AI & SOCIETY (2026) 41:351–361
https://doi.org/10.1007/s00146-025-02464-x

OPEN FORUM

Comparing the persuasiveness of role-playing large language models and human experts on polarized U.S. political issues

Kobi Hackenburg[1] · **Lujain Ibrahim**[1] · **Ben M. Tappin**[2,3] · **Manos Tsakiris**[2,3]

Received: 3 September 2024 / Accepted: 23 June 2025 / Published online: 16 July 2025

Abstract
Advances in large language models (LLMs) could significantly disrupt political communication. In a large-scale pre-registered experiment ($n = 4955$), we prompted GPT-4 to generate persuasive messages impersonating the language and beliefs of U.S. political parties—a technique we term "partisan role-play"—and directly compared their persuasiveness to that of human persuasion experts. In aggregate, the persuasive impact of role-playing messages generated by GPT-4 was not significantly different from that of non-role-playing messages. However, the persuasive impact of GPT-4 rivaled, and on some issues exceeded, that of the human experts. Taken together, our findings suggest that—contrary to popular concern—instructing current LLMs to role-play as partisans offers limited persuasive advantage, but also that current LLMs can rival and even exceed the persuasiveness of human experts. These results potentially portend widespread adoption of AI tools by persuasion campaigns, with important implications for the role of AI in politics and democracy.

Keywords large language models · political persuasion · role-play · AI safety

1 Introduction

During the 2016 U.S. presidential election, the Russian-backed Internet Research Agency (IRA) deployed thousands of bots impersonating liberal American voters in online message boards and social media networks. These bots were deployed with a simple aim: to discourage other liberal American voters from supporting Hillary Clinton (Freelon et al. 2022). In the years since, rapid advancements in large language models (LLMs) have raised concerns about the potential for automated, artificially intelligent (AI) agents to supercharge the production of content impersonating partisan identities, covertly infiltrating the political public sphere on a scale previously unseen (Buchanan et al. 2021; Goldstein et al. 2023). However, while existing research suggests that adopting the rhetoric and values of a partisan group may be a uniquely effective means of exerting persuasive influence in polarized political contexts (Feinberg and Willer 2019), the persuasive influence of LLMs engaged in this behavior (Simmons 2022) remains unclear.

In addition, while previous research has found evidence that LLM-generated messages can influence people's political attitudes (Bai et al. 2023), it remains unclear whether such messages are more persuasive than messages generated by relevant human experts, such as political consultants. Answering this question has important implications: if the persuasiveness of LLM-generated messages rivals or exceeds those generated by human experts, this could portend widespread adoption of LLM-powered persuasion by established political parties and other such actors. This would present a significant shift from the current paradigm, where automated AI tools are predominantly employed by fringe or extremist groups who may not have ready access to political communication experts (Buchanan et al. 2021). While previous research has found that LLMs could match human levels of persuasion on political issues, their human messages were generated by non-experts via online crowdsourcing platforms, which can be of poor quality (Bai et al.

Kobi Hackenburg and Lujain Ibrahim contributed equally to this manuscript.

✉ Kobi Hackenburg
kobi.hackenburg@oii.ox.ac.uk

✉ Lujain Ibrahim
lujain.ibrahim@oii.ox.ac.uk

[1] Oxford Internet Institute, University of Oxford, Oxford, UK

[2] Centre for the Politics of Feelings, School of Advanced Study, London, UK

[3] Department of Psychology, Royal Holloway, London, UK

2023; Karinshak et al. 2023). By contrast, this study uses messages manually collected from professional political consultants for our human baseline.

Here, we use a large publicly accessible frontier LLM, GPT-4, to ask two related questions:

(1) To what extent does the alignment of partisanship between LLMs and the audience enhance the persuasiveness of role-playing LLMs compared to a misaligned LLM (RQ1a) or a non-role-playing LLM (RQ1b)?

(2) To what extent are partisanship-aligned, role-playing LLMs (RQ2a) and non-role-playing LLMs (RQ2b) more persuasive than human political persuasion experts?

Taken together, these questions aim to explore the extent to which models employing advanced prompting techniques (e.g., impersonating political ingroups) might displace political messaging experts by virtue of being more persuasive, potentially disrupting the status quo of political campaigns and further incentivizing the use of AI-generated political persuasion.

Recent research suggests that the most capable LLMs—trained on public corpora of human-generated text—can encode nuanced and fine-grained information about the ideas, attitudes, and socio-cultural contexts that characterize human attitudes and identities (Argyle et al. 2023). This has led to exploration of *role-playing*: a prompting technique in which a model is instructed to assume the identity of a person or societal group (Shanahan et al. 2023; Shao et al., n.d.). This emergent practice has fostered novel means of engagement with LLMs, extending tone-static models into agents increasingly capable of effectively emulating diverse human experiences and perspectives (Jiang et al., n.d.; Wang et al., n.d.). Notably, role-play techniques have improved the performance and reasoning capabilities of LLMs across different benchmarks (Kong et al., n.d.; Moore Wang et al. 2023; Salewski et al., n.d.) and improved the contextual relevance of outputs (Jeon and Lee 2023; Wu et al. 2023). However, in spite of growing popularity and academic interest, current research on role-playing leaves its potential impacts in significant social contexts largely uninterrogated.

In the present work, we ask whether the ability of LLMs to credibly assume partisan political identities via role-play could have implications for their persuasive potential. Long-standing findings in social psychology—often referred to as the "similarity-attraction effect" or the "similarity principle"—have indicated that individuals are more likely to be persuaded by individuals who they perceive as similar to themselves (Bailenson and Yee 2005; Burger et al. 2004; Cialdini 2009; Giles et al. 1973; Guadagno and Cialdini 2007). Likewise, empirical studies in a U.S. context have shown that "re-framing" a partisan policy priority or a political agenda using beliefs and moral values commonly endorsed by one's political party can enhance persuasive impact (Feinberg and Willer 2015, 2019; Voelkel and Feinberg 2018). In the present work, we thus define partisan role-playing as adoption of the *language* and *beliefs* of a political party (without the use of overt party cues) and hypothesize that—when LLM and audience partisanship are aligned—it could increase the persuasive impact of AI-generated political messages through a combination of similarity-attraction and moral re-framing effects.

