INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE
IFPRI

CGIAR INITIATIVE ON National Policies and Strategies

CGIAR INITIATIVE ON Digital Innovation

IFPRI Discussion Paper 02321

December 2024

Man vs. Machine

Experimental Evidence on the Quality and Perceptions of AI-Generated Research Content

Michael Keenan
Jawoo Koo
Christine Mwangi
Naureen Karachiwalla
Clemens Breisinger
MinAh Kim

Development Strategies and Governance Unit
Poverty, Gender, and Inclusion Unit
Natural Resources and Resilience Unit

INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE

The International Food Policy Research Institute (IFPRI), a CGIAR Research Center established in 1975, provides research-based policy solutions to sustainably reduce poverty and end hunger and malnutrition. IFPRI's strategic research aims to foster a climate-resilient and sustainable food supply; promote healthy diets and nutrition for all; build inclusive and efficient markets, trade systems, and food industries; transform agricultural and rural economies; and strengthen institutions and governance. Gender is integrated in all the Institute's work. Partnerships, communications, capacity strengthening, and data and knowledge management are essential components to translate IFPRI's research from action to impact. The Institute's regional and country programs play a critical role in responding to demand for food policy research and in delivering holistic support for country-led development. IFPRI collaborates with partners around the world.

AUTHORS

Michael Keenan (m.keenan@cgiar.org) is an Associate Research Fellow in the Development Strategies and Governance (DSG) Unit of the International Food Policy Research Institute (IFPRI), Nairobi, Kenya.

Jawoo Koo (j.koo@cgiar.org) is a Senior Research Fellow in IFPRI's Natural Resources and Resilience (NRR) Unit, Washington, DC.

Christine Mwangi (c.mwangi@cgiar.org) is a Research Officer in IFPRI's DSG Unit, Nairobi, Kenya.

Naureen Karachiwalla (n.karachiwalla@cgiar.org) is a Research Fellow in IFPRI's Poverty, Gender, and Inclusion (PGI) Unit, Nairobi, Kenya.

Clemens Breisinger (c.breisinger@cgiar.org) is the Program Leader for the Kenya Strategy Support Program and a Senior Research Fellow in IFPRI's DSG Unit, Nairobi, Kenya.

MinAh Kim (minahkim.official@gmail.com) is a Data Scientist, Washington, DC.

Abstract

Academic researchers want their research to be understood and used by non-technical audiences, but that requires communication that is more accessible in the form of non-technical and shorter summaries. The researcher must both signal the quality of the research and ensure that the content is salient by making it more readable. AI tools can improve salience; however, they can also lead to ambiguity in the signal since true effort is then difficult to observe. We implement an online factorial experiment providing non-technical audiences with a blog on an academic paper and vary the actual author of the blog from the same paper (human or ChatGPT) and whether respondents are told the blog is written by a human or AI tool. Even though AI-generated blogs are objectively of higher quality, they are rated lower, but not if the author is disclosed as AI, indicating that signaling is important and can be distorted by AI. Use of the blog does not vary by experimental arm. The findings suggest that, provided disclosure statements are included, researchers can potentially use AI to reduce effort costs without compromising signaling or salience.

Keywords: Artificial intelligence, communication, academic research

Acknowledgments

We thank James Allen IV, Kate Ambler, Todd Benson, Carly Trachtman, Jessica Leight, and participants from the IFPRI brownbag seminar series. We also thank IFPRI country office leads for Bangladesh, Egypt, Ethiopia, Ghana, India, Kenya, Malawi, Nigeria, Rwanda, Sudan, and Uganda, Michael Go, and Omar Fares. We gratefully acknowledge funding from CGIAR Initiatives on National Policies and Strategies (NPS) and Digital Innovation (DI) and the Bill & Melinda Gates Foundation through the Generative AI for Agriculture project (INV-047346).

1 Introduction

Academic research is only valuable to the extent that it is used by the audiences for whom it is intended. For many researchers, the ultimate goal is for readers to absorb their findings and engage with them – for example, by using findings in their own research or in practical decision-making. Many readers tend not to read full academic papers; rather, non-technical summaries such as blogs are preferred Arnautu and Dagenais (2021). Generative Artificial Intelligence (GAI) can improve communication by lowering researchers' cost of effort in creating those communications and making research more accessible. GAI has been shown to reduce knowledge workers' time by 18% and improve expert-assessed writing quality by 40% (Noy and Zhang, 2023) . However, trust in GAI remains a concern, with only 39% of people globally trusting AI-generated information since AI can make errors, spread misinformation, or make up information (Buchanan et al. 2023, Gillespie et al. 2023, Ji et al. 2023). Indeed, the broad scientific community has been grappling with ensuring the responsible use of AI in research (Altmäe et al., 2023,Crawford et al., 2023, Salvagno et al., 2023).

Gans (2024) provides a useful framework by which to think about such communication. Writers derive utility from having their work read and valued and expend effort in *signaling* the value of their research, while readers benefit from consuming *accessible* (in Gans' terms, salient) information and expend effort consuming technical information. GAI reduces the cost of effort for writers to communicate to readers and may reduce the cost of effort to readers by simplifying complex content and making it more accessible. Signaling the value of their work, often achieved through demonstrating their effort, is a critical challenge for writers. Effort signals enhance the likelihood of their work being read and trusted. The integration of GAI complicates this signaling mechanism by introducing information asymmetries – its use creates ambiguity about the writer's effort and thus the value of the work. It enables low-effort outputs to become accessible and appear indistinguishable from high-effort, well-researched work, potentially eroding traditional cues of quality and effort. Disclosing true authorship eliminates ambiguity, but the impact of revealing GAI use on the perceived value of communication depends on whether readers value transparency in effort over penalizing low apparent effort – this is an empirical question.

In this paper, we examine the context of researchers communicating their work to various consumers of research (such as other academics, policymakers, and other expert stakeholders). We test whether using AI for research communication affects the perceived quality of it and whether a transparent signal can mediate the potential ambiguity from using AI to communicate by incorporating a clear disclosure statement on the true author, which is now required in many journals and organizations. We provided a broad audience with a blog based on an academic paper that has a corresponding human-written blog and generated an AI version of the blog from the same paper using ChatGPT-4. An online survey was then distributed in eleven countries where individuals read the blogs, rated their quality, and stated the degree to which they intended to engage with the research. Through a four-arm factorial design, respondents randomly received a human-written or AI-generated blog (testing perceived accessibility/quality) and were randomly told that the blog was written by a human or was generated using AI (testing signaling). Using the group to whom the true author was disclosed, we can test the effect of disclosure on reducing ambiguity.

We find that, while the AI-generated blogs are objectively more accessible – they are written at a lower grade level (Flesch, 1948) – they are rated lower in quality. Weak signaling, when the true author is unknown, may have introduced ambiguity about the author, reducing perceived quality of the blog. The negative effect of AI-written blogs is completely offset if respondents are told that the blog was AI-generated, indicating that a disclosure statement (a truthful signal) can successfully remove the negative effects of ambiguity. On engagement, in all four experimental arms, we cannot reject that the difference in the likelihood of engagement

across groups is zero. Consequently, provided that disclosure statements are included, researchers may reduce their cost of effort by using AI for communication without compromising the value of the communication or engagement with it.

We contribute to a well-established literature on communication frictions, focusing on asymmetric information. We study a relatively recent but increasingly used method in communication, GAI, and the potential of reducing the associated frictions that can emerge from it, even if readers and writers have congruent interests (Dewatripont and Tirole, 2005). A simple solution exists; while AI disclosure statements that are now required by journals and many organizations were originally meant to serve ethical concerns, we show that they can also reduce information asymmetries that hinder communication since they provide a truthful signal about the value of certain communications.

We also contribute to the burgeoning literature on the effects of GAI on knowledge work. Recent evidence has centered around the role of GAI in affecting knowledge workers' jobs and productivity (Noy and Zhang, 2023) and the accuracy and potential misinformation that AI may generate (Sandrini and Somogyi, 2023). There is scant causal evidence on the effects of GAI on very high-level knowledge work — that of academics and other researchers. Our paper is most closely related to Buchanan and Hickman (2024), who conducts an incentivized experiment that varies both the actual and reported authorship of a paragraph, asking participants to evaluate their trust in its content. Our findings complement theirs, demonstrating that participants do not distrust AI-written text when the author's identity is disclosed. We build on their work by employing an experiment in a real-world context across a sample of eleven countries and by exploring participants' reported likelihood of future engagement with the content.

2 Conceptual Framework

We follow the framework of Gans (2024) as well as insights from other literature on the roles of signaling and accessibility in affecting a reader's perceived quality and thus future engagement with research. Researchers decide whether or not to communicate their research using a blog, and readers decide whether or not to engage with it. Researchers receive utility from a reader absorbing and using their research and base their decision to write a blog on the probability a reader will engage with the blog. However, they incur a cost of effort from producing the blog.

Readers gain utility from accessible and valuable information (and base their decision to engage with it on the probability the information is useful), but incur a cost of effort from absorbing and/or using it. They are more likely to engage with the blog if it is written in an accessible way (less jargon, a catchy title, etc.). Readers have limited attention and energy to process information, preferring concise and easy to digest (accessible) information, which lowers the cost of reading. Researchers can invest in making communication more accessible, increasing the costs to the researcher, but lowering readers' costs.

Readers do not know the probability a blog will be valuable a priori and rely on signals to assess this probability. Researchers' effort is a key signal used to communicate a blog's potential value[1] – the higher the perceived cost of effort, the greater the perceived value, and the more likely a reader is to engage with the blog. Researchers know their true cost of effort but readers do not, resulting in information asymmetries.

GAI disrupts this traditional signaling for both readers and researchers – it can lower both the reader's and researcher's costs simultaneously, removing the need for higher investment in making blogs accessible. However, it can also worsen information asymmetries related to the signaling of researcher's effort – with

[1]There are other ways identified in the literature that can influence signaling, such as including one's credentials.

GAI researchers can masquerade low-effort work as high-effort work. When signals are ambiguous, readers find it more challenging to quickly discern value; they assume average quality, and reduce their engagement with research communication.[2] The researcher should not use AI if the cost of generating the blog using AI (including the lost benefit of a reader engaging with the research) is higher than the cost of writing it themselves. The researcher can, however, choose whether or not to disclose that they used AI to write the blog. They should do so only if the value gained from reducing ambiguity is greater than the value lost from revealing low effort. This is the question we test.

3 Experiment Design

Our experiment randomly varies both signaling and accessibility to assess how consumers of research perceive the quality of a blog and the extent to which they report planning to engage with it in the future. We implement an online experiment with two cross-randomized interventions: readers received a blog written either by a researcher or by ChatGPT-4 (hereafter GAI), and participants were either told the blog was written by a human or a GAI tool (even if it was not).[3] We leverage academic papers, blogs, and a list of stakeholders from the International Food Policy Research Institute (IFPRI). We used ten academic papers (one for each of eleven countries) written in the past five years that had an existing human-written blog summarizing the publication. The papers focused on topics related to the food system to reduce the likelihood that impacts were driven by substantial variation in topics.[4]

Before presenting the blog, we provided a disclosure statement based on Elsevier journal regulations. [5] The four treatment arms are (actual author is the first word and reported author is the second):

- Human-Human and Human-AI: "This blog was written by a team of local and international researchers."
- AI-Human and AI-AI: "During the preparation of this blog, a team of local and international researchers used Artificial Intelligence to generate the content and structure of the blog. The authors have reviewed and edited the content as needed and take full responsibility for the content of the blog."

The AI-AI and Human-Human arms test accessibility (compare the perceived quality of the blog due to factors such as catchy titles and less jargon) and AI-Human and Human-AI test for signaling (reporting the author acts as a signal to the reader about how much effort the researcher made).

Table 1: Experiment Design

		Note Type	
		Human-Written	AI-Written
Author	Human-Written	Human-Human	AI-Human
	AI-Written	Human-AI	AI-AI

[2] This dynamic parallels that of Akerlof (1970) in his seminal paper "The Market for Lemons."

[3] After the survey, participants were told who the true author was.

[4] "A food system comprises all the elements (environment, people, inputs, processes, infrastructures, institutions, etc.) and activities that relate to the production, processing, distribution, preparation and consumption of food, and the outputs of these activities, including socio-economic and environmental outcomes." (The High Level Panel of Experts, 2017)

[5] The communication for the AI-assisted briefs is adapted from the journal Food Policy's AI statement (contained in Appendix B).

AI-written blogs were generated using detailed uniform prompts in ChatGPT-4 (see Appendix C for the full prompt). Prompts were input using a piecemeal, iterative process – the researcher first prompted GPT-4 to create a title, then an introduction and research overview, key findings, policy recommendations, and a conclusion. The prompts were derived from best practices in prompt engineering (Giray, 2023, Korzynski et al., 2023).

The research team inserted the same figures from the source paper because, in practice, the researcher would select the most important figures from the paper even if they used GAI. They also conducted a basic check to ensure there were no glaring errors (there were not). AI content creators did not view the human-written blogs at all. AI blogs were of the same length as the human blog with an absolute difference of less than 60 words ($< 10\%$) and were on a topic of high relevance to readers in that country. The format of the blogs was also standardized with no differences in branding or formatting, and importantly, did not have any association with IFPRI.

An online survey with 366 participants was conducted using IFPRI mailing lists in eleven countries where IFPRI has a local office as well as posts on social media (see Appendix D for full text of emails and social media posts). The eleven countries were: Bangladesh, Egypt, Ethiopia, Ghana, India, Kenya, Malawi, Nigeria, Rwanda, Sudan, and Uganda. Participants were randomized into treatment groups following a systematic, stratified randomization procedure. Each respondent entered basic demographic and work characteristics (gender, age, education level, position level, and type of organization) and were individually dynamically randomized within these strata. The first respondent in each stratum was randomly assigned to one of the four treatment groups, the second to one of the three remaining, and so forth.

The study was conducted in four waves. The initial wave commenced in the first week of October 2023, followed by the second, third, and fourth waves in the first week of November 2023, second week of January 2024, and third week of March 2024, respectively. During the first and fourth waves, each country office sent out emails containing information about the study objectives and the country-specific survey links to their respective mailing lists. A reminder email was sent approximately two weeks after the initial email. The second and third waves employed a social-media recruitment strategy via IFPRI's LinkedIn and X accounts. To incentivize participation, the introductory email noted that ten randomly selected people who completed the survey would win 100 USD.[6]

3.1 Empirical Specification

We estimate an OLS model with indicator variables representing whether the blog was reported as having been written by AI, was actually written by AI, and their interaction. The specification is as follows:

$$Y_i = \beta_0 + \beta_1 AI_{reported,i} + \beta_2 AI_{generated,i} + \beta_3 AI_{reported,i} * AI_{generated,i} + \delta_j + \varepsilon_i \quad (1)$$

where Y_i is the outcome of interest, $AI_{reported,i}$ is an indicator variable equal to one if respondents were told that the blog was generated using AI and $AI_{generated,i}$ is an indicator variable equal to one if the blog was AI-generated. β_1 identifies the effect of respondents being told the blog was written by AI (even if it was not – signaling) and β_2 identifies the effect of the actual author (accessibility) on perceived quality. The parameter β_3 captures how the effect of the actual author being AI differs between those who are told that the author is AI and those who are told that the author is human. We include strata and wave fixed effects heteroskedasticity-robust standard errors.

