

Measuring Human Contribution in AI-Assisted Content Generation

Yueqi Xie^{1*†}, Tao Qi^{2*†}, Jingwei Yi^{3*†}, Xiyuan Yang⁴, Ryan Whalen⁵,
Junming Huang¹, Qian Ding⁶, Yu Xie^{1,7}, Xing Xie⁶, Fangzhao Wu^{6†}

¹Princeton University, ²Tsinghua University,

³University of Science and Technology of China, ⁴University of Illinois Urbana-Champaign,

⁵The University of Hong Kong, ⁶Microsoft Research Asia, ⁷Peking University

Correspondence: yueqixie@princeton.edu, taoqi.qt@gmail.com, jwyi1029@gmail.com, fangzhu@microsoft.com

Abstract

With the growing prevalence of generative 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. To further enhance real-world applicability, we extend the framework to estimate the minimal necessary human contribution for any text without requiring human input and validate its effectiveness. We hope that this work lays a foundation for measuring human contributions in AI-assisted content generation in the era of generative AI.

1 Introduction

Recent advances in large language models (LLMs) have impacted our personal and working lives in significant ways, notably by changing the process of content generation (Wingström et al., 2024). AI “copilots” have emerged as a new and powerful content production tool across a variety of domains, such as lyrics creation (Zhang et al., 2022), office work (Microsoft, 2023; Zheng et al., 2022), academic writing (Dergaa et al., 2023), 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 (Fui-Hoon Nah et al., 2023; Wang et al., 2020; Hemmer et al., 2023). 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 (Garbett and Gatto, 2023; Fui-Hoon Nah et al., 2023). 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 education (Hutson, 2024), academic research (Yu, 2023; Nakadai et al., 2023; Kwon, 2024), and creative work (U.S. Copyright Office, 2023b). 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 (of America, 2023; Singer, 2023). Similarly, there is a growing debate, underscored by notable incidents (U.S. Copyright Office Review Board, 2022; U.S. Copyright Office, 2023a), concerning the copyright eligibility of AI-assisted works (U.S. Copyright Office, 2023b; Abbott and Rothman, 2023; Hristov, 2016).

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*

*Joint First Authors

†Corresponding Authors

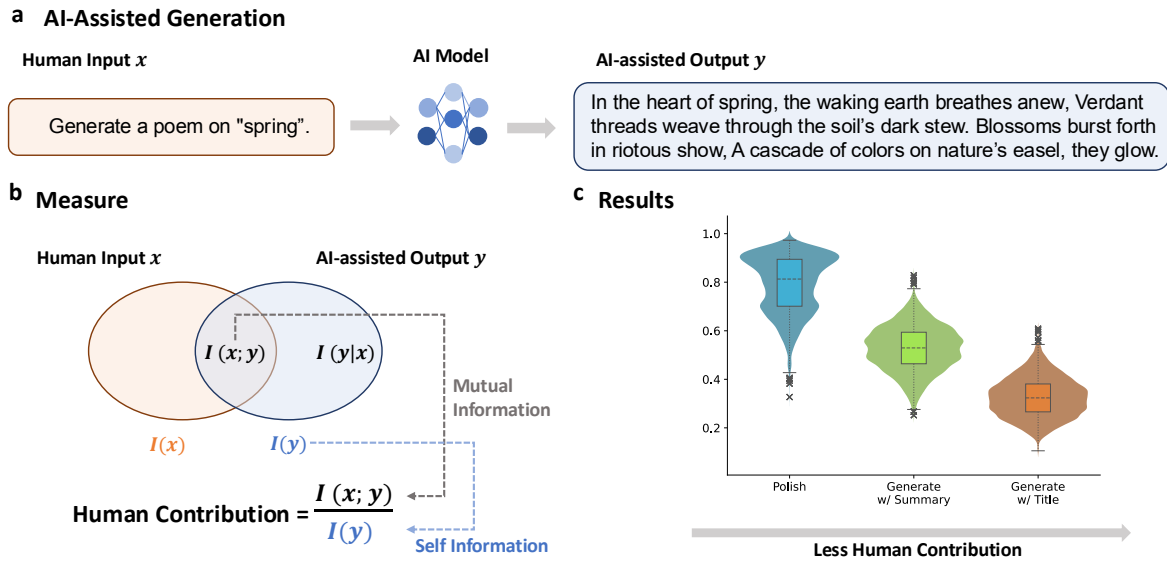


Figure 1: **a.** Illustration of AI-assisted content generation, where an AI model is prompted with human input and generates output. **b.** Overview of the proposed method for measuring human contribution, quantified by the ratio of mutual information between human input and AI-assisted output to the 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).

