Supporting Information for

Human Heuristics for AI-Generated Language Are Flawed

Maurice Jakesch\textsuperscript{a,b}, Jeffrey T Hancock\textsuperscript{c}, Mor Naaman\textsuperscript{a,b},

\textsuperscript{a} Cornell University, Ithaca, NY 14850, United States
\textsuperscript{b} Cornell Tech, New York, NY 10044, United States
\textsuperscript{c} Stanford University, Stanford, CA 94305, United States

* Maurice Jakesch

Email: mj32@cornell.edu

This PDF file includes:

- Supporting text
- Tables S1 to S6
- Figure S1 to S2
Supporting Text

Below we provide additional information on several aspects of our experiments. Table S1 summarizes the treatment, stimuli, and recruitment methods used across the six studies and three labeling tasks. Table S2 shows a sample of self-presentations for each study and treatment group.

Table S3 shows the results of an auxiliary analysis testing whether certain groups are better at detecting AI-generated language than others. Older participants were slightly more likely to detect AI-generated self-presentations, with participants older than 50 achieving an accuracy of 53% (compared to 51% for younger participants). No gender or ethnic group performed better than others. Participants with a university degree performed about 1% worse than those without, and self-reported technical knowledge was not correlated with more accurate ratings. Neither the time taken for the judgment nor the length of profiles predicted higher judgment accuracy. Across contexts, groups, and treatments, participants could not detect AI-generated self-presentations.

Table S4 and S5 provide further detail on the qualitative analysis of participants’ explanations of why they thought certain self-presentations were AI-generated or human-written. Two researchers independently coded a sample of responses into themes to provide an overview of participants’ self-reported heuristics. Table S4 presents an overview of recurring themes. Participants most commonly referred to the content of a self-presentation (blue-shaded regions in Table S4 representing 40% of responses). The participants reported associating specific content related to family and life experiences with language written by humans and generic or nonsensical content with AI-generated language. Participants also reported basing their decisions on grammatical cues (gray, 28%), where first-person pronouns and the mastery of grammar were mentioned as indicative of human-generated language. Some participants saw grammatical errors as associated with a subpar AI, but others claimed they associated them with fallible human authors. Another category of cues mentioned by participants was the tone (green, 24%). Participants reported associating warm and genuine language with humanity and impersonal, monotonous style with AI-generated language. The codebook, theme frequencies, and sample responses are shown in Table S5. Table S6 provides a complete overview of the developed language features and statistical summaries.

Prior research suggests that asking participants to explain their responses could have changed their subsequent evaluations or degraded performance (1,2). We thus conducted an analysis testing whether participants’ performance had changed after being asked to explain their judgment. The results are shown in Figure S1. There was no evidence for such change in our data as participants’ accuracy before and after the open-ended response did not change across any of the three contexts. Note that open-ended responses were only solicited for the three main experiments. The validation experiments did not include open-ended responses, showing similar outcomes and providing further evidence that participants’ ratings (and our findings) were not affected by the explanations.

Figure S2 shows how crowdworkers evaluated human-written and AI-generated self-presentations in a separate labeling task when asked whether the text was nonsensical, seemed repetitive, or had grammatical issues. Crowdworkers were significantly more likely to rate AI-generated self-presentations as nonsensical (13.6% vs. 9.6%, p<0.0001). This was the case in the hospitality context, in particular, where we had used the older GPT-2 model to generate self-presentations. Crowdworkers also rated generated self-presentations as more repetitive (12.7% vs. 7.11%, p<0.0001), particularly in the professional context. Finally, crowdworkers labeled generated self-presentations as having fewer grammatical issues than human-written text (14.8% vs. 19.6%, p<0.0001). This difference was most
pronounced in the dating and professional contexts where we had used the more advanced GPT-3 model to generate self-presentation.

**SI References**

1. T. D. Wilson, J. W. Schooler, *Thinking too much: introspection can reduce the quality of preferences and decisions*. J. Pers. Soc. Psychol. 60, 181 (1991).
2. T. D. Wilson, D. S. Dunn, D. Kraft, D. J. Lisle, "Introspection, attitude change, and attitude-behavior consistency: The disruptive effects of explaining why we feel the way we do" in *Advances in Experimental Social Psychology*, (Elsevier, 1989), pp. 287–343.

