Looking under the hood: How LLMs attempt political persuasion and microtargeting

Alex Lyman *Computer Science, Brigham Young University*
Ethan C. Busby *Political Science, Brigham Young University*
Lisa P. Argyle *Political Science, Purdue University*
Joshua R. Gubler *Political Science, Brigham Young University*
Bryce Hepner *Computer Science, Brigham Young University*
David Wingate *Computer Science, Brigham Young University*

Abstract: Significant academic and public attention has been paid to how generative AI tools can be used in political persuasion and microtargeting. A growing body of recent research finds that in many cases, effort to customize political messaging using an LLM yields little persuasive benefit. In this project, we explore the way LLMs construct persuasive messages, how such statements vary when the LLM is asked to microtarget individuals, and the degree to which the changes induced by mcicrotargeting lead to increased persuasion of human readers. We find variation in the degree to which different LLMs successfully comply with the microtargeting task. Furthermore, even for LLMs that produce more distinct messaging, the strategies are rarely systematically aligned with users' background features and do not increase the persuasiveness of the message in the aggregate. We discuss the implications of these findings for research on persuasion and generative AI.

Keywords: artificial intelligence, alignment, large language models, silicon sampling, computational social science

Conflict of interest: The authors declare no conflicts of interest.

Introduction

Generative large language models (LLMs) have great potential for social and political science. The potential range of social science applications of LLMs includes the capacity to generate research questions, ideas, and hypotheses (Meincke et al., 2024; Si, Yang and Hashimoto, 2024; Manning, Zhu and Horton, 2024; Wang et al., 2025; Qin, 2024); build on theories from a variety of social science disciplines (Xu, 2023; Kirkeby-Hinrup and Stenseke, 2025; Ke et al., 2025; Argyle, Busby, Gubler, Hepner, Lyman and Wingate, 2025); assist in coding (Nejjar et al., 2025); classify text and image data (Ornstein, Blasingame and Truscott, 2023; Törnberg, 2024; Heseltine and Clemm von Hohenberg, 2024; Wang, 2024; Timoneda and Vallejo Vera, 2026; Li and Jiang, 2026; Wu, 2024); act as interventions in experiments (Velez and Liu, 2025; Argyle et al., 2023; Tessler et al., 2024; Hackenburg and Margetts, 2024); simulate the attitudes and behavior of people and social networks (Argyle et al., 2023a; Dillion et al., 2023; Park et al., 2023; Chengxing Xie, 2024; Sreedhar et al., 2025; Bojić et al., 2025; Wang et al., 2025); and represent treatment effects in experiments (Dillion et al., 2023; Hewitt et al., 2024; Manning, Zhu and Horton, 2024; Cui, Li and Zhou, 2024).[1]

One growing application of LLMs in political science and political communication is in the study of persuasion. This topic is both a key area for science and central to the workings of democracy (Mutz, Sniderman and Brody, 1996). It is also a source of significant concern as political elites and malicious actors can use persuasive strategies to manipulate mass opinion for their own political purposes (Garsten, 2006; Hamilton, 1787; Druckman and Jacobs, 2015). Work focusing on generative AI and political persuasion indicates that LLMs can generate messages that are at least as persuasive as human-created messages

[1]These applications are separate from research on the governance of AI, an important area of study in computer and political science (Gasser and Almeida, 2017; Minkkinen and Mäntymäki, 2023; Lu et al., 2024; Woods, 2025; Batool, Zowghi and Bano, 2025)

(Bai et al., 2023; Palmer and Spirling, 2023; Hackenburg and Margetts, 2024; Matz et al., 2024), although perhaps with some penalties when people are aware the messages are AI-authored (Palmer and Spirling, 2023). Others have shown how LLMs offer important ways to explore core theoretical concepts in the study of persuasion, such as customization and elaboration (Argyle, Busby, Gubler, Lyman, Olcott, Pond and Wingate, 2025), the way different features of LLMs (such as their size in number of parameters) influence their ability to be persuasive (Hackenburg, Tappin, Röttger, and Margetts, 2025; Hackenburg et al., 2025), and how LLMs perform when given different persuasive tactics and post-training (Hackenburg et al., 2025).

The proficiency of LLMs at persuasion tasks raises additional normative concern for democratic institutions and societies (Summerfield et al., 2025). The concern is two-fold, based on potential misinformation and manipulation. On the one hand, LLMs can easily produce believable misinformation, whether as an incidental byproduct of their training (hallucinations), or because directed to by actors interested in propagating disinformation. In research analysis, both factual and artifactual claims in AI-generated texts make the output more persuasive (Hackenburg et al., 2025). Further concern about the potential negative impact of AI on public opinion arises from the potential for manipulation due to the ability of LLMs to produce personally targeted messaging at scale (Goldstein and Sastry, 2023; Simchom, Edwards and Lewandowsky, 2024; Goldstein et al., 2024). In the worst case scenario, individualized messaging at scale could exacerbate echo chambers and stymie critical thinking. The specter of LLMs as infinitely flexible and adaptive misinformation and manipulation tools poses daunting challenges to democratic societies around the world.

In this paper, we add to the existing work on political persuasion by LLMs by considering exactly what LLMs do when asked to target or customize persuasive appeals to citizens. Understanding the details of how LLMs approach a micro-targeting task can improve both scholarly and policy-making assessment of the risks, benefits, and potential

impacts of LLM-produced persuasive text on democratic outcomes. We begin by discussing the concept of customization from the social sciences broadly and possible ways that LLMs might customize persuasive appeals to individual people. We then provide two empirical studies evaluating customization and persuasiveness in LLM output.

In our first empirical study, we re-analyze data collected in published work on generative AI and persuasion (Argyle, Busby, Gubler, Lyman, Olcott, Pond and Wingate, 2025) and compare customized to uncustomized messages created by one specific LLM (OpenAI's GPT-4)(OpenAI, 2023). We consider both how customized and generic messages differ along a range of semantic and content features and how they elicit different responses from persuadees. In this analysis, we find that customized messages differ systematically from non-customized messages, but that these differences do not correspond with increases or decreases in how persuasive the messages are to their human targets. We then expand this analysis in our second empirical study, where we evaluate customized and non-customized messages generated by GPT-4 (from Study 1), GPT-5 (OpenAI)(OpenAI, 2025), Gemma 3 27b (Google)(Team, 2025), and Qwen3-Next-80b (Alibaba Cloud)(QwenTeam, 2025). We consider the same types of variables as in study 1 and present conclusions about customization that go beyond a single model.

We conclude with a discussion of the implications of these findings for democracy, theories of persuasion, and using LLMs as a tool to tailor messages, survey questions, and experiments to specific individuals (Argyle et al., 2023; Velez and Liu, 2024; Argyle, Busby, Gubler, Lyman, Olcott, Pond and Wingate, 2025). Specifically, we find that LLMs are at a baseline quite persuasive, but that they vary in their steerability — meaning the models demonstrate different levels of capacity to provide systematically different content across topics, directions, and targets. However, even when microtargeting induces relatively large content differences targeted to specific individuals, exhibiting the ability scholars and policy-makers have feared might lead to unusually high persuasive power, these changes in content or rhetorical strategy *are not correlated with systematically higher levels of*

persuasion in the human targets. On average, messages without any microtargeting are just as persuasive.

We note that our prompts instruct the models to microtarget using common sociodemographic characteristics, perhaps the most common approach to microtargeting. There might be more effective ways to microtarget using LLMs that need further exploration. Still, as we discuss further in our conclusion, our results raise questions about many of the assumptions underpinning arguments about microtargeting by suggesting that targeting individual characteristics does not necessarily lead to more persuasive power. More generally, while our results do not mitigate the need to be vigilant about future developments and deployments of AI messaging, they suggest that the scholarly and policymaking communities have at least some time to develop appropriate bulwarks against the potential negative future implications of AI microtargeting.

Persuasion and Customization

Persuasion involves changing or influencing the mental state of another, where the target of persuasion has at least some choice or agency (O'Keefe, 2016). Empirical and theoretical work on political persuasion emphasizes a variety of factors that are likely to make an appeal more successful. These include the source of a message or argument (Nicholson, 2012; Kam, 2020), the emotions evoked by a persuasive appeal (Brader, 2005; Albertson, Dun and Kushner Gadarian, 2020; Van Kleef, van den Berg and Heerdink, 2015), rhetorical strategies such as moral re-framing (Voelkel and Feinberg, 2018; Kalla, Levine and Broockman, 2022; Feinberg and Willer, 2019) or storytelling (Krause and Rucker, 2019; Kalla and Broockman, 2020; Green and Brock, 2000), the strength of a persuasive statement (Aarøe, 2011; Petersen and Arceneaux, 2020), the degree of elaboration on the part of the person being persuaded (Petty and Cacioppo, 1986; O'Keefe, 2013; Susmann et al., 2021; Wagner and Petty, 2022), and the amount of information provided in the persuasive appeal (Petty and Cacioppo, 1986; Coppock, 2023; Hackenburg et al., 2025; Sides, Vavreck

and Warshaw, 2022).

Many perspectives on persuasion posit that customized messages — appeals that adapt arguments or rhetorical approaches to the characteristics, values, beliefs, or personality of the target of persuasion — will be more successful than more generic arguments (Hirsh, Kang and Bodenhausen, 2012; Tappin et al., 2023; Dijkstra, 2008). This is because customization is thought to lead to messages that draw on concepts and experiences more familiar to participants than more generic messages. This should make customized message more likely generate the type of affect necessary for persuasion. In politics, customization is often referred to as "microtargeting"[2] and is widespread (Hersh, 2015; Votta et al., 2024).

One type of customization relies on matching messages to the sociodemographic characteristics of the target of persuasion. This approach assumes that groups of people — as defined by a set of social, political, and demographic characteristics - will react differently to persuasive appeals with content more directly related to their lived experience. Customization allows appeals to adapt messages based on group-level differences. Due to the promise of this idea, this type of messaging and campaigning has long been a part of political campaigns and has persisted (if not increased) in the time of digital media (Votta et al., 2024; Bär et al., 2024).

However, there are a number of reasons to be skeptical that this kind of customization provides consistent persuasive benefits. Many have noted that the effects of persuasive efforts are often remarkably homogeneous across groups (Hersch and Schaffner, 2013; Coppock, Hill and Vavreck, 2020; Coppock, 2023). Others find that persuasive efforts

[2]We make no distinction between the terms customization and microtageting, only to note that customization is used more broadly outside of political science. Some distinguish targeted from microtageting by the number of criteria used to select and send messages (Votta et al., 2024). From that perspective, we examine microtageting here as multiple characteristics of the targets of persuasion are used in combination.

may be most effective when they use fewer and less complex criteria for customization (Tanusondjaja et al., 2023; Tappin et al., 2023), raising questions about the processes behind this element of persuasion. Customization may also work better when based on attitudinal or psychological traits, such as moral foundations or Big 5 Personality traits, than based on more coarse demographic categories (Matz et al., 2024; Nezhad, Kisomi and Gholinezhad, 2025; Timm, Talele and Haimes, 2025). Some who have specifically explored customization from LLMs have noted that targeting messages to respondents' characteristics neither diminishes *nor* improves the persuasiveness of LLM-generated appeals (Hackenburg and Margetts, 2024; Argyle, Busby, Gubler, Lyman, Olcott, Pond and Wingate, 2025; Hackenburg et al., 2025; Lin et al., 2025), although other research finds people report a preference for messages that align with their predispositions (Matz et al., 2024).

