LARGE LANGUAGE MODELS

The levers of political persuasion with conversational artificial intelligence

Kobi Hackenburg[1,2]*†, Ben M. Tappin[3]*†, Luke Hewitt[4], Ed Saunders[1], Sid Black[1], Hause Lin[5], Catherine Fist[1], Helen Margetts[2], David G. Rand[5,6]*, Christopher Summerfield[1,7]*

There are widespread fears that conversational artificial intelligence (AI) could soon exert unprecedented influence over human beliefs. In this work, in three large-scale experiments (N = 76,977 responses from 42,357 people), we deployed 19 large language models (LLMs)—including some post-trained explicitly for persuasion—to evaluate their persuasiveness on 707 political issues. We then checked the factual accuracy of 466,769 resulting LLM claims. We show that the persuasive power of current and near-future AI is likely to stem more from post-training and prompting methods—which boosted persuasiveness by as much as 51 and 27%, respectively—than from personalization or increasing model scale, which had smaller effects. We further show that these methods increased persuasion by exploiting LLMs' ability to rapidly access and strategically deploy information and that, notably, where they increased AI persuasiveness, they also systematically decreased factual accuracy.

Box 1. Glossary of abbreviations and key terms (with definitions as used in this paper).

FLOPs Floating-point operations; here used to index model scale through "effective pretraining compute."

Effective compute The total FLOPs used to pretrain a model.

Post-training Any training or adaptation applied after pretraining to shape model behavior (e.g., generic chat-tuning, SFT, RM, or SFT+RM).

PPT Persuasion post-training: Post-training specifically to increase persuasiveness (operationalized through SFT, RM, or SFT+RM).

SFT Supervised fine-tuning on curated persuasive dialogues to teach the model to mimic successful persuasion patterns.

RM Reward modeling: A separate model scores candidate replies for how persuasive they will be; the system then selects the top-scoring reply for giving to the human user (i.e., a best-of-k reranker at each turn).

SFT+RM Combined approach: An SFT model generates candidates; an RM selects the most persuasive one.

Base Model with no persuasion-specific post-training. For open-source models, this means generic chat-tuning; for closed-source models, out of the box.

Chat-tuned Open-source models fine-tuned for generic (nonpersuasive) open-ended dialogue to hold post-training constant across models.

Developer post-trained Closed-source frontier models post-trained by developers using heterogeneous, opaque methods.

Open-source versus proprietary (closed-source) models Open-source models are those that the authors could fine-tune; proprietary models could not be fine-tuned and were used out of the box (and, where applicable, with RM).

Frontier model Highly capable, developer post-trained proprietary model (e.g., GPT-4.5 or Grok-3 in this study's taxonomy).

Information density Number of fact-checkable claims made by AI in a conversation.

Academics, policy-makers, and technologists fear that artificial intelligence (AI) may soon be capable of exerting substantial persuasive influence over people (*1*–*13*). Large language models (LLMs) can now engage in sophisticated interactive dialogue, enabling a powerful mode of human-to-human persuasion (*14*–*16*) to be deployed at unprecedented scale. However, the extent to which this will affect society is unknown. We do not know how persuasive AI models can be, what techniques increase their persuasiveness, and what strategies they might use to persuade people. For example, as compute resources continue to grow, models could become ever more persuasive, mirroring the so-called scaling laws observed for other capabilities. Alternatively, specific choices made during model training, such as the use of highly curated datasets, tailored instructions, or user personalization, might be the key enablers of ever greater persuasiveness. In this study, we set out to understand what makes conversational AI persuasive and to define the horizon of its persuasive capability.

To do so, we examine three fundamental research questions (RQs) related to distinct risks. First, if the persuasiveness of conversational AI models increases at a rapid pace as models grow larger and more sophisticated, this could confer a substantial persuasive advantage to powerful actors who are best able to control or otherwise access the largest models, further concentrating their power. Thus, we ask, are larger models more persuasive (RQ1)? Second, because LLM performance in specific domains can be optimized by targeted post-training techniques (Box 1), as has been done in the context of general reasoning or mathematics (*17*–*19*), even small open-source models—many deployable on a laptop—could potentially be converted into highly persuasive agents. This could broaden the range of actors able to effectively deploy AI to persuasive ends, benefiting those who wish to perpetrate coordinated inauthentic behavior for ideological or financial gain, foment political unrest among geopolitical adversaries, or destabilize information ecosystems (*10*, *20*, *21*). So, to what extent can targeted post-training increase AI persuasiveness (RQ2)? Third, LLMs deployed to influence human beliefs could do so by leveraging a range of potentially harmful strategies, such as by exploiting individual-level data for personalization (*4*, *22*–*25*) or by using false or misleading information (*3*), with malign consequences for public discourse, trust, and privacy. So finally, what strategies underpin successful AI persuasion (RQ3)?

We answer these questions using three large-scale survey experiments, across which 42,357 people engaged in conversation with one of 19 open- and closed-source LLMs that had been instructed to persuade them on one of a politically balanced set of 707 issue stances. The sample of LLMs in our experiments spans more than four orders of magnitude in model scale and includes several of the most advanced ("frontier") models as of May 2025: GPT-4.5, GPT-4o, and Grok-3-beta. In addition to model scale, we examine the persuasive impact of eight different prompting strategies motivated by prevailing theories of persuasion and three different post-training methods—including supervised fine-tuning and reward modeling—explicitly designed to maximize AI persuasiveness. Using LLMs and professional human fact-checkers,

[1]UK AI Security Institute, London, UK. [2]Oxford Internet Institute, University of Oxford, Oxford, UK. [3]Department of Psychological and Behavioural Science, London School of Economics and Political Science, London, UK. [4]Department of Sociology, Stanford University, Stanford, CA, USA. [5]Sloan School of Management, Massachusetts Institute of Technology, Boston, MA, USA. [6]Department of Information Science, Marketing and Management Communications, and Department of Psychology, Cornell University, Ithaca, NY, USA. [7]Department of Experimental Psychology, University of Oxford, Oxford, UK. ***Corresponding author. Email: kobi.hackenburg@oii.ox.ac.uk (K.H.); b.tappin@lse.ac.uk (B.M.T.); dgr7@cornell.edu (D.G.R.); christopher.summerfield@psy.ox.ac.uk (C.S.)** †These authors contributed equally to this work.

Downloaded from https://www.science.org at EPF Lausanne on November 27, 2026

we then count and evaluate the accuracy of 466,769 fact-checkable claims made by the LLMs across more than 91,000 persuasive conversations. The resulting dataset is, to our knowledge, the largest and most systematic investigation of AI persuasion to date, offering an unprecedented window into how and when conversational AI can influence human beliefs. Our findings thus provide a foundation for anticipating how persuasive capabilities could evolve as AI models continue to develop and proliferate and help identify which areas may deserve particular attention from researchers, policy-makers, and technologists concerned about its societal impact.

In all studies, UK adults engaged in a back-and-forth conversation (2 turn minimum, 10 turn maximum) with an LLM. Before and after the conversation, they reported their level of agreement with a series of written statements expressing a particular political opinion relevant to the UK on a 0 to 100 scale [following the method used in a related recent work (*26*)]. In the treatment group, the LLM was prompted to persuade the user to adopt a prespecified stance on the issue, using a persuasion strategy randomly selected from one of eight possible strategies (materials and methods). Throughout, we measure the persuasive effect as the difference in mean posttreatment opinion between the treatment group and a control group in which there was no persuasive conversation (unless stated otherwise) in units of percentage points. Although participants were crowdworkers with no obligation to remain beyond two conversation turns to receive a fixed show-up fee, treatment dialogues lasted an average of seven turns and 9 min (see materials and methods for more detail), which implies that participants were engaged by the experience of discussing politics with AI.

Results

Before addressing our main research questions, we begin by validating key motivating assumptions of our work—that conversing with AI (i) is meaningfully more persuasive than exposure to a static AI-generated message and (ii) can cause durable attitude change. To validate the first assumption, we included two static-message conditions in which participants read a 200-word persuasive message written by GPT-4o (study 1) or GPT-4.5 (study 3) but did not engage in a conversation. As predicted, the AI was substantially more persuasive in conversation than through static message, both for GPT-4o (+2.94 percentage points, $P < 0.001$, +41% more persuasive than the static message effect of 6.1 percentage points) and GPT-4.5 (+3.60 percentage points, $P < 0.001$, +52% more persuasive than the static message effect of 6.9 percentage points). To validate the second assumption, in study 1, we conducted a follow-up 1 month after the main experiment, which showed that between 36% (chat 1, $P < 0.001$) and 42% (chat 2, $P < 0.001$) of the immediate persuasive effect of GPT-4o conversation was still evident at recontact—demonstrating durable changes in attitudes (see supplementary materials, section 2.2, for complete output).