**Table S1: Overview of experiments**

| Context          | Stimuli                                                                 | Treatment                                                                 | Recruitment                                      |
|------------------|------------------------------------------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------|
| Main study 1:    | 1,500 self-presentations from Airbnb and 1,500 generated by GPT-2; 30-60 words each; 16 per subject | Within-subject variation of self-presentation type                        | N = 2,000 US-representative sample via Lucid      |
| Hospitality      |                                                                        |                                                                          |                                                  |
| Main study 2:    | 1,000 self-presentations from OkCupid and 1,000 generated by GPT-3; 60-90 words; 12 per subject | Within-subject variation of self-presentation type and between-subject bonus payments for correct ratings | N = 1,000 gender-balanced sample via Prolific    |
| Dating           |                                                                        |                                                                          |                                                  |
| Main study 3:    | 1,000 self-presentations from Guru and 1,000 generated by GPT-3; 60-90 words each; 12 per subject | Within-subject variation of self-presentation type and between-subject feedback on answers | N = 1,000 gender-balanced sample via Prolific    |
| Professional     |                                                                        |                                                                          |                                                  |
| Validation study 1:    | 100 self-presentations from Airbnb, 100 generated by GPT-2, and 100 optimized using the language model classifier; 16 per subject | Within-subject variation of self-presentation type                        | N = 250 US-representative sample via Lucid       |
| Hospitality      |                                                                        |                                                                          |                                                  |
| Validation study 2:    | 100 self-presentations from OkCupid, 100 generated by GPT-3, and 100 optimized by the regression classifier; 16 per subject | Within-subject variation of self-presentation type                        | N = 200 gender-balanced sample via Prolific      |
| Dating           |                                                                        |                                                                          |                                                  |
| Validation study 3:    | 100 self-presentations from Guru, 100 generated by GPT-3, and 100 optimized using an ensemble classifier; 16 per subject | Within-subject variation of self-presentation type                        | N = 200 gender-balanced sample via Prolific      |
| Professional     |                                                                        |                                                                          |                                                  |
| Labeling task 1:  | 1,500 self-presentations from Airbnb and 1,500 generated by GPT-2; 30-60 words each; 12-16 per crowdworker | None                                                                     | N = 600 US-representative sample via Lucid       |
| Hospitality      |                                                                        |                                                                          |                                                  |
| Labeling task 2: Dating | 1,000 self-presentations from OkCupid and 1,000 generated by GPT-3; 60-90 words; 12 per crowdworker | None | N = 350 gender-balanced sample via Prolific |
|-------------------------|-----------------------------------------------|------|-------------------------------------------|
| Labeling task 3: Professional | 1,000 self-presentations from Guru and 1,000 generated by GPT-3; 60-90 words each; 12 per crowdworker | None | N = 350 gender-balanced sample via Prolific |

Table S2: Self-presentation examples

| Context       | Source               | Example                                                                                                                                                                                                                                                                                                                                 |
|---------------|----------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Hospitality   | Human                | My family has lived in DC for the past several years. Some of our favorite things about living on Capitol Hill are running through the neighborhood, exploring all the museums and exhibits that are walking distance from our home, and having a variety of great food offerings only steps away. |
| Hospitality   | Generated (GPT-2)     | A teacher and young entrepreneur, I love to ski and travel. My wife & I have lived in Vermont for the past 10 years and love the beauty and the snow that we get to ski during the summer.                                                                                                           |
| Hospitality   | Generated (GPT-2) & optimized (regression) | My husband and I have lived in Denver for 20 years. A few summers ago we visited my two brothers who live elsewhere so we decided to make our home available for others to enjoy as well. We love traveling in Europe, South America and anywhere new! Welcome to your home away from home. |
| Dating        | Human                | i'm an elementary school social worker and find my job both fulfilling and frustrating. an la native, i've also lived in the midwest and new england. i've been in sf for about 6 years now and love the people, politics, and food here. but, i do miss having seasons and look forward to my annual vacations back in the midwest, which generally involve lounging on a lake and drinking bell's beer. i enjoy being fit, active, and healthy, though i do eat ice cream for dinner on occasion. |
| Dating        | Generated (GPT-3)     | i just moved to the city last august and really don't know many people here yet. i'm interested in hanging out and maybe even finding someone special. i would love to be able to spend time together without any drama and want to get to know each other better. i'd love to find someone that i can share all of these exciting things in life with like art galleries, theatre, dinner, etc... |
| Dating        | Generated (GPT-3) & optimized (GPT-2) | hey i moved to sf about 2 years ago, it's such a great city..i like to explore the city, always trying to find new hangouts and food... i've travelled a lot around the world and would love to travel more. i'm easy going and down to earth, i know what i want in life and am working towards my goals. message me if you want to know more :) |
| Professional  | Human                | I have 19 years of journalism experience. My work has appeared in daily and weekly newspapers, international trade magazines and textbooks. I also have worked in broadcast news, and my reporting has been picked up by the Associated Press. For six years, my... |
interviews focused on C-level execs at Fortune 500 power companies, tech startups and government. In 2015, I became managing editor of a publication in the petroleum and fluid handling equipment industry.

