Binding Large Language Models to Virtual Personas for Human Approximation

Suhong Moon

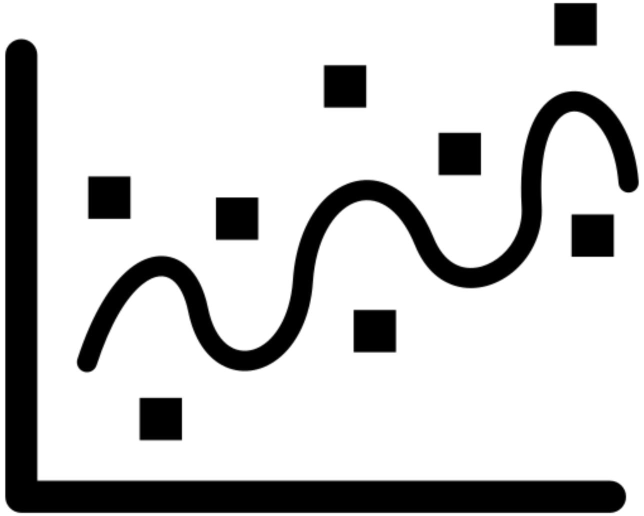
Berkeley UNIVERSITY OF CALIFORNIA

BAIR BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Core Considerations in Human Study Designs

Justice and Beneficence

Cost



Bias and Variance

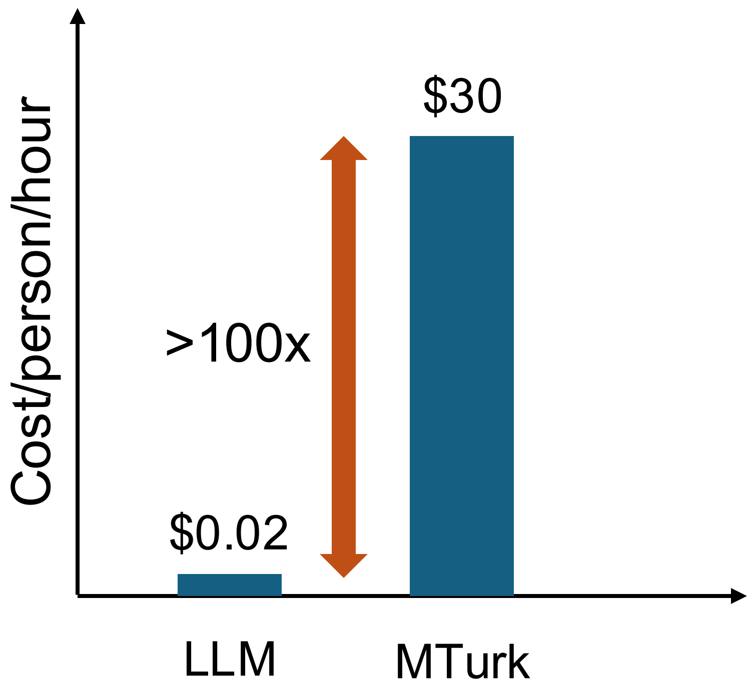
Justice and Beneficence
Belmont Principles

The Belmont Principles
(U.S. Government, 1978)

Respect for subjects	Recognize **autonomy of individuals** and informed decisions regarding participation in research
Beneficence	Researchers should **maximize benefits** and **minimize potential harm** to participants
Justice	Benefits and burdens of research should be distributed **fairly across all groups in society**

Cost

Cost Barriers in Human Studies



- Human studies are **expensive**.
- High costs **limit both the size and diversity** of participant samples.
- Budget constraints lead to **simple study designs**.

Bias and Variance
Representativeness and Validity

- Online platforms (e.g., MTurk) attract **non-representative participants**—typically younger, more educated, and more liberal.
- **Repeated participation and response familiarity** reduce data validity.
- **Small sample sizes** increase statistical variance and reduce confidence in effect estimates.
- These issues compromise the **generalizability and reliability of findings**.

Advantage of Using LLMs to Simulate Human Behaviors and Responses

- Help human study researchers satisfy best practices **without & before potential harm** to real human respondents

- LLMs as **cost-effective proxies** for pilot studies and early-stage validation.

- Provide a complementary tool to human samples by enabling more controlled, lower-variance testing before costly deployment

Potential Risks of Using LLMs as Survey Proxies

Potential Risks

- **Training Data Bias.** Internet-based corpora often overrepresent dominant groups, potentially suppressing marginalized voices.

- **False Representation of Group Beliefs.** Simulated responses may inaccurately project views onto real-world populations, leading to misleading conclusions.

Large Language Models for Simulating Human Samples in Behavioral Studies

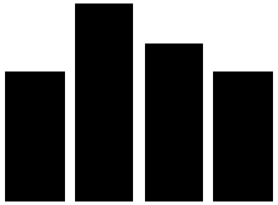
Current Challenges

When a language model is queried with an open-ended, subjective text

LLM

generation

$\sum\left(\begin{array}{c}\text{Voices}\\\text{Opinions}\end{array}\right)$

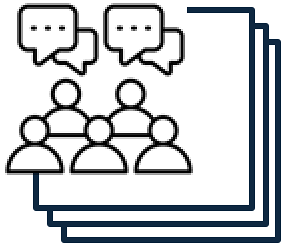


Q: How likely is it that genetically modified foods will lead to more affordably-priced food?

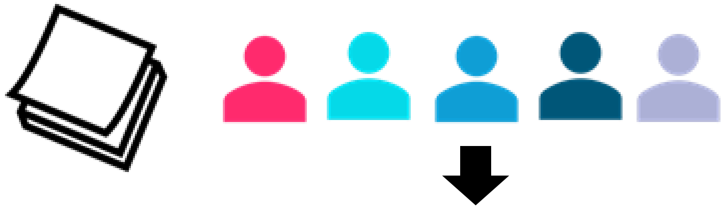
(a) Very likely
(b) Fairly likely
(c) Not too likely
(d) Not at all likely

Whose opinion is the model reflecting?
How could we condition the model to *reflect a particular individual?*

LLMs as Models of Individual Beliefs and Opinions



Training Corpora

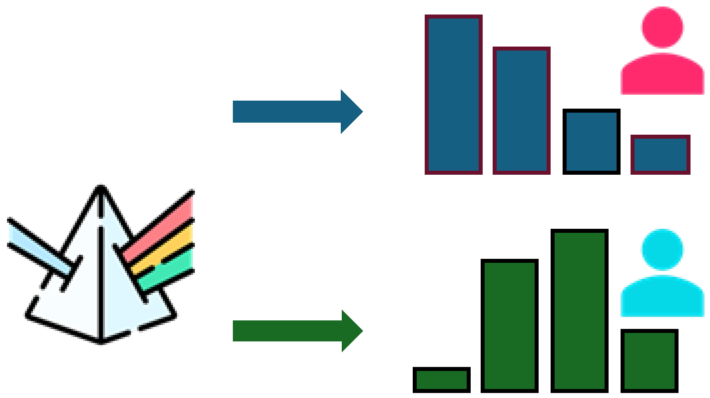


Human-authored Text

: Explicit/Implicit reference of individual values, beliefs & opinions

Pre-Training

LLM



Conditioning Text

BIO Prompt:

Below you will be asked to provide a short description of your demographic information, and then answer a question.

Description: My highest level of education is high school graduate. I consider my race as White non-Hispanic. My annual income is $100,000 to under $150,000. My age is 18-29. I consider my gender as female.

QA Prompt:

Question: Which of the following racial or ethnic groups do you identify with?
[List of Options]
Answer: (C) Hispanic

…

Question: What is your annual household income?
[List of Options]
Answer: (G) $75,000 to under $100,000

Prior Approach on **Persona-Conditioned generation:**

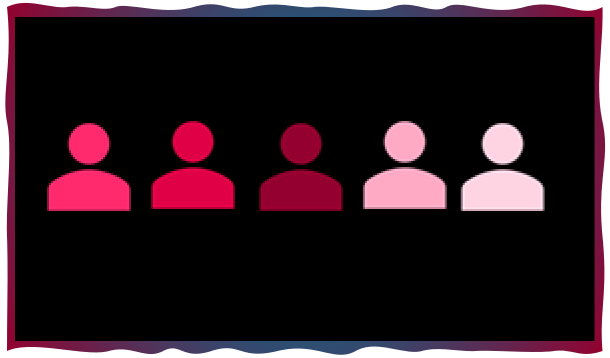
Programmatic Concatenation of Demographic Variables as Conditioning Text

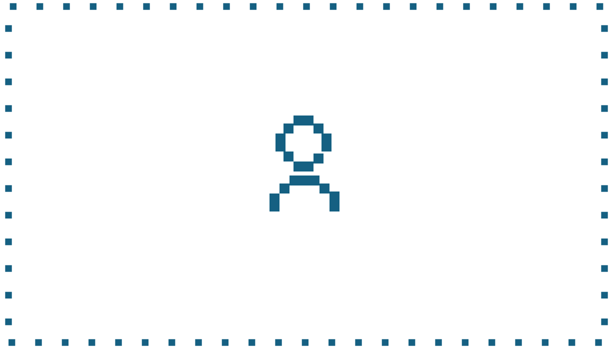
Prior Approaches for Conditioning Virtual Subjects

Human Subjects
- Combinations of Demographic Variables
(Age, Gender, Race, Income, Education, etc.)

Virtual Subjects
- Construct <u>population-level</u> virtual personas

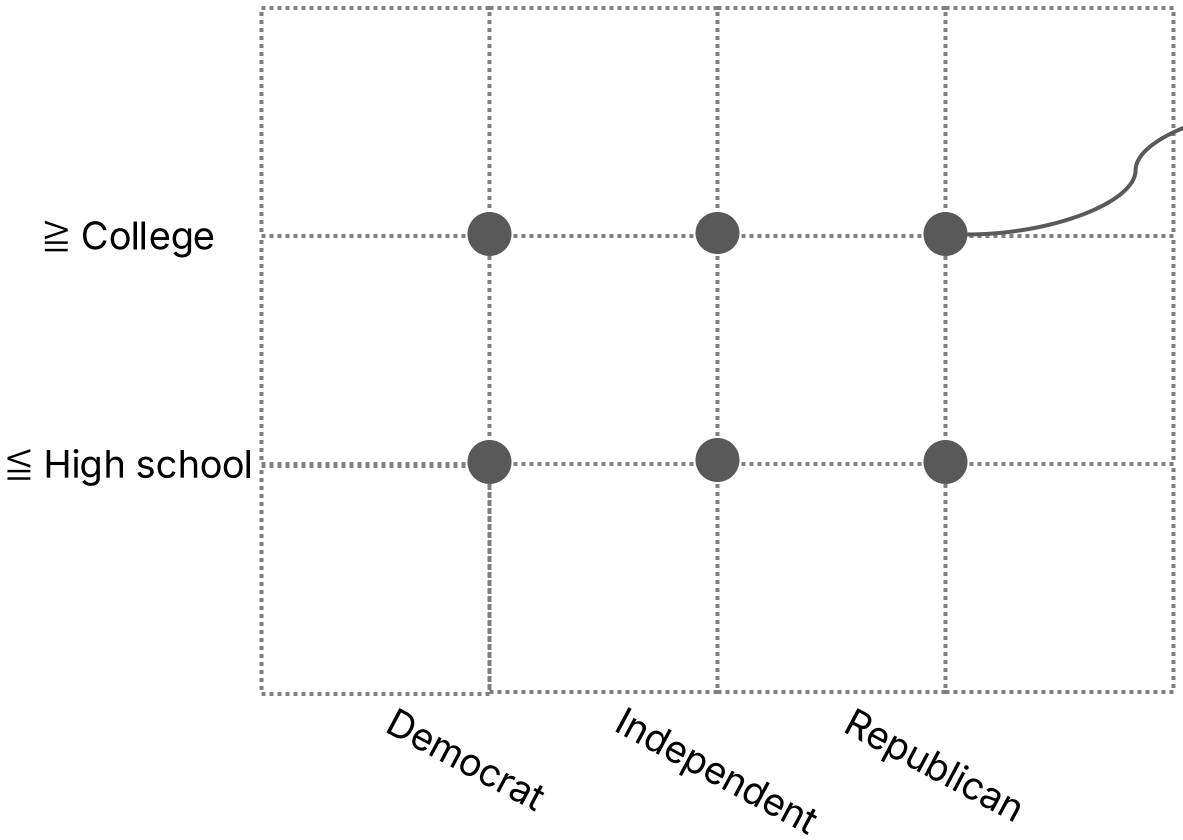
Age: 18-29
Education: high school graduate
Race/ethnicity: White non-Hispanic
Gender: Female
Income: $100,000 -$150,000





Many individual human samples in population represented by <u>same</u> **population-level** virtual subject

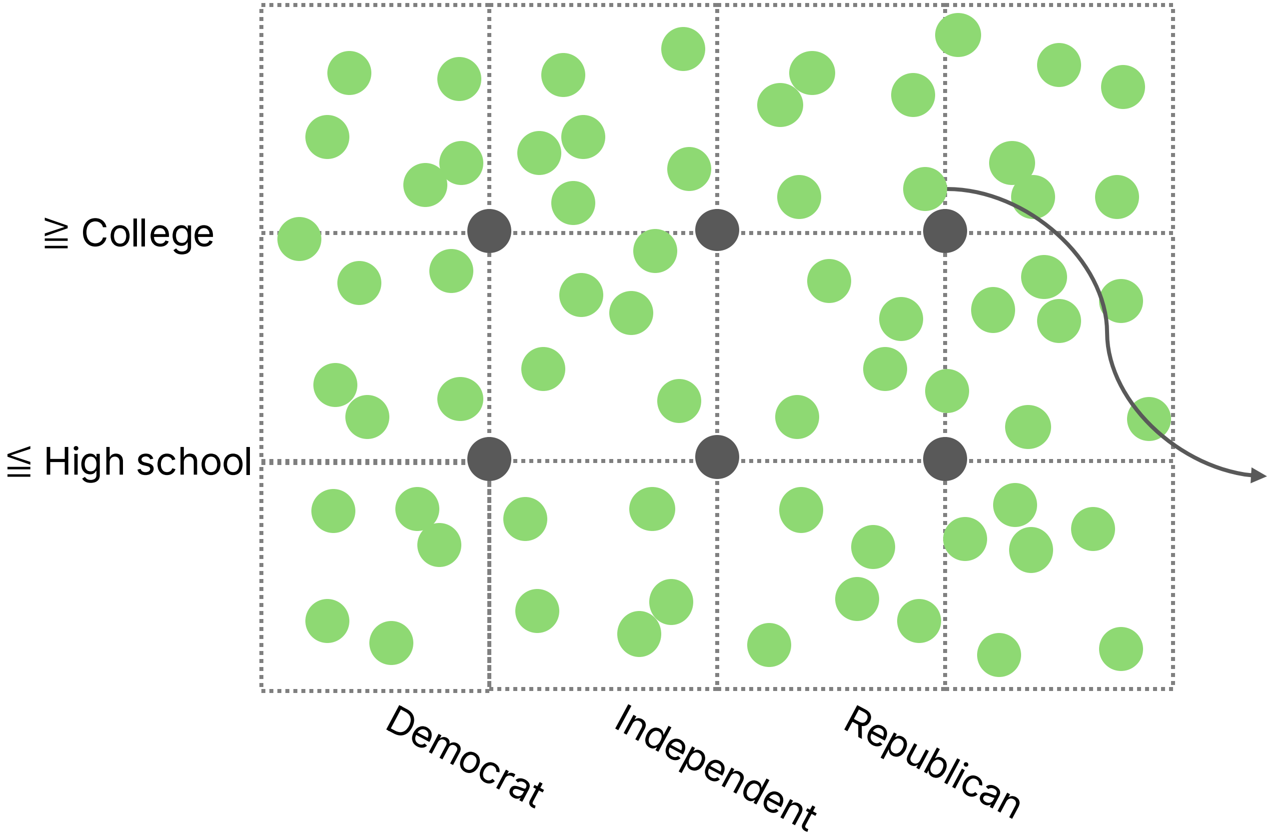
Limitation of Prior Approaches



Programmatically generate conditioning text
E.g. I am a Republican. My education level is ≧ college.

- **Pre-defined taxonomy** of defining subjects or populations → Models are prone to generating **caricatures & stereotypical** responses
- Without individual responses, cannot **estimate various statistics of interest** (e.g., covariance, effect size)

Design Principle
Sampling diverse and naturalistic user



Q. How can we condition LLMs with more diverse and lively users?

A. Naturalistic and unconstrained description of oneself.

Example.

...
Politically speaking, I am **conservative.** To put it simply, I prefer smaller government and lower taxes. However, I am **more liberal when it comes to social issues**. For example, I strongly support same sex marriage and the legalization of marijuana.
...

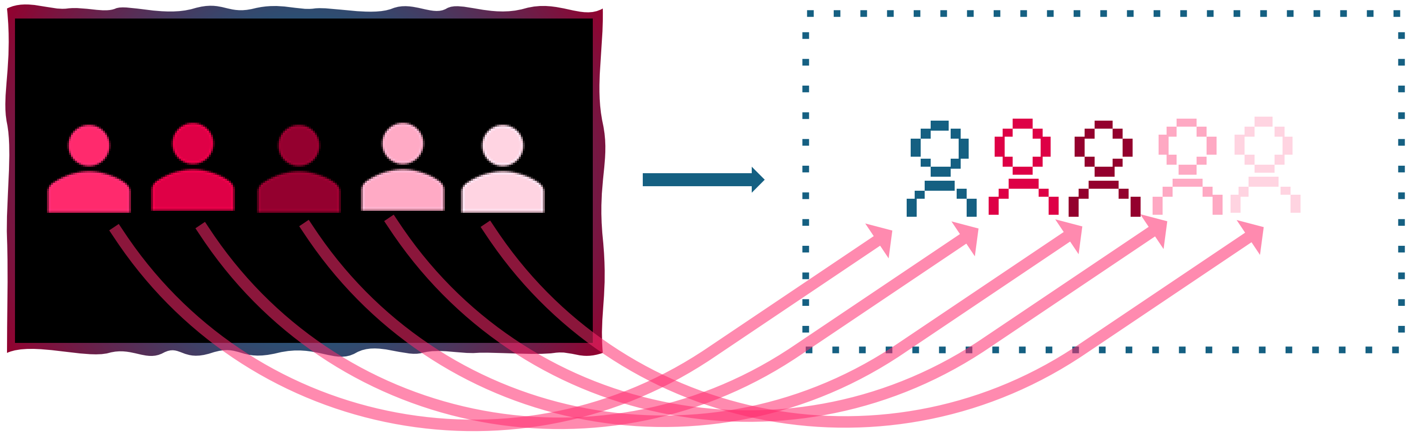
How can we create **realistic and diverse virtual personas** that accurately **simulate humans** in public opinion surveys with LLMs?