We build on this latter body of research to consider exactly what LLMs do (or do not) do when asked to customize persuasive appeals to people based on their sociodemographic characteristics. In addition, we consider whether certain customization strategies used by LLMs are more persuasive than others. This will help researchers understand the ways LLMs approach customization and the possible consequences of increasing LLM use by political actors and campaigns across society. Unlike humans, LLMs lack an intrinsic intent to persuade, but are optimized through a variety of procedures to respond to users' instructions (Ouyang et al., 2022). Consistent with other work on persuasion (Durmus et al., 2024; Hackenburg and Margetts, 2024; Schoenegger et al., 2025; Palmer and Spirling, 2023; Hackenburg et al., 2025; Lin et al., 2025), we begin with the idea that humans can provide persuasive goals to LLMs through their prompts and instructions to the models, and that models are programmed to complete these tasks to the best of their ability. Thus, while this differs from human's intrinsic intent to persuade, models can adopt this intent through their prompting.

We consider a range of things that the generative AI tools might vary when asked

to customize a persuasive message. These come from established work on persuasion broadly and include the following:

- **Persuasive strategies:** Persuasive actors have a range of strategies to choose from, including storytelling (Krause and Rucker, 2019; Kalla and Broockman, 2020; Green and Brock, 2000), personal narratives (Naunov, Rueda-Cañòn and Ryan, 2025), moral reframing (Voelkel and Feinberg, 2018; Kalla, Levine and Broockman, 2022; Feinberg and Willer, 2019), debate-based strategies (Blumenau and Lauderdale, 2024), perspective-taking or active listening (Kalla and Broockman, 2016), referencing norms or the beliefs of others (Cialdini, Kallgren and Reno, 1991; Gerber, Green and Larimer, 2008), and more. Existing research on customization does not provide complete or systematic evidence that some of these strategies are consistently more effective for some demographic groups. We consider whether LLMs use some of these types of rhetorical strategies more than others when asked to customize persuasive appeals.

- **Emotional tone:** Other work evaluates the degree to which persuasion intersects with emotions in different ways, finding that the emotions evoked by an appeal are related to its effectiveness (Brader, 2005; Albertson, Dun and Kushner Gadarian, 2020; Van Kleef, van den Berg and Heerdink, 2015). Some work, outside the LLM context, has found that targeted appeals do not differ in their overall negativity from more general appeals (Ortega, 2022), but this has not been connected to other notions of emotions or linked to LLMs' efforts to customize. Accordingly, we consider if LLMs evoke emotions in different ways when asked to customize persuasive statements.

- **Informational content and density:** A key element for some existing work on persuasion is the information present in the appeal, both in terms of density and diversity of appeals (Coppock, 2023; Hackenburg et al., 2025; Lin et al., 2025). Generally, more information-dense appeals (i.e., persuasion that involves more arguments

or fact-like statements) appear to be more successful.[3] Some have also noted that customized messages tend to be more diverse in their content than more generic appeals (Ortega, 2022); this diversity may in turn result in more persuasiveness. Given this work, we consider if customization by LLMs changes information diversity, density, or both.

- **Values-based appeals:** As mentioned, one underlying element of the logic of customization is that it works by tailoring the structure and content of an appeal to respondents' experiences and perspectives. One way that this might occur is through appealing to different values, or goals that people have that transcend specific situations, serve as larger guiding principles, and vary in their importance (Schwartz, 1992; Schwartz et al., 2012). Existing studies of political communication have found that persuasiveness can be influenced by the correspondence between the values in a persuasive message and the values held by the target of persuasion (Nelson and Garst, 2005; Barker, 2005; Gordon and Miller, 2004), further indicating possible connections between values and customization.

- **Obvious targeting/pandering:** One thing that LLMs might do when asked to customize their messages is obvious targeting of specific respondent characteristics. Others have called this group-centric "pandering" and noted that it has contingent and sometimes ironic effects (Hersch and Schaffner, 2013). Nonetheless, LLMs may make direct or clearly inferrable references to specific groups or attitudes respondents have when asked to customize persuasive messages.

- **Structural features of the message:** A final element we consider are the structural features of the message itself - elements like length and textual complexity. There

[3]Some have found a more complex relationship between arguments/facts and persuasion, concluding that narratives and perspective taking are more consistently persuasive than facts (Naunov, Rueda-Cañòn and Ryan, 2025).

are a number of ways this might relate to persuasiveness, for example one way to customize a message for someone with less education might be to reduce the text complexity. However, because there is scant work connecting these kinds of features to customization directly we consider these elements of persuasive appeals without strong expectations from existing research.

We look for changes in these elements in persuasive messages generated by LLMs that do and do not attempt to customize those messages to specific respondents. In the section below, we explain the data we use to consider these factors and how we use these data to explore persuasion and customization. In doing so, we use an openly inductive approach - while previous research makes suggestions about what elements to consider (as discussed above), this work does not indicate what LLMs may or may not do in this context. As such, we explore the data presented below, hoping to gain insights for other studies and theories about persuasion, customization, and generative AI.

Data and Methods

We use two different empirical studies to examine customization and persuasion by LLMs. The first evaluates what one LLM (GPT-4) does when asked to customize persuasive appeals to specific individuals, how those individuals perceive customized and generic messages, and how the messages from the LLM correspond with actual persuasion. The second builds on these findings and considers three other, recently released LLMs - one closed-weight and two open-weight - to expand our conclusions beyond a single LLM. We describe the data and approach we use for each below:

Study 1:

In the first study, we re-analyze data collected in the United States from a recently published paper on persuasion and generative AI (Argyle, Busby, Gubler, Lyman, Olcott,

Pond and Wingate, 2025). As part of this project, the authors prompted OpenAI's GPT-4 model to craft persuasive messages under different conditions. Two of these conditions — which we re-analyze here — had the LLM generate paragraph-long persuasive statements for respondents that worked to persuade respondents away from their initial views on a topic. For a randomly selected half of these respondents, the LLM was provided with a description of the respondent and told to customize its appeal to that person.[4] This study also involved two different topics, with one data collection focusing on whether the United States should increase the amount of legal immigration (hereafter, "immigration"), and the other on whether the federal government should increase regulation related to K-12 teachers sharing their personal social and political views in the classroom (hereafter, "K-12 education"). The original study was conducted in May of 2024; Table 1 lists the number of respondents (and therefore LLM-generated appeals) for the different conditions in these data.

	Immigration	K-12 education	**Totals**
Generic	306	302	**608**
Customized	313	300	**613**
Totals	**619**	**602**	**1221**

Table 1: Count of messages by type and topic

In general, LLM output is strongly and directly influenced by the prompt used. The task of microtargeting is very broad and can encompass a variety of approaches or customization criteria; in this case, the LLM was asked to target messages on the following

[4]There were four other conditions in the original study that we do not consider here as they differ in their structure and objective. This included a dynamic conversation, a more introspective, motivational-interviewing conversation, a dynamic control discussion about board games, and a single static control message about board games. See the original study for more details about these conditions (Argyle, Busby, Gubler, Lyman, Olcott, Pond and Wingate, 2025).

variables: age, religion, marital status, education, occupation, location, political party, and political ideology. The model was then presented with the topic on which it was to persuade the respondent, followed by the respondent's demographics. It was also given the respondents' initial view on that topic and told to persuade the person away from that position. Then, the model was given a reasoning template that instructed the model to think about the respondent and construct a persuasive message that is "as micro-targeted as possible" **without** explicitly referencing the respondent's demographics. The exact prompts and other technical details can be found in online Appendix A.

We analyze these messages for the elements discussed in the previous section. In doing so, we consider patterns across the whole sample (all 1221 messages) and within each topic (N=619 for immigration and N=602 for K-12 education). We also make corrections for multiple comparisons in these tests (using a Bonferroni comparison), given that our analyses involve more than 100 statistical comparisons.

The following list describes the features of these messages and how those elements were analyzed. All of the persuasive strategies, emotional appeals, and fear-based appeals in the list were identified using an LLM (GPT-4o)(OpenAI, 2024). This process involved randomly sampling 50 messages from each of the combinations of topic (immigration or K-12 education) and argument direction (in favor of more or less regulation of those topics). This gave us two balanced sets of 100 messages, one for each topic. For each set of 100, we used GPT-4o to read the messages and extract the persuasive strategies used across the messages, with specific emphasis on strategies that were present in some but not all messages (to ensure the strategies would discriminate between messages.)[5] Because the model was tasked with finding strategies independent of message content, the lists were very similar across topic and direction. We used a reasoning model, R1 1776 (AI, 2025), to distill the two lists into a single, final list of 11 strategies. We describe the relevant strategies from this list below.

[5]The text of this and all prompts used in this study can be found in the appendix.

- **Persuasive strategies:** In comparing these messages, we evaluate a range of persuasive strategies, including:
  - *Storytelling:* This involves the use anecdotes (e.g., immigrant success stories) and efforts to humanize issues and foster emotional investment.
  - *Refutation of counter-arguments:* This strategy acknowledges opposing views (e.g., "Some say diversity divides, but...") to preempt criticism. It aims to strengthen persuasiveness by demonstrating thoroughness and disarming potential counter-arguments.
  - *Use of rhetorical questions:* This tactic provokes reflection ("Shouldn't we value diverse viewpoints?") to nudge audiences toward desired conclusions. It also engages readers by prompting active thought.
  - *Analogies or metaphors:* This approach compares complex ideas to relatable concepts (e.g., classrooms as "marketplaces of ideas," immigration policies as "gardens needing balance"). It attempts to clarify abstract arguments through vivid imagery.
  - *Credibility statements:* This tactic aligns arguments with trusted sources (historical precedents, democratic ideals, expert data) in an attempt to bolster trustworthiness and authority.
  - *Hypothetical scenarios:* Messages using this tactic paint vivid "what if" futures (e.g., societal collapse vs. utopia due to immigration levels). One of the goals is to leverage imagination to amplify stakes and outcomes.
  - *Logical arguments:* This strategy uses cause-effect arguments (e.g., "immigration drives innovation"), problem-solution, or cost-benefit analysis. It also supports arguments with structured reasoning and empirical/logical consistency (logos).

  - *Pragmatic appeals:* Messages employing this approach use centrist rhetoric to lightly push the reader. They frame issues as pragmatic, realistic, or grounded in the real world.

- **Emotional tone:** We also assess the messages for emotion in different ways, including:

  - *Emotional appeals:* This tactic evokes empathy (immigrant stories), pride ("American Dream"), or fear (resource strain). It creates personal connections to abstract policies.

  - *Fear-based appeals:* This strategy highlights risks (indoctrination, cultural erosion) to motivate caution and is often paired with urgency ("Without action, we lose...").

  - *Sentiment:* Sentiment was evaluated using NLTK's VADER package (Valence Aware Dictionary and sEntiment Reasoner) Hutto and Gilbert (2014), a rule-based sentiment analysis tool. This package generates positive, negative, and neutral sentiment scores for each passage of text, which can be aggregated into a composite score.

- **Informational content and density:** We use multiple measures of informational content and density as we compare these messages. This includes the following:

  - *Semantic embeddings:* We consider differences in the substance of messages in the different cells of Table 1 through semantic embeddings. Specifically, we use Qwen3-Embedding-4b(Zhang et al., 2025) with an embedding dimension of 2560 to create the embeddings. In this space, messages (represented as vectors) that are closer have more similar meanings than messages that are more distant. For visualization and presentational purposes, we then projected these points into two dimensions using UMAP.

- *Number of topics:* We also used a LLM to extract the topics being used across the persuasive messages. Similar to the process used to identify persuasive strategies and appeals, we randomly sampled 100 immigration messages (50 messages in each direction), and used GPT-4o to extract and distill a list of 10 topics used across the messages. We repeated this process for the K-12 education topic. With these lists, we use GPT-4o to code each persuasive message, identifying whether each listed topic is present in the message. We sum the number of topics present in each message. Higher values on this variable indicate that the LLM brought up more distinct topics in the message.

- **Values-based appeals:** There are many competing notions of values that may be at play in persuasive appeals. Here we consider one prominent values system - moral foundations theory (Graham, Haidt and Nosek, 2009; Graham et al., 2013)- but we recognize that this is only one of a number of possible value sets we could examine. Further, there are a number of important concerns and debates about this particular approach to moral values (Suhler and Churchland, 2011; Haidt and Joseph, 2011; Gray and Keeney, 2015). Nonetheless, this approach to values is commonly used both by academics and many in the public sphere.

