Negative Belief Shifts after LLM-Written Arguments and the Limits of Lexical Cues: A Reanalysis of Two Public Benchmarks

Onur Aksoy
Independent Researcher, Ankara, Türkiye
ORCID: 0009-0001-4481-1759 Correspondence: dev@onuraksoy.com.tr

Abstract

Comparisons of human and LLM persuasion usually report average attitude change. We reanalyse the Anthropic persuasion dataset (3,939 ratings of 1,313 arguments on 56 policy claims, written by people or by five Claude models) at the level of arguments, because each argument was rated three times. On average, Claude 3 Opus arguments shifted support by 0.48 Likert points and human arguments by 0.54; the difference was not significant ($p = 0.37$), and equivalence held only for a ± 0.20 margin. The distributions differed sharply. Of 1,120 Claude arguments, 221 moved at least one reader away from the advocated position, against 1 of 174 human arguments. The gap appeared in 54 of the 56 claims (Cochran–Mantel–Haenszel $p = 6 \times 10^{-10}$), with matched starting attitudes and similar rates across all five models. We then asked whether word-level cues often recommended for persuasive and search-visible writing (statistics, research terms, certainty, inclusive language) explain either outcome. They did not predict negative shifts, their associations with average change largely disappeared once prompt type and initial attitude were controlled, and neither word lists nor a full bag-of-words model predicted change on unseen claims (out-of-fold $R^2 \leq 0.003$). In GEO-bench (4,996 Google top-five pages for 1,000 queries), the same cues varied little with rank and mainly reflected query intent. Average persuasion scores can hide frequent opposite-direction responses to LLM text, and these responses are not visible in simple word counts.

Keywords: persuasion; large language models; negative attitude change; belief change; lexical features; equivalence testing; generative engine optimization

1 Introduction

Studies with thousands of participants now show that text written by large language models (LLMs) can move people's views on policy questions about as much as text written by people [1, 2, 3, 4]. A natural next question is what in the text does the work. Practical guidance has ready answers: add numbers and references to research, use an authoritative register, address the reader as part of a shared "we" [5]. Generative engine optimization (GEO) makes a parallel claim for search, reporting that adding statistics, citations and quotations raises a source's visibility in LLM-generated answers [6].

Most of these comparisons summarise persuasion as an average change in support. An average can hide a mixture of readers who move towards a position and readers who move away from it, and the second group matters when LLM text is deployed at scale. The original report on the Anthropic persuasion dataset describes average effects and does not discuss movement away from the advocated position [4].

The cue recommendations concern surface features that can be counted, and counting them is easy. Interpreting the counts is harder, for three reasons. First, the same LLM produces very different text depending on how it is prompted, so a cue may correlate with persuasion only because a particular prompt produced both the cue and a stronger argument. Second, benchmark datasets often contain repeated ratings of the same text, and treating these as independent observations

inflates precision. Third, a corpus of documents that a search engine has already ranked tells us about the pages that were retrieved, not about the pages that were not.

We reanalyse two public benchmarks with these problems in mind. Corpus A is the Anthropic persuasion dataset [4], in which crowd workers rated their support for a claim before and after reading one argument. Corpus B is the test split of GEO-bench [6], which pairs search queries with the top five Google results. The two corpora have different outcomes, but the question is the same: do the cues that optimization advice recommends vary with the outcome of interest, once the structure of the data is respected? We ask:

RQ1 How do average belief change and text profiles differ between human-written arguments and arguments from five Claude models?

RQ2 How often do human and Claude arguments move readers away from the advocated position, and does this depend on claim, model, prompt or text features?

RQ3 Which lexical associations with belief change survive control for prompt type and initial attitude, and how well does text predict belief change on claims not seen in training?

RQ4 In GEO-bench, do the same cues vary with Google rank, or mainly with query intent?

2 Background

2.1 Persuasion cues

Dual-process accounts separate careful evaluation of arguments from reliance on cues such as source credibility or emotional tone [7, 8]. Source credibility has been studied since the 1950s [9, 10]; framing outcomes as losses changes choices [11, 12]; and compliance research lists liking, authority and social proof among recurring influence principles [13, 5]. Messages that feel pushy can also provoke reactance and move people away from the advocated position [14]. Neuroscientific work has linked persuasion and valuation to specific brain systems [15, 16], but a text corpus cannot measure these processes; word counts describe the message, not how it was processed [17].

2.2 LLMs as persuaders

People respond socially to computers even when they know the program is simple [18, 19], and attribute human traits to agents with human-like cues [20, 21]. LLMs trained with human feedback tend to agree with users and mirror their views [22, 23]. In controlled studies, LLM-written messages have persuaded people on health [24], misinformation [25], propaganda [26], political microtargeting [27], policy [1] and conspiracy beliefs [28], and in live debates and voter conversations [2, 3]. Gains from model size in single messages appear to level off [29], while post-training and prompting methods matter considerably in conversation [30]. Personalised psychological targeting [31] and concerns about manipulation [32] make it practically important to know which message features matter.

2.3 Generative search

Aggarwal et al. [6] introduced GEO-bench and showed that editing sources, for example by adding statistics or citations, changed how often an LLM-based answer engine cited them. Their sources are the top five Google results for each query. This design lets us ask a simpler, descriptive question: do the recommended cues already differ between higher- and lower-ranked pages in the pool that the engine draws on?

3 Data and methods

3.1 Datasets

Corpus A. The Anthropic/persuasion dataset [4, 33] contains 3,939 ratings by 3,832 crowd workers of 1,313 unique arguments. Each argument was rated by exactly three workers. Arguments address 56 contested policy claims and 19 control claims. For each policy claim, the dataset contains arguments written by people (174 arguments, 522 ratings) and by five Claude models (Claude 1.3, Claude Instant 1.2, Claude 2, Claude 3 Haiku and Claude 3 Opus; 224 arguments and 672 ratings each). Model arguments were generated with four prompts (Expert Writer Rhetorics, Compelling Case, Logical Reasoning and Deceptive; 280 arguments each). According to the original report, human arguments were written by participants randomly assigned to a claim (three per claim; two claims have six), who were asked for about 250 words and offered a bonus for the most persuasive argument; no selection or editing of the human arguments is described [4]. Workers reported support for the claim on a 7-point scale before and after reading; the outcome is the difference (final minus initial).

Corpus B. We used the test split of GEO-bench [6, 34] (1,000 queries; the dataset also has 8,000 training and 1,000 validation queries). Each query comes with up to five Google results (4,996 pages: 1,000 each at ranks 1–3, 999 at rank 4 and 997 at rank 5); we used the cleaned page text. Queries carry several tags each. We assigned a query to *Transactional* if it was tagged "transactional" (97 queries), otherwise to *Opinion/Debate* if tagged "opinion" or "debate" (221), and otherwise to *Informational* (682). No query had both a transactional and an opinion or debate tag. Page domains were grouped into institutional (.gov, .edu, .ac, .int, .mil), encyclopedic and dictionary sites, scholarly publishers, e-commerce retailers and other sites; the lists are in the code.

