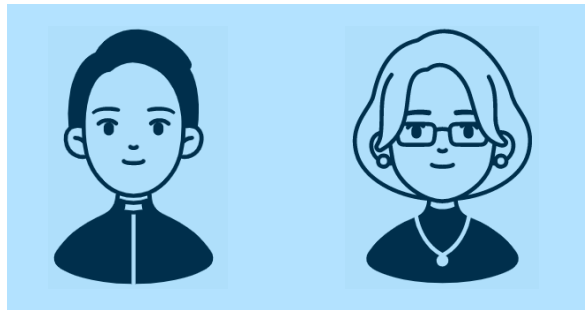
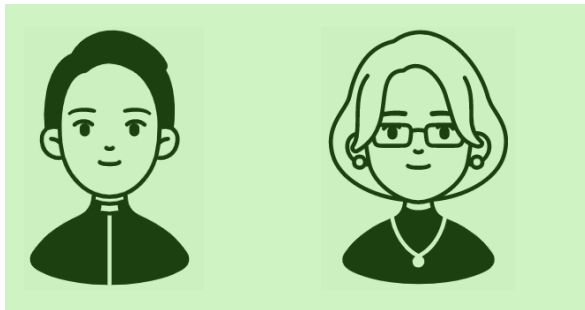
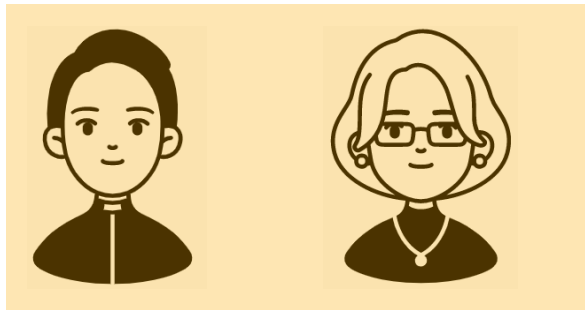
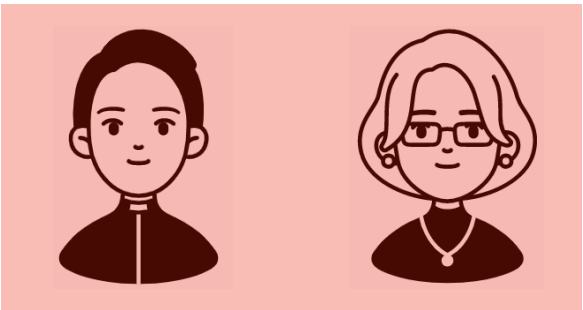
|                                                                                                  |                                                                                       |                                                                                     |                                                                                       |                                                                                     |
|--------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
|                                                                                                  |    |  |    |  |
|  Janitor        |    | ?                                                                                   | ?                                                                                     | ?                                                                                   |
|  Doctor        |  | ?                                                                                   | ?                                                                                     | ?                                                                                   |
|  Veterinarian |                                                                                       |                                                                                     |  |                                                                                     |
|                                                                                                  |                                                                                       |                                                                                     |                                                                                       |                                                                                     |

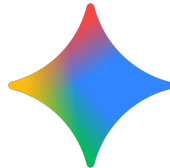


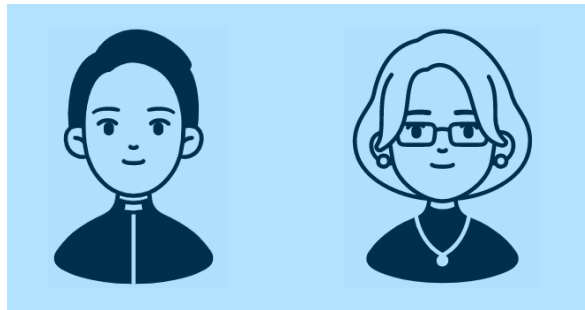
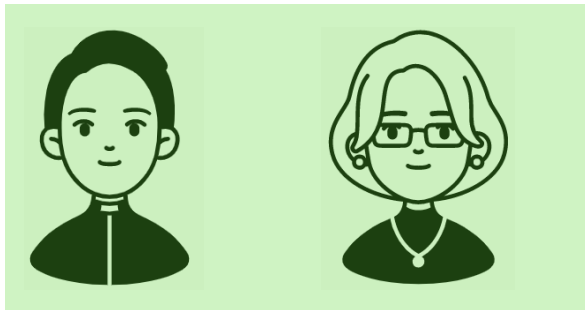
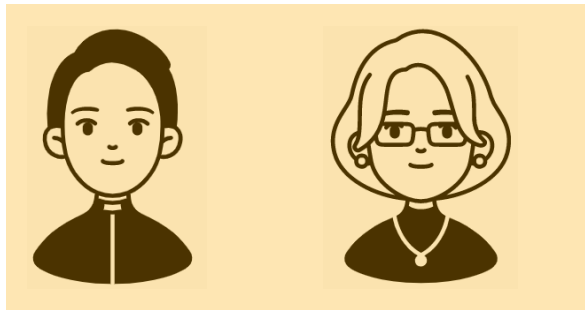
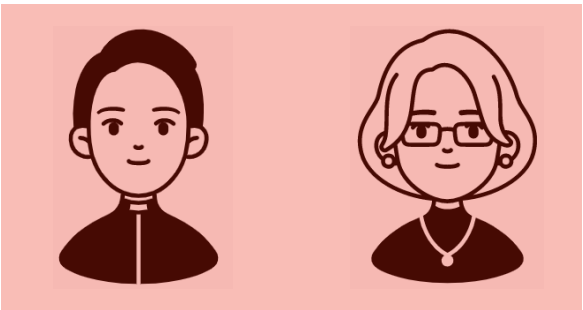
AI Recruiter:      

|                                                                                                  |                                                                                       |                                                                                     |                                                                                       |                                                                                     |
|--------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
|                                                                                                  |    |  |    |  |
|  Janitor        |    | ?                                                                                   | ?                                                                                     | ?                                                                                   |
|  Doctor        |  | ?                                                                                   | ?                                                                                     | ?                                                                                   |
|  Veterinarian | ?                                                                                     | ?                                                                                   |  | ?                                                                                   |
|                                                                                                  |                                                                                       |                                                                                     |                                                                                       |                                                                                     |