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 theory (Shannon, 1948), 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*, $I(y)$. The numerator, $I(x; y)$, is the portion of self-information $I(y)$ that is shared

with the total information content from human input, $I(x)$, which we define as *mutual information*. The difference between the two is the *conditional self-information* in AI-assisted output given user input, $I(y|x)$, 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, writing style, model temperature, 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.

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	-

To broaden the applicability of our human contribution evaluation framework to real-world scenarios, where humans may engage in complex content generation processes or where human input is unknown, we propose an extension that estimates human contribution based on any text alone, without requiring access to its generation process. The contributions are summarized as follows:

- We formulate a novel research question aimed at quantitatively evaluating informational human contribution in AI-assisted generation.
- We propose simple yet effective information-theoretic measures to approach this problem.
- We construct a dataset of AI-assisted generation with varying levels of human contribution for rigorous evaluation.
- We systematically evaluate the proposed methods on this dataset, examining overall effectiveness, the impact of content length, writing style, model temperature, resilience to adaptive attacks, and generalization performance.
- We extend the framework to estimate the human contribution of any text alone, without requiring access to the generation process.

2 Related Work

The research problem most closely related to evaluating human contribution is the detection of content generated by LLMs (Yang et al., 2023b; Wu et al., 2025). 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 (Pan et al., 2023; Xie et al., 2023), and academic integrity (Bin-Nashwan et al., 2023), among other fields (Yang et al., 2023b; Kumar et al., 2023; Yang et al., 2023c). 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 (OpenAI, 2023; Zhan et al., 2023; Hao et al., 2025), employing features for zero-shot detection (Lavergne et al., 2008; Mitchell et al., 2023; Yang et al., 2023a; Hans et al., 2024), or incorporating specific watermarks during content generation (Kirchenbauer et al., 2023; Zhao et al., 2023; Hou et al., 2023).

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. 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.

3 Dataset Construction

To verify the validity and reliability of the proposed measure of human contribution, we construct a dataset of AI-assisted generation data with known varying levels of human contribution. By design, our dataset spans a very large range of human contributions 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 section A.1.

We further process the data into a uniform structure with decreasing levels of information: *content*, *summary*, *title*, and *subject* (except poems, because of less-informative titles), with missing parts of the raw data supplemented using GPT-3.5. 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 section A.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-weight LLMs Llama-3 (Touvron et al., 2023) and Mixtral (Clark et al., 2023) and the proprietary models GPT-3.5 (OpenAI, 2022), GPT-4o, Gemini-2.0 Flash (Team et al., 2023), and Claude-3.5 Haiku (Anthropic, 2023). The specific API versions of the models are detailed in section A.3. We generate five times for a human input with the temperature set as 0.7 for diverse outputs.

4 Methods

4.1 Defining Human Contribution in AI-Assisted Generation

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 theory (Shannon, 1948); 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 an event, reflecting the probability of that event occurring. In this context,

generating content that is less probable is considered more informative event. 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 event, given the occurrence of another event. Here, we represent the conditional self-information of the generated output \mathbf{y} given “human input is \mathbf{x} ” as follows:

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

where $p_\theta(\mathbf{y} | \mathbf{x})$ is the probability that the content \mathbf{y} is generated conditioned on human input \mathbf{x} . Specifically, considering the generative process of the autoregressive LLM M_θ , these probabilities can be computed by multiplying the probabilities of each token y_i being sampled from the generative distribution, conditioned on the previously generated tokens (and the human input \mathbf{x} for conditional probability with human input):

$$p_\theta(\mathbf{y}) = \prod_{i=1}^N p_\theta(y_i | \mathbf{y}_{<i}), \quad (3)$$

$$p_\theta(\mathbf{y} | \mathbf{x}) = \prod_{i=1}^N p_\theta(y_i | \mathbf{y}_{<i}, \mathbf{x}), \quad (4)$$

where N is the total number of tokens in \mathbf{y} .