3.2 Text features

We counted words from nine non-overlapping word lists, each expressed per 100 words. A token matched a list if it equalled a listed word or that word plus "s" or "es".

- *Logical connectives*: because, therefore, consequently, thus, hence, since, whereas, although, furthermore, moreover.
- *Evidential nouns*: evidence, research, data, study, studies, findings, analysis, report, reported, statistics, survey.
- *Hedges and qualifiers*: may, might, could, often, suggests, likely, tends, however, potential, partially, sometimes, uncertain, nuanced, typically, generally.
- *Collective-reference terms*: we, our, us, ourselves, everyone, society, people, citizens, humanity, community, communities, together. As a sensitivity check we also counted the pronouns *we, our, us, ourselves* alone.
- *Affective terms*: feel, feeling, hope, fear, pain, care, trust, compassion, empathy, family, children, dignity.
- *Loss/threat terms*: risk, threat, danger, jeopardize, harm, crisis, destroy, catastrophic, vulnerable, exploit.
- *Certainty terms*: must, should, undoubtedly, clearly, always, never, essential, crucial, proven, inevitable.
- *Authority terms*: university, journal, institute, researchers, scientists, according, federal, clinical, official, regulatory.
- *Commercial terms*: buy, purchase, price, pricing, discount, sale, shop, shipping, affordable, guarantee, coupon, retail, checkout.

We also counted numeric tokens (numbers, percentages, amounts), citation markers (URLs, DOIs and bracketed numbers), question marks, argument length, the standard deviation of sentence length in words, Guiraud's index (types divided by the square root of tokens) and an approximate Flesch reading-ease score [37]. Two summary indices were formed for Corpus A: an affective/relational index (collective + affective + certainty + loss/threat) and an analytical index (connectives + evidential + numeric + hedges). These lists are a transparent operationalisation rather than validated dictionaries such as LIWC [17].

3.3 Statistical analysis

Unit of analysis. Because every argument in Corpus A was rated three times, we used the argument (mean of its three ratings) as the unit for all comparisons of belief change between sources. Confidence intervals for means and rates were obtained by resampling arguments (10,000 bootstrap draws). Trial-level Wilson intervals [38] are reported for descriptive negative-shift rates.

Comparisons. Groups were compared with two-sided Mann–Whitney tests [39] and Cliff's δ [40]; more than two groups with Kruskal–Wallis tests [41]. Within each family of tests we controlled the false discovery rate with the Benjamini–Hochberg procedure [42] and report q values. Balance in initial attitudes was checked with a two-sample Kolmogorov–Smirnov test.

Negative shifts. An argument was counted as producing a negative shift if at least one of its three ratings decreased. Human and Claude arguments were compared with Fisher's exact test, with a Cochran–Mantel–Haenszel test stratified by claim [35], and with a permutation test that shuffled the human/Claude label among arguments of the same claim (20,000 permutations). Differences among models and prompts were tested with χ^2 tests. To ask whether text features predict negative shifts, we fitted a logistic generalised estimating equation (GEE) [46] to Claude ratings whose reader could move down (initial rating ≥ 2), with the ten standardised text features, initial rating, prompt and model as predictors and an exchangeable correlation within arguments.

Equivalence. To test whether Claude 3 Opus and human arguments were similarly persuasive, we used two one-sided tests (TOST) [43, 44] on argument means (Welch standard errors). No smallest effect size of interest was fixed in advance, so we report results for margins of ± 0.10, ± 0.15 and ± 0.20 Likert points (about 0.11, 0.17 and 0.22 standard deviations of the rating change across human and Claude ratings), together with a version that pairs Opus and human arguments by claim.

Lexical associations. For the 1,120 unique Claude arguments, we computed Spearman correlations [45] between features and mean belief change, first pooled and then within each prompt type ($n = 280$ each). We then fitted ordinary least squares models to the 3,360 Claude ratings with standardised features, standardised initial rating and prompt-type fixed effects, using cluster-robust standard errors by argument [46]. A second model added claim fixed effects.

Prediction. We predicted belief change for the 3,882 human and Claude ratings with five-fold cross-validation grouped by claim, so that no claim appeared in both training and test folds. We compared ridge regression [47] ($\alpha = 50$) with histogram-based gradient boosting [48, 49] on three input sets: text features only, initial rating only, and both. For the binary outcome "positive shift" we report the area under the ROC curve [50], using logistic regression and gradient-boosting classifiers. To check that the results do not depend on our word lists, we repeated the prediction with a bag-of-words representation of the full argument text (TF-IDF weighted unigrams and bigrams occurring in at least three arguments [36]; ridge regression with the penalty chosen inside each training fold, and logistic regression for positive and negative shifts).

Corpus B. Ranks were compared within queries with Friedman tests [51] on the 996–997 queries with five pages, with Kendall's W as the effect size; rank 1 and rank 5 were also compared with Wilcoxon signed-rank tests [52]. Intent groups were compared with Mann–Whitney tests on query-level means, so that each query counts once.

Software and AI assistance. Analyses used Python with pandas, SciPy [53], statsmodels [54] and scikit-learn [55]; the random seed was fixed. Calculations were carried out with the help

Table 1: Belief change and text profiles by source (Corpus A). Mean shift intervals resample arguments; negative-shift intervals are trial-level Wilson intervals. Text indices are means over unique arguments (per 100 words; sentence-length SD in words).

Source	Ratings / args	Mean shift [95% CI]	Negative shifts, % [Wilson 95% CI]	Args with ≥1 neg.	Aff./rel.	Analytic	Sent. SD
Control	57 / 19	0.070 [0.00, 0.18]	1 (1.75%) [0.31, 9.29]	1	3.54	2.02	5.98
Claude 1.3	672 / 224	0.344 [0.27, 0.42]	57 (8.48%) [6.60, 10.83]	49	4.03	1.82	6.49
Claude Instant 1.2	672 / 224	0.335 [0.27, 0.41]	53 (7.89%) [6.08, 10.17]	50	4.70	1.96	6.58
Claude 2	672 / 224	0.424 [0.36, 0.50]	46 (6.85%) [5.17, 9.01]	43	4.90	1.78	5.61
Claude 3 Haiku	672 / 224	0.435 [0.37, 0.50]	38 (5.65%) [4.15, 7.67]	37	3.58	1.95	6.75
Claude 3 Opus	672 / 224	0.482 [0.40, 0.56]	46 (6.85%) [5.17, 9.01]	42	4.00	1.78	7.21
Human	522 / 174	0.540 [0.46, 0.63]	1 (0.19%) [0.03, 1.08]	1	3.29	1.83	8.97

of Gemini Carbon (Google) and cross-checked with Claude Opus 5.5 (Anthropic). The author reviewed the code and all reported values.

4 Results

4.1 RQ1: Human and Claude arguments

Mean belief change rose from 0.34 for Claude 1.3 and Claude Instant 1.2 to 0.48 for Claude 3 Opus (Table 1, Figure 1). The five models differed (Kruskal–Wallis on argument means, $H = 12.36$, $p = 0.015$), and Opus arguments were more persuasive than Claude 1.3 arguments ($\delta = 0.13$, $q = 0.027$). The ordering does not follow release date: Claude 3 Haiku (0.43) was close to Claude 2 (0.42).