  In a similar process to the rest of the text annotation, GPT-4o was used to code each persuasive message for moral foundations. We prompted the annotator to evaluate on a three-point scale whether each moral foundation was

  - Not referred to
  - Referred to weakly/implicitly
  - Referred to strongly or multiple times

  The annotating model was provided with a list of the six moral foundations (authority, care, equality, loyalty, proportionality, and purity), as well as a description

and definition of each foundation. For each foundation, the model provided a justification and score for each foundation, and a subset of the model's justifications and scores were reviewed by a human reviewer to validate performance. This LLM coding was nonexclusive; in other words, a message could make reference to one, several, or all of these values.

- **Obvious targeting/pandering:** In the prompt used to generate persuasive messages, models were explicitly instructed not to reference the user's demographics in the message. Consequently, very few messages refer to the microtargeted variables. However, roughly 10 percent of messages contain an explicitly targeted analogy based on the user's occupation. (In one instance, while crafting a message to an IT manager, the model likens America's economy and immigration system to a well-engineered network requiring scalable capacity.) Once again, we used GPT-4o to identify microtargeted analogies, which were reviewed by a human annotator.

  While analogies based on occupation were the only instances of microtargeting that humans could discern, it is possible that the model performed some type of microtargeting discernible only to itself. There is evidence to support the idea that LLMs can generate text that is meaningful to the LLM, but whose meaning is opaque to others (Zolkowski et al., 2025; Roger and Greenblatt, 2023). To test this, we had GPT-4o, the same model that generated the persuasive messages, read each message and attempt to guess the target's demographics, given a list of options.

- **Structural features of the message:** We also consider two structural features of these messages. This includes:

  - *Message length:* We used different indicators of message length — the number of characters in the statement generated by the LLM, the number of words in the LLM message, and the amount of time the human respondent spent reading the message (measured in seconds).

- *Textual difficulty:* Textual difficulty was measured using the Flesch Reading Ease Flesch (1948), a simple metric based on average sentence length and number of syllables per word. We also consider Type-token ratio, or TTR, as an alternative measure. TTR is an indicator of lexical diversity that is the ratio of the number of unique words to the total number of words in a document or statement. Higher values of TTR indicates greater diversity in a text.

In addition to these strategies, and as mentioned in the prior section, we also considered the impact of generic and customized persuasive appeals on respondents' perceptions and attitudes. We consider four variables, all of which correspond to the participants' reported attitudes after exposure to the persuasive appeal.

- **Actual persuasion:** We evaluate the degree to which respondents moved in their attitudes related to the persuasive appeal from the LLM. These were derived by comparing respondents' pre and post-persuasive appeal attitudes on the topic of persuasion (immigration or K-12 education). We construct this measure in the same way reported in the original analysis of this experiment (Argyle, Busby, Gubler, Lyman, Olcott, Pond and Wingate, 2025); more positive values of this variable indicate more movement towards the persuasive appeal.

- **Perceived persuasion:** The original experiment also measured respondents' *perceptions* of persuasion. This involved a survey question after the persuasive appeal which stated "Earlier, you read some arguments that attempted to convince you that [persuasive position here]. How persuasive did you find those arguments?". We include this because some other research focuses on perceptions of persuasion (Simchom, Edwards and Lewandowsky, 2024) and because there is significant debate and disagreement about how much perceived and actual persuasion correspond (Dillard, Weber and Vail, 2007; O'Keefe, 2018, 2020).

- **Confidence:** As an alternative, attitude-based measure, we consider respondents'

self-reported confidence or certainty in their views following the persuasive appeals. This is in recognition that attitudes have many features, only one of which is the direction or specific position. Persuasion may have an impact not just on changing people's views but shifting their underlying confidence in or strength of those views (Crano and Prislin, 2006; Visser, Bizer and Krosnick, 2006; Tormala, 2016).

- **Source perceptions:** One key element in persuasion can be the source of a persuasive appeal (Hartman and Weber, 2009; Kam, 2020). This has become important in the domain of AI-empowered persuasion, as some work has found that the persuasive effects of AI-generated messages can be impacted by knowledge of its source (Palmer and Spirling, 2023; Lu, Tormala and Duhachek, 2025). We therefore consider the degree to which respondents' perceived the persuasive messages as coming from an AI source. We assessed this using a question near the end of the survey, asking respondents to guess the source of the persuasive appeal; one of these options was "A chatbot (like ChatGPT)".

Study 2:

To expand the analysis in Study 1, we recreated the same messages from those data with three other LLMs that were released more recently than GPT-4: GPT-5, Gemma 3, and Qwen3. Two of these - Gemma 3 and Qwen3 - are open-weight and one — GPT-5 — is closed-weight. To generate the messages, we provided these LLMs with the same respondent backgrounds (in the customized conditions) and instructions and prompted them to create the corresponding persuasive appeals. As such, we have a set of 1221 messages for each of these models, with the same breakdown as shown in Table 1.

To generate these new messages we use exactly the same prompts deployed with GPT-4 in Study 1 with each of these three models. There are likely gains to be made by tailoring prompts to each of these models, as different LLMs often react differently to the same instructions (Argyle, Busby, Gubler, Hepner, Lyman and Wingate, 2025; Zhuo et al., 2024).

However, if we were to do this, we would also run the risk of distilling different definitions of customization to each of these LLMs and generating different kinds of empirical patterns due to unintentional changes in how we conceptualize customization. To avoid this, we use identical prompts for all of these models, acknowledging that this comes with the tradeoff of failing to optimize prompt performance for each of these different LLMs.

We then analyzed these persuasive messages using the same variables that we discussed for Study 1. The exceptions to this are the perception/human-derived variables; for Study 2, we do not have human assessments of the messages created by these three different LLMs. As such, we focus on the observable elements of the messages themselves and rely on Study 1 for conclusions about effects on persuasion and human perceptions.

Results

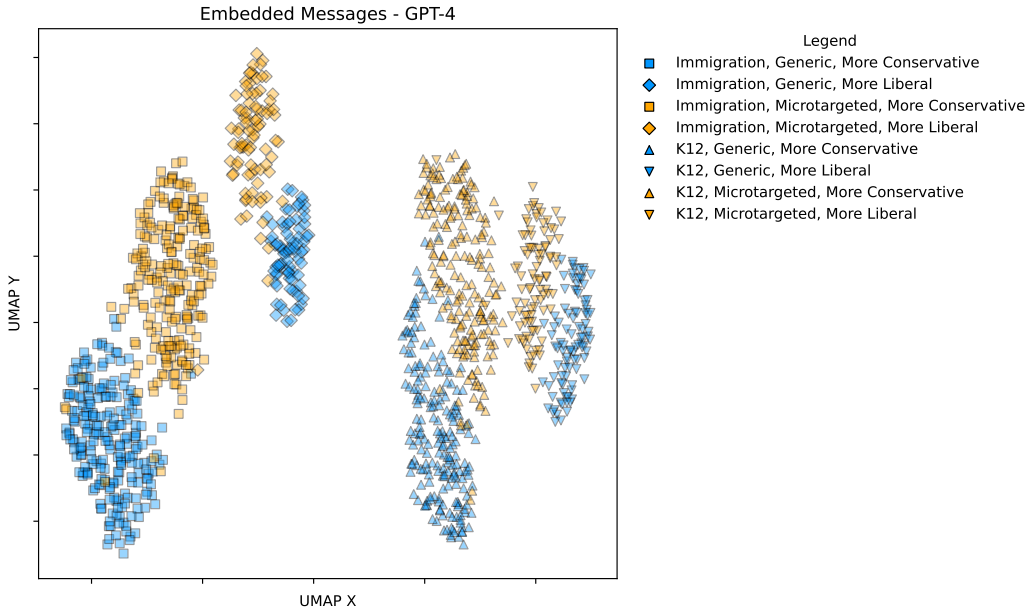
Study 1:

We begin by describing differences in the customized versus generic messages from GPT-4 used in the original study. We consider both analyses of the statements themselves and evaluations of how human targets of persuasion reacted differently to the two types of messages. In all the analyses that follow, we look at the aggregated results across topics and break out the patterns for the two different issue areas (immigration and K-12 education).

We begin by comparing the semantic embeddings for the customized and generic messages across both topics. Figure 1 presents these results.

These findings suggest that (1) there is clear separation based on the topic of the persuasive appeal (as we would hope) and (2) generic and customized messages differ from one another. Although these gaps are smaller than the gaps by topic, this analysis suggests that the generic and customized messages are semantically distinct from one another.

Figure 1: Embedded Messages



When we consider the length of the messages (see Figure 2), we see that customized or microtargeted messages are significantly longer than generic messages. At the same time, respondents do not spend more time *reading* these longer messages. The LLM generates longer statements, but not statements that result in more in-depth engagement from respondents. This may be explained by a general lack of a difference in the difficultly of reading the messages between the two message types (see Figure 3).

We next evaluate the different strategies used in the persuasive statements and information density; the corresponding results of comparisons across the customized and generic messages can be found in figures 4 and 5. Although not all strategies vary by the message type, these comparisons suggests some areas of difference between the two kinds of messages. Customized messages tend to include fewer preemptive refutations of counterarguments than generic messages, more analogies, more hypotheticals, and fewer logical strategies. Taken together, these suggest that the LLM draws from a different set of persuasive tactics when asked to customize its persuasive appeals. Figure 4 also includes estimates on the number of topics referenced in each appeal and speaks to information

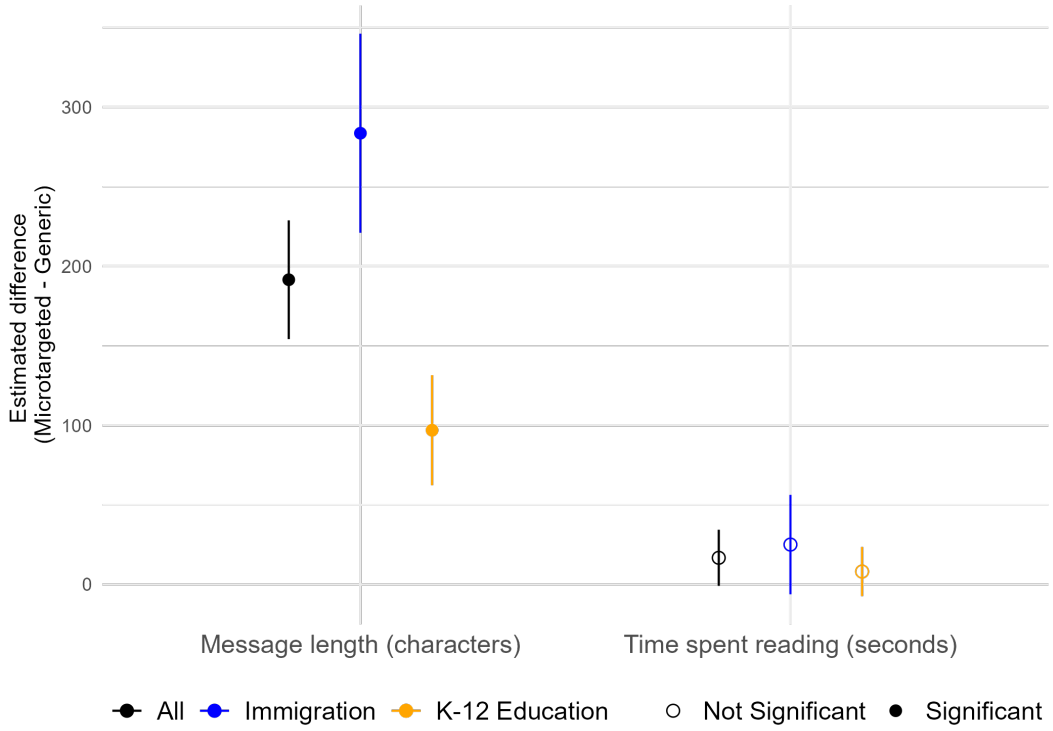


Figure 2: Length in words and time spent reading the different messages. The y-axis shows the differences between the microtargeted and generic messages; positive values indicate more of the variable in the microtargeted messages.

density. In this case, the customized messages reference fewer topics and are, by extension, lower in information density than the generic appeals.