Based on these two information concepts, we define *mutual information* between the generated content \mathbf{y} and the human input \mathbf{x} as the reduction in surprisal when human input \mathbf{x} is known for generating content \mathbf{y} . This indicates how much information content in \mathbf{y} can be attributed to human input \mathbf{x} . Formally, it is defined as:

$$I(\mathbf{x}; \mathbf{y}) = I(\mathbf{y}) - I(\mathbf{y} | \mathbf{x}). \quad (5)$$

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) *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 $I(\mathbf{y})$: $\phi = \frac{I(\mathbf{x}; \mathbf{y})}{I(\mathbf{y})}$.*

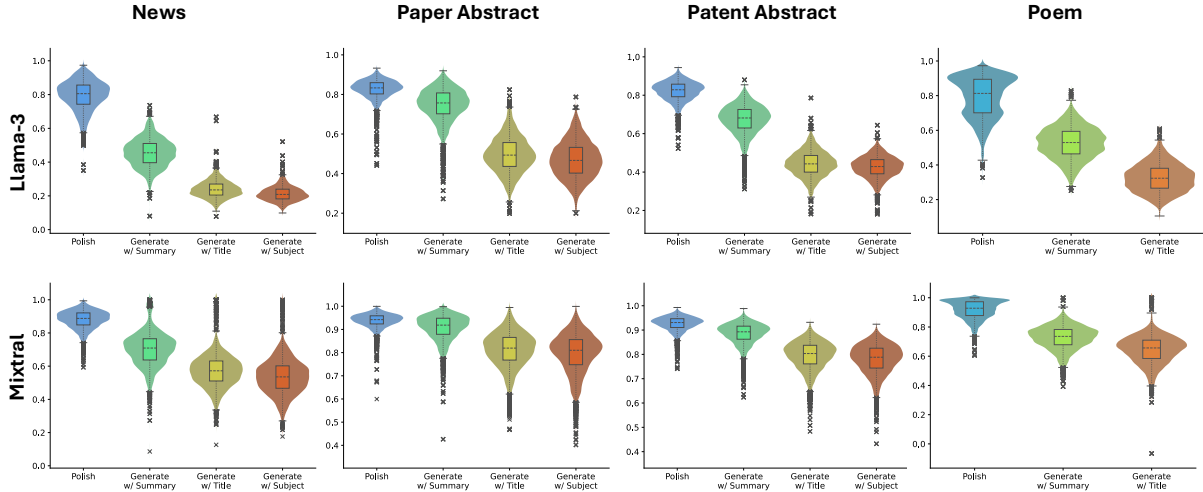


Figure 2: The distribution of the outcomes of the **proposed measure** for the constructed dataset. Overall, the proposed measure exhibits the expected trend that lower values are obtained for the generated content with less human contribution.

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. This definition can be directly employed for *originality authentication*, where the human author can provide both the AI model’s generative distribution and the original human input (certified with appropriate documentation) 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.

4.2 Estimating Human Contribution Without Human Input

To better address real-world scenarios where the AI-assisted generation process is unknown, we extend our method to estimate the human contribution of any given text alone without knowledge of human input and output generation processes. We have less information now than the situation considered before. This loss of information leads to a degree of uncertainty, as the same output can be generated from inputs with varying levels of contribution. Therefore, we redefine our goal to estimate the proportion of information content in the AI-assisted output that is *necessarily* provided by humans. This is also referred to as the *minimum human contribution* for human input x to plausibly generate the AI-assisted output y . This focus on minimal human contribution stems from sev-

eral practical considerations, such as evaluating the unique original informativeness value that LLMs cannot easily generate, as detailed in section B.