Human arguments produced a mean change of 0.54. Opus arguments were not detectably different ($\delta = -0.05$, $p = 0.37$), whereas the five Claude models taken together were somewhat less persuasive than people (0.40 vs. 0.54; $\delta = -0.12$, $q = 0.024$). Absence of a significant difference is not evidence of similarity, so we tested equivalence directly. The Opus–human difference was −0.058 (90% CI −0.159 to 0.042). Equivalence was supported for a ±0.20 margin ($p_{\mathrm{TOST}} = 0.010$) but not for ±0.15 ($p = 0.066$) or ±0.10 ($p = 0.25$); pairing by claim gave the same pattern ($p = 0.014$, 0.085 and 0.30). Treating the 1,194 individual ratings as independent would have suggested equivalence at ±0.15 ($p = 0.042$), which illustrates how repeated ratings can overstate precision.

The texts also differed. Compared with human arguments, unique Claude arguments had less variable sentence length (SD 6.53 vs. 8.97 words; $\delta = -0.53$, $q = 1.2 \times 10^{-28}$), about half as many logical connectives (0.14 vs. 0.29 per 100 words; $\delta = -0.23$), about twice as many evidential nouns (0.43 vs. 0.21; $\delta = 0.26$), more loss/threat terms (0.71 vs. 0.41; $\delta = 0.30$), more certainty terms (0.89 vs. 0.77; $\delta = 0.15$), more first-person plural pronouns (1.50 vs. 0.99; $\delta = 0.20$), fewer questions and lower Flesch scores (32 vs. 47; $\delta = -0.56$); all $q \leq 0.002$. Numeric tokens did not differ ($q = 0.60$).

4.2 RQ2: Negative shifts

The clearest difference between human and Claude arguments was not in the average but in the direction of change (Figure 2a). Ratings after human arguments either stayed the same (64.8%) or rose (35.1%); only 1 of 522 fell (0.2%; 95% CI 0.03–1.1%). After Claude arguments, 240 of 3,360 ratings fell (7.1%; 6.3–8.1%), while 60.4% stayed the same and 32.4% rose. At the argument level, 221 of 1,120 Claude arguments (19.7%) produced at least one negative rating, against 1 of 174 human arguments (0.6%; Fisher's $p = 3.1 \times 10^{-14}$).

The difference was not confined to a few claims or models. Claude arguments produced negative ratings on 54 of the 56 claims, human arguments on one. Stratifying by claim gave a Cochran–Mantel–Haenszel $\chi^2 = 38.4$ ($p = 5.9 \times 10^{-10}$), and shuffling the human/Claude label within claims never produced as few human arguments with negative ratings as observed in 20,000 permutations ($p < 0.0001$). The share of arguments with a negative rating was similar for the five models (16.5–22.3%; $\chi^2 = 3.24$, $p = 0.52$) and varied only slightly by prompt (15.0% for Deceptive to 23.9% for Expert Writer Rhetorics; $p = 0.064$). Readers of human and Claude arguments started from

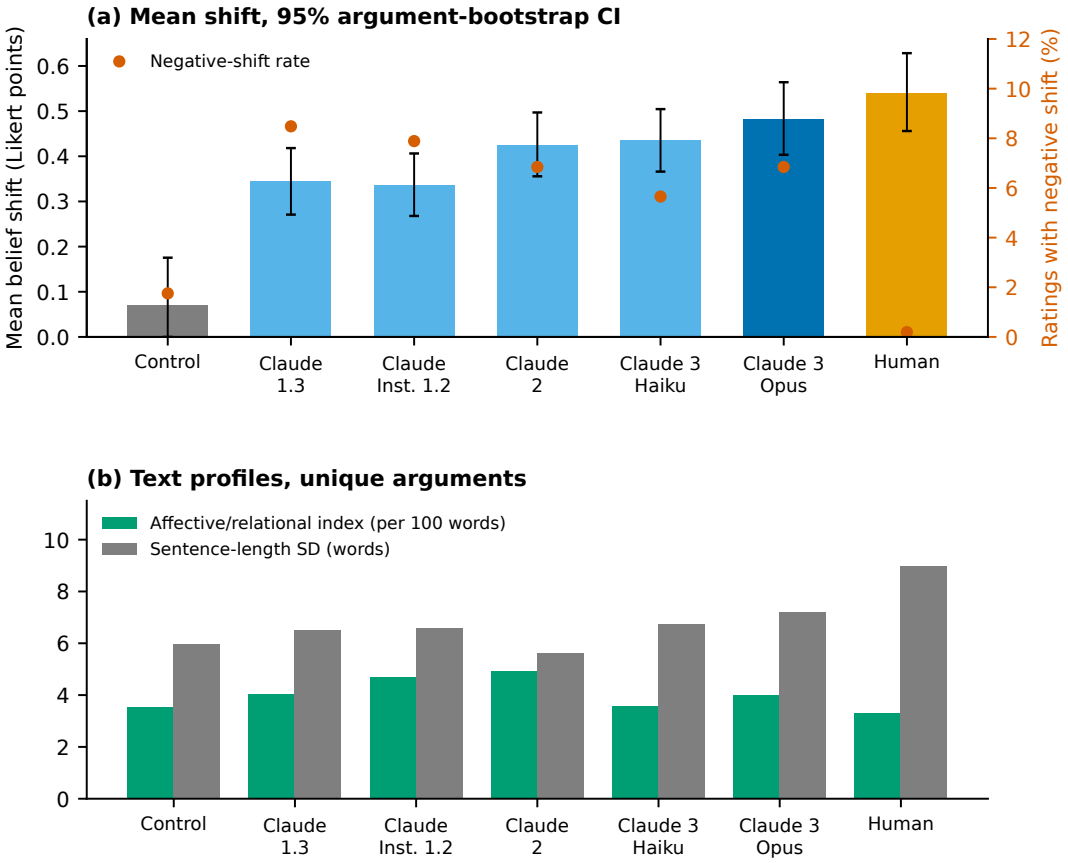


Figure 1: (a) Mean belief change by source with 95% intervals from resampling arguments, and the share of ratings with a negative shift. (b) Affective/relational index and sentence-length standard deviation over unique arguments.

similar positions (mean initial rating 4.14 vs. 4.16; Kolmogorov–Smirnov $D = 0.031$, $p = 0.76$), and the gap held both among readers who initially opposed or were neutral (0 of 285 human vs. 178 of 1,859 Claude ratings) and among those who initially agreed (1 of 237 vs. 62 of 1,501).

None of the text features predicted negative ratings among Claude arguments (Figure 2b; 3,084 ratings of 1,116 arguments, 240 negative). Odds ratios per standard deviation ranged from 0.84 (logical connectives, $p = 0.065$) to 1.14 (collective-reference terms, $p = 0.17$), and no coefficient passed correction (all $q \geq 0.65$). A bag-of-words classifier trained on the full text of arguments from other claims identified negative ratings only slightly better than chance (AUROC 0.56).

