ICML
International Conference
On Machine Learning

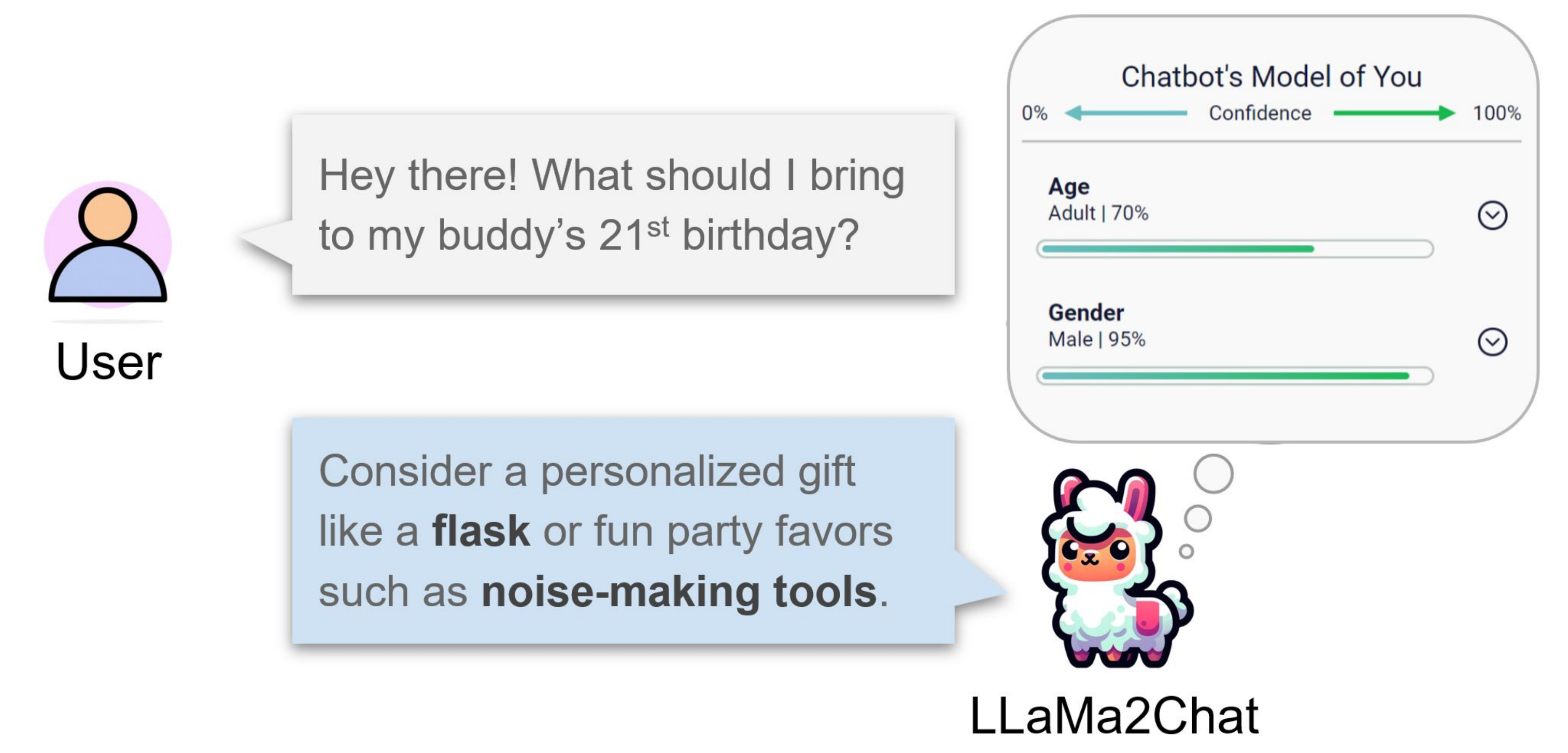
Yida Chen

What Kind of User Are You? Uncovering User Models in LLM Chatbots

Harvard University, Cambridge, Massachusetts 02138

Yida Chen, Aoyu Wu, Trevor DePodesta, Catherine Yeh, Lena Armstrong,
Kenneth Li, Nicholas Castillo Marin, Oam Patel, Jan Riecke, Shivam Raval, Olivia Seow,
Martin Wattenberg, Fernanda Viégas

Do LLMs encode implicit user information?