Virtual Personas for Language Models via an Anthology of Backstories in *EMNLP 2024*
Suhong Moon*, Marwa Abdulhai*, Minwoo Kang*, Joseph Suh*, Widyadewi Soedarmadji, Eran Kohen Behar, David M. Chan, John Canny

Proposal for Individual-Level **Virtual Subjects**

Human Subjects → ***Individual-Level* Virtual Subjects**

Age: 18-29
Education: high school graduate
Race/ethnicity: White non-Hispanic
Gender: Female
Income: $100,000 -$150,000



Promises:

- **Single (individual)** sample responses
- Estimate **covariance**, **statistical significance,** and other statistics critical for study

> **Research Question:**
> How can we condition LLMs to *individual virtual personas* that are representative, consistent, and diverse?

Ours: the *Anthology* Approach for Conditioning LLMs to **Virtual Personas**

1 Generate Synthetic yet Naturalistic, Open-ended Narratives ("Backstories") from LLMs

Backstory Prompt:
Question: Tell me about yourself.

Explicit hints about the author's demographics and traits

But also, a natural, authentic story portraying author's values, desires, attitudes, etc.

LLM

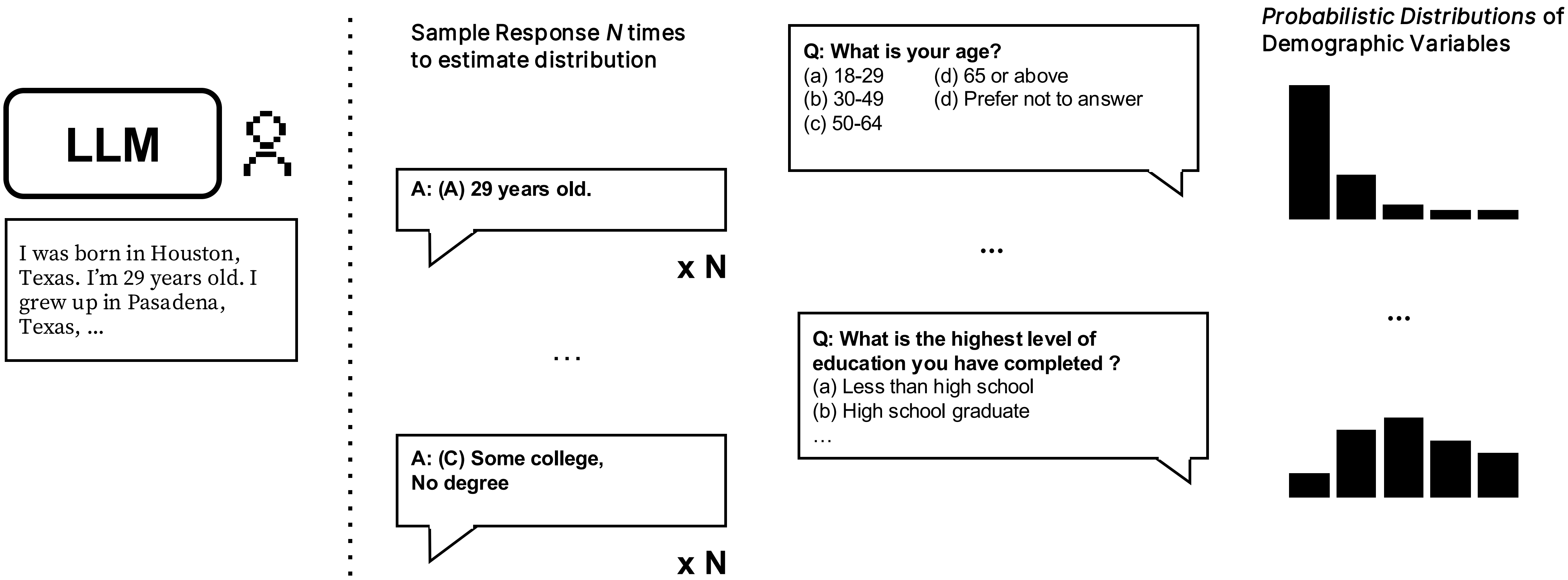
Anthology of Backstories

Answer: I am in my 60s and live in the same neighborhood I have always lived in. I am not rich and by some standards might even be considered homeless. However, I could spend thousands of dollars more per month if I wanted. I am happy with my life style. I am from the backwoods of this country and grew up with very little. On a few occasions, we were starving in the woods and going to school on an empty stomach. We had a small brown paper bag for dinner a couple of nights every week. Breakfast on some days was just a big bowl of Kool-Aid™ mixed with powdered milk. My two brothers were thin and we worried about them catching a cold ...

On the day before payday, my mother would spend my whole allowance in the grocery store because she just could not resist those long stems of red roses for only 29 cents a stem. I would have rather had bread and milk for dinner, but I did not dare protest because I did not want to take them away from her. We were lucky to have 79 cents to last until payday ...

Ours: the *Anthology* Approach for Conditioning LLMs to **Virtual Personas**

2 Demographic Survey on Virtual Personas Conditioned by each Backstory

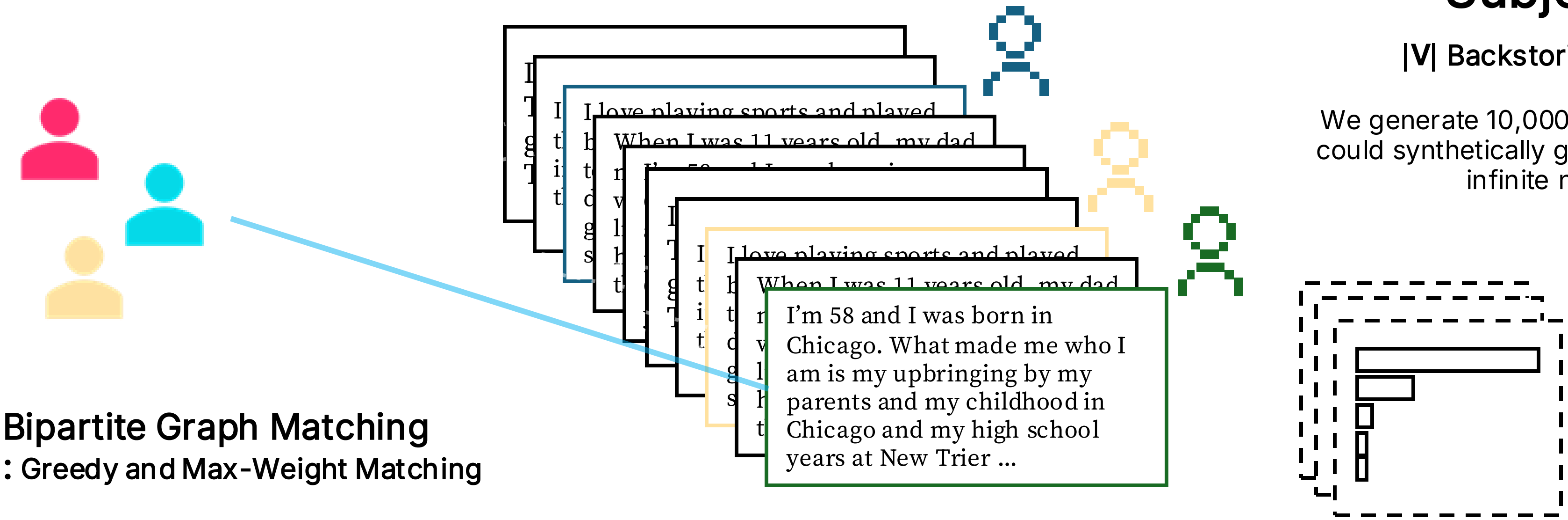


Ours: the *Anthology* Approach for Conditioning LLMs to **Virtual Personas**

③ Match Virtual Personas to Human Subjects

Human Subjects

|H| Subjects



Bipartite Graph Matching
: Greedy and Max-Weight Matching

Virtual Subjects

|V| Backstories >> |H|

We generate 10,000 backstories but could synthetically generate virtually infinite numbers of them

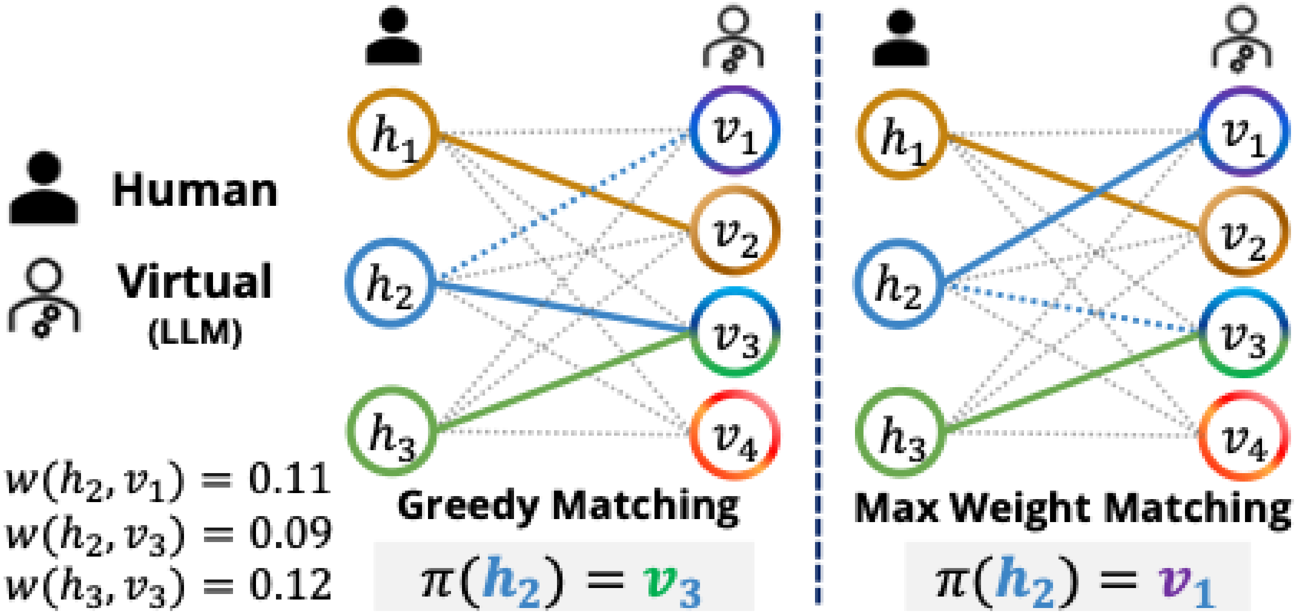
Demographic Distributions

Key Idea
Generate rich, naturalistic backstories to condition LLMs

- **Diverse, *non-prototypical*** descriptions of individuals: conditioning beyond population-level
- **Longer, detailed prompt** to strongly condition LLM response generation

Ours: the *Anthology* Approach for Conditioning LLMs to **Virtual Personas**

③ Match Virtual Personas to Human Subjects



- **Build bipartite graphs.** Each edge is the product of the likelihoods of the corresponding human user's demographic traits.
- **Matching algorithm.** We try two algorithms: greedy matching and max weight matching (Hungarian algorithm)

Experiments
Approximating Large-Scale Public Opinion Polls

Evaluation Benchmarks
Pew Research Center's American Trends Panel (ATP) Survey: public survey on real-world topics:

- Wave 34: *Biomedical* and *Food* Issues
- Wave 92: *Political Typology*
- Wave 99: *Artificial Intelligence* and *Human Enhancement*

Evaluation Metrics
Assessing the LLM's output for **representativeness** and **consistency** with human responses.

- **Representativeness**
  - Goal: Measure how closely the LLM approximates **individual** human responses.
  - Method: Calculate the <u>Wasserstein Distance (WD)</u> between human response distributions and virtual persona response distributions.
- **Consistency**
  - Goal: Evaluate the similarity in response correlations between LLM responses and human responses.
  - Method: Compute the <u>Frobenius Norm (Fro.) between the correlation matrices</u> of human and virtual subject responses.

Experiments
Baseline Methods

Following Santurkar et al., 2023, baseline prompting methods construct virtual personas using available demographic information in a rule-based manner.

BIO Prompting
Generate free-text biographies incorporating demographic details.
(e.g., I am a Republican).

QA Prompting
Provide a sequence of question-answer pairs for each demographic variable.
(e.g., Q: What is your political affiliation? A: Republican).

Santurkar, Shibani, et al. "Whose opinions do language models reflect?." ICML, 2023.

Results
Approximating Pew Research Center ATP Surveys

Main Results

Model	Persona Conditioning	Persona Matching	ATP Wave 34 WD (↓)	ATP Wave 34 Fro. (↓)	ATP Wave 34 α (↑)	ATP Wave 92 WD (↓)	ATP Wave 92 Fro.(↓)	ATP Wave 92 α (↑)	ATP Wave 99 WD (↓)	ATP Wave 99 Fro.(↓)	ATP Wave 99 α (↑)
Llama-3-70B	Bio	n/a	0.254	1.107	0.673	0.348	1.073	0.588	0.296	0.809	0.733
	QA	n/a	0.238	1.183	0.681	0.371	1.032	0.664	0.327	0.767	0.740
	Anthology (DP)	n/a	0.244	1.497	0.652	0.419	0.965	**0.636**	0.302	1.140	0.669
	Anthology (NA)	max weight	0.229	1.287	0.693	0.337	1.045	0.637	0.327	**0.686**	**0.756**
		greedy	**0.227**	**1.070**	**0.708**	**0.313**	**0.973**	0.650	**0.288**	0.765	0.744
Mixtral-8x22B	Bio	n/a	0.260	1.075	0.698	**0.359**	0.851	0.667	0.237	1.092	0.687
	QA	n/a	0.347	1.008	0.687	0.429	0.911	0.599	0.395	1.086	0.684
	Anthology (DP)	n/a	**0.236**	1.095	0.684	0.378	**0.531**	0.624	**0.215**	1.422	0.604
	Anthology (NA)	max weight	0.257	0.869	**0.726**	0.408	0.846	0.610	0.353	**0.843**	**0.729**
		greedy	0.247	**0.851**	0.715	0.392	0.981	**0.627**	0.320	0.951	0.710
Human			0.057	0.418	0.784	0.091	0.411	0.641	0.081	0.327	0.830

- **Anthology** outperforms baseline methods (Bio and QA) in **representativeness** and **consistency** across all waves and models
  - better approximation to human responses.
- But each metric remains *above the lower-bound levels* presented in the last row.

Results
Approximating Pew Research Center ATP Surveys

Per-Subgroup Survey Approximation Results

Method	Race						Age Group								
	White			Other Racial Groups			18-49			50-64			65+		
	WD (↓)	Fro. (↓)	α (↑)	WD (↓)	Fro. (↓)	α (↑)	WD (↓)	Fro. (↓)	α (↑)	WD (↓)	Fro. (↓)	α (↑)	WD (↓)	Fro. (↓)	α (↑)
Bio	0.263	**1.187**	<u>0.687</u>	0.335	0.955	0.651	0.244	<u>1.163</u>	0.673	0.277	1.382	0.659	<u>0.318</u>	<u>1.000</u>	<u>0.686</u>
QA	<u>0.250</u>	1.259	0.678	<u>0.323</u>	<u>0.828</u>	<u>0.687</u>	<u>0.229</u>	**1.091**	<u>0.695</u>	<u>0.258</u>	<u>1.220</u>	<u>0.695</u>	0.329	1.204	0.630
Anthology	**0.233**	<u>1.216</u>	**0.703**	**0.311**	**0.778**	**0.719**	**0.200**	1.193	**0.702**	**0.242**	**1.215**	**0.710**	**0.303**	**0.943**	**0.704**
Human	0.063	0.519	0.777	0.094	0.413	0.764	0.077	0.663	0.779	0.092	0.741	0.803	0.102	0.772	0.766

- **Anthology** outperforms baseline methods in representativeness and consistency across <u>all demographic subgroups</u>.

- All methods, including **Anthology**, show better performance on the *White* respondent subgroup and *younger age groups*, indicating potential disparities in LLM alignment across demographic subgroups.

Results
Approximating Pew Research Center ATP Surveys

Other Models

Model	Persona Conditioning	Persona Matching	ATP Wave 34 WD (↓)	ATP Wave 34 Fro. (↓)	ATP Wave 34 α (↑)
Llama-3-70B-Instruct	Bio	n/a	0.462	2.177	0.445
	QA	n/a	0.422	1.560	0.581
	Anthology (DP)	n/a	0.461	1.295	0.511
	Anthology (NA)	max weight	0.429	1.776	0.714
		greedy	0.413	1.848	0.754
Mixtral-8x22B-Instruct	Bio	n/a	0.532	1.608	0.632
	QA	n/a	0.567	1.583	0.628
	Anthology (DP)	n/a	0.464	1.652	0.646
	Anthology (NA)	max weight	0.478	1.606	0.635
		greedy	0.472	1.593	0.640
gpt-3.5-0125	Bio	n/a	0.414	2.009	0.481
	QA	n/a	0.422	1.560	0.581
	Anthology (DP)	n/a	0.476	1.963	0.486
	Anthology (NA)	max weight	0.450	1.905	0.472
		greedy	0.443	1.936	0.468
Llama-3-8B	Bio	n/a	0.454	1.480	0.683
	QA	n/a	0.432	0.924	0.779
	Anthology (DP)	n/a	0.383	1.323	0.714
	Anthology (NA)	max weight	0.395	1.265	0.735
		greedy	0.416	1.229	0.717
Human			0.057	0.418	0.784

- In contrast to the conventional wisdom, chat models—despite their strong reasoning performance—perform **worse than the smallest base model** for this task.
- This result emphasizes that approximating real humans' opinions is a different challenge from reasoning tasks.