My name is Gary Stauch and I have been in the computer and electronics business for over 30 years. I have a A.S. in electronics, a B.S. in computer science and I am a registered professional engineer in Texas. In addition to my own company, I have worked for several others in the design and deployment of large scale network infrastructure in the data center and enterprise server market. I have designed and developed server platforms, workstations, servers, switches, routers and other devices that are part of large scale networks.

I am a mother of three and a grandmother of two. I live in beautiful Iowa and have been here all my life. I enjoy doing different things but I am a master at none. I love to tell stories and make people smile with laughter. I am very well at reading people and knowing what to do to get the job done. I am very good at multi-tasking. I am very organized and very well at using my time.

Table S3: Regression coefficients predicting the accuracy of a judgment based on treatment, social context, and participant demographics. No group performed much above chance level.

| Dependent variable: | Likelihood of accurate assessment OR (95% CIs) |
|---------------------|---------------------------------------------|
| Context: Dating profiles | 0.974 (0.882, 1.065) |
| Context: Professional profiles | 0.926 (0.845, 1.007) |
| Treatment: Feedback | 1.038 (0.966, 1.110) |
| Treatment: Incentives | 1.022 (0.944, 1.100) |
| Age | **1.002** (1.001, 1.003) |
| Gender: Female | 1.002 (0.967, 1.036) |
| Gender: Non-binary | 1.010 (0.834, 1.186) |
| Race: African American | 0.959 (0.895, 1.022) |
| Race: Asian | 1.055 (0.976, 1.134) |
| Race: Hispanic | 1.005 (0.940, 1.069) |
| Race: Other | 0.973 (0.887, 1.059) |
| Level of education | **0.986** (0.976, 0.996) |
| Technical knowledge | 1.006 (0.982, 1.030) |
| Rating: Time taken | 1.000 (1.000, 1.001) |
| Profile: Word count | 1.000 (0.998, 1.002) |
Table S4. Themes in participants’ explanations of why they thought a self-presentation was human or generated language. N = 800, tile areas correspond to theme prevalence reported in brackets. Heuristics are classified by whether they refer to the content (blue), tone (green), grammar (gray), or form (red) of a self-presentation. Lighter tiles show cue that were associated with generated language.

Table S5: Examples themes and codes in participants’ explanations of judgments

| Category                  | Code                      | Freq. | Example                                                                 |
|---------------------------|---------------------------|-------|-------------------------------------------------------------------------|
| Content cues for AI       | Nonsensical content       | 7%    | “travel here from around the world' in third sentence doesn't make sense” |
| Content cues for AI       | Generic content           | 6%    | “seems just a bit to generic and a bit random”                          |
| Content cues for AI       | Unlikely content          | 4%    | “A full time manager at a nuclear plant doesn't travel frequently enough to care about hotel amenities.” |
| Content cues for Humanity | Specific content          | 14%   | “How detailed descriptions were”                                       |
| Content cues for Humanity | Family and biography      | 6%    | I determine this is a person because he says him and his wife and son travel and go places on there free time” |
| Feature Name for Humanity | Consistent | 3% | “Based primarily on the content, and whether each part of the statement made sense logically and thematically with the rest.” |
|---------------------------|------------|----|---------------------------------------------------------------------------------------------------------------------------------|
| Form cues for AI          | Repetitive | 2% | “the repetition of the sentences make the whole thing sound lifeless and robotic.” |
| Form cues for AI          | Template-like | 2% | “I looked for a stock template response for AI, or for signs of a disjointed copy and paste from real user statements.” |
| Grammar cues for AI       | Errors     | 7% | “If things are worded incorrectly.” |
| Grammar cues for AI       | Unusual punctuation | 7% | “There should be a comma after ‘I’m Kellie’” |
| Grammar cues for Humanity | Errors     | 5% | “Believe there was a grammar error where it should have been knowledgeable” |
| Grammar cues for Humanity | 1st person speech | 4% | “Using I, me, we language” |
| Grammar cues for Humanity | Good grammar | 3% | “The English is good, but not great. It possibly is written by someone who is ESL.” |
| Grammar cues for Humanity | Rare words | 3% | “Certain words that were unusual.” |
| Tone cues for AI          | Strange and unpersonal | 6% | “The personal touch is very unnatural sounding.” |
| Tone cues for AI          | Monotonous | 3% | “most people either put in little or more thought and AI just feels like a perfect monotone read” |
| Tone cues for Humanity    | Genuinely personal | 10% | “one can have a few replies per question and then have the AI Place together; but this isn’t random.. it is Genuine” |
| Tone cues for Humanity    | Warm and welcoming | 6% | “Its how the phrase comes across, An AI Having Emotion…” |