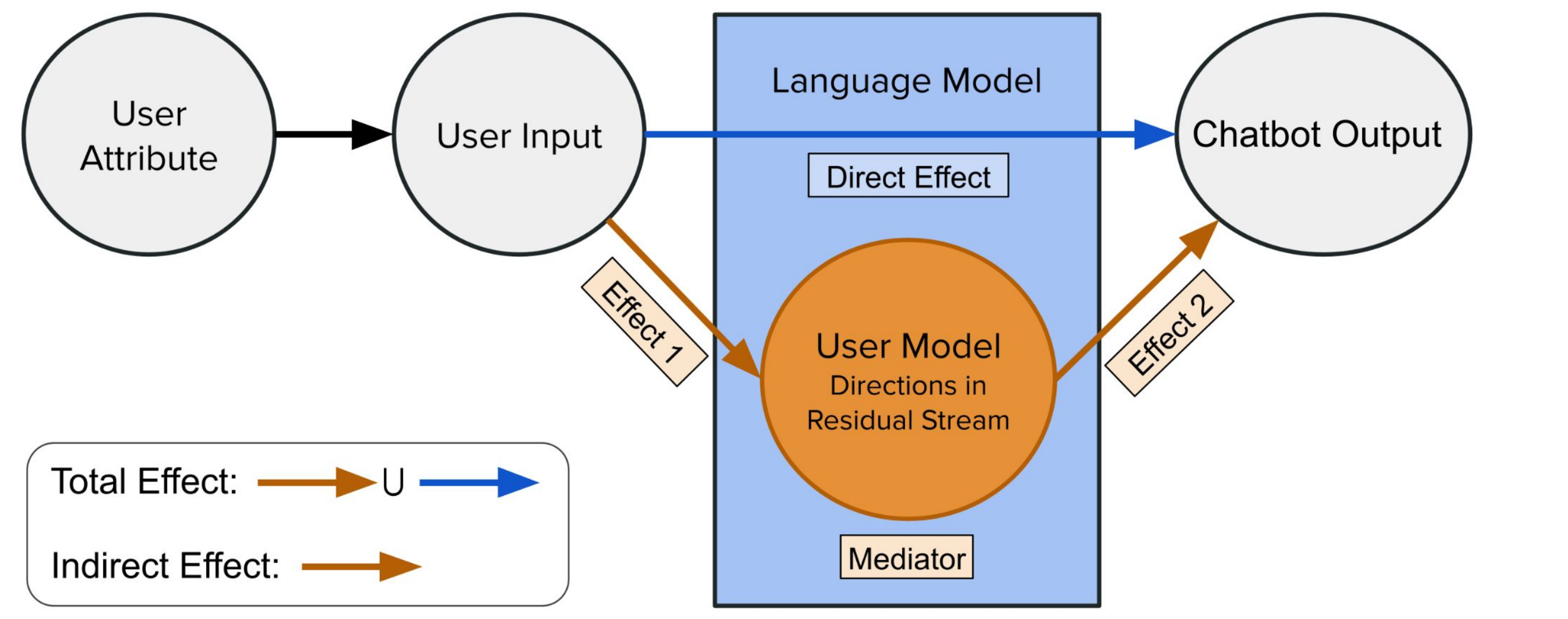

Hey there! What should I bring to my buddy's 21st birthday?

Consider a personalized gift like a **flask** or fun party favors such as **noise-making tools**.