To compute this, we adapt the original formula $\frac{I(x;y)}{I(y)}$. While the denominator $I(y)$ can still be directly calculated, the numerator $I(x;y) = I(y) - I(y|x)$ is estimated as its minimum possible value for a valid hypothetical x that plausibly generates y . Intuitively, if a hypothetical x plausibly generates y , the likelihood of y being generated based on x , $p_\theta(y | x)$, should be sufficiently high (have a lower bound). This implies that the conditional information (or surprisal) $I(y|x) = -\log(p_\theta(y | x))$ of generating y given x should have an upper bound. We leverage this upper bound to compute the minimum possible value of $I(x;y)$. Due to limited space, the detailed motivation for estimating minimum human contribution, along with the calculation method, is provided in section B.

5 Experiments

5.1 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 y and human input x are known, and the AI model M_θ ’s output probability is available in evaluation.

Overall Trend Analysis: Figure 2 illustrates the human contribution results of two open-weight 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. 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). These differences are reasonable because the same generation mode does not necessarily equate to a similar percentage of human contribution across different domains. For example, in our dataset, the News category consists of full articles with detailed and vivid descriptions, while Paper Abstracts are typically concise paragraph. However, the summaries, titles, and subjects used for generation in our dataset for both categories are of comparable information content. Consequently, when generating a news article versus a paper abstract from a summary, title, or subject, the expected human contribution distributions differ due to the inherent difference in the information content of the data. 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.

Human-annotation Dataset: To further validate the reliability and consistency of our quantitative analysis, we construct a dataset with human annotations. As discussed earlier, obtaining a dataset with clear, real-valued labels remains a significant challenge, as there is no established method for quantifying human contribution when both human input and output are known. Inspired by evaluation platforms such as Chatbot Arena (Chiang et al., 2024), we collect human annotation data by pairwise comparisons. Specifically, we sample 1,500 pairs of input-output data points generated by Llama-3 across all four domains and categories

with the condition that the measured human contribution gap between the two points exceeds 0.1 using our method. This threshold is chosen to ensure the meaningfulness of the comparisons and to reduce ambiguity during annotation. Pairs of input-output data points are so annotated by three human annotators as to reflect which one contains a higher degree of human contribution. The annotation process is detailed in section A.4. Using this annotated dataset, we evaluate how well the relative magnitudes derived from our measurements align with human judgments in these comparisons. The results indicate that **95.93%** of the data exhibit consistency between our measurements and human annotations. Moreover, as shown in Figure 8, the inconsistent pairs, i.e., cases where the human annotation disagrees with the measurement’s ranking, are predominantly concentrated in cases where the measured human contribution gap between two data points is less than 0.2. This further lends support to the robustness and validity of our measurement, demonstrating its consistency across different domains and setups.

Due to space limitations, we defer the analysis of **the impact of content length (C.1)**, **model temperature (C.2)**, and **resilience to adaptive attacks (C.3)** to the Appendix.

5.2 Impact of Writing Style

Note that our proposed framework is based on token-level probability calculations. As a result, semantically similar content may yield different measured values of human contribution due to variations in writing style. In this section, we conduct an ablation study to examine the impact of writing style on the outcomes of our measurement using Llama-3. Specifically, we explore four stylistic variations within the news domain by appending the following prompts: “Generate in an informal style.”, “Generate in a formal style.”, “Generate in a vivid story-telling style.”, and “Generate in an objective news reporting style.” These prompts are applied while keeping the original input content fixed, thereby inducing markedly different wording styles without altering other human contributed information. This setup allows us to assess the extent to which semantically-preserving stylistic variation influences the outcomes of our measurement.

The results are shown in Figure 3. The results demonstrate that the proposed measurement robustly captures differences in human informational contribution, even when writing style varies. This