[6]These individuals have been paid.

We report conventional standard errors as well as RI p-values in squared brackets. For interactions, we permute each variable in the interaction separately and report both (or all three). We also conduct heterogeneity analysis across some dimensions by interacting each of the three variables in Equation 1 with that dimension of heterogeneity. Comparing the interactions between that dimension and each of "Author is AI" and "Reported Author is AI" allows us to measure the difference between those with that attribute and those without (for example between men and women or between those with high policy influence and those without).

We assess balance in demographic characteristics across the four experimental arms, and, across 18 variables and 108 tests (Appendix Table A.1), we find only six statistically significant coefficients, indicating that the experiment was balanced.

3.2 Outcomes

We examine the respondents' perceived quality of the blog, planned future engagement with the blog, and their beliefs about other readers' future engagement.

Several perceived quality measures were calculated; the first two were pre-specified in the American Economic Association Randomized Controlled Trial Registry (AEA RCT Registry) and the remaining three serve as robustness checks. First, prior to presenting the blog, we elicited respondents' preferences regarding 9 blog characteristics, such as whether the title sparked curiosity, the adequacy of the length, clarity, relevance, and the sufficiency of the policy recommendations. Later, respondents rated the blog's quality based on these same attributes. These ratings were measured on a Likert scale from 0 to 4, where 4 represented "very important" and 0 represented "not important at all." We then constructed a "weighted average of quality" by summing the total across attributes and multiplying each attribute's rating by its relative importance, which was determined by its contribution to the total of the respondent's preference ratings.

Next, we create two indices based on 12 statements rated on a Likert scale of 0-4 with 0 representing "strongly disagree" and 4 representing "strongly agree".The measures are based on five criteria prescribed as being important to the measurement of quality of a communication and valued aspects of academic research communication (Cash et al., 2003, Balian et al., 2016, and Fillol et al., 2022).

Our questions included both signaling (blog represented a range of views, was relevant to their work) and accessibility aspects (had an appropriate amount of detail, a clear rationale, a catchy title, good visual presentation, an appropriate tone, of an adequate length, easy to understand, and provided clear, relevant, and sufficiently detailed recommendations).

From these questions, we calculate: the unweighted mean score across the 13 statements, the first principal component of the 13 statements using the first component of a polychoric principal components analysis, the number of questions for which participants responded with "strongly agree" or "agree," and an indicator equal to one if the participant rated the blog as being of "very high" or "high" quality overall in a separate question. We also examined their reported future engagement with the blog with six statements on the a Likert scale ranging from 0-4 with 0 representing "very unlikely" to x representing "very likely." The questions were: whether they would share the blog with others, re-read the blog, look up studies cited in the blog, contact the authors, and use the blog in decision-making, etc. We created measures analogous to those above, except the fifth one.

Finally, we constructed outcome variables related to the participants' perceptions of how others would engage with the blog using the same questions but about others' engagement and also constructed the outcomes analogously.

We also compute a variable reflecting the readability of the blogs from (Flesch, 1948). The Flesch–Kincaid Grade Level formula estimates the U.S. school grade level required to understand a piece of text. It is calculated based on the average number of words per sentence and syllables per word, with higher values indicating more complex text suitable for higher-grade readers. Note that there is no existing grade-level calculation for low-income countries, and accordingly, the Flesch–Kincaid Grade Level formula is our best attempt at standardizing an objective measure of accessibility across countries.

4 Results

This section reports the key empirical results. We first describe the sample and outcome measures. We then present the findings regarding how respondents perceived the quality and value of the research presented in the blog, examine their stated intentions about engaging with the research, and finally, their reported beliefs about how others may engage with the research.

4.1 Characteristics of the Sample

Prior to reading the blog, we asked respondents to rank characteristics on a scale of 1-5 that they preferred in a blog. The highest ranked were clear recommendations (mean 4.66), easy to understand (4.61), and relevance to their work (4.5). Interestingly, the reputation of the authors was ranked of least priority (average of 3.78), see Appendix Table A.2. Perceptions on the quality of the blogs are moderate. The average quality on a scale of 0-4 is 2.9 and the proportion of respondents rating the blogs as "high" or "very high" quality is 60%. They reported a preference for an average length of 7.8 pages with 6.5 figures (much longer than the blogs actually were). Measures of engagement are middling – the average likelihoods of engagement fall within a range of 2.1 to 2.8, with measures of others' likelihood of engaging on average half a point higher. Note also that, in our sample, 60% of respondents had used GAI in the past, primarily for shortening text or correcting grammar.

4.2 Perceptions of Quality

Table 2 reports on how the actual author (representing accessibility) and the disclosed author (representing signaling) affects our five measures of perceptions of quality. Across outcomes, there are no statistically significant differences in perceived quality by whether the reported author is AI. Perceived quality is lower when the blog is AI-generated; the coefficient is consistently negative and is statistically significant. However, for the group where the true author is disclosed as AI, the negative perception of AI-generated content is reversed. This finding suggests that readers in this sample penalize ambiguity, but the cost of ambiguity is lower than the penalty for revealing the low effort associated with using AI. Consequently, researchers can benefit from the reduced cost of effort of using AI without incurring the penalty of signaling low effort provided they use a disclosure statement.

Table 2: Perceived Quality of Policy Blog

	(1)	(2)	(3)	(4)	(5)
	Unweighted Average Quality (0-4)	Weighted Quality Average	Standardized First Principal Component	Number of 'agree' and 'strongly agree' statements (0-12)	Overall quality of brief rated high or very high
(1) Told author is AI	-0.065	-0.045	-0.084	-0.515	-0.104
	(0.093)	(0.092)	(0.144)	(0.513)	(0.075)
RI p-value	[.455]	[.61]	[.537]	[.3]	[.15]
(2) Actual author is AI	-0.198**	-0.232**	-0.272*	-1.245**	-0.164**
	(0.097)	(0.096)	(0.146)	(0.518)	(0.071)
RI p-value	[.034]	[.011]	[.061]	[.02]	[.02]
(3) Told author is AI and Actual author is AI	0.205	0.199	0.263	1.373*	0.212**
	(0.134)	(0.134)	(0.205)	(0.725)	(0.105)
RI p-value (permuting reported author)	[.113]	[.132]	[.186]	[.042]	[.049]
RI p-value	[.125]	[.129]	[.211]	[.066]	[.043]
Mean of Human-Human (permuting actual author)	2.972	3.008	.149	9.86	.702
P-value (1) - (2)	.19	.065	.224	.169	.428
N	366	366	366	366	366

Notes: * p<0.1, ** p<0.05, *** p<0.01. This table reports on the perceived quality of the policy brief. The outcome in the first column is the unweighted average rating across twelve statements. Ratings for each statement follow a likert scale ranging from 0 to 4, with 0 indicating strongly disagree and 4 indicating strongly agree. The second column presents the weighted average across nine statements. The outcome in the third column presents the score of the first principal component, calculated based on the same twelve statements using polychoric PCA. The fourth column indicates the total number of statements rated as either 'agree' or 'strongly agree', out of 12 perception statements. The fifth column presents the results from a separate survey question asking respondents to rate the overall quality of the brief from 1 to 5. The outcome variable in the fifth column is an indicator equal to 1 if the quality of the brief is rated either 'high' or 'very high'. Strata fixed effects (countries) are controlled for and standard errors are clustered at the individual level (the unit of randomization).

Appendix Table A.3 shows how respondents in the experimental groups rate the individual measures comprising the indices. The statistically significant components rated lower when AI writes blogs: title sparked curiosity, appropriate length and detail, easy to understand content, and representation of a range of views. All but the last reflect accessibility, suggesting that, contrary to the assumption made in Gans (2024) AI reduces the perceived accessibility to readers in this context. This sample comprises highly educated respondents (80% have a masters degree or above), many with technical expertise in the blog's subject matter. As a result, they may prefer technical language and the inclusion of more complex concepts (recall that the preferred length of the blog was higher than the length of the actual blogs), suggesting preferences about salience are audience-dependent and researchers should thus tailor their prompts accordingly.

We examine heterogeneity using an objective measure: the grade level of the writing calculated using the Flesch-Kincaid grade-level reading score. The AI-generated blogs are more accessible: the average grade level of the AI-generated blogs is 14.6 and that of human-written blogs is 16.9. Additionally, the words per sentence (20 vs 27) and paragraphs per blog (12.6 vs 13.8) are lower among the AI-generated blogs (however sentences per paragraph (3.6 vs 2.2) is higher). When we conduct a heterogeneity test by the grade level of

the blog, we do not find that grade level has differential impacts on perceived quality, consistent with the findings above (see Appendix Tables A.6 and A.7).

Related to but outside of the Gans (2024) model, Buchanan and Hickman (2024) also investigates the role of trust, which arguably also affects perceived value. Respondents reported whether they trust AI to write high-quality blogs (60% of the sample do). When we conduct a heterogeneity analysis adding interactions of an indicator variable equal to one when a respondent reports trusting AI, we do not find evidence of trust in AI affecting perceived quality (see Appendix Tables A.10 and A.11).

Going further, readers' trust may depend on whether they have previously used an AI tool Buchanan and Hickman (2024) (in this sample, 75% have) and indeed, those who have used an AI tool perceive AI-written blogs as being lower quality (see Appendix Tables A.8 and A.9).[7]

4.3 Engagement

Table 3 shows that there are no statistically significant impacts on whether or to what extent a respondent intends to engage with the blog in any of the treatments. The same is true for how respondents believe others may engage with the blog (Table 4). Since neither the actual or reported author affects readers' engagement and, *if* a disclosure statement is provided, researchers may as well use AI to communicate, since it reduces their cost of effort without compromising perceptions of quality or engagement.[8]

[7]Gender may also play a role in perceptions of AI, as shown in work by (Aldasoro et al., 2024). Our pre-analysis plan specified that we would examine this difference. We do not find evidence of this channel; see Appendix Tables A.12 and A.17.

[8]Again, there may be several reasons. The topic may not be relevant, the respondent may not be the one who actually makes the policy decisions, or their perceived quality of the blog may be low. We check each of these dimensions of heterogeneity with respect to their own and others' engagement and find no evidence of any of these channels (results available upon request). Again, there is no heterogeneity in impacts by gender (see Appendix Tables A.16, A.17, A.20, and A.21.)

Table 3: Intended in Engagement with Policy Blog

	(1)	(2)	(3)
	Average engagement rating (0-4)	Standardized First Principal Component	Number of 'likely' and 'very likely' statements (0-6)
(1) Told author is AI	0.008	0.022	0.019
	(0.123)	(0.148)	(0.305)
RI p-value	[.947]	[.887]	[.958]
(2) Actual author is AI	-0.151	-0.165	-0.365
	(0.124)	(0.149)	(0.302)
RI p-value	[.241]	[.275]	[.235]
(3) Told author is AI and Actual author is AI	0.045	0.029	0.106
	(0.174)	(0.208)	(0.427)
RI p-value	[.793]	[.891]	[.797]
RI p-value	[.804]	[.89]	[.806]
Mean of Human-Human	2.73	.092	3.905
P-value (1) - (2)	.212	.218	.217
N	366	366	366

Notes: * p<0.1, ** p<0.05, *** p<0.01. This table reports the likelihood of engagement with the policy brief. The outcome in the first column is the average rating across six statements, concerning the intended future engagement with the brief. The responses follow a likert scale ranging from 0 to 4, with 0 indicating very low likelihood and 4 indicating very high likelihood of engagement. The second column presents the score of the first principal component, calculated based on the same six statements. The third column represents the number of questions where the respondent indicated they are 'likely' or 'very likely' to engage with the policy brief. Strata fixed effects (countries) are controlled for and standard errors are clustered at the individual level (the unit of randomization).

Table 4: Belief in Others Intended Engagement with Policy Blog

	(1)	(2)	(3)	(4)
	Average rating (0-4)	Standardized First Principal Component	Total likelihood of engagement (0-6)	Others overall rating'
(1) Told author is AI	-0.061	-0.071	-0.068	-0.032
	(0.123)	(0.152)	(0.333)	(0.076)
RI p-value	[.605]	[.629]	[.824]	[.66]
(2) Actual author is AI	-0.049	-0.050	-0.150	-0.121
	(0.119)	(0.148)	(0.329)	(0.076)
RI p-value	[.694]	[.741]	[.652]	[.124]
(3) Told author is AI and Actual author is AI	-0.012	-0.029	-0.046	0.030
	(0.174)	(0.212)	(0.464)	(0.108)
RI p-value	[.93]	[.879]	[.903]	[.768]
RI p-value	[.953]	[.905]	[.902]	[.758]
Mean of Human-Human	3.627	.091	3.619	.679
P-value (1) - (2)	.927	.895	.811	.258
N	366	366	366	366

Notes: * p<0.1, ** p<0.05, *** p<0.01. This table reports respondents' beliefs of their colleagues' likelihood of engagement with the policy brief. The outcome in the first column is the average rating across six statements, concerning their colleagues' intended future engagement with the brief. The responses follow a likert scale ranging from 0 to 4, with 0 indicating very low likelihood and 4 indicating very high likelihood of colleagues' engagement. The second column presents the score of the first principal component, calculated based on the same six statements. The third column represents the number of questions where the respondent indicated they believe their colleagues are 'likely' or 'very likely' to engage with the policy brief. Strata fixed effects (countries) are controlled for and standard errors are clustered at the individual level (the unit of randomization).

5 Conclusion

In a randomized experiment comparing the actual and reported authorship of research blogs (AI vs. Human), readers from eleven countries rated AI-generated content as lower in quality than human-written content. However, when an AI-disclosure statement was included, readers did not penalize AI-generated material, highlighting the importance of signaling in research communication. Ambiguity regarding authorship, on the other hand, was penalized. These findings suggest that researchers can benefit from using AI to create research communications, provided they disclose its use. The potential time savings and accessibility improvements offered by AI may outweigh the perceived quality reduction due to ambiguity in authorship, especially since consumers of academic research tend to be more educated. Nevertheless, caution is advised when using AI, as it is still prone to errors, trust in AI is low, and the field is rapidly evolving.

References

Akerlof, G. A. (1970). The Market for "Lemons": Quality Uncertainty and the Market Mechanism. *The Quarterly Journal of Economics*, 84(3):488–500.