**Table S7**: Overview of language features and their correlations with participants’ judgments.

| Feature Name (manual labels) | Mean  | SD.   | Min | Max  | Cor. with ratings | Cor. with source |
|-----------------------------|-------|-------|-----|------|-------------------|------------------|
| Nonsensical                 | 0.117 | 0.233 | 0   | 1    | 0.086             | 0.114            |
| Category                | Value1 | Value2 | Count1 | Value3 | Value4 | Value5 |
|-------------------------|--------|--------|--------|--------|--------|--------|
| Repetitive (manual labels) | 0.099  | 0.222  | 0      | 1      | 0.127  | 0.057  |
| Grammatical issues (manual) | 0.172  | 0.281  | 0      | 1      | -0.086 | 0.057  |
| LIWC Achieve            | 2.325  | 2.535  | 0      | 17.72  | 0.037  | -0.009 |
| LIWC Acquire            | 0.492  | 0.941  | 0      | 9.72   | -0.012 | 0.01   |
| LIWC Adjective          | 6.968  | 3.797  | 0      | 30.95  | 0.029  | -0.007 |
| LIWC Adverb             | 3.898  | 3.038  | 0      | 22.22  | -0.094 | 0.023  |
| LIWC Affect             | 7.368  | 4.983  | 0      | 34.48  | 0.028  | 0.024  |
| LIWC Affiliation        | 3.15   | 4.238  | 0      | 25.81  | 0.033  | 0.032  |
| LIWC Allnone            | 0.854  | 1.374  | 0      | 12.9   | 0.017  | 0.001  |
| LIWC Allpunc            | 17.037 | 8.73   | 0      | 257.14 | -0.009 | -0.046 |
| LIWC Allure             | 9.614  | 4.967  | 0      | 32.35  | -0.056 | 0.09   |
| LIWC Analytic           | 56.357 | 27.358 | 1      | 99     | 0.098  | -0.051 |
| LIWC Apostro            | 1.75   | 2.249  | 0      | 21.67  | -0.109 | 0.03   |
| LIWC Article            | 5.938  | 3.066  | 0      | 20.45  | 0.013  | 0.088  |
| LIWC Assent             | 0.035  | 0.259  | 0      | 8.82   | -0.047 | -0.013 |
| LIWC Attention          | 0.615  | 1.259  | 0      | 10.64  | 0.003  | 0.019  |
| LIWC Auditory           | 0.336  | 0.976  | 0      | 11.54  | -0.011 | -0.036 |
| LIWC   | Total | Mean | N | Mean | Lower | Upper |
|--------|-------|------|---|------|-------|-------|
| Authentic | 72.388 | 30.232 | 1 | 99 | -0.197 | 0.031 |
| Auxverb | 7.599 | 3.585 | 0 | 25 | -0.077 | 0.112 |
| Bigwords | 20.104 | 8.673 | 0 | 68.42 | 0.123 | -0.128 |
| Cause | 0.94 | 1.367 | 0 | 9.52 | 0.033 | -0.014 |
| Certitude | 0.35 | 0.89 | 0 | 9.3 | -0.038 | -0.01 |
| Clout | 33.423 | 35.196 | 1 | 99 | 0.162 | 0.01 |
| Cognition | 8.361 | 5.307 | 0 | 36.67 | -0.012 | -0.011 |
| Cogproc | 7.457 | 5.005 | 0 | 36.67 | -0.017 | -0.01 |
| Comm | 1.236 | 1.758 | 0 | 17.65 | -0.035 | -0.011 |
| Comma | 5.566 | 4.722 | 0 | 42.11 | 0.024 | -0.075 |
| Conflict | 0.033 | 0.248 | 0 | 5 | -0.02 | -0.019 |
| Conj | 8.083 | 3.071 | 0 | 25.3 | -0.033 | 0.055 |
| Conversation | 0.24 | 0.801 | 0 | 21.05 | -0.089 | -0.026 |
| Culture | 0.988 | 1.961 | 0 | 19.05 | 0.06 | -0.02 |
| Curiosity | 0.983 | 1.601 | 0 | 12.5 | -0.007 | 0.022 |
| Death | 0.02 | 0.19 | 0 | 3.61 | -0.007 | -0.012 |
| Det | 11.627 | 4.017 | 0 | 27.66 | -0.021 | 0.06 |