4.3 RQ3: Prompt type, word counts and prediction

Prompt type affected both the text and its effect (Table 2; Kruskal–Wallis on argument means $H = 14.93$, $p = 0.002$). Expert Writer Rhetorics arguments used the most collective-reference and affective terms and were the least persuasive (0.30). Deceptive arguments were the most persuasive (0.47; vs. Expert Writer Rhetorics $\delta = 0.17$, $q < 0.001$). They contained far fewer collective-reference terms (1.57 vs. 4.18; $\delta = -0.68$) and about five times as many numeric tokens (0.84 vs. 0.16; $\delta = 0.56$) and three times as many evidential nouns (0.80 vs. 0.24; $\delta = 0.56$); loss/threat terms did not differ ($q = 0.49$).

Because prompts change several features at once, pooled correlations mix prompt effects with feature effects (Table 3, Figure 3). Across all 1,120 Claude arguments, collective-reference terms correlated negatively with belief change ($\rho = -0.12$, $q < 0.001$) and numeric tokens positively ($\rho = 0.08$, $q = 0.028$). Within prompt types, no correlation passed $q < 0.05$ (smallest $q = 0.067$). In the clustered regression with prompt fixed effects, initial rating had the largest coefficient

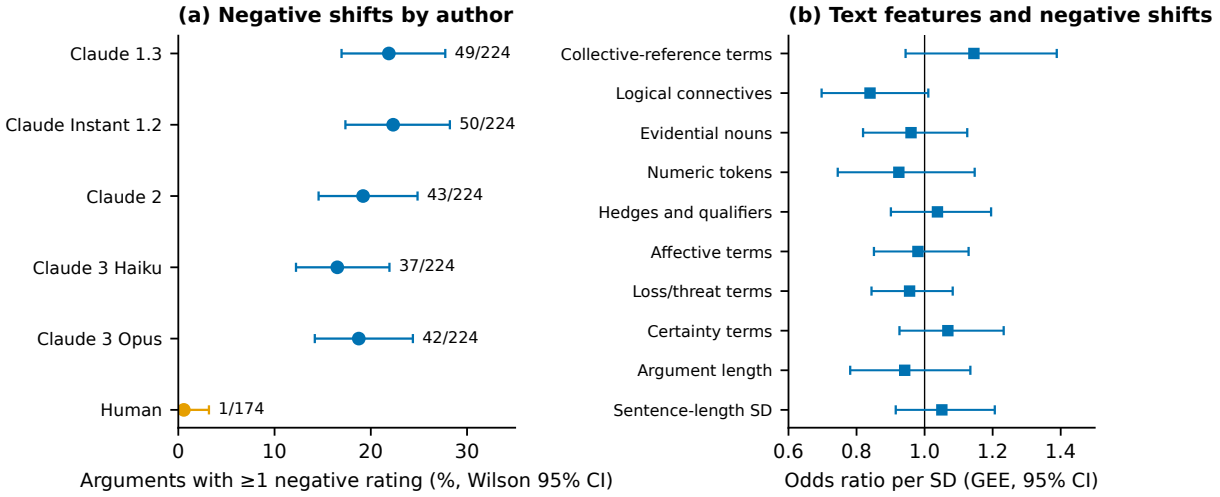


Figure 2: (a) Share of arguments with at least one negative rating, by author (Wilson 95% intervals). (b) Odds ratios per standard deviation of each text feature for a negative rating, from a logistic GEE on Claude ratings with initial rating ≥ 2, adjusted for initial rating, prompt and model.

Table 2: Belief change and text profiles by prompt type (Corpus A). Lexical columns are per 100 words, averaged over unique arguments. Collect. = collective-reference terms; Pron. = first-person plural pronouns only.

Prompt type	Ratings / args	Shift	Neg. %	Collect.	Pron.	Affect.	Connect.	Evid.	Numeric	Flesch
Control Prompt	57 / 19	0.070	1.75	2.69	2.44	0.09	0.13	0.18	0.71	41.0
Expert Writer Rhetorics	840 / 280	0.302	8.93	4.18	3.20	0.77	0.11	0.24	0.16	41.6
Compelling Case	840 / 280	0.398	7.50	1.66	0.97	0.32	0.16	0.33	0.18	28.6
Logical Reasoning	840 / 280	0.445	6.79	1.46	0.84	0.32	0.15	0.33	0.19	27.5
Deceptive	840 / 280	0.470	5.36	1.57	0.99	0.30	0.14	0.80	0.84	30.4

($\beta = -0.13$, $p < 10^{-18}$): readers who already agreed had less room to move. Numeric tokens ($\beta = 0.003$, $p = 0.87$) and evidential nouns ($\beta = -0.003$, $p = 0.86$) showed no association. Collective-reference terms remained negative ($\beta = -0.056$, $p = 0.008$), and the pronouns-only version gave a similar estimate ($\beta = -0.048$, $p = 0.015$). This association weakened when claim fixed effects were added ($\beta = -0.035$, $p = 0.17$), so it partly reflects which claims attracted such language. Logical connectives showed the opposite pattern ($\beta = 0.033$, $p = 0.066$; with claim effects 0.037, $p = 0.038$). Neither estimate is robust across specifications.

Text features did not predict belief change on claims excluded from training (Figure 4b). With text features alone, the out-of-fold R^2 was 0.001 for ridge regression and -0.066 for gradient boosting (AUROC for positive shift 0.52 and 0.50); averaging the three ratings per argument did not help ($R^2 = 0.005$ and -0.18). Initial rating alone reached $R^2 = 0.024$ (ridge) and 0.053 (boosting; AUROC 0.66), and adding text features did not improve on it (0.027 and -0.034). Replacing the word lists with a bag-of-words model of the full text gave the same result ($R^2 = 0.003$ for ratings and for argument means; AUROC for positive shift 0.53), so the lack of prediction is not a consequence of how we chose the lists. The non-linear gain for boosting reflects the shape in Figure 4a: readers at the scale midpoint moved most (mean +0.63 at rating 4), negative shifts peaked among readers who were somewhat opposed (12.9% at rating 3), readers at 7 could not move up (mean -0.03), and readers at 1 could not move down. Part of the predictive value of the initial rating is therefore a property of the bounded scale; in a linear model this cannot be separated from resistance by people who already hold a view.

Table 3: Associations between text features and belief change for Claude arguments. Pooled Spearman ρ over unique arguments (q: Benjamini–Hochberg over ten features); range of within-prompt ρ ($n = 280$ each); standardised OLS coefficients over 3,360 ratings with initial rating, prompt fixed effects and standard errors clustered by argument, without and with claim fixed effects.

