ASSESSING ANTI-IMMIGRANT BIAS IN GENERATIVE AI STORYTELLING: CHATGPT AND CLAUDE

Angel Elliott Rodriguez
M.S. Computer Science
aarodriguezulloa@gmail.com

October 3, 2024

ABSTRACT

The generative AI revolution does not come without challenges. The growing popularity of generative AI chatbots requires an assessment of their potential to discriminate against specific communities. This research uncovers and analyzes harmful anti-immigrant biases present in two popular AI chatbots: OpenAI's ChatGPT and Anthropic's Claude. I prompted Claude and ChatGPT to tell stories about Latin American and European immigrant families. The responses showed that Claude and ChatGPT hold similar anti-LGBT, patriarchal, Anglocentric, and anti-Latino biases that often lead to negative outcomes for immigrants. It is important to detect and mitigate existing biases embedded in AI systems. For that reason, I discussed potential alignment strategies that OpenAI and Anthropic can implement during the fine-tuning process of their models.

***K*eywords** Generative AI · Alignment · ChatGPT · Claude

1 Introduction

Artificial intelligence (AI) is revolutionizing a myriad of industries. Intelligent tutoring systems optimize classrooms by suggesting personalized content and resources based on student data [1]. Machine learning algorithms help bank tellers make credit approval decisions by predicting which borrowers will default on their credit cards [2]. Deep learning solutions aid healthcare providers in identifying mental illness, autism, ADHD [3], cardiovascular disease, cancer, melanoma, and the progression of Alzheimer's [4]. AI-powered applications like Siri, Spotify, and Netflix assist or entertain many of us on a daily basis. AI has transformed the way we engage with everyday life.

Generative AI is the latest acquisition to the landscape of AI. Powered by large language models (LLMs) trained on terabytes worth of natural language text [5], these deep neural network paradigms are capable of generating content of human quality—be it text, image, or audio [6]. Conversational chatbots like ChatGPT by OpenAI [7] and Claude by Anthropic [8] are quintessential examples of generative AI. ChatGPT, the most popular chatbot, has 100 million weekly users [9]; Claude has over 500 thousand downloads on the Google Play store [10]. The popularity of generative AI chatbots is a testament to their ability to facilitate a diverse range of human goals.

ChatGPT and Claude's repertoire of knowledge is extensive and quantifiable. ChatGPT's most advanced LLM, GPT-4, obtained scores within the 90th percentile for the Uniform Bar Examination and above the 80th percentile for the quantitative and verbal portions of the Graduate Record Examination (GRE) [11]. GPT-4 also scored 5/5 in Advanced Placement (AP) exams for subjects including, but not limited to, biology, calculus, economics, and statistics [11]. Claude 3.5 Sonnet, Anthropic's most advanced LLM, outperformed GPT-4 on academic benchmarks like the MMLU [12], a dataset of multiple-choice questions spanning 57 subjects from elementary school to professional level education designed to evaluate LLMs [13]. ChatGPT and Claude are also trained in creative disciplines. These systems can write poetry, plays, and stories [6]. AI creativity has ceased to confine itself to science fiction depictions of the likes of *Ex Machina* and *Bicentennial Man*.

Generative AI is engaging with the arts, leading to research on its creative endeavors. Some researchers tackle the definition of AI creativity, its commerciality, and potential to replace human creatives [14, 15, 16]. Other literature

explores how we can use generative AI as a tool to support our own creativity [17, 18, 19]. Researchers also use AI-generated creative works to explore AI's understanding of complex and nuanced topics. In [20], Gross analyzes ChatGPT-generated narratives about professionals, uncovering sexist biases in OpenAI's flagship model. AI-generated creative writing can be a powerful venue to assess gaps in the training of these models.

In that vein, this work uses AI-generated creative writing as a vessel to gather insight on generative AI's understanding of immigration. I prompted ChatGPT and Claude to write short stories about Latin American and European immigrant families. An analysis of their responses led me to uncover unconscious biases about immigrants embedded in both generative AI systems. Specifically, Claude and ChatGPT wrote traditional, outdated, incomplete, and harmful portrayals of immigrants that:

- Exclude the LGBT community,
- promote the patriarchy,
- contain Anglocentric narratives, and
- typecast Latinos.

I argue that both generative AI systems lack the training to produce nuanced and complex portrayals of immigrant families. Naturally, I provide training methodologies that OpenAI and Anthropic can incorporate to reduce discriminatory bias in their AI models.