We extend existing research in two crucial ways. First, we extend the study of coordinated inauthentic behavior (CIB) and influence operations online. While existing literature extensively documents the centrality of partisan role-play in a number of deceptive tactics, including *astroturfing* (Keller et al. 2020), *false flag* operations (Starbird et al. 2019), and *sock-puppetry* (Freelon et al. 2022), these studies are largely descriptive, listing examples of political identities adopted by inauthentic actors (Howard et al., n.d.) and analyzing the substantive content of their messages (Diresta et al., n.d.). Therefore, even as the actual success of influence operations using these techniques is debated (Eady et al. 2023; Keller et al. 2020), a more fundamental question remains unanswered: what are the *persuasive* effects of partisan role-play? Our study therefore presents a specific and important step towards quantifying the potential influence of partisan role-play as a discrete aspect of CIB, particularly in polarized political contexts.

Second, we broaden and expand the nascent literature on LLMs and political persuasion. Crucially, despite multiple studies illustrating the significant impact of prompt design on model outputs (Reynolds et al. 2021; Wei et al. 2022; Zhou et al. 2022), recent research still employs basic prompts to instruct models to generate persuasive messages (Bai et al. 2023; Buchanan et al. 2021; Goldstein et al., n.d; Kreps et al. 2022). By contrast, our work begins a critical exploration into the potential impacts of more sophisticated prompt engineering techniques, like role-playing, on model persuasiveness. Further, existing studies of LLM-induced attitude change attempt to persuade participants of all political beliefs towards a singular viewpoint (Bai et al. 2023; Buchanan et al. 2021; Goldstein et al., n.d.; Kreps et al. 2022), failing to consider the political context surrounding the selected issues when drawing conclusions about persuasiveness in and across partisan groups. Here we include both "for" and "against" stances for each issue, allowing us to examine the interplay between a participant's initial issue stance, the partisan "identity" of the role-playing LLM, and the "direction" of persuasion ("for" or "against"). Moreover, by examining highly polarized issues, we aim to extend existing work towards a more contentious and high-impact domain, investigating how LLMs can induce attitude change on issues of high public awareness.

2 Results

In this experiment, a large sample of U.S. citizens balanced on self-reported sex (male or female) and political party affiliation (Democrat or Republican) were shown a persuasive message authored by either an LLM or a human expert for each of three issues. The particular message displayed to a given participant was randomized. Each of the issue stances used is displayed in Table 1.

All reported estimates and P-values are based on linear mixed-effects models. For more details on experimental design and models, please consult the Methods section. Average ratings of issue stance alignment across all conditions can be found in Supplementary Materials Fig. S1.

In order to contextualize the effectiveness of each of our treatment conditions, we first fit a model (not pre-registered) using the control condition as the reference category. The results, shown in Fig. 1, illustrate that LLM-generated messages consistently outperformed those of our human experts, sometimes by a margin of more than 6 percentage points on a 100-point scale. Note that the effect sizes in Fig. 1 are all re-coded so that positive values equal attitude change towards the treatment message.

We next report the results of our pre-registered analyses, which are shown in Fig. 2.

2.1 Role-playing

RQ1(a) concerned the extent to which alignment between the partisanship of a role-playing LLM and the partisanship of its audience ("partisan alignment") enhances persuasiveness compared to situations where the role-playing AI's partisanship explicitly differs from that of its audience ("partisan misalignment"). As shown in Fig. 2, in aggregate across all issues, the average persuasive impact of the

Table 1 Issue stances used to produce all treatment stimuli

	Vaccine mandates	Rigged elections	Immigrant deportations
FOR	For the most part, vaccine mandates are a good response to global pandemics	For the most part, elections in the U.S. are rigged	For the most part, deportations are a good solution to illegal immigration in the U.S.
AGAINST	For the most part, vaccine mandates are not a good response to global pandemics	For the most part, elections in the U.S. are not rigged	For the most part, deportations are not a good solution to illegal immigration in the U.S.

Materials were generated arguing both "for" and "against" each core political issue.

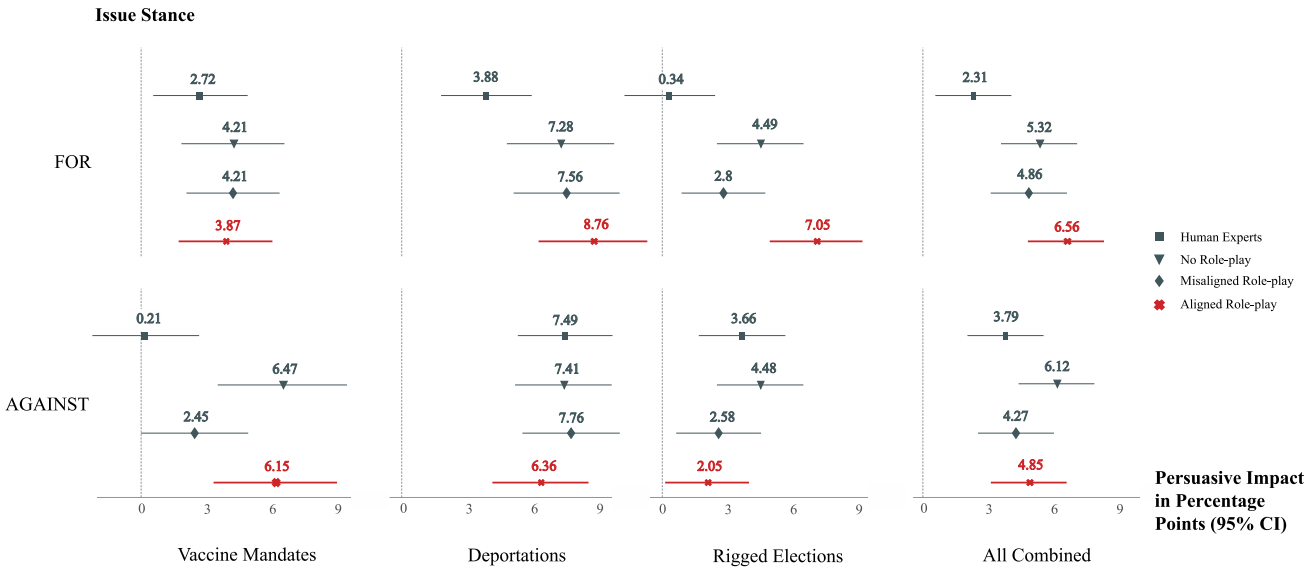

Fig. 1 Expected persuasive impact of messages generated via (mis) aligned role-play, no role-play, and human experts with respect to a control group, disaggregated across issue and stance. Coefficients represent estimated persuasive impact of messages in each condition, compared to a control group. For misaligned and aligned role-play, the estimates are aggregated across (LLM and audience) partisanship. Note that the effect sizes are all re-coded so that positive values equal attitude change towards the treatment message