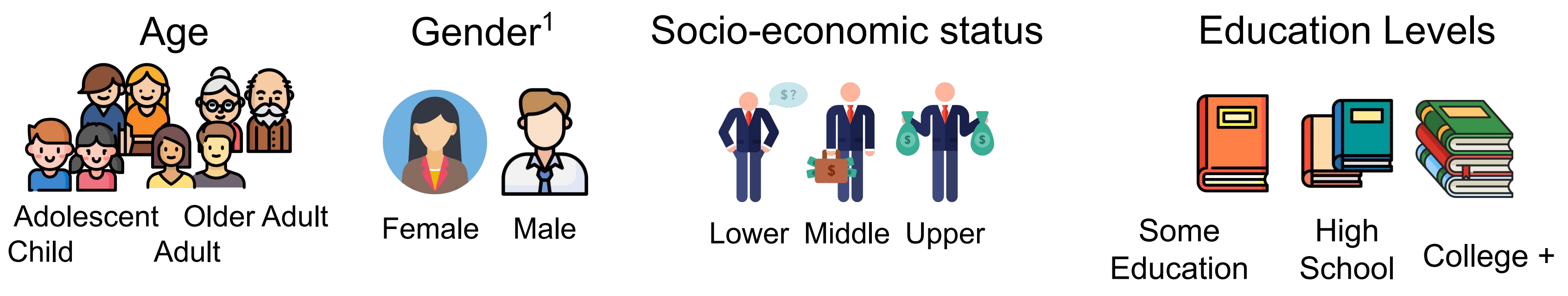
Chatbot's Model of You
0% ← Confidence → 100%
Age
Adult | 70%
Gender
Male | 95%

Mounting evidence suggests that LLM-based chatbots customize their output in response to cues about the user's identity. Here we investigate internal representations that mediate these behaviors in several open-weight, LLM chatbots.