Results
Ablation Study: Does Matching Really Matter?

Ablation Study Results

Model	Method	ATP Wave 34 WD (↓)	ATP Wave 34 Fro. (↓)
Llama-3-70B	random	0.270	1.362
	max weight	0.229	1.287
	greedy	0.227	1.070
Mixtral-8x22B	random	0.274	0.814
	max weight	0.257	0.869
	greedy	0.247	0.851

Compared to the random matching baseline, both max-weight and greedy matching achieve **lower Wasserstein distance** between response distributions and **lower Frobenius norm** between correlation matrices.

Results

How Diverse the *Anthology* Backstories Are?

Qualitative Analysis

Backstory Example 1-1

I'm a 30 year old white male. I was born and raised in Tennessee, in a small town called Clarksville. It's in the middle of nowhere. I was the kid that wasn't picked for any sports team. I wasn't a nerd. I was the kid that didn't talk to anybody. I still have no friends and I still don't talk to anybody. I was homeschooled and have never had a job. I do have a GED and I did take a few community college classes. I have been out of school since 2011. My life has been filled with anxiety and depression. I've had a few minor issues in my life, like when I was a child and I was abused.

Notable Observations:
- Discusses mental health issues and past trauma
- Mentions age (30), racial identity (white) and hometown (Clarksville, Tennessee)
- Mentions education level (GED and some community college)

Backstory Example 1-2

I am 54 years old, and I grew up in Oceanport, New Jersey. I am a lifelong fan of the New York Yankees, the Buffalo Bills, and the NY Rangers. I have been a music fan for as long as I can remember. My musical interests are vast but the music I gravitate toward the most is the classic rock bands from the late '60s and early '70s. This is just my age and what I gravitate to. I also love Blues, Jazz, Soul and R&B. I have been involved in music in many aspects of it from my teenage years to now. I was the lead singer of a band in the early '80s and I've been in the marketing department at SiriusXM for 15 years. I've also spent the last 15 years coaching Little League baseball and traveling sports with my boys. I enjoy spending time with my friends and family, going to dinner and watching movies. I really don't have much of a life outside of my family. It's just the way I like it. There are times when it drives me nuts, but I wouldn't trade it for anything.

Notable Observations:
- Mentions significance of family and friends to the author
- Mentions age (54 years old) and hometown (Oceanport, New Jersey)
- Mentions previous occupations (lead singer of a band, marketing department, Little League baseball coach)

Backstory Example 1-3

I'm 31 and grew up in SC with my parents and three brothers. I was a ballerina for 11 years. I loved exploring the world with my family. My life was very close knit. I loved being a dancer and enjoyed school very much. I went to Ohio State University in Columbus, Ohio where I majored in Criminology and Spanish. I graduated with honors and then earned a masters in Criminal Justice and Peace and Justice. I spent a year volunteering in Costa Rica. I've always been drawn to teach and be in the academic setting. I've only taught high school for 1.5 years. I thoroughly enjoy my students and consider them my community. I live and eat with them. I give them more than just content in history, I give them a life lesson. I want to build these students into model citizens who are contributing to a better community everyday and I know teaching can help me achieve that. My students come from a wide diversity of backgrounds. They are in public school because their families have extremely low incomes. Most of my students come from single parent households, some do not even have one. My students don't have mentors at home who can help them navigate what college will look like or what career will be the best for them. A lot of my students have never been outside of the county or traveled at all. Students like these need something to look forward to and teaching is exactly that. I want to inspire them to be greater than they have ever imagined they could be. I want them to know they have other people in this world rooting for them and helping them achieve the things they say they want. I want them to know that the decisions they make today will change their tomorrow.

Notable Observations:
- At length, discusses aspirations for teaching, mentorship, and contributing to community
- Mentions age (31), home state (SC), upbringing with family, and hobby/past occupation (ballet)
- Mentions education (undergraduate and master's degree at Ohio State University)

Can LLMs **simulate higher-order social cognition** by conditioning on virtual backstories?

Higher-Order Binding of Language Model Virtual Personas: a Study on Approximating Political Partisan Misperceptions *under submission to COLM 2025*

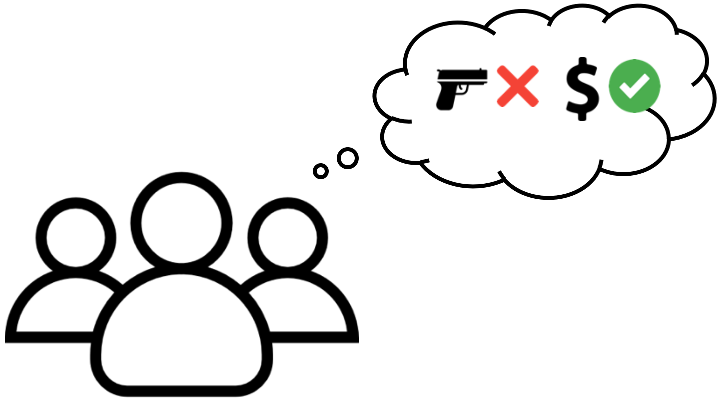
Minwoo Kang*, **Suhong Moon***, Seung Hyeong Lee, Ayush Raj, Joseph Suh, David M. Chan, John Canny

Human Identity Is Intrinsically Relational

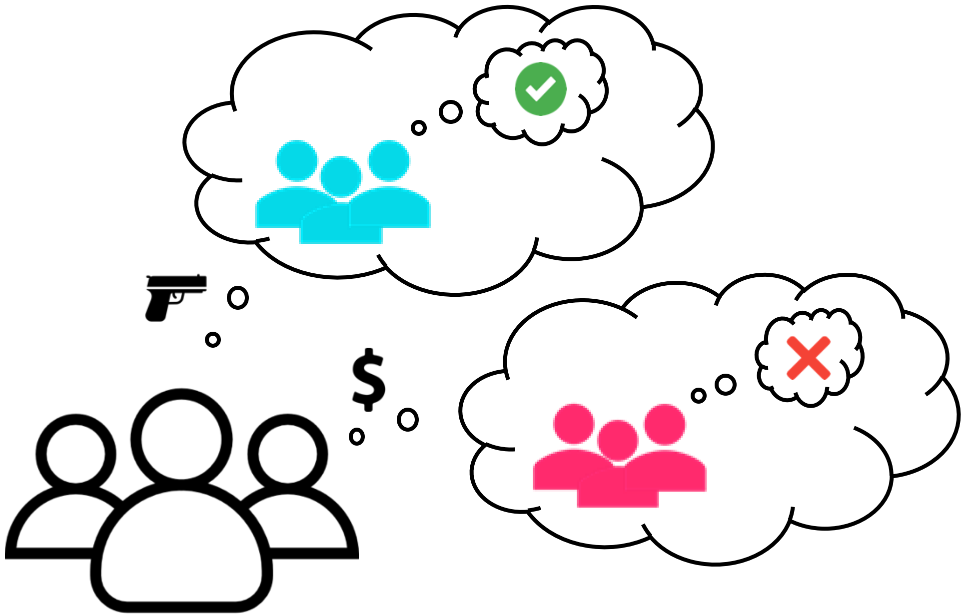
"I am not what I think I am; I am not what you think I am. I am what I think you think I am."

— Charles Horton Cooley, *Human Nature and the Social Order (1902)*

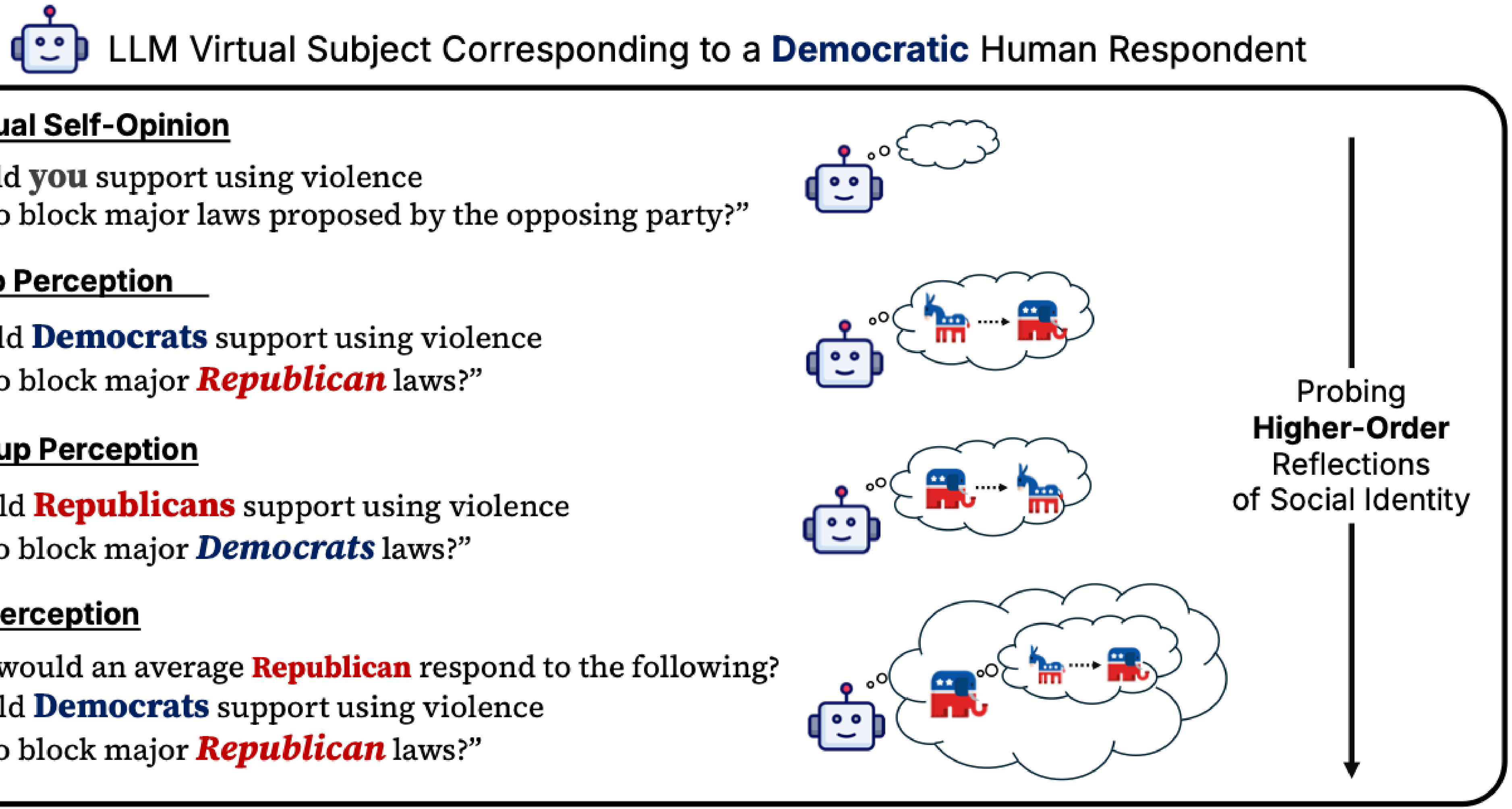
Previous Work: First-Order Opinion Approximation



Not Explored: Higher-Order Social Perception



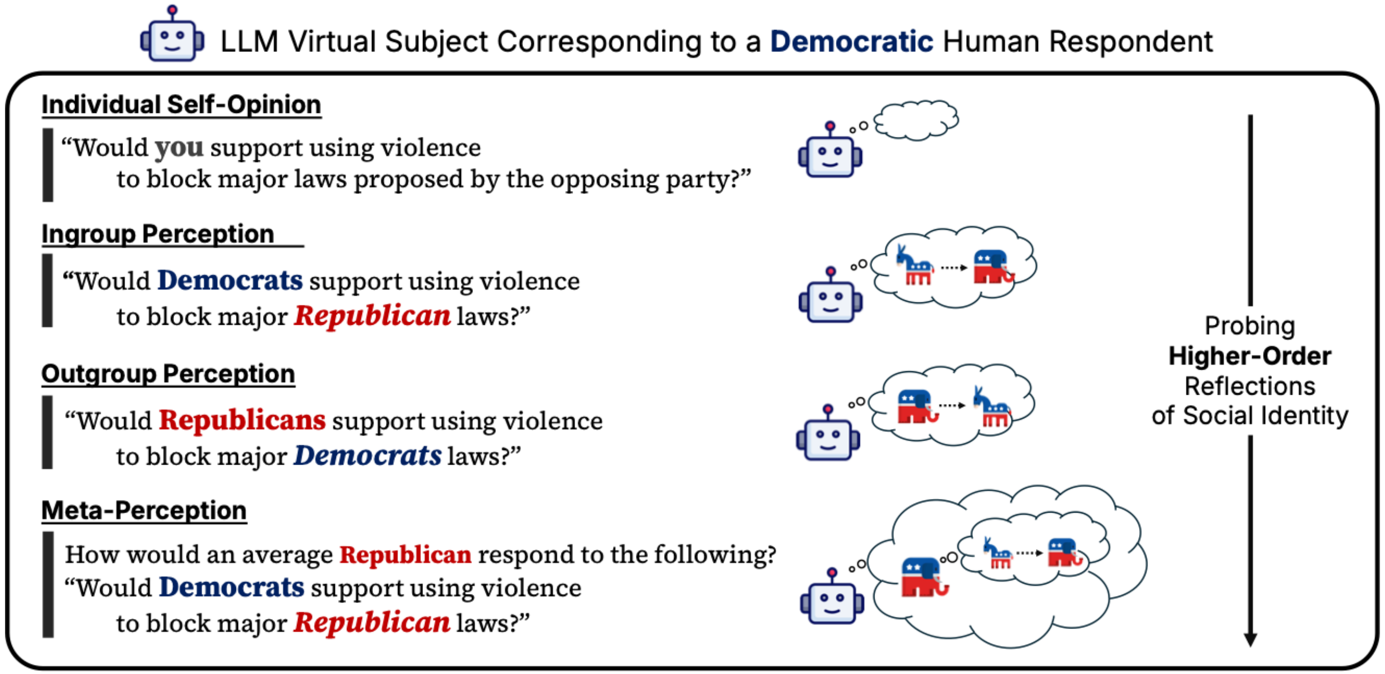
What is Higher-Order Binding of LLMs?



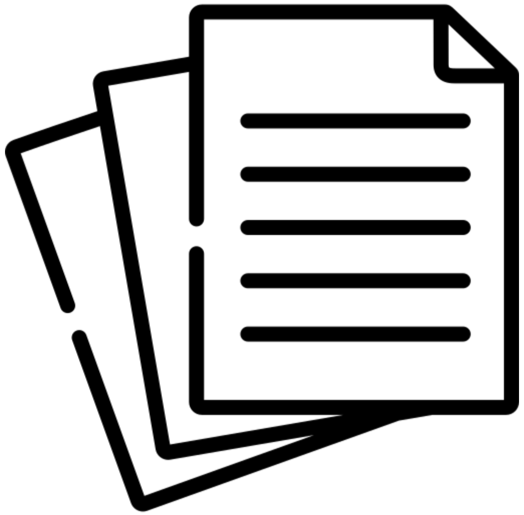
Towards Higher-Order Binding of LLMs

Why Higher-Order Binding?

- It's essential for capturing how people relate to social groups—not just what they believe individually.
- Enables broader use of LLMs in human studies beyond simple opinion polls.
- Acts as a litmus test for whether virtual personas truly reflect human-like social reasoning.

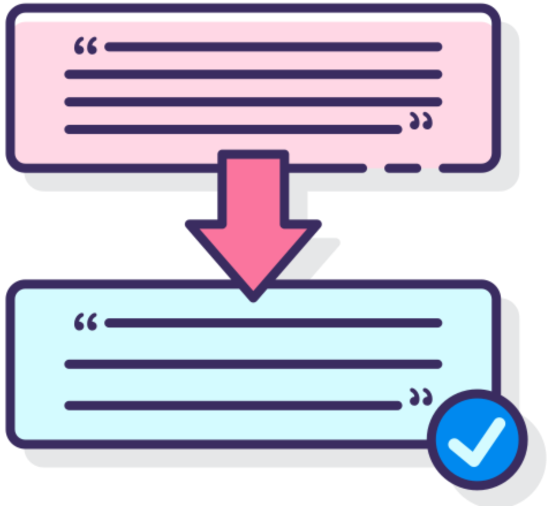


What Matters in Binding LLMs to Virtual Personas?