Persuasive returns to model scale

We now turn to RQ1: the impact of scale on AI model persuasiveness. To investigate this, we evaluate the persuasiveness of 17 distinct base LLMs (Table 1), spanning four orders of magnitude in scale [measured in effective pretraining compute (*27*); materials and methods]. Some of these models were open-source models, which we uniformly post-trained for open-ended conversation [using 100,000 examples from Ultrachat (*28*) or "chat-tuned" models; see materials and methods for

Table 1. Parameters, pretraining tokens, effective compute, and post-training (open-source, frontier, and PPT) for all base models across the three studies. Ranks are within each study; values marked "≈" are approximate.

Study	Rank	Model name	Parameters	Pretraining tokens (T)	Effective compute (FLOPS, 1×10^{21})	Post-training
1	1	Qwen1.5-0.5B	0.5 billion	2.4	7.20	Open-source
	2	Qwen1.5-1.8B	1.8 billion	2.4	25.92	Open-source
	3	Qwen1.5-4B	4 billion	2.4	57.60	Open-source
	4	Qwen1.5-7B	7 billion	4.0	168.00	Open-source
	5	Llama3-8B	8 billion	15.0	720.00	Open-source
	6	Qwen1.5-14B	14 billion	4.0	336.00	Open-source
	7	Qwen1.5-32B	32 billion	4.0	768.00	Open-source
	8	Llama3-70B	70 billion	15.0	6300.00	Open-source
	9	Qwen1.5-72B	72 billion	3.0	1296.00	Open-source
	10	Qwen1.5-72B-chat	72 billion	3.0	1296.00	Frontier
	11	Qwen1.5-110B-chat	110 billion	4.0	1980.00	Frontier
	12	Llama3-405B	405 billion	15.0	36,450.00	Open-source
	13	GPT-4o	Unknown	Unknown	≈38,100.00*	Frontier
2	1	Llama-3.1-8B	8 billion	15.6	748.80	Open-source + PPT
	2	GPT-3.5-turbo	≈20 billion*	Unknown	≈2578.00*	Frontier + PPT
	3	Llama-3.1-405B	405 billion	15.6	37,908.00	Open-source + PPT
	4	GPT-4o	Unknown	Unknown	≈38,100.00*	Frontier + PPT
	5	GPT-4.5	Unknown	Unknown	≈210,000.00†	Frontier + PPT
3	1	GPT-4o-old (6 August 2024)	Unknown	Unknown	≈38,100.00*	Frontier + PPT
	2	GPT-4o-new (27 March 2025)	Unknown	Unknown	≈38,100.00*	Frontier + PPT
	3	GPT-4.5	Unknown	Unknown	≈210,000.00†	Frontier + PPT
	4	Grok-3-beta	Unknown	Unknown	≈464,000.00*	Frontier + PPT

*Effective compute estimates from Epoch AI (*71*). †Industry insiders suggest that GPT-4.5 was pretrained on ≈10 × the compute of GPT-4. Multiplying Epoch AI's GPT-4 estimate (2.1×10^{25} FLOPs) by 10 yields 2.1×10^{26}.

Downloaded from https://www.science.org at EPF Lausanne on November 27, 2026

details]. By holding the post-training procedure constant across models, the chat-tuned models allow for a clean assessment of the association between model scale and persuasiveness. We also examined a number of closed-source models (such as GPT-4.5 from OpenAI and Grok-3-beta from xAI) that have been extensively post-trained by well-resourced frontier laboratories using opaque, heterogeneous methods ("developer post-trained" models). Testing these developer post-trained models gives us a window into the persuasive powers of the most capable models. However, because they are post-trained in different (and unobservable) ways, model scale may be confounded with post-training for these models, which makes it more difficult to assess the association between scale and persuasiveness.

In Fig. 1, we show the estimated persuasive impact of a conversation with each LLM. Pooling across all models (our preregistered specification), we find a positive linear association between persuasive impact and the logarithm of model scale (Fig. 1, dashed black line), which suggests a reliable persuasive return to model scale: +1.59 percentage points {Bayesian 95% confidence interval (CI), [1.07, 2.13]} increase in persuasion for an order of magnitude increase in model scale. Notably, we find a positive linear association of similar magnitude when we restrict to chat-tuned models only (+1.83 percentage points [1.42, 2.25]; Fig. 1, purple), where post-training is held constant by design. Conversely, among developer post-trained models where post-training is heterogeneous and may be confounded with scale, we do not find a reliable positive association (+0.32 percentage points [−1.18, 1.85]; Fig. 1, green; significant difference between chat-tuned and developer post-trained models, −1.39 percentage points [−2.72, −0.11]). For example, GPT-4o (27 March 2025) is more persuasive (11.76 percentage points) than models thought to be considerably larger in scale, GPT-4.5 (10.51 percentage points, difference test $P = 0.004$) and Grok-3 (9.05 percentage points, difference test $P < 0.001$), as well as models thought to be equivalent in scale, such as GPT-4o with alternative developer post-training (6 August 2024) (8.62 percentage points, difference test $P < 0.001$) (see supplementary materials, section 2.3, for full output tables).

Overall, these results imply that model scale may deliver reliable increases in persuasiveness (although it is hard to assess the effect of scale among developer post-training because of heterogeneous post-training). Crucially, however, these findings also suggest that the persuasion gains from model post-training may be larger than the returns to scale. For example, our best-fitting curve (pooling across models and studies) predicts that a model trained on 10× or 100× the compute of current frontier models would yield persuasion gains of +1.59 percentage points and +3.19 percentage points, respectively (relative to a baseline current frontier persuasion of 10.6 percentage points). This is smaller than the difference in persuasiveness that we observed between two equal-scale deployments of GPT-4o in study 3 that otherwise varied only in their post-training: 4o (March 2025) versus 4o (August 2024) (+3.50 percentage points in a head-to-head difference test, $P < 0.001$; supplementary materials, section 2.3.2). Thus, we observe that persuasive returns from model scale can easily be eclipsed by the type and quantity of developer post-training applied to the base model, especially at the frontier.

Persuasive returns to model post-training

Given these results, we next more systematically examine the effect of post-training on persuasiveness. We focus on post-training that is specifically designed to increase model persuasiveness [we called this persuasiveness post-training (PPT)] (RQ2). In study 2, we test two PPT methods. First, we used supervised fine-tuning (SFT) using a curated subset of the 9000 most persuasive dialogues from study 1 (see materials and methods for inclusion criteria) to encourage the model to copy previously successful conversational approaches. Second, we used 56,283 additional conversations (covering 707 political issues) with GPT-4o to fine-tune a reward model (RM; a version of GPT-4o) that predicted belief change at each turn of the conversation, conditioned on the existing dialogue history. This allowed us to enhance persuasiveness by sampling a minimum of 12 possible AI responses at each dialogue turn, and choosing the response that the RM predicted would be most persuasive (materials and methods). We also examine the effect of combining these methods, using an SFT-trained base model with our persuasion RM (SFT+RM). Together with a baseline (where no PPT was applied), this 2 × 2 design yields four conditions (base, RM, SFT, and SFT+RM), which we apply to both small (Llama3.1-8B) and large (Llama3.1-405B) open-source models.

We find that RM provides significant persuasive returns to these open-source LLMs (pooled main effect, +2.32 percentage points, $P < 0.001$, relative to a baseline persuasion effect of 7.3 percentage points; Fig. 2A). By contrast, there were no significant persuasion gains from SFT (+0.26 percentage points, $P = 0.230$), and no significant

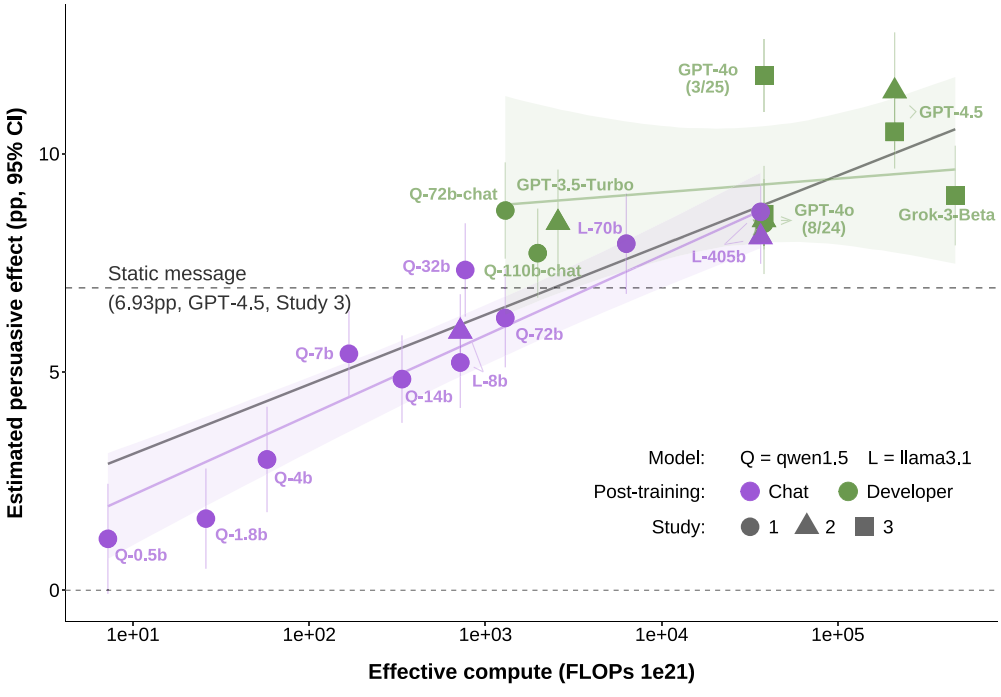

Fig. 1. Persuasiveness of conversational AI increases with model scale. The persuasive impact in percentage points (versus control group) is shown on the *y* axis, plotted against effective pretraining compute [floating-point operations (FLOPs)] on the *x* axis (logarithmic scale). Point estimates are raw average treatment effects with 95% CIs. The black solid line represents the association across all models assuming a linear relationship, and colored lines show separate fits for models that we uniformly chat-tuned for open-ended conversation (purple) and models that were post-trained using heterogeneous, opaque methods by frontier AI developers (green). For proprietary models (GPT-3.5, GPT-4o, GPT-4.5, and Grok-3), where true scale is unknown, we used scale estimates published by research organization Epoch AI (*59*). pp, percentage points.

