

Measuring Human Contribution in AI-Assisted Content Generation

Yueqi Xie^{1,2*†}, Tao Qi^{3*†}, Jingwei Yi^{4*†}, Ryan Whalen⁵, Junming Huang², Qian Ding⁶, Yu Xie^{2,7}, Xing Xie⁶, and Fangzhao Wu^{6†}

¹Hong Kong University of Science and Technology, Hong Kong

²Paul and Marcia Center on Contemporary China, Princeton University, Princeton, NJ 08544, United States

³Tsinghua University, Beijing, 100084, China

⁴University of Science and Technology of China, Hefei 230026, China

⁵The University of Hong Kong, Hong Kong

⁶Microsoft Research Asia, Beijing 100080, China

⁷Center for Social Research, Guanghua School of Management, Peking University, Beijing 100871, China

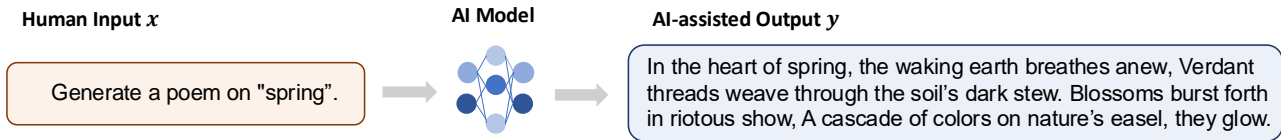
*Joint First Authors

†Correspondence: yxieay@connect.ust.hk, taoqi.qt@gmail.com, yjw1029@mail.ustc.edu.cn, fangzhu@microsoft.com

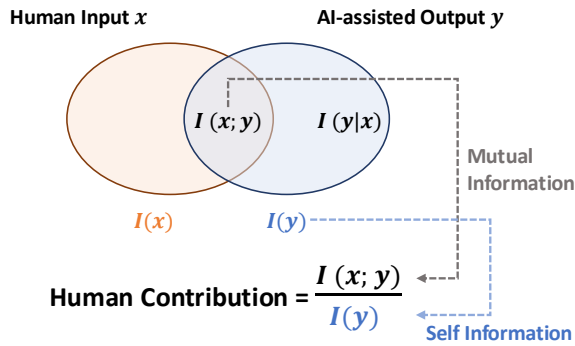
ABSTRACT

With the growing prevalence of generative artificial intelligence (AI), an increasing amount of content is no longer exclusively generated by humans but by generative AI models with human guidance. This shift presents notable challenges for the delineation of originality due to the varying degrees of human contribution in AI-assisted works. This study raises the research question of measuring human contribution in AI-assisted content generation and introduces a framework to address this question that is grounded in information theory. By calculating mutual information between human input and AI-assisted output relative to self-information of AI-assisted output, we quantify the proportional information contribution of humans in content generation. Our experimental results demonstrate that the proposed measure effectively discriminates between varying degrees of human contribution across multiple creative domains. We hope that this work lays a foundation for measuring human contributions in AI-assisted content generation in the era of generative AI.

a AI-Assisted Generation



b Measure



c Results

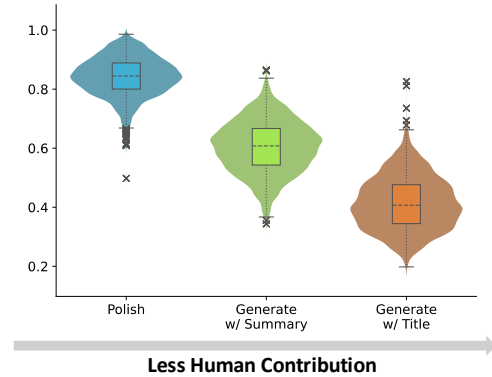


Figure 1. a. Illustration of AI-assisted content generation, where an AI model is prompted with human input and generates output. b. Overview of the proposed method for measuring human contribution, quantified by the ratio of mutual information between human input and AI-assisted output to the total self-information of the AI-assisted output. c. Outcomes of our proposed measure across various poem generation scenarios using Llama-3, involving varying degrees of human contribution (polishing a human poem, generation with the summary, in other words, key ideas, of a human poem, and generation with a poem title). The center line represents the median, the box limits indicate the upper and lower quartiles, the whiskers extend to 1.5x the interquartile range, and the points are outliers. Our measure effectively differentiates varying degrees of human contribution across the scenarios.

Recent advances in large language models (LLMs)^{1–5} have impacted our personal and working lives in significant ways, most notably by changing the process of content generation.^{6,7} Artificial intelligence (AI) “copilots” have emerged as a new and powerful content production tool across a variety of domains^{8–10}, such as lyrics creation¹¹, office work^{9,10}, academic writing⁸, etc. Consequently, an increasing amount of new content being generated is no longer solely created by humans but is rather the result of AI-assisted creation^{12–14}. In this new creative modality, humans contribute by providing prompts to AI models, resulting in the generation of “AI-assisted output”, as illustrated in Figure 1a.

This development has raised debates about determining the originality and corresponding regulation of content generated with AI assistance^{12,15}. The varying degrees of human contribution in AI-assisted generation complicate the attribution of intellectual contribution to AI-assisted outputs. This issue is particularly pertinent in fields that prioritize originality, such as education¹⁶, academic research^{17–19}, and creative work²⁰. For example, universities face a dilemma in whether to ban or embrace AI. Administrators and instructors are concerned that students might use AI to create materials for evaluation with varying levels of originality, potentially compromising educational fairness and effectiveness^{21,22}. Similarly, there is a growing debate, underscored by notable incidents^{23,24}, concerning the copyright eligibility of AI-assisted works^{20,25,26}.