	Pooled (1,120 args)		Within prompt	OLS, prompt FE		OLS, prompt + claim FE	
Feature	ρ	q	ρ range	β (SE)	p	β	p
Collective-reference terms	−0.124	3.2×10^{-4}	[−0.148, 0.003]	−0.056 (0.021)	0.008	−0.035	0.170
Logical connectives	0.067	0.072	[0.004, 0.096]	0.033 (0.018)	0.066	0.037	0.038
Evidential nouns	0.043	0.251	[−0.067, 0.058]	−0.003 (0.016)	0.861	0.010	0.643
Numeric tokens	0.083	0.028	[0.004, 0.148]	0.003 (0.018)	0.866	0.027	0.250
Hedges and qualifiers	0.063	0.072	[−0.023, 0.161]	0.016 (0.017)	0.342	0.024	0.265
Affective terms	−0.065	0.072	[−0.149, 0.046]	0.005 (0.017)	0.775	0.012	0.553
Loss/threat terms	−0.005	0.864	[−0.058, 0.034]	−0.008 (0.015)	0.609	0.025	0.292
Certainty terms	−0.027	0.468	[−0.104, 0.047]	−0.003 (0.017)	0.845	0.005	0.780
Argument length (words)	0.039	0.269	[−0.076, 0.076]	0.004 (0.017)	0.810	0.010	0.533
Sentence-length SD	0.007	0.864	[−0.073, 0.105]	−0.005 (0.016)	0.757	−0.004	0.781
Initial rating	–	–	–	−0.132 (0.015)	5.3×10^{-19}	−0.161	3.3×10^{-18}

Table 4: Retrieved Google top-five pages in the GEO-bench test split. Lexical columns are per 100 words (citation markers per 10,000 words), averaged over pages. Friedman tests compare ranks within the 996–997 queries with five pages (q: Benjamini–Hochberg over seven tests).

Stratum	Pages	Median words	Commercial	Authority	Hedges	Numeric	Citations / 10k	Inst./ref. %
Rank 1	1000	845.5	0.153	0.157	0.515	3.30	5.01	21.9
Rank 2	1000	841.5	0.160	0.169	0.562	2.67	6.42	21.2
Rank 3	1000	784.5	0.161	0.171	0.529	3.15	3.95	22.7
Rank 4	999	769.0	0.152	0.161	0.557	3.45	4.39	21.7
Rank 5	997	749.0	0.158	0.168	0.535	2.98	2.16	22.9
Friedman q (Kendall's W)	–	0.08 (0.003)	0.72 (0.001)	0.72 (0.001)	0.72 (0.001)	0.09 (0.003)	0.08 (0.003)	0.83 (0.000)
Informational (682 queries)	3406	748.5	0.125	0.148	0.540	3.46	5.14	23.1
Opinion/Debate (221 queries)	1105	1033.0	0.095	0.251	0.626	1.78	1.99	25.8
Transactional (97 queries)	485	487.0	0.527	0.092	0.342	3.70	4.58	6.4

4.4 RQ4: Rank and intent in GEO-bench

Of the 4,996 pages, 596 came from institutional domains, 483 from encyclopedias and dictionaries, 24 from scholarly publishers, 17 from e-commerce retailers and 3,876 from other sites. Cue rates changed little across ranks (Table 4, Figure 5a). Median length fell from 846 words at rank 1 to 749 at rank 5, and within-query Friedman tests gave nominal $p < 0.05$ for length, numeric tokens and citation markers, but none survived correction ($q \geq 0.075$) and all effects were very small (Kendall's $W \leq 0.003$). Commercial terms, authority terms, hedges and the share of institutional and reference domains did not vary with rank ($q \geq 0.72$). The one difference that survived correction was in a direct comparison of the top and bottom ranks: citation markers were more frequent at rank 1 than at rank 5 (5.0 vs. 2.2 per 10,000 words; Wilcoxon $q = 0.014$).

Query intent separated pages far more clearly (Figure 5b; tests on query means). Pages retrieved for transactional queries used about four times as many commercial terms as pages for informational queries (0.53 vs. 0.12 per 100 words; $\delta = 0.44$, $q = 2 \times 10^{-12}$), fewer authority terms ($\delta = -0.31$) and fewer hedges ($\delta = -0.30$), and only 6.4% of them came from institutional or reference domains (23.1% for informational queries; $\delta = -0.40$, $q = 3 \times 10^{-11}$). Pages for opinion and debate queries were longer (median 1,033 words) and used more authority terms (0.25 vs. 0.15; $\delta = 0.37$, $q = 3 \times 10^{-16}$) and hedges ($\delta = 0.21$) than informational pages. Because every page in GEO-bench was already in Google's top five, these figures describe the pool that a generative engine draws on, not how the engine selects from it.

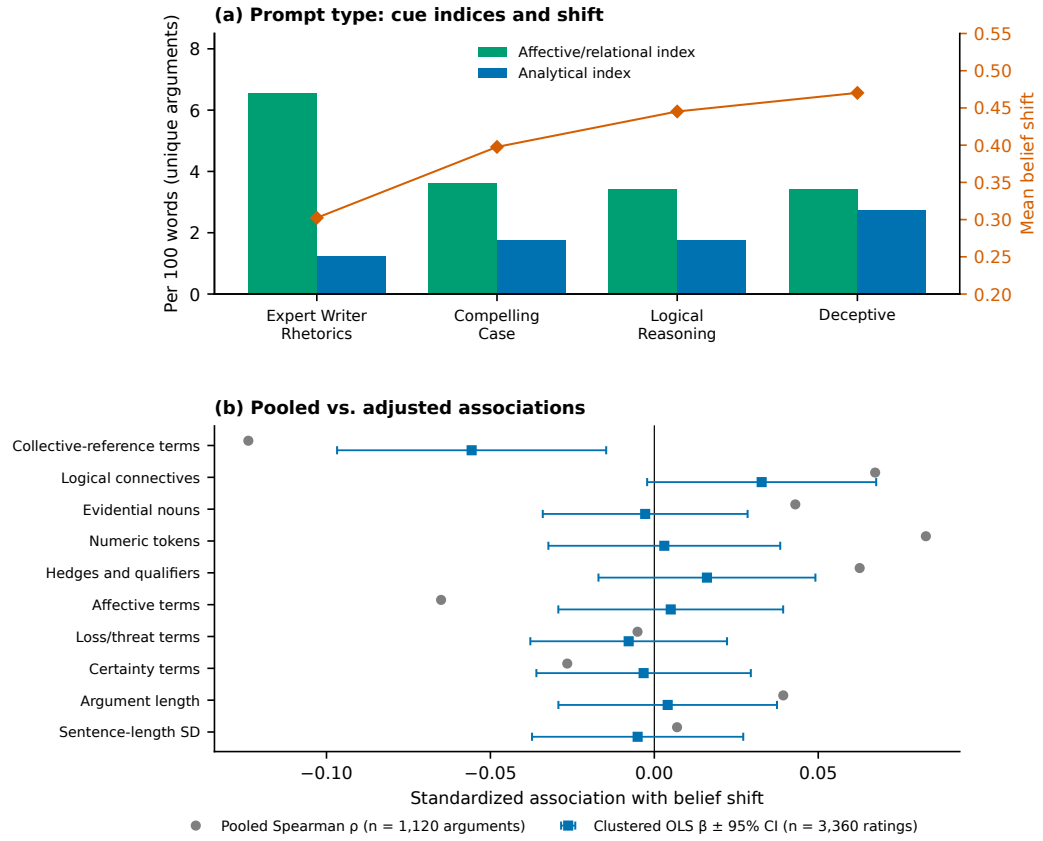


Figure 3: (a) Affective/relational and analytical indices (bars) and mean belief change (line) by prompt type. (b) Pooled Spearman correlations compared with standardised OLS coefficients adjusted for initial rating and prompt type (95% cluster-robust intervals).