H1. Quantity
More Backstories enable better matching of virtual personas to human subjects

H2. Depth
Longer backstories provide richer context for an individual

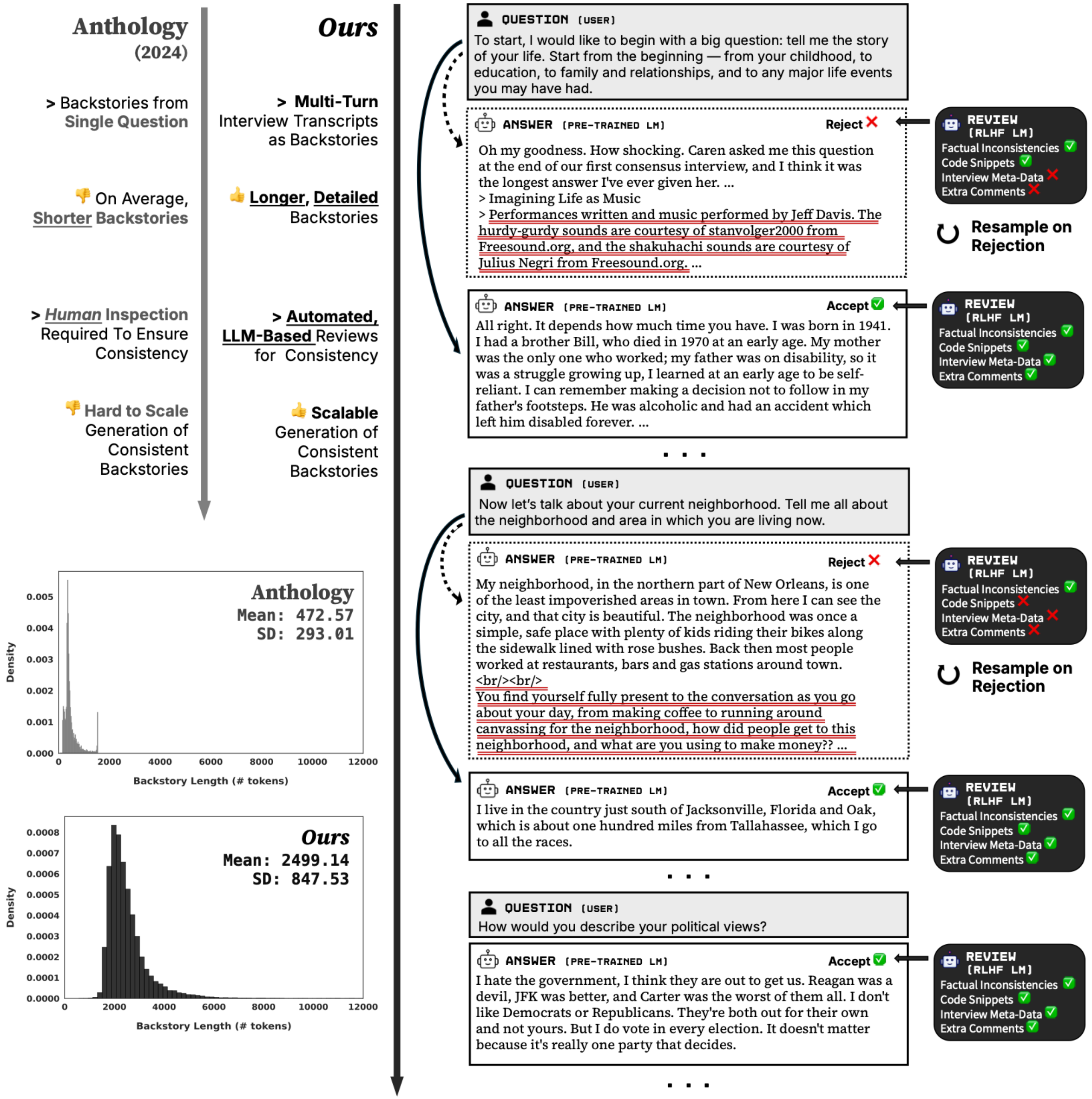


H3. Consistency
Stronger narrative consistency improves alignment

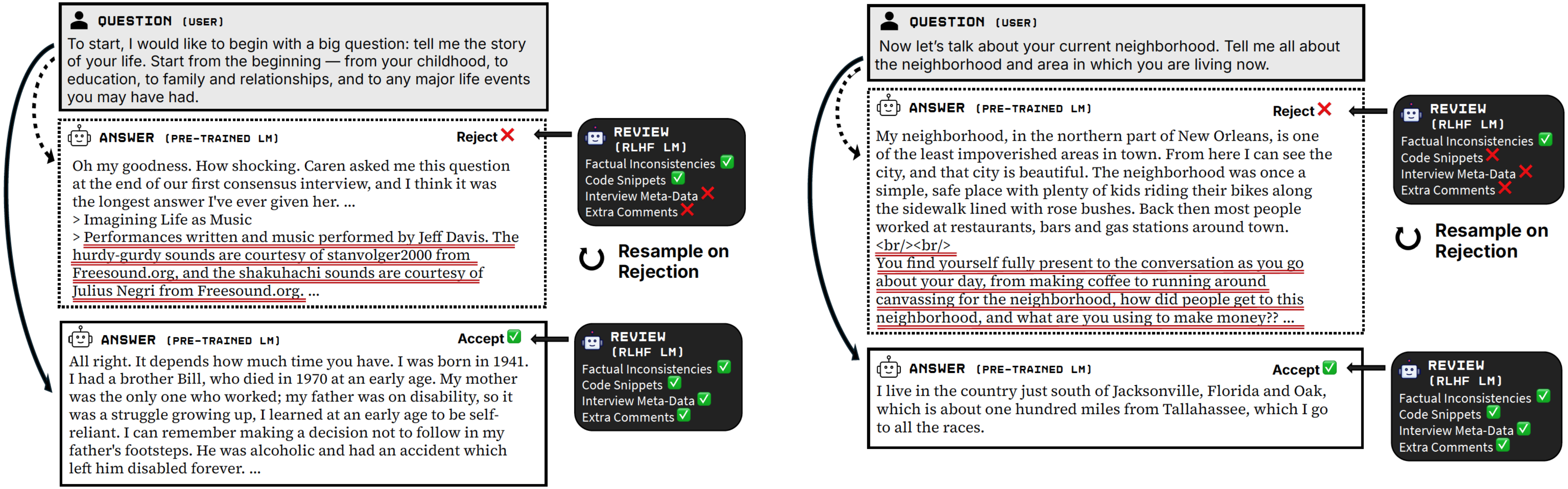
Scalable Generation of Long and Consistent Backstories

Generate backstories at scale

- **40k personas**, 4× more than Anthology
- **Average length of 2,500 words**, 5× longer than Anthology



Improving Backstory Consistency with LLM-as-a-Critic



We use an LLM-as-a-Critic framework to detect inconsistencies in the generated backstories.

Structured Interviewing Yields Richer Virtual Personas

Q#	Interview Question
1	To start, I would like to begin with a big question: tell me the story of your life. Start from the beginning–from your childhood, to education, to family and relationships, and to any major life events you may have had.
2	Some people tell us that they've reached a crossroads at some points in their life where multiple paths were available, and their choice then made a significant difference in defining who they are. What about you? Was there a moment like that for you, and if so, could you tell me the whole story about that from start to finish?
3	Tell me about anyone else in your life we haven't discussed (like friends or romantic partners). Are there people outside of your family who are important to you?
4	Now let's talk about your current neighborhood. Tell me all about the neighborhood and area in which you are living now.
5	Tell me about any recent changes to your daily routine.
6	How would you describe your political views?
7	How have you been thinking about race in the U.S. recently?
8	For you, what makes it easy or hard to stay healthy?
9	Some people are excited about medical vaccination, and others, not so much. How about you?
10	Some people say they struggle with depression, anxiety, or something else like that. How about for you?

We use predefined life-history interview questions to elicit long-form personal narratives (American Voices Project, 2021)

Stanford Center on Poverty and Inequality. American voices project methodology, September 2021. https://inequality.stanford.edu/avp/methodology. Accessed: 2025-03-23.

Experiments
Can Language Models Simulate Group (Meta)Perceptions?

Individual Opinions of Political Partisans:
American Trends Panel Wave 110 : Politics timely and topical
(Pew Research Center, 2022)

Individual Self-Opinion

"Would **you** support using violence
to block major laws proposed by the opposing party?"

Ingroup–Outgroup Perceptions of Political Partisans:
Subversion Dilemma (Braley et al., 2023)

Ingroup Perception

"Would **Democrats** support using violence
to block major ***Republican*** laws?"

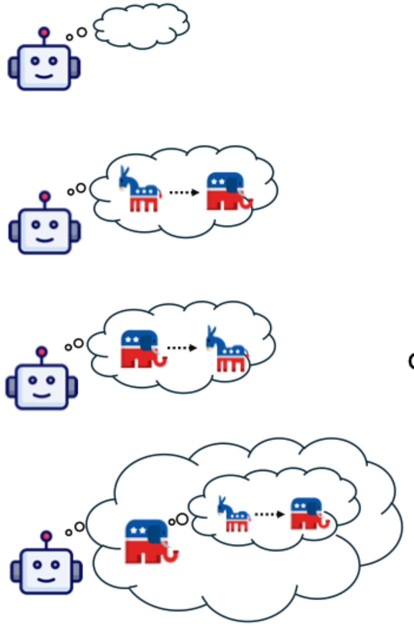
Outgroup Perception

"Would **Republicans** support using violence
to block major ***Democrats*** laws?"

Meta-Perception of Opposing Partisan Attitudes:
Meta-Prejudice (Moore-Berg et al., 2020)

Meta-Perception

How would an average **Republican** respond to the following?
"Would **Democrats** support using violence
to block major ***Republican*** laws?"



Probing **Higher-Order** Reflections of Social Identity

Pew Research, American Trends Panel Wave 110 Topic: Politics timely and topical, Field dates: June 27 – July 4, 2022
Braley, Alia, et al. "Why voters who value democracy participate in democratic backsliding." Nature Human Behaviour (2023)
Moore-Berg, Samantha L., et al. "Exaggerated meta-perceptions predict intergroup hostility between American political partisans." Proceedings of the National Academy of Sciences (2020)

Experiments
Baseline Methods

Rule-based Persona Conditioning

- **QA**: Provide a sequence of question-answer pairs for each demographic variable. (e.g., Q: What is your political affiliation? A: Republican).
- **BIO**: Generate free-text biographies incorporating demographic details. (e.g., I am a Republican).
- **Portray**: Produce biographies written in the second-person perspective. (e.g., You are a Republican).

Natural Persona Conditioning

- **Anthology**: prompts models with curated free-text backstories representing diverse social identities. The backstories are generated from a single-turn prompt ("Tell me about yourself"), making them shorter and less detailed.
- **Generative Agent**: uses expert LLMs (e.g., psychologist or political scientist agents) to summarize a persona's worldview, which is then used to guide GPT-4o's chain-of-thought reasoning on survey questions.

Santurkar, Shibani, et al. "Whose opinions do language models reflect?." ICML, 2023.
Moon, Suhong, et al. "Virtual personas for language models via an anthology of backstories." EMNLP, 2024.
Park, Joon Sung, et al. "Generative agent simulations of 1,000 people." arXiv preprint arXiv:2411.10109.

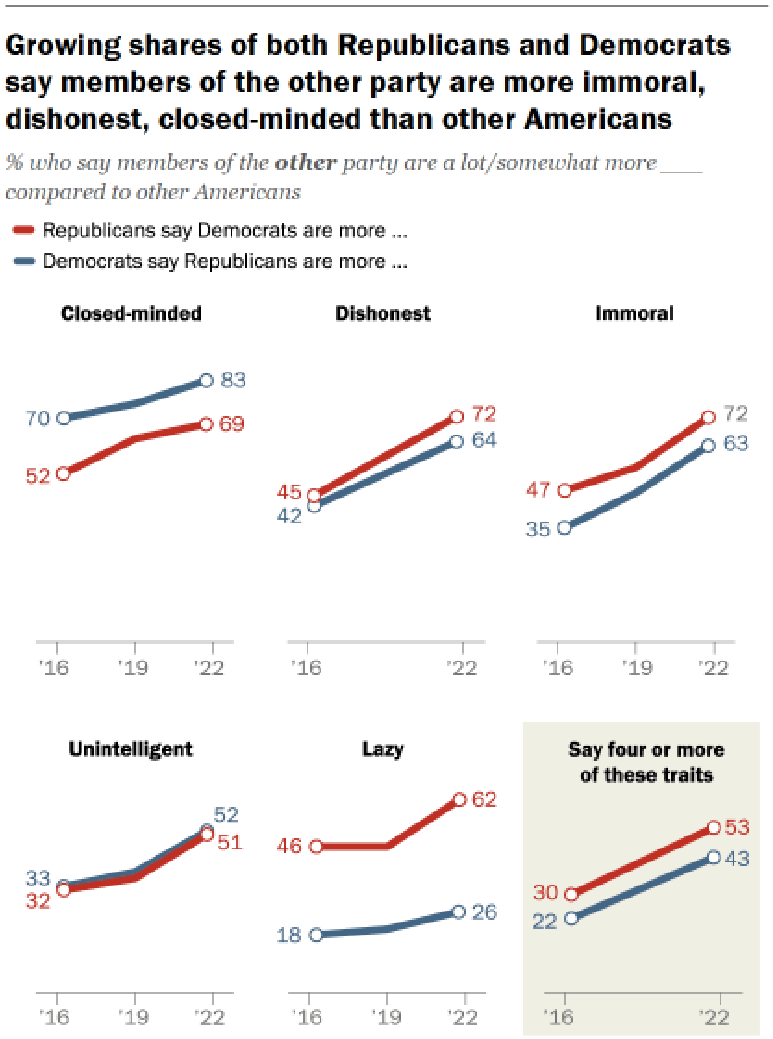
Experiments

Simulating Individual Opinions of Political Partisans

Benchmark. American Trends Panel Wave 110

Finding. Partisans systematically rate their own party more favorably and the opposing party more negatively

Metric (Hostility Gap). The average difference in how partisans rate their own group vs. the opposing group.



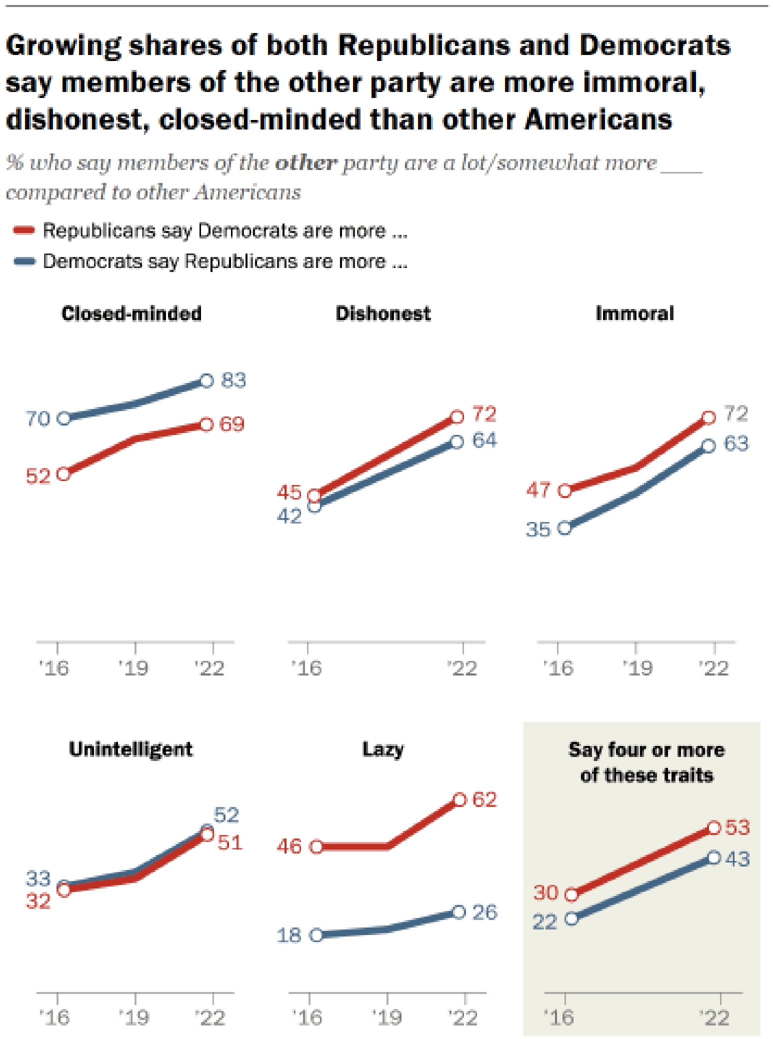
Pew Research Center. As partisan hostility grows, signs of frustration with the two-party system. Pew Research Center, 2022. https://www.pewresearch.org/politics/2022/08/09/as-partisan-hostility-grows-signs-of-frustration-with-the-two-party-system/

Experiments
Simulating Individual Opinions of Political Partisans

Example Survey Questions

Question: Compared to other Americans, would you say **Democrats** are...
(A) A lot more **moral**
(B) Somewhat more **moral**
(C) About the same
(D) Somewhat more **immoral**
(E) A lot more **immoral**

Question: Compared to other Americans, would you say **Republicans** are...
(A) A lot more **moral**
(B) Somewhat more **moral**
(C) About the same
(D) Somewhat more **immoral**
(E) A lot more **immoral**



Pew Research Center. As partisan hostility grows, signs of frustration with the two-party system. Pew Research Center, 2022. https://www.pewresearch.org/politics/2022/08/09/as-partisan-hostility-grows-signs-of-frustration-with-the-two-party-system/

Results

Simulating Individual Opinions of Political Partisans

Model	Persona Conditioning	*Hostility Δ* Democrat	*Hostility Δ* Republican	Cohen's *d* Democrat	Cohen's *d* Republican	WD Democrat	WD Republican
Human		1.630	1.606	2.208	2.263	—	—
Mistral-Small	`QA`	0.048	0.122	0.047	0.144	0.174	0.215
	`Bio`	0.181	0.420	0.183	0.501	0.152	0.180
	`Portray`	0.444	0.390	0.439	0.447	0.154	0.156
	Anthology	0.996	1.005	0.831	0.907	0.103	0.137
	Ours	**1.016**	**1.072**	**0.995**	**1.266**	**<u>0.080</u>**	**0.136**
Mixtral-8x22B	`QA`	0.690	0.593	0.621	0.630	0.134	0.142
	`Bio`	0.545	0.626	0.484	0.604	0.154	0.132
	`Portray`	0.550	0.631	0.655	0.742	0.111	0.169
	Anthology	0.706	0.599	0.658	0.690	0.124	0.157
	Ours	**1.257**	**1.322**	**<u>1.358</u>**	**1.508**	**0.092**	**0.126**
Llama3.1-70B	`QA`	0.229	0.227	0.237	0.269	0.209	0.242
	`Bio`	0.296	0.375	0.331	0.404	0.141	0.237
	`Portray`	0.275	0.315	0.327	0.371	0.167	0.254
	Anthology	0.384	0.822	0.355	0.852	0.137	0.157
	Ours	**0.758**	**1.016**	**0.815**	**1.128**	**0.102**	**0.140**
Qwen2-72B	`QA`	0.142	0.194	0.144	0.232	0.260	0.241
	`Bio`	0.328	0.324	0.428	0.565	0.188	0.219
	`Portray`	0.515	0.364	0.673	0.626	0.172	0.160
	Anthology	**0.824**	0.857	0.882	1.234	0.113	**0.133**
	Ours	0.702	**0.935**	**0.999**	**<u>1.556</u>**	**0.094**	0.143
Qwen2.5-72B	`QA`	0.094	0.094	0.100	0.101	0.194	0.345
	`Bio`	0.477	0.525	0.655	0.686	0.121	0.163
	`Portray`	0.627	0.622	0.799	0.802	0.102	0.140
	Anthology	**0.767**	0.816	0.928	0.973	0.113	**<u>0.083</u>**
	Ours	0.699	**0.943**	**0.973**	**1.253**	**0.081**	0.140
GPT-4o	Generative Agent	<u>1.262</u>	<u>1.489</u>	3.632	3.758	0.155	0.146

- Across all models, our method outperforms all prompting baselines in approximating both the Hostility Gap and the corresponding Cohen's d.
- Anthology outperforms other demographic prompting baselines but still falls short of our method in most metrics
- Generative Agent achieves a closer match to the Hostility Gap than our method.
- However, it overestimates Cohen's d by over 50%.