Downloaded from https://www.science.org at EPF Lausanne on November 27, 2026

interaction between SFT and RM ($P = 0.558$) (Fig. 2A). Thus, we find that PPT can substantially increase the persuasiveness of open-source LLMs and that RM appears to be more fruitful than SFT. Notably, applying RM to a small open-source LLM (Llama3.1-8B) increased its persuasive effect from model GPT-4o (August 2024) (see supplementary materials, section 2.4, for full output tables).

Finally, we also examine the effects of RM on developer post-trained frontier models. (Many of these models are closed-source, rendering SFT infeasible.) Specifically, we compare base versus RM-tuned models for GPT-3.5, GPT-4o (August 2024), and GPT-4.5 in study 2 and for GPT-4o (August 2024 and March 2025), GPT-4.5, and Grok-3 in study 3. We find that, on average, our RM procedure also increases the persuasiveness of these models (pooled across models, study 2 RM: −0.08 percentage points, $P = 0.864$; study 3 RM: +0.80 percentage points, $P < 0.001$; precision-weighted average across studies: +0.63 percentage points, $P = 0.003$, relative to an average baseline persuasion effect of 9.8 percentage points; Fig. 2, B and C), although the effect increase is smaller than what we found for the open-source models. This could be a result of models with frontier post-training generating more consistent responses and thus offering less-variable samples for the RM to select between (supplementary materials, section 2.10).

How do models persuade?

Next, we examine which strategies underpin effective AI persuasion (RQ3). First, given widespread concern that AI systems will be able to microtarget their arguments to increase their persuasiveness for specific individuals (*4*, *22*–*25*), we consider the effect of providing the LLM with information about the user (personalization). We test three personalization methods across studies: (i) prompt-based personalization, where participants' initial attitude score (0 to 100 scale) and an open-ended reflection explaining their initial attitude were appended to model prompts; (ii) SFT on personalized data, where models were post-trained on datasets containing participants' attitudes and reflections plus nine pieces of demographics and political information (e.g., age, gender, political ideology, and party affiliation); and (iii) personalized RM, where an RM was trained to select persuasive responses using all of the aforementioned personalization data during both training and inference (see materials and methods for further details). Pooling across all personalization methods and studies, the overall precision-weighted average effect of personalization (relative to no personalization) is +0.43 percentage points (95% CI [0.22, 0.64]) (for a baseline comparison, the corresponding non-personalized persuasive effect is 8.5 percentage points). Furthermore, across individual methods and studies, the effect of personalization is similarly sized and never exceeds +1 percentage point (supplementary materials, section 2.5). Thus, although we do find some evidence of persuasive returns to personalization, they are small compared with the magnitude of the persuasion gains that we document from model scale and post-training.

We then examined how the model's rhetorical strategy affected persuasive success. In each conversation, we randomized the LLM's prompt to instruct it to use one of eight theoretically motivated strategies for persuasion, such as moral reframing (*15*, *29*, *30*), storytelling (*16*, *31*, *32*), deep canvassing (*14*, *33*), and information-based argumentation (in which an emphasis is placed on providing facts and evidence) (*34*, *35*), as well as a basic prompt (only instruction: "Be as persuasive as you can"). The persuasive impact of each strategy, relative to the basic prompt, is shown in Fig. 3A. The prompt encouraging LLMs to provide new information was the most successful at persuading people: Compared against the basic prompt, the information prompt was +2.29 percentage points [1.84, 2.75] more persuasive, whereas the next-best prompt was only +1.37 percentage points [0.92, 1.81] more persuasive than the basic prompt (these are precision-weighted averages across studies; see supplementary materials, section 2.6.1, for breakdown by study). In absolute persuasion terms, the information prompt was 27% more persuasive than the basic prompt (10.60 percentage points versus 8.34 percentage points, $P < 0.001$). Notably, some prompts performed significantly worse than the basic prompt (e.g., moral reframing and deep canvassing; Fig. 3A). This suggests that LLMs may be successful persuaders insofar as they are encouraged to pack their conversation with facts and evidence that appear to support their arguments—that is, to pursue an information-based persuasion mechanism (*35*)—more so than using other psychologically informed persuasion strategies.

To further investigate the role of information in AI persuasion, we combined GPT-4o and professional human fact-checkers to count the number of fact-checkable claims made in the 91,000 persuasive conversations ("information density") (materials and methods). {In a validation test, the counts provided by GPT-4o and human fact-checkers were correlated at correlation coefficient (r) = 0.87, 95% CI [0.84, 0.90]; see materials and methods and supplementary materials, section 2.8, for further details.} As expected, information density is consistently largest under the information prompt relative to the other rhetorical strategies (Fig. 3B). Notably, we find that information density for each rhetorical strategy is, in turn, strongly associated with how persuasive the model is when using that strategy (Fig. 3B), which implies that information-dense AI messages are more persuasive. The average correlation between information density and persuasion is $r = 0.77$ (Bayesian 95% CI [0.09, 0.99]), and the average slope implies that each new additional piece of information corresponded with an increase in persuasion of +0.30 percentage points [0.23, 0.38] (Fig. 3B) (see materials and methods for analysis details).

Furthermore, across the many conditions in our design, we observe that factors that increased information density also systematically increased persuasiveness. For example, the most persuasive models in our sample (GPT-4o March 2025 and GPT-4.5) were at their most

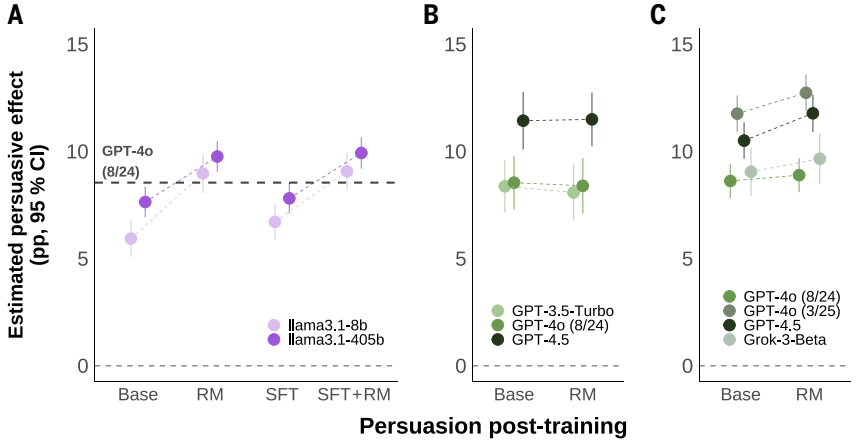

Fig. 2. PPT can substantially increase the persuasiveness of conversational AI. (**A**) Persuasive impact of Llama3.1-8B and Llama3.1-405B models under four conditions: SFT, RM, combined SFT + RM, and Base (no PPT). (**B**) Persuasive impact of Base and RM in study 2. (**C**) Persuasive impact of Base and RM in study 3. All panels show persuasive impact in percentage points (versus control group) with 95% CIs. In (A), Base refers to open-source versions of a model fine-tuned for open-ended dialogue but with no persuasion-specific post-training; in (B) and (C), it refers to unmodified closed-source models deployed out of the box with no additional post-training. Models were prompted with one of a range of persuasion strategies. See materials and methods for training details.