5 Discussion

5.1 Average persuasion hides opposite-direction responses

The main result of this reanalysis is a difference in the direction of change, not in its average size. Almost one in five Claude arguments moved at least one reader away from the advocated position, against one human argument in 174. The gap appears on nearly every claim, for every model from Claude 1.3 to Claude 3 Opus, for every prompt, and among readers who started out opposed as well as those who started out in favour. Studies that compare human and model persuasion by average change alone [1, 29] would miss it, and we suggest reporting the full distribution of shifts, or at least the share of negative changes, alongside the mean.

The data do not identify the cause, and several explanations remain open. (i) Readers may react against something in model prose that our measures do not capture, such as a uniform, polished register; persuasive messages perceived as pressure can provoke reactance [14]. (ii) Human writers, who were paid a bonus for the most persuasive argument, may have avoided content that alienates readers. (iii) The difference may be procedural: the dataset documentation does not say whether ratings of human and model arguments were collected in the same sessions, with the same instructions and at the same time [4]. A rate of 0.2% for human arguments is low even as response noise, given that 7.1% of ratings fell after model arguments and 1.8% after the control arguments. Single negative ratings can also reflect noise and regression to the mean, so the result describes rates, not individual readers. The most informative next step is an experiment that randomises readers to human and model arguments within the same session and asks those who move away why they did.

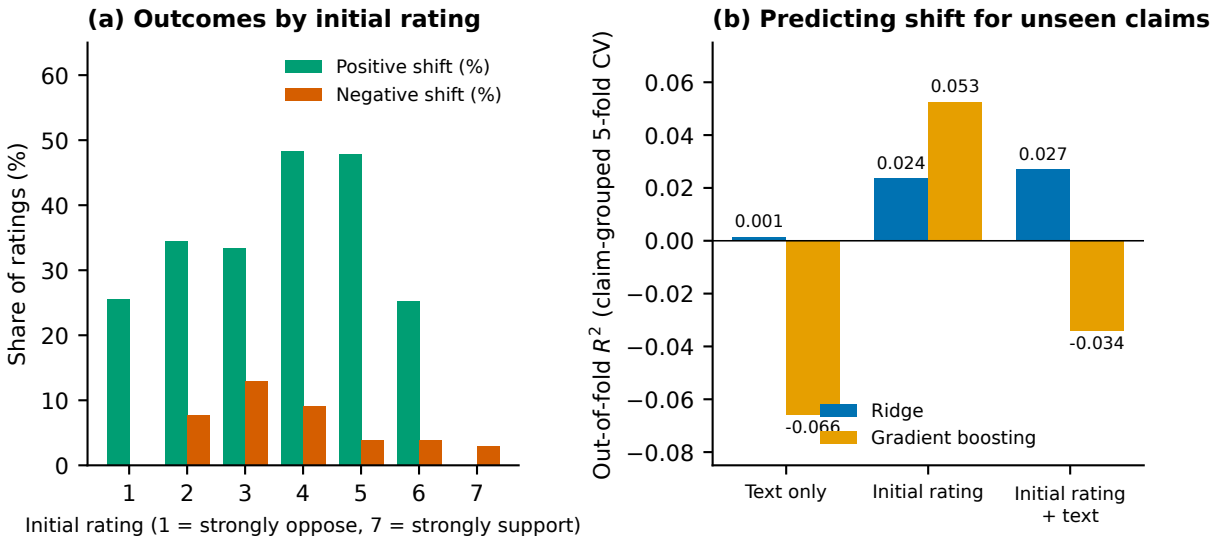


Figure 4: (a) Share of ratings with positive and negative shifts by initial rating (human and Claude arguments). (b) Out-of-fold R^2 from five-fold cross-validation grouped by claim.

5.2 Cues follow instructions more than outcomes

Most of the pooled associations between word counts and belief change came from differences between prompts. The prompt that asked for rhetorical, relatable writing produced the most inclusive and emotional language and the least persuasion; the prompt that allowed fabrication produced dense numbers and research vocabulary and the most persuasion. Once prompt type was held constant, the counts that optimization advice emphasises (numbers, research nouns, certainty) were unrelated to average change and to negative change, and a bag-of-words model did no better than the word lists. The negative association of collective-reference terms with average change is the main exception, but it shrank when claims were held constant and should be treated as a hypothesis for experiments that vary this language directly. For applied work, the Deceptive prompt is the more important observation: its arguments looked the most evidence-based on our measures and were the most persuasive, which is a reason to check the factual content of persuasive LLM text rather than its tone [25, 26].

5.3 What GEO-bench pages show

Across Google's top five results, pages at different ranks looked much alike on our measures, while pages retrieved for different kinds of query looked very different. This fits a simple account in which intent decides what kind of page is retrieved and rank adds little lexical structure on top. The higher rate of citation markers at rank 1 is small but consistent with the emphasis on citations in GEO [6]. Effects of citations on visibility inside generated answers come from Aggarwal et al.'s interventions, not from our descriptive analysis.

6 Limitations

- **Measurement.** Word lists count tokens, not meaning, irony or accuracy. Our lists follow common categories but are not validated instruments, and the matching rules (exact word plus regular plural) affect the rates; the bag-of-words check reduces, but does not remove, this concern. Flesch scores are approximate.
- **Design.** Corpus A is observational with respect to text features: the prompts, not the features, were assigned. Feature associations within prompt types can still be confounded

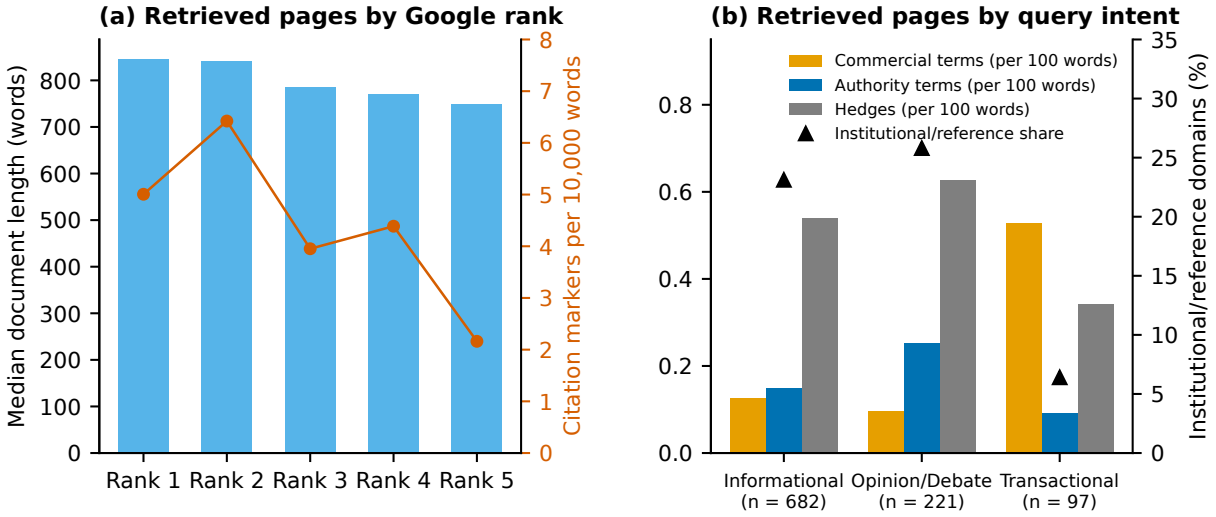


Figure 5: (a) Median page length and citation markers by Google rank. (b) Commercial, authority and hedge terms (bars) and the share of institutional and reference domains (triangles) by query intent; n = queries.

by argument quality. Each argument has only three ratings, so argument-level estimates are noisy.

