Evaluating the Effectiveness of Large Language Models for Targeted Political Messaging

Aseem Behl
University of Tübingen
Adrian Zarbock
University of Tübingen

Cite as:
Behl Aseem, Zarbock Adrian (2025), Evaluating the Effectiveness of Large Language Models for Targeted Political Messaging. *Proceedings of the European Marketing Academy*, 54th, (126399)

Paper from the 54th Annual EMAC Conference, Madrid, Spain, May 25-30, 2025

EMAC
EUROPEAN MARKETING ACADEMY

Evaluating the Effectiveness of Large Language Models for Targeted Political Messaging

Abstract: Recent advancements in AI, particularly with models like GPT-4o, have raised concerns about potential misuse in political campaigns. To address this, our research investigates the persuasive power of targeted political messages generated by large language models (LLMs). Specifically, our study explores whether targeted political messages created by generative language models are perceived as more persuasive than non-targeted messages. Building on prior research, we conducted an online experiment to compare the persuasiveness of targeted and non-targeted political messages. Unlike earlier work, this study exclusively measures persuasion, distinguishing it from confirmation bias. It innovates by using implicit data collection methods for targeting, avoiding the reliance on self-reported or sensitive information. Additionally, the study employs a multidimensional political spectrum model, offering greater nuance compared to traditional one-dimensional approaches.

Keywords: machine learning, political marketing, large language models
Track: Social Responsibility and Ethics

1 Introduction

Targeted voter engagement has been a cornerstone of political communication since the inception of modern election campaigns (Magin et al., 2017). However, the revelations surrounding the data-driven campaigns orchestrated by Cambridge Analytica and Facebook, particularly their alleged influence on the 2016 US Presidential election and the Brexit referendum (Codwalladr & Graham-Harrison, 2018), underscored the unprecedented impact of technology on democratic processes.

Recent advancements in generative artificial intelligence have reignited concerns about data-driven manipulation in elections. Large language models, such as OpenAI's GPT-4o, offer the potential for sophisticated, cost-effective, and precisely targeted political messaging at an unparalleled scale. This concern is further substantiated by reports of foreign actors, including the Russian Social Design Agency, leveraging generative AI tools to disseminate disinformation and exert political influence, notably in Germany, through social media platforms (Erb et al., 2024; Terberl & Schmitt, 2024).

The convergence of data-driven microtargeting and generative AI necessitates a deeper understanding of their synergistic potential, particularly in a year like 2024, which witnessed elections involving nearly half the world's population across over 70 countries (Robinson, 2024). This concern is echoed by OpenAI itself, which identifies persuasion as a medium-level risk—the only risk factor so categorized—in their analysis of GPT-4o (OpenAI, 2024).

Against this backdrop, this study explores a crucial question: Are targeted political messages crafted by generative language models perceived as more persuasive than comparable non-targeted messages? This investigation introduces an innovative approach to studying political microtargeting with LLMs, employing an online experiment to evaluate message effectiveness. A purpose-built interactive web application facilitates real-time message delivery and data collection, enabling a robust comparison of targeted and non-targeted messages within a between-subjects experimental design.

Ethical considerations constrained the study. To mitigate potential harm from persuading participants towards opposing viewpoints, controversial or emotionally charged topics, appeals, and misinformation were avoided. Messages focused on factual, unemotional topics. While this might reduce measured effect sizes compared to more provocative approaches, these restrictions were necessary to maintain ethical standards.

2 Related Work

Research on political microtargeting, particularly with large language models, is growing. Tappin et al. (2023) found a 70% persuasion increase with targeted messages on single issues using machine learning, but only when using one targeting variable (party affiliation). Bai et al. (2023) showed language model-generated arguments are as persuasive as human-written ones. Building on this, Hackenburg and Margetts (2024) found that both targeted and non-targeted language model-generated messages are persuasive, but targeting offered no significant advantage.

This study, similar to Hackenburg and Margetts (2024), examines targeted message persuasiveness but with key differences. First, it focuses solely on the *persuasive* effect (changing opinions) rather than the *confirmatory* effect (reinforcing existing opinions). Second, it employs implicit data collection based on established sociological methods for microtargeting, avoiding sensitive data and self-reported information biases. Third, it uses a multi-dimensional model of political space (four dimensions) for more precise classification, unlike previous one-dimensional approaches. Finally, it utilizes the more advanced GPT-4o language model. These differences offer a more nuanced approach to assessing whether large language models can create more persuasive targeted political messages.