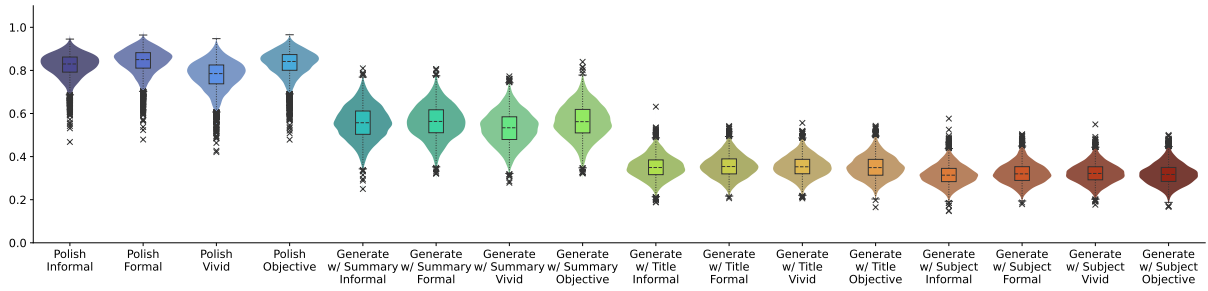


Figure 3: The distribution of the outcomes of the proposed measure for the constructed dataset of news in different writing styles using Llama-3. Overall, the writing styles have little influence on the measurement outcomes.

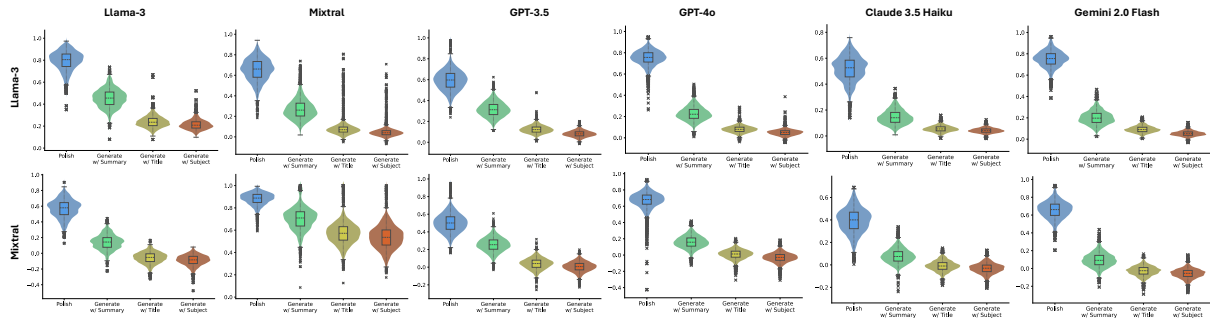


Figure 4: The distribution of the outcomes of the proposed measure for the constructed dataset of news articles for various generation models (columns) and surrogate models for measurement (rows). Overall, for each model pair, the proposed measure exhibits the expected pattern that lower measured values are obtained for the generated content with less human contribution.

is consistent with our expectation, as the variance introduced by such semantically-preserving stylistic fluctuations is relatively limited compared to the differences driven by variations in the amount of information provided in the input.

5.3 Generalization of Our Method

In real-world applications, the AI model’s generative probability p_θ may not be available. For instance, applications like ChatGPT don’t 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, GPT-3.5, GPT4o, Claude 3.5 Haiku, and Gemini 2.0 Flash) in the constructed dataset. Note that generation models can include chatbots or models that do not release their generative probabilities, while surrogate models require access to generative probability distributions for evaluation.

Figure 4 illustrates the effectiveness of our approach in the news domain for various combina-

tions of surrogate model (rows) and generation model (columns). Results for other domains are presented in section C.4. 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 similarity in 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.