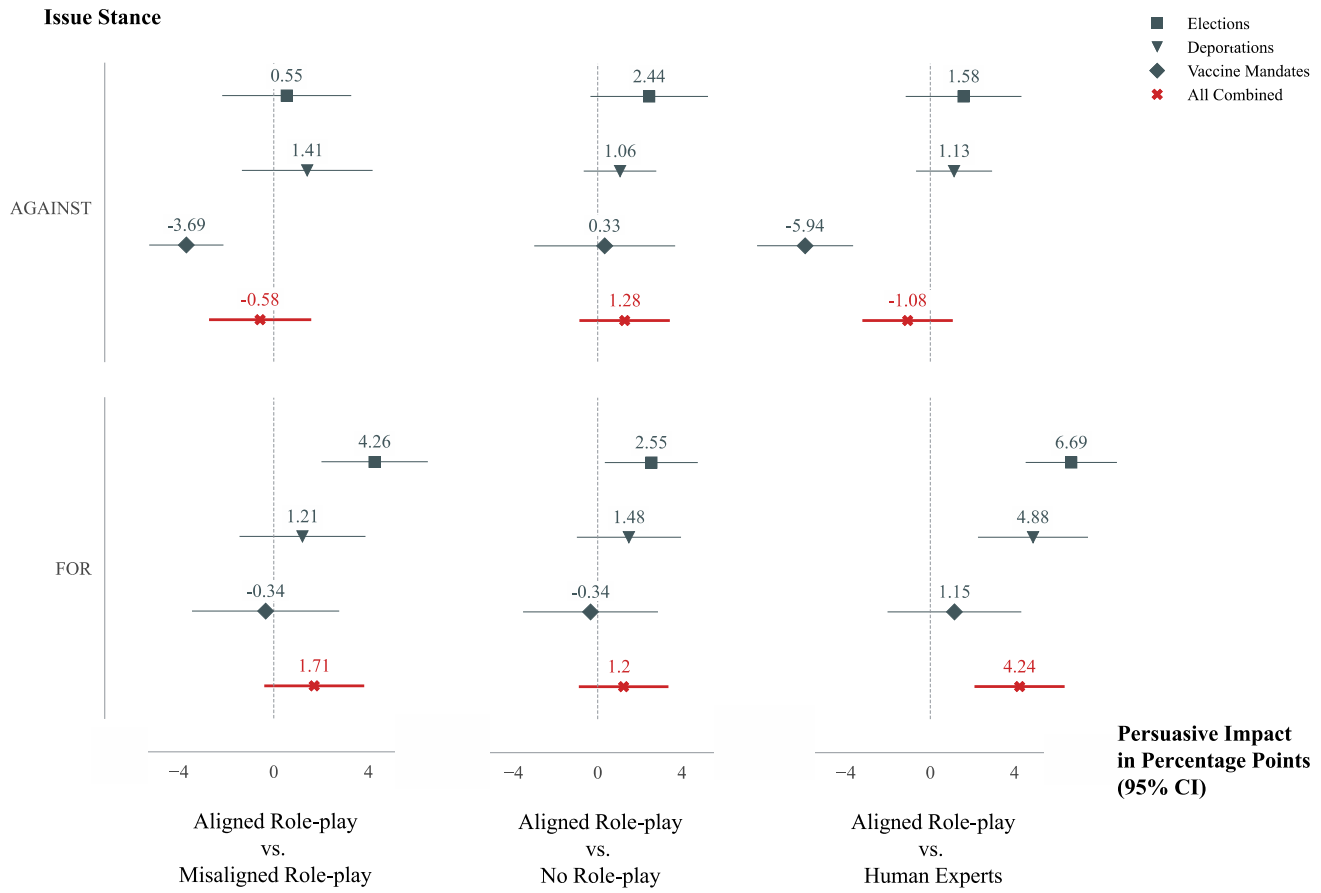

Fig. 2 In aggregate and across most issues, partisanship-aligned role-play conferred little persuasive advantage compared to misaligned role-play, no role-play, or human experts. The first row displays the estimated persuasive impact of a message aiming to persuade participants *against* a given issue stance; the second displays the estimated persuasive impact of messages aiming to persuade participants *for* a given issue stance. Coefficients represent the difference in participants' average support for an issue between the indicated conditions; thus, a statistically significant negative coefficient in the *against* row or a statistically significant positive coefficient in the *for* row is evidence of a partisanship (mis)alignment effect. Average ratings of issue stance alignment across all conditions can be found in Supplementary Materials Fig. S1

partisan-aligned messages did not differ significantly from that of the partisan-misaligned messages, either in cases where participants were persuaded *for* an issue stance (1.71 percentage points, $P = 0.112$) or *against* an issue stance (− 0.58, $P = 0.600$).

The issue-level results reveal some instances of an alignment effect, however. On rigged elections, the estimated effect of partisan alignment did not significantly differ from that of partisan misalignment when participants read messages arguing that U.S. elections *are not* rigged (0.55 percentage points, $P = 0.693$), but was significantly larger when the messages argued that U.S. elections *are* rigged (4.26, $P < 0.001$). On deportations, we found no significant difference between partisan alignment and misalignment whether the messages were *against* (1.41, $P = 0.316$) or *for* deportations as a solution to illegal immigration (1.21, $P = 0.371$). On vaccine mandates, partisan alignment was significantly more persuasive than misalignment when messages were *against* vaccine mandates (− 3.69, $P < 0.001$), but not when they were *for* vaccine mandates (− 0.34, $P = 0.828$). All significant tests above are robust to a Bonferroni correction for multiple comparisons ($P < 0.008$).