At the two extreme ends of the human–AI contribution spectrum, the attribution of originality is relatively clear. If a human author simply uses AI to polish their document, it should be considered the result of the author’s own work. Conversely, if a human uses a short, less-informative prompt to generate a large amount of text, it will not reflect much of the human’s intellectual conception. However, there remains a substantial grey area between these two extremes, in which determining originality requires insight into the degree of *human contribution* during the AI-assisted generation process. Hence, there is an urgent need for a credible measure by which to evaluate human contribution in AI-assisted content generation.

In this paper, we address the quantification of human contribution in AI-assisted content generation. We begin with the recognition that a major obstacle is the lack of a well-defined perspective, or medium, by which to ascertain the extent to which content output can be attributed to humans rather than the AI tools they have used. Towards this goal, we introduce a new general framework within which we provide a preliminary attempt to quantify human contribution in AI-assisted content generation. Our framework hinges on the concept of *information content* as a modeling medium. Utilizing principles from information theory²⁷, as depicted in Figure 1b, our approach centers on the quantification of the proportion of the information

content in the AI-assisted output that can be attributed to human input. Specifically, it is a ratio of two quantities. The denominator is the total/unconditional information content (surprisal) in AI-assisted output, calculated as the negative logarithm of the probability of generating the AI-assisted generated content, which we refer to as *self-information*, $I(\mathbf{y})$. The numerator, $I(\mathbf{x};\mathbf{y})$, is the portion of self-information $I(\mathbf{y})$ that is shared with the total information content from human input, $I(\mathbf{x})$, which we define as *mutual information*. The difference between the two is the *conditional self-information* in AI-assisted output given user input, $I(\mathbf{y}|\mathbf{x})$, calculated as the negative logarithm of the probability of generating the AI-assisted output conditional on human input.

We systematically validate the proposed method as a reliable measure of human contribution by evaluating its effectiveness, domain adaptivity, and model adaptivity. To achieve this, we construct a comprehensive dataset of AI-assisted content generation, encompassing various levels of human contribution, multiple creative domains, and outputs from different LLMs. For instance, Figure 1c illustrates the distribution of the outcomes of our proposed measure for AI-assisted poem generation, across three varying levels of human contribution, ranging from high to low, using the LLM Llama-3. Our proposed measure effectively discriminates between varying degrees of contribution, generally producing lower values for content with less human contribution. Additionally, we investigate the impact of content length, resilience to adaptive attacks, and generalization of our method in evaluation. We further apply our measure to real-world human-AI co-creation data, demonstrating its practical applicability. In brief, this paper poses a novel research question on quantitatively evaluating human contribution in AI-assisted generation and presents a simple yet effective information-based measure as a potential solution.

Results

Table 1. Detailed statistics of the constructed dataset.

Type	Corpus	# Content Words	# Summary Words	# Title Words	# Subject Words
Paper Abstract	Arxiv	134.24±63.07	68.38±20.59	9.63±3.79	2.87±1.37
News	News Articles	532.85±86.26	78.36±16.76	8.91±2.32	4.06±1.00
Patent Abstract	HUPD	171.65±24.22	59.89±13.85	8.57±5.28	3.91±0.79
Poem	Poetry Foundation	208.18±94.08	48.20±11.27	3.65±2.80	-

Dataset Construction

To verify the reliability of the proposed measure of human contribution, we construct a dataset of AI-assisted generation data with known varying levels of human contribution. Note that there is no absolute ground truth for assessing human contribution. Therefore, by design, our dataset spans a very large range of human contribution in AI-assisted output with distinct levels that are hardly controversial. For a comprehensive evaluation, we further vary three factors, beyond the level of human contribution: (1) domains, focusing on those where originality protection is crucial; (2) different LLMs; and (3) different random generation runs. Building this dataset primarily involves two steps: *raw information collection and processing* and *AI-assisted content generation*.

Raw Information Collection and Processing: First, we collect and process *multi-level information* in various domains. Specifically, we sample raw data from public datasets across the following domains: academic paper abstracts, news articles, patent abstracts, and poems. We sample 2,000 entries for each domain. For paper abstracts, each raw data entry includes content, title, and subject; for the other three domains, each raw entry includes content and title. Details of the original dataset and sampling process are provided in Supplemental Materials Section 1.1. We further process the data into a uniform structure with decreasing levels of information: *content*, *summary*, *title*, and *subject* (except poems, because of short titles), with missing parts of the raw data supplemented using GPT-3.5². The corresponding statistics are presented in Table 1.

AI-Assisted Content Generation: Next, we generate new content using LLMs with varying levels of *human input* constructed from the earlier process, categorized as follows: *polishing*, *generation with summary*, *generation with title*, *generation with subject* (where applicable). These inputs use information corresponding to content, summary, title, and subject, respectively. The detailed prompt constructions are shown in Supplementary Materials Section 1.2. These human inputs represent varying levels of human contribution, from high to low, based on the amount of information provided. To support a comprehensive analysis, we apply different LLMs, including the state-of-the-art open-source LLMs Llama-3¹ and Mixtral²⁸ and the chatbot GPT-3.5². We generate 5 times for a human input with the temperature set as 0.7 for diverse outputs.

Human Contribution Evaluation

We evaluate the effectiveness of the proposed measure using the constructed dataset. This section focuses on the original scenario where both the AI-assisted output \mathbf{y} and human input \mathbf{x} are known, and the AI model M_θ 's output probability is