3 Methodology

This section details the methodology employed in this study, encompassing the conceptual basis of microtargeting, the selection process for the chosen topics, the messaging strategy adopted, the experimental design implemented, and the statistical analysis applied.

3.1 Research design

Within political science, the nature of political space and its optimal dimensions for mapping have been debated for decades (Stokes, 1963). These dimensions measure political similarities and differences between agents within a political system (Benoit & Laver, 2012), such as the relationship between voters and parties. Statements like "Party A is closer to Party B than Party C" rely on implicit understandings of this space. While countless dimensions are theoretically possible, reducing them to a manageable number is practically challenging (Benoit & Laver, 2012).

This reduction is possible because many political dimensions are correlated. Similar attitudes can be grouped into superordinate dimensions (e.g., liberal-conservative) (Benoit & Laver, 2012). However, determining relevant dimensions remains subjective (Benoit & Laver, 2012), and these dimensions are constantly changing due to political competition (De Vries & Hobolt, 2012; Stokes, 1963). This explains why many studies simplify political space to a single dimension, such as the left-right spectrum. This simplification, while potentially suitable for the US two-party system (Wissenschaftliche Dienste, 2012), is inadequate for more complex, multi-party systems like Germany's (Wissenschaftliche Dienste, 2012), where multiple dimensions are necessary to capture the political landscape accurately (Bakker et al., 2012).

This study utilizes a multi-dimensional microtargeting approach. Drawing upon established sociological surveys (European Social Survey (2023) and German Longitudinal Election Study (2024)) and an expert interview with a political scientist, four key dimensions were identified: economic left-right, social left-right, global vs. national politics (expanding on the European policy dimension), and climate protection, acknowledging its increasing political relevance. Each dimension was operationalized using dual-good trade-offs (German Longitudinal Election Study, 2024): economic (low taxes vs. social safety net), social (individual freedom vs. traditional values), global-national (cooperation vs. sovereignty), and climate (protection vs. living standards). This method allows for precise political classification without requiring demographic or sensitive personal information.

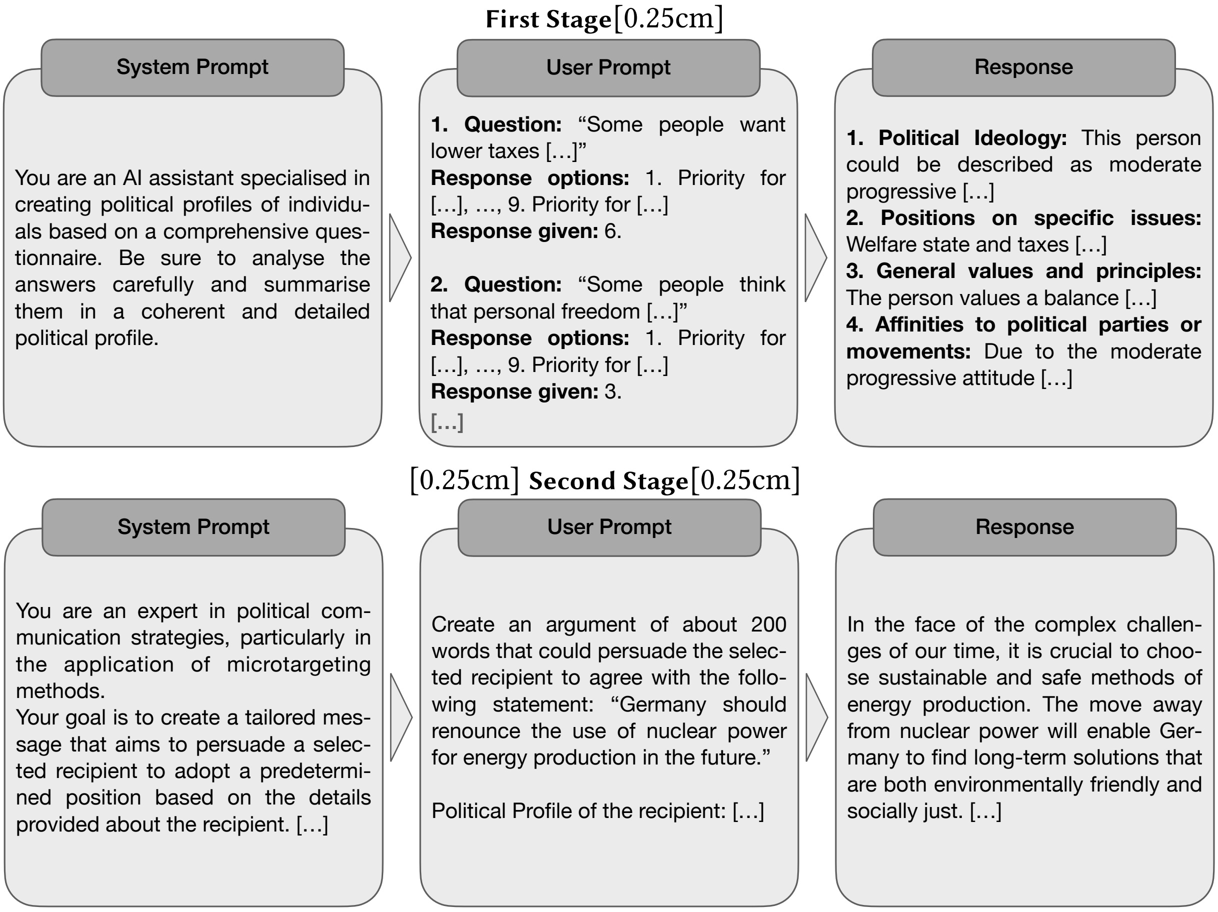
The selection of topics for the study adhered to specific criteria: existing public discourse, appropriateness of content (avoiding highly sensitive topics like military conflicts or migration due to ethical concerns related to potential LLM-generated misinformation), and general accessibility. Four topics met these criteria: continuation of nuclear power (nuclear power), ban on extremist parties (party bans), lifting the debt brake (debt brake), and general speed limit on German highways (general speed limit). These topics represent key areas of political debate in Germany and offer sufficient complexity for nuanced opinions.