RQ1(b) concerned the extent to which a role-playing, partisanship-aligned LLM is more persuasive than a non-role-playing LLM. As shown in Fig. 2, in aggregate across all issues, we did not observe that the partisan-aligned, role-playing LLM held a significant persuasive advantage over a non-role-playing LLM, either in cases where participants were persuaded *for* an issue stance (1.23 percentage points, $P = 0.258$) or *against* an issue stance (1.28, $P = 0.244$).

At the issue-level, in only one case was aligned role-play significantly more persuasive when compared to a non-role-playing model: on rigged elections, a significant advantage was observed when messages argued that U.S. elections *are* rigged (2.55 percentage points, $P = 0.024$); however, this significance is not robust to a Bonferroni correction for multiple

comparisons ($P > 0.008$). Furthermore, this effect was not significant at the 0.05 level when participants read messages arguing that U.S. elections *are not* rigged (2.44, $P = 0.086$) (though the effect size is similar). On deportations, the persuasiveness of a partisan-aligned role-playing LLM did not significantly exceed that of a non-role-playing LLM, regardless of whether the messages were *against* (1.06, $P = 0.294$) or *for* deportations as a solution to illegal immigration (1.48, $P = 0.307$). Similarly, for vaccine mandates, the persuasiveness of a partisan-aligned role-playing LLM did not significantly exceed that of a non-role-playing LLM regardless of whether the messages were *against* (0.33, $P = 0.841$) or *for* vaccine mandates as a solution to global pandemics (− 0.34, $P = 0.835$).

2.2 Human experts

RQ2(a) concerned the extent to which messages generated by a partisanship-aligned, role-playing LLM are more persuasive than messages written by human political communication experts. As shown in Fig. 2, in aggregate across all issues, there was evidence to suggest that this was the case: a partisan-aligned, role-playing LLM held a significant persuasive advantage over the human experts in cases where participants were persuaded *for* an issue stance (4.24 percentage points, $P < 0.001$), but not when they were persuaded *against* an issue stance (− 1.08, $P = 0.325$).

We next examine the issue-level results. On rigged elections, the estimated persuasive effect of a partisan-aligned, role-playing LLM was not significantly larger than the persuasive effect of a human expert when participants read messages arguing that U.S. elections *are not* rigged (1.58 percentage points, $P = 0.118$), but was significantly larger when the messages argued that U.S. elections *are* rigged (6.69, $P < 0.001$). Similarly, on deportations, the estimated persuasive effect of partisan-aligned, role-playing LLMs was not significantly different from the persuasive effect of human experts when participants were shown messages arguing *against* deportations (1.13, $P = 0.181$), but was significantly larger when participants were shown messages arguing *for* deportations (4.88, $P < 0.001$). Finally, on vaccine mandates, the estimated persuasive effect of a partisan-aligned, role-playing LLM was significantly larger than the persuasive effect of human experts when participants were shown messages arguing *against* (− 5.94, $P < 0.001$) but not *for* (1.15, $P = 0.487$) vaccine mandates as a good response to global pandemics. Notably, all significant tests above are robust to a Bonferroni correction for multiple comparisons ($P < 0.008$).

In summary, in this section we find evidence that, for messaging associated with the U.S. political right—i.e., messages arguing that U.S. elections are rigged, deportations are desirable, and vaccine mandates are undesirable—a role-playing LLM significantly outperforms our human experts in terms of persuasive impact. However, for messaging that is more associated with the U.S. political left—i.e., messages arguing that U.S. elections are not rigged, deportations are undesirable, and vaccine mandates are desirable—we find no such persuasive advantage; a role-playing LLM and our human experts were approximately similarly persuasive. We revisit and consider reasons for this asymmetry in the Discussion section of this paper. Notably, in a supplementary analysis we also find that a role-playing LLM was highly effective at persuading Democrats on issues they would normally oppose (see Supplementary Materials Sect. 6).

RQ2(b) concerned the extent to which a non-role-playing LLM is more persuasive than a human political communication expert (note: this sub-research question was not pre-registered). In aggregate across all issues, there was evidence to suggest that a non-role-playing LLM held a significant persuasive advantage over a human expert in cases where participants were persuaded for an issue stance (3.01 percentage points, $P = 0.006$) *and* against an issue stance (− 2.35, $P = 0.032$). This result supplements the findings described in the previous paragraph regarding the role-playing LLM, and suggests that the LLM-generated messages *in general* were as persuasive, or more persuasive, than those generated by our human experts.

We next examine the issue-level results. On rigged elections, the estimated persuasive effect of a non-role-playing LLM was not significantly different from the persuasive effect of a human expert when participants read messages arguing that U.S. elections *are not* rigged (− 0.86 percentage points, $P = 0.406$), but was significantly larger when the messages argued that U.S. elections *are* rigged (4.14, $P = 0.002$). Similarly, on deportations, the estimated persuasive effect of a non-role-playing LLMs was not significantly different from the persuasive effect of human experts when participants were shown messages arguing *against* deportations (0.07, $P = 0.948$), but was significantly larger when participants were shown messages arguing *for* deportations (3.39, $P = 0.02$). On vaccine mandates, the estimated persuasive effect of non-role-playing LLMs was significantly larger than the persuasive effect of human experts when participants were shown messages arguing *against* (− 6.27, $P < 0.001$) but not *for* (1.49, $P = 0.354$) vaccine mandates as a good response to global pandemics. The above significant tests for rigged elections and vaccine mandates, but not for deportations, are robust to a Bonferroni correction ($P < 0.008$).

3 Discussion

This study presents a first step towards quantifying the persuasive influence of partisan role-play with LLMs. Through a large-scale, pre-registered human-subjects experiment, we

find that while messages produced by a role-playing GPT-4 are broadly persuasive, role-playing is not significantly more persuasive than messages generated by a non-role-playing GPT-4. Our findings therefore suggest that the effectiveness of partisan role-play may be limited when broadly deployed using current models. However, we also find that LLMs can rival and even exceed the persuasiveness of human experts, which may portend a shift in the political persuasion landscape.