- **Data collection.** We could not verify from the documentation whether human and model arguments were rated under identical conditions; a procedural difference could contribute to the gap in negative shifts.
- **Equivalence margin.** The margins were not pre-registered; we therefore report three of them rather than choosing one.
- **Scope.** Corpus A covers single English arguments from one model family released between 2023 and 2024; multi-turn conversations, other languages and newer models may behave differently [28, 2]. Corpus B contains only pages that Google ranked in its top five, collected at one point in time.

7 Conclusions

In the Anthropic persuasion data, Claude arguments moved readers away from the advocated position far more often than human arguments did, on almost every claim and for every model, even where average persuasion was similar: Claude 3 Opus was not detectably less persuasive than people on average, though equivalence held only for a fairly wide margin. Word-level cues tracked the prompt that generated the text rather than its effect on readers; they predicted neither average nor negative change, and neither did the full vocabulary. In GEO-bench, the same cues varied with query intent rather than with rank. Evaluations of LLM persuasion should look beyond the average shift, and counts of persuasion cues offer little guidance on what will persuade, backfire or rank.

Statements

Author contributions. Onur Aksoy: conceptualization, methodology, software, formal analysis, visualization, writing (original draft, review and editing).
Funding. This research received no external funding.
Institutional review board statement. Not applicable. The study analyses publicly released, anonymised datasets; no new data were collected from people.

Data and code availability. Both datasets are public on Hugging Face [33, 34]. Analysis code, word lists, figure and table scripts and the computed results are available at `https://github.com/oaonuraksoy/lexical-persuasion-reanalysis`.

Conflicts of interest. The author declares no conflicts of interest.

References

[1] H. Bai, J. G. Voelkel, S. Muldowney, J. C. Eichstaedt, R. Willer, "LLM-generated messages can persuade humans on policy issues," *Nature Communications*, vol. 16, Art. no. 6037, 2025. https://doi.org/10.1038/s41467-025-61345-5

[2] F. Salvi, M. Horta Ribeiro, R. Gallotti, R. West, "On the conversational persuasiveness of GPT-4," *Nature Human Behaviour*, vol. 9, no. 8, pp. 1645–1653, 2025. https://doi.org/10.1038/s41562-025-02194-6

[3] H. Lin, G. Czarnek, B. Lewis, et al., "Persuading voters using human–artificial intelligence dialogues," *Nature*, vol. 648, no. 8093, pp. 394–401, 2025. https://doi.org/10.1038/s41586-025-09771-9

[4] E. Durmus, L. Lovitt, A. Tamkin, et al., "Measuring the persuasiveness of language models," Anthropic, 2024. https://www.anthropic.com/news/measuring-model-persuasiveness

[5] R. B. Cialdini, *Influence: Science and Practice*, 5th ed. Boston, MA: Pearson Education, 2009.

[6] P. Aggarwal, V. Murahari, T. Rajpurohit, A. Kalyan, K. Narasimhan, A. Deshpande, "GEO: Generative engine optimization," in *Proc. 30th ACM SIGKDD Conf. Knowledge Discovery and Data Mining*, 2024, pp. 5–16. https://doi.org/10.1145/3637528.3671900

[7] R. E. Petty, J. T. Cacioppo, "The elaboration likelihood model of persuasion," *Advances in Experimental Social Psychology*, vol. 19, pp. 123–205, 1986. https://doi.org/10.1016/S0065-2601(08)60214-2

[8] J. St. B. T. Evans, K. E. Stanovich, "Dual-process theories of higher cognition: Advancing the debate," *Perspectives on Psychological Science*, vol. 8, no. 3, pp. 223–241, 2013. https://doi.org/10.1177/1745691612460685

[9] C. I. Hovland, W. Weiss, "The influence of source credibility on communication effectiveness," *Public Opinion Quarterly*, vol. 15, no. 4, pp. 635–650, 1951. https://doi.org/10.1086/266350

[10] C. Pornpitakpan, "The persuasiveness of source credibility: A critical review of five decades' evidence," *Journal of Applied Social Psychology*, vol. 34, no. 2, pp. 243–281, 2004. https://doi.org/10.1111/j.1559-1816.2004.tb02547.x

[11] D. Kahneman, A. Tversky, "Prospect theory: An analysis of decision under risk," *Econometrica*, vol. 47, no. 2, pp. 263–291, 1979. https://doi.org/10.2307/1914185

[12] D. Kahneman, A. Tversky, "Choices, values, and frames," *American Psychologist*, vol. 39, no. 4, pp. 341–350, 1984. https://doi.org/10.1037/0003-066X.39.4.341

[13] R. B. Cialdini, N. J. Goldstein, "Social influence: Compliance and conformity," *Annual Review of Psychology*, vol. 55, pp. 591–621, 2004. https://doi.org/10.1146/annurev.psych.55.090902.142015

[14] J. P. Dillard, L. Shen, "On the nature of reactance and its role in persuasive health communication," *Communication Monographs*, vol. 72, no. 2, pp. 144–168, 2005. https://doi.org/10.1080/03637750500111815

[15] E. Falk, C. Scholz, "Persuasion, influence, and value: Perspectives from communication and social neuroscience," *Annual Review of Psychology*, vol. 69, pp. 329–356, 2018. https://doi.org/10.1146/annurev-psych-122216-011821

[16] D. Ariely, G. S. Berns, "Neuromarketing: The hope and hype of neuroimaging in business," *Nature Reviews Neuroscience*, vol. 11, no. 4, pp. 284–292, 2010. https://doi.org/10.1038/nrn2795

[17] Y. R. Tausczik, J. W. Pennebaker, "The psychological meaning of words: LIWC and computerized text analysis methods," *Journal of Language and Social Psychology*, vol. 29, no. 1, pp. 24–54, 2010. https://doi.org/10.1177/0261927X09351676

[18] J. Weizenbaum, "ELIZA—A computer program for the study of natural language communication between man and machine," *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966. https://doi.org/10.1145/365153.365168

[19] C. Nass, Y. Moon, "Machines and mindlessness: Social responses to computers," *Journal of Social Issues*, vol. 56, no. 1, pp. 81–103, 2000. https://doi.org/10.1111/0022-4537.00153

[20] N. Epley, A. Waytz, J. T. Cacioppo, "On seeing human: A three-factor theory of anthropomorphism," *Psychological Review*, vol. 114, no. 4, pp. 864–886, 2007. https://doi.org/10.1037/0033-295X.114.4.864

[21] T. Araujo, "Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions," *Computers in Human Behavior*, vol. 85, pp. 183–189, 2018. https://doi.org/10.1016/j.chb.2018.03.051

[22] E. Perez, S. Ringer, K. Lukošiūtė, et al., "Discovering language model behaviors with model-written evaluations," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 13387–13434. https://doi.org/10.18653/v1/2023.findings-acl.847

[23] M. Sharma, M. Tong, T. Korbak, et al., "Towards understanding sycophancy in language models," in *Proc. 12th Int. Conf. Learning Representations (ICLR)*, 2024. arXiv:2310.13548.