3.2 Sample

Study participants (n=271, after filtering for completion and attention checks) were recruited from a German university and an external online survey platform to ensure a broader demographic distribution. Eligibility criteria included being at least 18 years old and possessing C1 German language proficiency.

3.3 *Messaging strategy*

Existing research on political microtargeting with LLMs often uses explicit demographic and political affiliation data (Santurkar et al., 2023). This study adopts a novel two-stage approach, diverging from this reliance on potentially inaccurate self-reported information (Staab et al., 2023). For targeted messages, the first stage creates a detailed political profile based on the participant's stance on the four political dimensions. The second stage generates the message using this profile (Figure 1). Non-targeted messages proceed directly to message generation without profiling (Figure 2). This approach aims for more nuanced and potentially more persuasive messaging.



First Stage[0.25cm]

System Prompt

You are an AI assistant specialised in creating political profiles of individuals based on a comprehensive questionnaire. Be sure to analyse the answers carefully and summarise them in a coherent and detailed political profile.

User Prompt

1. Question: "Some people want lower taxes […]"
Response options: 1. Priority for […], …, 9. Priority for […]
Response given: 6.

2. Question: "Some people think that personal freedom […]"
Response options: 1. Priority for […], …, 9. Priority for […]
Response given: 3.
[…]

Response

1. Political Ideology: This person could be described as moderate progressive […]
2. Positions on specific issues: Welfare state and taxes […]
3. General values and principles: The person values a balance […]
4. Affinities to political parties or movements: Due to the moderate progressive attitude […]

[0.25cm] **Second Stage**[0.25cm]

System Prompt

You are an expert in political communication strategies, particularly in the application of microtargeting methods.
Your goal is to create a tailored message that aims to persuade a selected recipient to adopt a predetermined position based on the details provided about the recipient. […]

User Prompt

Create an argument of about 200 words that could persuade the selected recipient to agree with the following statement: "Germany should renounce the use of nuclear power for energy production in the future."