| LIWC   |   |   |   |   |   |
|-------|---|---|---|---|---|
| LIWC Dic | 88.99 | 6.611 | 36.84 | 100 | -0.092 |
| LIWC Differ | 2.054 | 2.199 | 0 | 14.71 | -0.04 |
| LIWC Discrep | 1.208 | 1.687 | 0 | 12.2 | -0.005 |
| LIWC Drives | 6.244 | 4.802 | 0 | 29.41 | 0.069 |
| LIWC Emo Anger | 0.026 | 0.233 | 0 | 5.88 | -0.022 |
| LIWC Emo Anx | 0.033 | 0.277 | 0 | 8.22 | -0.015 |
| LIWC Emo Neg | 0.132 | 0.56 | 0 | 9.09 | -0.032 |
| LIWC Emo Pos | 2.502 | 2.662 | 0 | 17.65 | -0.012 |
| LIWC Emo Sad | 0.016 | 0.173 | 0 | 5.08 | 0.023 |
| LIWC Emotion | 2.679 | 2.747 | 0 | 20.59 | -0.018 |
| LIWC Ethnicity | 0.122 | 0.675 | 0 | 16.39 | 0.002 |
| LIWC Exclam | 0.76 | 1.68 | 0 | 26.58 | -0.007 |
| LIWC Family | 0.602 | 1.465 | 0 | 12.9 | -0.083 |
| LIWC Fatigue | 0.014 | 0.164 | 0 | 4 | -0.022 |
| LIWC Feeling | 0.267 | 0.738 | 0 | 6.67 | 0.018 |
| LIWC Female | 0.426 | 1.197 | 0 | 19.35 | -0.008 |
| LIWC Filler | 0.005 | 0.098 | 0 | 4.11 | -0.015 |
| LIWC Focusfuture | 0.919 | 1.624 | 0 | 16.67 | 0.022 | -0.001 |
| LIWC Focuspast  | 2.345 | 2.636 | 0 | 15.38 | -0.111 | 0.008 |
| LIWC Focuspresent | 5.2 | 2.984 | 0 | 24.14 | 0.003 | 0.072 |
| LIWC Food       | 0.737 | 1.657 | 0 | 19.05 | -0.01 | 0.009 |
| LIWC Friend     | 0.466 | 1.053 | 0 | 14.29 | 0.025 | 0.039 |
| LIWC Fulfill    | 0.153 | 0.527 | 0 | 5.56 | 0.036 | -0.017 |
| LIWC Function   | 51.71 | 8.185 | 1.32 | 79.41 | -0.129 | 0.162 |
| LIWC Health     | 0.31 | 1.037 | 0 | 17.86 | -0.006 | -0.01 |
| LIWC Home       | 0.721 | 1.531 | 0 | 22.86 | 0.021 | -0.007 |
| LIWC I me       | 7.962 | 4.525 | 0 | 24.39 | -0.212 | 0.031 |
| LIWC Illness    | 0.024 | 0.249 | 0 | 6.25 | 0.019 | -0.026 |
| LIWC Insight    | 1.674 | 1.994 | 0 | 15 | 0.022 | -0.029 |
| LIWC Ipron      | 2.301 | 2.483 | 0 | 22.06 | -0.005 | 0.029 |
| LIWC Lack       | 0.051 | 0.381 | 0 | 6.9 | -0.003 | -0.021 |
| LIWC Leisure    | 1.975 | 2.788 | 0 | 19.35 | -0.043 | -0.004 |
| LIWC Lifestyle  | 8.156 | 5.838 | 0 | 40 | 0.025 | -0.013 |
| LIWC Linguistic | 66.286 | 9.102 | 6.58 | 91.18 | -0.122 | 0.156 |
| LIWC Category   | Value1 | Value2 | Value3 | Value4 | Value5 |
|-----------------|--------|--------|--------|--------|--------|
| LIWC Male       | 0.572  | 1.223  | 0      | 15.69  | -0.004 | 0.009  |
| LIWC Memory     | 0.031  | 0.259  | 0      | 4.76   | 0.02   | -0.021 |
| LIWC Mental     | 0.022  | 0.246  | 0      | 8      | -0.022 | 0.011  |
| LIWC Money      | 1.089  | 2.075  | 0      | 20.51  | 0.065  | -0.012 |
| LIWC Moral      | 0.204  | 0.69   | 0      | 8.11   | -0.016 | -0.04  |
| LIWC Motion     | 2.131  | 2.293  | 0      | 16.13  | -0.032 | 0.018  |
| LIWC Need       | 0.282  | 0.86   | 0      | 8.86   | 0.022  | -0.009 |
| LIWC Negate     | 0.521  | 1.082  | 0      | 12.5   | -0.057 | -0.019 |
| LIWC Netspeak   | 0.184  | 0.686  | 0      | 21.05  | -0.082 | -0.028 |
| LIWC Nonflu     | 0.022  | 0.196  | 0      | 4.35   | -0.022 | 0.003  |
| LIWC Number     | 1.364  | 1.965  | 0      | 27.27  | -0.031 | -0.026 |
| LIWC Otherp     | 1.901  | 3.802  | 0      | 163.77 | 0.014  | -0.043 |