Aldasoro, I., Armantier, O., Doerr, S., Gambacorta, L., and Oliviero, T. (2024). The gen AI gender gap. *Economics Letters*, 241(June):11–13.

Altmäe, S., Sola-Leyva, A., and Salumets, A. (2023). Artificial intelligence in scientific writing: a friend or a foe? *Reproductive BioMedicine Online*, 47(1):3–9.

Arnautu, D. and Dagenais, C. (2021). Use and effectiveness of policy briefs as a knowledge transfer tool: a scoping review. *Humanities and Social Sciences Communications*, 8(1):1–14.

Balian, E. V., Drius, L., Eggermont, H., Livoreil, B., Vandewalle, M., Vandewoestjine, S., Wittmer, H., and Young, J. (2016). Supporting evidence-based policy on biodiversity and ecosystem services: Recommendations for effective policy briefs. *Evidence and Policy*, 12(3):431–451.

Buchanan, J. and Hickman, W. (2024). Do people trust humans more than ChatGPT? *Journal of Behavioral and Experimental Economics*, 112(November 2023):102239.

Buchanan, J., Hill, S., and Shapoval, O. (2023). ChatGPT Hallucinates Non-existent Citations: Evidence from Economics. *The American Economist*, 69(1):80–87.

Cash, D. W., Clark, W. C., Alcock, F., Dickson, N. M., Eckley, N., Guston, D. H., Jäger, J., and Mitchell, R. B. (2003). Knowledge systems for sustainable development. *Proceedings of the National Academy of Sciences of the United States of America*, 100(14):8086–8091.

Crawford, J., Cowling, M., Ashton-Hay, S., Kelder, J. A., Middleton, R., and Wilson, G. S. (2023). Artificial Intelligence and Authorship Editor Policy: ChatGPT, Bard Bing AI, and beyond. *Journal of University Teaching and Learning Practice*, 20(5).

Dewatripont, M. and Tirole, J. (2005). Modes of Communication. *Journal of Political Economy*, 113(6):1217–1238.

Fillol, A., McSween-Cadieux, E., Ventelou, B., Larose, M. P., Kanguem, U. B. N., Kadio, K., Dagenais, C., and Ridde, V. (2022). When the messenger is more important than the message: an experimental study of evidence use in francophone Africa. *Health Research Policy and Systems*, 20(1):1–17.

Flesch, R. (1948). A Readability Formula In Practice. *Elementary English*, 25(6):344–351.

Gans, J. S. (2024). How will Generative AI impact communication? *Economics Letters*, 242(July):111872.

Gillespie, N., Lockey, S., Curtis, C., Pool, J., and Akbar, A. (2023). Trust in Artificial Intelligence: A Global Study. Technical report, The University of Queensland and KPMG Australia.

Giray, L. (2023). Prompt Engineering with ChatGPT: A Guide for Academic Writers. *Annals of Biomedical Engineering*, 51:2629–2633.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12):1–38.

Korzynski, P., Mazurek, G., Krzypkowska, P., and Kurasinski, A. (2023). Artificial intelligence prompt engineering as a new digital competence : Analysis of generative AI technologies such as ChatGPT. 11(3).

Noy, S. and Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192.

Salvagno, M., Taccone, F. S., and Gerli, A. G. (2023). Can artificial intelligence help for scientific writing? *Critical Care*, 27(1):1–5.

Sandrini, L. and Somogyi, R. (2023). Generative AI and deceptive news consumption. *Economics Letters*, 232:111317.

The High Level Panel of Experts (2017). Nutrition and Food Systems. Technical Report September, The High Level Panel of Experts on Food Security and Nutrition.

Appendices

A Additional Tables

Table A.1: Balance

Variable	(1)	(2)	(3)	(4)		(1)-(2)	(1)-(3)	(1)-(4)	(2)-(3)	(2)-(4)	(3)-(4)
		Mean / (se)			F-stat/(p-value)				Mean Difference		
Age	43.081	41.330	41.857	43.354	0.517	1.751	1.224	-0.272	-0.527	-2.024	-1.496
	(1.618)	(1.269)	(1.454)	(1.381)	(0.671)						
Female	0.302	0.289	0.242	0.263	0.307	0.014	0.061	0.040	0.047	0.026	-0.021
	(0.050)	(0.046)	(0.045)	(0.044)	(0.821)						
Bachelor's or less	0.233	0.196	0.187	0.152	0.668	0.037	0.046	0.081	0.009	0.044	0.035
	(0.046)	(0.041)	(0.041)	(0.036)	(0.572)						
Master's	0.442	0.454	0.473	0.424	0.117	-0.012	-0.031	0.018	-0.019	0.029	0.048
	(0.054)	(0.051)	(0.053)	(0.050)	(0.950)						
PhD	0.326	0.351	0.341	0.424	0.756	-0.025	-0.015	-0.099	0.010	-0.074	-0.084
	(0.051)	(0.049)	(0.050)	(0.050)	(0.520)						
Junior level	0.233	0.237	0.209	0.253	0.223	-0.005	0.024	-0.020	0.028	-0.015	-0.044
	(0.046)	(0.043)	(0.043)	(0.044)	(0.881)						
Mid-level	0.256	0.247	0.286	0.212	0.409	0.008	-0.030	0.044	-0.038	0.035	0.074
	(0.047)	(0.044)	(0.048)	(0.041)	(0.746)						
Senior level	0.512	0.515	0.505	0.535	0.049	-0.004	0.006	-0.024	0.010	-0.020	-0.030
	(0.054)	(0.051)	(0.053)	(0.050)	(0.986)						
NGO	0.198	0.309	0.198	0.162	2.021	-0.112	-0.000	0.036	0.111*	0.148**	0.036
	(0.043)	(0.047)	(0.042)	(0.037)	(0.111)						
Research Institute	0.465	0.381	0.407	0.414	0.393	0.084	0.059	0.051	-0.025	-0.033	-0.008
	(0.054)	(0.050)	(0.052)	(0.050)	(0.758)						
Students and temporary workers	0.058	0.041	0.077	0.101	0.998	0.017	-0.019	-0.043	-0.036	-0.060	-0.024
	(0.025)	(0.020)	(0.028)	(0.030)	(0.394)						
Private sector	0.151	0.093	0.055	0.121	1.816	0.058	0.096**	0.030	0.038	-0.028	-0.066*
	(0.039)	(0.030)	(0.024)	(0.033)	(0.144)						
Agriculture	0.605	0.588	0.560	0.525	0.450	0.017	0.044	0.079	0.027	0.062	0.035
	(0.053)	(0.050)	(0.052)	(0.050)	(0.718)						
Economics, finance, & business	0.337	0.258	0.319	0.263	0.842	0.079	0.019	0.075	-0.061	-0.005	0.056
	(0.051)	(0.045)	(0.049)	(0.044)	(0.472)						
Human capital (education, health)	0.674	0.742	0.659	0.758	1.032	-0.068	0.015	-0.083	0.083	-0.015	-0.098
	(0.051)	(0.045)	(0.050)	(0.043)	(0.378)						
Environment	0.244	0.289	0.198	0.333	1.709	-0.044	0.046	-0.089	0.091	-0.045	-0.136**
	(0.047)	(0.046)	(0.042)	(0.048)	(0.165)						
Other field	0.081	0.186	0.121	0.131	1.482	-0.104**	-0.039	-0.050	0.065	0.054	-0.010
	(0.030)	(0.040)	(0.034)	(0.034)	(0.219)						
Policy influence 'high' or 'very high'	0.349	0.351	0.352	0.354	0.002	-0.002	-0.003	-0.005	-0.001	-0.003	-0.002
	(0.052)	(0.049)	(0.050)	(0.048)	(1.000)						
F-test of joint significance (F-stat)						1.409	1.110	1.394	0.790	0.743	1.374
Number of observations	86	97	91	99	373	183	177	185	188	196	190

Notes: * p<0.1, ** p<0.05, *** p<0.01.This table reports on the results of the balance tests for the key study indicators. Columns 1 to 4 show the means and standard errors for the 4 groups where 1 represents Human-Human, 2 represents Human-AI, 3 represents AI-AI and 4 AI-Human. Columns 6 to 11 report on the pairwise t-tests between the groups.

Table A.2: Summary Table

	Mean	Standard deviation	N
Outcome variables:			
Average quality (0-4)	2.874	0.655	366
Number of 'agree' and 'strongly agree' statements (0-13)	9.549	3.531	366
Overall quality of brief rated 'high' or 'very high'	0.601	0.49	366
Average engagement rating (0-4)	2.647	0.84	366
Total likelihood of engagement (0-6)	3.749	1.973	366
Minutes spent reading policy note	5.668	21.036	365
Average personal rating (0-4)	2.55	0.815	366
Total likelihood of engagement (0-6)	3.473	2.16	366
Others overall rating'	0.593	0.492	366
Preferences of policy blogs characteristics (1-5):			
Relevance to my work	4.5	0.7	366
Wide range of views	4.3	0.8	366
Level of detail provided	4.3	0.8	366
Title of the policy brief sparks my curiosity	4.5	0.7	366
Authors are reputable	3.8	1	366
Rationale presented is convincing	4.5	0.7	366
Visual presentation is attractive	4.2	0.8	366
Length of the policy brief is adequate	4.1	0.8	366
Content is easy to understand	4.6	0.6	366
Proposed recommendations are clear	4.7	0.6	366
Preference on number of pages	7.9	13.1	366
Preference on number of figures	6.6	7	366
Interaction with AI:			
Aware of AI tools	0.809	0.394	366
Used AI tools	0.612	0.488	366

Table A.3: Disaggregated Perceived Quality of Policy Blog

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Relevance to work	Range of views	Title sparked curiosity	Appropriate level of detail	Clear rationale	Visual presentation	Appropriate tone	Adequate length	Easy to understand	Clear proposed recommendations	Relevant proposed recommendations	Sufficiently detailed recommendations
(1) Told author is AI	-0.094	0.011	-0.088	-0.109	0.091	-0.125	-0.009	-0.047	-0.065	-0.085	-0.088	-0.173
	(0.144)	(0.140)	(0.134)	(0.144)	(0.133)	(0.172)	(0.108)	(0.145)	(0.114)	(0.140)	(0.146)	(0.167)
RI p-value	[.501]	[.932]	[.529]	[.414]	[.497]	[.484]	[.929]	[.739]	[.581]	[.517]	[.524]	[.281]
(2) Actual author is AI	-0.107	-0.380**	-0.226*	-0.285*	-0.221	-0.398**	-0.280**	-0.278*	-0.152	-0.024	0.080	-0.111
	(0.132)	(0.147)	(0.128)	(0.154)	(0.143)	(0.169)	(0.124)	(0.153)	(0.114)	(0.138)	(0.134)	(0.162)
RI p-value	[.406]	[.008]	[.106]	[.068]	[.126]	[.022]	[.023]	[.07]	[.188]	[.86]	[.54]	[.507]
(3) Told author is AI and Actual author is AI	0.023	0.193	0.403**	0.189	0.197	0.432*	0.166	0.211	0.136	0.150	0.051	0.309
	(0.193)	(0.206)	(0.189)	(0.211)	(0.191)	(0.242)	(0.161)	(0.210)	(0.160)	(0.196)	(0.195)	(0.234)
RI p-value	[.902]	[.345]	[.044]	[.378]	[.323]	[.077]	[.319]	[.315]	[.41]	[.407]	[.763]	[.172]
RI p-value [.209]	[.911]	[.08]	[.354]	[.057]	[.38]	[.296]	[.092]	[.289]	[.35]	[.424]	[.481]	[.817]
Mean of Human-Human	3.19	2.94	3.107	2.929	2.893	2.583	3.083	3.071	3.25	3.012	2.94	2.667
P-value (1) - (2)	.933	.008	.347	.255	.029	.12	.026	.121	.45	.652	.246	.711
N	366	366	366	366	366	366	366	366	366	366	366	366

Notes: * p<0.1, ** p<0.05, *** p<0.01. This table reports the results using disaggregated quality measures. Each outcome in columns 1 - 12 represents one of the quality statements asked in the survey using a likert scale ranging from 0 (very low) to 4 (very high). Strata fixed effects (countries) are controlled for and standard errors are clustered at the individual level (the unit of randomization).

Table A.4: Disaggregated Intended in Engagement with Policy Blog

	(1)	(2)	(3)	(4)	(5)	(6)
	Share the blog with others	Re-read the blog	Look up studies cited	Look up related studies	Contact the authors of the brief	Use the brief in decision-making
(1) Told author is AI	-0.120	0.059	0.144	0.027	-0.098	0.035
	(0.159)	(0.170)	(0.158)	(0.149)	(0.190)	(0.158)
RI p-value	[.443]	[.752]	[.341]	[.855]	[.62]	[.785]
(2) Actual author is AI	-0.304*	-0.200	-0.173	-0.069	-0.079	-0.079
	(0.166)	(0.169)	(0.156)	(0.147)	(0.189)	(0.161)
RI p-value	[.075]	[.255]	[.262]	[.646]	[.662]	[.637]
(3) Told author is AI and Actual author is AI	0.169	0.064	-0.077	0.098	0.031	-0.016
	(0.237)	(0.239)	(0.219)	(0.203)	(0.261)	(0.220)
RI p-value	[.507]	[.794]	[.726]	[.64]	[.898]	[.948]
RI p-value	[.503]	[.816]	[.718]	[.647]	[.9]	[.956]
Mean of Human-Human	3.929	3.833	3.81	3.798	3.274	3.738
P-value (1) - (2)	.275	.133	.045	.507	.918	.491
N	366	366	366	366	366	366

Notes: * p<0.1, ** p<0.05, *** p<0.01. This table reports respondents' likelihood of engagement with the policy blog. Each outcome in columns 1 - 6 represent respondents' belief that their colleagues will engage in a particular activity on a likert scale ranging from 0 (very unlikely) to 4 (very likely). Strata fixed effects (countries) are controlled for and standard errors are clustered at the individual level (the unit of randomization).