We also observe differences in emotions and value-based appeals. With respect to affect and emotional appeals, we observe that customized messages are lower in negative sentiment than generic messages and make fewer emotional appeals (see figures 6 and 7. Figure 8 suggests that customized appeals make fewer references to care and loyalty values, despite the fact that these values are not explicitly referenced in any of the prompts given to the LLM. On other values, we do not observe differences across the message types.

We next consider how respondents reacted to both kinds of messages - this can be found in Figure 9. Despite the fact that we observe some differences in the strategies, length, emotions, and information density of the customized and generic messages, these

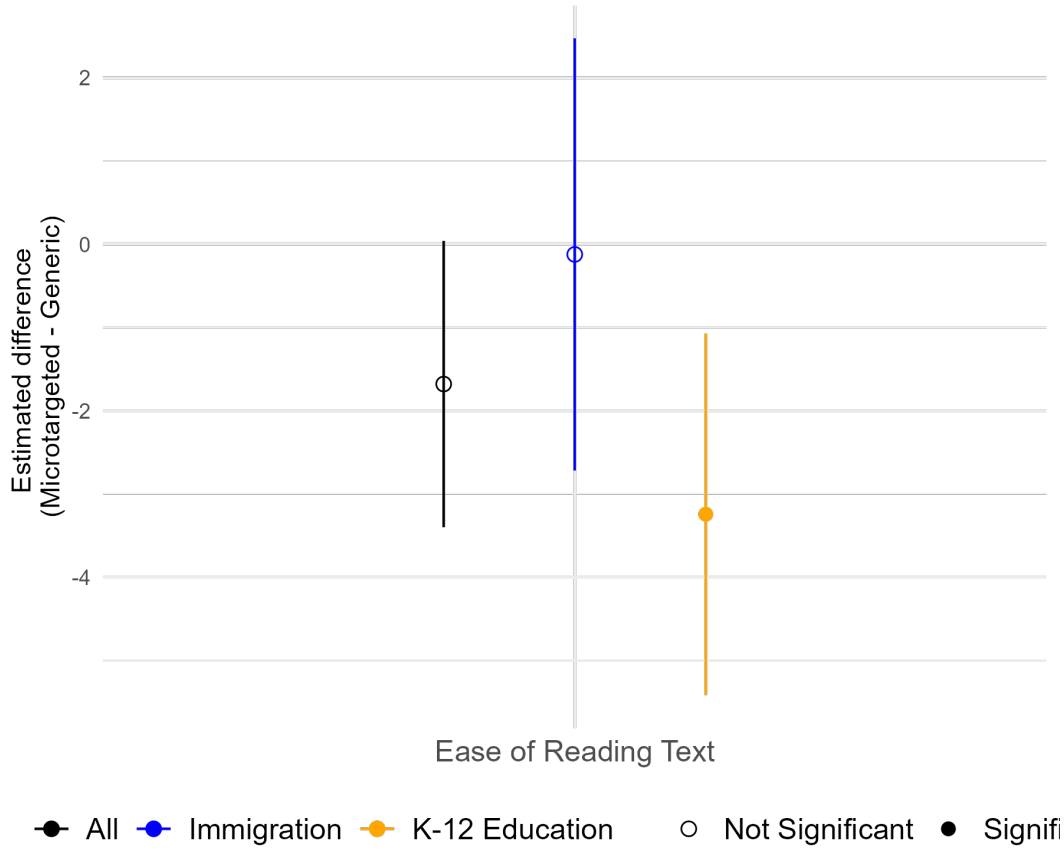


Figure 3: Textual ease of reading (Flesch Reading Ease) of the different messages. The y-axis shows the differences between the microtargeted and generic messages; positive values indicate more of the variable in the microtargeted messages. Higher values indicate easier texts to read.

do not seem to carry over to different perceptions of the messages by the human respondents. We see no differences in the perceived or actual persuasiveness of the messages (distinguished by self-reported persuasiveness and actual attitude change). The different messages also do not generate weaker or less confident attitudes, and customized messages do not provoke more concerns or awareness of the AI source of the appeals. Connecting this to the analyses of the other attributes of the messages suggests that while this LLM constructs messages differently when asked to customize, those differences do not result in measurable changes in human reactions to the persuasive appeals.[6]

[6]Given the possibility of heterogeneity in response to political persuasion, we also estimated the results in Figure 9 for ideologues (as measured by extreme ideological self-

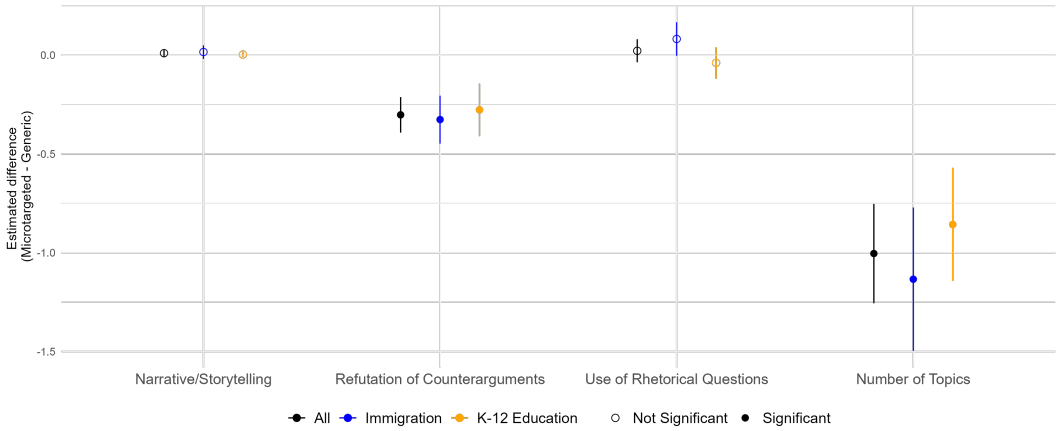


Figure 4: The figure shows different strategies in the different messages. The y-axis shows the differences between the microtargeted and generic messages; positive values indicate more of the variable in the microtargeted messages.

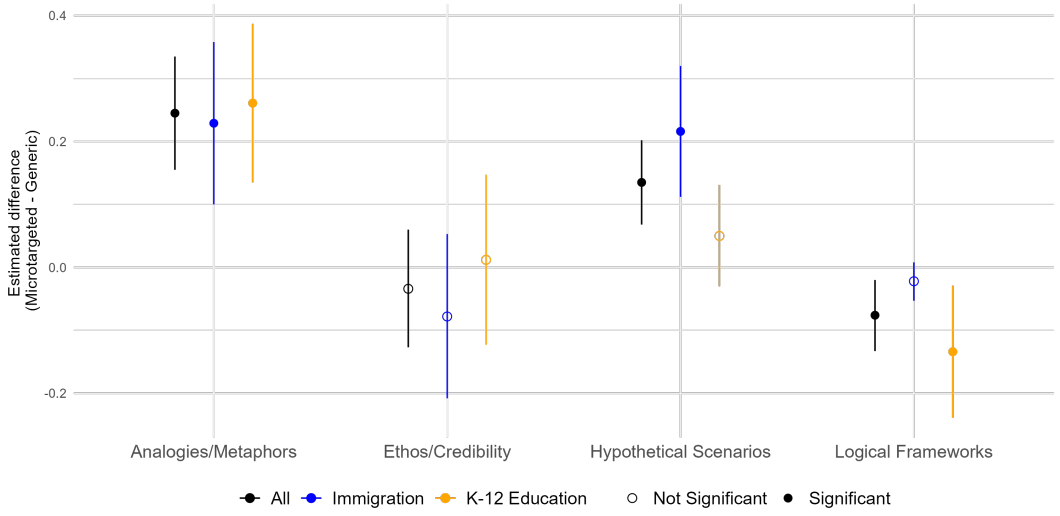


Figure 5: The figure shows different strategies in the different messages. The y-axis shows the differences between the microtargeted and generic messages; positive values indicate more of the variable in the microtargeted messages.

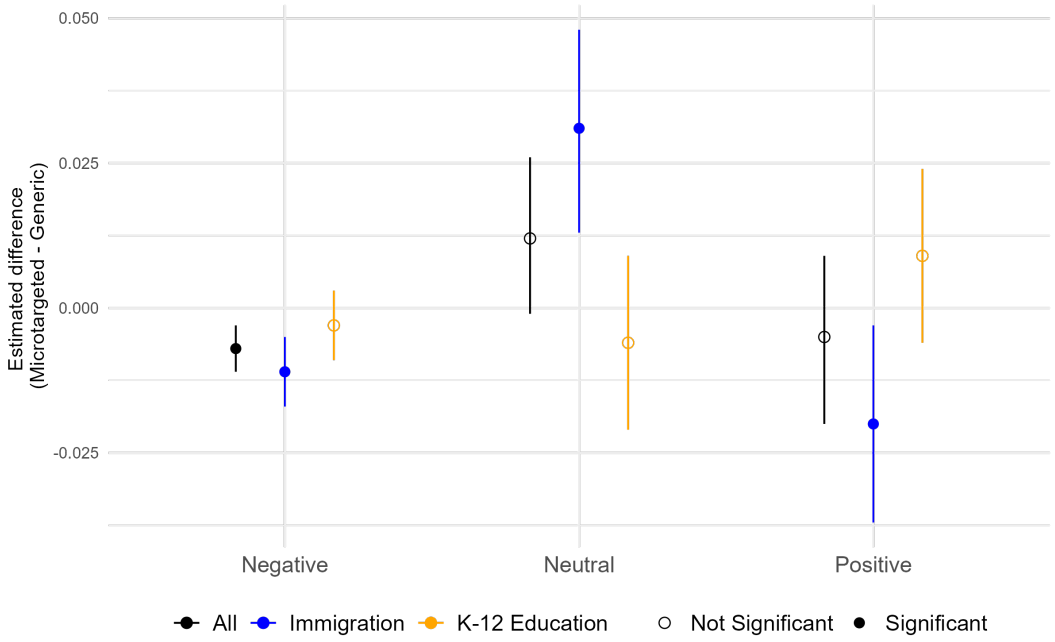


Figure 6: Sentiment in the different messages. The y-axis shows the differences between the microtargeted and generic messages; positive values indicate more of the variable in the microtargeted messages.

One possible explanation for the combination of these results — that the LLM is changing its tactics when microtargeting but that this does not lead to more persuasion or changes in perceptions — is that changes the LLM introduces do not correspond in systematic ways to the demographic groups involved in the targeting. In other words, the variations introduced by the LLM when it is asked to microtarget may not consistently be aligned with targets' different social and political characteristics.

We consider this possibility in different ways. First, we evaluate evidence of signals

placement), non-ideologues, those without a four year college degree, those with at least a four year college degree, those who reported being interested in politics, and those who reported not being interested in politics. Across all of these subgroups, we fail to observe statistically significant differences in reactions to the microtargeted and generic messages. This homogeneity is consistent with larger research about political persuasion (Coppock, Hill and Vavreck, 2020; Coppock, 2023).