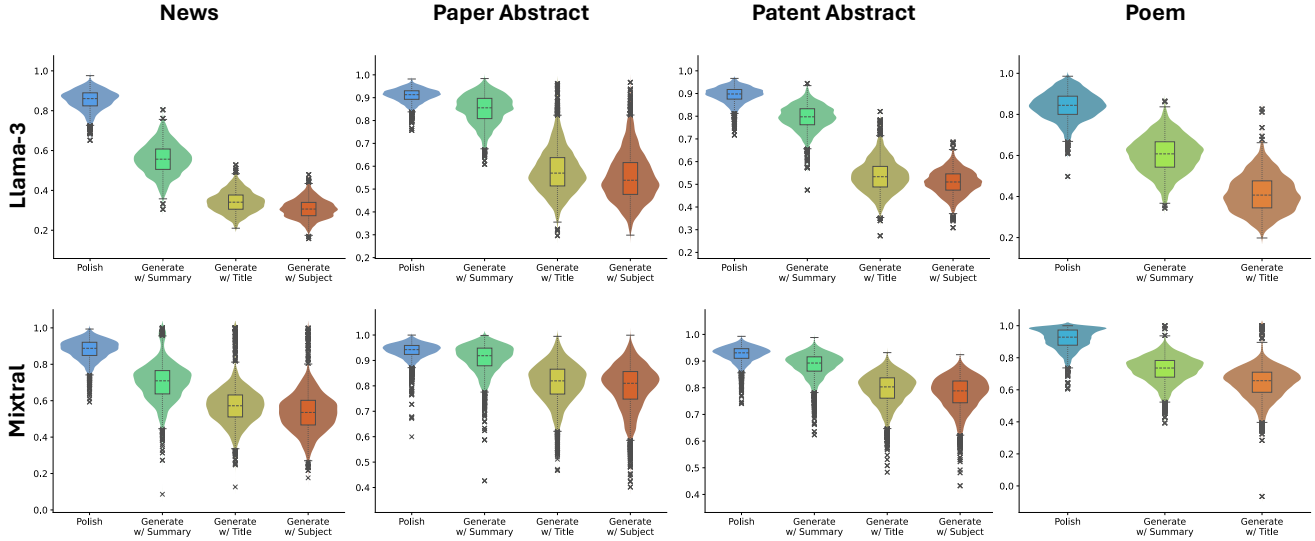


Figure 2. The distribution of the outcomes of the proposed measure for the constructed dataset. The center line represents the median, the box limits indicate the upper and lower quartiles, the whiskers extend to 1.5x the interquartile range, and the points are outliers. Overall, the proposed measure exhibits the expected trend that lower values are obtained for the generated content with less human contribution.

available in evaluation. As a result, the generative probability $p_{\theta}(y)$ and the conditional generative probability $p_{\theta}(y|x)$ for generating the content can be calculated. In this context, we can directly apply the previously discussed formulas to compute human contribution. Particularly, as shown in Figure 1b, human contribution is calculated as the ratio of the mutual information between human input and AI-assisted output to the total self-information of the AI-assisted output. The detailed calculation is provided in the Methods section. Real-world scenarios corresponding to this situation include *originality authentication*, where the human author can provide both the AI model’s generative distribution and the original human input for authentication and evaluation, as well as cases where *model service providers* directly apply the definition during the generation process to calculate a measurement of human contribution.

Figure 2 illustrates the human contribution results of two state-of-the-art open-source LLMs, Llama-3 and Mixtral, on the constructed dataset across various domains. From the results, we make the following observations. First, for each combination of model and data domain, varying levels of human contribution yield different distributions for measured human input in the expected direction: the lower the human author’s informational contribution in AI-assisted generation (from polishing, to generation with summary, to generation with title, and finally to generation with subject), the smaller the proposed metric’s value. For instance, when generating news articles with Llama-3, polished content demonstrates an average human contribution measurement of 85.37%, while content generated with a subject shows an average human contribution measurement of 30.83%. This indicates that our proposed measure can effectively distinguish different levels of human contribution in AI-assisted generation, providing useful measurements from an informational perspective.

Second, we observe variability in the outcomes of the proposed measure across different data domains for a specific generation mode (e.g., generation with summary). For instance, the average score for content generated with a summary using Llama-3 is 55.69% for news but 84.94% for paper abstracts. These differences are reasonable because the same generation mode does not necessarily equate to a similar percentage of human contribution across different domains. As discussed in the Dataset Construction section, without an absolute ground truth, we can only generate content with approximate varying levels of human contribution. For example, the ratio of summary length to content length is significantly higher for paper abstracts than for news articles, as shown in Table 1, indicating a variation in ground truth human contribution. Therefore, in our evaluation, the primary consideration is to verify whether our measure can consistently reflect the overall pattern of varying levels of human contribution for each generation model and creative domain. This consistency would validate the reliability of our proposed measure for distinguishing different levels of human contribution.

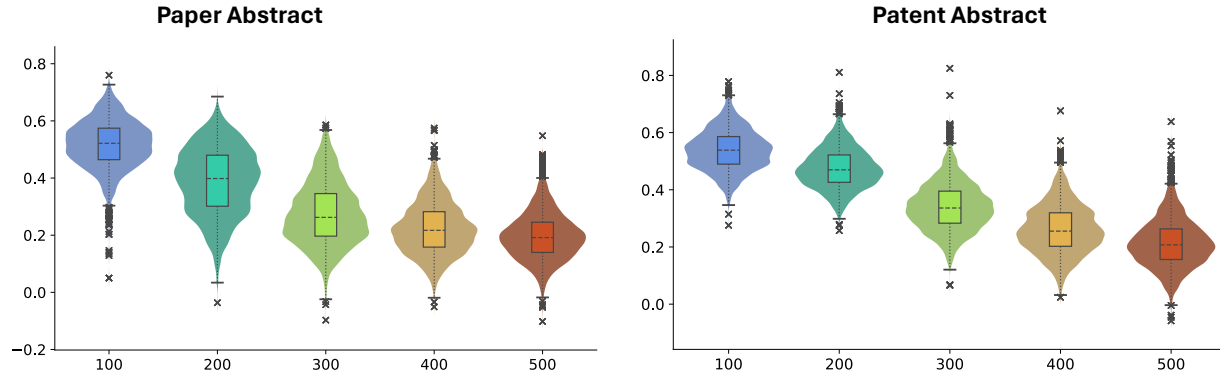


Figure 3. The distribution of the outcomes of the proposed measure for academic paper and patent abstracts of different lengths, generated with titles using Llama-3. The center line represents the median, the box limits indicate the upper and lower quartiles, the whiskers extend to 1.5x the interquartile range, and the points are outliers. Overall, the results align with our expectation that with the same human input information the longer the AI-assisted output, the smaller the measured human contribution.