| LIWC Perception | 11.3   | 5.608  | 0      | 43.24  | -0.021 | 0.016  |
| LIWC Period     | 7.001  | 4.89   | 0      | 245.71 | 0.003  | 0.019  |
| LIWC Physical   | 1.785  | 2.432  | 0      | 23.81  | -0.022 | -0.006 |
| LIWC Polite     | 0.38   | 0.995  | 0      | 10     | 0.043  | -0.044 |
| LIWC Politic    | 0.185  | 0.802  | 0      | 13.64  | 0.031  | -0.005 |
| LIWC       | Mean | Std Dev | Min | Median | Max | Skew | Kurtosis |
|------------|------|---------|-----|--------|-----|------|----------|
| LIWC Power | 0.855| 1.535   | 0   | 15.66  | 0.073| -0.054|
| LIWC Ppron | 11.167| 4.369 | 0 | 27.91 | -0.133| 0.064|
| LIWC Prep  | 13.841| 4.042 | 0 | 29.51 | -0.013| 0.035|
| LIWC Pronoun | 13.468| 5.147 | 0 | 32.65 | -0.115| 0.068|
| LIWC Prosocial | 0.887| 1.507 | 0 | 13.33 | 0.089| -0.019|
| LIWC Qmark | 0.061| 0.431 | 0 | 13.24 | -0.016| -0.002|
| LIWC Quantity | 3.614| 2.857 | 0 | 18.82 | -0.067| -0.018|
| LIWC Relig | 0.085| 0.561 | 0 | 17.65 | -0.001| -0.029|
| LIWC Reward | 0.228| 0.682 | 0 | 6.67 | 0.043| -0.012|
| LIWC Risk | 0.094| 0.432 | 0 | 7.69 | 0.028| -0.045|
| LIWC Sexual | 0.026| 0.246 | 0 | 7.81 | -0.032| -0.041|
| LIWC Shehe | 0.131| 0.767 | 0 | 13.89 | 0.08| -0.005|
| LIWC Socbehav | 4.371| 3.262 | 0 | 23.33 | 0.032| -0.019|
| LIWC Social | 11.563| 6.541 | 0 | 48.72 | 0.074| 0.028|
| LIWC Socrefs | 6.542| 5.332 | 0 | 36.17 | 0.07| 0.045|
| LIWC Space | 7.688| 4.578 | 0 | 30.3 | -0.016| 0.02|
| LIWC Substances | 0.084| 0.465 | 0 | 10.2 | 0.019| 0.018|
| LIWC         | Value1 | Value2 | Value3 | Value4 | Value5 | Value6 |
|--------------|--------|--------|--------|--------|--------|--------|
| LIWC Swear   | 0.025  | 0.213  | 0      | 4.23   | -0.058 | -0.02  |
| LIWC Tech    | 0.682  | 1.653  | 0      | 19.05  | 0.055  | -0.007 |
| LIWC Tentat  | 1.583  | 2.181  | 0      | 15.79  | -0.041 | 0.009  |
| LIWC They    | 0.283  | 0.863  | 0      | 10     | 0.035  | 0.024  |
| LIWC Time    | 3.959  | 2.977  | 0      | 24.39  | -0.088 | -0.01  |
| LIWC Tone    | 79.83  | 26.516 | 1      | 99     | 0      | 0.03   |
| LIWC Tone Neg| 0.318  | 0.921  | 0      | 9.38   | -0.043 | -0.045 |
| LIWC Tone Pos| 6.986  | 4.917  | 0      | 31.03  | 0.039  | 0.034  |
| LIWC Verb    | 15.177 | 5.054  | 0      | 36     | -0.09  | 0.119  |
| LIWC Visual  | 0.775  | 1.351  | 0      | 10.81  | 0.009  | 0.001  |
| LIWC Want    | 0.321  | 0.829  | 0      | 8.99   | -0.001 | -0.007 |
| LIWC Wordcount| 60.942 | 17.212 | 28     | 97     | -0.087 | -0.006 |
| LIWC We      | 1.479  | 3.23   | 0      | 22.58  | 0.04   | 0.029  |
| LIWC Wellness| 0.117  | 0.584  | 0      | 9.09   | -0.001 | -0.018 |
| LIWC Work    | 4.9    | 5.389  | 0      | 40     | 0.039  | -0.006 |
| LIWC Words per sentence | 15.624 | 6.985  | 3.47   | 97     | 0.014  | -0.059 |
| LIWC You     | 0.987  | 1.863  | 0      | 16.67  | 0.082  | 0.009  |
| Part Of Speech CC   | 3.646  | 1.772  | 0  | 21   | -0.042 | 0.056 |
|-------------------|--------|--------|----|------|---------|------|