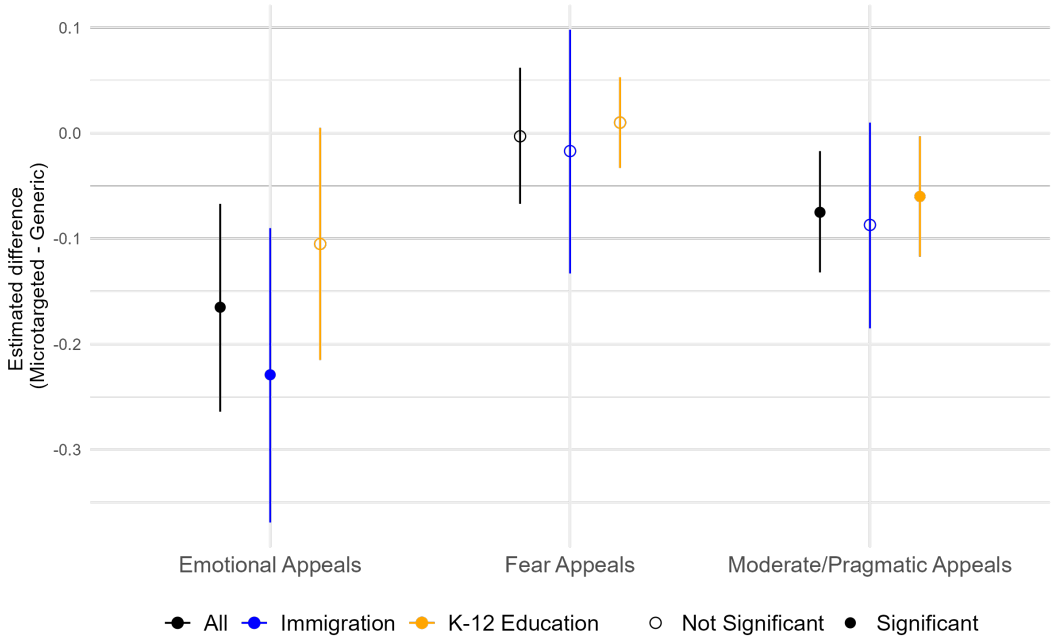


Figure 7: Various appeals in the different messages. The y-axis shows the differences between the microtargeted and generic messages; positive values indicate more of the variable in the microtargeted messages.

about the characteristics of the target that can be inferred from the text of the persuasive appeal (we call this obvious pandering in the preceding sections). Because it is possible that the model is customizing messages in an advanced or nuanced way that is difficult to detect by an outside observer, we ask whether a different instance of the same LLM, exposed only to a microtargeted or generic message, can infer any of the respondent's demographics from the message. We then compared these guesses to the respondents' actual characteristics and coded for correct and incorrect inferences.

When considering inferences from the microtargeted messages, we find that the inferences from the LLM (GPT-4, the same one used to create the messages) significantly underperform random chance. This provides additional evidence that systematic microtargeting isn't occurring in any discernible way. We also compare the rate of correct inferences in the generic and microtargeted messages; figure 10 shows the results of this analysis. Here again we see no clear evidence of signals about persuasive targets in the targeted messages. The exception to this is ideology, which seems to be more evident in the

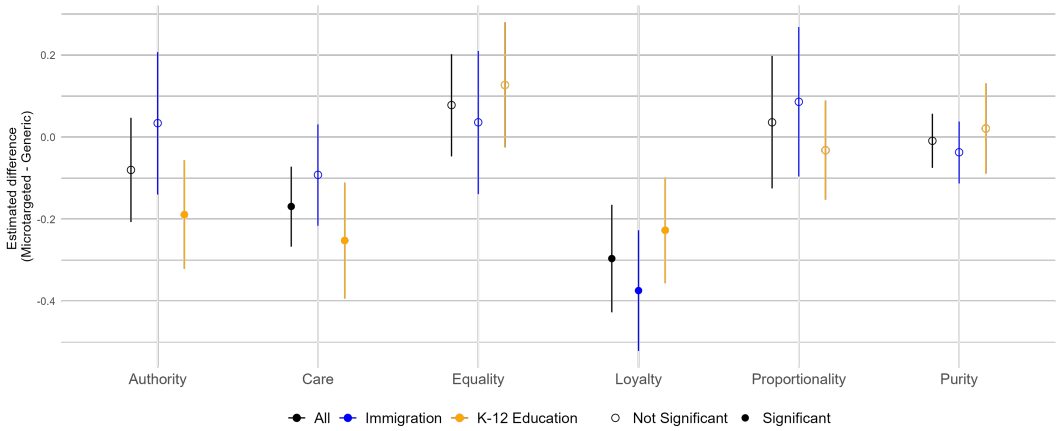


Figure 8: Moral foundation use in the different messages. The y-axis shows the differences between the microtargeted and generic messages; positive values indicate more of the variable in the microtargeted messages.

customized messages in the immigration topic. However, this seems, at best, weak and inconsistent evidence that detectable signals about respondents can be discerned from the messages. Moreover, we note that the model may be able to infer certain aspects of the user, simply given the direction of persuasion. For example, the model may be able to infer from prior knowledge about party platforms that a Republican is more likely to be the recipient of a message on decreasing defense spending.

We also considered the degree to which the changes in the rhetorical and persuasive strategies noted earlier correspond with the characteristics of the respondents. This would still be evidence of successful targeting - using metaphors or emotional appeals for men more than for women, for example, would indicate a coherent (even if unsuccessful) attempt to microtarget - even if it is less obvious or explicit than the signaling in Figure 10. As a test of this, we evaluate the correlations between different respondent characteristics and the strategies we document in the sections above.

One thing we consider in this analysis is that for all of our messages - generic and microtargeted - the LLM was provided with the respondents' initial position on the subject of persuasion. This was necessitated by the study design, which had the LLM persuade respondents away from their initial positions. At the same time, this might introduce

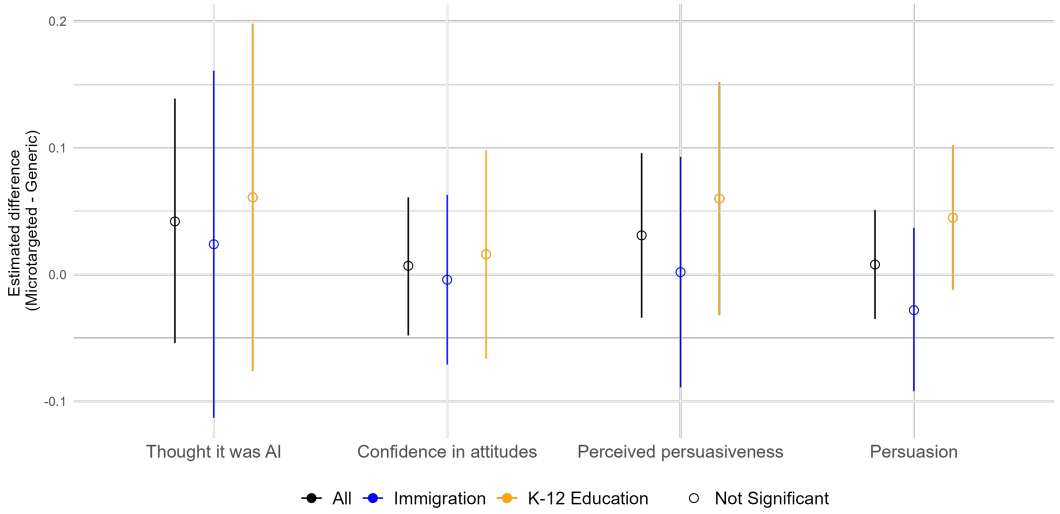


Figure 9: Subjective outcomes in the different messages. The y-axis shows the differences between the microtargeted and generic messages; positive values indicate more of the variable in the microtargeted messages.

some changes in persuasive tactic that may, in a second-order way, then correspond to respondents' characteristics. To account for this possibility, we compare the correlations between the tactics and respondent characteristics in the microtargeted *and* static conditions; specifically, we subtract the pairwise correlations from the generic conditions from the targeted conditions to estimate what increase or decrease in correlation between persuasive tactics and respondent characteristics is provided from the microtargeting instructions and information (as opposed to just knowing about respondents' initial positions on immigration or K-12 education). We would expect increases in the correlations if the LLM is using the microtargeting information to deploy different persuasive approaches to different groups of respondents.

We provide additional information on these results, and figures of these differences in correlations, in the appendix (their size and the number of variables being compared make them difficult to see in the main text). Our conclusion from these results is that the difference in correlation in the microtargeted messages is inconsistent, small, and more likely to be negative than positive. This suggests evidence against the claim that the LLM is targeting with different approaches to persuasion. In other words, microtarget-

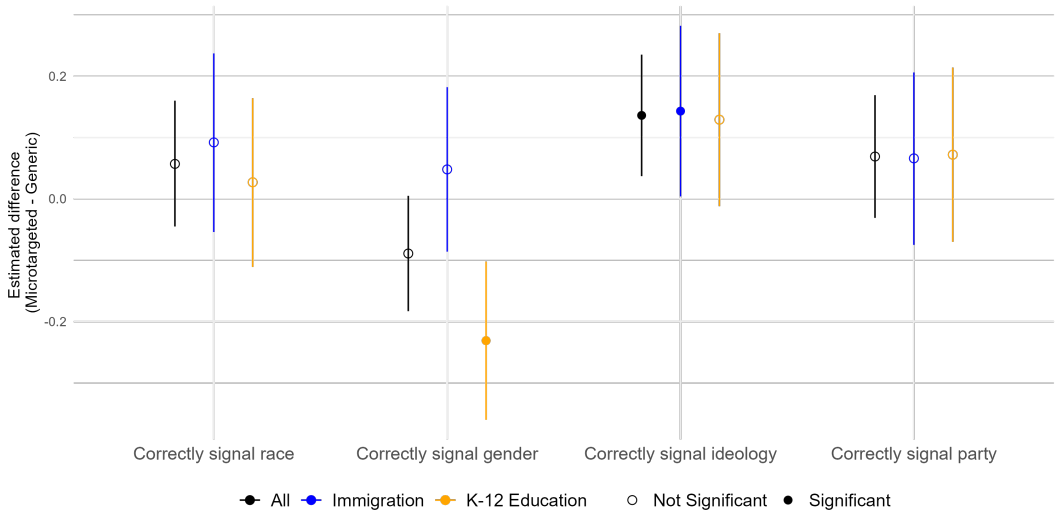


Figure 10: Signals found in the different messages. Correct signals occur when the coding LLM (GPT-4o) correctly guesses the characteristic based on the message. The y-axis shows the differences between the microtargeted and generic messages; positive values indicate more of the variable in the microtargeted messages.

ing strategies and approaches are generally not tied systematically to any of the target's demographics.

A final possibility is that the LLM is inconsistently effective, and some of the microtargeting tactics are more persuasive to the targets but this effect is lost in the aggregate results depicted in Figure 9. In order to test for this possibility, we conducted a series of OLS regressions predicting attitude change in the human respondents on the basis of each set of the measured text features. None of the features have a significant relationship with measured attitude change. Notably, however, several of the features have a positive impact on the respondent's subjective rating of whether the message was persuasive. [7] In other words, some of these features feel more persuasive to the reader, but they do not result in additional attitude change - a pattern that corresponds with existing results in the

[7] All results provided in the appendix. The features that had a positive effect on subjective ratings of persuasiveness include: Equality and Proportionality moral foundations, correct prediction by an LLM of ideology, party, and ethnicity; fear appeals and arguments about resource strain had negative impacts on persuasiveness.

literature (Hackenburg and Margetts, 2024; Matz et al., 2024). This suggests that while the LLM might be able to make changes that sound appealing on the surface, it is not fundamentally selecting arguments or strategies that result in more real-world attitude change on the part of the respondents.