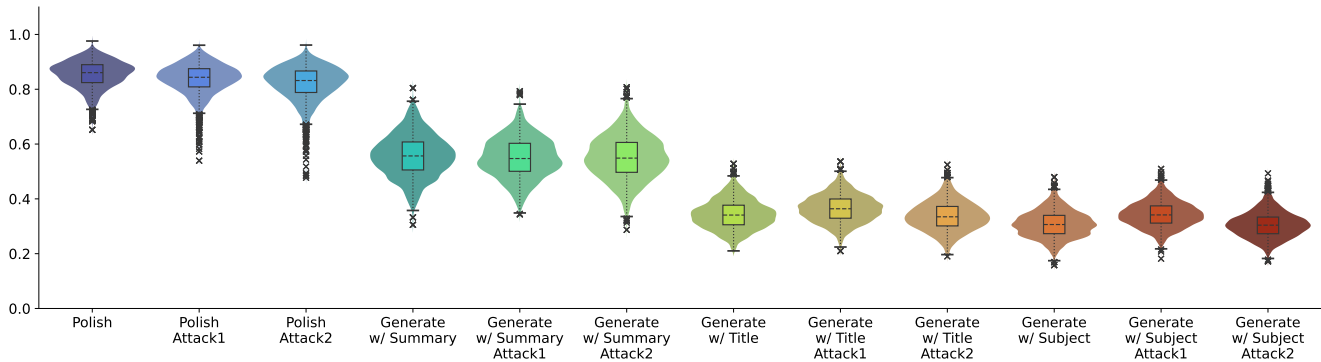


Figure 4. The distribution of the outcomes of the proposed measure for the constructed dataset of news with and without adaptive attacks using Llama-3. The center line represents the median, the box limits indicate the upper and lower quartiles, the whiskers extend to 1.5x the interquartile range, and the points are outliers. Overall, the adaptive attacks have little to no influence on the measurement outcomes.

Impact of Content Length

In addition to the varying levels of human contribution present in our constructed dataset, we further validate our method by varying the length of the AI-assisted content that is generated. This helps us determine whether our method adequately evaluates the proportion of human contribution in AI-assisted generated content when the same human input information yields AI-assisted outputs of different lengths. Intuitively, when human informational input remains constant, the longer the generated content, the smaller the measured human contribution should be. To verify this, we use Llama-3 to generate AI-assisted outputs of varying lengths from titles by specifying the length of the AI-assisted output in the prompt.

Figure 3 shows the results for paper and patent abstracts with different lengths. The results for news and poems are included in Supplementary Materials Section 2.1. The results align with our initial expectation: as we require AI-assisted output to be longer, human informational contribution relative to total informational content decreases, as does our measurement of human contribution.

Resilience to Adaptive Attacks

We further investigate whether adaptive attacks could be employed in real-world applications to artificially inflate measured human contribution. To this end, we design two adaptive attacks: we separately append two instructions to the original input that do not provide additional information but do guide the AI’s generation process to potentially increase the measured human contribution. These instructions are to 1) always choose words you rarely use and 2) mimic human writing. The first instruction

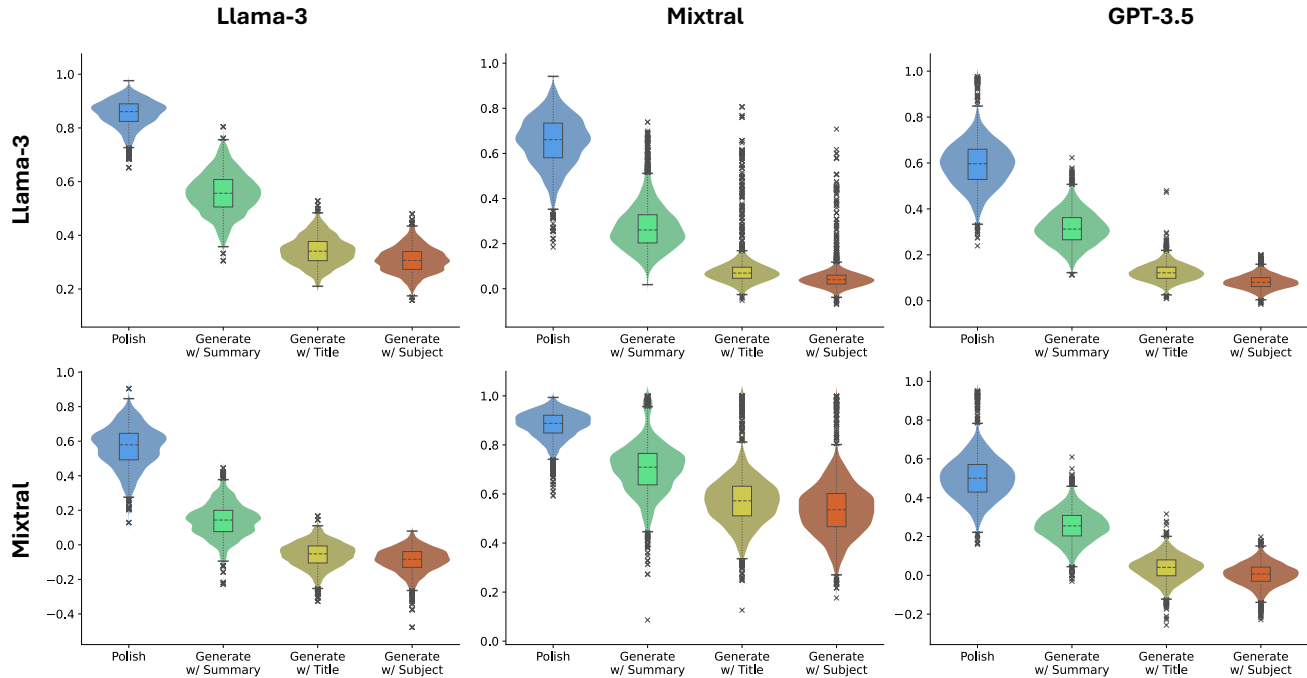


Figure 5. The distribution of the outcomes of the proposed measure for the constructed dataset of news articles for various generation models (columns) and surrogate models for measurement (rows). The center line represents the median, the box limits indicate the upper and lower quartiles, the whiskers extend to 1.5x the interquartile range, and the points are outliers. Overall, for each model pair, the proposed measure exhibits the expected pattern that lower measured values are obtained for the generated content with less human contribution.