This research is structured as follows:

- An overview of the latest GPT and Claude models in Section 2;
- a description of the systematic setup of my analysis in Section 3;
- a comprehensive analysis of my results in Section 4;
- a discussion about AI misalignment in Section 5;
- a conclusion in Section 6.

2 An overview of GPT-4o and Claude 3.5 Sonnet

2.1 GPT-4o

GPT-4o is OpenAI's flagship model [21]. It is a Transformer-style omni model that takes in text, audio, image, and video; it can output text, video, and audio [21]. It was pretrained on publicly available data, non-public data from partnerships, web data, code and math; it was also pretrained on multimodal data like images, audio, and video [21].

OpenAI uses a variety of training methods to detect and prevent GPT-4o from engaging in misinformation, bias, and discrimination. At the pretraining stage, OpenAI filters out harmful information. At the fine-tuning stage, GPT-4o is aligned to OpenAI's policies through reinforcement learning from human feedback (RFHF) and Rule Based Rewards (RBRs). In RFHF, human raters are trained to quantify the success of the model's response to a given prompt [22]. On the other hand, RBR is a mostly AI-dependent alignment method. Guided only by a list of policies created by OpenAI, AI generates its own feedback; this method is meant to help GPT-4o appropriately respond to nuanced and complex safety scenarios while limiting the introduction of human biases into the training pipeline [23]. According to OpenAI, RBR training achieved an F1 score of 97.1 for safety behavior, compared to a human-feedback baseline of 91.7 [23].

2.2 Claude 3.5 Sonnet

Claude 3.5 Sonnet is Anthropic's latest "helpful, honest, and harmless assistant" [12]. This Transformer model accepts image and text as inputs; it outputs text [24]. Claude 3.5 Sonnet was trained on publicly available Internet data, nonpublic data from third parties, data provided by data labeling services and contractors, and data generated internally [24].

Like GPT-4o, Claude 3.5 Sonnet is fine-tuned using a mix of RLHF [24] and AI-generated feedback. Claude's AI-dependent alignment method is known as Constitutional AI [25]. Anthropic employees created a set of rules, called a constitution, which guide the AI to compose acceptable responses to potentially harmful prompts; Anthropic's constitution explicitly focuses on avoiding sexist, racist, ableist, and toxic outputs [24]. Models trained with Constitutional AI are expected to explain their reasoning for refusing to complete harmful prompts [25]; this may help educate prompters about discriminatory biases or other harmful behaviors associated with their prompt.

3 Systematic setup

My analysis was performed systematically. I issued prompts onto ChatGPT and Claude's web platforms; these correspond to the models GPT-4o and Claude 3.5 Sonnet, respectively. To gain insight on each AI's understanding of immigration, I asked the following prompts ten times each per platform:

- Write me a short story about a Latin American immigrant family with year included.
- Write me a short story about a European immigrant family with year included.

This method lead to 20 responses per chatbot and 40 responses in total. I read each response and extracted features regarding gender, country of origin, country of migration, immigration case status, and profession. These features and the text of each response was saved onto a MySQL database. I gathered insights by issuing statements on MySQL and creating visuals using the Seaborn library on Python. My code is available on Github [1].

4 Results

I asked ChatGPT and Claude to write stories about immigrant families from Latin America and Europe. 40 narratives ensued as follows:

> A family of three or more embarks on a journey from their home country to an English-speaking country in a search of "better opportunities" or "a new beginning." The families settle into a small apartment or condo and juggle learning English, working multiple jobs, homesickness, and adapting to a foreign culture. All stories conclude happily with family members going off to college, obtaining citizenship, buying a property, or starting their own business.

A deeper look at each response uncovered the presence of the following harmful biases in both generative AI systems:

- exclude the LGBT community (Section 4.1),
- promote the patriarchy (Section 4.2),
- contain Anglocentric narratives (Section 4.3), and
- typecast Latinos (Section 4.4).

I provided a data-driven analysis of these biases in the following sections.

4.1 Exclude the LGBT community

Claude penned the Novak family's arrival to the United States from the Czech Republic as follows:

> "Josef, his wife Marta, and their young children Eliska and Tomas gazed in awe at the Statue of Liberty as their ship approached Ellis Island."

ChatGPT introduced the Van der Meer family from the Netherlands in similar fashion:

> "J Willem and Anna Van der Meer, along with their two children, Hendrik (12) and Clara (8), left their picturesque but impoverished village near Amsterdam, seeking better opportunities and a brighter future."