Downloaded from https://www.science.org at EPF Lausanne on November 27, 2026

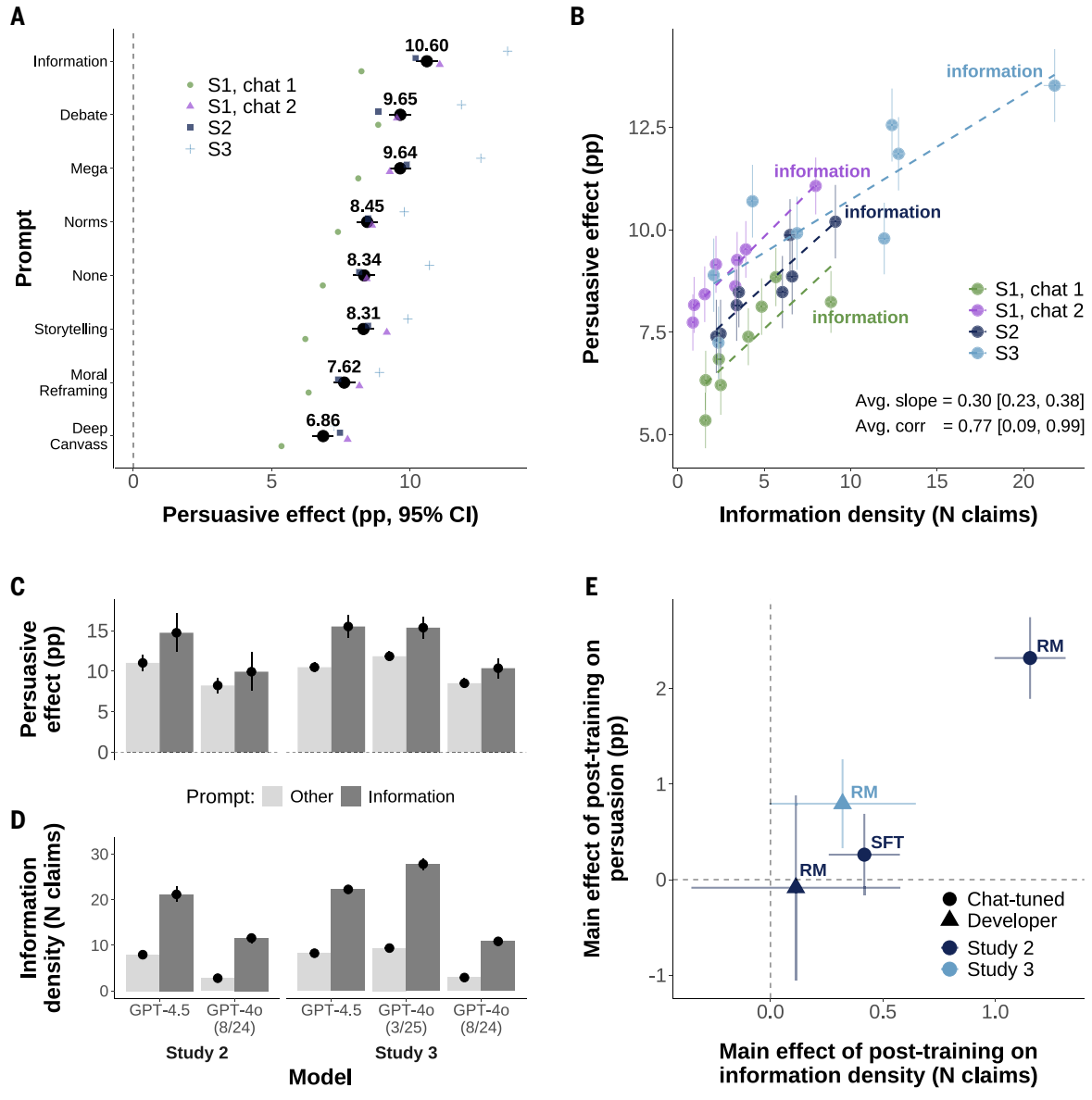

Fig. 3. Persuasion increases with information density. (**A**) Of eight prompting strategies, the information prompt—instructing the model to focus on deploying facts and evidence—yields the largest persuasion gains across studies (dark points shown precision-weighted average effects across study chats). (**B**) Mean policy support and mean information density (number of fact-checkable claims per conversation) for each of our eight prompts in each study chat. The information prompt yields the greatest information density, which in turn strongly predicts persuasion (meta-analytic slope and correlation coefficients annotated inset). (**C**) The persuasive advantage of the most persuasive models (GPT-4o March 2025, GPT-4.5) over GPT-4o (August 2024) is largest when they are information prompted (see supplementary materials, section 2.6.2, for interaction tests). (**D**) Information prompting also causes a disproportionate increase in information density among the most persuasive models (see supplementary materials, section 2.6.2, for interaction tests). (**E**) Main effects of persuasion post-training (versus Base) on both information density and persuasion. Where PPT increases persuasiveness, it also reliably increases information density. In all panels, error bars are 95% CIs.

persuasive when prompted to use information (Fig. 3C). This prompting strategy caused GPT-4o (March 2025) to generate more than 25 fact-checkable claims per conversation on average, compared with <10 for other prompts ($P < 0.001$) (Fig. 3D). Similarly, we find that our RM PPT reliably increased the average number of claims made by our chat-tuned models in study 2 (+1.15 claims, $P < 0.001$; Fig. 3E), where we also found it clearly increased persuasiveness (+2.32 percentage points, $P < 0.001$). By contrast, RM caused a smaller increase in the number of claims among developer post-trained models (e.g., in study 3: +0.32 claims, $P = 0.053$), and it had a correspondingly smaller effect on persuasiveness there (+0.80 percentage points, $P < 0.001$) (Fig. 3E). Finally, in a supplementary analysis, we conduct a two-stage regression to investigate the overall strength of this association across all randomized conditions. We estimate that information density explains 44% of the variability in persuasive effects generated by all of our conditions and 75% when restricting to developer post-trained models (see materials and methods for further details). We find consistent evidence that factors that most increased persuasion—whether through prompting or post-training—tended to also increase information density, which suggests that information density is a key variable driving the persuasive power of current AI conversation.

Downloaded from https://www.science.org at EPF Lausanne on November 27, 2026

How accurate is the information provided by the models?

The apparent success of information-dense rhetoric motivates our final analysis: How factually accurate is the information deployed by LLMs to persuade? To test this, we used a web search–enabled LLM (GPT-4o Search Preview) tasked with evaluating the accuracy of claims (on a 0 to 100 scale) made by AI in the large body of conversations collected across studies 1 to 3. The procedure was independently validated by comparing a subset of its ratings with ratings generated by professional human fact-checkers, which yielded a correlation of $r = 0.84$ (95% CI [0.79, 0.88]) (see materials and methods and supplementary materials, section 2.8, for details).

Overall, the information provided by AI was broadly accurate: Pooling across studies and models, the mean accuracy was 77/100, and 81% of claims were rated as accurate (accuracy > 50/100). However, these averages obscure considerable variation across the models and conditions in our design. In Fig. 4A, we plot the estimated proportion of claims rated as accurate against model scale (in the supplementary materials, section 2.7, we show that the results below are substantively identical if we instead analyze average accuracy on the full 0 to 100 scale). Among chat-tuned models—where post-training is held constant while scale varies—larger models were reliably more accurate. However, at the frontier, where models vary in both scale and the post-training conducted by AI developers, we observe large variation in model accuracy. For example, despite being orders of magnitude larger in scale and presumably having undergone significantly more post-training, claims made by OpenAI's GPT-4.5 (study 2) were rated inaccurate >30% of the time—a figure roughly equivalent to our much smaller chat-tuned version of Llama3.1-8B. Unexpectedly, we also find that GPT-3.5—a model released more than 2 years earlier than GPT-4.5—made ~13 percentage points fewer inaccurate claims (Fig. 4A).

We document another disconcerting result: Although the biggest predictor of a model's persuasiveness was the number of fact-checkable claims (information) that it deployed, we observe that the models with the highest information density also tended to be less accurate on average. First, among the most persuasive models in our sample, the most persuasive prompt—that which encouraged the use of information—significantly decreased the proportion of accurate claims made during conversation (Fig. 4B). For example, GPT-4o (March 2025) made substantially fewer accurate claims when prompted to use information (62%) versus a different prompt (78%; difference test, $P < 0.001$). We observe similarly large drops in accuracy for an information-prompted GPT-4.5 in study 2 (56% versus 70%; $P < 0.001$) and study 3 (72% versus 82%; $P < 0.001$). Second, although applying RM PPT to chat-tuned models increased their persuasiveness (+2.32 percentage points, $P < 0.001$), it also increased their proportion of inaccurate claims (−2.22 percentage points fewer accurate claims, $P < 0.001$) (Fig. 4C). Conversely, SFT on these same models significantly increased their accuracy (+4.89 percentage points, $P < 0.001$) but not their persuasiveness (+0.26 percentage points, $P = 0.230$). Third, we previously showed that new developer post-training on GPT-4o (March 2025 versus August 2024) markedly increased its persuasiveness (+3.50 percentage points, $P < 0.001$; Fig. 1); it also substantially increased its proportion of inaccurate claims (−12.53 percentage points fewer accurate claims, $P < 0.001$; Fig. 4A).