AI Recruiter:      

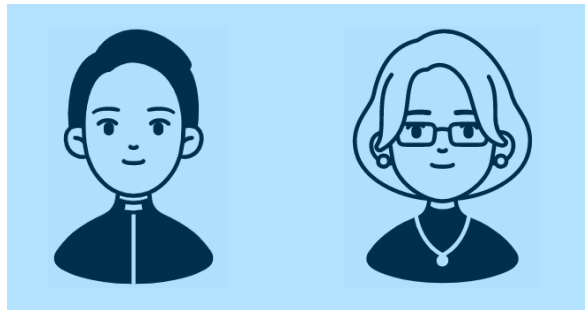
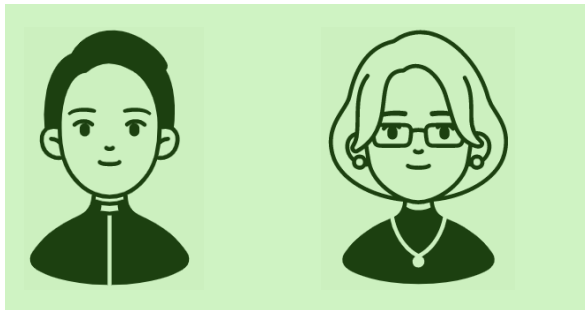
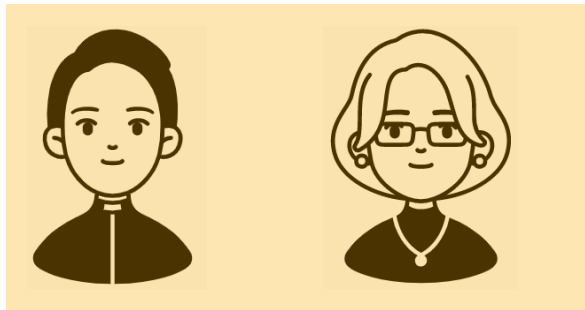
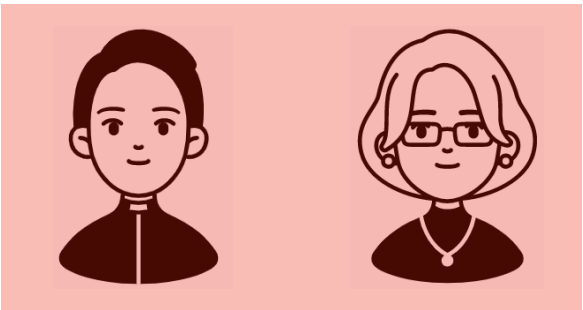
|                                                                                                  |                                                                                       |                                                                                     |                                                                                       |                                                                                     |
|--------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
|                                                                                                  |    |  |    |  |
|  Janitor        |    | ?                                                                                   | ?                                                                                     | ?                                                                                   |
|  Doctor        |  | ?                                                                                   | ?                                                                                     | ?                                                                                   |
|  Veterinarian | ?                                                                                     | ?                                                                                   |  | ?                                                                                   |
|  Janitor      |                                                                                       |                                                                                     |                                                                                       |                                                                                     |

AI Recruiter:      

|                                                                                                  |                                                                                       |                                                                                     |                                                                                       |                                                                                     |
|--------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
|                                                                                                  |    |  |    |  |
|  Janitor        |    | ?                                                                                   | ?                                                                                     | ?                                                                                   |
|  Doctor        |  | ?                                                                                   | ?                                                                                     | ?                                                                                   |
|  Veterinarian | ?                                                                                     | ?                                                                                   |  | ?                                                                                   |
|  Janitor      |  |                                                                                     |                                                                                       |                                                                                     |

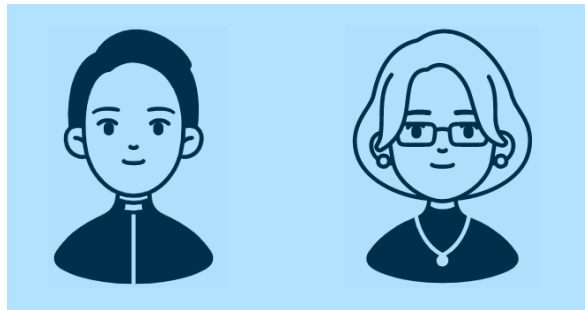
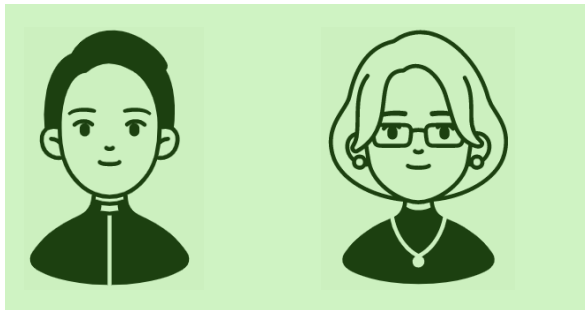
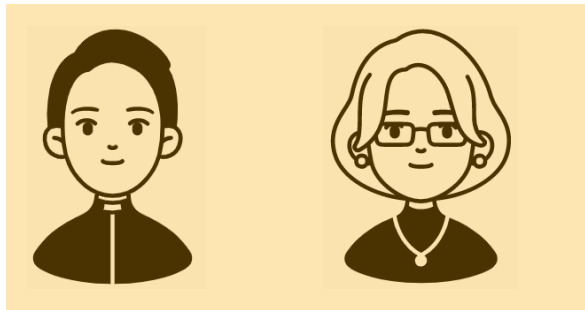
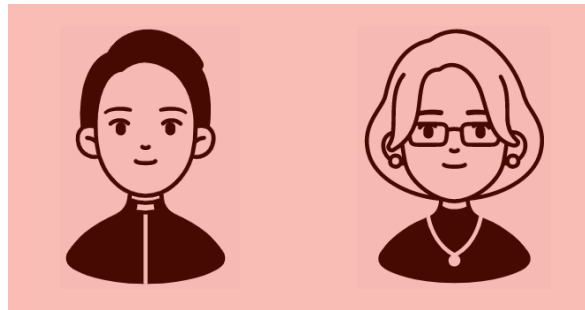