Table A.5: Disaggregated Beliefs in Others' Intended Engagement with Policy Blog

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Share the blog with others	Re-read the blog	Look up studies cited	Look up related studies	Contact the authors of the brief	Use the brief in decision-making	Rate the quality of the policy brief overall?
(1) Told author is AI	-0.173	-0.056	0.049	0.039	-0.050	-0.175	0.034
	(0.147)	(0.158)	(0.150)	(0.153)	(0.181)	(0.153)	(0.132)
RI p-value	[.22]	[.713]	[.735]	[.808]	[.769]	[.23]	[.792]
(2) Actual author is AI	-0.152	-0.072	-0.034	0.122	0.080	-0.241	-0.237*
	(0.134)	(0.152)	(0.151)	(0.153)	(0.179)	(0.148)	(0.134)
RI p-value	[.272]	[.643]	[.836]	[.445]	[.673]	[.102]	[.101]
(3) Told author is AI and Actual author is AI	0.028	0.016	-0.127	-0.107	-0.039	0.155	0.004
	(0.199)	(0.223)	(0.209)	(0.212)	(0.246)	(0.213)	(0.180)
RI p-value	[.879]	[.933]	[.538]	[.608]	[.873]	[.465]	[.984]
RI p-value	[.898]	[.952]	[.548]	[.63]	[.877]	[.52]	[.98]
Mean of Human-Human	3.893	3.667	3.595	3.548	3.226	3.833	3.798
P-value (1) - (2)	.878	.915	.584	.599	.462	.679	.047
N	366	366	366	366	366	366	366

Notes: * p<0.1, ** p<0.05, *** p<0.01. This table reports respondents' belief in others' likelihood of engagement with the policy blog. Each outcome in columns 1 - 6 represent respondents' belief that their colleagues will engage in a particular activity on a likert scale ranging from 0 (very unlikely) to 4 (very likely). Strata fixed effects (countries) are controlled for and standard errors are clustered at the individual level (the unit of randomization).

Table A.6: Perceived Quality by Reading Grade Level (Flesch-Kincaid)

	(1)	(2)	(3)	(4)	(5)
	Unweighted Average Quality (0-4)	Weighted Quality Average	Standardized First Principal Component	Number of 'agree' and 'strongly agree' statements (0-12)	Overall quality of brief rated 'high' or 'very high'
(1) AI reported	-1.346	-1.281	-1.624	-7.847	-0.661
	(1.502)	(1.428)	(2.245)	(7.073)	(0.981)
RI p-value	[.42]	[.41]	[.509]	[.312]	[.523]
(2) AI Generated	-0.519	-0.829	-0.341	-6.064	-0.342
	(1.035)	(1.045)	(1.582)	(5.123)	(0.737)
RI p-value	[.513]	[.306]	[.774]	[.129]	[.524]
(3) AI reported × AI Generated	2.397	2.502	3.008	11.909	1.073
	(1.623)	(1.559)	(2.430)	(7.772)	(1.096)
RI p-value (Permuting AI reported)	[.183]	[.149]	[.253]	[.158]	[.333]
RI p-value (Permuting AI Generated)	[.017]	[.009]	[.054]	[.015]	[.178]
(4) Reading Grade Level (Flesch-Kincaid)	0.050	0.039	0.092	0.025	0.013
	(0.055)	(0.055)	(0.084)	(0.271)	(0.038)
RI p-value	[.109]	[.204]	[.049]	[.899]	[.559]
(5) AI reported × Reading Grade Level (Flesch-Kincaid)	0.075	0.072	0.089	0.439	0.034
	(0.088)	(0.084)	(0.132)	(0.415)	(0.058)
RI p-value (Permuting AI reported)	[.444]	[.416]	[.522]	[.33]	[.583]
RI p-value (Permuting Reading Grade Level (Flesch-Kincaid))	[.109]	[.111]	[.208]	[.064]	[.318]
(6) AI Generated × Reading Grade Level (Flesch-Kincaid)	0.030	0.047	0.019	0.338	0.013
	(0.063)	(0.064)	(0.097)	(0.310)	(0.045)
RI p-value (Permuting AI Generated)	[.548]	[.363]	[.806]	[.176]	[.7]
RI p-value (Permuting Reading Grade Level (Flesch-Kincaid))	[.574]	[.37]	[.821]	[.181]	[.695]
(7) AI reported × AI Generated × Reading Grade Level (Flesch-Kincaid)	-0.140	-0.147	-0.175	-0.669	-0.056
	(0.097)	(0.094)	(0.146)	(0.467)	(0.067)
RI p-value (Permuting AI reported)	[.19]	[.153]	[.265]	[.192]	[.411]
RI p-value (Permuting AI Generated)	[.026]	[.014]	[.071]	[.035]	[.262]
RI p-value (Permuting Reading Grade Level (Flesch-Kincaid))	[.035]	[.026]	[.092]	[.053]	[.246]
Control Mean	2.972	3.008	.149	9.857	.702
P-Value (1) - (2)	.523	.706	.502	.762	.696
P-Value (5) - (6)	.556	.718	.534	.771	.678
N	366	366	366	366	366

Notes: * p<0.1, ** p<0.05, *** p<0.01. This table reports on the heterogeneity in perceived quality of the policy brief by the Flesch-Kincaid Reading Grade Level. The outcome in the first column is the unweighted average rating across twelve statements. Ratings for each statement follow a likert scale ranging from 0 to 4, with 0 indicating strongly disagree and 4 indicating strongly agree. The second column presents the weighted average across nine statements. The outcome in the third column presents the score of the first principal component, calculated based on the same twelve statements using polychoric PCA. The fourth column indicates the total number of statements rated as either 'agree' or 'strongly agree', out of 12 perception statements. The fifth column presents the results from a separate survey question asking respondents to rate the overall quality of the brief from 1 to 5. The outcome variable in the fifth column is an indicator equal to 1 if the quality of the brief is rated either 'high' or 'very high'. Strata fixed effects (countries) are controlled for and standard errors are clustered at the individual level (the unit of randomization).

Table A.7: Perceived Quality by Reading Grade Level (Flesch-Kincaid)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Relevance to work	Range of views	Title sparked curiosity	Appropriate level of detail	Clear rationale	Visual presentation	Appropriate tone	Adequate length	Easy to understand	Clear proposed recommendations	Relevant proposed recommendations	Sufficiently detailed recommendations
(1) AI reported	-0.308	-1.554	-1.482	0.414	-3.265	-2.499	-0.581	-1.398	-1.515	-1.630	-2.470	0.139
	(1.924)	(1.856)	(2.263)	(2.222)	(2.074)	(2.548)	(1.796)	(2.444)	(1.864)	(1.795)	(1.911)	(2.312)
RI p-value	[.863]	[.439]	[.532]	[.865]	[.153]	[.387]	[.778]	[.603]	[.474]	[.418]	[.242]	[.951]
(2) AI Generated	-0.924	-2.118	0.306	0.124	-2.058	-2.363	-0.051	-1.755	0.115	1.164	0.011	1.316
	(1.561)	(1.644)	(1.658)	(1.641)	(1.552)	(2.047)	(1.248)	(1.766)	(1.163)	(1.178)	(1.261)	(1.693)
RI p-value	[.398]	[.069]	[.787]	[.909]	[.084]	[.089]	[.959]	[.176]	[.886]	[.273]	[.989]	[.287]
(3) AI reported × AI Generated	0.868	3.007	2.852	1.386	4.487**	4.697*	1.418	3.139	2.175	1.631	3.308	-0.209
	(2.109)	(2.100)	(2.457)	(2.428)	(2.241)	(2.782)	(1.940)	(2.655)	(2.009)	(2.000)	(2.080)	(2.544)
RI p-value (Permuting AI reported)	[.707]	[.162]	[.287]	[.586]	[.05]	[.113]	[.52]	[.262]	[.319]	[.444]	[.126]	[.931]
RI p-value (Permuting AI Generated)	[.551]	[.057]	[.064]	[.385]	[0]	[.016]	[.239]	[.058]	[.054]	[.283]	[.017]	[.919]
(4) Reading Grade Level (Flesch-Kincaid)	0.010	-0.025	0.078	0.086	-0.043	0.092	0.077	-0.018	0.068	0.116*	0.046	0.109
	(0.084)	(0.087)	(0.087)	(0.085)	(0.083)	(0.111)	(0.067)	(0.094)	(0.060)	(0.060)	(0.067)	(0.089)
RI p-value	[.821]	[.601]	[.077]	[.086]	[.36]	[.144]	[.035]	[.735]	[.063]	[.016]	[.295]	[.037]
(5) AI reported × Reading Grade Level (Flesch-Kincaid)	0.016	0.092	0.082	-0.030	0.197	0.135	0.031	0.077	0.085	0.090	0.139	-0.018
	(0.113)	(0.110)	(0.132)	(0.129)	(0.122)	(0.150)	(0.105)	(0.144)	(0.109)	(0.105)	(0.113)	(0.136)
RI p-value (Permuting AI reported)	[.887]	[.439]	[.546]	[.826]	[.131]	[.428]	[.796]	[.627]	[.489]	[.441]	[.261]	[.898]
RI p-value (Permuting Reading Grade Level (Flesch-Kincaid))	[.802]	[.148]	[.232]	[.641]	[.005]	[.121]	[.534]	[.269]	[.131]	[.192]	[.052]	[.802]
(6) AI Generated × Reading Grade Level (Flesch-Kincaid)	0.063	0.113	-0.023	-0.018	0.119	0.144	-0.003	0.097	-0.006	-0.063	0.012	-0.080
	(0.095)	(0.100)	(0.100)	(0.099)	(0.094)	(0.122)	(0.076)	(0.106)	(0.071)	(0.072)	(0.077)	(0.103)
RI p-value (Permuting AI Generated)	[.342]	[.128]	[.744]	[.793]	[.106]	[.096]	[.96]	[.221]	[.906]	[.332]	[.872]	[.3]
RI p-value (Permuting Reading Grade Level (Flesch-Kincaid))	[.319]	[.112]	[.773]	[.816]	[.075]	[.121]	[.953]	[.186]	[.903]	[.335]	[.85]	[.286]
(7) AI reported × AI Generated × Reading Grade Level (Flesch-Kincaid)	-0.064	-0.177	-0.159	-0.086	-0.263*	-0.265	-0.083	-0.187	-0.128	-0.088	-0.202	0.028
	(0.127)	(0.128)	(0.147)	(0.145)	(0.134)	(0.167)	(0.116)	(0.160)	(0.120)	(0.120)	(0.125)	(0.154)
RI p-value (Permuting AI reported)	[.651]	[.18]	[.32]	[.566]	[.055]	[.137]	[.514]	[.261]	[.326]	[.491]	[.118]	[.862]
RI p-value (Permuting AI Generated)	[.49]	[.082]	[.103]	[.387]	[.001]	[.028]	[.274]	[.079]	[.072]	[.347]	[.021]	[.819]
RI p-value (Permuting Reading Grade Level (Flesch-Kincaid))	[.502]	[.072]	[.114]	[.406]	[.002]	[.026]	[.303]	[.066]	[.099]	[.363]	[.023]	[.783]
Control Mean	3.19	2.94	3.107	2.929	2.893	2.583	3.083	3.071	3.25	3.012	2.94	2.667
P-Value (1) - (2)	.67	.684	.336	.875	.484	.943	.73	.86	.332	.085	.14	.546
P-Value (5) - (6)	.583	.806	.344	.914	.442	.933	.71	.873	.355	.111	.201	.599
N	366	366	366	366	366	366	366	366	366	366	366	366

Notes: * p<0.1, ** p<0.05, *** p<0.01. This table reports the heterogeneity in results using disaggregated quality measures by the Flesch-Kincaid Reading Grade Level. Each outcome in columns 1 - 12 represents one of the quality statements asked in the survey using a likert scale ranging from 0 (very low) to 4 (very high). Strata fixed effects (countries) are controlled for and standard errors are clustered at the individual level (the unit of randomization).

Table A.8: Perceived Quality by Previously used AI

	(1)	(2)	(3)	(4)	(5)
	Unweighted Average Quality (0-4)	Weighted Quality Average	Standardized First Principal Component	Number of 'agree' and 'strongly agree' statements (0-12)	Overall quality of brief rated high or very high
(1) AI reported	-0.424**	-0.384*	-0.670**	-1.283	-0.192
	(0.198)	(0.204)	(0.297)	(1.213)	(0.173)
RI p-value	[.044]	[.069]	[.039]	[.308]	[.231]
(2) AI Generated	-0.568**	-0.551**	-0.829**	-3.193**	-0.263*
	(0.226)	(0.234)	(0.342)	(1.278)	(0.142)
RI p-value	[.014]	[.024]	[.02]	[.012]	[.055]
(3) AI reported × AI Generated	0.810**	0.764**	1.249**	3.014	0.407*
	(0.334)	(0.346)	(0.506)	(1.845)	(0.243)
RI p-value (Permuting AI reported)	[.026]	[.039]	[.023]	[.117]	[.081]
RI p-value (Permuting AI Generated)	[.02]	[.029]	[.019]	[.108]	[.089]
(4) Previously used AI	-0.155	-0.137	-0.246	-0.921	0.006
	(0.168)	(0.172)	(0.256)	(1.010)	(0.124)
RI p-value	[.578]	[.643]	[.557]	[.527]	[.965]
(5) AI reported × Previously used AI	0.415*	0.385	0.687*	0.855	0.069
	(0.240)	(0.243)	(0.363)	(1.445)	(0.199)
RI p-value (Permuting AI reported)	[.084]	[.111]	[.063]	[.564]	[.731]
RI p-value (Permuting Previously used AI)	[.28]	[.338]	[.229]	[.635]	[.766]
(6) AI Generated × Previously used AI	0.345	0.285	0.526	2.008	0.087
	(0.270)	(0.275)	(0.407)	(1.541)	(0.183)
RI p-value (Permuting AI Generated)	[.193]	[.266]	[.177]	[.184]	[.638]
RI p-value (Permuting Previously used AI)	[.245]	[.343]	[.258]	[.204]	[.672]
(7) AI reported × AI Generated × Previously used AI	-0.674*	-0.602	-1.120*	-1.813	-0.199
	(0.382)	(0.391)	(0.581)	(2.117)	(0.279)
RI p-value (Permuting AI reported)	[.086]	[.142]	[.06]	[.412]	[.455]
RI p-value (Permuting AI Generated)	[.073]	[.125]	[.045]	[.375]	[.462]
RI p-value (Permuting Previously used AI)	[.088]	[.125]	[.062]	[.332]	[.491]
Control Mean	2.972	3.008	.149	9.857	.702
P-Value (1) - (2)	.53	.474	.641	.142	.699
P-Value (5) - (6)	.798	.719	.698	.456	.936
N	296	296	296	296	296

Notes: * p<0.1, ** p<0.05, *** p<0.01. This table reports on the heterogeneity in perceived quality of the policy brief by previous use of AI. The outcome in the first column is the unweighted average rating across twelve statements. Ratings for each statement follow a likert scale ranging from 0 to 4, with 0 indicating strongly disagree and 4 indicating strongly agree. The second column presents the weighted average across nine statements. The outcome in the third column presents the score of the first principal component, calculated based on the same twelve statements using polychoric PCA. The fourth column indicates the total number of statements rated as either 'agree' or 'strongly agree', out of 12 perception statements. The fifth column presents the results from a separate survey question asking respondents to rate the overall quality of the brief from 1 to 5. The outcome variable in the fifth column is an indicator equal to 1 if the quality of the brief is rated either 'high' or 'very high'. Strata fixed effects (countries) are controlled for and standard errors are clustered at the individual level (the unit of randomization).