Political Profile of the recipient: […]

Response

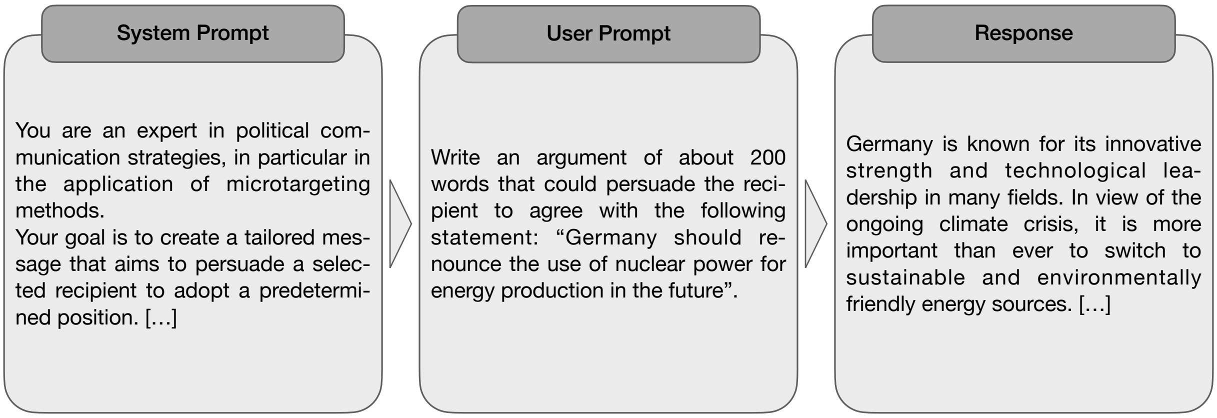
In the face of the complex challenges of our time, it is crucial to choose sustainable and safe methods of energy production. The move away from nuclear power will enable Germany to find long-term solutions that are both environmentally friendly and socially just. […]

Figure 1: Targeted Message Generation Process

3.4 *Experimental design*

The experimental design involved four sequential treatments, one for each topic, presented in random order. Each message presented an argument contrary to the participant's pre-existing stance on that issue. Participants were randomly assigned to one of three treatment groups: non-targeted, targeted, or false-targeted (using an inverted political profile). A separate control group was deemed unnecessary as the baseline persuasiveness of LLM-generated messages is well-established. The false-targeted group served to assess potential negative effects from inaccurate targeting.

The pre-treatment phase consisted of: 1) demographic data collection, 2) assessing participants' positions on the four political dimensions using dual-good trade-off questions, and 3) determining their stance on the four topics using binary (agree/disagree) questions to avoid anchoring effects. During the treatment phase, participants evaluated each message



You are an expert in political communication strategies, in particular in the application of microtargeting methods.
Your goal is to create a tailored message that aims to persuade a selected recipient to adopt a predetermined position. […]

Write an argument of about 200 words that could persuade the recipient to agree with the following statement: “Germany should renounce the use of nuclear power for energy production in the future”.

Germany is known for its innovative strength and technological leadership in many fields. In view of the ongoing climate crisis, it is more important than ever to switch to sustainable and environmentally friendly energy sources. […]

Figure 2: Non-Targeted Message Generation Process[1]

based on argument quality and agreement with the presented (opposing) viewpoint. The post-treatment phase measured perceived target audience.

4 Results

The central research question of this study investigates whether targeted political messages exhibit greater persuasiveness compared to non-targeted or false-targeted messages. In this context, persuasiveness is operationalized as the level of agreement a message recipient expresses with a thesis they initially disagreed with (and vice versa).

It is important to note that the chosen experimental design precludes measuring individual-level persuasion effects. To mitigate potential anchoring effects, the precise level of agreement or disagreement with each issue *before* treatment was deliberately not recorded. Consequently, the analysis of persuasiveness relies solely on comparing aggregated agreement levels between the different treatment groups. The measured effect, therefore, reflects the post-treatment agreement levels and not the difference between pre- and post-treatment opinions. This approach is considered reliable because the treatment groups demonstrate no significant differences in relevant demographic or political characteristics. This similarity allows us to attribute any significant differences in post-treatment agreement levels to the treatment itself.

Examining the overall results presented in Figure 3, non-targeted messages emerge as the most persuasive. With an average agreement level of 3.22 (on a scale from 1 = *do not agree at all* to 9 = *fully agree*), non-targeted messages elicited significantly higher agreement than false-targeted messages (3.22 vs. 2.95, $p = 0.0249$). While non-targeted messages also achieved higher agreement than targeted messages (3.22 vs. 3.05, $p = 0.1264$), this difference is not statistically significant. Thus, the central hypothesis that targeted messages are generally more persuasive is not confirmed.