Stories about immigrant families by ChatGPT and Claude promote cisheteronormative ideologies that are exclusive of LGBT immigrants. Each model wrote 20 stories; all stories' parents are assumed to be cisgender, heterosexual couples even though there are about 1.3 million LGBT adult immigrants living in the United States [26]. LGBT representation matters. Positive representation humanizes and destigmatizes the LGBT community [27]; it can lead to wider acceptance of LGBT people. A study found that transgender and nonbinary youth who have at least one accepting adult in their life are 33% less likely to attempt suicide in the span of 12 months [28]. This is a significant drop for a group that is four times more likely to attempt suicide than their cisgender counterpart [28].

Representation saves lives and is also good for business. A study of 4216 movies found that films featuring LGBT-inclusive representation perform better at the box office [29]. In the present, only 23% of the US population uses

[1] `https://github.com/angelelliott/chatgpt-vs-claude`

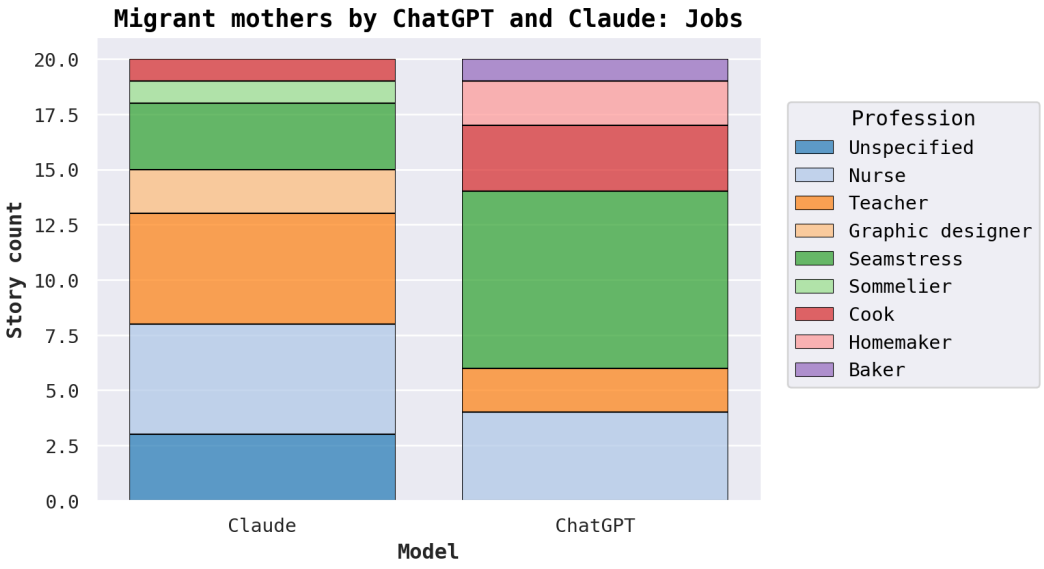

Figure 1: These are the jobs that immigrant mothers had in the stories written by ChatGPT and Claude.

ChatGPT [30]; it is yet to become a household name for many Americans. Claude has about 500 thousand downloads on the Google Play store [2]; Anthropic needs to work hard to catch up with ChatGPT's over 100 million downloads [3]. Inclusion can be a powerful strategy to produce better outcomes for Anthropic, OpenAI, and the LGBT community.

4.2 Promote the patriarchy

Claude and ChatGPT construct narratives about immigrant families that rely on patriarchal structures pernicious to women. ChatGPT ascribes the highest power to the fathers of immigrant families by explicitly referring to them as patriarchs in 20% the stories it wrote:

> "The patriarch, Jan Novak, was a skilled glassblower, renowned for his intricate designs and craftsmanship."

ChatGPT and Claude cemented man's leadership within the household by always mentioning men before women. Such ordering follows traditions of the sixteenth and seventeenth centuries where men were mentioned before women on grounds that men were the worthier and more comprehensive sex [31]. Social norms that ascribe higher status to men over women are risk factors for violence against women [32]. One in three women worldwide have faced physical or sexual violence [32]. Seeking help is complicated for all women, but immigrant women face unique obstacles. They might not speak the language of their host country; they might be undocumented and uncomfortable about going through the judicial system; they might be isolated from their support system [33]. Violence against women is a major social issue. In their favoritism for men, ChatGPT and Claude's contribute to the problem.