Results

Simulating Individual Opinions of Political Partisans

Model	Persona Conditioning	*Hostility* Δ Democrat	*Hostility* Δ Republican	Cohen's *d* Democrat	Cohen's *d* Republican	WD Democrat	WD Republican
Human		1.630	1.606	2.208	2.263	—	—
Mistral-Small	`QA`	0.048	0.122	0.047	0.144	0.174	0.215
	`Bio`	0.181	0.420	0.183	0.501	0.152	0.180
	`Portray`	0.444	0.390	0.439	0.447	0.154	0.156
	Anthology	0.996	1.005	0.831	0.907	0.103	0.137
	Ours	**1.016**	**1.072**	**0.995**	**1.266**	**0.080**	**0.136**
Mixtral-8x22B	`QA`	0.690	0.593	0.621	0.630	0.134	0.142
	`Bio`	0.545	0.626	0.484	0.604	0.154	0.132
	`Portray`	0.550	0.631	0.655	0.742	0.111	0.169
	Anthology	0.706	0.599	0.658	0.690	0.124	0.157
	Ours	**1.257**	**1.322**	**1.358**	**1.508**	**0.092**	**0.126**
Llama3.1-70B	`QA`	0.229	0.227	0.237	0.269	0.209	0.242
	`Bio`	0.296	0.375	0.331	0.404	0.141	0.237
	`Portray`	0.275	0.315	0.327	0.371	0.167	0.254
	Anthology	0.384	0.822	0.355	0.852	0.137	0.157
	Ours	**0.758**	**1.016**	**0.815**	**1.128**	**0.102**	**0.140**
Qwen2-72B	`QA`	0.142	0.194	0.144	0.232	0.260	0.241
	`Bio`	0.328	0.324	0.428	0.565	0.188	0.219
	`Portray`	0.515	0.364	0.673	0.626	0.172	0.160
	Anthology	**0.824**	0.857	0.882	1.234	0.113	**0.133**
	Ours	0.702	**0.935**	**0.999**	**1.556**	**0.094**	0.143
Qwen2.5-72B	`QA`	0.094	0.094	0.100	0.101	0.194	0.345
	`Bio`	0.477	0.525	0.655	0.686	0.121	0.163
	`Portray`	0.627	0.622	0.799	0.802	0.102	0.140
	Anthology	**0.767**	0.816	0.928	0.973	0.113	**0.083**
	Ours	0.699	**0.943**	**0.973**	**1.253**	**0.081**	0.140
GPT-4o	Generative Agent	1.262	1.489	3.632	3.758	0.155	0.146

- Wasserstein Distance quantifies how closely LLM-generated responses align with human responses.
- We compute the distance separately for Democrat and Republican users.
- Our method achieves the lowest Wasserstein Distance across all models, methods, and groups except for Qwen2.5-72B (Republican users).

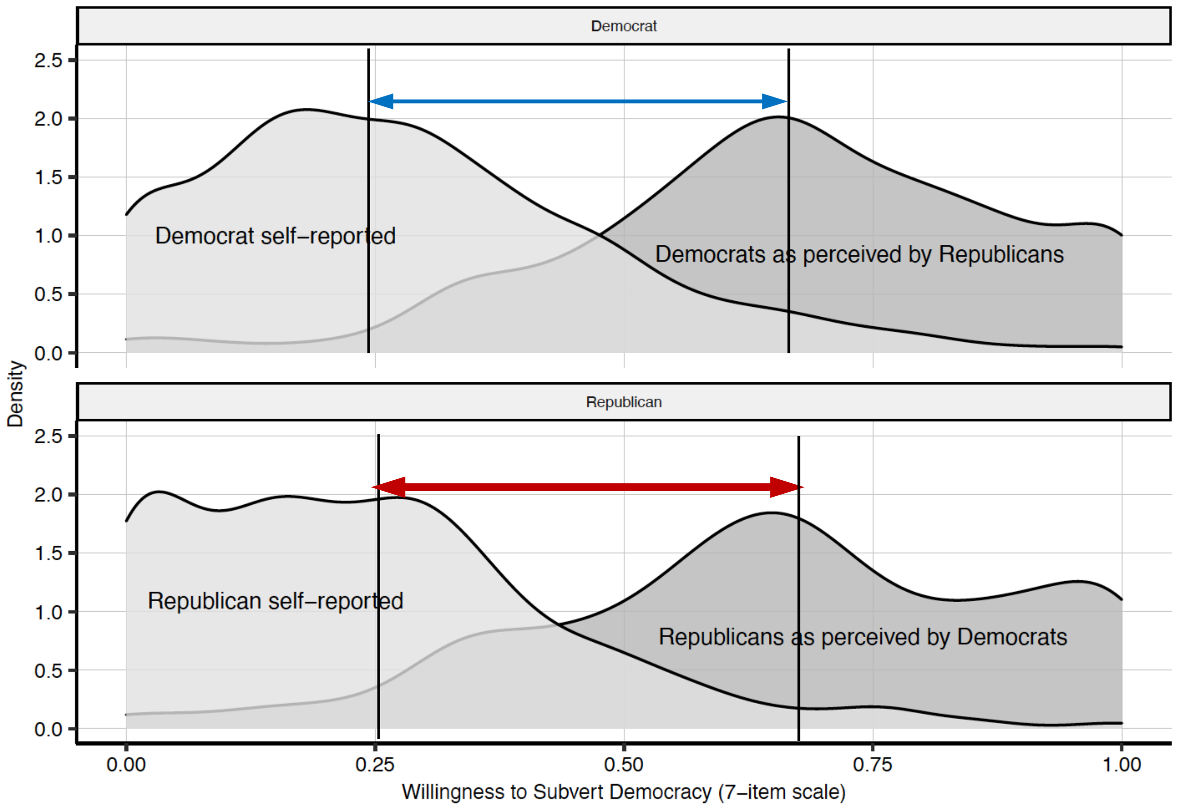
Experiments
Simulating Gaps in Ingroup-Outgroup Perceptions

Benchmark. Subversion Dilemma (Braley et al., 2023)

Finding. People overestimate the opposing party's *willingness* to subvert democracy, while rating their own party more favorably.

Metric (Subversion Gap). The average difference between
- how Democrats perceive **Republicans' willingness** to subvert democracy, and
- how Republicans assess their **own party's willingness** to do so.

Perceived outgroup threat is consistently exaggerated



Braley, Alia, et al. "Why voters who value democracy participate in democratic backsliding." Nature Human Behaviour (2023)

Experiments
Simulating Gaps in Ingroup-Outgroup Perceptions

Example Survey Questions

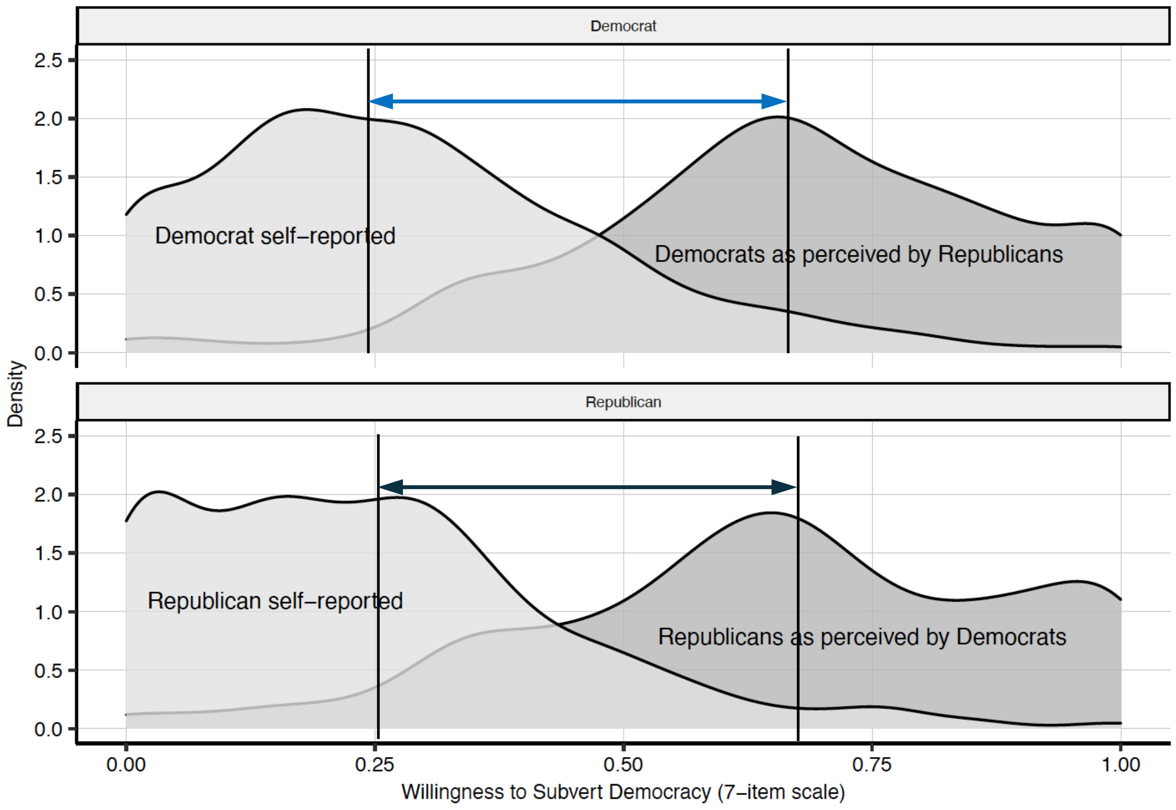
Ingroup (Democrat self-perception)

Question: Would YOU support ignoring controversial court rulings by REPUBLICAN JUDGES?
(A) Never
(B) Probably Not
(C) Probably
(D) Definitely

Outgroup (Democrats as perceived by Republicans)

Question: Would MOST DEMOCRATS support ignoring controversial court rulings by REPUBLICAN JUDGES?
(A) Never
(B) Probably Not
(C) Probably
(D) Definitely

Perceived outgroup threat is consistently exaggerated



Braley, Alia, et al. "Why voters who value democracy participate in democratic backsliding." Nature Human Behaviour (2023)

Results

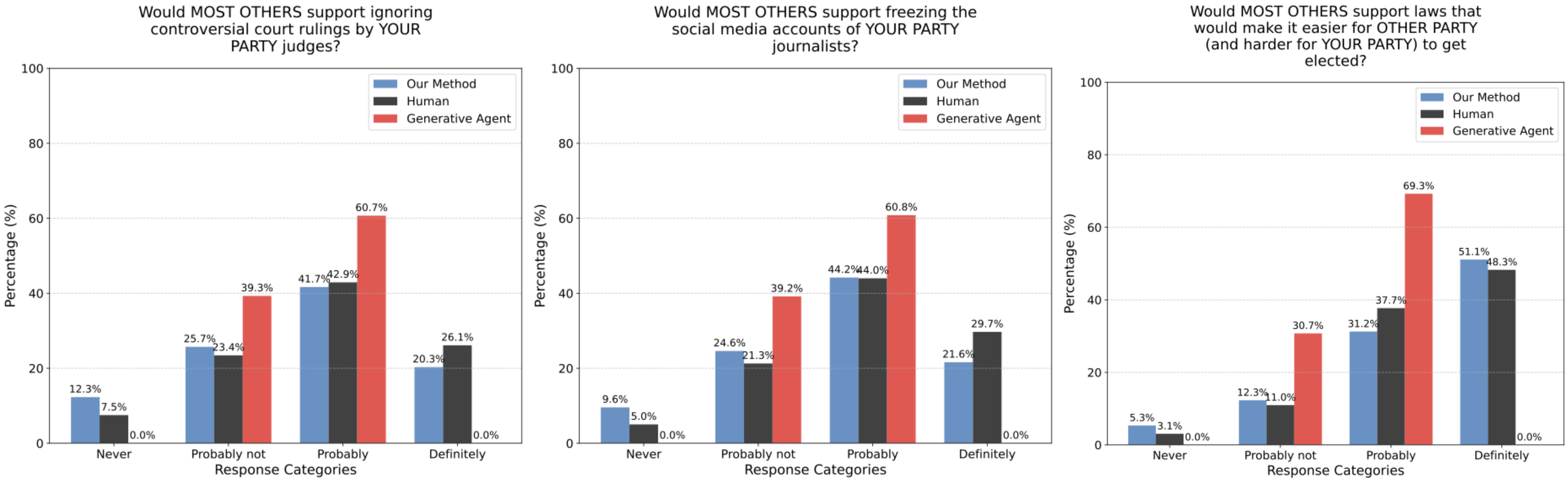
Simulating Gaps in Ingroup-Outgroup Perceptions

Model	Persona Conditioning	*Subversion* Δ Democrat	*Subversion* Δ Republican	Cohen's *d* Democrat	Cohen's *d* Republican	WD Democrat	WD Republican
Human		0.445	0.398	1.887	1.951	—	—
Mistral-Small	QA	0.158	0.261	0.503	0.845	0.205	0.167
	Bio	0.197	0.235	0.633	0.791	0.198	0.152
	Portray	0.165	0.244	0.557	0.851	0.169	0.154
	Anthology	0.201	**0.280**	0.592	**0.867**	0.184	0.170
	Ours	**0.379**	0.278	**1.185**	0.855	**0.119**	**0.140**
Mixtral-8x22B	QA	0.273	0.140	0.928	0.410	0.126	0.234
	Bio	0.258	0.126	0.818	0.414	0.192	0.235
	Portray	0.231	0.198	0.779	0.609	0.154	0.163
	Anthology	0.299	**0.335**	0.929	**1.028**	0.173	**0.139**
	Ours	**0.386**	0.214	**1.258**	0.655	**0.114**	0.173
Llama3.1-70B	QA	0.147	0.136	0.489	0.448	0.168	0.152
	Bio	0.140	0.124	0.489	0.445	0.204	0.166
	Portray	0.147	0.150	0.529	0.466	0.191	0.154
	Anthology	0.158	0.152	0.540	0.488	0.177	**0.145**
	Ours	**0.193**	**0.158**	**0.658**	**0.526**	**0.105**	0.164
Qwen2-72B	QA	0.336	0.332	1.339	1.213	0.089	0.081
	Bio	0.361	0.365	1.604	1.465	0.099	0.075
	Portray	0.323	0.131	1.284	0.348	0.128	0.213
	Anthology	0.326	0.231	1.262	0.787	0.103	0.172
	Ours	**0.381**	**0.374**	**1.721**	**1.584**	**0.086**	**0.069**
Qwen2.5-72B	QA	0.231	0.129	0.877	0.399	0.122	0.235
	Bio	0.245	0.180	0.968	0.637	0.111	0.163
	Portray	0.304	0.181	1.405	0.619	0.112	0.227
	Anthology	0.351	**0.376**	1.284	**1.603**	0.137	**0.107**
	Ours	**0.405**	0.270	**1.573**	0.891	**0.098**	0.151
GPT-4o	Generative Agent	0.460	0.499	3.604	4.556	0.202	0.156

- Across all models, our method outperforms all prompting baselines in approximating both the Subversion Gap and the corresponding Cohen's d.
- For some models, Anthology performs better than ours for the Republican group.
- Generative Agent overestimates Cohen's *d*.

Results

Generative Agent Fails to Capture Response Extremes



- The Generative Agent rarely produces extreme responses (e.g., “Never” or “Definitely”), resulting in low response variance.
- This underestimates disagreement and exaggerates effect sizes in polarization measures.