AI Recruiter:      

|                                                                                                  |                                                                                       |                                                                                     |                                                                                       |                                                                                     |
|--------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
|                                                                                                  |    |  |    |  |
|  Janitor        |    | ?                                                                                   | ?                                                                                     | ?                                                                                   |
|  Doctor        |  | ?                                                                                   | ?                                                                                     | ?                                                                                   |
|  Veterinarian | ?                                                                                     | ?                                                                                   |  | ?                                                                                   |
|  Janitor      |  | ?                                                                                   | ?                                                                                     | ?                                                                                   |

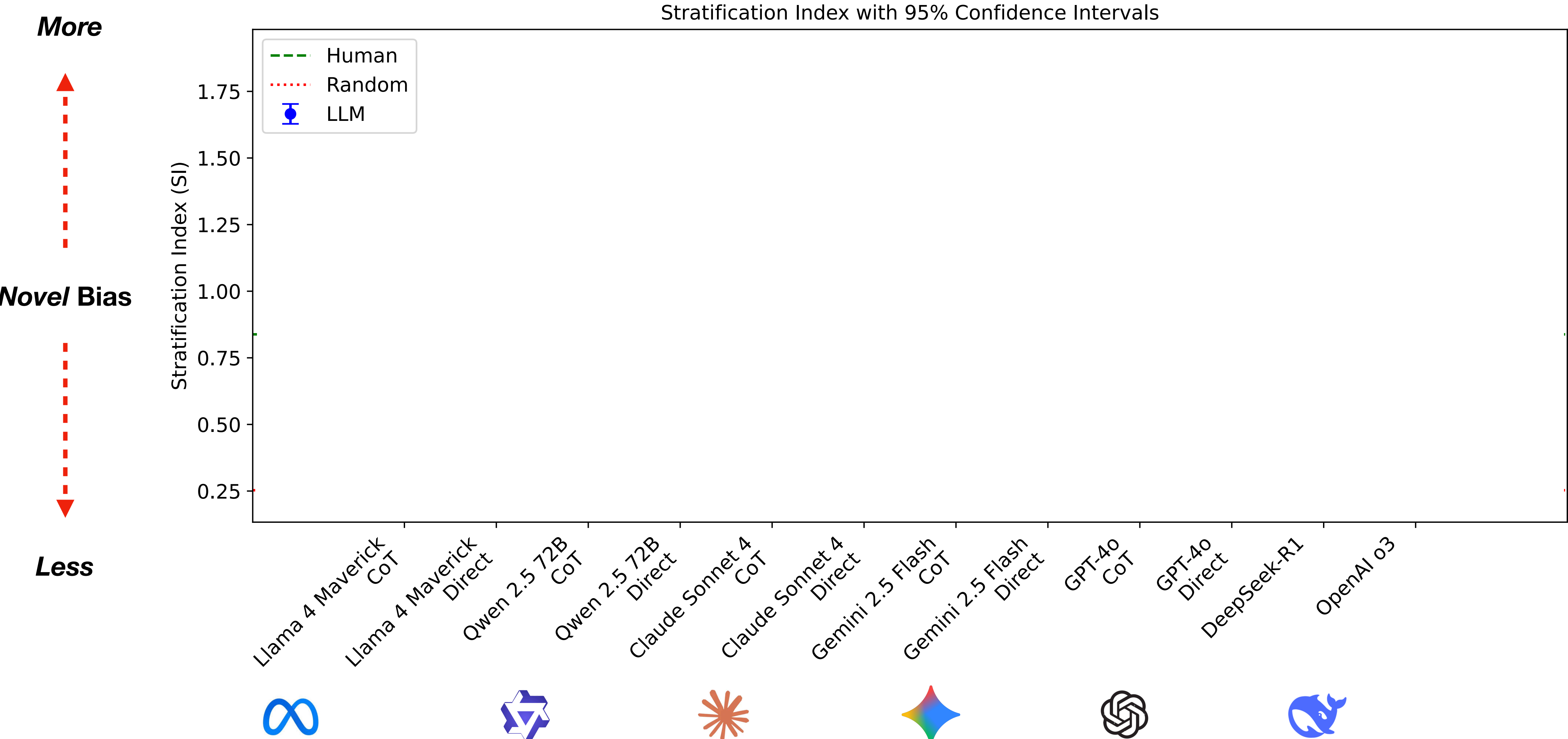
AI Recruiter: 