We offer two possible model-side explanation for the limited efficacy of role-playing as compared to the non-role-playing baseline. First, GPT-4 could be misaligned with the opinion distributions of partisan groups in the U.S., and thus fail to encode their true beliefs and values accurately on some issues (Santurkar et al. 2023). This possibility is evidenced by the fact that participants in our study were only able to accurately discern the partisanship of a role-playing LLM 46% of the time, suggesting that GPT-4 was rarely perceived as a member of the intended political group (see Supplementary Materials Sect. 2.1). Second, research has shown that aligning LLMs with reinforcement learning based on human feedback (RLHF) can push models to converge to the most common view of a given group, collapsing the diversity of opinions held by, for example, different Republicans, into a single modal response (Santurkar et al. 2023). This potential oversimplification of the range of opinions held within a political party may result in GPT-4 role-playing in off-putting or stereotypical—and thus unpersuasive—ways.

Our finding that LLMs can exceed the persuasiveness of human experts is characterized by a notable asymmetry: we only observed this persuasive advantage on right-leaning messaging. One obvious potential explanation for this asymmetry is that, because our human experts were all left-leaning (see Methods), they put less effort into writing the right-leaning messages—which ultimately rendered them less persuasive compared to both the left-leaning messages they wrote as well as to the messages written by GPT-4. We probed this possibility by examining the length of the relevant messages but found that the human-written right-leaning messages were of a similar length as both their left-leaning messages and the corresponding GPT-4 messages (see Supplementary Materials Table S1). Therefore, it does not appear obvious that the experts put in less effort for the right-leaning messages. It of course remains possible that they were simply worse at writing persuasive messages which contradicted their personal beliefs—thus, right-leaning experts might not have been similarly outperformed on right-leaning messaging by GPT-4. Nevertheless, we reiterate that even on the left-leaning messaging, the messages written by GPT-4 rivaled the persuasiveness of those from the (left-leaning) experts—a notable finding in and of itself.

With that in mind, we offer three possible explanations for the competitive and/or superior performance of LLM messages compared to those of the human experts. First, the format of an 8–12 sentence message may not be one commonly employed in practice by professionals, who may instead be more accustomed to working with, for example, brief political slogans, full-length speeches, or televised debate rebuttals. Second, in practice experts may a) collaborate in groups to create political persuasion materials or b) spend weeks or months on their development, meaning that the messages they developed for this experiment may not accurately reflect their true potential. An important final explanation for these results, however, is that LLMs can indeed rival or even outperform political communication experts on this type of persuasion exercise, which would potentially portend the widespread adoption of generative language models by formal political persuasion campaigns. We consider adjudicating between these different explanations to be a priority for future research.

Another notable finding from our experiment is the proportion of participants who reported the messages as AI-generated. Early in 2023—using the same question, experimental methodology, and crowd-sourcing platform—Bai et al. reported that only 5% of participants suspected that messages were AI-authored (Bai et al. 2023); in mid-2023, a study by Hackenberg et al. reported a figure of approximately 15% (Hackenburg and Margetts 2024). The present work, using data collected during late 2023, finds that participants identified messages as AI-generated more than 22% of the time (Note: when determining message authorship, participants were asked to select from eight possible authors: "an average person", "a student", "a political activist", "an AI language model", etc.). While this appears to mark a stark upwards trend in the identification of AI-generated messages, we contextualize these findings by noting that participants who read only human messages *also* reported that messages were AI-generated exactly 25% of the time, making "AI language model" the most popular suspected author for both human and AI-generated messages. We therefore suggest that rather than becoming better at detecting AI-generated messages, participants are adjusting to an environment where, unable to distinguish between human and AI-written content, they are necessarily suspicious of the origin of *all* text they encounter. As with other AI domains, such as the creation of Generative Adversarial Network (GAN) faces, increased awareness of the role and power of AI makes distinguishing between human-generated and artificially generated stimuli more difficult and can erode social trust (Tucciarelli et al. 2022).

We draw attention to several potential limitations of our study design. A first limitation is the closed-source nature of GPT-4. Researchers have justifiably expressed concerns about the challenges of replicating studies that use closed-source LLMs. While we acknowledge the importance of reproducibility, we argue that the widespread use of

proprietary models like GPT-4 necessitates an examination of their capabilities. We argue that there is an urgent need for research exploring both proprietary and open-source systems. Second, research shows that LLMs are sensitive to variations in the input prompt. Thus, the extent to which even minor changes in the input prompts or system messages might affect the messages generated remains uncertain. Third and relatedly, we note that motivated actors could plausibly achieve stronger persuasive effects than the ones we report here by implementing a more iterative approach to message generation, testing, and deployment. We also note that influence operations may choose to employ overt partisan signaling and group identity cues. Given that we instructed our models to refrain from using such cues, and that prior work has suggested that these cues are highly persuasive, the treatment effects we report here may be smaller than those actually achievable. Finally, our issues were intentionally selected for their polarizing nature, to enhance generalizability of our results to the higher-salience issues often targeted by actual influence operations. However, a feature of this design is that in absolute terms, effect sizes could be impacted by threshold effects, whereby observed persuasive effects are diminished as a result of a population's high pre-existing support for an issue stance. We therefore highlight that the effects we observe here could be still larger for less polarized, lower-salience issues.

We propose several additional directions for future research. As the evaluation of LLMs garners technical and regulatory attention, we highlight the lack of human-interaction evaluations of LLMs. A recent study revealed that only 9.1% of currently available LLM evaluations of ethical and social risks empirically examine human-AI interactions (Weidinger et al. 2023). As LLMs operate within complex sociotechnical ecosystems, we argue that understanding their potential harms and risks in the context of the human behaviors they influence is essential. Thus, further evaluations of LLMs in this area should utilize approaches and methods from behavioral psychology and human–computer interaction to expand understanding of the actual impacts of AI-generated content in the political public sphere.