However, a more nuanced analysis reveals topic-specific variations in persuasiveness. Targeted messages on *party bans* and *debt brake* showed higher agreement levels compared to non-targeted messages, although these differences were not statistically significant (*party bans*: 3.69 vs. 3.61, $p = 0.3923$; *debt brake*: 3.63 vs. 3.50, $p = 0.3209$). More pronounced differences emerged when comparing targeted messages to false-targeted messages. Agreement with targeted messages on *party bans* was significantly higher than with false-targeted messages (3.69 vs. 3.08, $p = 0.0091$). A similar trend was observed for *debt brake*, with targeted messages receiving higher agreement than false-targeted messages, although the difference was not statistically significant (3.63 vs. 3.27, $p = 0.0834$).

Interestingly, targeted messages performed worse on two topics: *nuclear power* and *general speed limit*. On the topic of *nuclear power*, targeted messages elicited lower agreement

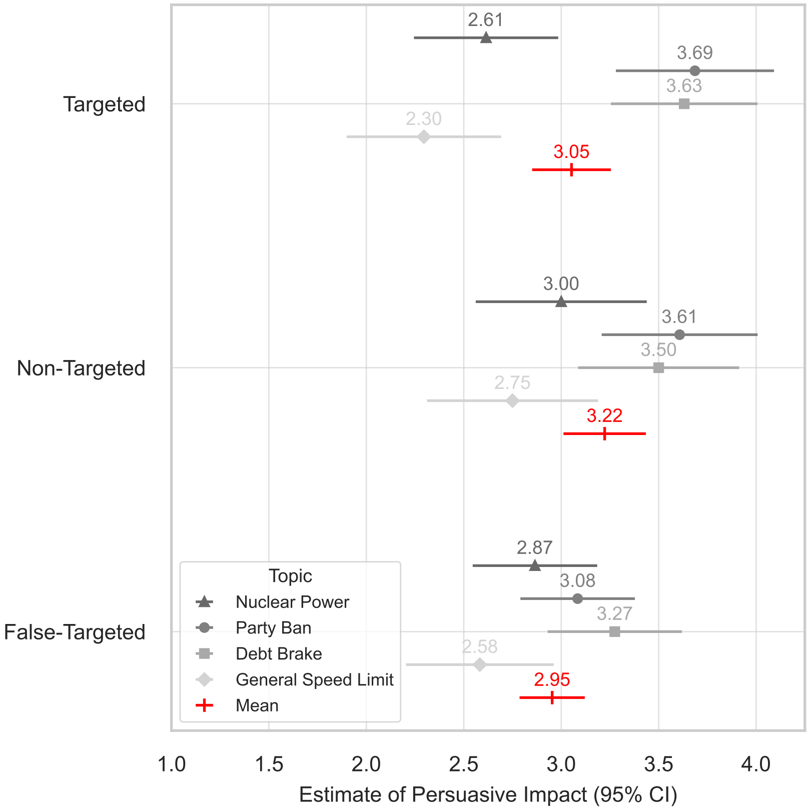


Figure 3: Persuasive Impact of Messages

than both non-targeted messages (2.61 vs. 3.00, $p = 0.0916$) and false-targeted messages (2.61 vs. 2.87, $p = 0.1547$). The pattern was similar for *general speed limit*, with targeted messages again performing worse than both non-targeted messages (2.30 vs. 2.75, $p = 0.0640$) and false-targeted messages (2.30 vs. 2.58, $p = 0.1499$). None of these differences reached statistical significance.

These findings suggest a topic-specific sensitivity influencing the persuasiveness of targeted messages. While non-targeted and false-targeted message agreement levels remained relatively consistent across topics, targeted messages exhibited greater variability. Specifically, targeted messages on *party bans* and *debt brake* performed better, while those on *nuclear power* and *general speed limit* performed worse. This highlights the topic-dependent nature of targeted message effectiveness.

To gain a deeper understanding, two additional metrics were collected: perceived argument quality and perceived target audience. As shown in Figure 4, no significant differences in perceived quality were observed between the three treatment groups across all topics. On average, targeted messages were rated marginally lower than non-targeted messages (4.68 vs. 4.69, $p = 0.4763$) and marginally higher than false-targeted messages (4.68 vs. 4.66, $p = 0.4361$), with neither difference being statistically significant.

Topic-level analysis of perceived quality revealed some minor, non-significant variations, with targeted messages on *party bans* tending to receive higher ratings. Overall, participants perceived argument quality as similar across treatment groups.