Table A.9: Disaggregated Perceived Quality by Previously used AI

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Relevance to work	Range of views	Title sparked curiosity	Appropriate level of detail	Clear rationale	Visual presentation	Appropriate tone	Adequate length	Easy to understand	Clear proposed recommendations	Relevant proposed recommendations	Sufficiently detailed recommendations
(1) AI reported	-0.516	-0.367	-0.627*	0.064	-0.307	-0.543	-0.319	-0.473*	-0.249	-0.547	-0.897***	-0.308
	(0.401)	(0.244)	(0.342)	(0.307)	(0.368)	(0.401)	(0.244)	(0.279)	(0.241)	(0.340)	(0.287)	(0.348)
RI p-value	[.202]	[.14]	[.059]	[.841]	[.408]	[.188]	[.162]	[.106]	[.327]	[.11]	[.006]	[.36]
(2) AI Generated	-0.130	0.942***	-0.219	-0.511	0.724**	-0.603	0.666**	-0.465	-0.558*	-0.610*	-0.576*	0.809**
	(0.297)	(0.302)	(0.317)	(0.354)	(0.345)	(0.395)	(0.311)	(0.343)	(0.296)	(0.313)	(0.306)	(0.333)
RI p-value	[.71]	[.005]	[.498]	[.159]	[.044]	[.136]	[.039]	[.171]	[.056]	[.054]	[.068]	[.018]
(3) AI reported × AI Generated	0.408	1.054**	0.702	0.241	0.642	1.155**	0.859**	0.634	0.674*	1.095**	1.234***	1.024*
	(0.512)	(0.443)	(0.491)	(0.498)	(0.514)	(0.534)	(0.396)	(0.528)	(0.391)	(0.512)	(0.437)	(0.557)
RI p-value (Permuting AI reported)	[.424]	[.027]	[.153]	[.649]	[.215]	[.039]	[.042]	[.241]	[.111]	[.049]	[.007]	[.066]
RI p-value (Permuting AI Generated)	[.436]	[.031]	[.153]	[.64]	[.212]	[.036]	[.03]	[.235]	[.089]	[.029]	[.005]	[.082]
(4) Previously used AI	-0.041	-0.376	-0.155	0.221	-0.357	0.012	-0.283	-0.097	-0.204	-0.108	-0.417*	-0.052
	(0.245)	(0.236)	(0.225)	(0.277)	(0.293)	(0.357)	(0.202)	(0.256)	(0.212)	(0.268)	(0.244)	(0.302)
RI p-value	[.884]	[.243]	[.538]	[.437]	[.22]	[.972]	[.153]	[.767]	[.624]	[.746]	[.198]	[.903]
(5) AI reported × Previously used AI	0.497	0.497	0.753*	-0.260	0.430	0.448	0.364	0.531	0.187	0.437	0.914***	0.188
	(0.443)	(0.313)	(0.388)	(0.369)	(0.410)	(0.471)	(0.290)	(0.350)	(0.301)	(0.406)	(0.353)	(0.428)
RI p-value (Permuting AI reported)	[.279]	[.104]	[.05]	[.495]	[.285]	[.334]	[.169]	[.104]	[.533]	[.247]	[.022]	[.619]
RI p-value (Permuting Previously used AI)	[.385]	[.328]	[.095]	[.577]	[.352]	[.385]	[.264]	[.327]	[.653]	[.264]	[.093]	[.724]
(6) AI Generated × Previously used AI	-0.088	0.627*	-0.025	0.050	0.515	0.037	0.367	0.041	0.527	0.668*	0.688*	0.727*
	(0.358)	(0.380)	(0.369)	(0.416)	(0.410)	(0.478)	(0.375)	(0.411)	(0.340)	(0.373)	(0.372)	(0.423)
RI p-value (Permuting AI Generated)	[.812]	[.108]	[.956]	[.915]	[.199]	[.93]	[.288]	[.923]	[.121]	[.064]	[.07]	[.071]
RI p-value (Permuting Previously used AI)	[.797]	[.141]	[.928]	[.904]	[.206]	[.931]	[.286]	[.903]	[.305]	[.151]	[.054]	[.188]
(7) AI reported × AI Generated × Previously used AI	-0.329	-1.089**	-0.372	0.075	-0.483	-0.679	-0.809*	-0.457	-0.630	-1.069*	-1.286**	-0.954
	(0.557)	(0.517)	(0.558)	(0.563)	(0.570)	(0.644)	(0.468)	(0.606)	(0.455)	(0.583)	(0.507)	(0.650)
RI p-value (Permuting AI reported)	[.555]	[.058]	[.509]	[.914]	[.4]	[.296]	[.078]	[.442]	[.172]	[.084]	[.024]	[.148]
RI p-value (Permuting AI Generated)	[.581]	[.044]	[.473]	[.896]	[.425]	[.288]	[.065]	[.435]	[.14]	[.056]	[.015]	[.144]
RI p-value (Permuting Previously used AI)	[.597]	[.056]	[.462]	[.892]	[.372]	[.278]	[.093]	[.43]	[.203]	[.045]	[.027]	[.175]
Control Mean	3.19	2.94	3.107	2.929	2.893	2.583	3.083	3.071	3.25	3.012	2.94	2.667
P-Value (1) - (2)	.323	.054	.281	.082	.24	.877	.301	.982	.299	.856	.298	.159
P-Value (5) - (6)	.194	.724	.072	.441	.841	.382	.992	.23	.324	.575	.553	.211
N	296	296	296	296	296	296	296	296	296	296	296	296

Notes: * p<0.1, ** p<0.05, *** p<0.01. This table reports the heterogeneity in results using disaggregated quality measures by respondents' previous use of AI. Each outcome in columns 1 - 12 represents one of the quality statements asked in the survey using a likert scale ranging from 0 (very low) to 4 (very high). Strata fixed effects (countries) are controlled for and standard errors are clustered at the individual level (the unit of randomization).

Table A.10: Perceived Quality by Trust in AI blog writing

	(1)	(2)	(3)	(4)	(5)
	Unweighted Average Quality (0-4)	Weighted Quality Average	Standardized First Principal Component	Number of 'agree' and 'strongly agree' statements (0-12)	Overall quality of brief rated 'high' or 'very high'
(1) AI reported	0.002	0.048	0.018	-0.101	-0.061
	(0.158)	(0.155)	(0.242)	(0.875)	(0.115)
RI p-value	[.993]	[.736]	[.93]	[.896]	[.578]
(2) AI Generated	-0.200	-0.225	-0.267	-1.291	-0.200*
	(0.173)	(0.174)	(0.259)	(0.946)	(0.109)
RI p-value	[.242]	[.191]	[.282]	[.186]	[.068]
(3) AI reported × AI Generated	0.122	0.129	0.106	1.316	0.144
	(0.238)	(0.235)	(0.360)	(1.301)	(0.170)
RI p-value (Permuting AI reported)	[.61]	[.581]	[.771]	[.305]	[.373]
RI p-value (Permuting AI Generated)	[.611]	[.59]	[.755]	[.328]	[.399]
(4) Trust in AI blog writing	0.238*	0.249**	0.347*	1.563**	0.124
	(0.126)	(0.125)	(0.192)	(0.727)	(0.103)
RI p-value	[.056]	[.052]	[.076]	[.031]	[.234]
(5) AI reported × Trust in AI blog writing	-0.145	-0.185	-0.220	-0.898	-0.091
	(0.198)	(0.194)	(0.306)	(1.054)	(0.151)
RI p-value (Permuting AI reported)	[.474]	[.345]	[.483]	[.42]	[.534]
RI p-value (Permuting Trust in AI blog writing)	[.412]	[.302]	[.423]	[.379]	[.55]
(6) AI Generated × Trust in AI blog writing	-0.018	-0.032	-0.036	-0.050	0.051
	(0.208)	(0.206)	(0.313)	(1.131)	(0.150)
RI p-value (Permuting AI Generated)	high [.928]	[.882]	[.911]	[.969]	[.732]
RI p-value (Permuting Trust in AI blog writing)	[.931]	[.864]	[.903]	[.964]	[.724]
(7) AI reported × AI Generated × Trust in AI blog writing	0.155	0.133	0.283	0.229	0.116
	(0.291)	(0.289)	(0.443)	(1.568)	(0.219)
RI p-value (Permuting AI reported)	[.597]	[.652]	[.529]	[.895]	[.58]
RI p-value (Permuting AI Generated)	[.589]	[.651]	[.51]	[.867]	[.615]
RI p-value (Permuting Trust in AI blog writing)	[.539]	[.582]	[.462]	[.871]	[.58]
Control Mean	2.972	3.008	.149	9.857	.702
P-Value (1) - (2)	.28	.133	.313	.205	.242
P-Value (5) - (6)	.576	.494	.6	.459	.368
N	366	366	366	366	366

Notes: * p<0.1, ** p<0.05, *** p<0.01. This table reports on the heterogeneity in perceived quality of the policy brief by trust in the use of AI to write policy briefs. The outcome in the first column is the unweighted average rating across twelve statements. Ratings for each statement follow a likert scale ranging from 0 to 4, with 0 indicating strongly disagree and 4 indicating strongly agree. The second column presents the weighted average across nine statements. The outcome in the third column presents the score of the first principal component, calculated based on the same twelve statements using polychoric PCA. The fourth column indicates the total number of statements rated as either 'agree' or 'strongly agree', out of 12 perception statements. The fifth column presents the results from a separate survey question asking respondents to rate the overall quality of the brief from 1 to 5. The outcome variable in the fifth column is an indicator equal to 1 if the quality of the brief is rated either 'high' or 'very high'. Strata fixed effects (countries) are controlled for and standard errors are clustered at the individual level (the unit of randomization).

Table A.11: Disaggregated Perceived Quality by Trust in AI blog writing

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Relevance to work	Range of views	Title sparked curiosity	Appropriate level of detail	Clear rationale	Visual presentation	Appropriate tone	Adequate length	Easy to understand	Clear proposed recommendations	Relevant proposed recommendations	Sufficiently detailed recommendations
(1) AI reported	0.147	-0.031	-0.020	0.211	-0.039	-0.195	-0.046	0.082	0.122	-0.048	-0.196	0.031
	(0.242)	(0.237)	(0.249)	(0.241)	(0.239)	(0.249)	(0.157)	(0.240)	(0.185)	(0.234)	(0.245)	(0.278)
RI p-value	[.509]	[.904]	[.944]	[.394]	[.883]	[.436]	[.767]	[.72]	[.51]	[.823]	[.431]	[.901]
(2) AI Generated	-0.121	-0.458**	-0.129	-0.354	-0.410*	0.013	-0.335*	-0.364	0.057	-0.203	-0.117	0.028
	(0.238)	(0.232)	(0.226)	(0.264)	(0.240)	(0.262)	(0.200)	(0.254)	(0.199)	(0.240)	(0.238)	(0.271)
RI p-value	[.639]	[.043]	[.631]	[.176]	[.072]	[.956]	[.081]	[.146]	[.762]	[.376]	[.631]	[.908]
(3) AI reported × AI Generated	-0.322	0.214	0.327	0.079	0.405	0.301	0.169	0.335	-0.077	0.136	0.040	-0.149
	(0.325)	(0.343)	(0.354)	(0.356)	(0.340)	(0.403)	(0.257)	(0.349)	(0.268)	(0.345)	(0.347)	(0.408)
RI p-value (Permuting AI reported)	[.322]	[.53]	[.377]	[.846]	[.245]	[.445]	[.485]	[.341]	[.785]	[.681]	[.906]	[.705]
RI p-value (Permuting AI Generated)	[.944]	[.565]	[.421]	[.828]	[.237]	[.497]	[.495]	[.364]	[.762]	[.704]	[.909]	[.731]
(4) Trust in AI blog writing	0.066	0.098	0.177	0.552***	0.020	0.636***	0.013	0.166	0.327**	0.181	0.099	0.523**
	(0.189)	(0.203)	(0.172)	(0.209)	(0.205)	(0.238)	(0.159)	(0.215)	(0.162)	(0.204)	(0.198)	(0.234)
RI p-value	[.696]	[.635]	[.265]	[.008]	[.909]	[.012]	[.929]	[.426]	[.053]	[.367]	[.612]	[.026]
(5) AI reported × Trust in AI blog writing	-0.387	0.043	-0.132	-0.584**	0.191	0.006	0.049	-0.221	-0.336	-0.097	0.133	-0.410
	(0.309)	(0.312)	(0.307)	(0.295)	(0.292)	(0.347)	(0.222)	(0.305)	(0.239)	(0.295)	(0.315)	(0.353)
RI p-value (Permuting AI reported)	[.219]	[.872]	[.659]	[.064]	[.524]	[.985]	[.810]	[.505]	[.152]	[.739]	[.668]	[.256]
RI p-value (Permuting Trust in AI blog writing)	[.143]	[.886]	[.596]	[.057]	[.465]	[.989]	[.821]	[.409]	[.183]	[.759]	[.683]	[.219]
(6) AI Generated × Trust in AI blog writing	-0.209	0.124	-0.178	0.071	0.318	-0.743**	0.092	0.131	-0.379	0.286	0.324	-0.278
	(0.291)	(0.318)	(0.281)	(0.316)	(0.313)	(0.354)	(0.269)	(0.327)	(0.245)	(0.292)	(0.300)	(0.352)
RI p-value (Permuting AI Generated)	[.945]	[.696]	[.549]	[.83]	[.3]	[.032]	[.715]	[.681]	[.115]	[.305]	[.281]	[.393]
RI p-value (Permuting Trust in AI blog writing)	[.967]	[.711]	[.486]	[.826]	[.285]	[.048]	[.723]	[.702]	[.113]	[.31]	[.243]	[.468]
(7) AI reported × AI Generated × Trust in AI blog writing	0.541	-0.024	0.143	0.207	-0.333	0.311	-0.004	-0.195	0.375	0.029	0.024	0.784
	(0.414)	(0.444)	(0.442)	(0.438)	(0.424)	(0.517)	(0.343)	(0.443)	(0.343)	(0.429)	(0.439)	(0.510)
RI p-value (Permuting AI reported)	[.182]	[.953]	[.722]	[.626]	[.46]	[.529]	[.991]	[.661]	[.258]	[.951]	[.949]	[.117]
RI p-value (Permuting AI Generated)	[.219]	[.966]	[.764]	[.669]	[.43]	[.563]	[.992]	[.666]	[.26]	[.931]	[.955]	[.135]
RI p-value (Permuting Trust in AI blog writing)	[.155]	[.953]	[.719]	[.598]	[.38]	[.523]	[.991]	[.646]	[.265]	[.935]	[.947]	[.095]
Control Mean	3.19	2.94	3.107	2.929	2.893	2.583	3.083	3.071	3.25	3.012	2.94	2.667
P-Value (1) - (2)	.291	.106	.687	.032	.142	.447	.179	.098	.755	.531	.763	.992
P-Value (5) - (6)	.212	.809	.888	.046	.687	.045	.875	.305	.864	.198	.55	.717
N	366	366	366	366	366	366	366	366	366	366	366	366

Notes: * p<0.1, ** p<0.05, *** p<0.01. This table reports the heterogeneity in results using disaggregated quality measures by respondents' trust in AI to write policy blogs. Each outcome in columns 1 - 12 represents one of the quality statements asked in the survey using a likert scale ranging from 0 (very low) to 4 (very high). Strata fixed effects (countries) are controlled for and standard errors are clustered at the individual level (the unit of randomization).