Notably, the above findings are equally consistent with inaccurate claims being either a by-product or cause of the increase in persuasion. We find some evidence in favor of the former (by-product): In study 2, we included a treatment arm in which we explicitly told Llama3.1-405B to use fabricated information (Llama3.1-405B- deceptive-info; Fig. 4A). This increased the proportion of inaccurate claims versus the standard information prompt (+2.51 percentage points, $P = 0.006$) but did not significantly increase persuasion (−0.73 percentage points, $P = 0.157$). Furthermore, across all conditions in our study, we do not find evidence that persuasiveness was positively associated with the number of inaccurate claims after controlling for the total number of claims (see materials and methods for details).

The impact of pulling all the persuasion levers at once

Finally, we examined the effect of a conversational AI designed for maximal persuasion considering all features—or "levers"—examined in our study (model, prompt, personalization, and post-training). This can shed light on the potential implications of the inevitable use of frontier LLMs for political messaging "in the wild," where actors may do whatever they

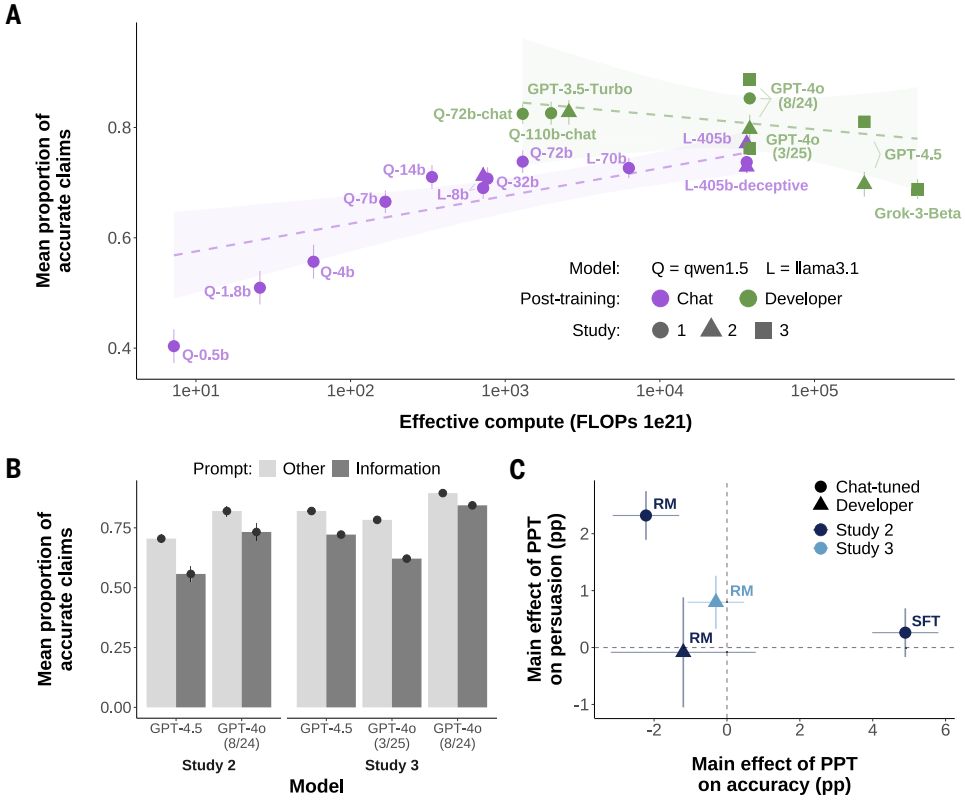

Fig. 4. Factors that made conversational AI more persuasive tended to decrease factual accuracy. (**A**) Proportion of AI claims rated as accurate (>50 on 0 to 100 scale) as a function of model scale. Chat-tuned models (purple) show increasing accuracy with scale, whereas developer post-trained models (green) exhibit high variance despite frontier scale. Notably, GPT-4.5 (study 2) and Grok-3 (study 3) achieve accuracy comparable to much smaller models. Note that some model labels have been removed for clarity. (**B**) The information prompt—the most effective persuasion strategy—causes substantial accuracy decreases relative to other prompts, and disproportionate decreases among the most persuasive models (GPT-4o March 2025 and GPT-4.5) compared with GPT-4o August 2024 (see supplementary materials, section 2.6.2, for interaction tests). (**C**) Main effects of persuasion post-training (versus Base) on both accuracy and persuasion. Where PPT increases persuasiveness, it tends to decrease accuracy. In all panels, error bars are 95% CIs.

Downloaded from https://www.science.org at EPF Lausanne on November 27, 2026

can to maximize persuasion. For this analysis, we used a cross-fit machine learning approach to (i) identify the most persuasive conditions and then (ii) estimate their joint persuasive impact out-of-sample (see materials and methods for details). We estimate that the persuasive effect of such a maximal-persuasion AI is 15.9 percentage points on average (which is 69.1% higher than the 9.4 percentage points average condition that we tested) and 26.1 percentage points among participants who initially disagreed with the issue (74.3% higher than the 15.2 percentage points average condition). These effect sizes are substantively large, even relative to those observed in other recent work on conversational persuasion with LLMs (*36*, *37*). We further estimate that, in these maximal-persuasion conditions, AI made 22.5 fact-checkable claims per conversation (versus 5.6 average) and that 30.0% of these claims were rated inaccurate (versus 16.0% average). Together, these results shed light on the level of persuasive advantage that could be achieved by actors in the real world seeking to maximize AI persuasion under current conditions. They also highlight the risk that AI models designed for maximum persuasion—even without explicitly seeking to misinform—may wind up providing substantial amounts of inaccurate information.

Discussion

Despite widespread concern about AI-driven persuasion (*1*–*13*), the factors that determine the nature and limits of AI persuasiveness have remained unknown. In this work, across three large-scale experiments involving 76,977 responses from 42,357 people, 707 political issues, and 19 LLMs, we systematically examined how model scale and post-training methods may contribute to the persuasiveness of current and future conversational AI systems. Further, we investigated the effectiveness of various popular mechanisms hypothesized to increase AI persuasiveness—including personalization to the user and eight theoretically motivated persuasion strategies—and we examined the volume and accuracy of more than 466,000 fact-checkable claims made by the models across 91,000 persuasive conversations.

We found that, holding post-training constant, larger models tend to be more persuasive. Notably, however, the largest persuasion gains from frontier post-training (+3.50 percentage points between different GPT-4o deployments) exceeded the estimated gains from increasing model scale 10×—or even 100×—beyond the current frontier (+1.59 percentage points and +3.19 percentage points, respectively). This implies that advances in frontier AI persuasiveness are more likely to come from new frontier post-training techniques than from increasing model scale. Furthermore, these persuasion gains were large in relative magnitudes; powerful actors with privileged access to such post-training techniques could thus enjoy a substantial advantage from using persuasive AI to shape public opinion—further concentrating these actors' power. At the same time, we found that subfrontier post-training (in which an RM was trained to predict which messages will be most persuasive) applied to a small open-source model (Llama-8B) transformed it into an as- or more-effective persuader than frontier model GPT-4o (August 2024). Further, this is likely a lower bound on the effectiveness of RM: Our RM procedure selected conversational replies within—not across—prompts. Although this allowed us to isolate additional variance (in the persuasiveness of conversational replies) not accounted for by prompt, it also reduced the variance available in replies for the RM to capitalize on. RM selecting across prompts could likely perform better. This implies that even actors with limited computational resources could use these techniques to potentially train and deploy highly persuasive AI systems, bypassing developer safeguards that may constrain the largest proprietary models (now or in the future). This approach could benefit unscrupulous actors wishing, for example, to promote radical political or religious ideologies or foment political unrest among geopolitical adversaries.

We uncovered a key mechanism driving these persuasion gains: AI models were most persuasive when they packed their dialogue with information—fact-checkable claims potentially relevant to their argument. We found clear evidence that insomuch as factors such as model scale, post-training, or prompting strategy increased the information density of AI messages, they also increased persuasion. Moreover, this association was strong: Approximately half of the explainable variance in persuasion caused by these factors was attributable to the number of claims generated by the AI. The evidence was also consistent across different ways of measuring information density, emerging for both (i) the number of claims made by AI (as counted by LLMs and professional human fact-checkers) and (ii) participants' self-reported perception of how much they learned during the conversation (supplementary materials, section 2.6.3).

Our result, documenting the centrality of information-dense argumentation in the persuasive success of AI, has implications for key theories of persuasion and attitude change. For example, theories of politically motivated reasoning (*38*–*41*) have expressed skepticism about the persuasive role of facts and evidence, highlighting instead the potential of psychological strategies that better appeal to the group identities and psychological dispositions of the audience. As such, scholars have investigated the persuasive effect of various such strategies, including storytelling (*16*, *31*, *32*), moral reframing (*15*, *29*, *30*), deep canvassing (*14*, *33*), and personalization (*4*, *22*–*25*), among others. However, a different body of work instead emphasizes that exposure to facts and evidence is a primary route to political persuasion—even if it cuts against the audience's identity or psychological disposition (*35*, *36*, *42*, *43*). Our results are consistent with fact- and evidence-based claims being more persuasive than these various popular psychological strategies (at least as implemented by current AI), thereby advancing this ongoing theoretical debate over the psychology of political information processing.