influences the model’s generation probabilities to produce less frequently used words, thereby attempting to increase the perceived information content (surprisal). The second instruction guides the model to generate text that closely resembles human writing, thereby attempting to increase the perceived human contribution.

Figure 4 shows the results of our measure with and without attacks using Llama-3 in the news domain. The results for other domains are included in Supplementary Materials Section 2.2. We can observe that our measure remains robust against the adaptive attacks. This aligns with our expectation, as we measure human contribution by utilizing the ratio of the mutual information between human input and AI-assisted output to self-information of the AI-assisted output itself. These non-informational instructions for manipulating the output do not significantly affect our measure.

Generalization of Our Method

In real-world applications, the AI model’s generative probability p_{θ} may not be available. For instance, generative applications like ChatGPT do release generative probabilities to users. This section demonstrates whether a surrogate model with generative probability p'_{θ} can be employed for our assessment when the AI model’s generative probability p_{θ} is unknown. Specifically, in this experiment, we use Llama-3 and Mistral as the surrogate models and use their generative probability p'_{θ} to assess the content generated by various LLMs (Llama-3, Mistral, and ChatGPT) in the constructed dataset.

Figure 5 illustrates the effectiveness of our approach in the news domain for various combinations of surrogate model (rows) and generation model (columns). Results for other domains are presented in Supplementary Materials Section 2.3. We observe that even without using the original AI model for evaluation, our proposed measure captures the expected trend in human contribution across various surrogate and generative model combinations. This validates the applicability of our measure when generation model information is unavailable. This effectiveness may be attributed to the similar generative distributions across LLMs, stemming from the universal knowledge they share during training. The gradient in human contribution across varying levels of human input is far more pronounced than the differences between the distributions themselves, indicating that our method is a robust assessment tool.

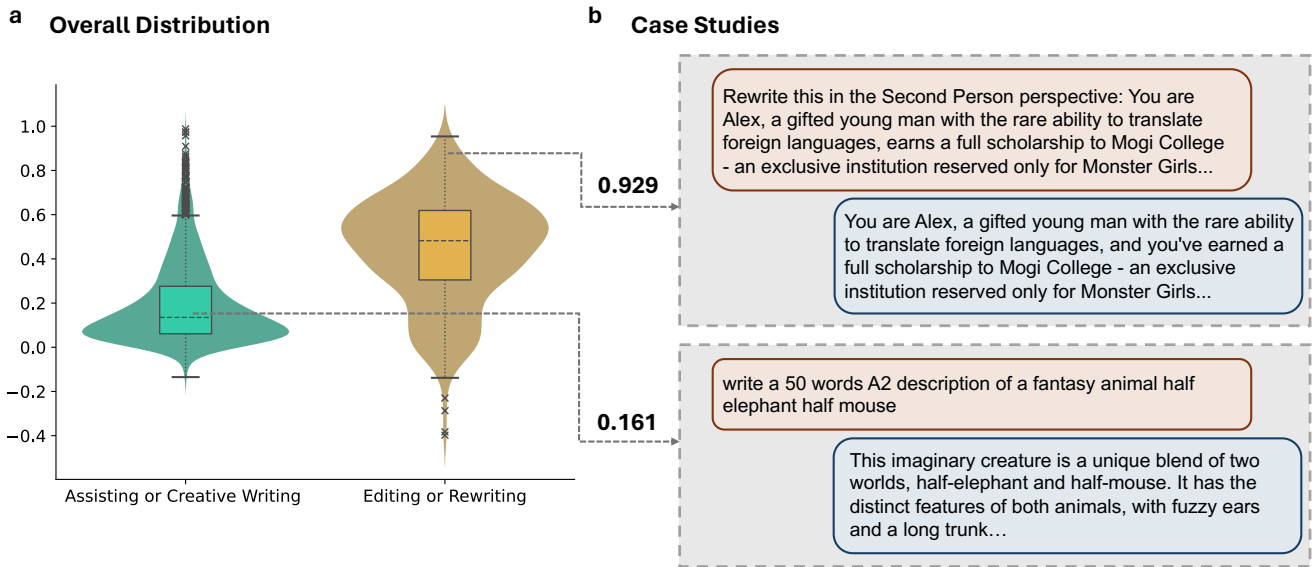


Figure 6. a. The distribution of the outcomes of the proposed measure across various classes of real-world AI-assisted generation, sampled from WildChat. The center line represents the median, the box limits indicate the upper and lower quartiles, the whiskers extend to 1.5x the interquartile range, and the points are outliers. The “editing or rewriting” class typically exhibits a higher degree of measured human contribution. **b.** Case studies on data samples from two distinct classes in WildChat.

Applications to Real-World AI-Assisted Generation

The aforementioned experiments were conducted on a synthetic dataset with known varying levels of human contribution, allowing us to verify the reliability of our measurement method. To test its real-world applicability, we apply our method to real-world scenarios involving user interactions with LLMs. Specifically, we sample cases from the WildChat-1M dataset²⁹ and classify them using a prompt classification tool³⁰. We then sample data from two prompt classes related to AI-assisted generation: “assisting or creative writing” (2,000 entries) and “editing or rewriting” (500 entries), according to their counts in the dataset. The evaluation surrogate model is Llama-3, while the contents were generated with ChatGPT.