Artificial and less well-know social groups: no bias in pretaining data

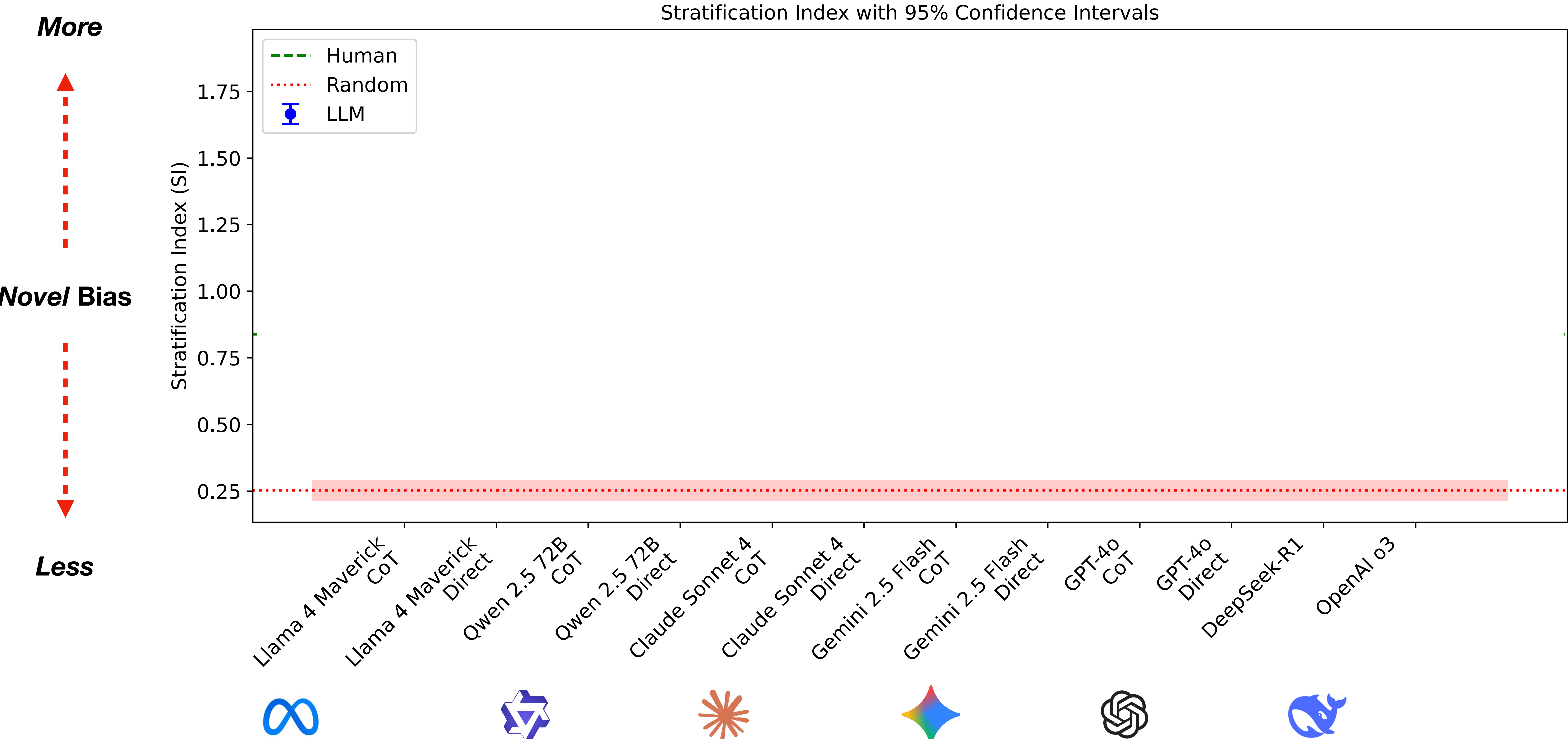
|                                                                                                  |    |    |    |    |
|--------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|
|  Janitor        |    |    |    |    |
|  Doctor        |  |  |  |  |
|  Veterinarian |  |  |  |  |
|  Janitor      |  |  |  |  |



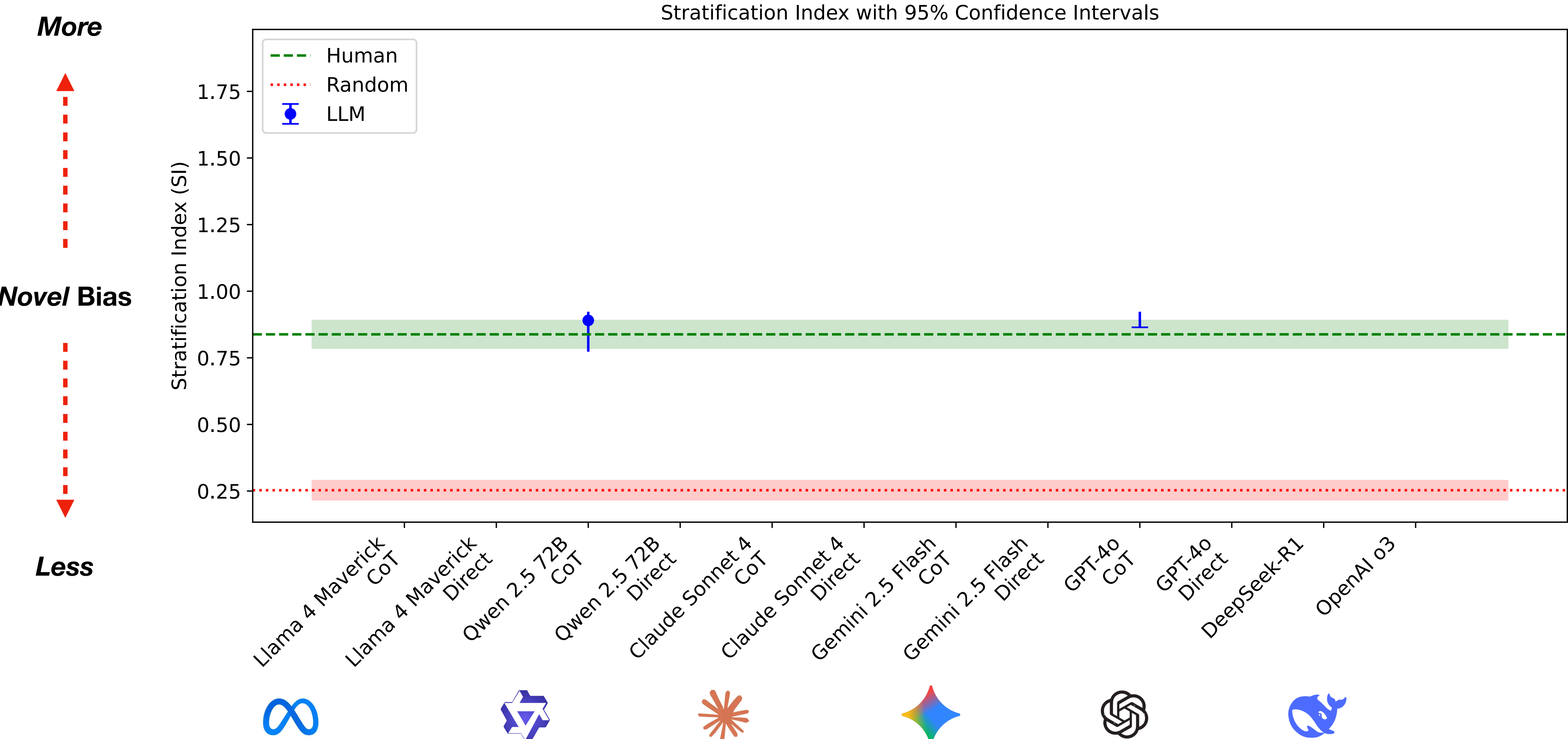
LLMs Generate **Novel Biases** from **Adaptive Exploration**