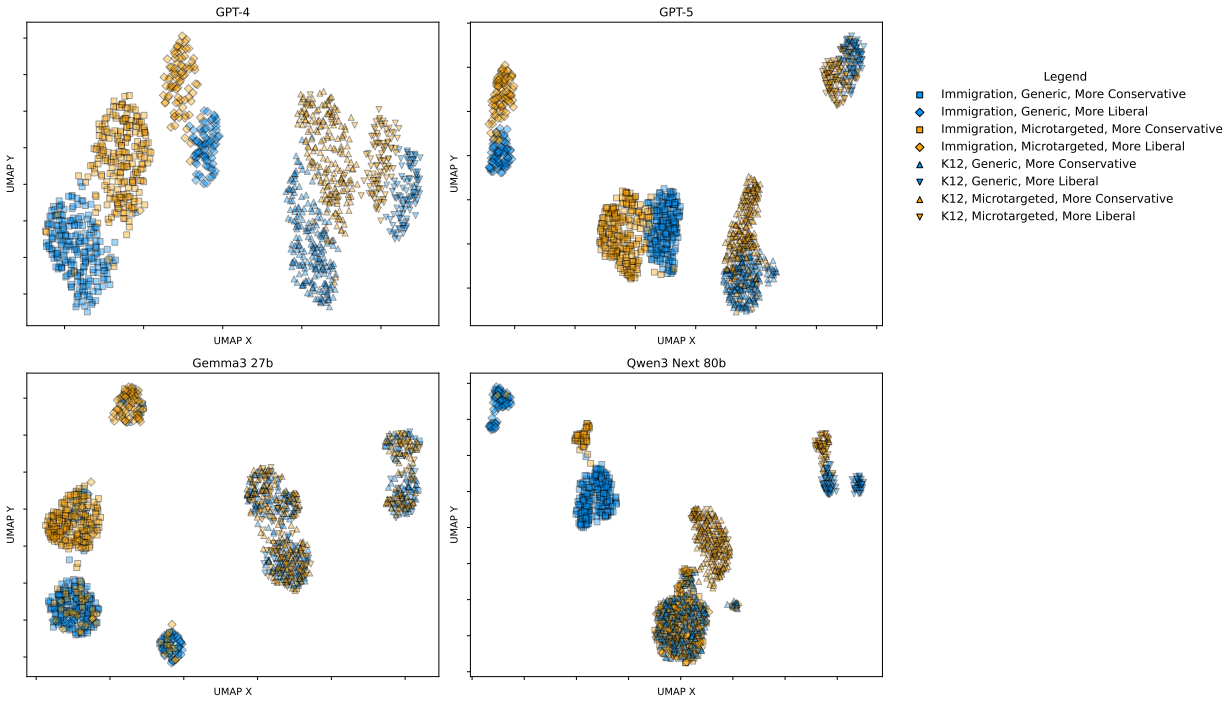
Study 2:

We next expand the findings from Study 1 to look beyond a single LLM and consider the customized and generic messages created by the three other models mentioned earlier (GPT-5, Gemma 3, and Qwen3). As mentioned earlier, the variables we analyze here are the objective characteristics of the messages we can evaluate, rather than the perceptions or responses to the messages by human participants. We summarize these analyses here and point readers to the appendix for additional details.

Across the models, we see a number of similarities that hint at patterns that may go beyond a single LLM. As an example, Figure 11 shows the semantic embeddings for all four models. The separation between the blue and yellow points in these graphs indicates that all four LLMs create customized messages that are different semantically from the generic messages. When it comes to length in characters, all of the LLMs generated longer messages when asked to customize their persuasive appeals (as compared to generic appeals). With respect to values, all of the models made fewer appeals to loyalty when customizing, especially on the topic of immigration. Three of the four models (all except Gemma 3) used more analogies and fewer topics (or lower information density) when constructing customized or microtargeted messages. We also observe small and typically statistically insignificant differences in the positive and negative sentiment of customized and generic messages. These patterns support the idea that LLMs generally employ different tactics and write messages differently when customizing persuasive appeals.

At the same time, some patterns seem model-specific. Referring back to Figure 11, the separation based on customization is stronger for GPT-4 and GPT-5 than for either

Figure 11: Embedded Conversations, All Models



of the other models. Additionally, Gemma 3 and GPT-4, use fewer appeals to emotions when microtargeting; the same is not true for GPT-5 or Qwen3. Qwen3 and GPT-4 made fewer references to logical frameworks when customizing, but this finding did not emerge for GPT-5 or Gemma 3. Customized messages by Gemma 3 and GPT-4 were harder to read than generic messages, but the opposite was true for Qwen3 and GPT-5. GPT-4 was the only LLM to use more hypothetical scenarios when customizing persuasive messages. Results along these lines suggest that the specific ways that LLMs change their customized appeals varies by the model being deployed. Combined with the similar patterns discussed previously, this suggests that LLMs do construct persuasive appeals differently when asked to target those appeals to specific people, but do not necessarily use the same tactics as other LLMs when customizing. Those interested in studying customization by LLMs will likely need to consider a range of LLMs to understand precisely how this occurs (rather than inferring from a specific LLM to a whole class of models).

Conclusion and Implications

The preceding analysis suggests two main patterns with implications for LLM persuasion. First, prompting an LLM to customize or microtarget a persuasive appeal to a specific individual pushes the model to create substantively different messages than when prompting generic persuasive statements. We see this across studies 1 and 2, although the precise way in which the messages changes can depend on the LLM being used. One possible explanation (among others) as to why LLMs are responsive to the instruction to microtarget has to do with how they are constructed and the process of Reinforcement Learning from Human Feedback (RLHF) that nearly all LLMS undergo, including each of the LLMs used in this study. RLHF is explicitly designed to provide "reinforcement signals tied to... success [however defined]" (Bai et al., 2022) and has been shown to demonstrably improve models' abilities in many areas (Ouyang et al., 2022; Stiennon et al., 2020), including goal-directedness, reasoning and persuasion (Pan et al., 2023; Lightman et al., 2023; Hackenburg et al., 2025). We believe this is part of the reason why these models do complete the narrow task assigned them: customizing messages in a variety of different ways.

The second pattern we observe is that these changes in the behavior of LLMs do not result in more persuasion or correlate in detectable ways with respondent characteristics (such as gender, race, and partisanship). These findings, combined with published work concluding that demographic-based microtargeting does not provide persuasive boosts (such as Argyle, Busby, Gubler, Lyman, Olcott, Pond and Wingate, 2025; Hackenburg and Margetts, 2024; Tappin et al., 2023), suggest that while LLMs are generally capable of successful persuasion, they at present have no special skill when it comes to crafting personalized messages based on demographic variables. This is true even for messages that have features that make users subjectively rate them as more persuasive.

In a more narrow way, our findings (primarily in Study 2) also add to the body of

work in computational social science that considers variation in performance across models. When considering four LLMs from different families (GPT-4, GPT-5, Gemma 3, and Qwen3), we find that LLMs vary in their response to microtargeting instructions. All models meet a baseline of steerability, meaning that they introduce changes to their normal output in response to the task they are asked to do. However, different models complete the task in different ways and to varying extents. Some only adjust a few characteristics of the message, while others change a more extensive set of features resulting in bigger semantic differences. When studying LLMs as a persuasive tool (and for other kinds of tasks), it remains important to evaluate different LLMs to establish larger claims about the capacity of this class of generative AI tools.

One interpretation of these results is to suggest that, although LLMs will instantly comply with the task, the vast array of potential information and arguments at the disposal of an LLM does not provide it with a master key to persuasive strategizing. While the LLMs produce systematically different messages, the messages produced are not more persuasive than generic messages, and individual features do not correspond systematically to more persuasion. The black box of an LLM does not seem to hold superhuman persuasive skill – for now, at least. At present, this helps to clarify the boundaries of the behavior of several recent and widely used LLMs.

There are a number of possible explanations for what we find here that should motivate future work in this area, prompting more discovery and productive research into this topic. Our results generalize to a specific context: LLM microtargeting attempts using basic sociodemographic data in the context of hot-button political issues. While we successfully replicate these results across a variety of different types of models and across two different political topics, future work should vary both model type and microtargeting type to further explore the bounds of these findings. For example, perhaps the reason we (and other authors) do not see big improvements from demographic microtargeting is because LLMs are not able to systematically and meaningfully target messages on the

basis of the types of features we have provided. They might, however, be able to target individuals more successfully if given different kinds of personal information or fewer constraints on how to employ that information in the text. For example, LLMs might become more persuasive if their efforts focused on detailed reasoning for people's specific beliefs rather than relatively generic features of partisanship and ideology. It is also possible that LLMs may perform differently in this area if asked to craft persuasive appeals in other, nonpolitical domains. Alternatively, perhaps these results point to a more fundamental theoretical insight suggesting that the psychological assumptions behind microtargeting need further questioning.

More broadly, these results point out present limits to the impact of hyper-focused LLM political targeting. On the one hand, the pace of development and change in this technological space suggests a continued need for vigilance and continued evaluation about the impact and processes of AI-empowered messaging and persuasion. On the other hand, our present understanding, including the results shown here, suggest that scholarly and policymaking communities have some space, albeit of an unknown duration, to develop bulwarks and barriers against more negative and potent applications of AI microtargeting.

References

Aarøe, Lene. 2011. "Investigating Frame Strength: The Case of Episodic and Thematic Frames." *Political Communication* 28(2):207–226.

AI, Perplexity. 2025. "R1 1776: An Uncensored Version of DeepSeek-R1." https://huggingface.co/perplexity-ai/r1-1776. Open-source language model post-trained to remove censorship constraints. Accessed June 2025.

Albertson, Bethany, Lindsay Dun and Shana Kushner Gadarian. 2020. The Emotional Aspects of Political Persuasion. In *The Oxford Handbook of Electoral Persuasion*, ed. Elizabeth Suhay, Bernard Grofman and Alexander H. Trechsel. New York: Oxford University Press pp. 169–183.

Argyle, Lisa P., Christopher A. Bail, Ethan C. Busby, Joshua R. Gubler, Thomas Howe, Christopher Rytting, Taylor Sorensen and David Wingate. 2023. "Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale." *Proceedings of the National Academy of the Sciences* 120(41).

Argyle, Lisa P, Ethan C Busby, Joshua R Gubler, Alex Lyman, Justin Olcott, Jackson Pond and David Wingate. 2025. "Testing Theories of Political Persuasion Using Artificial Intelligence." *Proceedings of the National Academy of the Sciences* 122(18):e2412815122.

Argyle, Lisa P., Ethan C. Busby, Joshua R. Gubler, Bryce Hepner, Alex Lyman and David Wingate. 2025. "Arti-'fickle'intelligence: using LLMs as a tool for inference in the political and social sciences." *Nature Computational Science* 5:737–744.

Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting and David Wingate. 2023a. "Out of One, Many: Using Language Models to Simulate Human Samples." *Political Analysis* 3(3):337–351.
URL: *https://doi.org/10.1371/journal.pclm.0000429*

Bai, Hui, Jan Voelkel, Johannes Eichstaedt and Robb Willer. 2023. "Artificial Intelligence Can Persuade Humans on Political Issues.".
URL: *https://doi.org/10.21203/rs.3.rs-3238396/v1*

Bai, Yuntao, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann and Jared Kaplan. 2022. "Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback.".
URL: *https://arxiv.org/abs/2204.05862*

Barker, David C. 2005. "Values, Frames, and Persuasion in Presidential Nomination Campaigns." *Political Behavior* 27(4):375–394.
URL: *http://link.springer.com/10.1007/s11109-005-8145-4*

Batool, Amna, Didar Zowghi and Muneera Bano. 2025. "AI governance: a systematic literature review." *AI and Ethics* 5(3):3265–3279.
URL: *https://doi.org/10.1007/s43681-024-00653-w*

Blumenau, Jack and Benjamin E. Lauderdale. 2024. "The Variable Persuasiveness of Political Rhetoric." *American Journal of Political Science* 68(1):255–270.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12703*

Bojić, Ljubiša, Velibor Ilić, Veljko Prodanović and Vuk Vuković. 2025. "An Agent-Based Simulation of Politicized Topics Using Large Language Models: Algorithmic Personalization and Polarization on Social Media." *Chinese Political Science Review* .

Brader, Ted. 2005. "Striking a responsive chord: How political ads motivate and persuade voters by appealing to emotions." *American Journal of Political Science* 49(2):388–405.

Bär, Dominik, Francesco Pierri, Gianmarco De Francisci Morales and Stefan Feuerriegel. 2024. "Systematic discrepancies in the delivery of political ads on Facebook and Instagram." *PNAS Nexus* .