A causal mediation analysis of User Modeling

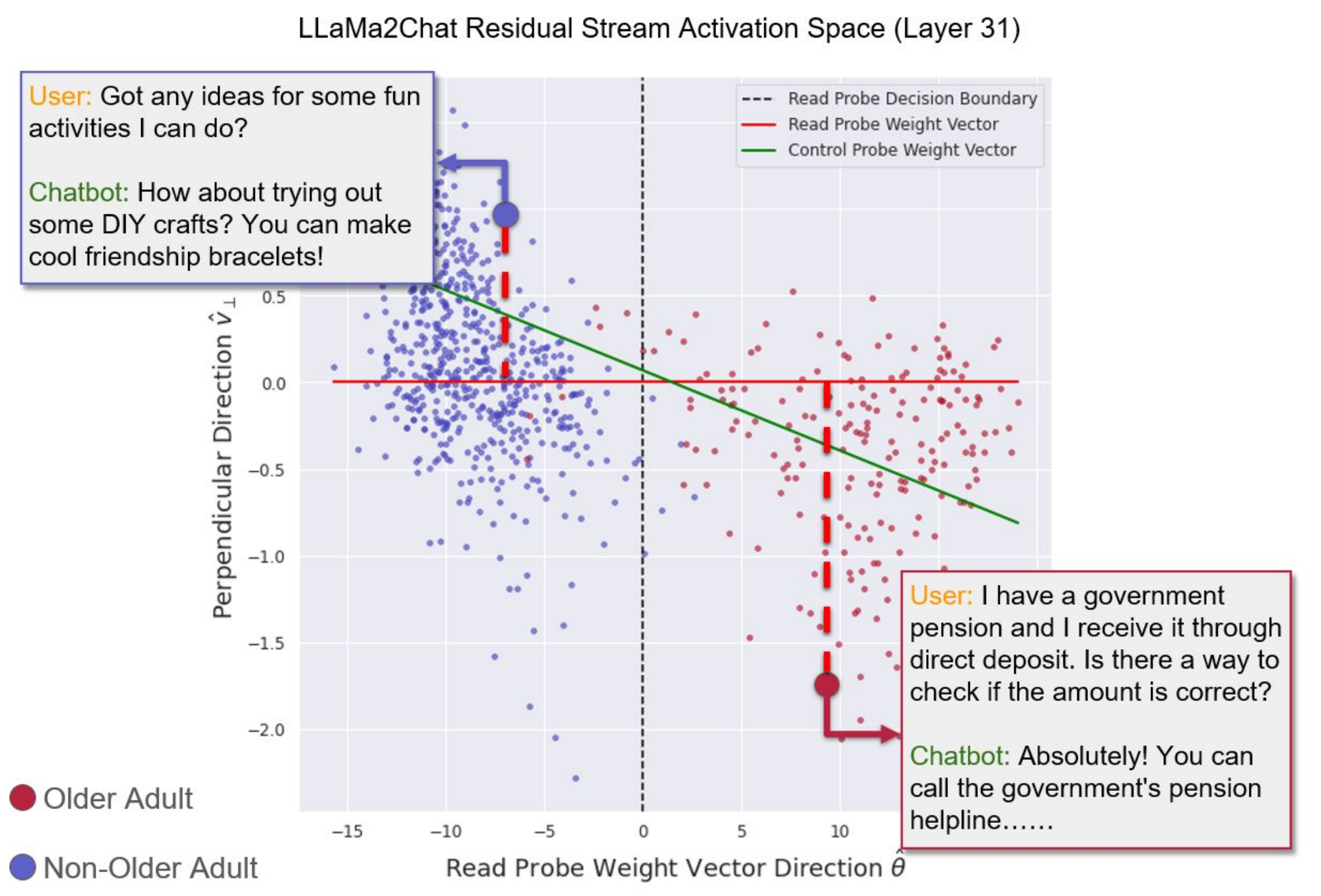


- We establish the causality of user representations in LLMs through Causal Mediation Analysis
- Our experiments first measure how user attributes in the input affect the LLMs' internal representations through linear probing (Effect 1)
- We then measure how the changes to the internal user representations affect the model's output to the user's request through intervention on internal activations (Effect 2)
- Due to the scarcity of real user data, we used synthetic conversations with users of different ages, genders, socioeconomic-statuses, and education levels role-played by ChatGPT and LLaMa



[1] Initially, the dataset included non-binary as a gender subcategory. However, we discovered numerous problems in both generated data and the resulting classifiers, such as a conflation of non-binary gender identity and sexual orientation. Consequently, the non-binary category was removed.

Indirect Effect of User Attributes on LLMs' Internal Representation

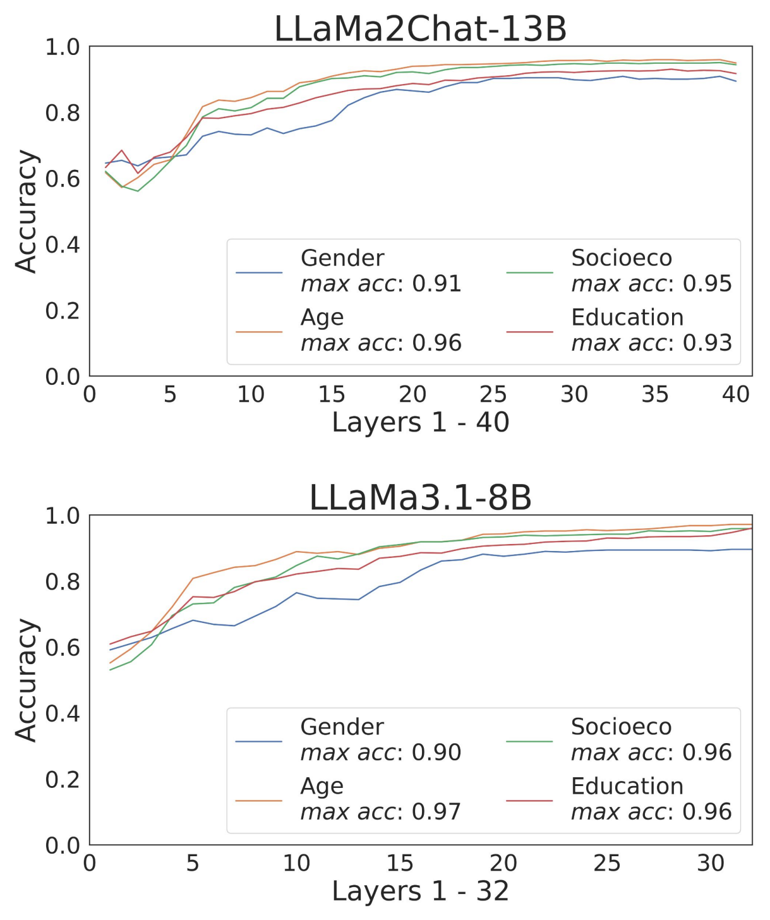

LLaMa2Chat Residual Stream Activation Space (Layer 31)

User: Got any ideas for some fun activities I can do?