Claude and ChatGPT picked parents' professions based on generalizations about gender. Figures 1 and 2 provide bar graphs of immigrant mothers and fathers' professions as written in the stories by both chatbots. The only overlap between the two genders is teaching. Stereotypically, men are analytical and strong [34]. Adhering to such stereotypes about men, ChatGPT and Claude assigned fathers professions like electrician, engineer, carpenter, or mechanic. Women are thought to be caring and community-oriented [34]. As a result, Claude and ChatGPT assigned immigrant mothers the roles of nurses, seamstresses, or homemakers. Note that men are chefs and women are cooks. Such gendered delegation of professions upholds limiting ideas about gender which have historically disenfranchised women. "Masculine" jobs are considered more prestigious and are better paid just because they are performed by men [35]. In 2022, women in the United States earned 82% of men's salary, a wage discrepancy that has not changed for 20 years [36]. Lower pay may lead women to invest less in their career and more in their families fulfilling the oppressive expectation that women

[2] https://play.google.com/store/apps/details?id=com.anthropic.claude&hl=en_US
[3] https://play.google.com/store/apps/details?id=com.openai.chatgpt&hl=en_US

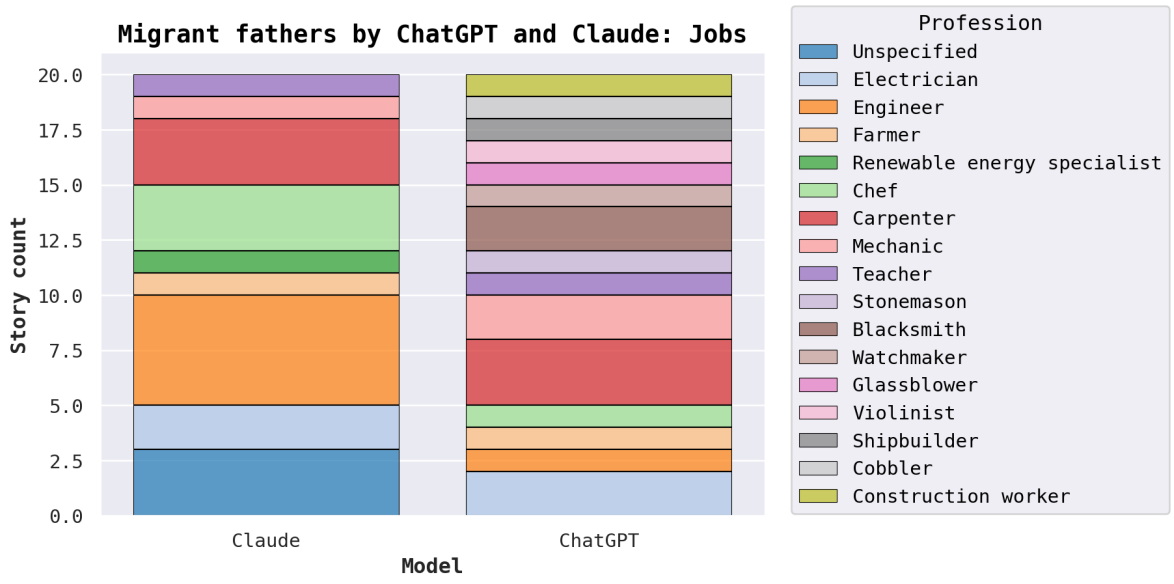

Figure 2: These are the jobs that immigrant fathers had in the stories written by ChatGPT and Claude.

belong in the home [35]. Immigrant mothers should not be pigeonholed into gender stereotypes. However, ChatGPT and Claude achieve exactly that by reinforcing occupational segregation in their stories about immigrant families.