| Part Of Speech CD  | 0.668  | 0.998  | 0  | 16   | -0.041  | -0.039 |
| Part Of Speech DT  | 4.359  | 2.361  | 0  | 17   | -0.036  | 0.06  |
| Part Of Speech EX  | 0.038  | 0.201  | 0  | 2    | 0.007   | 0.016 |
| Part Of Speech FW  | 0.019  | 0.167  | 0  | 7    | -0.012  | -0.03  |
| Part Of Speech IN  | 6.393  | 3.068  | 0  | 22   | -0.074  | -0.001 |
| Part Of Speech JJ  | 6.444  | 3.134  | 0  | 23   | -0.021  | -0.059 |
| Part Of Speech LS  | 0      | 0.012  | 0  | 1    | -0.007  | -0.012 |
| Part Of Speech MD  | 0.523  | 0.833  | 0  | 7    | -0.011  | 0.021 |
| Part Of Speech NN  | 18.628 | 6.965  | 3  | 51   | -0.024  | -0.076 |
| Part Of Speech PD  | 0.048  | 0.229  | 0  | 2    | 0.001   | 0.031 |
| Part Of Speech PO  | 0.092  | 0.339  | 0  | 6    | -0.007  | -0.002 |
| Part Of Speech PR  | 3.171  | 2.373  | 0  | 20   | -0.014  | 0.022 |
| Part Of Speech RB  | 3.026  | 2.459  | 0  | 20   | -0.124  | -0.01  |
| Part Of Speech RP  | 0.254  | 0.538  | 0  | 4    | -0.066  | 0.006 |
| Part Of Speech SY  | 0.003  | 0.053  | 0  | 2    | 0.01    | -0.005 |
| Part Of Speech TO  | 1.992  | 1.508  | 0  | 13   | -0.013  | 0.056 |
|                              | Value1 | Value2 | Value3 | Value4 | Value5 | Value6 |
|------------------------------|--------|--------|--------|--------|--------|--------|
| Part Of Speech UH           | 0.011  | 0.11   | 0      | 3      | -0.019 | 0      |
| Part Of Speech VB           | 11.998 | 4.558  | 0      | 29     | -0.135 | 0.068  |
| Part Of Speech WD           | 0.194  | 0.474  | 0      | 6      | 0.018  | 0.038  |
| Part Of Speech WP           | 0.286  | 0.599  | 0      | 5      | -0.018 | 0.016  |
| Part Of Speech WR           | 0.218  | 0.503  | 0      | 4      | -0.025 | 0.018  |
| Contains List               | 2.25   | 2.125  | 0      | 26     | 0.024  | -0.093 |
| Number Negations            | 0.165  | 0.449  | 0      | 5      | -0.055 | -0.013 |
| Number Of Addresses         | 0.003  | 0.053  | 0      | 1      | 0.005  | 0.021  |
| Number Of Names             | 0      | 0.012  | 0      | 1      | 0.024  | 0.012  |
| Number Of Numbers           | 0.783  | 1.559  | 0      | 30     | -0.006 | -0.034 |
| Number Of Punctuation       | 8.255  | 5.18   | 0      | 174    | -0.019 | -0.043 |
| Number Of Question Marks    | 0.04   | 0.277  | 0      | 9      | -0.026 | -0.005 |
| Number Of Symbols           | 0.108  | 1.48   | 0      | 107    | 0.005  | -0.005 |
| URL Count                   | 0.004  | 0.083  | 0      | 4      | 0.001  | -0.021 |
| Flesch Kincaid Grade Level  | 7.363  | 3.386  | 0      | 32.9   | 0.088  | -0.113 |
| Flesch Reading Ease Level   | 69.988 | 16.841 | -23.45 | 111.78 | -0.119 | 0.129  |
| Sentiment AFINN             | 8.642  | 6.309  | -16    | 44     | 0.014  | 0.047  |
| Sentiment          | NRC 1  | NRC 2 | 0  | 0.22 | -0.018 | -0.032 |
|--------------------|--------|-------|----|------|--------|--------|
| NRC Anger          | 0.009  | 0.019 | 0  | 0.22 | -0.018 | -0.032 |