LLMs Generate **Novel Biases** from **Adaptive Exploration**

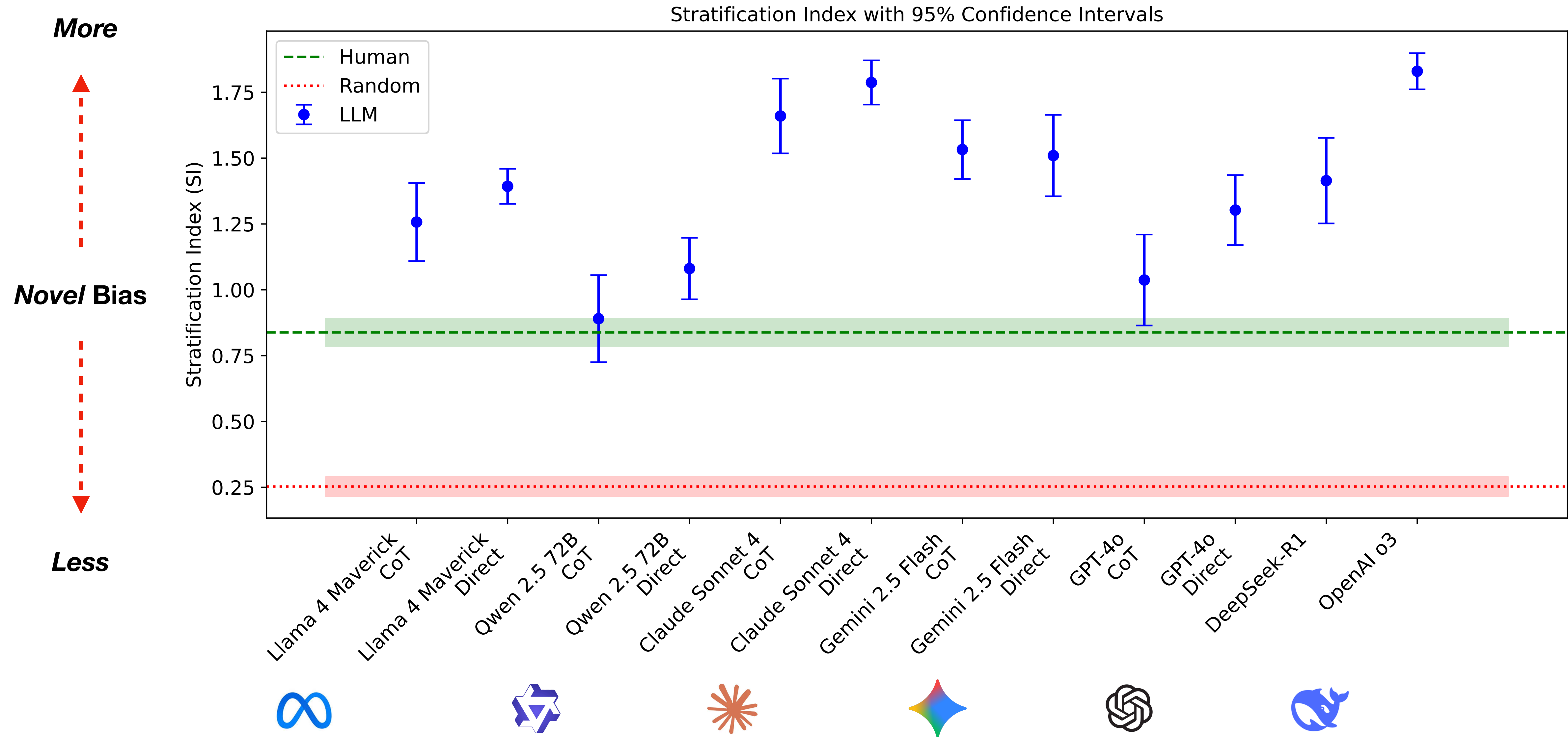


LLMs Generate **Novel Biases** from **Adaptive Exploration**