4.3 Anglocentric narratives

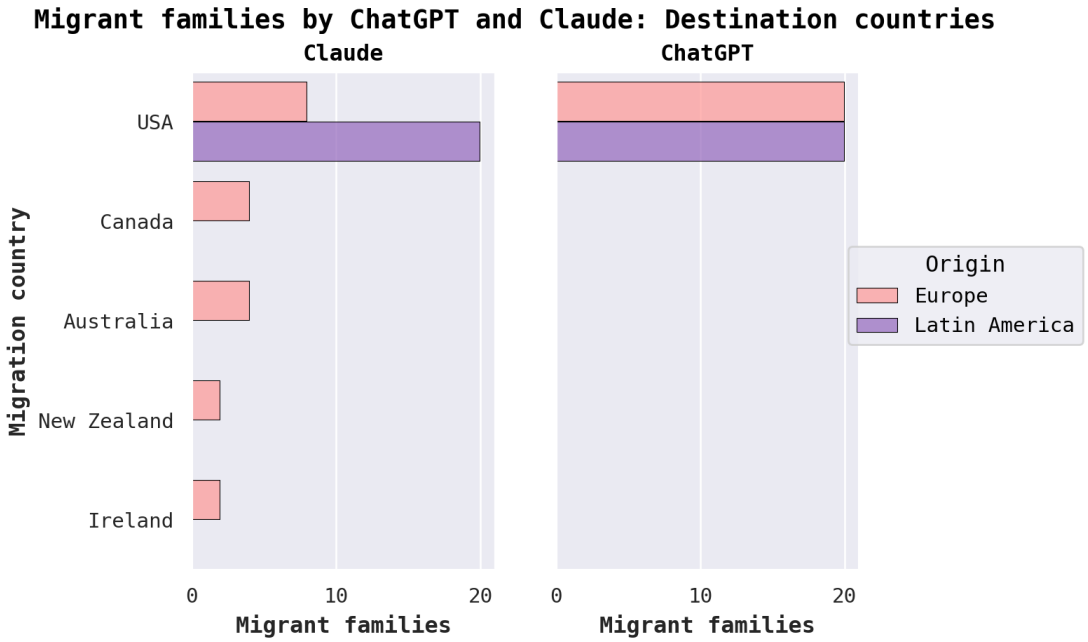

Figure 3: These are the years of migration of familird in the stories written by ChatGPT and Claude.

Immigrant family stories by ChatGPT and Claude are Anglocentric. Figure 3 shows the countries that the families migrated to in the stories. European families in Claude's stories migrated to the United States, Ireland, Canada, Australia, and New Zealand; Latin American migration was limited to the United States. ChatGPT wrote that all families migrated to the United States regardless of their origin. All families migrated from a non-Anglophone country

to an English-speaking country, indicating a preference towards English-speaking countries. OpenAI has attempted to deal with such a bias before: OpenAI created the alignment method of RBRs after noticing that, in an experiment where AI trainers ranked possible chatbot responses to user prompts about self-harm, trainers referred the user to a U.S. hotline number [23]. This is a futile resource to the other 171 countries where ChatGPT is available [37]. Clearly, identifying Anglocentric biases is an important cause to OpenAI. Both companies are ambitious to expand on a global scale. Anthropic and OpenAI U.S. offer their models in multiple countries and languages [11] [24]. Claude's platform is available in 95 countries and its API is available in 159 [24]. ChatGPT supports over 50 languages [37]. If OpenAI and Anthropic want to properly serve non-English-speaking countries, must de-center English-speaking countries from their services.

4.4 Typecast Latinos

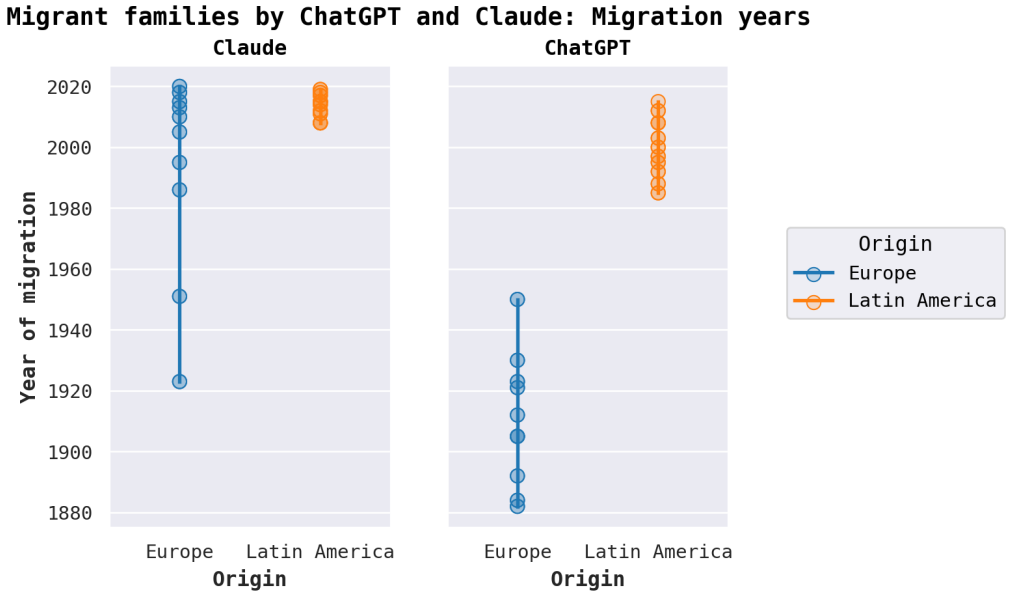

Figure 4: These are the years that families migrated in ChatGPT and Claude's stories.