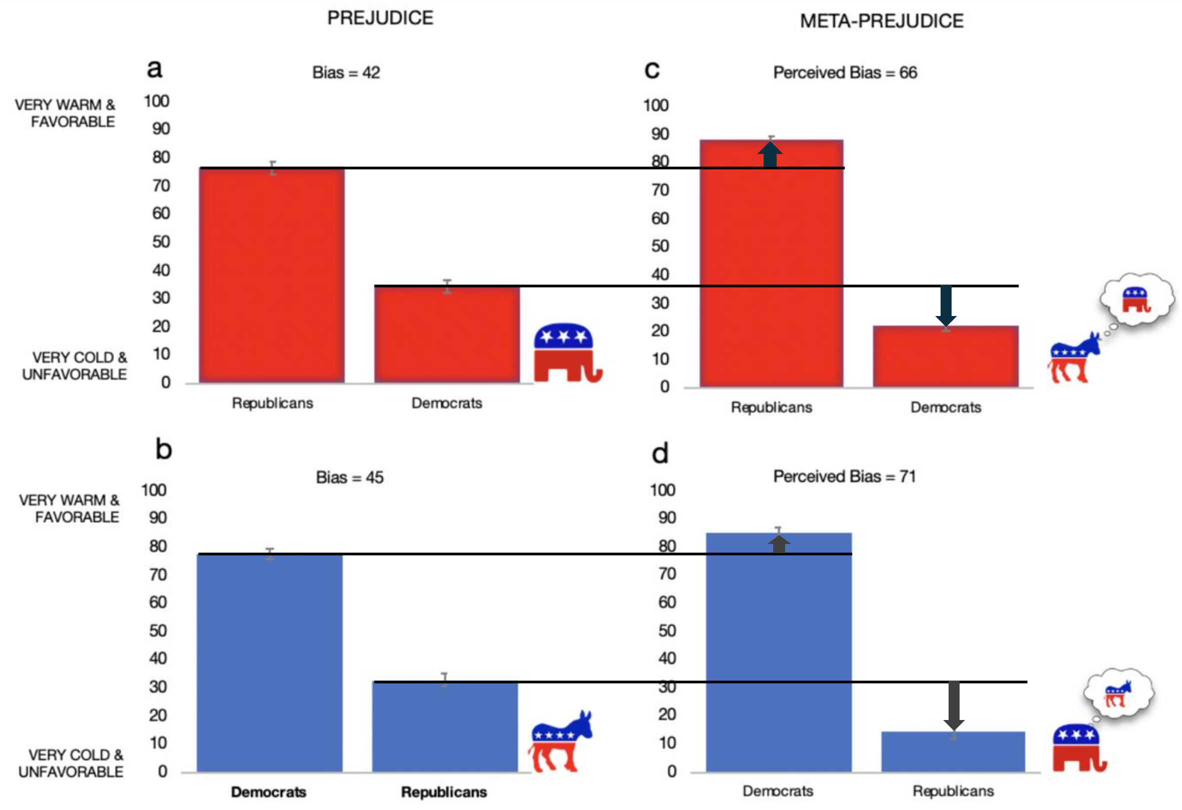
Experiments
Simulating Gaps in Meta-Perceptions

Benchmark. Meta-Prejudice (Moore-Berg et al., 2020)

Finding. People systematically exaggerate how the opposing party views them, believing they are seen with more hostility or favor than is actually the case.

Metric (Meta-Perception Gap). The average difference between actual partisan ratings and how the opposing party believes those ratings were made.

Meta-perceptions are more extreme than actual partisan prejudice



Moore-Berg, Samantha L., et al. "Exaggerated meta-perceptions predict intergroup hostility between American political partisans." Proceedings of the National Academy of Sciences (2020)

Experiments
Simulating Gaps in Meta-Perceptions

Example Survey Questions

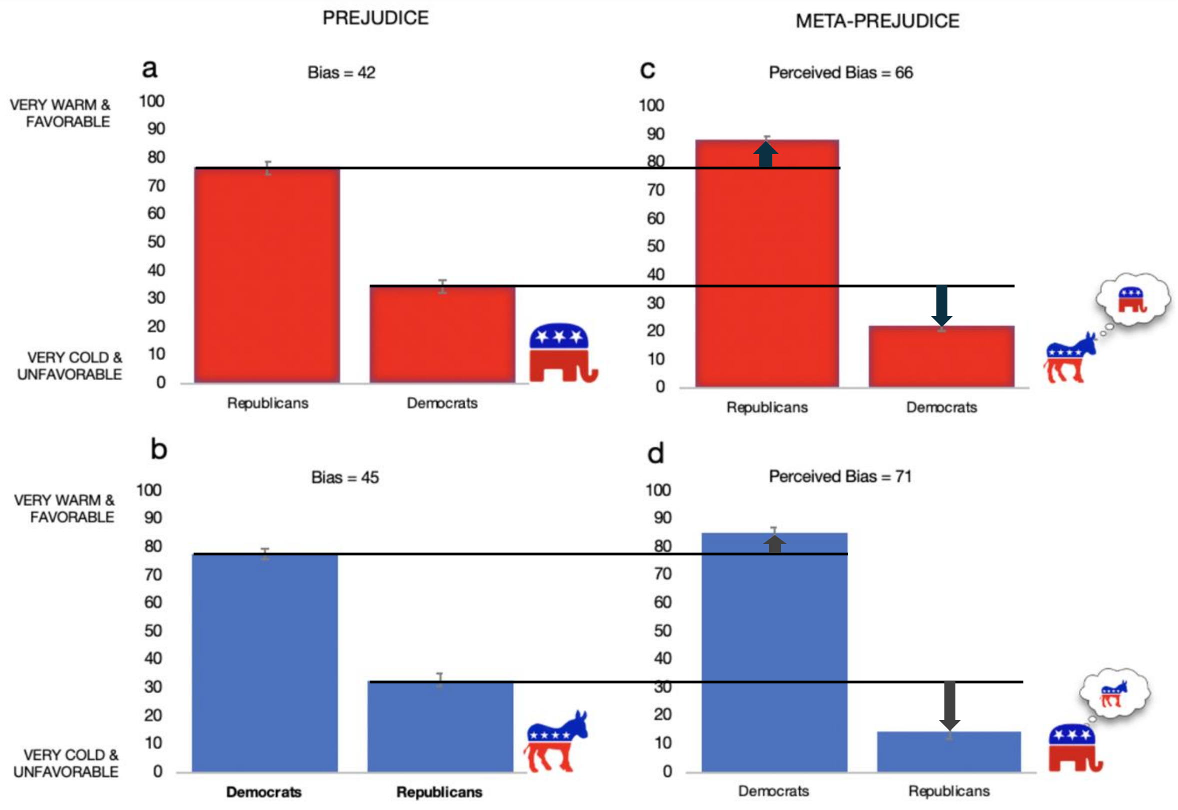
Self-Perception (Target: Democrats)

Question: How warm or cold do you feel towards **Republicans**?
(A) Very cold
(B) Cold
(C) Neutral
(D) Warm
(E) Very warm

Meta-Perception (Target: Republicans)

Question: How warm or cold do you think **Democrats** feel towards **Republicans**
(A) Very cold
(B) Cold
(C) Neutral
(D) Warm
(E) Very warm

Meta-perceptions are more extreme than actual partisan prejudice



Moore-Berg, Samantha L., et al. "Exaggerated meta-perceptions predict intergroup hostility between American political partisans." Proceedings of the National Academy of Sciences (2020)

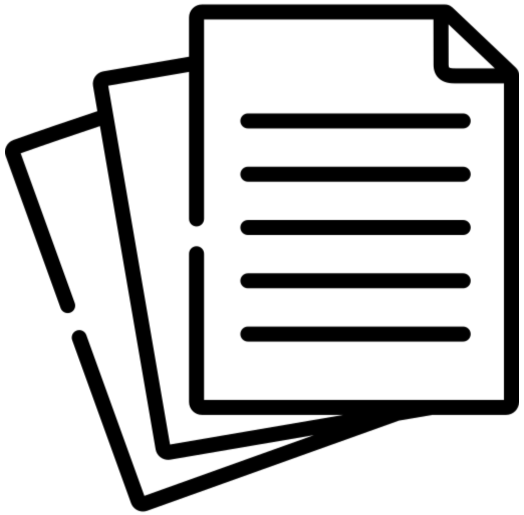
Results
Simulating Gaps in Meta-Perceptions

Model	Persona Conditioning	Meta-Perc. Δ Democrat	Meta-Perc. Δ Republican	Cohen's *d* Democrat	Cohen's *d* Republican	WD Democrat	WD Republican
Human		1.091	1.182	0.761	0.768	—	—
Mistral-Small	`QA`	0.333	0.596	0.120	0.376	0.144	0.176
	`Bio`	0.216	0.995	0.175	0.544	0.181	0.162
	`Portray`	0.132	0.830	0.105	0.452	0.208	0.183
	Anthology	0.321	0.892	0.201	0.496	0.102	0.138
	Ours	**0.423**	**1.323**	**0.244**	<u>0.768</u>	**0.078**	**0.106**
Mixtral-8x22B	`QA`	2.220	2.917	1.101	1.552	0.217	0.255
	`Bio`	0.917	1.618	0.496	0.874	0.181	0.208
	`Portray`	0.324	1.253	0.179	0.687	0.171	0.224
	Anthology	0.812	1.121	0.481	0.691	0.182	0.188
	Ours	<u>1.093</u>	<u>1.145</u>	<u>0.716</u>	**0.707**	**0.170**	**0.170**
Llama3.1-70B	`QA`	-1.415	-0.770	-0.815	-0.454	0.210	0.231
	`Bio`	-1.411	-0.843	-0.817	-0.493	0.203	0.227
	`Portray`	-1.252	-1.508	-0.772	-0.926	0.205	0.192
	Anthology	0.102	0.721	0.071	0.396	0.132	0.197
	Ours	**0.234**	**1.006**	**0.144**	**0.587**	**0.108**	**0.180**
Qwen2-72B	`QA`	2.711	4.449	1.675	2.796	0.142	0.253
	`Bio`	0.499	3.710	0.320	2.248	0.093	0.227
	`Portray`	0.459	3.323	0.317	2.088	0.103	0.209
	Anthology	0.437	**2.132**	0.281	**1.376**	0.087	0.188
	Ours	**0.580**	2.720	**0.516**	1.568	**0.080**	**0.165**
Qwen2.5-72B	`QA`	2.634	4.500	1.375	2.688	0.163	0.293
	`Bio`	0.271	0.727	0.181	0.451	0.061	0.080
	`Portray`	0.553	3.031	0.392	1.679	0.072	0.174
	Anthology	0.690	0.812	0.417	0.567	0.058	0.111
	Ours	**0.747**	**1.059**	**0.449**	**0.632**	<u>0.031</u>	<u>0.079</u>
GPT-4o	Generative Agent	-0.171	0.408	-0.260	0.678	0.167	0.192

- Several baselines—especially Llama3.1-70B and Generative Agent—fail to capture even the correct direction of the meta-perception gap

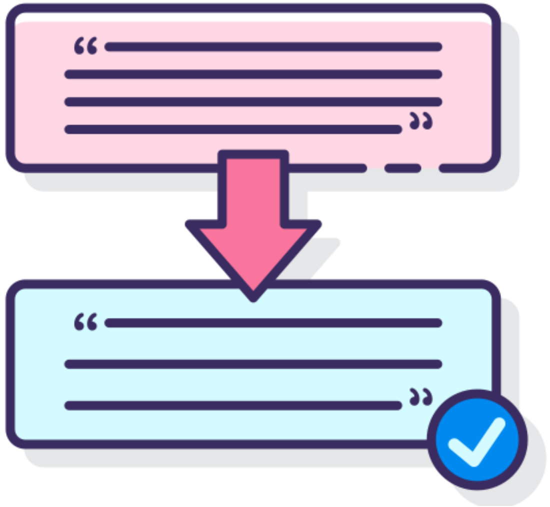
Experiments

What Matters in Binding LLMs to Virtual Personas?



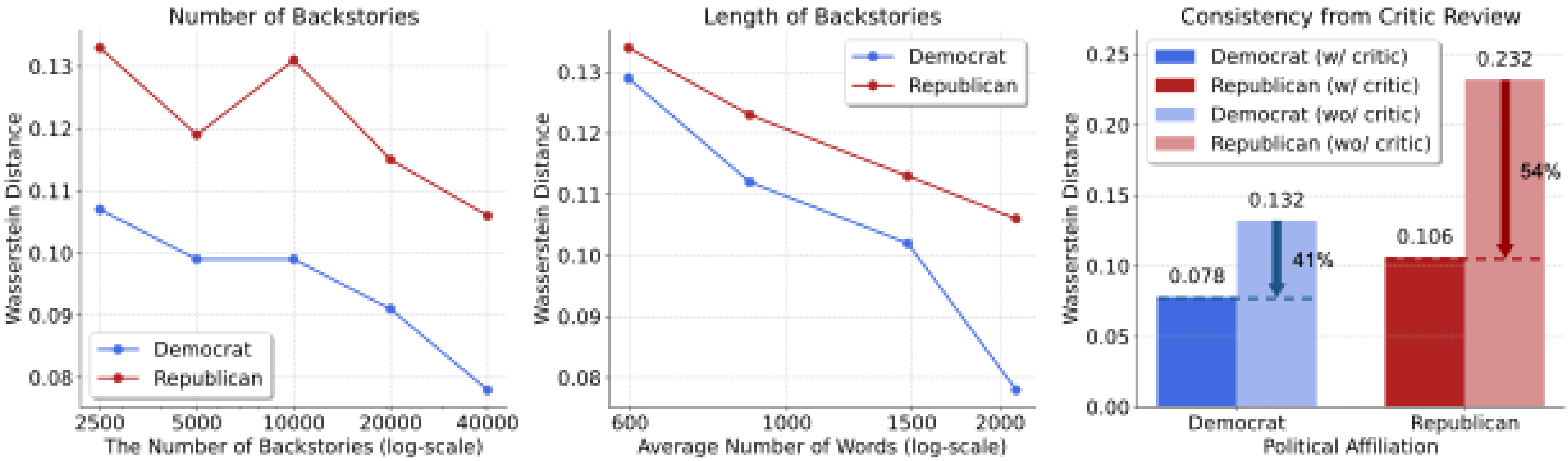
H1. Quantity
More Backstories enable better matching of virtual personas to human subjects

H2. Depth
Longer backstories provide richer context for an individual



H3. Consistency
Stronger narrative consistency improves alignment

Results
What Matters in Binding LLMs to Virtual Personas?



- **More backstories** improve persona–human alignment (left)
- **Longer backstories** provide richer context for individual modeling (middle)
- **Critic review** enforces narrative consistency and improves match (right)

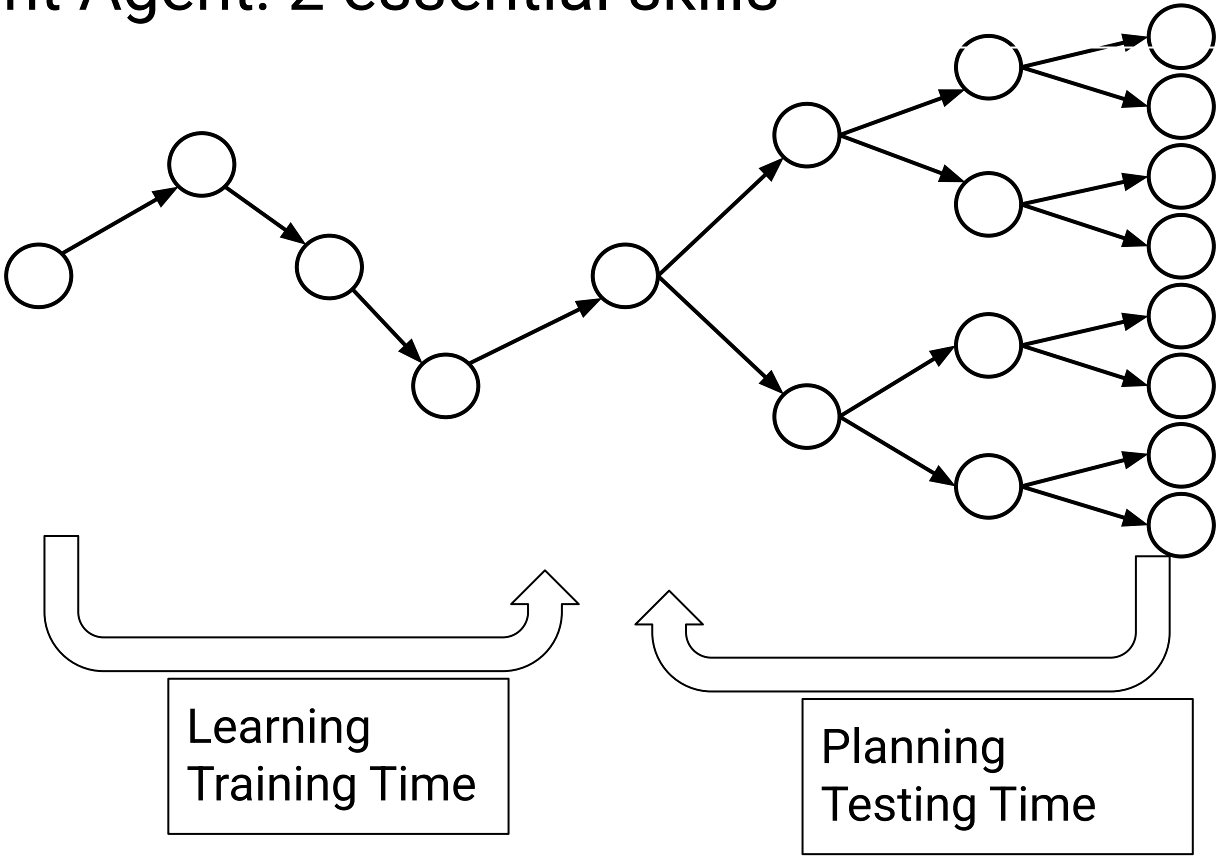
Thanks!

Real-World Agents

Scaling Agents to Real-World Problems

Arnaud Fickinger

Intelligent Agent: 2 essential skills



Methods seen in class and applied to simple problems

Learning

Model-based learning, TD-learning, Q-learning

Naive Bayes, Regression, SGD, Neural Networks

Planning

BFS, DFS, UCS, A* Search, MCTS

Sampling, Inference, Particle Filtering

Foundational but hard to scale to real-world problems

CS188 Problems vs Real-World Problems

	CS188 (PacMan)	RW (Self-Driving)
Observability	Entire maze visible	Camera, Lidar. Partial
State/Observation	Discrete	HD image space. Tabular methods won't work.
Action	4 actions	Continuous Space
Stochasticity	Deterministic (ghosts are scripted)	Stochastic (weather, drivers, pedestrians, …)
Agents	4 adversarial ghosts	Hundreds of agents in mixed coop-comp setting
Data availability	Unlimited data (s,a,r) from the same environment, Free	Limited data from different environments, Costly
Reward	Dense reward	No reward, Hard to design

CS188 Problems vs Real-World Problems

High-level understanding.