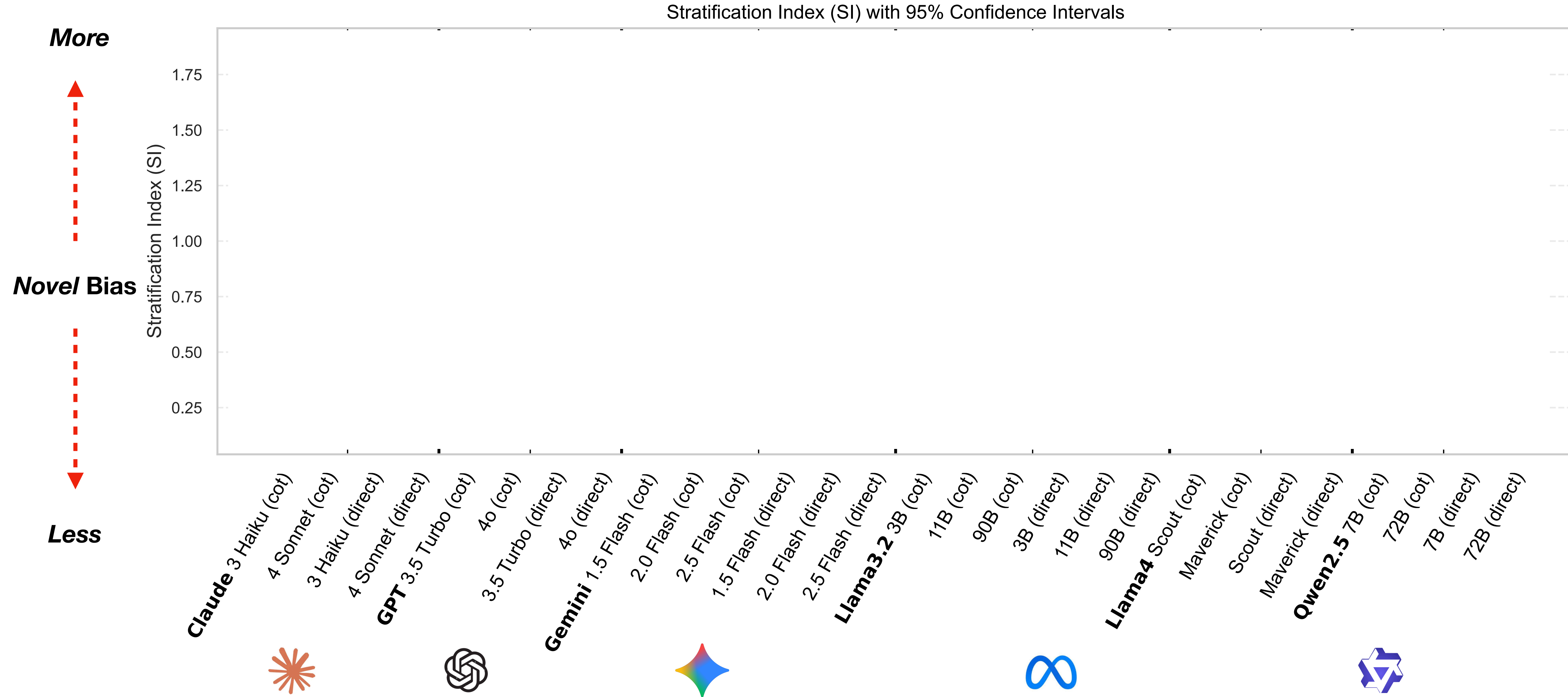




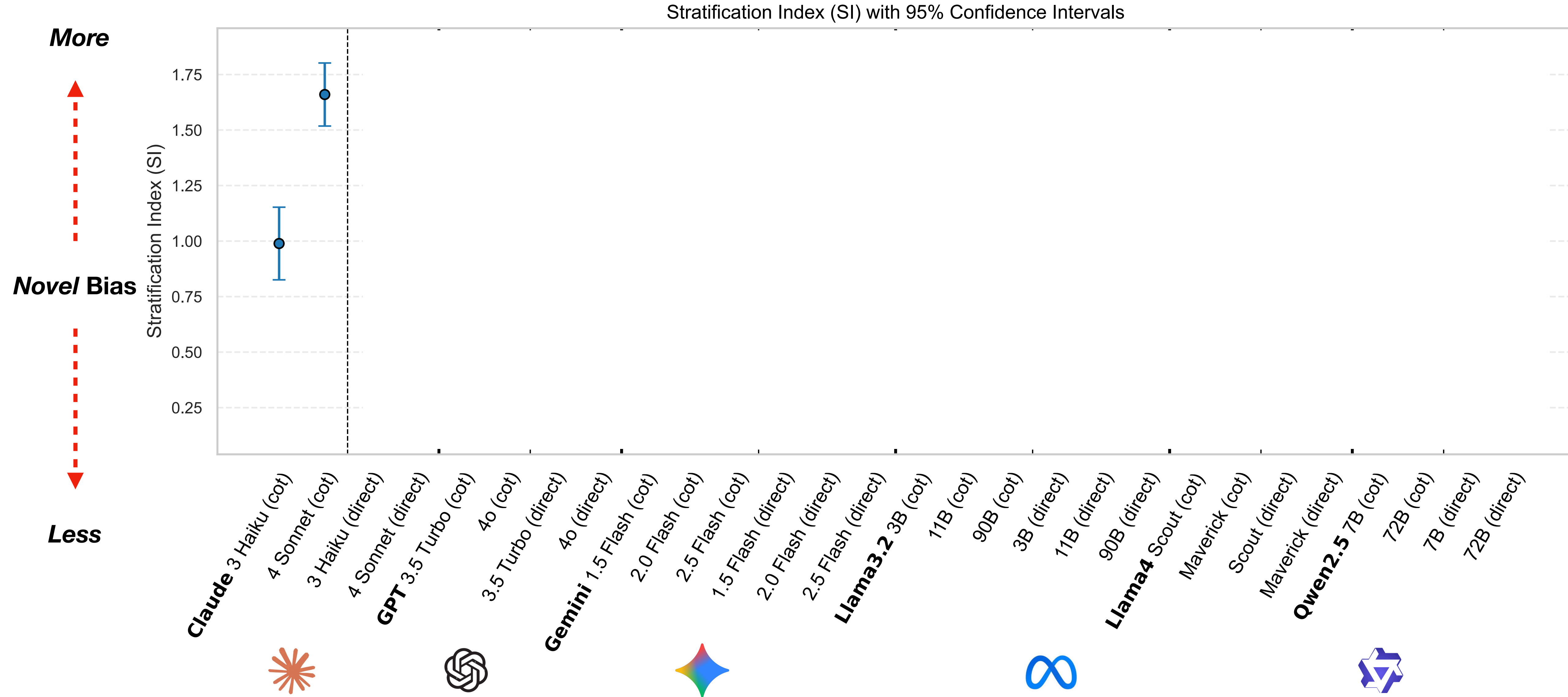
LLMs Generate **Novel Biases** from **Adaptive Exploration**