| NRC Anticipation   | 0.077  | 0.054 | 0  | 0.316| -0.005 | 0.01   |
| NRC Disgust        | 0.007  | 0.017 | 0  | 0.22 | 0.002  | -0.023 |
| NRC Fear           | 0.012  | 0.023 | 0  | 0.22 | 0.007  | -0.049 |
| NRC Joy            | 0.107  | 0.076 | 0  | 0.5  | -0.016 | 0.049  |
| NRC Negative       | 0.021  | 0.03  | 0  | 0.304| -0.012 | -0.055 |
| NRC Positive       | 0.195  | 0.085 | 0  | 0.571| 0.06   | 0.036  |
| NRC Sadness        | 0.017  | 0.026 | 0  | 0.222| -0.016 | -0.014 |
| NRC Surprise       | 0.025  | 0.032 | 0  | 0.286| 0.004  | -0.012 |
| NRC Trust          | 0.093  | 0.062 | 0  | 0.429| 0.061  | -0.001 |
| Polarity           | 0.262  | 0.161 | -0.443 | 1     | 0.025  | 0.018  |
| Subjectivity       | 0.51  | 0.148 | 0  | 1    | -0.003 | 0.005  |
| Vader              | 0.812  | 0.265 | -0.895 | 0.998 | -0.021 | 0.015  |
| Lexical Diversity  | 0.755  | 0.079 | 0.167 | 1     | -0.016 | -0.202 |
| Character Count    | 341.203 | 107.151 | 126 | 705  | -0.025 | -0.058 |
| Contractions Count | 1.021  | 1.439 | 0  | 12   | -0.152 | 0.02   |
| Line Break Count   | 0.986  | 1.76 | 0  | 26   | 0.05   | 0.041  |
|                                | 1.973 | 1.249 | 1     | 45    | 0.071 | 0.126 |
|--------------------------------|-------|-------|-------|-------|-------|-------|
| Longest Repetition Length     | 16    | 7.546 | 3.737 | 89    | 0.01  | -0.066|
| Mean Sentence Length          |       |       |       |       |       |       |
| Mean Word Length              | 4.565 | 0.559 | 3.265 | 7.933 | 0.142 | -0.157|
| Number Of Exclamation Marks   | 0.39  | 0.843 | 0     | 21    | -0.033| -0.03 |
| Number Of Unique Words        | 46.039| 11.648| 15    | 77    | -0.113| -0.089|
| Percentage Common 2-grams     | 0.048 | 0.055 | 0     | 0.385 | -0.046| 0.106 |
| Percentage Common 3-grams     | 0.029 | 0.046 | 0     | 0.375 | -0.025| 0.113 |
| Percentage Common 4-grams     | 0.011 | 0.043 | 0     | 1     | -0.039| 0.092 |
| Percentage Common Words       | 0.156 | 0.096 | 0     | 0.688 | -0.04 | 0.12  |
| Percentage Rare 2-grams       | 0.691 | 0.153 | 0     | 1     | 0.082 | -0.207|
| Percentage Rare Words         | 0.065 | 0.066 | 0     | 0.529 | 0.069 | -0.223|
| Percentage Stop Words         | 0.476 | 0.075 | 0     | 0.733 | -0.127| 0.181 |
| Word Density                  | 0.183 | 0.018 | 0.112 | 0.241 | -0.159| 0.151 |
| LDA Topic Vectors             | Various techniques incl. structural topic models were explored but not used due to robustness and interpretability issues. |
Figure S1. Participants’ performance in identifying generated self-presentations did not change throughout the experiment. Error bars represent 95% confidence intervals for 6,000–16,000 judgments of 2,000–3,000 self-presentations per bar.

Figure S2. Participants in a separate labeling task rated AI-generated self-presentations as nonsensical and repetitive more often than human-written self-presentations. Error bars represent 95% confidence intervals for 1,898–4,704 judgments of 1,000–1,500 self-presentations per bar.