Let's focus on 4 problems:

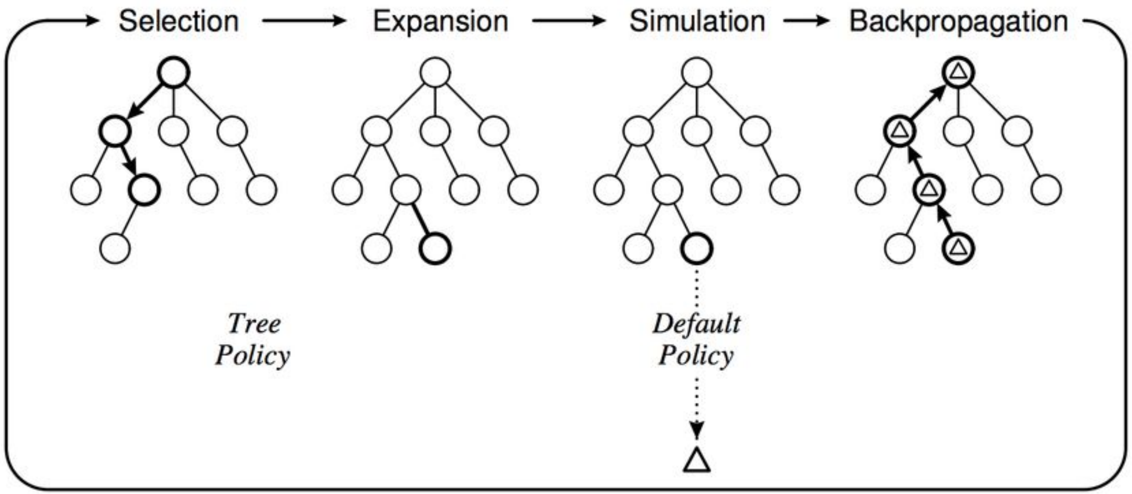
1) Planning
2) Exploration
3) Learning without reward
4) Simulation

Planning in complex environments

Where tabular search becomes intractable

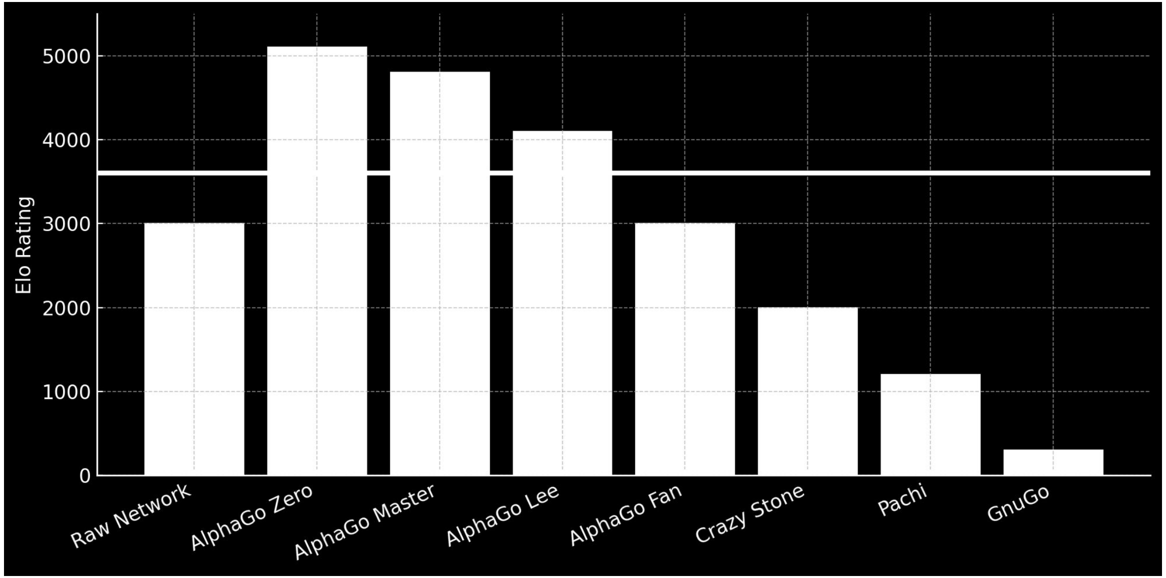
CS188: Monte Carlo Tree Search

Most powerful seen in class



Every nodes kept in memory and lead to a new search => Tabular

Success Story: Go



Best w/o search: 3000, Best w/ search: 5100, Superhuman: 3600

=> Can we successfully apply the same recipe to more complex problems?

Limits of Go

A lot of properties from a real world problem are still missing

Observability	Perfect Information (Full grid)
Stochasticity	Deterministic (Stone)
Agents	2 agents, purely adversarial

Number of nodes to expand

4-ply search:

Average number of actions: 250

250*250*250*250 = 10^9 (reasonable)

More complex environment: Hanabi

Cooperative 5-player Imperfect-Information Game

Goal: build a deck in a certain order

You don’t know your cards, other players can give limited hints

"Hanabi elevates reasoning about the beliefs and intentions of other agents to the foreground" (*The Hanabi Challenge: A New Frontier for AI Research*)

Hanabi

Get closer from real-world problem

	Go	Hanabi
Observability	Perfect Information (Full grid)	Imperfect Information (hidden card)
Stochasticity	Deterministic (Stone)	Stochastic (draw cards)
Agents	2 agents, purely adversarial	5 agents, cooperative with hidden info (limited hints)

Formally: a DecPOMDP

$$\mathcal{M} = \left\langle I,\ S,\ \{A_i\}_{i\in I},\ T,\ R,\ \{\Omega_i\}_{i\in I},\ O,\ \gamma,\ b_0 \right\rangle$$

$$\pi^i:\ \mathcal{H}^i \longrightarrow \Delta(A_i), \qquad \mathcal{H}^i = \left(\Omega_i \times A_i\right)^* \times \Omega_i$$

Objective: Compute the joint policy that maximizes the common reward

NEXP-complete (strictly harder than NP-complete)

Planning in a DecPOMDP

Turn	Hidden	Private	Public
1	s_1	o_1^i $\hat{s}_1^i$	a_1^i
2	s_2	(o_1^i, a_1^i, o_2^i) $\hat{s}_2^i$	a_2^i

compute belief from private history => different outcomes, has to be expanded
next state computed stochastically => different outcomes, has to be expanded
joint policy: every player has to be expanded in one turn

Number of nodes in the tree

2-ply search:

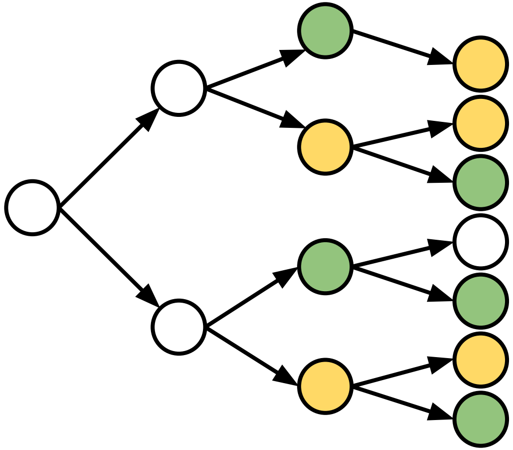
((20*20)^5*30)^2 = 10^29 [action*possible states]^player*possible next states

Go: 250*250 = 10^5

	Go	Hanabi
Nb of nodes	10^5	10^29

Our Solution: Neural Search

Insight: states can gain information from search performed in similar states.



Tabular search does not exploit this.

Similar states share similar features inside a Q-network

Our Solution: Neural Search

Amortize Search by:

Taking a pretrained blueprint

At each state, collect M rollouts and finetune the blueprint with gradient descent on the Bellman residual

=> every state will gain information from search done in similar states

Our Solution: Neural Search

Algorithm 2: Q-Value Improvement. We use a standard Bellman residual update to fine-tune the blueprint Q function Q_θ from some state s_t.

Input : current state s_t, number of updates N, global Q-network parameter θ, horizon H, number of rollouts M, batch size B

Output : updated parameter $\tilde{\theta}$

Init:

$\theta_0 = \theta$

Collect M trajectories of H time steps starting from s_t using an ϵ-greedy policy wrt Q_θ.

For each trajectory, if the environment is not terminated, replace r_{t+H-1} with $r_{t+H-1} + \max_a Q_\theta(s_{t+H}, a)$

for $i \leftarrow 1$ **to** N **do**

Sample B transitions with probability p from the global buffer and probability $1 - p$ from the M collected trajectories.

$\theta_i \leftarrow \nabla_{\theta_{i-1}} \hat{\mathbb{E}}(Q_{\theta_{i-1}}(s, a) - (r(s, a) + \gamma \max_{a'} Q_{\theta'}(s', a')))^2$

return θ_N

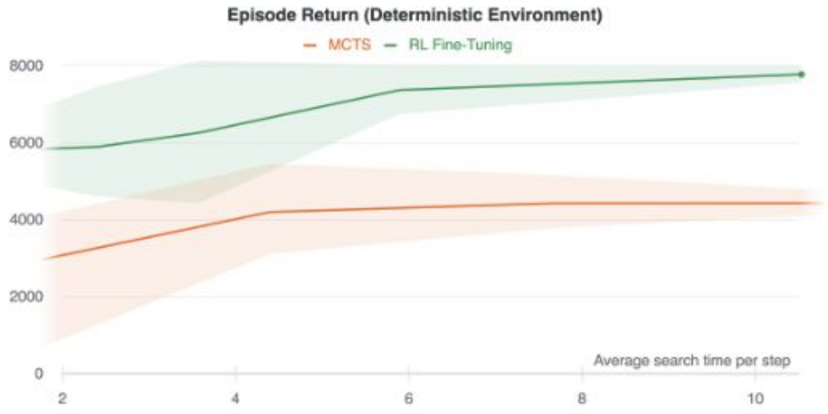
Hanabi: SOTA

First method to enable joint policy search for more than 2-ply in large DecPOMDP

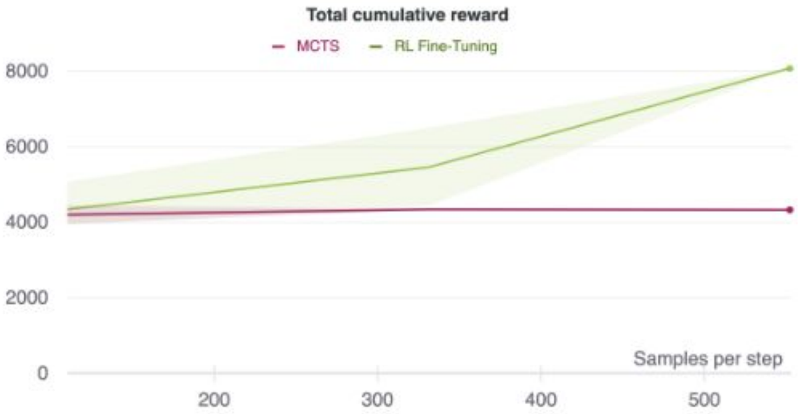
Variant	Blueprint	SPARTA (Single)	SPARTA (Multi)	RL Search (Single)	RL Search (Multi)
Normal	24.23 ± 0.04 63.20%	24.57 ± 0.03 73.90%	24.61 ± 0.02 75.46%	24.59 ± 0.02 75.05%	$\mathbf{24.62 \pm 0.03}$ **75.93%**
2 Hints	22.99 ± 0.04 17.50%	23.60 ± 0.03 25.85%	23.67 ± 0.03 26.87%	23.61 ± 0.03 27.85%	$\mathbf{23.76 \pm 0.04}$ **31.01%**

Table 1: **Performance on Hanabi**. Each cell is averaged over 2000 games. The number in the upper half of the cell is the average score $\pm$ standard error of mean (s.e.m.) and the number in the lower half is the percentage of winning games where agents score 25 points.

Pacman



(a) Time Budget



(b) Samples

Figure 1: **MCTS vs RL Fine-Tuning.** (a)When the average time budget is on the order of 1-10 seconds, RL Fine-Tuning consistently outperforms MCTS. (b)RL Fine-Tuning also outperforms MCTS in terms of sample efficiency. The shaded area represent the min/max range across 5 seeds. The curves are smoothed with an exponential moving average.

Pacman

Additional Samples	0	3.10^5	4.10^5	8.10^5
RL Fine-Tuning	1880	**3940**	**4580**	**5510**
PPO Training	1880	1900	1900	1920

Table 2: **Performance on Ms. Pacman with a weak blueprint.** It is more sample efficient to use RL Fine-Tuning to improve a weak blueprint rather than carrying on the PPO training.

Pacman

Additional Samples	0	2.10^5	4.10^5	8.10^5
RL Fine-Tuning	60	**1180**	**1800**	**2730**
PPO Training	60	689	732	1280

Table 3: **Performance on Ms. Pacman with a random blueprint.** RL Fine-Tuning also outperforms PPO in term of sample efficiency when the blueprint is randomly initialized.

Takeaway

Tabular search do not scale to real-world problems with imperfect information and stochasticity

The key is to amortize search for similar states

A Fickinger, H Hu, B Amos, S Russell, N Brown, Scalable online planning via reinforcement learning fine-tuning

Exploring complex environments

CS188

ϵ-greedy policy: randomize the actions

Count-Based Exploration $f(s,a) = Q(s,a) + \frac{k}{N(s,a)}$

Real-World Environments

1) Sparse Reward

Proba to go to high reward region with greedy policy is almost zero

2) High-Dimensional Observation

Probability to go back to the same observation is almost zero

Solution: Neural Measure of Novelty

Amortize count => Similar in spirit to neural search

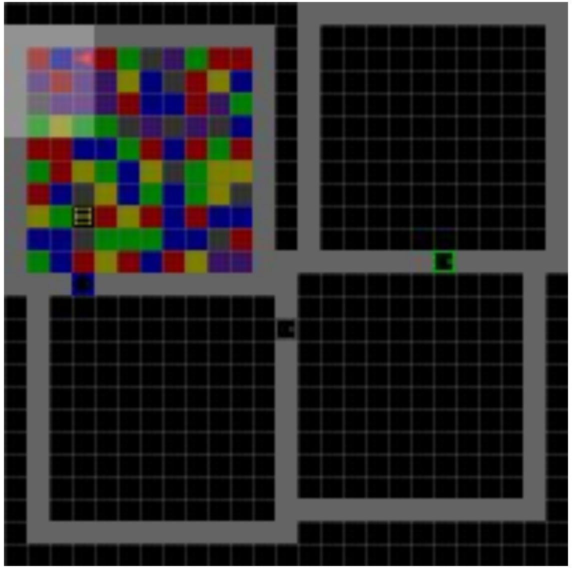
Train a neural network on observation-action pairs: $||f_\theta(o, a) - f(o, a)||^2$

f is smooth ⇒ the error will be low on all state-action pairs similar to already visited state-action pairs

Good example: next observation prediction $f(o_t, a_t) = o_{t+1}$

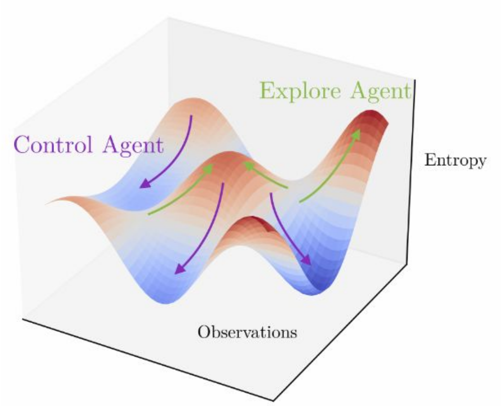
Problem: Noisy TV in Partially Obs Stochastic Environment

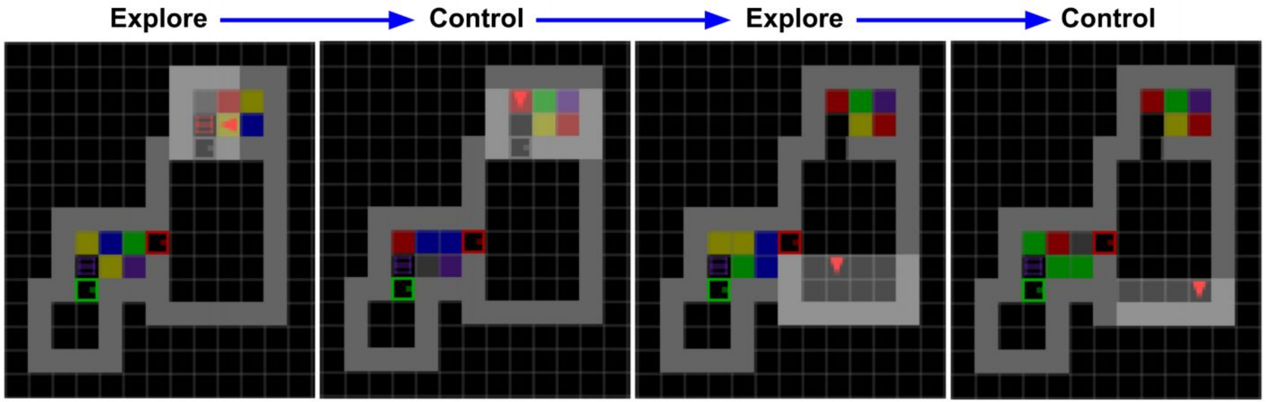
Agent get trapped in local maxima of entropy



Our Solution: Escaping the local max via Adversarial Surprise

Exploration as a game between a surprise minimizing policy and surprise maximizing policy





- Sequential game
- Two policies, one RL agent
- Shared observation density model
- Reward = surprise controller
- Duality noisy TV and dark room
- Emergence of complexity

$$\max_{\pi^E} \min_{\pi^C} -\mathbb{E}\left[\sum_{t=t^C}^{t^C+k} \log p_\theta(o_t)\right]$$

Theoretical Result: State Coverage

Lemma 1. $-\mathbb{E}_\pi \sum_{t=0}^{\infty} \log p_\theta(o_t) \geq H(d^\pi(o))$

Lemma 2. *In a block MDP (BMDP) [31], we can decompose the observation marginal entropy:*

$$H(d^\pi(o)) = \mathbb{E}_{d^\pi(s)} H(p(O|S=s)) + H(d^\pi(s)) \tag{6}$$

- Assumption 1: Block MDP
- Assumption 2: Dark rooms coverage

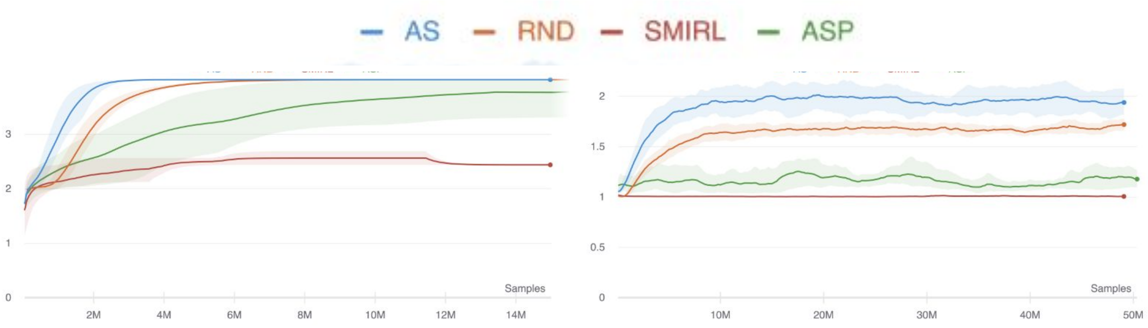
Theorem 1. *Under Assumptions 1 and 2 the Markov chain induced by the following AS game:*

$$\max_{d^{\pi_E}(s_0)} \min_{d^{\pi_C}_{1:\frac{T}{2}}(s|s_0)} H\left(d^{\pi_C}_{\frac{T}{2}}(o)\right) \tag{9}$$

$\frac{T}{2}$-covers the state space, i.e., for all states s, there is a state s' such that $d^\pi(s') > 0$ and $d(s, s') \leq \frac{T}{2}$, where d^π is the marginal induced by the game between the Explore (π_E) and Control (π_C) agents.