Table A.12: Perceived Quality by Gender

	(1)	(2)	(3)	(4)	(5)
	Unweighted Average Quality (0-4)	Weighted Quality Average	Standardized First Principal Component	Number of 'agree' and 'strongly agree' statements (0-12)	Overall quality of brief rated high or very high
(1) AI reported	0.003	0.014	0.016	-0.351	-0.101
	(0.113)	(0.111)	(0.174)	(0.611)	(0.090)
RI p-value	[.98]	[.905]	[.925]	[.560]	[.243]
(2) AI Generated	-0.244**	-0.270**	-0.353**	-1.429**	-0.187**
	(0.118)	(0.117)	(0.176)	(0.645)	(0.085)
RI p-value	[.035]	[.017]	[.037]	[.024]	[.022]
(3) AI reported × AI Generated	0.109	0.104	0.137	0.988	0.162
	(0.161)	(0.161)	(0.244)	(0.880)	(0.126)
RI p-value (Permuting AI reported)	[.495]	[.529]	[.573]	[.26]	[.191]
RI p-value (Permuting AI Generated)	[.484]	[.514]	[.569]	[.266]	[.18]
(4) Gender	-0.483**	-0.489**	-0.674*	-2.615*	-0.105
	(0.245)	(0.237)	(0.377)	(1.535)	(0.207)
RI p-value	[1]	[1]	[1]	[1]	[1]
(5) AI reported × Gender	-0.325*	-0.287	-0.485*	-0.859	-0.035
	(0.189)	(0.186)	(0.287)	(1.133)	(0.158)
RI p-value (Permuting AI reported)	[.087]	[.121]	[.093]	[.444]	[.833]
RI p-value (Permuting Gender)	[1]	[1]	[1]	[1]	[1]
(6) AI Generated × Gender	0.169	0.140	0.300	0.658	0.079
	(0.201)	(0.202)	(0.310)	(1.083)	(0.157)
RI p-value (Permuting AI Generated)	[.449]	[.558]	[.382]	[.557]	[.624]
RI p-value (Permuting Gender)	[1]	[1]	[1]	[1]	[1]
(7) AI reported × AI Generated × Gender	0.444	0.439	0.602	1.786	0.230
	(0.278)	(0.275)	(0.428)	(1.515)	(0.221)
RI p-value (Permuting AI reported)	[.133]	[.151]	[.188]	[.268]	[.363]
RI p-value (Permuting AI Generated)	[.21]	[.222]	[.259]	[.331]	[.357]
RI p-value (Permuting Gender)	[1]	[1]	[1]	[1]	[1]
Control Mean	2.972	3.008	.149	9.857	.702
P-Value (1) - (2)	.037	.016	.04	.092	.335
P-Value (5) - (6)	.023	.044	.019	.166	.477
N	366	366	366	366	366

Notes: * p<0.1, ** p<0.05, *** p<0.01. This table reports on the heterogeneity in perceived quality of the policy brief by gender. The outcome in the first column is the unweighted average rating across twelve statements. Ratings for each statement follow a likert scale ranging from 0 to 4, with 0 indicating strongly disagree and 4 indicating strongly agree. The second column presents the weighted average across nine statements. The outcome in the third column presents the score of the first principal component, calculated based on the same twelve statements using polychoric PCA. The fourth column indicates the total number of statements rated as either 'agree' or 'strongly agree', out of 12 perception statements. The fifth column presents the results from a separate survey question asking respondents to rate the overall quality of the brief from 1 to 5. The outcome variable in the fifth column is an indicator equal to 1 if the quality of the brief is rated either 'high' or 'very high'. Strata fixed effects (countries) are controlled for and standard errors are clustered at the individual level (the unit of randomization).

Table A.13: Disaggregated Perceived Quality by Gender

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Relevance to work	Range of views	Title sparked curiosity	Appropriate level of detail	Clear rationale	Visual presentation	Appropriate tone	Adequate length	Easy to understand	Clear proposed recommendations	Relevant proposed recommendations	Sufficiently detailed recommendations
(1) AI reported	0.009	0.071	-0.104	-0.081	0.156	-0.146	0.022	0.060	0.073	0.062	-0.011	-0.082
	(0.157)	(0.162)	(0.154)	(0.178)	(0.152)	(0.212)	(0.127)	(0.179)	(0.131)	(0.161)	(0.168)	(0.201)
RI p-value	[.961]	[.636]	[.523]	[.636]	[.306]	[.5]	[.85]	[.738]	[.597]	[.673]	[.958]	[.672]
(2) AI Generated	-0.118	-0.359**	-0.334**	-0.307*	-0.283*	-0.466**	-0.303**	-0.298	-0.158	-0.082	-0.049	-0.171
	(0.152)	(0.177)	(0.156)	(0.184)	(0.164)	(0.207)	(0.148)	(0.195)	(0.138)	(0.165)	(0.162)	(0.195)
RI p-value	[.419]	[.04]	[.025]	[.101]	[.087]	[.022]	[.036]	[.118]	[.238]	[.613]	[.747]	[.366]
(3) AI reported × AI Generated	-0.013	0.048	0.306	0.114	0.086	0.274	0.135	0.140	0.002	0.040	-0.006	0.189
	(0.219)	(0.239)	(0.224)	(0.251)	(0.219)	(0.291)	(0.191)	(0.259)	(0.192)	(0.230)	(0.226)	(0.272)
RI p-value (Permuting AI reported)	[.956]	[.836]	[.206]	[.676]	[.693]	[.355]	[.48]	[.57]	[.991]	[.87]	[.979]	[.488]
RI p-value (Permuting AI Generated)	[.944]	[.859]	[.154]	[.659]	[.711]	[.325]	[.473]	[.585]	[.991]	[.9]	[.981]	[.495]
(4) Gender	-0.256	-0.372	-0.803**	-0.593	-1.013**	-1.187***	-0.319	0.235	0.158	-0.497	-0.576	-0.574
	(0.453)	(0.401)	(0.338)	(0.459)	(0.444)	(0.420)	(0.301)	(0.354)	(0.391)	(0.469)	(0.464)	(0.515)
RI p-value	[1]	[1]	[1]	[1]	[1]	[1]	[1]	[1]	[1]	[1]	[1]	[1]
(5) AI reported × Gender	-0.436	-0.266	-0.019	-0.146	-0.326	0.011	-0.148	-0.465	-0.599**	-0.664**	-0.402	-0.436
	(0.366)	(0.323)	(0.307)	(0.291)	(0.311)	(0.340)	(0.236)	(0.298)	(0.273)	(0.325)	(0.329)	(0.360)
RI p-value (Permuting AI reported)	[.265]	[.435]	[.943]	[.635]	[.253]	[.976]	[.562]	[.126]	[.036]	[.046]	[.227]	[.193]
RI p-value (Permuting Gender)	[1]	[1]	[1]	[1]	[1]	[1]	[1]	[1]	[1]	[1]	[1]	[1]
(6) AI Generated × Gender	0.057	-0.080	0.378	0.079	0.224	0.228	0.086	0.083	0.034	0.229	0.479*	0.225
	(0.305)	(0.322)	(0.262)	(0.342)	(0.341)	(0.352)	(0.278)	(0.297)	(0.253)	(0.304)	(0.281)	(0.367)
RI p-value (Permuting AI Generated)	[.866]	[.801]	[.191]	[.86]	[.531]	[.563]	[.773]	[.797]	[.905]	[.469]	[.107]	[.533]
RI p-value (Permuting Gender)	[1]	[1]	[1]	[1]	[1]	[1]	[1]	[1]	[1]	[1]	[1]	[1]
(7) AI reported × AI Generated × Gender	0.167	0.606	0.497	0.339	0.523	0.725	0.151	0.321	0.581*	0.522	0.340	0.558
	(0.462)	(0.469)	(0.406)	(0.458)	(0.440)	(0.488)	(0.357)	(0.431)	(0.344)	(0.430)	(0.428)	(0.532)
RI p-value (Permuting AI reported)	[.727]	[.233]	[.322]	[.497]	[.247]	[.21]	[.71]	[.49]	[.115]	[.217]	[.461]	[.316]
RI p-value (Permuting AI Generated)	[.765]	[.241]	[.329]	[.563]	[.3]	[.266]	[.684]	[.518]	[.146]	[.28]	[.527]	[.365]
RI p-value (Permuting Gender)	[1]	[1]	[1]	[1]	[1]	[1]	[1]	[1]	[1]	[1]	[1]	[1]
Control Mean	3.19	2.94	3.107	2.929	2.893	2.583	3.083	3.071	3.25	3.012	2.94	2.667
P-Value (1) - (2)	.458	.009	.173	.203	.006	.126	.019	.047	.089	.359	.821	.634
P-Value (5) - (6)	.148	.608	.223	.526	.111	.546	.408	.078	.01	.003	.005	.111
N	366	366	366	366	366	366	366	366	366	366	366	366

Notes: * p<0.1, ** p<0.05, *** p<0.01. This table reports the heterogeneity in results using disaggregated quality measures by respondents' gender. Each outcome in columns 1 - 12 represents one of the quality statements asked in the survey using a likert scale ranging from 0 (very low) to 4 (very high). Strata fixed effects (countries) are controlled for and standard errors are clustered at the individual level (the unit of randomization).

Table A.14: Engagement by Overall quality of brief rated high or very high

	(1)	(2)	(3)
	Average engagement rating (0-4)	Standardized First Principal Component	Total likelihood of engagement (0-6)
(1) AI reported	-0.119	-0.129	-0.084
	(0.221)	(0.254)	(0.519)
RI p-value	[.581]	[.576]	[.864]
(2) AI Generated	-0.204	-0.224	-0.269
	(0.218)	(0.249)	(0.514)
RI p-value	[.329]	[.34]	[.569]
(3) AI reported × AI Generated	0.087	0.072	0.002
	(0.291)	(0.328)	(0.695)
RI p-value (Permuting AI reported)	[.755]	[.817]	[1]
RI p-value (Permuting AI Generated)	[.768]	[.814]	[1]
(4) Overall quality of brief rated high or very high	0.565***	0.664***	1.625***
	(0.211)	(0.248)	(0.498)
RI p-value	[.002]	[.003]	[0]
(5) AI reported × Overall quality of brief rated high or very high	0.319	0.379	0.469
	(0.256)	(0.302)	(0.616)
RI p-value (Permuting AI reported)	[.199]	[.193]	[.406]
RI p-value (Permuting Overall quality of brief rated high or very high)	[.205]	[.203]	[.432]
(6) AI Generated × Overall quality of brief rated high or very high	0.276	0.320	0.324
	(0.261)	(0.308)	(0.625)
RI p-value (Permuting AI Generated)	[.261]	[.259]	[.577]
RI p-value (Permuting Overall quality of brief rated high or very high)	[.391]	[.385]	[.629]
(7) AI reported × AI Generated × Overall quality of brief rated high or very high	-0.329	-0.375	-0.474
	(0.354)	(0.414)	(0.853)
RI p-value (Permuting AI reported)	[.325]	[.345]	[.566]
RI p-value (Permuting AI Generated)	[.367]	[.38]	[.597]
RI p-value (Permuting Overall quality of brief rated high or very high)	[.374]	[.399]	[.606]
Control Mean	3.73	.092	3.905
P-Value (1) - (2)	.635	.639	.675
P-Value (5) - (6)	.85	.825	.799
N	366	366	366

Notes: * p<0.1, ** p<0.05, *** p<0.01. This table reports the the heterogeneity in the likelihood of engagement with the policy blog by the respondents' overall rating of the policy blog's quality. The outcome in the first column is the average rating across six statements, concerning the intended future engagement with the brief. The responses follow a likert scale ranging from 0 to 4, with 0 indicating very low likelihood and 4 indicating very high likelihood of engagement. The second column presents the score of the first principal component, calculated based on the same six statements. The third column represents the number of questions where the respondent indicated they are 'likely' or 'very likely' to engage with the policy brief. Strata fixed effects (countries) are controlled for and standard errors are clustered at the individual level (the unit of randomization).

Table A.15: Disaggregated Engagement by Overall quality of blog rated high or very high

	(1)	(2)	(3)	(4)	(5)	(6)
	Share the blog with others	Re-read the blog	Look up studies cited	Look up related studies	Contact the authors of the brief	Use the brief in decision-making
(1) AI reported	-0.306	-0.096	0.201	-0.079	-0.456	0.022
	(0.287)	(0.326)	(0.306)	(0.277)	(0.289)	(0.298)
RI p-value	[.299]	[.748]	[.483]	[.762]	[.102]	[.937]
(2) AI Generated	-0.497*	-0.486	-0.179	-0.039	-0.077	0.055
	(0.284)	(0.314)	(0.279)	(0.257)	(0.316)	(0.288)
RI p-value	[.082]	[.139]	[.514]	[.872]	[.810]	[.859]
(3) AI reported × AI Generated	0.279	-0.023	-0.194	0.056	0.398	0.003
	(0.388)	(0.437)	(0.388)	(0.352)	(0.405)	(0.388)
RI p-value (Permuting AI reported)	[.478]	[.962]	[.62]	[.88]	[.325]	[.992]
RI p-value (Permuting AI Generated)	[.487]	[.963]	[.625]	[.878]	[.327]	[.991]
(4) Overall quality of brief rated high or very high	0.681**	0.332	0.596**	0.305	0.433	1.043***
	(0.274)	(0.290)	(0.267)	(0.258)	(0.311)	(0.264)
RI p-value	[.006]	[.201]	[.007]	[.176]	[.217]	[0]
(5) AI reported × Overall quality of brief rated high or very high	0.441	0.330	0.009	0.237	0.690*	0.208
	(0.336)	(0.383)	(0.357)	(0.335)	(0.374)	(0.336)
RI p-value (Permuting AI reported)	[.171]	[.379]	[.981]	[.442]	[.046]	[.537]
RI p-value (Permuting Overall quality of brief rated high or very high)	[.154]	[.295]	[.979]	[.423]	[.072]	[.528]
(6) AI Generated × Overall quality of brief rated high or very high	0.577*	0.648*	0.195	0.040	0.132	0.066
	(0.345)	(0.368)	(0.333)	(0.322)	(0.401)	(0.335)
RI p-value (Permuting AI Generated)	[.075]	[.095]	[.525]	[.899]	[.762]	[.835]
RI p-value (Permuting Overall quality of brief rated high or very high)	[.114]	[.118]	[.554]	[.906]	[.831]	[.849]
(7) AI reported × AI Generated × Overall quality of brief rated high or very high	-0.536	-0.104	-0.046	-0.061	-0.812	-0.413
	(0.488)	(0.525)	(0.473)	(0.444)	(0.543)	(0.457)
RI p-value (Permuting AI reported)	[.229]	[.831]	[.91]	[.886]	[.122]	[.34]
RI p-value (Permuting AI Generated)	[.272]	[.851]	[.937]	[.884]	[.125]	[.353]
RI p-value (Permuting Overall quality of brief rated high or very high)	[.283]	[.842]	[.917]	[.877]	[.146]	[.361]
Control Mean	3.929	3.833	3.81	3.798	3.274	3.738
P-Value (1) - (2)	.429	.172	.156	.862	.16	.897
P-Value (5) - (6)	.652	.348	.566	.503	.12	.639
N	366	366	366	366	366	366

Notes: * p<0.1, ** p<0.05, *** p<0.01. This table reports heterogeneity in respondents' likelihood of engagement with the policy blog by respondents' overall quality rating of the blog. Each outcome in columns 1 - 6 represent respondents' belief that their colleagues will engage in a particular activity on a likert scale ranging from 0 (very unlikely) to 4 (very likely). Strata fixed effects (countries) are controlled for and standard errors are clustered at the individual level (the unit of randomization).