Furthermore, our results on this front build on a wider theoretical and empirical foundation of understanding about how people persuade people. Long-standing theories of opinion formation in psychology and political science, such as the elaboration likelihood model (*34*) and the receive-accept-sample model (*44*), posit that exposure to substantive information can be especially persuasive. Moreover, the importance that such theoretical frameworks attach to information-based routes to persuasion is increasingly borne out by empirical work on human-to-human persuasion. For example, recent large-scale experiments support an "informational (quasi-Bayesian) mechanism" of political persuasion: Voters are more persuadable when provided with information about candidates who they know less about, and messages with richer informational content are more persuasive (*42*). Similarly, other experiments have shown that exposure to new information reliably shifts people's political attitudes in the direction of the information, largely independent of their starting beliefs, demographics, or context (*35*, *43*, *45*). Our work advances this prior theoretical and empirical research on human-to-human persuasion by showing that exposure to substantive information is a key mechanism driving successful AI-to-human persuasion. Moreover, the fact that our results are grounded in this prior work increases confidence that the mechanism that we identify will generalize beyond our particular sample of AI models and political issues. Insofar as information density is a key driver of persuasive success, this implies that AI could exceed the persuasiveness of even elite human persuaders, given their ability to generate large quantities of information almost instantaneously during conversation.

Our results also contribute to the ongoing debate over the persuasive impact of AI-driven personalization. Much concern has been expressed about personalized persuasion after the widely publicized claims of microtargeting by Cambridge Analytica in the 2016 European Union (EU) referendum and US presidential election (*46*–*48*). In light of these concerns, there is live scientific debate about the persuasive effect of AI-driven personalization, with scholars emphasizing its outsized power and thus danger (*22*–*24*), whereas others find limited,

Downloaded from https://www.science.org at EPF Lausanne on November 27, 2026

context-dependent, or no evidence of the effect of personalization (*25*, *49*, *50*) and argue that current concerns are overblown (*51*, *52*). Our findings push this debate forward in several ways. First, we examined various personalization methods, from basic prompting [as in prior work, e.g., (*36*)] to more advanced techniques that integrated personalization with model post-training. Second, by using a much larger sample size than past work, we were able to demonstrate a precise significant effect of personalization that is ~+0.5 percentage points on average—thereby supporting the claim that personalization does make AI persuasion more effective [and even that small effects such as this can have important effects at scale; see, for example, (*53*)]. Third, however, we are also able to place this effect of personalization in a crucial context by showing the much larger effect on persuasiveness of other technical and rhetorical strategies that can be implemented by current AI. In addition, given that the success of personalization depends on treatment effect heterogeneity—that is, different people responding in different ways to different messages (*54*)—our findings support theories that assume small amounts of heterogeneity and challenge those that assume large heterogeneity (*35*). So, although our results suggest that personalization can contribute to the persuasiveness of conversational AI, other factors likely matter more.

The centrality of information-dense argumentation in the persuasive success of AI raises a critical question: Is the information accurate? Across all models and conditions, we found that persuasive AI-generated claims achieved reasonable accuracy scores (77/100, where 0 = completely inaccurate and 100 = completely accurate), with only 19% of claims rated as predominantly inaccurate (≤50/100). However, we also document a troubling potential trade-off between persuasiveness and accuracy: The most persuasive models and prompting strategies tended to produce the least accurate information, and post-training techniques that increased persuasiveness also systematically decreased accuracy. Although in some cases these decreases were small (−2.22 percentage points: RM versus base among Llama models), in other cases they were large (−13 percentage points: GPT-4o March 2025 versus GPT-4o August 2024). Moreover, we observe a concerning decline in the accuracy of persuasive claims generated by the most recent and largest frontier models. For example, claims made by GPT-4.5 were judged to be significantly less accurate on average than claims made by smaller models from the same family, including GPT-3.5 and the version of GPT-4o released in the summer of 2024, and were no more accurate than substantially smaller models such as Llama3.1-8B. Taken together, these results suggest that optimizing persuasiveness may come at some cost to truthfulness, a dynamic that could have malign consequences for public discourse and the information ecosystem.

Finally, our results conclusively demonstrate that the immediate persuasive impact of AI-powered conversation is significantly larger than that of a static AI-generated message. This contrasts sharply with the results of recent smaller-scale studies (*55*) and suggests a potential transformation of the persuasion landscape, where actors seeking to maximize persuasion could routinely turn to AI conversation agents in place of static one-way communication. This result also validates the predictions of long-standing theories of human communication that posit that conversation is a particularly persuasive format (*56*–*58*) and extends prior work on scaling AI persuasion by suggesting that conversation could enjoy greater returns to scale compared with static messages (*26*).

What do these results imply for the future of AI persuasion? Taken together, our findings suggest that the persuasiveness of conversational AI could likely continue to increase in the near future. However, several important constraints may limit the magnitude and practical effect of this increase. First, the computational requirements for continued model scaling are considerable: It is unclear whether or how long investments in compute infrastructure will enable continued scaling (*59*–*61*). Second, influential theories of human communication suggest that there are hard psychological limits to human persuadability (*57*, *58*, *62*, *63*); if so, this may limit further gains in AI persuasiveness. Third, and perhaps most relevantly, real-world deployment of AI persuasion faces a critical bottleneck: Although our experiments show that lengthy, information-dense conversations are most effective at shifting political attitudes, the extent to which people will voluntarily sustain cognitively demanding political discussions with AI systems outside of a survey context remains unclear [e.g., owing to lack of awareness or interest in politics and competing demands on attention (*64*–*66*)]. Preliminary work suggests that the very conditions that make conversational AI most persuasive—sustained engagement with information-dense arguments—may also be those most difficult to achieve in the real world (*66*). Thus, although our results show that more capable AI systems may achieve greater persuasive influence under controlled conditions, the upper limit and practical impact of these increases is an important topic for future work.

We note several limitations. First, our sample of participants was a convenience sample and not representative of the UK population. Although this places some constraints on the generalizability of our estimates, we do not believe that these are strong constraints, for several reasons. First, applying census weights in our key analyses to render the sample representative of the UK along age, sex, and education yields substantively identical results as the unweighted analysis (supplementary materials, section 2.3.3). Second, previous work has indicated that treatment effects estimated in survey samples of crowdworkers correlate strongly with those estimated in nationally representative survey samples (*67*, *68*). This suggests that, even if absolute effect sizes do not generalize well, the relative effect sizes of different treatment factors (e.g., prompting, post-training, personalization, etc.) are likely to do so. Third, the sample of participants is just one (albeit important) dimension affecting the generalizability of a study's results. Other important dimensions in our context include, for example, the sample of political issues on which persuasion is happening and the sample of AI models doing the persuasion—and our design incorporates an unusually large and diverse sample of both political issues (700+ spanning a wide breadth of issue areas) and AI models (19 LLMs, spanning various model families and versions) [for further discussion, see (*69*)]. A second limitation is that, although we found that the persuasive effects of various psychological strategies (such as storytelling and deep canvassing) were smaller than instructing the model to deploy information, it is possible that these psychological strategies are at a specific disadvantage when implemented by AI (versus humans)—for example, if people perceive AI as less empathic (*70*). Furthermore, and relatedly, we emphasize that our evidence does not demonstrate that these psychological strategies are less effective in general, but, rather, that they are just less effective as implemented by the LLMs in our context. A third limitation is that some recent work has suggested that LLMs are already experiencing diminishing returns from model scaling (*26*); thus, the observed effect of model scale on persuasiveness may well have been more pronounced in earlier generations of LLMs and may increase in magnitude as new architectures emerge.

Our findings clarify where the real levers of AI persuasiveness lie—and where they do not. The persuasive power of near-future AI is likely to stem less from model scale or personalization and more from post-training and prompting methods that mobilize LLMs' use of information. As both frontier and subfrontier models grow more capable, ensuring that this power is used responsibly will be a critical challenge.

Materials and methods are available in the supplementary materials.

REFERENCES AND NOTES

1. F. Luciano, *Hypersuasion – On AI's Persuasive Power and How to Deal with It*. *Philos. Technol.* **37**, 64 (2024). doi: 10.1007/s13347-024-00756-6
2. M. Burtell, T. Woodside, Artificial Influence: An Analysis Of AI-Driven Persuasion. arXiv:2303.08721 [cs.CY] (2023).