Figure 6a demonstrates the overall distribution of measured human contributions across the two classes. We expect that the “editing or rewriting” class will involve more human contribution than “assisting or creative writing.” Consistent with this, the measured human contributions are generally higher for “editing or rewriting.” We present two specific cases in Figure 6b, with additional cases detailed in Supplementary Materials Section 2.4. Overall, the measured human contributions align with our expectations. For instance, the “editing or rewriting” case is measured as having 92.86% human contribution, while the “assisting or creative writing” case is measured at 16.14%. These distribution and case study results further support the validity of our method in measuring human contribution in real-world AI-assisted generation contexts.

Discussion

Technological advancements in generative AI have significantly altered the content production process, resulting in the generation of a vast amount of AI-assisted content^{8,10,11}. This proliferation of AI-assisted generated content poses challenges to the delineation of originality of these works, sparking intense debates regarding the application of AI in areas that prioritize originality, such as education, academic research, and copyright^{23–25,31–33}. It has been increasingly recognized that defining the intellectual originality of AI-assisted generated content cannot be approached with a one-size-fits-all solution; instead, it requires consideration of the extent of human contribution to the work^{20,22,32}. Consequently, to facilitate the generative AI era, decision makers will need credible methods to measure human contribution in AI-assisted generation across various contexts.

This study frames this challenge and presents a method to measure human contribution to AI-assisted generation that is grounded in information theory. We propose using information content to quantify the percentage of information in the AI-assisted output that is attributable to human input. By measuring the ratio of mutual information between human input and AI-assisted output to the self-information of AI-assisted output, we quantify the human contribution in AI-assisted generation. Our measure is validated through experiments conducted on multi-domain AI-assisted generation datasets using multiple LLMs. In raising this research question and proposing a new framework, we seek to measure human contribution quantitatively, inspire further research, and help advance the refinement of relevant originality delineation and content regulation in the future.

While it serves to frame the question and provide a preliminary method for measuring human contribution, our work has several limitations. First, we currently focus on scenarios where human input is available and reliable for evaluation. Further research is necessary to measure human contributions in scenarios where human input is unknown. Second, the current framework focuses on textual output from LLMs. However, originality issues related to AI-assisted generation are not limited to text; they also extend to images, audio, video, computer code, etc. Incorporating non-textual output raises even more complex problems due to the change in modality between human input and AI-assisted output. We aim to explore originality issues in such scenarios in future research. Third, this research does not include experiments which entailing multiple rounds of human-AI interaction, prompting and content generation. In the more complex situation where content is generated with multiple prompting rounds, human contribution should be measured by considering the human’s intent and understanding of the generated content, which may inform the contribution assessment. Fourth, caution is warranted when measuring human contribution in AI-assisted content generation in the copyright domain. Human edition, selection, and compilation of AI-assisted content may provide significant creative input, which could be relevant when assessing authorship for copyright purposes.

Ethical and Societal Impact

The objective of this study is to pose a research question and propose a framework for measuring human contribution in AI-assisted content generation. This question and framework aim to facilitate originality delineation in the era of creation with the assistance of AI. Simultaneously, this work seeks to inspire more research on technical methods that can support the enhancement of relevant regulations in the context of widespread AI utilization in various scenarios. A potential risk is that in real-world applications of the proposed framework, there might be targeted adaptive attacks aimed at manipulating the results to artificially elevate the assessed level of human contribution. Although this paper examines two adaptive attacks and verifies the robustness of the proposed measure against them, more sophisticated and advanced attacks may arise in real-world scenarios. We hope to further understand and mitigate such risks in future work.

The authors of this paper introduce a mere method to technically measure the human contribution in AI-assist content generation which can be potentially used in various scenarios. However, the paper does not intend to discuss the complex copyright legal and policy issues related to “originality” or “eligibility,” nor it reflect any of Microsoft’s legal and policy positions on the copyright issues.

Methods

Related Work

The research problem most closely related to evaluating human contribution is the detection of content generated by LLMs^{34–36}. As the performance of LLMs continues to improve, the risk of being unable to distinguish between content generated by LLMs and humans becomes increasingly apparent, with attendant threats in security, fraud prevention^{37,38}, and academic integrity³⁹, among other fields^{34,40,41}. Consequently, researchers are increasingly directing their efforts towards the detection of LLM-generated content, specifically ascertaining whether a given text is primarily the product of AI. These research efforts entail training detection models^{42–44}, employing features for zero-shot detection^{36,45,46}, or incorporating specific watermarks during content generation^{35,47,48}.

While the current body of research predominantly focuses on identifying content substantially generated by AI, thus optimized for binary detection, real-world AI-assisted generation often involves varying degrees of human contribution. In many practical application scenarios, it is not sufficient to merely detect content primarily generated by AI, rather it is crucial to discern the extent of human contribution. Therefore, distinct from the detection of AI-generated content, our emphasis is on reliably quantifying human contribution within AI-assisted generation from an informational perspective.

Defining Human Contribution in AI-Assisted Generation

In contrast to a binary classifier determining whether content is primarily generated by AI, our aim is to derive a quantitative measurement indicating the extent of human contribution in AI-assisted content generation. This is a novel and previously unexplored issue. Our core idea revolves around utilizing *information content* as a medium for gauging the contributions of humans and AI. Particularly, we define human contribution in AI-assisted generation as the ratio of mutual information between human input and AI-assisted output relative to the total self-information of the AI-assisted output, as illustrated in Figure 1b.