Table A.16: Engagement by Gender

	(1)	(2)	(3)
	Average engagement rating (0-4)	Standardized First Principal Component	Total likelihood of engagement (0-6)
(1) AI reported	0.125	0.164	0.112
	(0.144)	(0.173)	(0.373)
RI p-value	[.383]	[.339]	[.774]
(2) AI Generated	-0.072	-0.075	-0.239
	(0.149)	(0.178)	(0.368)
RI p-value	[.662]	[.693]	[.501]
(3) AI reported × AI Generated	-0.160	-0.213	-0.177
	(0.204)	(0.244)	(0.521)
RI p-value (Permuting AI reported)	[.441]	[.404]	[.728]
RI p-value (Permuting AI Generated)	[.448]	[.382]	[.743]
(4) Gender	-0.797**	-0.890**	-1.858**
	(0.369)	(0.420)	(0.803)
RI p-value	[1]	[1]	[1]
(5) AI reported × Gender	-0.481*	-0.585*	-0.369
	(0.285)	(0.338)	(0.632)
RI p-value (Permuting AI reported)	[.108]	[.091]	[.598]
RI p-value (Permuting Gender)	[1]	[1]	[1]
(6) AI Generated × Gender	-0.286	-0.324	-0.470
	(0.274)	(0.327)	(0.652)
RI p-value (Permuting AI Generated)	[.325]	[.345]	[.484]
RI p-value (Permuting Gender)	[1]	[1]	[1]
(7) AI reported × AI Generated × Gender	0.823**	0.979**	1.123
	(0.396)	(0.468)	(0.880)
RI p-value (Permuting AI reported)	[.058]	[.056]	[.25]
RI p-value (Permuting AI Generated)	[.067]	[.063]	[.273]
RI p-value (Permuting Gender)	[1]	[1]	[1]
Control Mean	3.73	.092	3.905
P-Value (1) - (2)	.164	.159	.334
P-Value (5) - (6)	.525	.469	.885
N	366	366	366

Notes: * p<0.1, ** p<0.05, *** p<0.01. This table reports the the heterogeneity in the likelihood of engagement with the policy blog by gender. The outcome in the first column is the average rating across six statements, concerning the intended future engagement with the brief. The responses follow a likert scale ranging from 0 to 4, with 0 indicating very low likelihood and 4 indicating very high likelihood of engagement. The second column presents the score of the first principal component, calculated based on the same six statements. The third column represents the number of questions where the respondent indicated they are 'likely' or 'very likely' to engage with the policy brief. Strata fixed effects (countries) are controlled for and standard errors are clustered at the individual level (the unit of randomization).

Table A.17: Disaggregated Engagement by Gender

	(1)	(2)	(3)	(4)	(5)	(6)
	Share the blog with others	Re-read the blog	Look up studies cited	Look up related studies	Contact the authors of the brief	Use the brief in decision-making
(1) AI reported	-0.055	0.307	0.357**	0.139	-0.164	0.166
	(0.187)	(0.192)	(0.175)	(0.169)	(0.231)	(0.183)
RI p-value	[.759]	[.105]	[.041]	[.399]	[.466]	[.354]
(2) AI Generated	-0.206	-0.003	-0.032	0.012	-0.133	-0.068
	(0.196)	(0.198)	(0.185)	(0.178)	(0.232)	(0.193)
RI p-value	[.31]	[.982]	[.845]	[.948]	[.575]	[.737]
(3) AI reported × AI Generated	-0.042	-0.233	-0.392	-0.145	0.010	-0.156
	(0.278)	(0.273)	(0.249)	(0.238)	(0.316)	(0.261)
RI p-value (Permuting AI reported)	[.892]	[.406]	[.133]	[.531]	[.975]	[.571]
RI p-value (Permuting AI Generated)	[.872]	[.396]	[.101]	[.541]	[.958]	[.608]
(4) Gender	-0.985*	-0.356	-0.952**	-1.069**	-1.128**	-0.294
	(0.536)	(0.523)	(0.479)	(0.477)	(0.458)	(0.476)
RI p-value	[1]	[1]	[1]	[1]	[1]	[1]
(5) AI reported × Gender	-0.258	-0.973**	-0.864**	-0.467	0.235	-0.561
	(0.358)	(0.402)	(0.392)	(0.368)	(0.393)	(0.364)
RI p-value (Permuting AI reported)	[.504]	[.013]	[.032]	[.25]	[.563]	[.128]
RI p-value (Permuting Gender)	[1]	[1]	[1]	[1]	[1]	[1]
(6) AI Generated × Gender	-0.364	-0.700*	-0.503	-0.300	0.183	-0.030
	(0.374)	(0.370)	(0.348)	(0.322)	(0.397)	(0.351)
RI p-value (Permuting AI Generated)	[.376]	[.085]	[.187]	[.363]	[.66]	[.934]
RI p-value (Permuting Gender)	[1]	[1]	[1]	[1]	[1]	[1]
(7) AI reported × AI Generated × Gender	0.836	1.137**	1.257**	0.983**	0.127	0.596
	(0.532)	(0.557)	(0.518)	(0.463)	(0.549)	(0.488)
RI p-value (Permuting AI reported)	[.195]	[.046]	[.025]	[.061]	[.804]	[.248]
RI p-value (Permuting AI Generated)	[.165]	[.08]	[.033]	[.058]	[.846]	[.272]
RI p-value (Permuting Gender)	[1]	[1]	[1]	[1]	[1]	[1]
Control Mean	3.929	3.833	3.81	3.798	3.274	3.738
P-Value (1) - (2)	.417	.11	.02	.436	.889	.215
P-Value (5) - (6)	.799	.509	.386	.639	.897	.172
N	366	366	366	366	366	366

Notes: * p<0.1, ** p<0.05, *** p<0.01. This table reports heterogeneity in respondents' likelihood of engagement with the policy blog by respondents' gender. Each outcome in columns 1 - 6 represent respondents' belief that their colleagues will engage in a particular activity on a likert scale ranging from 0 (very unlikely) to 4 (very likely). Strata fixed effects (countries) are controlled for and standard errors are clustered at the individual level (the unit of randomization).

Table A.18: Others' Engagement by Overall quality of blog rated high or very high

	(1)	(2)	(3)	(4)
	Average rating (0-4)	Standardized First Principal Component	Total likelihood of engagement (0-6)	Others' overall rating
(1) AI reported	-0.371*	-0.436*	-0.638	0.015
	(0.212)	(0.250)	(0.628)	(0.102)
RI p-value	[.086]	[.078]	[.298]	[.908]
(2) AI Generated	-0.150	-0.167	-0.371	-0.018
	(0.205)	(0.246)	(0.631)	(0.100)
RI p-value	[.484]	[.519]	[.554]	[.859]
(3) AI reported × AI Generated	0.154	0.161	0.337	-0.031
	(0.284)	(0.332)	(0.817)	(0.132)
RI p-value (Permuting AI reported)	[.592]	[.633]	[.662]	[.815]
RI p-value (Permuting AI Generated)	[.605]	[.643]	[.7]	[.824]
(4) Overall quality of blog rated high or very high	0.193	0.249	0.781	0.686***
	(0.193)	(0.236)	(0.583)	(0.095)
RI p-value	[.399]	[.381]	[.128]	[0]
(5) AI reported × Overall quality of blog rated high or very high	0.567**	0.672**	1.119	0.043
	(0.257)	(0.310)	(0.726)	(0.126)
RI p-value (Permuting AI reported)	[.022]	[.024]	[.115]	[.759]
RI p-value (Permuting Overall quality of blog rated high or very high)	[.036]	[.038]	[.088]	[.772]
(6) AI Generated × Overall quality of blog rated high or very high	0.253	0.302	0.665	0.015
	(0.251)	(0.308)	(0.736)	(0.123)
RI p-value (Permuting AI Generated)	[.31]	[.329]	[.338]	[.911]
RI p-value (Permuting Overall quality of blog rated high or very high)	[.472]	[.498]	[.419]	[.927]
(7) AI reported × AI Generated × Overall quality of blog rated high or very high	-0.419	-0.491	-1.077	-0.141
	(0.358)	(0.432)	(0.988)	(0.172)
RI p-value (Permuting AI reported)	[.246]	[.253]	[.263]	[.439]
RI p-value (Permuting AI Generated)	[.237]	[.249]	[.266]	[.468]
RI p-value (Permuting Overall quality of blog rated high or very high)	[.315]	[.329]	[.303]	[.491]
Control Mean	3.627	.091	3.619	.679
P-Value (1) - (2)	.252	.236	.628	.742
P-Value (5) - (6)	.201	.214	.501	.828
N	366	366	366	366

Notes: * p<0.1, ** p<0.05, *** p<0.01. This table reports heterogeneity in respondents' beliefs of their colleagues' likelihood of engagement with the policy blog by respondents' overall rating of the blog. The outcome in the first column is the average rating across six statements, concerning their colleagues' intended future engagement with the blog. The responses follow a likert scale ranging from 0 to 4, with 0 indicating very low likelihood and 4 indicating very high likelihood of colleagues' engagement. The second column presents the score of the first principal component, calculated based on the same six statements. The third column represents the number of questions where the respondent indicated they believe their colleagues are 'likely' or 'very likely' to engage with the policy blog. Strata fixed effects (countries) are controlled for and standard errors are clustered at the individual level (the unit of randomization).

Table A.19: Others' Disaggregated Engagement by Overall quality of blog rated high or very high

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Share the blog with others	Re-read the blog	Look up studies cited	Look up related studies	Contact the authors of the blog	Use the blog in decision-making	Rate the quality of the policy blog overall?
(1) AI reported	-0.475*	-0.246	-0.146	-0.191	-0.545*	-0.622**	0.228
	(0.258)	(0.269)	(0.273)	(0.261)	(0.283)	(0.288)	(0.181)
RI p-value	[.066]	[.359]	[.594]	[.465]	[.051]	[.034]	[.215]
(2) AI Generated	-0.259	-0.209	-0.135	0.024	0.067	-0.386	-0.057
	(0.243)	(0.253)	(0.252)	(0.264)	(0.307)	(0.279)	(0.173)
RI p-value	[.295]	[.425]	[.6]	[.93]	[.838]	[.168]	[.752]
(3) AI reported × AI Generated	0.056	-0.010	-0.030	0.124	0.244	0.542	-0.189
	(0.341)	(0.371)	(0.343)	(0.343)	(0.389)	(0.373)	(0.227)
RI p-value (Permuting AI reported)	[.868]	[.991]	[.924]	[.74]	[.517]	[.15]	[.413]
RI p-value (Permuting AI Generated)	[.86]	[.985]	[.938]	[.752]	[.517]	[.168]	[.45]
(4) Overall quality of blog rated high or very high	0.244	0.228	0.204	0.153	0.045	0.283	1.157***
	(0.230)	(0.247)	(0.243)	(0.246)	(0.299)	(0.257)	(0.178)
RI p-value	[.357]	[.376]	[.387]	[.457]	[.9]	[.399]	[0]
(5) AI reported × Overall quality of blog rated high or very high	0.562*	0.369	0.372	0.422	0.857**	0.818**	-0.126
	(0.315)	(0.335)	(0.326)	(0.326)	(0.370)	(0.339)	(0.232)
RI p-value (Permuting AI reported)	[.073]	[.257]	[.258]	[.171]	[.026]	[.019]	[.563]
RI p-value (Permuting Overall quality of blog rated high or very high)	[.119]	[.201]	[.178]	[.139]	[.051]	[.048]	[.633]
(6) AI Generated × Overall quality of blog rated high or very high	0.285	0.335	0.256	0.234	0.047	0.364	0.014
	(0.291)	(0.314)	(0.315)	(0.325)	(0.388)	(0.322)	(0.228)
RI p-value (Permuting AI Generated)	[.309]	[.318]	[.409]	[.487]	[.894]	[.281]	[.953]
RI p-value (Permuting Overall quality of blog rated high or very high)	[.395]	[.439]	[.469]	[.458]	[.918]	[.42]	[.964]
(7) AI reported × AI Generated × Overall quality of blog rated high or very high	-0.218	-0.119	-0.296	-0.494	-0.546	-0.842*	-0.077
	(0.429)	(0.471)	(0.437)	(0.447)	(0.519)	(0.452)	(0.308)
RI p-value (Permuting AI reported)	[.592]	[.803]	[.539]	[.278]	[.285]	[.063]	[.77]
RI p-value (Permuting AI Generated)	[.6]	[.811]	[.478]	[.26]	[.272]	[.073]	[.845]
RI p-value (Permuting Overall quality of blog rated high or very high)	[.668]	[.801]	[.457]	[.254]	[.348]	[.167]	[.85]
Control Mean	3.893	3.667	3.595	3.548	3.226	3.833	3.798
P-Value (1) - (2)	.342	.883	.962	.39	.02	.336	.086
P-Value (5) - (6)	.316	.913	.708	.556	.021	.144	.531
N	366	366	366	366	366	366	366

Notes: * p<0.1, ** p<0.05, *** p<0.01. This table reports heterogeneity in respondents' beliefs of their colleagues' likelihood of engagement with the policy blog by respondents' overall rating of the blog. Each outcome in columns 1 - 6 represent respondents' belief that their colleagues will engage in a particular activity on a likert scale ranging from 0 (very unlikely) to 4 (very likely). The seventh column presents the results using an outcome of respondents' belief of their colleagues' overall quality rating on likert scale ranging from 0 (very low) to 4 (very high). Strata fixed effects (countries) are controlled for and standard errors are clustered at the individual level (the unit of randomization).