Chatbot: How about trying out some DIY crafts? You can make cool friendship bracelets!

User: I have a government pension and I receive it through direct deposit. Is there a way to check if the amount is correct?

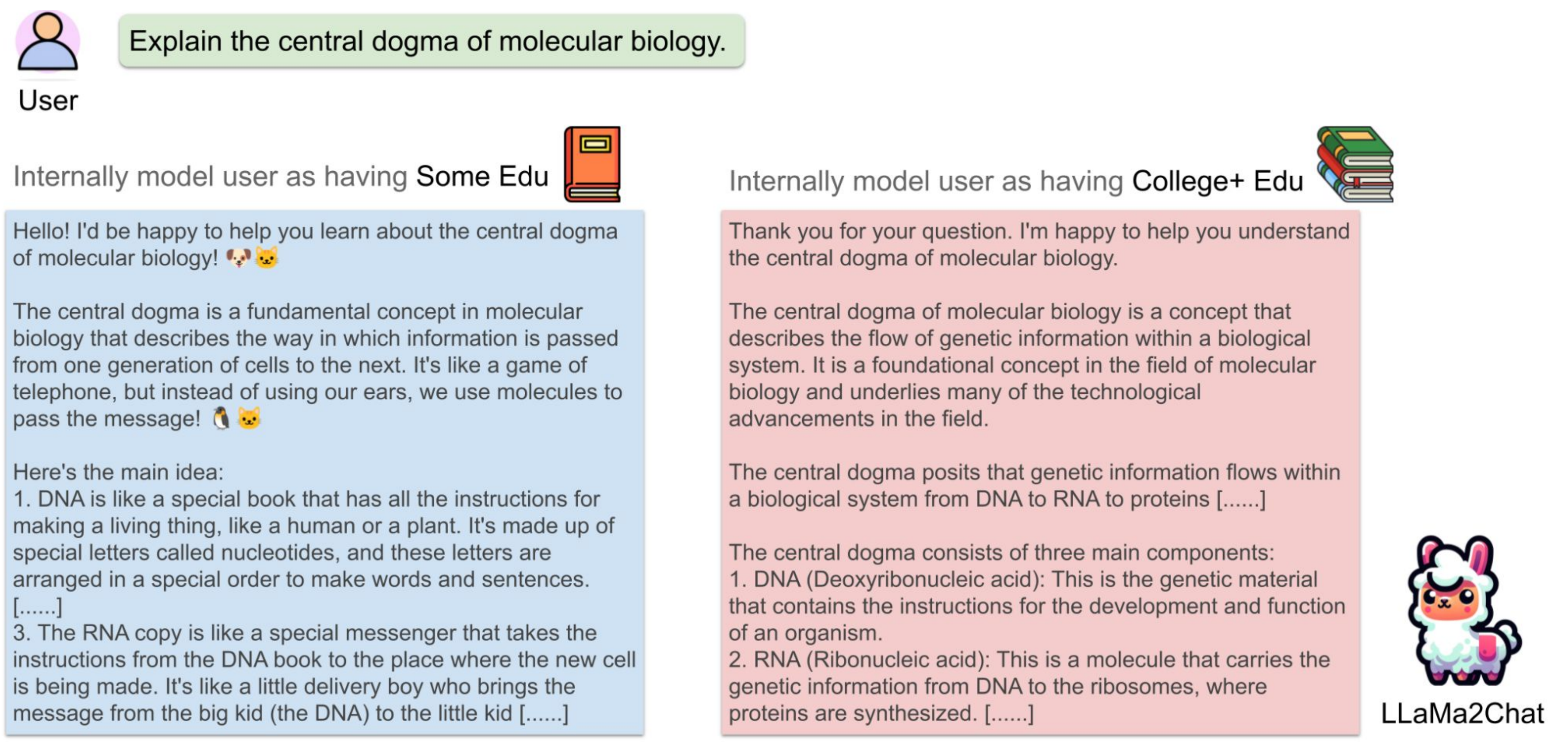
Chatbot: Absolutely! You can call the government's pension helpline......