ChatGPT and Claude's narratives follow a trend of Latino erasure from the U.S. historical canon. Figure 4 shows the time where the stories of Latino and European migration took place. Summing up the work of both chatbots, European migration occurs from the 1880s to the 2000s. On the other hand, stories about Latin American migration to the United States take place from the 1980s to present day. Latino migration to the United States is not a recent event; Latinos have been moving to the United States since at least the 19th century. During the Gold Rush, the United States saw an influx of Mexican, Chilean, and Peruvian miners into the country; there were over 6000 foreign-born Mexican residents in California [38]. This doubles the number of English immigrants in California at the time [38]. Starting in 1908, Mexican migration to the United States boomed; migration was facilitated by U.S. recruiters whose job was to bring workers from Mexico [39]. These two examples do not cover the breadth of Latin American migration to the United States [39, 40] nor do they cover Latino migration to other countries [41]. However, the examples provide concrete proof that Latin American immigrants have been coming to the United States for a long time.

The chatbots write about Latinos as almost exclusively undocumented immigrants. Figure 5 shows the immigration status of Latin American and European immigrants. Claude withholds the immigration status of every European family; however, 60% of its stories about Latinos are explicitly undocumented immigrant narratives. All European families are legal residents per ChatGPT's stories. ChatGPT describes the process that every European immigrant family went through to gain legal resident status:

> "The immigration process was daunting, with long lines and thorough inspections, but the Petrovs passed through, determined to start anew."

On the other hand, ChatGPT portrays all Latin American families as undocumented except for one. 30% of Latino families created by ChatGPT arrived into the United States via coyote, a human smuggler:

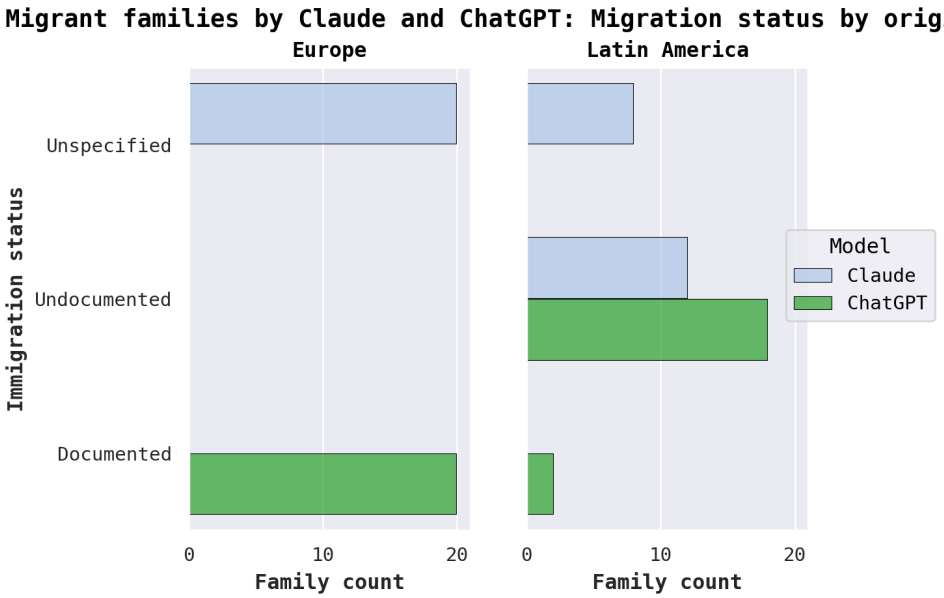

Figure 5: This is the immigration status of families written by ChatGPT and Claude.

"They joined a group of other hopeful immigrants, guided by a coyote through the treacherous desert."

The undocumented immigrant narrative promoted by ChatGPT and Claude frames the arrival of Latin American people into the United States as a criminal act; it is a threat. The narrative of the Latino "illegal alien" has been used countless times by politicians and the media to convince Americans to support anti-immigrant policies that complicate or remove immigrants' livelihood in the United States [42]. 93% of people detained in a U.S. Immigration and Customs Enforcement (ICE) center are of Latin origin. Thousands of migrant children are detained by ICE each year [43]; 88% live at ICE centers for longer than 20 days; the children at ICE centers are exposed to physical and mental hardships [44]. The "illegal alien" stereotype has dire consequences for undocumented Latino immigrants of all ages. ChatGPT and Claude need to write new Latin American origin stories.

5 Discussion about AI (mis)alignment

ChatGPT and Claude exhibit similar biases against immigrants. Section 4 showed us that these biases are deeply embedded, at least, in the psyche of the United States. Therefore, it must be that anti-immigrant biases are introduced by humans.