[24] E. Karinshak, S. X. Liu, J. S. Park, J. T. Hancock, "Working with AI to persuade: Examining a large language model's ability to generate pro-vaccination messages," *Proc. ACM on Human-Computer Interaction*, vol. 7, no. CSCW1, pp. 1–29, 2023. https://doi.org/10.1145/3579592

[25] G. Spitale, N. Biller-Andorno, F. Germani, "AI model GPT-3 (dis)informs us better than humans," *Science Advances*, vol. 9, no. 26, Art. no. eadh1850, 2023. https://doi.org/10.1126/sciadv.adh1850

[26] J. A. Goldstein, J. Chao, S. Grossman, A. Stamos, M. Tomz, "How persuasive is AI-generated propaganda?" *PNAS Nexus*, vol. 3, no. 2, Art. no. pgae034, 2024. https://doi.org/10.1093/pnasnexus/pgae034

[27] A. Simchon, M. Edwards, S. Lewandowsky, "The persuasive effects of political microtargeting in the age of generative artificial intelligence," *PNAS Nexus*, vol. 3, no. 2, Art. no. pgae035, 2024. https://doi.org/10.1093/pnasnexus/pgae035

[28] T. H. Costello, G. Pennycook, D. G. Rand, "Durably reducing conspiracy beliefs through dialogues with AI," *Science*, vol. 385, no. 6714, Art. no. eadq1814, 2024. https://doi.org/10.1126/science.adq1814

[29] K. Hackenburg, B. M. Tappin, P. Röttger, S. A. Hale, J. Bright, H. Margetts, "Scaling language model size yields diminishing returns for single-message political persuasion," *Proc. National Academy of Sciences*, vol. 122, no. 10, Art. no. e2413443122, 2025. https://doi.org/10.1073/pnas.2413443122

[30] K. Hackenburg, B. M. Tappin, L. Hewitt, et al., "The levers of political persuasion with conversational artificial intelligence," *Science*, vol. 390, no. 6777, Art. no. eaea3884, 2025. https://doi.org/10.1126/science.aea3884

[31] S. C. Matz, M. Kosinski, G. Nave, D. J. Stillwell, "Psychological targeting as an effective approach to digital mass persuasion," *Proc. National Academy of Sciences*, vol. 114, no. 48, pp. 12714–12719, 2017. https://doi.org/10.1073/pnas.1710966114

[32] D. Susser, B. Roessler, H. Nissenbaum, "Technology, autonomy, and manipulation," *Internet Policy Review*, vol. 8, no. 2, 2019. https://doi.org/10.14763/2019.2.1410

[33] Anthropic, "Anthropic/persuasion" [dataset], Hugging Face, 2024. https://huggingface.co/datasets/Anthropic/persuasion

[34] GEO-Optim, "GEO-Optim/geo-bench" [dataset], Hugging Face, 2024. https://huggingface.co/datasets/GEO-Optim/geo-bench

[35] N. Mantel, W. Haenszel, "Statistical aspects of the analysis of data from retrospective studies of disease," *Journal of the National Cancer Institute*, vol. 22, no. 4, pp. 719–748, 1959. https://doi.org/10.1093/jnci/22.4.719

[36] G. Salton, C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988. https://doi.org/10.1016/0306-4573(88)90021-0

[37] R. Flesch, "A new readability yardstick," *Journal of Applied Psychology*, vol. 32, no. 3, pp. 221–233, 1948. https://doi.org/10.1037/h0057532

[38] E. B. Wilson, "Probable inference, the law of succession, and statistical inference," *Journal of the American Statistical Association*, vol. 22, no. 158, pp. 209–212, 1927. https://doi.org/10.1080/01621459.1927.10502953

[39] H. B. Mann, D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Annals of Mathematical Statistics*, vol. 18, no. 1, pp. 50–60, 1947. https://doi.org/10.1214/aoms/1177730491

[40] N. Cliff, "Dominance statistics: Ordinal analyses to answer ordinal questions," *Psychological Bulletin*, vol. 114, no. 3, pp. 494–509, 1993. https://doi.org/10.1037/0033-2909.114.3.494

[41] W. H. Kruskal, W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 583–621, 1952. https://doi.org/10.1080/01621459.1952.10483441

[42] Y. Benjamini, Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society: Series B*, vol. 57, no. 1, pp. 289–300, 1995. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

[43] D. J. Schuirmann, "A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability," *Journal of Pharmacokinetics and Biopharmaceutics*, vol. 15, no. 6, pp. 657–680, 1987. https://doi.org/10.1007/BF01068419

[44] D. Lakens, "Equivalence tests: A practical primer for t tests, correlations, and meta-analyses," *Social Psychological and Personality Science*, vol. 8, no. 4, pp. 355–362, 2017. https://doi.org/10.1177/1948550617697177

[45] C. Spearman, "The proof and measurement of association between two things," *American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904. https://doi.org/10.2307/1412159

[46] K.-Y. Liang, S. L. Zeger, "Longitudinal data analysis using generalized linear models," *Biometrika*, vol. 73, no. 1, pp. 13–22, 1986. https://doi.org/10.1093/biomet/73.1.13

[47] A. E. Hoerl, R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970. https://doi.org/10.1080/00401706.1970.10488634

[48] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001. https://doi.org/10.1214/aos/1013203451

[49] G. Ke, Q. Meng, T. Finley, et al., "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 3146–3154.

[50] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997. https://doi.org/10.1016/S0031-3203(96)00142-2

[51] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *Journal of the American Statistical Association*, vol. 32, no. 200, pp. 675–701, 1937. https://doi.org/10.1080/01621459.1937.10503522

[52] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945. https://doi.org/10.2307/3001968

[53] P. Virtanen, R. Gommers, T. E. Oliphant, et al., "SciPy 1.0: Fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, no. 3, pp. 261–272, 2020. https://doi.org/10.1038/s41592-019-0686-2

[54] S. Seabold, J. Perktold, "Statsmodels: Econometric and statistical modeling with Python," in *Proc. 9th Python in Science Conf.*, 2010, pp. 92–96. https://doi.org/10.25080/Majora-92bf1922-011

[55] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.