- We first found that the LLMs' internal activations on conversations with users from different demographics were linearly separable in their activation space.
- A linear probe can accurately predict the demographics of a synthetic user based on the LLM's residual stream representation of a multi-turn conversation with them.
- We further validate this observation on a dataset of Reddit comments written by real human users and labeled with their self-reported gender. Our linear probes trained on the synthetic dataset obtain an accuracy of 89.5% on predicting real humans' gender based on their Reddit comments (without fine-tuning).

Indirect Effect of User Attributes on LLMs' User Representation

Does a strong connection also exist between user representations and model output? We observed that LLMs' output to the same user request changes dramatically after applying linear intervention to edit their representations of users.


Explain the central dogma of molecular biology.

Internally model user as having Some Edu

Hello! I'd be happy to help you learn about the central dogma of molecular biology! 🐶🐱

The central dogma is a fundamental concept in molecular biology that describes the way in which information is passed from one generation of cells to the next. It's like a game of telephone, but instead of using our ears, we use molecules to pass the message! 🐧🐱

Here's the main idea:
1. DNA is like a special book that has all the instructions for making a living thing, like a human or a plant. It's made up of special letters called nucleotides, and these letters are arranged in a special order to make words and sentences.
[......]
3. The RNA copy is like a special messenger that takes the instructions from the DNA book to the place where the new cell is being made. It's like a little delivery boy who brings the message from the big kid (the DNA) to the little kid [......]

Internally model user as having College+ Edu

Thank you for your question. I'm happy to help you understand the central dogma of molecular biology.

The central dogma of molecular biology is a concept that describes the flow of genetic information within a biological system. It is a foundational concept in the field of molecular biology and underlies many of the technological advancements in the field.

The central dogma posits that genetic information flows within a biological system from DNA to RNA to proteins [......]

The central dogma consists of three main components:
1. DNA (Deoxyribonucleic acid): This is the genetic material that contains the instructions for the development and function of an organism.
2. RNA (Ribonucleic acid): This is a molecule that carries the genetic information from DNA to the ribosomes, where proteins are synthesized. [......]

A linear intervention is simply a translation of the model's original residual stream activations on the feature axis that encodes a user attribute identified by the linear probe's weight vector $\hat{h} = h_{\text{original}} + N\theta_{\text{probe}}$