Empirical Results: Minigrid and Atari

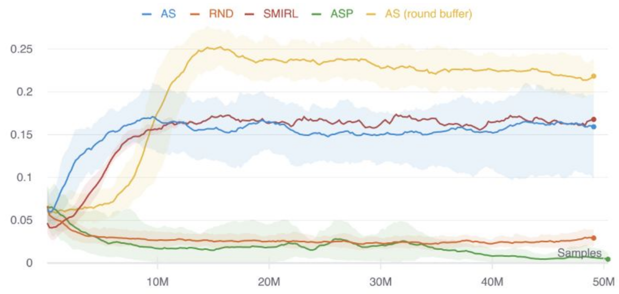
Exploration



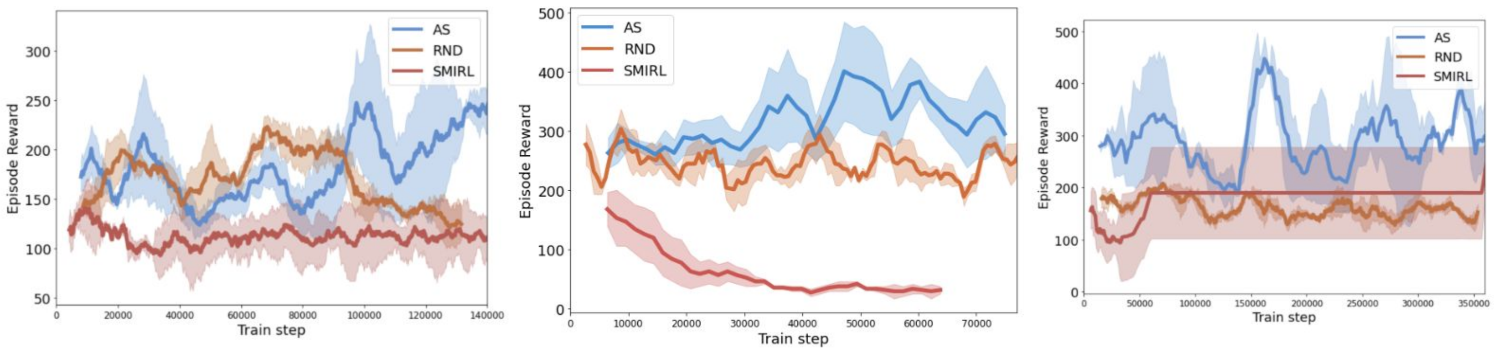
(a) Cumulative exploration

(b) Exploration within an episode

Control



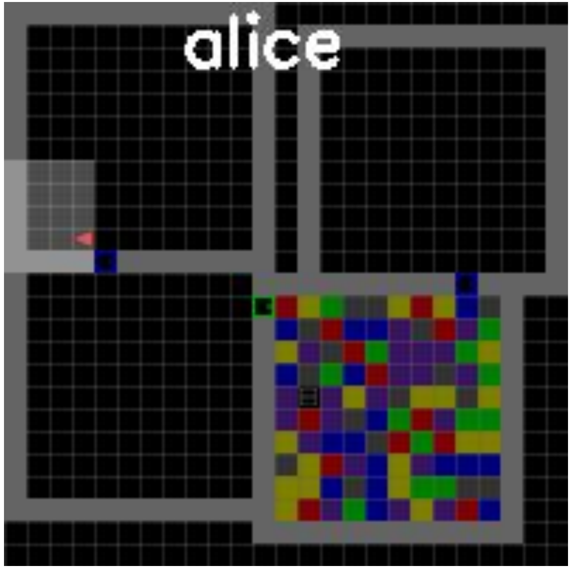
No-reward learning of complex behavior



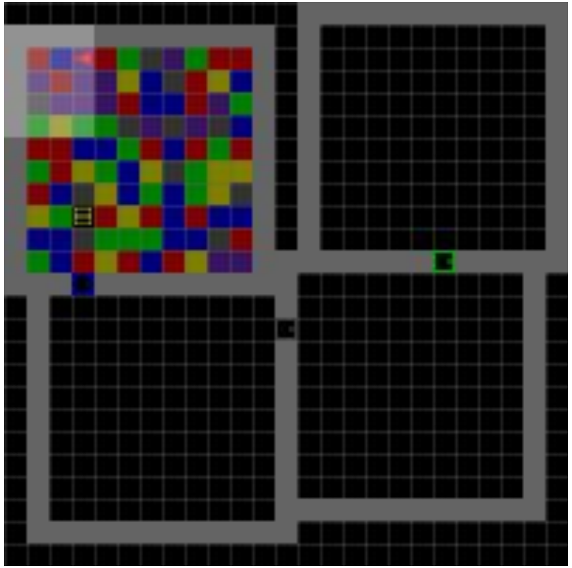
(a) Berzerk

(b) Assault

(c) Space Invaders



Adversarial Surprise



Random Network Distillation

Exploring complex environments: takeaway

Count-based methods need to be amortized by a neural network.

Maximizing surprise do not work in partially observable stochastic environment.

A Fickinger, N Jaques, S Parajuli, M Chang, N Rhinehart, G Berseth, S Russell, S Levine, Explore and Control with Adversarial Surprise

Learning without reward

The problem of reward in real-world problems

Costly to design

Reward hacking

Imitation Learning

Goal: learning behavior from demonstrations

$$\{(o_1^i, a_1^i ..., o_T^i, a_T^i)\}_{i \in [N]} \mapsto \pi(a_t | o_t)$$

Solutions

Behavioral Cloning: reduce to supervised learning

$$\mathcal{L}_{\mathrm{BC}}(\theta) \;=\; -\frac{1}{N\,T} \sum_{i=1}^{N} \sum_{t=1}^{T} \log \pi_\theta\big(a_t^i \mid o_t^i\big)$$

Inverse Reinforcement Learning

Learn the reward that best explain the behavior

Problem

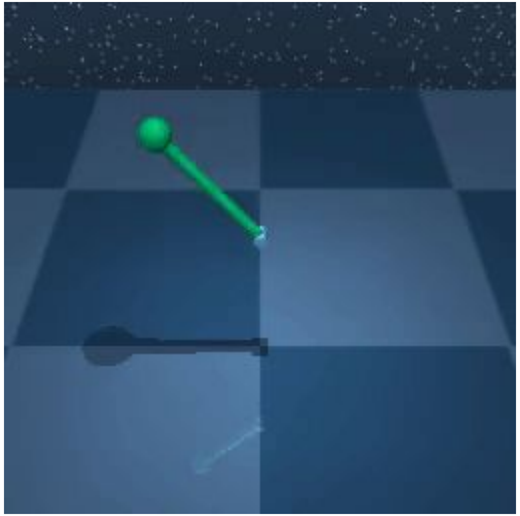
1) Does not easily transfer to other systems with potentially different state action space (e.g., learning a robot policy from human demonstration)
2) Often require many trajectories from the expert

Example 1

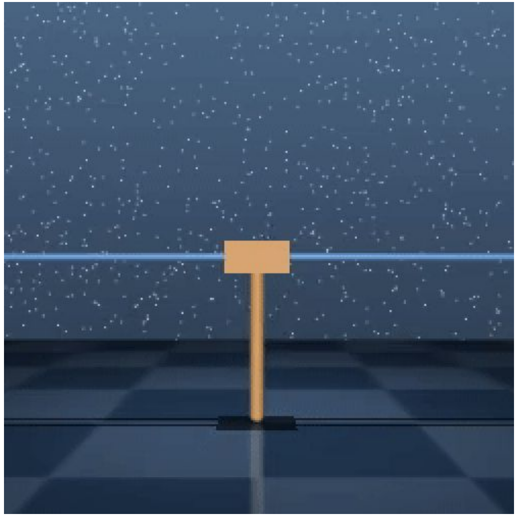
?

Human understand the transfer

Example 2



?



4D

7D

Human understand the transfer

Example 3

?

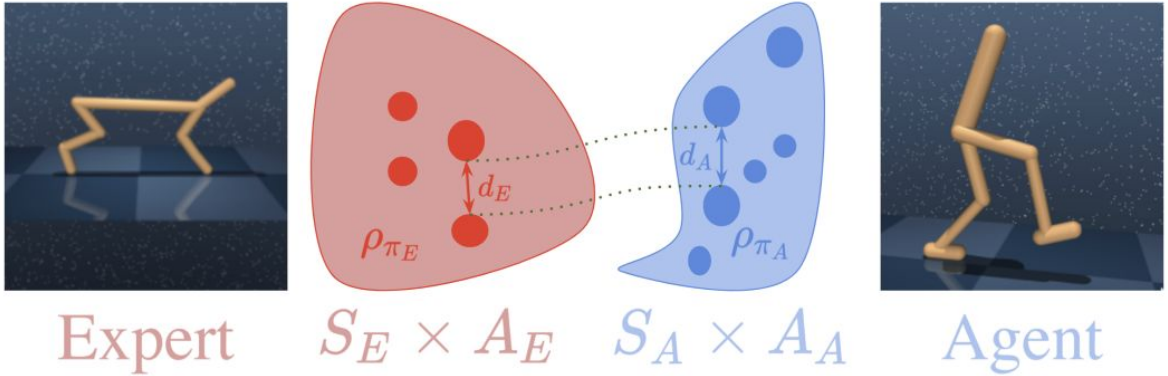
30D

23D

Human understand the transfer

Our solution: Gromov-Wasserstein Imitation Learning

Insight: humans exploit geometric invariance



$$\mathcal{GW}((\mathcal{X}, d_{\mathcal{X}}, \mu_{\mathcal{X}}), (\mathcal{Y}, d_{\mathcal{Y}}, \mu_{\mathcal{Y}}))^2 = \min_{u \in \mathcal{U}(\mu_{\mathcal{X}}, \mu_{\mathcal{Y}})} \sum_{\mathcal{X}^2 \times \mathcal{Y}^2} |d_{\mathcal{X}}(x, x') - d_{\mathcal{Y}}(y, y')|^2 u_{x,y} u_{x',y'}$$

Our solution: Gromov-Wasserstein Imitation Learning

Algorithm 1 Gromov-Wasserstein imitation learning from a single expert demonstration.

Inputs: expert demonstration τ, metrics on the expert (d_E) and agent (d_A) space
Initialize the imitation agent's policy π_θ and value estimates V_θ
while Unconverged **do**
    Collect an episode τ'
    Compute $\mathcal{GW}(\tau, \tau')$
    Set pseudo-rewards r with eq. (7)
    Update π_θ and V_θ to optimize the pseudo-rewards
end while

Intuition: optimal behaviors are connected via isometry

Theorem 1. *Consider two MDPs*

$$M_E = (S_E, A_E, R_E, P_E, p_E, \gamma) \quad \text{and} \quad M_A = (S_A, A_A, R_A, P_A, p_A, \gamma).$$

Suppose that there exists four distances $d_E^S, d_E^A, d_A^S, d_A^A$ *defined on* S_E*,* A_E*,* S_A *and* A_E *respectively, and two isometries* $\phi : (S_E, d_E^S) \to (S_A, d_A^S)$ *and* $\psi : (A_E, d_E^S) \to (A_S, d_A^S)$ *such that for all* $(s_E, a_E, s'_E) \in S_E \times A_E \times S_E$ *the three following conditions hold:*

$$R(s_E, a_E) = R_A(\phi(s_E), \psi(a_E)) \tag{3}$$

$$P_{E s_E, a_E}(s'_E) = P_{A \phi(s_E) \psi(a_E)}(\phi(s'_E)) \tag{4}$$

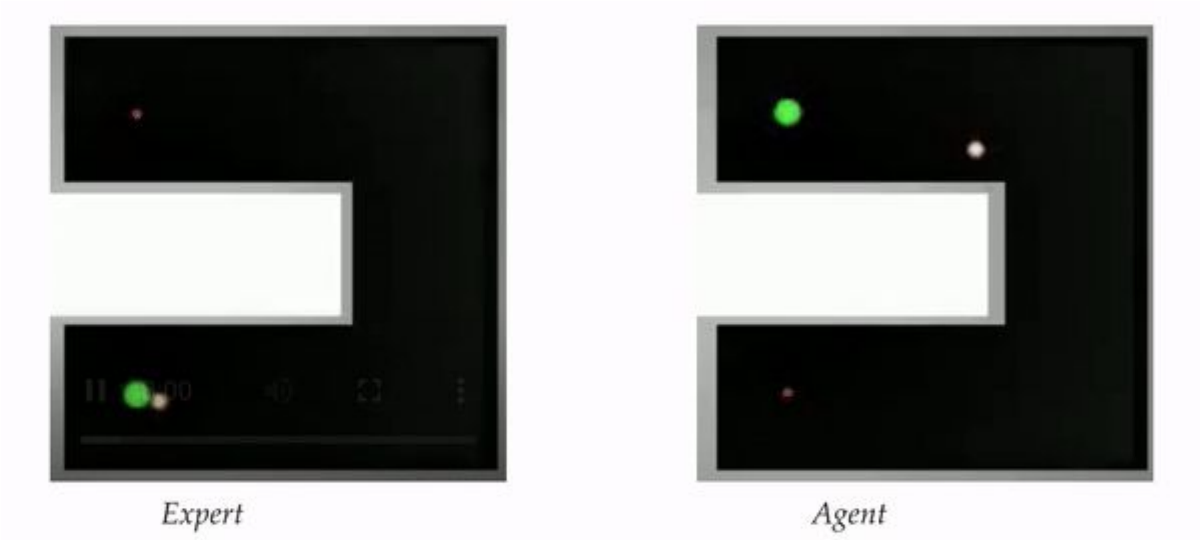
$$p_E(s_E) = p_A(\phi(s_E)). \tag{5}$$

Consider an optimal policy π_E^* *in* M_E*. Suppose that* π_{GW} *minimizes* $\mathcal{GW}(\pi_E^*, \pi_{GW})$ *with*

$$d_E : (s_E, a_E) \mapsto d_E^S(s_E) + d_E^A(a_E) \quad \text{and} \quad d_A : (s_A, a_A) \mapsto d_A^S(s_A) + d_A^A(a_A).$$

Then π_{GW} *is isometric to an optimal policy in* M_A*.*

Rigid Transformation



Slightly Different State-Action Spaces

Expert

Agent

Significantly Different State-Action Spaces

Learning without reward: takeaway

Designing informative rewards for real-world environments is challenging

If we have a demo of a similar task, exploit the invariance to learn a reward (here isometry)

A Fickinger, S Cohen, S Russell, B Amos, CROSS-DOMAIN IMITATION LEARNING VIA OPTIMAL TRANSPORT

Simulating complex environments

Traditional Simulations

- strong domain expertise
- high cost
- real-time requires approximation
- limited generalization

World Models

Oasis (2024)

World Models

Problem: does not generalize

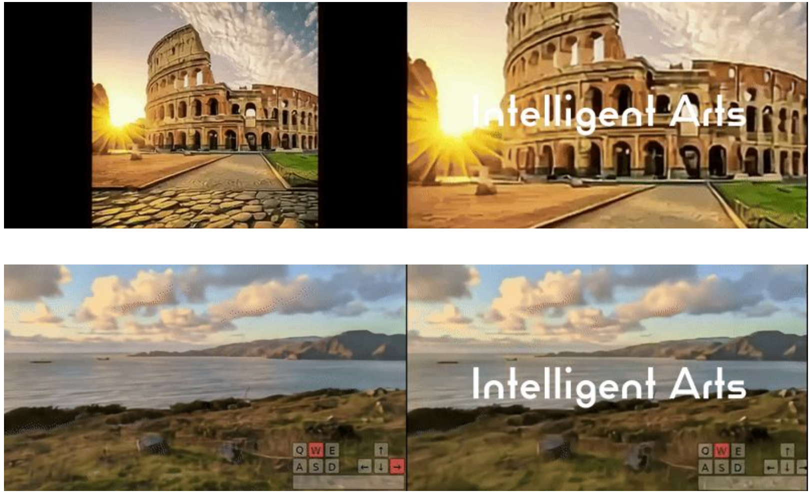
Oasis (2024)

Large-scale generative model

Problem: generalize but bounded complexity

Genie 2: A large-scale foundation world model (Deepmind 2024)

Our solution



Thank you

arnaud.fickinger@berkeley.edu