In this section, we first introduce related concepts derived from information theory²⁷; we then provide our definition of human contribution. In the following definition, we consider an AI model M_θ , its generative distribution p_θ , human input \mathbf{x} , and AI-assisted output \mathbf{y} . First, we quantify the information content within the generated output \mathbf{y} through the concept of *self-information*. Self-information measures the level of surprisal associated with the outcome of a random variable, which is related to the probability of that outcome occurring. In this context, content that is less probable in its generation is deemed

more informative. We represent the self-information of the generated output \mathbf{y} as follows:

$$I(\mathbf{y}) = -\log(p_{\theta}(\mathbf{y})), \quad (1)$$

where $p_{\theta}(\mathbf{y})$ is the probability that the content \mathbf{y} is generated without any condition.

On the other hand, when conditioned on human input \mathbf{x} , the information content within the generated output \mathbf{y} transforms into *conditional self-information*. Conditional self-information quantifies the information contained in an outcome of a random variable, given the occurrence of another event. Here, we represent the conditional self-information of the generated output \mathbf{y} given the human input \mathbf{x} as follows:

$$I(\mathbf{y}|\mathbf{x}) = -\log(p_{\theta}(\mathbf{y}|\mathbf{x})), \quad (2)$$

where $p_{\theta}(\mathbf{y})$ is the probability that the content \mathbf{y} is generated conditioned on human input \mathbf{x} .

Based on these two information concepts, we define the *mutual information* between the generated content \mathbf{y} and the human input \mathbf{x} as the information gain during generation when the human input is known. This signifies the reduction in surprisal when human input \mathbf{x} is for generating content \mathbf{y} is provided, defined as follows:

$$I(\mathbf{x};\mathbf{y}) = I(\mathbf{y}) - I(\mathbf{y}|\mathbf{x}). \quad (3)$$

Building upon the aforementioned definition of information within the AI-assisted generation process, we proceed to establishing the definition of human contribution in AI-assisted generation.

Definition 1 (Human contribution in AI-assisted generation). *Given an AI model M_{θ} and human input \mathbf{x} , where \mathbf{y} represents the AI-assisted generated content, the human contribution ϕ is defined as the ratio of mutual information $I(\mathbf{x};\mathbf{y})$ to self-information $I(\mathbf{y})$.*

This definition of human contribution pertains to the proportion of the information content within the generated output that can be attributed to human input, relative to the total information content of the generated output.

Data Availability

The Arxiv dataset is available at <https://huggingface.co/datasets/gfissore/arxiv-abstracts-2021>. The News Articles dataset is available at <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/GMFCTR>. The HUPD dataset is available at <https://huggingface.co/datasets/HUPD/hupd/blob/main/data/2018.tar.gz>. The Poetry Foundation dataset is available at <https://www.kaggle.com/datasets/tgdivy/poetry-foundation-poems>. The WildChat-1M dataset is available at <https://huggingface.co/datasets/allenai/WildChat-1M>.

Code Availability

The code applied in the experiments is publicly available at <https://github.com/xyq7/Human-Contribution-Measurement>.

References

1. Touvron, H. *et al.* Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
2. OpenAI. Chatgpt. <https://openai.com/blog/chatgpt> (2022).
3. Chowdhery, A. *et al.* Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.* **24**, 240:1–240:113 (2023). URL <http://jmlr.org/papers/v24/22-1144.html>.
4. OpenAI. Gpt-4 system card. <https://cdn.openai.com/papers/gpt-4-system-card.pdf> (2023).
5. Anthropic. Model card and evaluations for claude models (2023). URL <https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf>.
6. Wu, Z. *et al.* Ai creativity and the human-ai co-creation model. In *Human-Computer Interaction. Theory, Methods and Tools: Thematic Area, HCI 2021, Held as Part of the 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings, Part I 23*, 171–190 (Springer, 2021).
7. Wingström, R., Hautala, J. & Lundman, R. Redefining creativity in the era of ai? perspectives of computer scientists and new media artists. *Creat. Res. J.* **36**, 177–193 (2024).

8. Dergaa, I., Chamari, K., Zmijewski, P. & Saad, H. B. From human writing to artificial intelligence generated text: examining the prospects and potential threats of chatgpt in academic writing. *Biol. Sport* **40**, 615–622 (2023).
9. Microsoft. Introducing microsoft 365 copilot – your copilot for work. <https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/> (2023).
10. Zheng, C., Wang, D., Wang, A. Y. & Ma, X. Telling stories from computational notebooks: Ai-assisted presentation slides creation for presenting data science work. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–20 (2022).
11. Zhang, R. *et al.* Youling: an ai-assisted lyrics creation system. *arXiv preprint arXiv:2201.06724* (2022).
12. Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K. & Chen, L. Generative ai and chatgpt: Applications, challenges, and ai-human collaboration (2023).
13. Wang, D. *et al.* From human-human collaboration to human-ai collaboration: Designing ai systems that can work together with people. In *Extended abstracts of the 2020 CHI conference on human factors in computing systems*, 1–6 (2020).
14. Hemmer, P. *et al.* Human-ai collaboration: The effect of ai delegation on human task performance and task satisfaction. In *Proceedings of the 28th International Conference on Intelligent User Interfaces, IUI '23*, 453–463 (Association for Computing Machinery, New York, NY, USA, 2023). URL <https://doi.org/10.1145/3581641.3584052>. DOI 10.1145/3581641.3584052.
15. Gatto, T. G. J. Generative ai and copyright – some recent denials and unanswered questions. <https://www.intellectualpropertylawblog.com/archives/generative-ai-and-copyright-some-recent-denials-and-unanswered-questions/> (2023).
16. Hutson, J. Rethinking plagiarism in the era of generative ai. *J. Intell. Commun.* **4**, 20–31 (2024).
17. Yu, H. Reflection on whether chat gpt should be banned by academia from the perspective of education and teaching. *Front. Psychol.* **14**, 1181712 (2023).
18. Nakadai, R., Nakawake, Y. & Shibasaki, S. Ai language tools risk scientific diversity and innovation. *Nat. Hum. Behav.* **7**, 1804–1805 (2023).
19. Kwon, D. Ai is complicating plagiarism. how should scientists respond? *Nat.* (2024).
20. U.S. Copyright Office. Copyright registration guidance: Works containing material generated by artificial intelligence (2023). URL <https://www.federalregister.gov/documents/2023/03/16/2023-05321/copyright-registration-guidance-works-containing-material-generated-by-artificial-intelli>
21. of America, V. Schools ban chatgpt amid fears of artificial intelligence-assisted cheating. *VOA News* (2023). URL <https://www.voanews.com/a/schools-ban-chatgpt-amid-fears-of-artificial-intelligence-assisted-cheating-/6958125.html>. Accessed: 2024-06-27.
22. Singer, N. Ban or embrace? colleges wrestle with a.i.-generated admissions essays. *The New York Times* (2023). URL <https://www.nytimes.com/2023/09/01/business/college-admissions-essay-ai-chatbots.html>. Accessed: 2024-06-27.
23. U.S. Copyright Office Review Board. Decision affirming refusal of registration of a recent entrance to paradise (2022). URL <https://www.copyright.gov/rulings-filings/review-board/docs/a-recent-entrance-to-paradise.pdf>. At 2.
24. U.S. Copyright Office. Cancellation decision re: Zarya of the dawn (vau001480196) (2023). URL <https://www.copyright.gov/docs/zarya-of-the-dawn.pdf>. At 2.
25. Abbott, R. & Rothman, E. Disrupting creativity: Copyright law in the age of generative artificial intelligence. *Fla. L. Rev.* **75**, 1141 (2023).
26. Hristov, K. Artificial intelligence and the copyright dilemma. *Idea* **57**, 431 (2016).
27. Shannon, C. E. A mathematical theory of communication. *The Bell system technical journal* **27**, 379–423 (1948).
28. Clark, K. *et al.* Mixtral of experts. *arXiv preprint arXiv:2401.04088* (2023).
29. Zhao, W. *et al.* Wildchat: 1m chatgpt interaction logs in the wild. *arXiv preprint arXiv:2405.01470* (2024).
30. Valentina Pyatkin. Prompt classification (2024). URL <https://huggingface.co/valpy/prompt-classification>.