Analysis of the perceived target audience revealed that participants were generally unable to discern between targeted and non-targeted messages ($p = 0.1286$). As illustrated in Figure 5, perceived target audience assessments were relatively evenly distributed across treatment groups, suggesting a general perception of messages being addressed to a broad

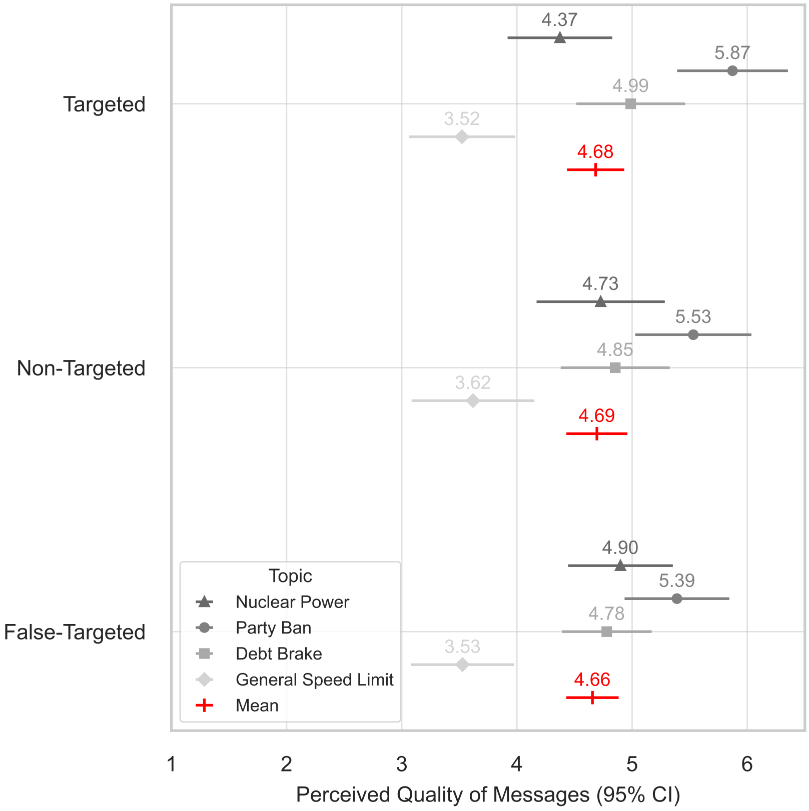


Figure 4: Perceived Quality of Messages

audience.

Finally, message length was examined as a potential confounding factor. While statistically significant differences were found (targeted messages slightly shorter than non-targeted messages, 130.47 vs. 134.04 words, $p = 0.0037$), the magnitude of these differences was small (Figure 6). These minor variations in length are unlikely to have substantially influenced message perception or effectiveness.

5 Discussion

The study's results suggest that targeted political messages are not more persuasive than non-targeted messages. Several explanations are explored, considering the study's limitations.

The sample was not representative of the population. However, sample size alone cannot explain the lack of a significant persuasive advantage for targeted messages; a fundamental reversal of effects would be needed.

One explanation questions the effectiveness of political microtargeting. Hackenburg and Margetts (2024) found no persuasive advantage of targeted messages, while Tappin et al. (2023) found an advantage only when targeting by party affiliation. This raises doubts about microtargeting's effectiveness.

Another explanation is that the targeting information was unsuitable, or the language model (GPT-4o) could not process it effectively. However, this is partly contradicted by the topic-dependent variation in persuasiveness. Targeted messages performed worse on two topics but were most persuasive on the other two, suggesting topic sensitivity.

A third explanation relates to the experimental design: a single interaction with short messages. Repeated exposure over a longer period in a realistic campaign might reveal