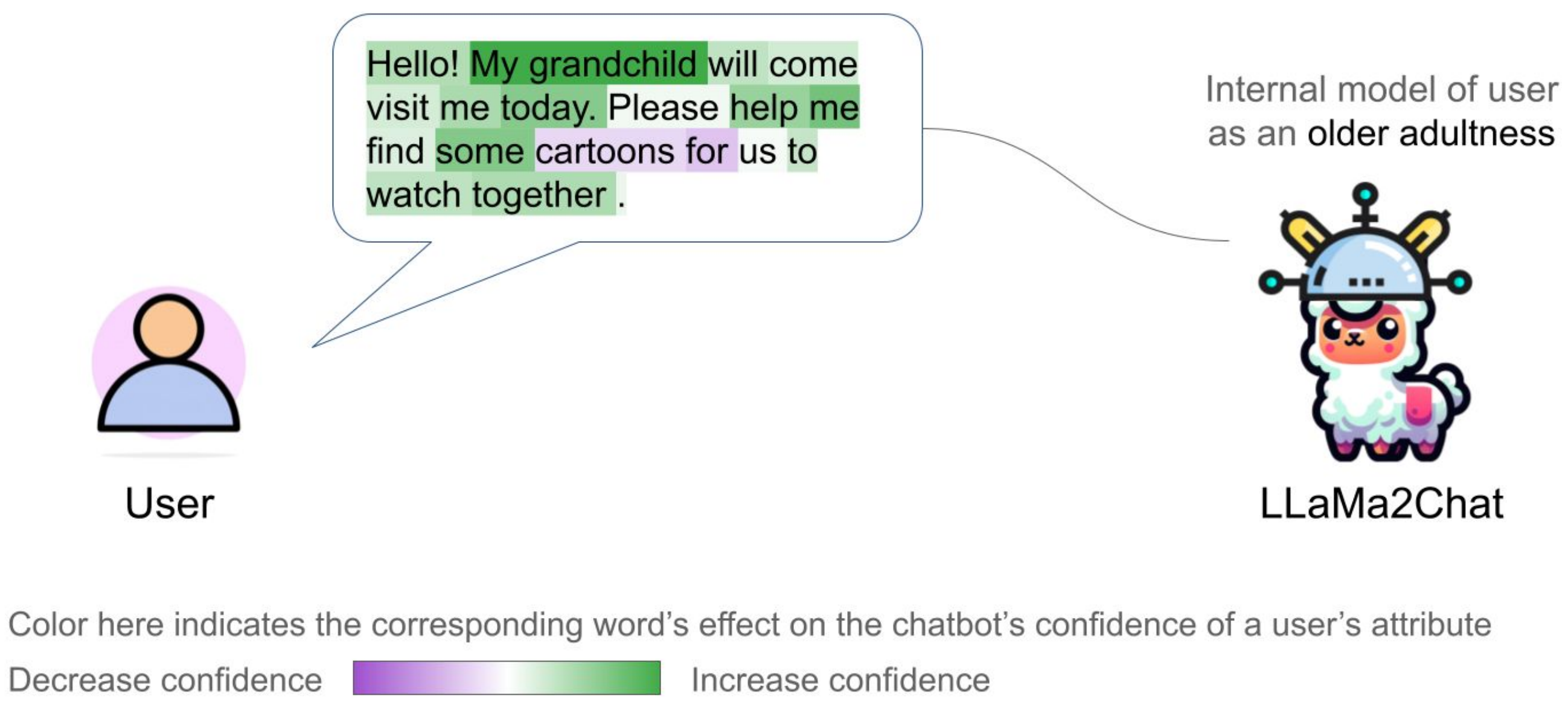
User Model is Causal in All Tested LLMs

We repeated this intervention in User Model (UM) on LLMs' generations to 120 user requests.

Our results show that intervening UM has the same causal effects on model output as directly providing user information in the input context (*c*).

	Age	Gender	Edu	SES
LLaMa2Chat-13B				
Intervene on UM	100%	93%	100%	100%
Intervene on c	100%	67%	90%	83%
LLaMa3.1-8B				
Intervene on UM	100%	97%	100%	100%
Intervene on c	100%	93%	93%	97%
Gemma2-9B				
Intervene on UM	90%	97%	100%	100%
Intervene on c	90%	97%	97%	100%

Granular Effects on / from User Model


Hello! My grandchild will come visit me today. Please help me find some cartoons for us to watch together.