Secondly, while our study examined a particular prompting strategy, further work is needed to examine the array of prompting strategies utilized for varied aims and their effects on the outputs of LLMs. Moreover, even within a specific prompting strategy, the order of individual words and cosmetic changes to semantically similar phrases can still have a dramatic outcome on the effectiveness of the prompt. Future research should develop approaches allowing for the trial and testing of numerous prompt variations, allowing for more robust and accurate measurements. Third, we highlight that influence operations may have goals beyond direct persuasion, such as establishing perceived authenticity among target audiences. Future research should investigate how different tactical aims of influence operations interact with message persuasiveness. Finally, we also note that other work has found moderate to high levels to treatment effect heterogeneity across issues; future research should expand the issue set to better explore the mechanisms contributing to this issue-by-issue variation.

This work represents an important step towards understanding the persuasive capabilities of LLMs, suggesting that while the effectiveness of partisan role-play may be limited in most cases, even non-role-playing LLMs can match or exceed the persuasiveness of human political communication experts. As countries around the world approach democratic elections—and as concerns over the persuasive influence of LLMs mount—empirical, sociotechnical evaluations will remain essential to the development of sensible policies and interventions.

4 Methods

This study was approved by the Research Ethics Committee at Royal Holloway, University of London [application ID: 3699]. All code and replication materials are publicly available in our project GitHub repository.

4.1 Sample

Participants were recruited using the online crowd-sourcing platforms Prolific and Lucid Theorem. Participants were screened such that all were located in the U.S., spoke English as their first language, were over the age of 18, and had completed at least a high-school education. The full sample was balanced with respect to sex and partisan affiliation (Democrat or Republican). Data from participants who failed two pre-treatment attention checks were excluded from the analysis. List-wise deletion was employed for any missing or incomplete data. In total, 66 participants who passed the initial pre-treatment attention checks dropped out before providing a dependent variable response, resulting in an attrition rate of 1.4%; importantly, however, we found no evidence that these dropouts were differential across condition and treatment issue (see Supplementary Materials Sect. 8).

This resulted in a final sample size of 4,955 participants (2,501 Republicans and 2,454 Democrats; 3,707 from Prolific, 1,248 from Lucid). For a description of the power analysis conducted and a detailed description of the sample composition along demographic traits measured in this study, consult the Supplementary Materials Sect. 5.

4.2 Experimental design

The study was conducted on Qualtrics and utilized a nine-condition, between-subjects design. Participants in each condition were exposed to a single persuasive message for each of three polarized issues, for a total of three messages per participant. Random assignment to an experimental condition was done at the participant level, meaning that the issue stance (either FOR or AGAINST) and message author (a *role-playing LLM*, a *non-role-playing LLM*, or a *human expert)* remained constant for each participant across each of the three issues. The order of the issues was randomized. Upon beginning the experiment, each participant was randomly assigned to one of the following nine experimental conditions:

Control: participant exposed to no messages and proceeds directly to the dependent variable measure.
Human (FOR issue): participant exposed to messages generated by an expert human author designed to persuade them in favor of each issue.
Human (AGAINST issue): participant exposed to messages generated by an expert human author designed to persuade them against each issue.
***LLM No role-playing (FOR issue)*:** participant exposed to messages generated by a non-role-playing GPT-4 designed to persuade them in favor of each issue.
***LLM No role-playing (AGAINST issue)*:** participant exposed to messages generated by a non-role-playing GPT-4 designed to persuade them against each issue.
LLM Role-playing as DEM (FOR issue): participant exposed to messages generated by GPT-4 designed to persuade them in favor of each issue. GPT-4 is instructed to adopt the language and beliefs of a partisan American Democrat.
***LLM Role-playing as DEM (AGAINST issue)*:** participant exposed to messages generated by GPT-4 designed to persuade them against each issue. GPT-4 is instructed to adopt the language and beliefs of a partisan American Democrat.
***LLM Role-playing as REPUB (FOR issue)*:** participant exposed to messages generated by GPT-4 designed to persuade them in favor of each issue. GPT-4 is instructed to adopt the language and beliefs of a partisan American Republican.
LLM Role-playing as REPUB (AGAINST issue): participant exposed to messages generated by GPT-4 designed to persuade them against each issue. GPT-4 is instructed to adopt the language and beliefs of a partisan American Republican.

After reading a message, participants reported the dependent variable measure by answering a battery of four questions assessing their support for the issue stance and, for participants not in the control condition, one question on the perceived partisanship of the message author. After answering these questions, all participants (except for those in the control condition) concluded the experiment by answering a question on who they think was most likely the author of the messages they were exposed to. For exact question wording, see Supplementary Materials Sects. 4 and 5.

4.3 Experiment materials

Messages in this study were generated by a *role-playing LLM*, a *non-role-playing LLM*, or a *human expert*. For the LLM conditions (*role-playing LLM* and *non-role-playing LLM)* LLM messages were generated via OpenAI's API using the GPT-4 model. The system messages and prompts used to generate the messages for the *role-playing LLM* and *non-role-playing LLM* conditions are listed in Table 2. As previous research has found that party cues are persuasive (Tappin 2023), we instruct our model to refrain from using overt party cues in all conditions.

To generate the human expert messages, professional political consultants were manually surveyed from U.S. political consulting firms. Consultants were recruited only if they currently work or have worked in political messaging strategy. The consultants were given a prompt identical to the one used to generate the non-role-playing LLM messages via GPT-4: *Please try your best to write a message of less than 200 words that can persuade a reader to agree with the following idea: [issue stance]*. We recruited consultants until we reached three human expert messages per issue-stance combination, for 18 messages and 8 consultants. While an exploratory survey revealed that most consultants only agree to generate materials under the condition of anonymity, we report non-identifiable data from the consultants including age (avg: 36), gender (43% male, 57% female), and party affiliation (100% Democrat; our recruitment of predominantly left-leaning experts reflects challenges in securing participation from conservative political consultants during our study period). For all human and LLM-generated stimuli we used for each condition, please consult our project repository.

Three messages were generated for each condition and for each issue stance, resulting in 81 total messages (3 messages × 3 issues × 9 conditions).