31. Chesterman, S. Good models borrow, great models steal: intellectual property rights and generative ai. *Policy Soc.* puae006 (2024).
32. Fenwick, M. & Jurcys, P. Originality and the future of copyright in an age of generative ai. *Comput. Law & Secur. Rev.* **51**, 105892 (2023).
33. Smits, J. & Borghuis, T. Generative ai and intellectual property rights. In *Law and artificial intelligence: regulating AI and applying ai in legal practice*, 323–344 (Springer, 2022).
34. Yang, X. *et al.* A survey on detection of llms-generated content. *arXiv preprint arXiv:2310.15654* (2023).
35. Kirchenbauer, J. *et al.* A watermark for large language models. *arXiv preprint arXiv:2301.10226* (2023).
36. Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D. & Finn, C. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *arXiv preprint arXiv:2301.11305* (2023).
37. Pan, Y. *et al.* On the risk of misinformation pollution with large language models. *arXiv preprint arXiv:2305.13661* (2023).
38. Xie, Y. *et al.* Defending chatgpt against jailbreak attack via self-reminders. *Nat. Mach. Intell.* **5**, 1486–1496 (2023).
39. Bin-Nashwan, S. A., Sadallah, M. & Bouteraa, M. Use of chatgpt in academia: Academic integrity hangs in the balance. *Technol. Soc.* **75**, 102370 (2023).
40. Kumar, S., Balachandran, V., Njoo, L., Anastasopoulos, A. & Tsvetkov, Y. Language generation models can cause harm: So what can we do about it? an actionable survey. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 3299–3321 (Association for Computational Linguistics, Dubrovnik, Croatia, 2023). URL <https://aclanthology.org/2023.eacl-main.241>.
41. Yang, X. *et al.* Shadow alignment: The ease of subverting safely-aligned language models. *arXiv preprint arXiv:2310.02949* (2023).
42. Tian, E. Gptzero: An ai text detector (2023). URL <https://gptzero.me/>.
43. OpenAI. AI text classifier (2023). URL <https://beta.openai.com/ai-text-classifier>.
44. Zhan, H., He, X., Xu, Q., Wu, Y. & Stenetorp, P. G3detector: General gpt-generated text detector. *arXiv preprint arXiv:2305.12680* (2023).
45. Lavergne, T., Urvoy, T. & Yvon, F. Detecting fake content with relative entropy scoring. *PAN* **8**, 27–31 (2008).
46. Yang, X., Cheng, W., Petzold, L., Wang, W. Y. & Chen, H. Dna-gpt: Divergent n-gram analysis for training-free detection of gpt-generated text. *arXiv preprint arXiv:2305.17359* (2023).
47. Zhao, X., Ananth, P. V., Li, L. & Wang, Y.-X. Provable robust watermarking for ai-generated text. *ArXiv abs/2306.17439* (2023).
48. Hou, A. B. *et al.* Semstamp: A semantic watermark with paraphrastic robustness for text generation. *arXiv preprint arXiv:2310.03991* (2023).

Author Contributions

Y.X. conceived the idea of this work, implemented the models, analyzed the results, and contributed to the writing of this manuscript. T.Q. conceived the idea of this work, implemented the models, analyzed the results, and contributed to the writing of this manuscript. J.Y. conceived the idea of this work, implemented the models, analyzed the results, and contributed to the writing of this manuscript. R.W. contributed to the writing of this manuscript. J.H. contributed to the writing of this manuscript. Q.D. contributed to the writing of this manuscript. Y.X. contributed to the writing of this manuscript and coordinated the research project. X.X. contributed to the writing of this manuscript and coordinated the research project. F.W. conceived the idea of this work, analyzed the results, contributed to the writing of this manuscript, and coordinated the research project.

Additional Information

Supplementary Information accompanies this manuscript in the attached supplementary information file.

Competing Interests: The authors declare no competing interests.