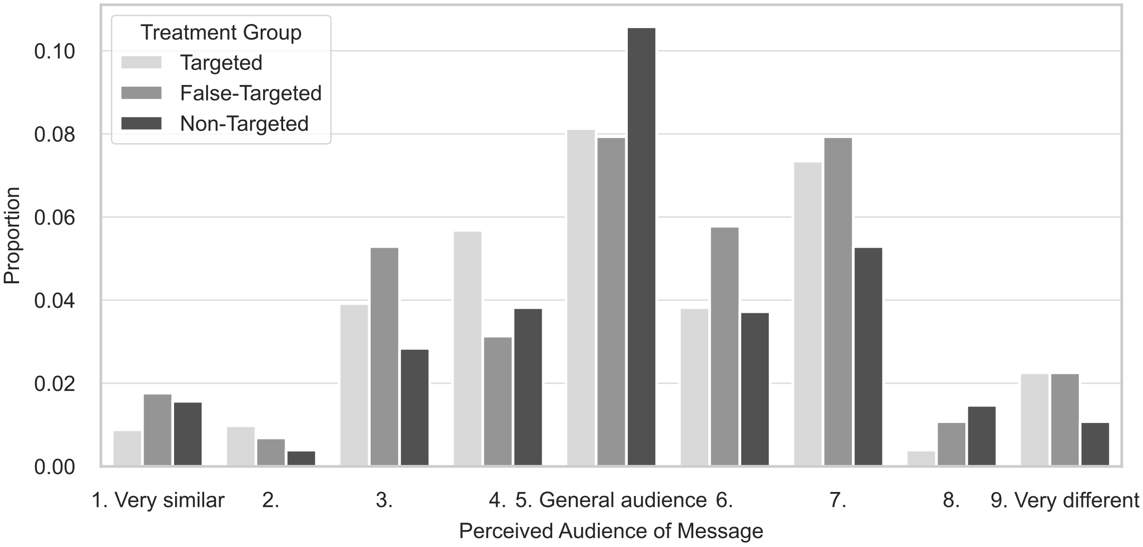


Figure 5: Perceived Audience of Messages

microtargeting's benefits. Furthermore, two-way interaction, a unique feature of microtargeting with large language models, was not explored.

The study's political focus may have attracted politically interested individuals, potentially leading to a self-selection bias. This could mean participants were less likely to be persuaded.

The study excluded confirmatory effects, which occur when messages support existing opinions. As Knobloch-Westerwick et al. (2020) showed, political messages are particularly persuasive when they confirm beliefs. This exclusion, and the ethical avoidance of controversial topics or emotional language, may have weakened the measured effects.

Overall, targeted messages were not found to be more persuasive. This should not be interpreted as a complete negation of the research question, but rather as a cautious assessment of political microtargeting with large language models. More significant effects might be seen in other contexts.

6 Conclusion

The research question of whether large language model-generated political messages are more persuasive with microtargeting was not confirmed. However, this is a conservative estimate of current capabilities. Limitations, both external and self-imposed, played a role, and not all political actors face such restrictions.

This work contributes a novel methodological approach for generating targeted political messages based solely on political attitudes and beliefs, offering a valuable alternative to methods using demographic data. This approach is versatile and can be applied in different political contexts, broadening our understanding of microtargeting with large language models. Practically, the study shows that while targeted persuasion is limited, even generic AI-generated messages are highly persuasive, posing challenges for policymakers regarding misinformation and manipulation.

Future research could explore continuous interaction between language models and recipients, repeated message exposure, or realistic field experiments to assess long-term effects. Investigating which topics are best suited for persuasion, refining models through reinforcement learning with human feedback (RLHF), and integrating generative visual content are also promising areas.