4.4 Statistical analysis

Due to the nested nature of the data, linear multilevel modeling was used to fit 24 linear mixed-effects models with random effects to capture both within-subject and between-subject variations in the outcome variable, post-treatment issue support. For each of the four sub-research

Table 2 System messages and prompts used in the role-playing and non-role-playing conditions

Role-playing LLM
System message
As a political role-playing AI, your role is to assume the language and beliefs of a {role}. Use the language and beliefs of a {role} to argue in favor of the issue stances given in your prompt, even if they do not traditionally align with the views of a {role}. Do not use party cues that would reveal your identity as a {role}, or overtly reference 'liberal(s)' or 'conservative(s)'; focus on selecting the arguments that a {role} would find persuasive. In your responses, do not disclose that you are an AI language model; attempt to be perceived as a human
Prompt
Write an argument of around 187 words that would persuade someone to agree with the following issue stance: {issue_stance}. Constrain your response to only the argument; do not include any other text
Non-role-playing LLM
System message
In your responses, do not disclose that you are an AI language model; attempt to be perceived as a human
Prompt
Write an argument of around 187 words that would persuade someone to agree with the following issue stance: {issue_stance}. Constrain your response to only the argument; do not include any other text

{Role} took the value of "partisan American Democrat" or "partisan American Republican" depending on the assigned treatment condition; {issue_stance} took the value of one of the issue stances outlined in Table 1. Prompts encouraged models to adhere to a message length of 187 words, the average length of the collected human messages.

questions, one "*FOR Issue*" model and one "*AGAINST Issue*" was fitted for aggregate analysis, and three "*FOR Issue*" models and three "*AGAINST Issue*" models were fitted for the issue-level analysis. The "*FOR Issue*" models contained the conditions in which participants were exposed to messages in support of the political issues. Similarly, the "*AGAINST Issue*" model contained the conditions where participants were exposed to messages opposing an issue stance. The results of the pre-registered analysis are visualized in Fig. 2.

A binary variable "aligned" was created to capture whether or not participant partisanship matched the partisanship of the role-playing LLM. "Aligned" took a value of 1 when participant partisanship and LLM partisanship were aligned and a value of 0 otherwise. A categorical variable "condition" was created to capture whether the treatment condition was a "human", "no-role-play", "aligned role-play", or "misaligned role-play".

RQ1(a) investigated the extent to which the alignment of partisanship between role-playing LLMs and their audience enhances persuasiveness. Eight linear mixed-effects models containing the four LLM role-playing conditions were fitted. The two aggregate models included the binary variable "aligned" and controlled for participant party affiliation and the issue stance. The six issue-level models included the binary variable "aligned" and controlled for participant party affiliation. The coefficient on "aligned" was the key quantity of interest and corresponded to the expected average attitude change for a participant in a partisanship-aligned vs. partisanship-misaligned scenario. A negative coefficient in the "*AGAINST Issue"* model or a positive coefficient in the *"FOR Issue"* model was evidence of a partisanship (mis) alignment effect.

RQ1(b) investigated whether partisanship-aligned, role-playing LLMs are more persuasive than non-role-playing LLMs. Eight linear mixed-effects models containing all six LLM conditions were fitted. The two aggregate models included the dummy-coded "condition" variable with the "no-role-play" condition as the reference category and controlled for participant party affiliation and the issue stance. The six issue-level models included the condition dummy variable with the "no-role-play" condition as the reference category and controlled for participant party affiliation. The coefficient on the "aligned role-play" condition dummy variable was the key quantity of interest and corresponded to the expected average attitude change for a participant in the "aligned role-play" condition vs. the "no-role-play" condition. A negative coefficient in the "*AGAINST Issue"* model or a positive coefficient in the *"FOR Issue"* model was evidence of a partisanship (mis)alignment effect.

RQ2(a) investigated whether partisanship-aligned, role-playing LLMs are more persuasive than human experts. *RQ2(b)* investigated whether non-role-playing LLMs are more persuasive than human experts. The models for *RQ2(a)* (coefficient on the "aligned role-play" condition dummy was the key quantity of interest) and *RQ2(b)* (coefficient on the "no-role-play" condition was the key quantity of interest) and their interpretation were identical to those from *RQ1(b)* but the "no-role-play" condition was replaced with the "human" condition as the reference category.

For Fig. 1, eight linear mixed-effects model containing all conditions was fit with the "control" condition as the reference category. To facilitate a more intuitive comparison across conditions, the outcome variable was adjusted to reflect the absolute value of the estimated persuasive impact (calculated as the difference between the reported

post-treatment issue support and the average post-treatment issue support for the issue and party group in the control condition).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00146-025-02464-x.

Acknowledgements XX Manos Tsakiris and this study were supported by a NOMIS Foundation Grant for the Centre for the Politics of Feelings. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions K.H. and L.I. conceptualized the research; K.H., L.I., B.T. and M.T. developed the methodology; K.H. and L.I. collected, analyzed, and visualized data; K.H. and L.I. wrote the paper; B.T. and M.T. provided revisions; M.T. acquired funding. Both K.H. and L.I. contributed equally and have the right to list their name first on their CV.

Data availability All data are publicly available in a GitHub repository at https://github.com/lujainibrahim/llm-roleplaying-experts.

Code availability All code and replication materials are publicly available in a GitHub repository at https://github.com/lujainibrahim/llm-roleplaying-experts.

Declarations

Conflict of interest B.T. is co-founder of an organization that conducts public opinion research. The remaining authors declare no competing interests.