Downloaded from https://www.science.org at EPF Lausanne on November 27, 2026

3. C. R. Jones, B. K. Bergen, Lies, Damned Lies, and Distributional Language Statistics: Persuasion and Deception with Large Language Models. arXiv:2412.17128 [cs.CL] (2024).
4. A. Rogiers, S. Noels, M. Buyl, T. De Bie, Persuasion with Large Language Models: A Survey. arXiv:2411.06837 [cs.CL] (2024).
5. S. El-Sayed *et al.*, A Mechanism-Based Approach to Mitigating Harms from Persuasive Generative AI. arXiv:2404.15058 [cs.CY] (2024).
6. K. Grace *et al.*, Thousands of AI Authors on the Future of AI. arXiv:2401.02843 [cs.CY] (2025).
7. J. Nosta, "AI's Superhuman Persuasion," *Psychology Today*, 27 October 2023; https://www.psychologytoday.com/intl/blog/the-digital-self/202310/ais-superhuman-persuasion.
8. Y. Bengio *et al.*, Managing extreme AI risks amid rapid progress. *Science* **384**, 842–845 (2024). doi: 10.1126/science.adn0117; pmid: 38768279
9. T. Hsu, S. A. Thompson, "Disinformation Researchers Raise Alarms About A.I. Chatbots," *New York Times*, 8 February 2023; https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html.
10. J. A. Goldstein *et al.*, Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. arXiv:2301.04246 [cs.CY] (2023).
11. L. MacKenzie, M. Scott, "How people view AI, disinformation and elections — In charts," *POLITICO*, 16 April 2024; https://www.politico.eu/article/people-view-ai-disinformation-perception-elections-charts-openai-chatgpt/.
12. A. Dudding, "Global Views on AI and Disinformation," *Ipsos*, 19 November 2023; https://www.ipsos.com/en-nz/global-views-ai-and-disinformation.
13. E. Durmus *et al.*, "Measuring the Persuasiveness of Language Models," *Anthropic*, 9 April 2024; https://www.anthropic.com/news/measuring-model-persuasiveness.
14. D. Broockman, J. Kalla, Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science* **352**, 220–224 (2016). doi: 10.1126/science.aad9713; pmid: 27124458
15. J. L. Kalla, A. S. Levine, D. E. Broockman, Personalizing Moral Reframing in Interpersonal Conversation: A Field Experiment. *J. Polit.* **84**, 1239–1243 (2022). doi: 10.1086/716944
16. J. L. Kalla, D. E. Broockman, Reducing Exclusionary Attitudes through Interpersonal Conversation: Evidence from Three Field Experiments. *Am. Polit. Sci. Rev.* **114**, 410–425 (2020). doi: 10.1017/S0003055419000923
17. L. Ouyang *et al.*, "Training language models to follow instructions with human feedback" in *Advances in Neural Information Processing Systems*, S. Koyejo *et al.*, Eds. (Curran Associates, Inc., 2022), pp. 27730–27744.
18. A. Lewkowycz *et al.*, "Solving Quantitative Reasoning Problems with Language Models" in *Advances in Neural Information Processing Systems*, S. Koyejo *et al.*, Eds. (Curran Associates, Inc., 2022), pp. 3843–3857.
19. J. Wei *et al.*, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models" in *Advances in Neural Information Processing Systems*, S. Koyejo *et al.*, Eds. (Curran Associates, Inc., 2022), pp. 24824–24837.
20. J. A. Goldstein, J. Chao, S. Grossman, A. Stamos, M. Tomz, How persuasive is AI-generated propaganda? *PNAS Nexus* **3**, pgae034 (2024). doi: 10.1093/pnasnexus/pgae034
21. M. Wack, C. Ehrett, D. Linvill, P. Warren, Generative propaganda: Evidence of AI's impact from a state-backed disinformation campaign. *PNAS Nexus* **4**, pgaf083 (2025). doi: 10.1093/pnasnexus/pgaf083
22. F. Salvi, M. H. Ribeiro, R. Gallotti, R. West, On the conversational persuasiveness of GPT-4. *Nat. Hum. Behav.* **9**, 1645–1653 (2025). doi: 10.1038/s41562-025-02194-6
23. S. C. Matz *et al.*, The potential of generative AI for personalized persuasion at scale. *Sci. Rep.* **14**, 4692 (2024). doi: 10.1038/s41598-024-53755-0; pmid: 38409168
24. A. Simchon, M. Edwards, S. Lewandowsky, The persuasive effects of political microtargeting in the age of generative artificial intelligence. *PNAS Nexus* **3**, pgae035 (2024). doi: 10.1093/pnasnexus/pgae035
25. K. Hackenburg, H. Margetts, Evaluating the persuasive influence of political microtargeting with large language models. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2403116121 (2024). doi: 10.1073/pnas.2403116121
26. K. Hackenburg *et al.*, Scaling language model size yields diminishing returns for single-message political persuasion. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2413443122 (2025). doi: 10.1073/pnas.2413443122; pmid: 40053360
27. J. Kaplan *et al.*, Scaling Laws for Neural Language Models. arXiv:2001.08361 [cs.LG] (2020).
28. N. Ding *et al.*, Enhancing Chat Language Models by Scaling High-quality Instructional Conversations. arXiv:2305.14233 [cs.CL] (2023).
29. J. G. Voelkel, M. Feinberg, Morally Reframed Arguments Can Affect Support for Political Candidates. *Soc. Psychol. Personal. Sci.* **9**, 917–924 (2017). doi: 10.1177/1948550617729408; pmid: 30595808
30. M. Feinberg, R. Willer, Moral reframing: A technique for effective and persuasive communication across political divides. *Soc. Personal. Psychol. Compass* **13**, e12501 (2019). doi: 10.1111/spc3.12501
31. M. C. Green, T. C. Brock, The role of transportation in the persuasiveness of public narratives. *J. Pers. Soc. Psychol.* **79**, 701–721 (2000). doi: 10.1037/0022-3514.79.5.701; pmid: 11079236
32. A. Hamby, D. Brinberg, K. Daniloski, Reflecting on the journey: Mechanisms in narrative persuasion. *J. Consum. Psychol.* **27**, 11–22 (2017). doi: 10.1016/j.jcps.2016.06.005
33. E. Santoro, D. E. Broockman, J. L. Kalla, R. Porat, Listen for a change? A longitudinal field experiment on listening's potential to enhance persuasion. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2421982122 (2025). doi: 10.1073/pnas.2421982122; pmid: 39977324
34. R. E. Petty, J. T. Cacioppo, "The Elaboration Likelihood Model of Persuasion" in *Communication and Persuasion: Central and Peripheral Routes to Attitude Change*, R. E. Petty, J. T. Cacioppo, Eds. (Springer, 1986), pp. 1–24.
35. A. Coppock, *Persuasion in Parallel: How Information Changes Minds about Politics* (Univ. Chicago Press, 2022).
36. T. H. Costello, G. Pennycook, D. G. Rand, Durably reducing conspiracy beliefs through dialogues with AI. *Science* **385**, eadq1814 (2024). doi: 10.1126/science.adq1814; pmid: 39264999
37. P. Schoenegger *et al.*, Large Language Models Are More Persuasive Than Incentivized Human Persuaders. arXiv:2505.09662 [cs.CL] (2025).
38. D. M. Kahan, Ideology, motivated reasoning, and cognitive reflection. *Judgm. Decis. Mak.* **8**, 407–424 (2013). doi: 10.1017/S1930297500005271
39. D. M. Kahan, "The Politically Motivated Reasoning Paradigm, Part 1: What Politically Motivated Reasoning Is and How to Measure It" in *Emerging Trends in the Social and Behavioral Sciences*, R. A. Scott, S. M. Kosslyn, Eds. (Wiley, 2016), pp. 1–16.
40. C. S. Taber, M. Lodge, Motivated Skepticism in the Evaluation of Political Beliefs. *Am. J. Pol. Sci.* **50**, 755–769 (2006). doi: 10.1111/j.1540-5907.2006.00214.x
41. J. J. Van Bavel, A. Pereira, The Partisan Brain: An Identity-Based Model of Political Belief. *Trends Cogn. Sci.* **22**, 213–224 (2018). doi: 10.1016/j.tics.2018.01.004; pmid: 29475636
42. D. E. Broockman, J. L. Kalla, When and Why Are Campaigns' Persuasive Effects Small? Evidence from the 2020 U.S. Presidential Election. *Am. J. Pol. Sci.* **67**, 833–849 (2023). doi: 10.1111/ajps.12724
43. B. M. Tappin, A. J. Berinsky, D. G. Rand, Partisans' receptivity to persuasive messaging is undiminished by countervailing party leader cues. *Nat. Hum. Behav.* **7**, 568–582 (2023). doi: 10.1038/s41562-023-01551-7; pmid: 36864137
44. J. R. Zaller, *The Nature and Origins of Mass Opinion* (Cambridge Univ. Press, 1992).
45. A. Coppock, S. J. Hill, L. Vavreck, The small effects of political advertising are small regardless of context, message, sender, or receiver: Evidence from 59 real-time randomized experiments. *Sci. Adv.* **6**, eabc4046 (2020). doi: 10.1126/sciadv.abc4046; pmid: 32917601
46. C. Cadwalladr, "The great British Brexit robbery: How our democracy was hijacked," *The Guardian*, 7 May 2017; https://www.theguardian.com/technology/2017/may/07/the-great-british-brexit-robbery-hijacked-democracy.
47. M. Hu, Cambridge Analytica's black box. *Big Data Soc.* **7**, 2053951720938091 (2020). doi: 10.1177/2053951720938091
48. M. Scott, "Cambridge Analytica helped 'cheat' Brexit vote and US election, claims whistleblower," *POLITICO*, 27 March 2018; https://www.politico.eu/article/cambridge-analytica-chris-wylie-brexit-trump-britain-data-protection-privacy-facebook/.
49. E. D. Hersh, B. F. Schaffner, Targeted Campaign Appeals and the Value of Ambiguity. *J. Polit.* **75**, 520–534 (2013). doi: 10.1017/S0022381613000182
50. B. M. Tappin, C. Wittenberg, L. B. Hewitt, A. J. Berinsky, D. G. Rand, Quantifying the potential persuasive returns to political microtargeting. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2216261120 (2023). doi: 10.1073/pnas.2216261120; pmid: 37307486
51. F. M. Simon, S. Altay, "Don't Panic (Yet): Assessing the Evidence and Discourse Around Generative AI and Elections," Knight First Amendment Institute (2025); https://knightcolumbia.org/content/dont-panic-yet-assessing-the-evidence-and-discourse-around-generative-ai-and-elections.
52. F. M. Simon, S. Altay, H. Mercier, Misinformation reloaded? Fears about the impact of generative AI on misinformation are overblown. *Harv. Kennedy Sch. Misinformation Rev.* **4**, 1–11 (2023). doi: 10.37016/mr-2020-127
53. D. C. Funder, D. J. Ozer, Evaluating Effect Size in Psychological Research: Sense and Nonsense. *Adv. Methods Pract. Psychol. Sci.* **2**, 156–168 (2019). doi: 10.1177/2515245919847202
54. L. B. Hewitt, B. M. Tappin, Rank-heterogeneous effects of political messages: Evidence from randomized survey experiments testing 59 video treatments. PsyArXiv (2022); https://doi.org/10.31234/osf.io/xk6t3.
55. L. P. Argyle *et al.*, Testing theories of political persuasion using AI. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2412815122 (2025). doi: 10.1073/pnas.2412815122; pmid: 40314974
56. S. Altay *et al.*, Scaling up interactive argumentation by providing counterarguments with a chatbot. *Nat. Hum. Behav.* **6**, 579–592 (2022). doi: 10.1038/s41562-021-01271-w; pmid: 35165435
57. H. Mercier, *Not Born Yesterday: The Science of Who We Trust and What We Believe* (Princeton Univ. Press, 2020).
58. H. Mercier, D. Sperber, *The Enigma of Reason* (Harvard Univ. Press, 2018).
59. Epoch AI, Machine Learning Trends (2023); https://epoch.ai/trends.