Chengxing Xie, Canyu Chen, Feiran Jia Ziyu Ye Shiyang Lai Kai Shu Jindong Gu Adel Bibi Ziniu Hu David Jurgens James Evans Philip H.S. Torr Bernard Ghanem Guohao Li. 2024. Can Large Language Model Agents Simulate Human Trust Behavior? In *38th Conference on Neural Information Processing Systems*.

Cialdini, Robert B., Carl A. Kallgren and Raymond R. Reno. 1991. "A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior." *Advances in Experimental Social Psychology* 24:201–234.

Coppock, Alexander. 2023. *Persuasion in Parallel: How Information Changes Minds about Politics*. Chicago: Univeristy of Chicago Press.

Coppock, Alexander, Seth J. Hill and Lynn Vavreck. 2020. "The small effects of political advertising are small regardless of context, message, sender, or receiver: Evidence from 59 real-time randomized experiments." *Science Advances* 6:eabc4046.

Crano, William D. and Radmila Prislin. 2006. "Attitudes and Persuasion." *Annual Review of Psychology* 57(1):345–374.
URL: *https://www.annualreviews.org/doi/10.1146/annurev.psych.57.102904.190034*

Cui, Ziyan, Ning Li and Huaikang Zhou. 2024. "Can AI Replace Human Subjects? A Large-Scale Replication of Psychological Experiments with LLMs.".
URL: *https://arxiv.org/abs/2409.00128*

Dijkstra, Arie. 2008. "The psychology of tailoring-ingredients in computer-tailored persuasion." *Social and personality psychology compass* 2(2):765–784.

Dillard, James Price, Kirsten M. Weber and Renata G. Vail. 2007. "The relationship between the perceived and actual effectiveness of persuasive messages: A meta-analysis with implications for formative campaign research." *Journal of Communication* 57(4):613–631.

Dillion, Danica, Niket Tandon, Yuling Gu and 1Kurt Gray. 2023. "Can AI language models replace human participants?" *Trends in Cognitive Sciences* 27(7):597–600. Publisher: Elsevier.
URL: *https://doi.org/10.1016/j.tics.2023.04.008*

Druckman, James N. and Lawrence R. Jacobs. 2015. *Who Governs? Presidents, Public Opinion, and Manipulaton*. Chicago: University of Chicago Press.

Durmus, Esin, Liane Lovitt, Alex Tamkin, Stuart Ritchie, Jack Clark and Deep Ganguli. 2024. "Measuring the Persuasiveness of Language Models." https://www.anthropic.com/research/measuring-model-persuasiveness. Anthropic Research.

Feinberg, Matthew and Robb Willer. 2019. "Moral reframing: A technique for effective and persuasive communication across political divides." *Social and Personality Psychology Compass* 13(12).

Flesch, Rudolf. 1948. "A new readability yardstick." *Journal of Applied Psychology* 32(3):221–233.

Garsten, Bryan. 2006. *Saving Persuasion: A Defense of Rhetoric and Judgment*. Cambridge, MA: Harvard University Press.

Gasser, Urs and Virgilio A.F. Almeida. 2017. "A Layered Model for AI Governance." *IEEE Internet Computing* 21(6):58–62.

Gerber, Alan S., Donald P. Green and Christopher W. Larimer. 2008. "Social pressure and

voter turnout: Evidence from a large-scale field experiment." *American Political Science Review* 102(01):33–48.

Goldstein, Josh A and Girish Sastry. 2023. "The Coming Age of AI-powered Propaganda. How to Defend Against Supercharged Disinformation." *Foreign Affairs* 7.

Goldstein, Josh A, Jason Chao, Shelby Grossman, Alex Stamos and Michael Tomz. 2024. "How persuasive is AI-generated propaganda?" *PNAS Nexus* 3(2):pgae034.
URL: *https://doi.org/10.1093/pnasnexus/pgae034*

Gordon, Ann and Jerry L. Miller. 2004. "Values and Persuasion During the First Bush-Gore Presidential Debate." *Political Communication*, . Publisher: Taylor & Francis.

Graham, Jesse, Jonathan Haidt and Brian A. Nosek. 2009. "Liberals and Conservatives Rely on Different Sets of Moral Foundations." *Journal of Personality and Social Psychology* 96(5):1029–1046.

Graham, Jesse, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik and Peter H. Ditto. 2013. Chapter Two - Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. Vol. 47 of *Advances in Experimental Social Psychology* Academic Press pp. 55–130.
URL: *https://www.sciencedirect.com/science/article/pii/B9780124072367000024*

Gray, Kurt and Jonathan E. Keeney. 2015. "Disconfirming Moral Foundations Theory on Its Own Terms: Reply to Graham (2015)." *Social Psychological and Personality Science* 6(8):874–877.
URL: *https://doi.org/10.1177/1948550615592243*

Green, Melanie C. and Timothy C. Brock. 2000. "The role of transportation in the persuasiveness of public narratives." *Journal of Personality and Social Psychology* 79(5):701–721.

Hackenburg, Kobi, Ben M. Tappin, Paul Röttger, and Helen Margetts. 2025. "Scaling language model size yields diminishing returns for single-message political persuasion." *Proceedings of the National Academy of Sciences* 122(10):e2413443122.

Hackenburg, Kobi and Helen Margetts. 2024. "Evaluating the persuasive influence of political microtargeting with large language models." *Proceedings of the National Academy of Sciences* 121(24):e2403116121.

Hackenburg, Kobi et al. 2025. "The levers of political persuasion with conversational artificial intelligence." *Science* 390(6777):–.
URL: *https://doi.org/10.1126/science.aea3884*

Haidt, Jonathan and Craig Joseph. 2011. "How Moral Foundations Theory Succeeded in Building on Sand: A Response to Suhler and Churchland." *Journal of Cognitive Neuroscience* 23(9):2117–2122.
URL: *https://doi.org/10.1162/jocn.2011.21638*

Hamilton, Alexander. 1787. "Federalist No. 1." The Independent Journal. Part of *The Federalist Papers*.
URL: *https://founders.archives.gov/documents/Hamilton/01-04-02-0152*

Hartman, Todd K. and Christopher R. Weber. 2009. "Who Said What? The Effects of Source Cues in Issue Frames." *Political Behavior* 31(4):537–558.
URL: *https://doi.org/10.1007/s11109-009-9088-y*

Hersch, Eitan D. and Brian F. Schaffner. 2013. "Targeted Campaign Appeals and the Value of Ambiguity." *The Journal of Politics* 75(2):283–566.

Hersh, Eitan D. 2015. *Hacking the electorate: How campaigns perceive voters*. Cambridge University Press.

Heseltine, Michael and Bernhard Clemm von Hohenberg. 2024. "Large language models as a substitute for human experts in annotating political text." *Research & Politics* 11(1):20531680241236239.

Hewitt, Luke, Ashwini Ashokkumar, Isaias Ghezae1 and Robb Willer. 2024. "Predicting Results of Social Science Experiments Using Large Language Models.".
URL: *https://samim.io/dl/Predicting%20results%20of%20social%20science%20experiments%20using%20la*

Hirsh, Jacob B, Sonia K Kang and Galen V Bodenhausen. 2012. "Personalized persuasion: Tailoring persuasive appeals to recipients' personality traits." *Psychological science* 23(6):578–581.

Hutto, C.J. and E.E. Gilbert. 2014. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*. Ann Arbor, MI: pp. 216–225.

Kalla, Joshua L., Adam Seth Levine and David E. Broockman. 2022. "Personalizing moral reframing in interpersonal conversation: A field experiment." *The Journal of Politics* 84(2):1239–1243.

Kalla, Joshua L. and David E. Broockman. 2016. "Durably reducing transphobia: A field experiment on door-to-door canvassing." *Science* 352(6282):220–224.

Kalla, Joshua L. and David E. Broockman. 2020. "Reducing exclusionary attitudes through interpersonal conversation: evidence from three field experiments." *American Political Science Review* 114(2):410–425.

Kam, Cindy D. 2020. ""And Why Is That a Partisan Issue?" Source Cues, Persuasion, and School Lunches." *The Journal of Politics* 82(1):361–366.

Ke, Luoma, Song Tong, Peng Cheng and Kaiping Peng. 2025. "Exploring the frontiers of LLMs in psychological applications: a comprehensive review." *Artificial Intelligence*

Review 58(10):305.

URL: *https://doi.org/10.1007/s10462-025-11297-5*

Kirkeby-Hinrup, Asger and Jakob Stenseke. 2025. "The psychology of LLM interactions: the uncanny valley and other minds." *Journal of Psychology and AI* 1(1):2457627.

Krause, Rebecca J. and Derek D. Rucker. 2019. "Strategic Storytelling: When Narratives Help Versus Hurt the Persuasive Power of Facts." *Personality and Social Psychology Bulletin* 46(2):216–227.

Li, Rende and Ying Jiang. 2026. "LLM-Based Viewpoint Mining in the "Blame Game": How U.S. Media Frame China's Debt Debate." *Chinese Political Science Review* .

Lightman, Hunter, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever and Karl Cobbe. 2023. "Let's Verify Step by Step.".

URL: *https://arxiv.org/abs/2305.20050*

Lin, Hause, Gabriela Czarnek, Benjamin Lewis, Joshua P. White, Adam J. Berinsky, Thomas Costello, Gordon Pennycook, David G. Rand et al. 2025. "Persuading voters using human–artificial intelligence dialogues." *Nature* .

Lu, Louise, Zakary L. Tormala and Adam Duhachek. 2025. "How AI sources can increase openness to opposing views." *Scientific Reports* 15(1):17170.

Lu, Qinghua, Liming Zhu, Xiwei Xu, Jon Whittle, Didar Zowghi and Aurelie Jacquet. 2024. "Responsible AI Pattern Catalogue: A Collection of Best Practices for AI Governance and Engineering." *ACM Comput. Surv.* 56(7).

URL: *https://doi.org/10.1145/3626234*

Manning, Benjamin S., Kehang Zhu and John Horton. 2024. "Automated Social Science:

Language Models as Scientist and Subjects.”.
URL: *https://www.nber.org/papers/w32381*

Matz, S.C., J.D. Teeny, S.S. Vaid, H. Peters, G.M. Harari and M. Cerf. 2024. “The potential of generative AI for personalized persuasion at scale.” *Scientific Reports* 14:4692.

Meincke, Lennart, Karan Girotra, Gideon Nave, Christian Terwiesch and Karl T. Ulrich. 2024. “Using Large Language Models for Idea Generation in Innovation.” *The Wharton School Research Paper Forthcoming* . Available at SSRN: https://ssrn.com/abstract=4526071.

Minkkinen, Matti and Matti Mäntymäki. 2023. “Discerning Between the “Easy” and “Hard” Problems of AI Governance.” *IEEE Transactions on Technology and Society* 4(2):188–194.

Mutz, Diana Carole, Paul M Sniderman and Richard A Brody. 1996. *Political persuasion and attitude change*. Ann Arbor: University of Michigan Press.

Naunov, Martin, Carlos Rueda-Cañòn and Timothy J. Ryan. 2025. “Who’s Persausive? Understanding Citizen-to-Citizen Efforts to Change Minds.” *The Journal of Politics* .

Nejjar, Mohamed, Luca Zacharias, Fabian Stiehle and Ingo Weber. 2025. “LLMs for science: Usage for code generation and data analysis.” *Journal of Software: Evolution and Process* 37(1):e2723.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/smr.2723*

Nelson, Thomas E. and Jennifer Garst. 2005. “Values-based Political Messages and Persuasion: Relationships among Speaker, Recipient, and Evoked Values.” *Political psychology* 26(4):489–516.

Nezhad, Mansoureh Motahari, Maysam Avakh Kisomi and Fatemeh Gholinezhad. 2025. Adaptive Persuasion in Conversational AI: An LLM-Driven Framework for Dynamic

Strategy Switching via Personality and Sentiment Analysis. In *2025 11th International Conference on Web Research (ICWR)*. pp. 145–149.