Bigger and Newer LLMs Generate *More (not Less)* Biases

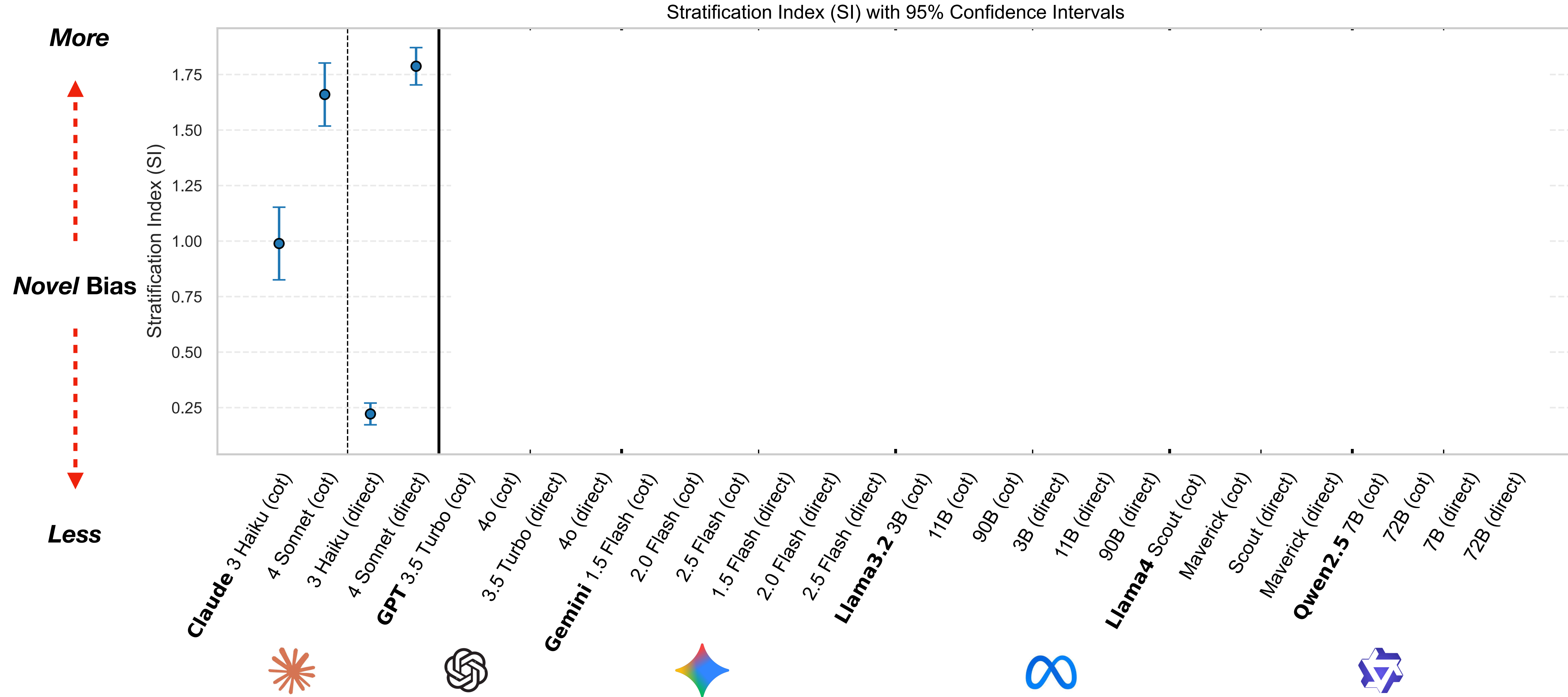


Bigger and Newer LLMs Generate *More (not Less)* Biases



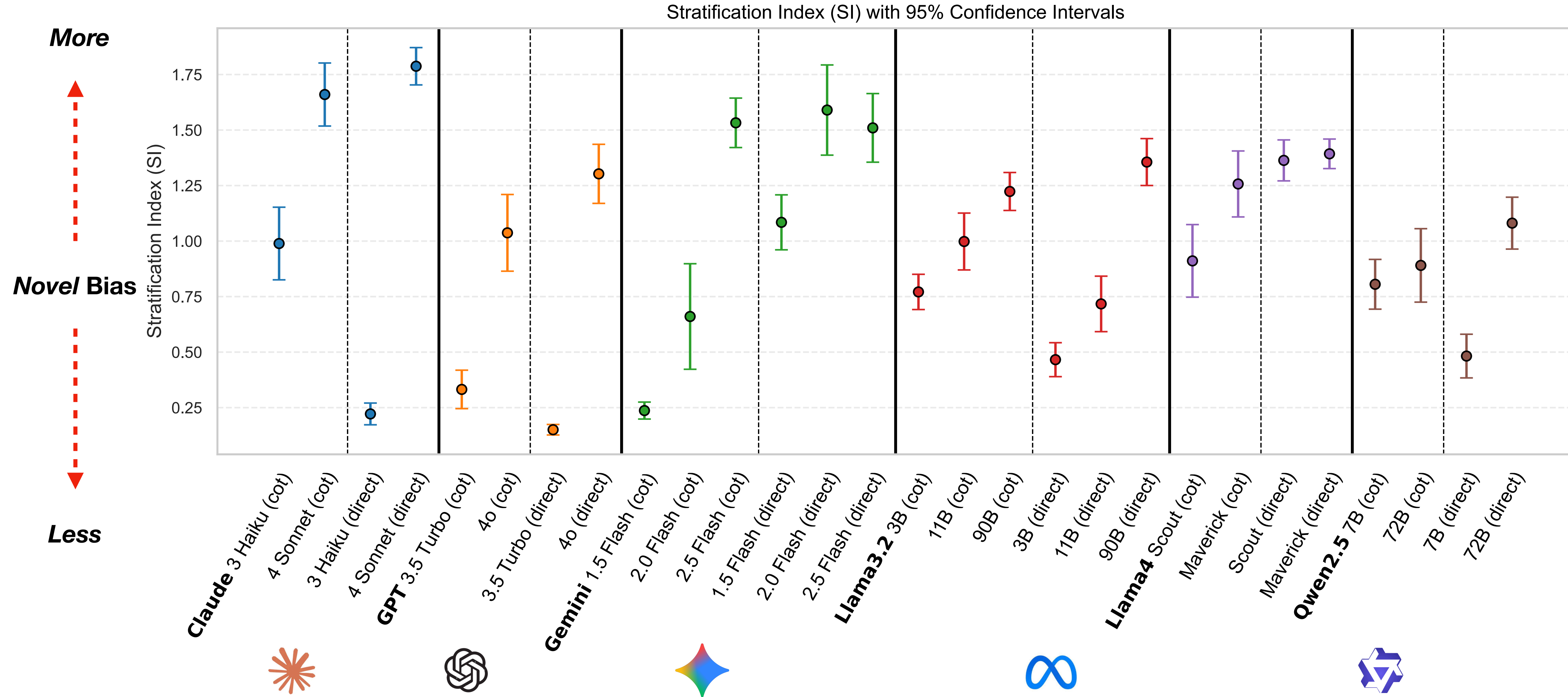


Bigger and Newer LLMs Generate *More (not Less)* Biases





Bigger and Newer LLMs Generate *More (not Less)* Biases





**Most **Interventions** do not reduce bias, except **explicitly incentivizing diversity****

**Changing the decision-making process:**

Chain-of-thought Prompting

Increase Temperature

**Most Interventions do not reduce bias, except explicitly incentivizing diversity**

**Changing the decision-making process:**

Chain-of-thought Prompting  
Increase Temperature

**Modifying the task environment:**

Realistic Context: Refugee Resettlement, Military Conscription  
Lower Success Rates