Downloaded from https://www.science.org at EPF Lausanne on November 27, 2026

60. J. Sevilla *et al.*, Can AI Scaling Continue Through 2030? (Epoch AI, 2024); https://epoch.ai/blog/can-ai-scaling-continue-through-2030.
61. K. F. Pilz, J. Sanders, R. Rahman, L. Heim, Trends in AI Supercomputers. arXiv:2504.16026 [cs.CY] (2025).
62. H. Mercier, How Gullible are We? A Review of the Evidence from Psychology and Social Science. *Rev. Gen. Psychol.* **21**, 103–122 (2017). doi: 10.1037/gpr0000111
63. D. Sperber *et al.*, Epistemic Vigilance. *Mind Lang.* **25**, 359–393 (2010).
64. M. Prior, *Post-Broadcast Democracy: How Media Choice Increases Inequality in Political Involvement and Polarizes Elections* (Cambridge Univ. Press, 2007).
65. M. X. Delli Carpini, S. Keeter, *What Americans Know about Politics and Why It Matters* (Yale Univ. Press, 1996).
66. Z. Chen *et al.*, A Framework to Assess the Persuasion Risks Large Language Model Chatbots Pose to Democratic Societies. arXiv:2505.00036 [cs.CL] (2025).
67. A. Coppock, Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach. *Political Sci. Res. Methods* **7**, 613–628 (2019). doi: 10.1017/psrm.2018.10
68. K. J. Mullinix, T. J. Leeper, J. N. Druckman, J. Freese, The Generalizability of Survey Experiments. *J. Exp. Political Sci.* **2**, 109–138 (2015). doi: 10.1017/XPS.2015.19
69. T. Yarkoni, The Generalizability Crisis. *Behav. Brain Sci.* **45**, e1 (2022). doi: 10.1017/S0140525X20001685
70. M. Rubin *et al.*, Comparing the value of perceived human versus AI-generated empathy. *Nat. Hum. Behav.* 10.1038/s41562-025-02247-w (2025). doi: 10.1038/s41562-025-02247-w; pmid: 40588597
71. Epoch AI, Data on AI Models (2024); https://epoch.ai/data/ai-models.
72. B. M. Tappin, Data and Analysis Code, *OSF* (2025). doi: 10.5525/gla.researchdata.1612

ACKNOWLEDGMENTS

D.G.R. and C.S. are co–senior authors. The authors acknowledge the use of resources provided by the Isambard-AI National AI Research Resource (AIRR). Isambard-AI is operated by the University of Bristol and is funded by the UK government's Department for Science, Innovation and Technology (DSIT), through UK Research and Innovation, and the Science and Technology Facilities Council (ST/AIRR/I-A-I/1023). For help during data collection, we thank L. Evans, M. Lee, S. Jones, and A. Price from Prolific. **Funding:** This study was supported by Leverhulme Trust Early Career Research Fellowship ECF-2022-244 (B.M.T.) and the UK Department for Science, Innovation and Technology. **Author contributions:** Conceptualization: K.H., B.M.T., D.G.R., C.S.; Data analysis: K.H., B.M.T., L.H.; Experiment design: K.H., B.M.T., L.H., D.G.R., C.S.; Model hosting: E.S., H.L.; Model training: K.H., L.H., S.B.; Project support: C.F.; Visualization: K.H., B.M.T.; Writing – original draft: K.H., B.M.T.; Writing – review & editing: K.H., B.M.T., L.H., H.L., H.M., D.G.R., C.S. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All aggregated data and analysis code necessary to reproduce the results are available in our project repository on GitHub (https://github.com/kobihackenburg/scaling-conversational-AI) and on the Open Science Framework (*72*). Raw human-AI conversation logs are not publicly available owing to privacy protections; please see the project repository on GitHub for up-to-date information about data access. **License information:** Copyright © 2025 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.aea3884
Materials and Methods; Supplementary Text; Figs. S1 to S10; Tables S1 to S170; References (*73*–*87*); MDAR Reproducibility Checklist

Submitted 7 July 2025; accepted 2 October 2025

10.1126/science.aea3884

Downloaded from https://www.science.org at EPF Lausanne on November 27, 2026

Corrected (4 February 2026): The authors have corrected the main text to clarify their descriptions of the sample size for this study. The distinction between the total number of responses versus the number of participants was originally reported only in the supplementary materials; this distinction is now clearly stated in the main text as well. Because some individuals participated in multiple experiments, references to "participant" numbers in four places (the Rationale section of the Research Article Summary, the Abstract in the main text, the third paragraph of the main text, and the first paragraph of the Discussion section of the main text) have been updated to distinguish between the total number of responses (76,977) and the number of individuals (42,357), rather than referring to 76,977 "participants." The scientific conclusions of the article are unaffected by these changes.

Science
AAAS

The levers of political persuasion with conversational artificial intelligence

Kobi Hackenburg, Ben M. Tappin, Luke Hewitt, Ed Saunders, Sid Black, Hause Lin, Catherine Fist, Helen Margetts, David G. Rand, and Christopher Summerfield

Science **390** (6777), eaea3884. DOI: 10.1126/science.aea3884

Editor's summary

Many fear that we are on the precipice of unprecedented manipulation by large language models (LLMs), but techniques driving their persuasiveness are poorly understood. In the initial "pretrained" phase, LLMs may exhibit flawed reasoning. Their power unlocks during vital "posttraining," when developers refine pretrained LLMs to sharpen their reasoning and align with users' needs. Posttraining also enables LLMs to maintain logical, sophisticated conversations. Hackenburg *et al.* examined which techniques made diverse, conversational LLMs most persuasive across 707 British political issues (see the Perspective by Argyle). LLMs were most persuasive after posttraining, especially when prompted to use facts and evidence (information) to argue. However, information-dense LLMs produced the most inaccurate claims, raising concerns about the spread of misinformation during rollouts of future models. —Ekeoma Uzogara

View the article online

https://www.science.org/doi/10.1126/science.aea3884

Permissions

https://www.science.org/help/reprints-and-permissions

Downloaded from https://www.science.org at EPF Lausanne on November 27, 2026

Use of this article is subject to the Terms of service

Science (ISSN 1095-9203) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science* is a registered trademark of AAAS.

Copyright © 2025 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works