Nicholson, Stephen P. 2012. ""Polarizing Cues." *American Journal of Political Science* 56(1):52–66.

O'Keefe, Daniel J. 2013. The Elaboration Likelihood Model. In *The SAGE Handbook of Persuasion*, ed. James Price Dillard and Lijiang Shen. Thousands Oaks, CA: Sage pp. 137–150.

O'Keefe, Daniel J. 2016. *Persuasion: Theory and Research*. Third ed. Thousand Oaks, CA: Sage Publications.

O'Keefe, Daniel J. 2018. "Message pretesting using assessments of expected or perceived persuasiveness: Evidence about diagnosticity of relative actual persuasiveness." *Journal of Communication* 68(1):120–142.

O'Keefe, Daniel J. 2020. "Message Pretesting Using Perceived Persuasiveness Measures: Reconsidering the Correlational Evidence." *Communication Methods and Measures* 14(1):25–37.

OpenAI. 2023. "GPT-4 System Card." https://cdn.openai.com/papers/gpt-4-system-card.pdf. OpenAI system card and documentation for GPT-4.

OpenAI. 2024. "Hello GPT-4o." https://openai.com/index/hello-gpt-4o/. Accessed: 2025-05-14.

OpenAI. 2025. "GPT-5 System Card." https://cdn.openai.com/gpt-5-system-card.pdf. OpenAI system card and documentation for GPT-5.

Ornstein, Joseph T, Elise N Blasingame and Jake S Truscott. 2023. "How to train your stochastic parrot: Large language models for political texts." *Political Science Research and Methods* pp. 1–18.

Ortega, Alberto López. 2022. "Are microtargeted campaign messages more negative and diverse? An analysis of Facebook Ads in European election campaigns." *European Political Science* 21:335–358.

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*. NIPS '22 Red Hook, NY, USA: Curran Associates Inc.

Palmer, Alexis and Arthur Spirling. 2023. "Large Language Models Can Argue in Convincing Ways About Politics, But Humans Dislike AI Authors: implications for Governance." *Political Science* 75(3).

Pan, Sarah, Vladislav Lialin, Sherin Muckatira and Anna Rumshisky. 2023. "Let's Reinforce Step by Step.".
URL: *https://arxiv.org/abs/2311.05821*

Park, Joon Sung, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*. pp. 1–22.

Petersen, Michael Bang and Kevin Arceneaux. 2020. "An Intuitionist Theory of Argument Strength in Politics: How Intuitive Cognitive Biases Produce Universally Strong Arguments." *Political Psychology* 41(6):1113–1131.

Petty, Richard E. and John T. Cacioppo. 1986. "The elaboration likelihood model of persuasion." *Advances in Experimental Social Psychology* 19.

Qin, Xuan. 2024. "Intelligent Technologies and Methodological Transformations in the Social Sciences." *Chinese Political Science Review* 9:1–17.

QwenTeam. 2025. "Qwen3 Model Card and Documentation." https://qwenlm.ai. Official documentation for Qwen 3 family.

Roger, Fabien and Ryan Greenblatt. 2023. "Preventing Language Models From Hiding Their Reasoning.".
URL: *https://arxiv.org/abs/2310.18512*

Schoenegger, Philipp, Francesco Salvi, Jiacheng Liu, Xiaoli Nan, Ramit Debnath, Barbara Fasolo, Evelina Leivada, Gabriel Recchia, Fritz Günther, Ali Zarifhonarvar, Joe Kwon, Zahoor Ul Islam, Marco Dehnert, Daryl Y. H. Lee, Madeline G. Reinecke, David G. Kamper, Mert Kobaş, Adam Sandford, Jonas Kgomo, Luke Hewitt, Shreya Kapoor, Kerem Oktar, Eyup Engin Kucuk, Bo Feng, Cameron R. Jones, Izzy Gainsburg, Sebastian Olschewski, Nora Heinzelmann, Francisco Cruz, Ben M. Tappin, Tao Ma, Peter S. Park, Rayan Onyonka, Arthur Hjorth, Peter Slattery, Qingcheng Zeng, Lennart Finke, Igor Grossmann, Alessandro Salatiello and Ezra Karger. 2025. "Large Language Models Are More Persuasive Than Incentivized Human Persuaders.".
URL: *https://arxiv.org/abs/2505.09662*

Schwartz, Shalom H. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in Experimental Social Psychology*. Vol. 25 San Diego: Academic Press pp. 1–65.

Schwartz, Shalom H., Jan Cieciuch, Michele Vecchione, Eldad Davidov, Ronald Rischer, Constanze Beierlein, Alice Ramos, Markku Verkasalo, Jan-Erik Lonnqvist, Kursad Demirutku, Ozlem Dirilen-Gumus and Mark Konty. 2012. "Refining the Theory of Basic Individual Values." *Journal of Personality and Social Psychology* 103(4):663–688.

Si, Chenglei, Diyi Yang and Tatsunori Hashimoto. 2024. "Can LLMs Generate Novel

Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers.".
URL: *https://doi.org/10.48550/arXiv.2409.04109*

Sides, John, Lynn Vavreck and Christopher Warshaw. 2022. "The Effect of Television Advertising in United States Elections." *American Political Science Review* 116(2):702–718.

Simchom, Almog, Matthew Edwards and Stephan Lewandowsky. 2024. "The persuasive effects of political microtargeting in the age of generative artificial intelligence." *PNAS Nexus* 3(2):1–5.

Sreedhar, Karthik, Alice Cai, Jenny Ma, Jeffrey V Nickerson and Lydia B Chilton. 2025. Simulating Cooperative Prosocial Behavior with Multi-Agent LLMs: Evidence and Mechanisms for AI Agents to Inform Policy Decisions. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*. pp. 1272–1286.

Stiennon, Nisan, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei and Paul Christiano. 2020. Learning to summarize from human feedback. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS '20 Red Hook, NY, USA: Curran Associates Inc.

Suhler, Christopher L. and Patricia Churchland. 2011. "Can Innate, Modular "Foundations" Explain Morality? Challenges for Haidt's Moral Foundations Theory." *Journal of Cognitive Neuroscience* 23(9):2103–2116.

Summerfield, Christopher, Lisa P Argyle, Michiel Bakker, Teddy Collins, Esin Durmus, Tyna Eloundou, Iason Gabriel, Deep Ganguli, Kobi Hackenburg, Gillian K Hadfield et al. 2025. "The impact of advanced AI systems on democracy." *Nature Human Behaviour* pp. 1–11.

Susmann, Mark W., Mengran Xu, Jason K. Clark, Laura E. Wallace, Kevin L. Blankenship, Aviva Z. Philipp-Muller, Andrew Luttrell andDuane T. Wegener and Richard E.

Petty. 2021. "Persuasion amidst a pandemic: Insights from the Elaboration Likelihood Model." *European Review of Social Psychology* 33(2):323–359.

Tanusondjaja, Arry, Aaron Michelon, Nicole Hartnett and Lara Stocchi. 2023. "Reaching Voters on Social Media: Planning Political Advertising on Snapchat." *International Journal of Market Research* 65(5):566–580.

Tappin, Ben M., Chloe Wittenberg, Luke B. Hewitt, Adam J. Berinsky and David G. Rand. 2023. "Quantifying the potential persuasive returns to political microtargeting." *Proceedings of the National Academy of Sciences* 120(25):e2216261120.

Team, Gemma. 2025. "Gemma 3 Technical Report." *arXiv preprint arXiv:2503.19786* .
URL: *https://arxiv.org/abs/2503.19786*

Tessler, Michael Henry, Michiel A. Bakker, Daniel Jarrett, Hannah Sheahan, Martin J. Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C. Parkes, Matthew Botvinick and Christopher Summerfield. 2024. "AI can help humans find common ground in democratic deliberation." *Science* 386(6719).

Timm, Jasper, Chetan Talele and Jacob Haimes. 2025. "Tailored truths: Optimizing llm persuasion with personalization and fabricated statistics." *arXiv preprint arXiv:2501.17273* .

Timoneda, Joan C. and Sebastián Vallejo Vera. 2026. "Rolling Memory: A New Approach to Annotation with Generative LLMs in Social and Political Research." *Chinese Political Science Review* .

Tormala, Zakary L. 2016. "The role of certainty (and uncertainty) in attitudes and persuasion." *Current Opinion in Psychology* 10:6–11. Consumer behavior.
URL: *https://www.sciencedirect.com/science/article/pii/S2352250X1530004X*

Törnberg, Petter. 2024. "Large language models outperform expert coders and supervised classifiers at annotating political social media messages." *Social Science Computer Review* p. 08944393241286471.

Van Kleef, Gerben A., Helma van den Berg and Marc W. Heerdink. 2015. "The persuasive power of emotions: Effects of emotional expressions on attitude formation and change." *Journal of Applied Psychology* 100(4):1124–1142.

Velez, Yamil and Patrick Liu. 2024. "Confronting Core Issues: A Critical Assessment of Attitude Polarization Using Tailored Experiments." *American Political Science Review* .

Velez, Yamil and Patrick Liu. 2025. "Confronting Core Issues: A Critical Assessment of Attitude Polarization Using Tailored Experiments." *American Political Science Review* 119(2):1036–1053.

Visser, Penny S., George Y. Bizer and Jon A. Krosnick. 2006. "Exploring the Latent Structure of Strength-related Attitude Attributes." *Advances in experimental social psychology* 38:1–67.
URL: *http://www.sciencedirect.com/science/article/pii/S006526010638001X*

Voelkel, Jan G. and Matthew Feinberg. 2018. "Morally reframed arguments can affect support for political candidates." *Social Psychology and Personality Science* 9(8):917–924.

Votta, Fabio, Simon Kruschinski, Mads Hove, Natali Helberger, Tom Dobber and Claes de Vreese. 2024. "Who Does (n't) Target You? Mapping the Worldwide Usage of Online Political Microtargeting." *Journal of Quantitative Description: Digital Media* 4.

Wagner, Benjamin C. and Richard E. Petty. 2022. The Elaboration Likelihood Model of Persuasion: Thoughtful and Non-Thoughtful Social Influence. In *Theories in Social Psychology, 2nd Edition*, ed. Derek Chadee. Hoboken, NJ: John Wiley & Sons pp. 120–142.

Wang, Yu. 2024. "LLMs in Political Science: Heralding a New Era of Visual Analysis." *arXiv preprint arXiv:2403.00154* .

Wang, Zhenyu, Dequan Wang, Yi Xu, Lingfeng Zhou and Yiqi Zhou. 2025. "Intelligent Computing Social Modeling and Methodological Innovations in Political Science in the Era of Large Language Models." *Journal of Chinese Political Science* .

Woods, Dwayne. 2025. "Stag Hunt in the Digital Wilds: Legitimizing Global AI Governance Amidst Diverse Terrains." *Fudan Journal of the Humanities and Social Sciences* .

Wu, Viviana Chiu Sik. 2024. "Leveraging Computational Methods for Nonprofit Social Media Research: A Systematic Review and Methodological Framework." *Journal of Chinese Governance* 9(3):303–327.

Xu, Yingjin. 2023. "What Can AI Learn from Psychology and When Can AI Neglect It?" *Fudan Journal of the Humanities and Social Sciences* 16:495–513.

Zhang, Yanzhao, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie et al. 2025. "Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models.".

Zhuo, Jingming, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin and Kai Chen. 2024. ProSA: Assessing and Understanding the Prompt Sensitivity of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics.
URL: *https://aclanthology.org/2024.findings-emnlp.108/*

Zolkowski, Artur, Kei Nishimura-Gasparian, Robert McCarthy, Roland S. Zimmermann and David Lindner. 2025. "Early Signs of Steganographic Capabilities in Frontier LLMs.".
URL: *https://arxiv.org/abs/2507.02737*