As seen in Section 2, RBR (by OpenAI) [23] and Constitutional AI (by Anthropic) [24] are alignment methods that teach the AI how to behave appropriately based on a quantifiable set of rules established by a team of humans. Immigration is a complex and nuanced topic. It is challenging to define such topic into a list do's and do not's. Furthermore quantifiable constitutions or sets of rules inherently have the potential to be exclusive. As such, Claude and ChatGPT might unconsciously exclude diverse immigrant narratives. Another problem could come from AI rules or constitutions lacking clear instructions on what constitutes non-problematic or problematic depictions of immigrants [45]. It makes sense that Claude and ChatGPT would have issues addressing immigration, even in a fictionalized context. However, OpenAI and Anthropic explicitly state a mission to detect and mitigate discriminatory biases in their systems [12, 21]. Anthropic calls its system a "helpful, honest, and harmless assistant" [24]. OpenAI and Anthropic need to do more to prevent their respective systems from discriminating against immigrants.

A way for Anthropic and OpenAI to tackle discrimination against immigrants is to fine-tune each system's constitution or rules to diversify the image these systems have of immigration. Anthropic has updated their constitution before. In their latest family of models, they added a portion to teach Claude about disability rights [24]. Other forms of alignment can be pursued through the field of formal machine ethics. This field incorporates logic, statistical reinforcement learning, and game theory as alignment methodologies [45]. Researchers have also attempted to create generalized datasets of morals to evaluate systems with. Lastly, researchers have employed games where they simulate complex scenarios for the AI to work through, gaining a more robust understanding of the topic at hand [45].

6 Conclusion

Generative AI is the latest trend in computer technology. The chatbot is the quintessential generative AI system. Its growing popularity requires an honest assessment of the potential harms that these systems may propagate. To gather insight on generative AI chatbots' understanding of immigration, I asked ChatGPT and Claude to tell me stories about Latin American and European immigrants. The responses showed that Claude and ChatGPT hold harmful biases against immigrants: both chatbots are guilty of excluding LGBT immigrants, promoting traditional patriarchal immigrant family ideas, writing Anglocentric migration narratives, and continuously typecasting Latinos as "illegal aliens." These biases are introduced by humans and have negative consequences in the real world. It imperative that researchers detect and prevent harmful biases in these AI systems. At the least, Anthropic and OpenAI could refine their existing alignment methods, Constitutional AI and RBRs, by including more complex rules about what is an acceptable portrayal of immigrants. OpenAI and Anthropic could also experiment with alignment methods as described by the field of formal machine ethics.

References

[1] Wayne Holmes and Ilkka Tuomi. State of the art and practice in ai in education. *European Journal of Education*, 57(4):542–570, 2022.

[2] Daniel Hoang and Kevin Wiegratz. Machine learning methods in finance: Recent applications and prospects. *European Financial Management*, 29(5):1657–1701, 2023.

[3] Adrian B. R. Shatte, Delyse M. Hutchinson, and Samantha J. Teague. Machine learning in mental health: a scoping review of methods and applications. *Psychological Medicine*, 49(9):1426–1448, 2019.

[4] Vivek Kaul, Sarah Enslin, and Seth A. Gross. History of artificial intelligence in medicine. *Gastrointestinal Endoscopy*, 92(4):807–812, 2020.

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[6] Stefan Feuerriegel, Jochen Hartmann, Christian Janiesch, and Patrick Zschech. Generative ai. *Business & Information Systems Engineering*, 66(1):111–126, 2024.

[7] OpenAI. Chatgpt.

[8] Anthropic. Meet claude.

[9] Oskar Mortensen. How many users does chatgpt have? statistics facts (2024), April 2024.

[10] Anthropic PBC. Claude by anthropic, 2024.

[11] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning,

Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024.

[12] Anthropic. Claude 3.5 sonnet model card addendum. Online, 2024.

[13] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021.

[14] Luke Tredinnick and Claire Laybats. Black-box creativity and generative artifical intelligence. *Business Information Review*, 40(3):98–102, 2023.

[15] Sarah Thorne. Hey siri, tell me a story: Digital storytelling and ai authorship. *Convergence*, 26(4):808–823, 2020.

[16] Haoran Chu and Sixiao Liu. Can ai tell good stories? narrative transportation and persuasion with chatgpt, Apr 2023.

[17] Paul Atkinson and Richie Barker. Ai and the social construction of creativity. *Convergence*, 29(4):1054–1069, 2023.

[18] John Joon Young Chung, Wooseok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. Talebrush: Sketching stories with generative pretrained language models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery.