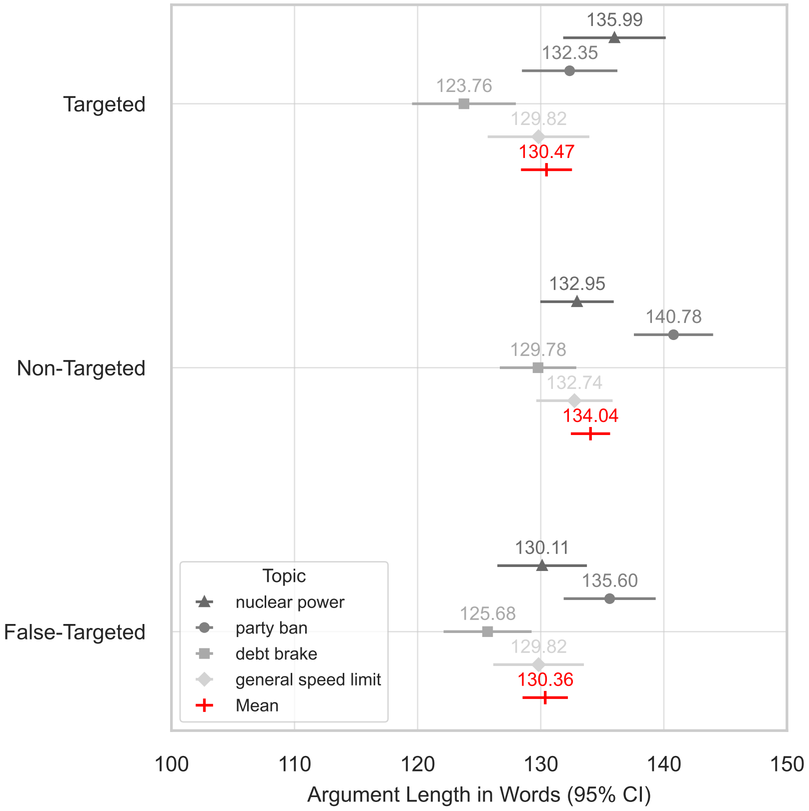


Figure 6: Length of Messages

Generative AI offers significant advances, but its use for political influence requires careful examination. While it can revolutionize political communication, it also carries risks of manipulation. Targeted AI-generated content represents a new dimension of political influence, the full impact of which is unknown. This work contributes to understanding this potential, providing a basis for further research. Policymakers, researchers, and society must ensure these technologies strengthen, rather than undermine, democratic discourse.

References

Bai, H., Voelkel, J. G., Eichstaedt, J. C., & Willer, R. (2023). Artificial intelligence can persuade humans on political issues.

Bakker, R., Jolly, S., & Polk, J. (2012). Complexity in the european party space: Exploring dimensionality with experts. *European Union Politics*, *13*(2), 219–245.

Benoit, K., & Laver, M. (2012). The dimensionality of political space: Epistemological and methodological considerations. *European Union Politics*, *13*(2), 194–218.

Codwalladr, C., & Graham-Harrison, E. (2018, March 17). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The Guardian.*

De Vries, C. E., & Hobolt, S. B. (2012). When dimensions collide: The electoral success of issue entrepreneurs. *European Union Politics*, *13*(2), 246–268.

Erb, S., Salem, S., Schmitt, J., Verschwele, L., & Weinmann, L. (2024, September 16). Propaganda vom Fließband. *Süddeutsche Zeitung.*

European Social Survey. (2023). ESS-10 2020 Documentation Report. Edition 3.0. *Bergen, European Social Survey Data Archive, Sikt - Norwegian Agency for Shared Services in Education and Research, Norway for ESS ERIC.*

German Longitudinal Election Study. (2024). GLES Panel 2023, Welle 25. *GESIS, Köln. ZA7731 Datenfile Version 1.0.0.*

Hackenburg, K., & Margetts, H. (2024). Evaluating the persuasive influence of political microtargeting with large language models. *Proceedings of the National Academy of Sciences*, *121*(24), e2403116121.

Knobloch-Westerwick, S., Mothes, C., & Polavin, N. (2020). Confirmation bias, ingroup bias, and negativity bias in selective exposure to political information. *Communication Research*, *47*(1), 104–124.

Magin, M., Podschuweit, N., Haßler, J., & Russmann, U. (2017). Campaigning in the fourth age of political communication. a multi-method study on the use of facebook by german and austrian parties in the 2013 national election campaigns. *Information, communication & society*, *20*(11), 1698–1719.

OpenAI. (2024). Gpt-4o system card.

Robinson, L. (2024, July 8). At least 70 countries have elections in 2024. A guide in maps and charts. *CNN.*

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose opinions do language models reflect? *International Conference on Machine Learning*, 29971–30004.

Staab, R., Vero, M., Balunović, M., & Vechev, M. (2023). Beyond memorization: Violating privacy via inference with large language models. *arXiv preprint arXiv:2310.07298.*

Stokes, D. E. (1963). Spatial models of party competition. *The American Political Science Review*, *57*(2), 368–377.

Tappin, B. M., Wittenberg, C., Hewitt, L. B., Berinsky, A. J., & Rand, D. G. (2023). Quantifying the potential persuasive returns to political microtargeting. *Proceedings of the National Academy of Sciences*, *120*(25), e2216261120.

Terberl, L., & Schmitt, J. (2024, September 25). Desinformation: Russlands hybrider Krieg [Audio Podcast Episode]. In *Das Thema.* Süddeutsche Zeitung.

Wissenschaftliche Dienste. (2012). *Die Wahlsysteme Deutschlands und der USA: Ein Vergleich.* (WD 1 – 3000/071/12). Deutscher Bundestag.