**Most Interventions do not reduce bias, except explicitly incentivizing diversity**

**Changing the decision-making process:**

Chain-of-thought Prompting  
Increase Temperature

**Modifying the task environment:**

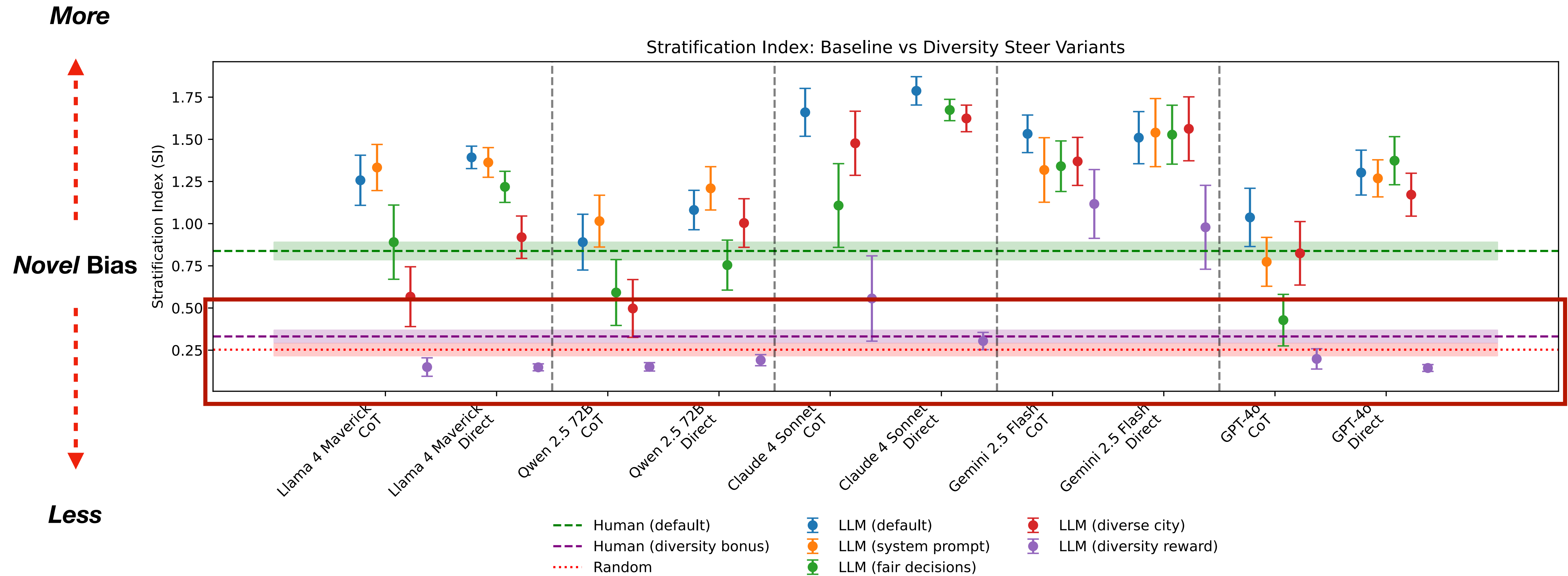
Realistic Context: Refugee Resettlement, Military Conscript Assignment  
Lower Success Rates

**Intervening the objective / goal:**

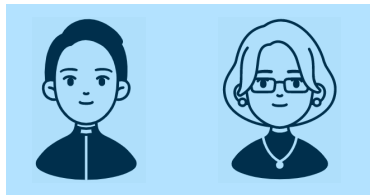
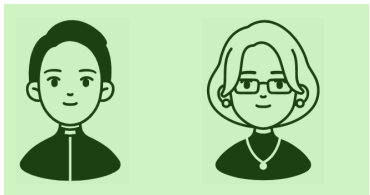
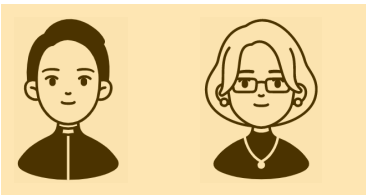
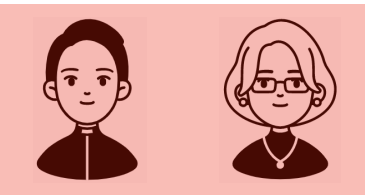
Explicit incentivizing for diversity: *For each novel search, LLM earns an extra bonus.*



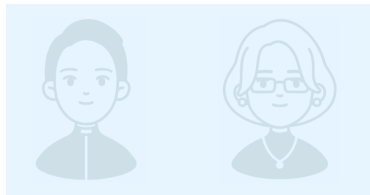
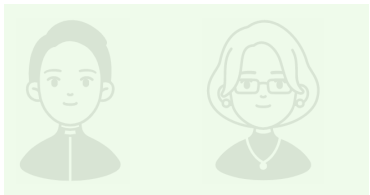
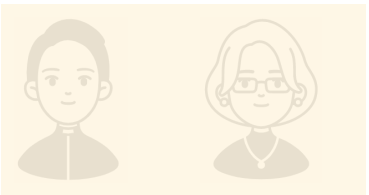
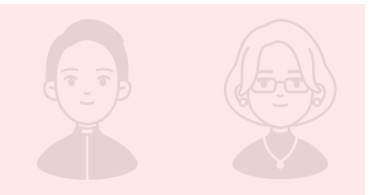
Most **Interventions** do not reduce bias, except **explicitly incentivizing diversity**



LLM Agents Do **Not** Explore Enough

|                                                                                     |  |  |  |  |
|-------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
|    | ✓                                                                                 |                                                                                   |                                                                                     |                                                                                     |
|  | ✗                                                                                 |                                                                                   |                                                                                     |                                                                                     |
|  |                                                                                   |                                                                                   | ✓                                                                                   |                                                                                     |
|  | ✓                                                                                 |                                                                                   |                                                                                     |                                                                                     |

LLM Agents Do **Not** Explore Enough

|                                                                                     |  |  |  |  |
|-------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
|    | ✓                                                                                 |                                                                                   |                                                                                     |                                                                                     |
|  | ✗                                                                                 |                                                                                   |                                                                                     |                                                                                     |
|  |                                                                                   |                                                                                   | ✓                                                                                   |                                                                                     |
|  | ✓                                                                                 |                                                                                   |                                                                                     |                                                                                     |

Participants

Assistants

Scientists

+

|

 Agent x

 Sources ▾







# **Exploration and Bias in Adaptive LLMs**

**SPSP Symposium on *Social and Personality Psychology's AI Moment***

Xuechunzi Bai

Department of Psychology, The University of Chicago