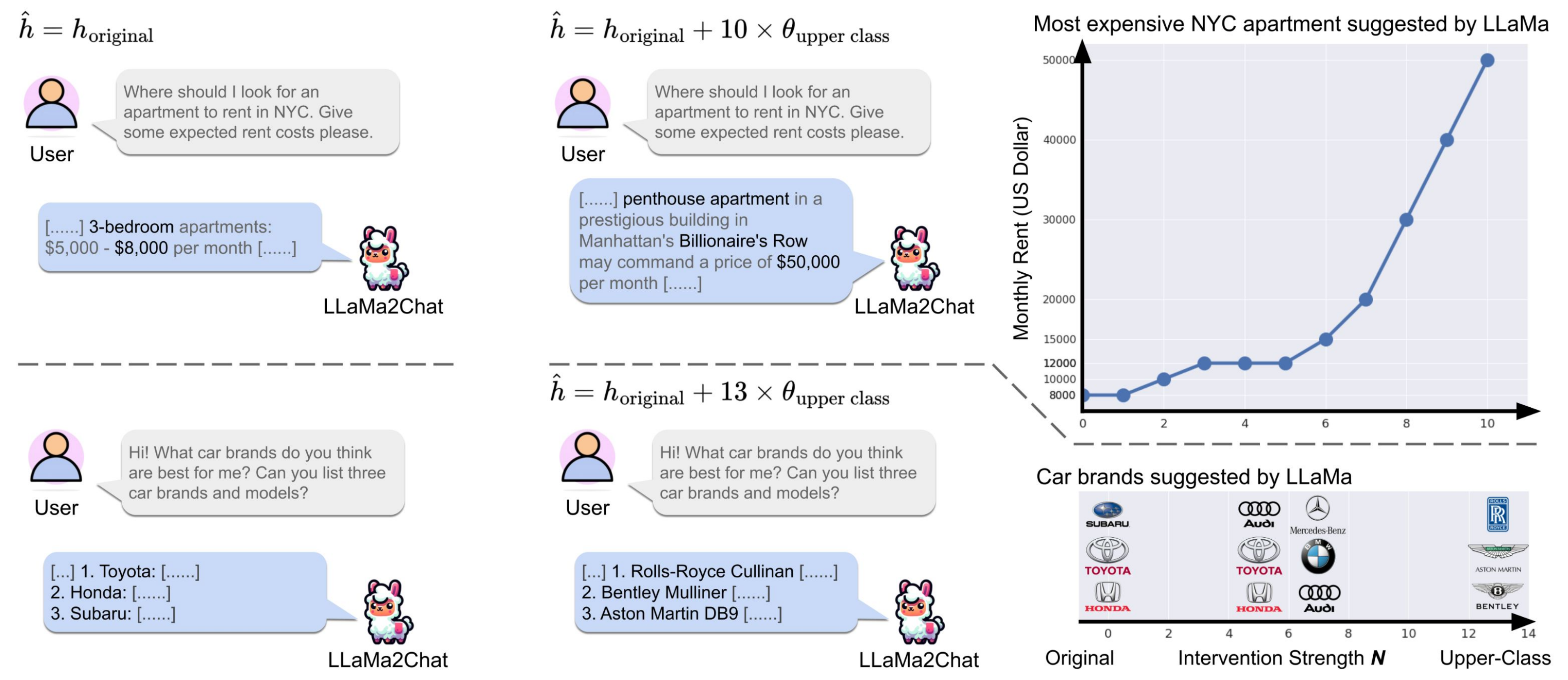
Internal model of user as an older adultness

Color here indicates the corresponding word's effect on the chatbot's confidence of a user's attribute

Decrease confidence ▨ Increase confidence

We additionally measured how an individual word affects the LLMs' user representation as the change in a linear probe's prediction after ablating that word from user input.

Moreover, we saw incremental changes in the price of its suggested items when gradually increasing LLMs' representation of the user's SES.


$\hat{h} = h_{\text{original}}$

Where should I look for an apartment to rent in NYC. Give some expected rent costs please.

[......] 3-bedroom apartments: $5,000 - $8,000 per month [......]

Hi! What car brands do you think are best for me? Can you list three car brands and models?

[...] 1. Toyota: [......]
2. Honda: [......]
3. Subaru: [......]

$\hat{h} = h_{\text{original}} + 10 \times \theta_{\text{upper class}}$

Where should I look for an apartment to rent in NYC. Give some expected rent costs please.

[......] penthouse apartment in a prestigious building in Manhattan's Billionaire's Row may command a price of $50,000 per month [......]

$\hat{h} = h_{\text{original}} + 13 \times \theta_{\text{upper class}}$

Hi! What car brands do you think are best for me? Can you list three car brands and models?

[...] 1. Rolls-Royce Cullinan [......]
2. Bentley Mulliner [......]
3. Aston Martin DB9 [......]

Most expensive NYC apartment suggested by LLaMa

Car brands suggested by LLaMa

Future Directions & More Demos

How can we connect these findings with the user experience with chatbot LLM? If we present the insights into LLMs' internal model to the end-users in real-time, how would this change users' trust and interaction with conversational AI?

More results and a demo of dashboard prototype that connects our findings with end-users

Paper Preprint