5.4 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 in-

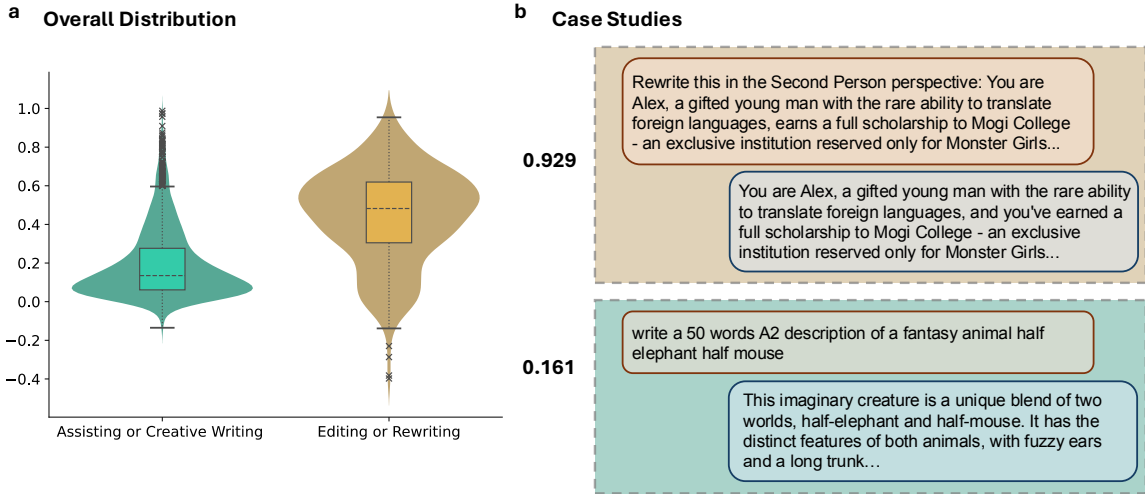


Figure 5: **a.** The distribution of the outcomes of the proposed measure across various classes of real-world AI-assisted generation, sampled from WildChat. The “editing or rewriting” class typically exhibits a higher degree of measured human contribution. **b.** Case studies on data samples from two distinct classes in WildChat.

teractions with LLMs. Specifically, we sample cases from the WildChat-1M dataset (Zhao et al., 2024) and classify them using a prompt classification tool (Valentina Pyatkin, 2024). 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 5a 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 for two classes respectively in Figure 5b. 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.

5.5 Human Contribution Estimation Without Human Input

In Figure 6, we present the human contribution estimation results for two state-of-the-art open-source LLMs, Llama-3 and Mixtral, applied to the con-

structed dataset across various domains. We make the following observations: First, the proposed estimation method produces results that follow the expected trend in our dataset: as the human author’s informational contribution decreases during AI-assisted generation, the estimated metric value correspondingly decreases. Second, as we suppress the diverse real values of $I(\mathbf{y} | \mathbf{x})$ to an estimated boundary value that is empirically chosen, the resulting distributions do not fully match human contribution evaluations where \mathbf{x} is known (Figure 2). Finally, since this is an estimation process of a lower bound, there are cases where the estimated values fall below zero. In our original framework, under the defined setting where both the human input and the generative distribution are known, the value range is generally between 0 and 1. However, due to the nature of lower-bound estimation, some values may fall below zero. These negative estimates can be interpreted as cases where the corresponding outputs could plausibly have been generated with minimal or negligible human contribution, under the minimal contribution assumption.

In addition, we use the annotated dataset same as section 5.1 to evaluate how well the relative magnitudes derived from our estimates align with human judgments in these comparisons. The results indicate that **92.87%** of the samples exhibit consistency between our estimates and human annotations, further supporting the validity of our estimates.