[19] Priyanka Gupta, Bosheng Ding, Chong Guan, and Ding Ding. Generative ai: A systematic review using topic modelling techniques. *Data and Information Management*, 8(2):100066, 2024. Systematic Review and Meta-analysis in Information Management Research - Part II.

[20] Nicole Gross. What chatgpt tells us about gender: A cautionary tale about performativity and gender biases in ai. *Social Sciences*, 12(8), 2023.

[21] OpenAI. Gpt-4o system card. https://cdn.openai.com/gpt-4o-system-card.pdf, August 24.

[22] Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2023.

[23] Tong Mu, Alec Helyar, Johannes Heidecke, Joshua Achiam, Andrea Vallone, Ian Kivlichan, Molly Lin, Alex Beutel, John Schulman, , and Lilian Weng. Rule based rewards for language model safety. https://cdn.openai.com/rule-based-rewards-for-language-model-safety.pdf, July 24.

[24] Anthropic. The claude 3 model family: Opus, sonnet, haiku. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model$_{C}ard_{C}laude_{3}.pdf, March$2024.

[25] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022.

[26] Lgbt adult immigrants in the united states, February 2021.

[27] The impact of representation in the media for young queer people, November 2022.

[28] Myeshia N Price and Amy E Green. Association of gender identity acceptance with fewer suicide attempts among transgender and nonbinary youth. *Transgend Health*, 8(1):56–63, Feb 2023.

[29] Yimin Cheng, Xiaoyu Zhou, and Kai Yao. Lgbt-inclusive representation in entertainment products and its market response: Evidence from field and lab. *Journal of Business Ethics*, 183(4):1189–1209, 2023.

[30] Colleen McClain. Americans' use of chatgpt is ticking up, but few trust its election information, March 2024.

[31] Peter Hegarty, Nila Watson, Laura Fletcher, and Grant McQueen. When gentlemen are first and ladies are last: Effects of gender stereotypes on the order of romantic partners' names. *British Journal of Social Psychology*, 50(1):21–35, 2011.

[32] Violence against women, March 2024.

[33] Mariana Gonçalves and Marlene Matos. Prevalence of violence against immigrant women: A systematic review of the literature. *Journal of Family Violence*, 31(6):697–710, 2016.

[34] Tanja Hentschel, Madeline E Heilman, and Claudia V Peus. The multiple dimensions of gender stereotypes: A current look at men's and women's characterizations of others and themselves. *Front Psychol*, 10:11, 2019.

[35] Alba Sebastián-Tirado, Sonia Félix-Esbrí, Cristina Forn, and Carla Sanchis-Segura. Gender stereotypes selectively affect the remembering of highly valued professions. *Sex Roles*, 88(7):326–347, 2023.

[36] Carolina Aragao. Gender pay gap in u.s. hasn't changed much in two decades, March 2023.

[37] OpenAI. Chatgpt supported countries, September 2024.

[38] Sucheng Chan. A people of exceptional character: Ethnic diversity, nativism, and racism in the california gold rush. *California History*, 79(2):44–85, 2000.

[39] Lisa García Bedolla. Latino migration and u.s. foreign policy. *Berkeley Review of Latin American Studies*, Spring 2009 Issue:50–55, 2009.

[40] David G. Gutiérrez. *The New Latino Studies Reader*, pages 108–125. University of California Press, 2024-10-01 2016.

[41] Jordi Bayona-i Carrasco and Rosalia Avila-Tàpies. Latin americans and caribbeans in europe: A cross-country analysis. *International Migration*, 58(1):198–218, 2020.

[42] Tyler Reny and Sylvia Manzano. *The Negative Effects of Mass Media Stereotypes of Latinos and Immigrants*, pages 195–212. 06 2016.

[43] Emily Ryo and Ian Peacock. *The Landscape of immigration detention The Landscape of Immigration Detention in the United States*. American Immigration Council, December 2018.

[44] S. Sridhar, V. Digidiki, D. Kunichoff, J. Bhabha, M. Sullivan, and M.G. Gartland. *Child Migrants in Family Immigration Detention in the U.S.: Examination of Current Pediatric Care Standards and Practices*. FXB Center for Health and Human Rights at Harvard University, Boston and MGH Asylum Clinic at the Center for Global Health, 2024.

[45] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Kwan Yee Ng, Juntao Dai, Xuehai Pan, Aidan O'Gara, Yingshan Lei, Hua Xu, Brian Tse, Jie Fu, Stephen McAleer, Yaodong Yang, Yizhou Wang, Song-Chun Zhu, Yike Guo, and Wen Gao. Ai alignment: A comprehensive survey, 2024.