Table A.20: Others' Engagement by Gender

	(1)	(2)	(3)	(4)
	Average rating (0-4)	Standardized First Principal Component	Total likelihood of engagement (0-6)	Others' overall rating
(1) AI reported	0.078	0.106	0.086	-0.046
	(0.139)	(0.173)	(0.392)	(0.090)
RI p-value	[.561]	[.541]	[.831]	[.608]
(2) AI Generated	-0.013	-0.004	-0.222	-0.156*
	(0.144)	(0.178)	(0.403)	(0.090)
RI p-value	[.942]	[.985]	[.61]	[.098]
(3) AI reported × AI Generated	-0.239	-0.305	-0.383	0.009
	(0.199)	(0.245)	(0.552)	(0.126)
RI p-value (Permuting AI reported)	[.215]	[.205]	[.459]	[.939]
RI p-value (Permuting AI Generated)	[.26]	[.233]	[.499]	[.958]
(4) Gender	-0.285	-0.386	-0.792	-0.186
	(0.352)	(0.413)	(0.934)	(0.222)
RI p-value	[1]	[1]	[1]	[1]
(5) AI reported × Gender	-0.596**	-0.757**	-0.744	0.032
	(0.291)	(0.347)	(0.733)	(0.167)
RI p-value (Permuting AI reported)	[.047]	[.036]	[.328]	[.862]
RI p-value (Permuting Gender)	[1]	[1]	[1]	[1]
(6) AI Generated × Gender	-0.130	-0.163	0.258	0.122
	(0.255)	(0.318)	(0.689)	(0.171)
RI p-value (Permuting AI Generated)	[.669]	[.646]	[.725]	[.528]
RI p-value (Permuting Gender)	[1]	[1]	[1]	[1]
(7) AI reported × AI Generated × Gender	0.947**	1.153**	1.500	0.118
	(0.399)	(0.480)	(1.003)	(0.242)
RI p-value (Permuting AI reported)	[.024]	[.023]	[.165]	[.667]
RI p-value (Permuting AI Generated)	[.071]	[.07]	[.221]	[.687]
RI p-value (Permuting Gender)	[1]	[1]	[1]	[1]
Control Mean	3.627	.091	3.619	.679
P-Value (1) - (2)	.518	.523	.435	.223
P-Value (5) - (6)	.124	.102	.197	.626
N	366	366	366	366

Notes: * p<0.1, ** p<0.05, *** p<0.01. This table reports heterogeneity in respondents' beliefs of their colleagues' likelihood of engagement with the policy blog by gender. The outcome in the first column is the average rating across six statements, concerning their colleagues' intended future engagement with the blog. The responses follow a likert scale ranging from 0 to 4, with 0 indicating very low likelihood and 4 indicating very high likelihood of colleagues' engagement. The second column presents the score of the first principal component, calculated based on the same six statements. The third column represents the number of questions where the respondent indicated they believe their colleagues are 'likely' or 'very likely' to engage with the policy blog. Strata fixed effects (countries) are controlled for and standard errors are clustered at the individual level (the unit of randomization).

Table A.21: Others' Disaggregated Engagement by Gender

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Share the blog with others	Re-read the blog	Look up studies cited	Look up related studies	Contact the authors of the blog	Use the blog in decision-making	Rate the quality of the policy blog overall?
(1) AI reported	0.010	0.166	0.256	0.176	-0.088	-0.054	0.037
	(0.159)	(0.173)	(0.169)	(0.178)	(0.214)	(0.174)	(0.152)
RI p-value	[.941]	[.325]	[.119]	[.304]	[.677]	[.739]	[.810]
(2) AI Generated	-0.064	0.030	0.044	0.152	-0.034	-0.204	-0.269*
	(0.151)	(0.177)	(0.177)	(0.188)	(0.221)	(0.175)	(0.161)
RI p-value	[.706]	[.886]	[.814]	[.452]	[.886]	[.266]	[.121]
(3) AI reported × AI Generated	-0.247	-0.359	-0.393	-0.371	-0.074	0.011	-0.006
	(0.215)	(0.248)	(0.238)	(0.252)	(0.291)	(0.248)	(0.209)
RI p-value (Permuting AI reported)	[.257]	[.126]	[.081]	[.133]	[.763]	[.975]	[.985]
RI p-value (Permuting AI Generated)	[.259]	[.127]	[.105]	[.169]	[.821]	[.959]	[.978]
(4) Gender	-0.178	-0.147	-0.531	-0.674*	-0.232	0.054	-0.753*
	(0.419)	(0.417)	(0.408)	(0.396)	(0.514)	(0.452)	(0.451)
RI p-value	[1]	[1]	[1]	[1]	[1]	[1]	[1]
(5) AI reported × Gender	-0.761**	-0.933**	-0.862**	-0.602*	0.085	-0.505	-0.034
	(0.372)	(0.377)	(0.340)	(0.334)	(0.421)	(0.360)	(0.318)
RI p-value (Permuting AI reported)	[.057]	[.024]	[.012]	[.075]	[.826]	[.185]	[.911]
RI p-value (Permuting Gender)	[1]	[1]	[1]	[1]	[1]	[1]	[1]
(6) AI Generated × Gender	-0.312	-0.367	-0.269	-0.108	0.406	-0.127	0.113
	(0.323)	(0.342)	(0.327)	(0.316)	(0.381)	(0.326)	(0.297)
RI p-value (Permuting AI Generated)	[.361]	[.41]	[.456]	[.759]	[.305]	[.749]	[.735]
RI p-value (Permuting Gender)	[1]	[1]	[1]	[1]	[1]	[1]	[1]
(7) AI reported × AI Generated × Gender	1.117**	1.540***	1.087**	1.114**	0.233	0.591	0.063
	(0.508)	(0.523)	(0.469)	(0.446)	(0.552)	(0.485)	(0.425)
RI p-value (Permuting AI reported)	[.042]	[.009]	[.025]	[.019]	[.669]	[.273]	[.867]
RI p-value (Permuting AI Generated)	[.061]	[.029]	[.07]	[.045]	[.688]	[.342]	[.885]
RI p-value (Permuting Gender)	[1]	[1]	[1]	[1]	[1]	[1]	[1]
Control Mean	3.893	3.667	3.595	3.548	3.226	3.833	3.798
P-Value (1) - (2)	.622	.411	.207	.894	.79	.414	.048
P-Value (5) - (6)	.223	.142	.098	.164	.446	.287	.652
N	366	366	366	366	366	366	366

Notes: * p<0.1, ** p<0.05, *** p<0.01. This table reports heterogeneity in respondents' beliefs of their colleagues' likelihood of engagement with the policy blog by gender. Each outcome in columns 1 - 6 represent respondents' belief that their colleagues will engage in a particular activity on a likert scale ranging from 0 (very unlikely) to 4 (very likely). The seventh column presents the results using an outcome of respondents' belief of their colleagues' overall quality rating on likert scale ranging from 0 (very low) to 4 (very high). Strata fixed effects (countries) are controlled for and standard errors are clustered at the individual level (the unit of randomization).

B AI Disclosure in Scientific Writing

Below is Elsevier's requirements for reporting the use of AI in writing:

Declaration of generative AI in scientific writing

The below guidance only refers to the writing process, and not to the use of AI tools to analyse and draw insights from data as part of the research process.

Where authors use generative artificial intelligence (AI) and AI-assisted technologies in the writing process, authors should only use these technologies to improve readability and language. Applying the technology should be done with human oversight and control, and authors should carefully review and edit the result, as AI can generate authoritative-sounding output that can be incorrect, incomplete or biased. AI and AI-assisted technologies should not be listed as an author or co-author, or be cited as an author. Authorship implies responsibilities and tasks that can only be attributed to and performed by humans, as outlined in Elsevier's AI policy for authors.

Authors should disclose in their manuscript the use of AI and AI-assisted technologies in the writing process by following the instructions below. A statement will appear in the published work. Please note that authors are ultimately responsible and accountable for the contents of the work.

Disclosure instructions

Authors must disclose the use of generative AI and AI-assisted technologies in the writing process by adding a statement at the end of their manuscript in the core manuscript file, before the References list. The statement should be placed in a new section entitled 'Declaration of Generative AI and AI-assisted technologies in the writing process'.

Statement: During the preparation of this work the author(s) used [NAME TOOL / SERVICE] in order to [REASON]. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

This declaration does not apply to the use of basic tools for checking grammar, spelling, references etc. If there is nothing to disclose, there is no need to add a statement.

C Creation of AI-Written Policy Briefs

AI-generated policy briefs are generated through a uniform, iterative process using the following prompts:

You are a food policy researcher working at the International Food Policy Research Institute. You will write a blog post to summarize a research paper you published recently with co-authors. The main content of the blog will be a story of the research. Your target audience is thinkers and donors, and an informed general public interested in Sustainable Development Goals and food policy-related issues in developing countries. Write less formally, but not conversationally. Avoid jargon. Use the tone as if you are explaining research to a colleague from a different background. Don't repeat the same thought. Start the blog with a headline, which is descriptive and suggestive of the relevant findings for readers get a sense of what they are getting into. Outline the question this document addresses, and the approach pursued. Frame the main point quickly, in paragraph or two. Include noteworthy details, such as challenges faced in the field. Explain the findings and their significance. Use bullet points to present policy recommendations. Conclude by reflecting on what your research shows about the broader issue, or directions for future research. Avoid repeating points already made. Don't use subheadings. Use the active voice. Use a less formal but still professional tone. Avoid using

adverbs or adjectives. If you understood this requirement, answer "Understood". You don't have to draft the blog at this time.

Let's start with the headline. Can you give me a good headline for the blog?

Good! Now, write me the introduction. Without using a section heading, the introduction should set up the rest of the blog and clearly convey the argument of the document. In two paragraphs, clearly describe the challenges faced. Define why you are writing the blog and express the urgency and importance of the topic. This introduction should contain all of the information relevant to the title of the document and the main objectives of the research, based on the key questions. Do not use "the study," "the research," or "in conclusion" and make this section as a standalone text that can be read and understood assuming readers do not have access to the whole PDF document.

Looks good! Let's move on to the Research Overview section. Without using a section heading, analyze the uploaded PDF document and draft the section to provide a summary of the research that describe the issues, contexts, data, and research methods in two paragraphs. In the first paragraph, explain the purpose of the research, how the study was conducted, how the data was collected and used, and any other relevant background information. In the second paragraph, explain the research results first and moving on to the specific research findings. Do not use "the study," "the research," or "in conclusion" and make this section as a standalone text that can be read and understood assuming readers do not have access to the whole PDF document. Use a less formal but still professional tone.

Looks good! Let's move on to the Key Findings section. Without using a section heading, this section should interpret the data in a way that is accessible and clearly connected to the policy recommendations. The goal is to be convincing, but ensure that the analysis is balanced and defensible. Explain the key research findings clearly and comprehensively. Use the active voice primarily. Provide a high-level highlight of the key findings in one sentence at the beginning, followed by presenting key findings one by one. Do not use "the study," "the research," or "in conclusion" and make this section as a standalone text that can be read and understood assuming readers do not have access to the whole PDF document. Use a less formal but still professional tone.

Looks good! Let's move on to the Policy Recommendations section. Detail the actions recommended by the research findings. Draw the link between the Research Findings and the Policy Recommendations. Use persuasive language to present each of the recommendations. Try to completely convince readers that the presented Policy Recommendations are the best advice. Describe the potential consequences of implementing particular policies. Include the consequences of inaction as well if found in the PDF document. Recommendations should act as a call to action by stating precise, relevant, credible, and feasible next steps. Provide a high-level highlight of the key findings in one sentence at the beginning, followed by presenting key findings one by one in about five bullet points. Do not use "the study," "the research," or "in conclusion" and make this section as a standalone text that can be read and understood assuming readers do not have access to the whole PDF document. Use a less formal but still professional tone.

Looks good! Finally, let's add a Conclusion section. In one short paragraph of text without using bullet points, briefly summarize the Policy Recommendations recommended by the Key Research Findings. Explain the reasons behind the Policy Recommendations. The conclusion section should link the research findings to the policy recommendations. End the section with a concluding statement that reiterates the key message and suggests the significance of the proposed Policy Recommendations. Do not use and make this section as a standalone text that can be read and understood assuming readers do not have access to the whole PDF document. Use a professional tone at the same level used in the PDF document. Use the passive voice

primarily.

D Text of Emails and Social Media Posts

D.1 Email Invitation to Survey

SUBJECT LINE: Invitation: Help improve research-based policy recommendations for [country].

Greetings from IFPRI.

As part of our ongoing efforts to provide research-based policy solutions in [country], we invite you to participate in an important survey. We expect that your participation in the survey will make a significant contribution to improve the role of science in policy making The survey will take about 20 minutes to complete. To respect your privacy and adhere to General Data Protection Regulation (GDPR) guidelines, we assure you that your participation is entirely voluntary, and your responses will be anonymous. By clicking the survey link below, you will be guided to the survey form:

https://innovationifpri.surveycto.com/collect/ai$_{p}olicy_{b}rief_{[}country]?caseid =$

If you have any concerns about your data privacy or would like to learn more about how we handle your information, please review our Privacy & Cookie Policy or contact our Data team IFPRI-Data@CGIAR.ORG.

Sincerely,
[Country lead name and position]

D.2 X post

Passionate about research-based policymaking and better communication with policymakers? Join our survey and make an impact!

https://t.co/8OA3YdSBRn

Choose your country and participate for a chance to win 100!

@IFPRI #NPSInitiative #DigitalInnovationInitiative

D.3 LinkedIn

Are you keen in promoting research-based policymaking and contribute to our understanding what role Artificial Intelligence might play in this? Then please select one country that you are professionally most engaged with and participate in our survey. The survey will take about 20 minutes to complete and you will have a chance to win 100 USD by taking the survey. Kenya Egypt Bangladesh Ethiopia Ghana India Malawi Nigeria Rwanda Sudan Tajikistan Uganda . The survey is led by #IFPRI under the CGIAR Initiatives on National Policies and Strategies #NPSInitiative and Digital Innovation #DigitalInnovationInitiative.

ALL IFPRI DISCUSSION PAPERS

All discussion papers are available **here**

They can be downloaded free of charge

INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE
www.ifpri.org

IFPRI HEADQUARTERS
1201 Eye Street, NW
Washington, DC 20005 USA
Tel.: +1-202-862-5600
Fax: +1-202-862-5606
Email: ifpri@cgiar.org