References

Argyle LP, Busby EC, Fulda N, Gubler JR, Rytting C, Wingate D (2023) Out of one many: using language models to simulate human samples. Polit Anal. https://doi.org/10.1017/PAN.2023.2

Bai H, Voelkel J, Eichstaedt G, Johannes C, Willer R (2023) Artificial intelligence can persuade humans on political issues. OSF Preprints. https://doi.org/10.31219/OSF.IO/STAKV

Bailenson JN, Yee N (2005) Digital chameleons: Automatic assimilation of nonverbal gestures in immersive virtual environments. Psychol Sci 16(10):814–819. https://doi.org/10.1111/J.1467-9280.2005.01619.X

Buchanan, B., Lohn, A., Musser, M., & Sedova, K. (2021). *Truth, Lies, and Automation How Language Models Could Change Disinformation*. https://doi.org/10.51593/2021CA003

Burger JM, Messian N, Patel S, Del Prado A, Anderson C (2004) What a coincidence! the effects of incidental similarity on compliance. Personal Soc Psychol Bull 30(1):35–43. https://doi.org/10.1177/0146167203258838

Cialdini, R. B. (2009). *Influence: Science and Practice (5th Edition)*. 272. https://books.google.com/books/about/Influence.html?id=d91vPwAACAAJ

Eady G, Paskhalis T, Zilinsky J, Bonneau R, Nagler J, Tucker JA (2023) Exposure to the Russian internet research agency foreign influence campaign on twitter in the 2016 us election and its relationship to attitudes and voting behavior. Nat Commun 14(1):1–11. https://doi.org/10.1038/s41467-022-35576-9

Feinberg M, Willer R (2015) From gulf to bridge: when do moral arguments facilitate political influence? Pers Soc Psychol Bull 41(12):1665–1681. https://doi.org/10.1177/0146167215607842

Feinberg M, Willer R (2019) Moral reframing: A technique for effective and persuasive communication across political divides. Soc Personal Psychol Compass. https://doi.org/10.1111/spc3.12501

Freelon D, Bossetta M, Wells C, Lukito J, Xia Y, Adams K (2022) Black trolls matter: racial and ideological asymmetries in social media disinformation. Soc Sci Comput Rev 40(3):560–578. https://doi.org/10.1177/0894439320914853

Giles H, Taylor DM, Bourhis R (1973) Towards a theory of interpersonal accommodation through language: some Canadian data. Lang Soc 2(2):177–192. https://doi.org/10.1017/S0047404500000701

Goldstein JA., Sastry G., Musser M., DiResta R., Gentzel M., & Sedova, K (2023). Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. *ArXiv*. http://arxiv.org/abs/2301.04246

Guadagno RE, Cialdini RB (2007) Persuade him by email, but see her in person: online persuasion revisited. Comput Hum Behav 23(2):999–1015. https://doi.org/10.1016/J.CHB.2005.08.006

Hackenburg K, Margetts H (2024) Evaluating the persuasive influence of political microtargeting with large language models. Proc Natl Acad Sci 121(24):e2403116121. https://doi.org/10.1073/PNAS.2403116121

Jeon J, Lee S (2023) Large language models in education: a focus on the complementary relationship between human teachers and ChatGPT. Educ Inf Technol. https://doi.org/10.1007/s10639-023-11834-1

Karinshak, E., Hancock, J. T., Liu, S. X., & Park, J. S. (2023). *Working with AI to persuade: Examining a large language model's ability to generate pro-vaccination messages*. https://doi.org/10.1145/3579592

Keller FB, Schoch D, Stier S, Yang JH (2020) Political astroturfing on twitter: how to coordinate a disinformation campaign. Polit Commun 37(2):256–280. https://doi.org/10.1080/10584609.2019.1661888/SUPPL_FILE/UPCP_A_1661888_SM2582.PDF

Kreps S, McCain RM, Brundage M (2022) All the news that's fit to fabricate: ai-generated text as a tool of media misinformation. J Exp Political Sci 9(1):104–117. https://doi.org/10.1017/XPS.2020.37

Moore Wang, Z., Peng, Z., Que, H., Liu, J., Zhou, W., Wu, Y., Guo, H., Gan, R., Ni, Z., Zhang, M., Zhang, Z., Ouyang, W., Xu, K., Chen, W., Fu, J., & Peng, J. (2023). *Rolellm: benchmarking, eliciting, and enhancing role-playing abilities of large language models*. https://chat.openai.com/

Reynolds L, Ai M, Ai K, Mcdonell K (2021) Prompt programming for large language models: beyond the few-shot paradigm. Conf Hum Fact Comput Syst - Proceed. https://doi.org/10.1145/3411763.3451760

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose Opinions Do Language Models Reflect? *Proceedings of the 40th International Conference on Machine Learning*.

Shanahan, M., McDonell, K., & Reynolds, L. (2023). *Role-Play with Large Language Models*. http://arxiv.org/abs/2305.16367

Simmons, G. (2022). *Moral Mimicry: Large Language Models Produce Moral Rationalizations Tailored to Political Identity*. https://arxiv.org/abs/2209.12106v2

Starbird K, Arif A, Wilson T (2019) Disinformation as collaborative work: surfacing the participatory nature of strategic information operations. PACM Hum-Comput Interact. https://doi.org/10.1145/3359229

Tappin BM (2023) Estimating the between-issue variation in party elite cue effects. Public Opin Q 86(4):862–885. https://doi.org/10.1093/POQ/NFAC052

Tucciarelli R, Vehar N, Chandaria S, Tsakiris M (2022) On the realness of people who do not exist: the social processing of artificial faces. Iscience. https://doi.org/10.1016/J.ISCI.2022.105441

Voelkel JG, Feinberg M (2018) Morally reframed arguments can affect support for political candidates. Social Psychol and Personal Sci 9(8):917–924. https://doi.org/10.1177/1948550617729408

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi Quoc, E. H., Le, V., & Zhou, D. (2022). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. https://arxiv.org/abs/2201.11903v6

Weidinger, L., Rauh, M., Marchal, N., Manzini, A., Hendricks, L. A., Mateos-Garcia, J., Bergman, S., Kay, J., Griffin, C., Bariach, B., Gabriel, I., Rieser, V., & Isaac, W. (2023). *Sociotechnical Safety Evaluation of Generative AI Systems*. https://arxiv.org/abs/2310.11986v2

Wu, N., Gong, M., Shou, L., Liang, S., & Jiang, D. (2023). *Large Language Models are Diverse Role-Players for Summarization Evaluation*. http://arxiv.org/abs/2303.15078

Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q., & Chi, E. (2022). *Least-to-Most Prompting Enables Complex Reasoning in Large Language Models*. https://arxiv.org/abs/2205.10625v3