We further demonstrate the effectiveness of the

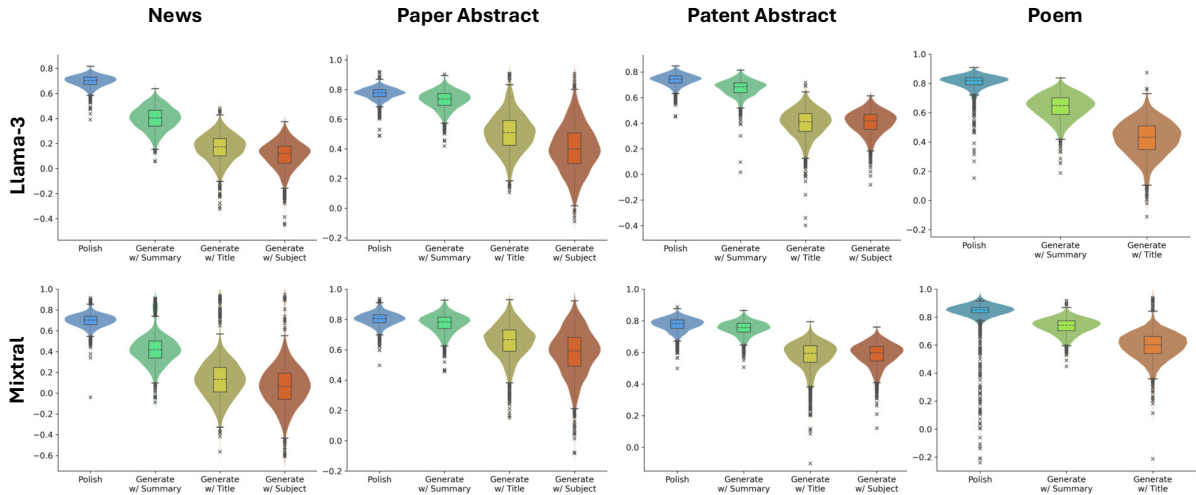


Figure 6: The distribution of the outcomes of the **proposed estimation** for the constructed dataset. Overall, the proposed estimation exhibits the expected trend that lower values are obtained for the generated content with less human contribution.

proposed estimation method in **multi-round generation scenarios** in section C.5 and **AI-generated content detection** in section C.6.

Overall, our method provides a meaningful solution for estimating human contribution in situations where direct input information and generation processes are unavailable.

6 Conclusion

This study formally frames the challenge of quantifying human contribution in AI-assisted generation and introduces a principled, information-theoretic method as a credible measure. We validate the proposed measure through extensive experiments on multi-domain AI-assisted generation datasets using multiple LLMs. To enhance real-world applicability, we further extend the framework to scenarios in which the original human input is unavailable and or the underlying generative distribution is inaccessible. By articulating this research question and proposing a principled measurement framework, this work aims to advance future efforts in originality delineation and the regulation of AI-assisted content in generative AI era.

Limitations

While it serves to frame the question and provide a preliminary method for measuring human contribution, our work has several limitations. First, 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. In other modalities, since data may no longer be discrete, this framework may not be directly applicable. However, as generative models are fundamentally grounded in certain generative probabilistic distributions, we hope that the information theory-based framework proposed in this work can serve as a foundation and inspiration for addressing originality challenges across diverse modalities. Second, 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. Finally, we acknowledge that “human contribution” can have different meanings. In this work, we focus specifically on measuring informational contribution, while leaving other interpretations, such as prompt-engineering effectiveness, for future exploration. Also, informational contribution should not be directly equated with creativity or originality. Our framework is designed to quantify the informational contribution contained in the input, rather than to provide a normative judgment about creative value or authorship.

Ethical Considerations

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 contexts. 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 method to technically measure human contribution in AI-assist content generation. The method is widely applicable in different situations. 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.

References

- Ryan Abbott and Elizabeth Rothman. 2023. Disrupting creativity: Copyright law in the age of generative artificial intelligence. *Fla. L. Rev.*, 75:1141.
- Anthropic. 2023. [Model card and evaluations for claude models](#).
- Saeed Awadh Bin-Nashwan, Mouad Sadallah, and Mohamed Bouteraa. 2023. Use of chatgpt in academia: Academic integrity hangs in the balance. *Technology in Society*, 75:102370.
- Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolos Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, and 1 others. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.
- Kevin Clark, Deepak Narayanan, Jiseung Kwon, Aaron Harlap, Matei Zaharia, and Christopher Re. 2023. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.
- Ismail Dergaa, Karim Chamari, Piotr Zmijewski, and Helmi Ben Saad. 2023. From human writing to artificial intelligence generated text: examining the prospects and potential threats of chatgpt in academic writing. *Biology of Sport*, 40(2):615–622.
- Fiona Fui-Hoon Nah, Ruilin Zheng, Jingyuan Cai, Keng Siau, and Langtao Chen. 2023. Generative ai and chatgpt: Applications, challenges, and ai-human collaboration.
- Tiana Garbett and James Gatto. 2023. 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/>.
- Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.
- Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting llms with binoculars: Zero-shot detection of machine-generated text. *arXiv preprint arXiv:2401.12070*.
- Wei Hao, Ran Li, Weiliang Zhao, Junfeng Yang, and Chengzhi Mao. 2025. [Learning to rewrite: Generalized LLM-generated text detection](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6421–6434, Vienna, Austria. Association for Computational Linguistics.
- Patrick Hemmer, Monika Westphal, Max Schemmer, Sebastian Vetter, Michael Vössing, and Gerhard Satzger. 2023. [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*, page 453–463, New York, NY, USA. Association for Computing Machinery.
- Abe Bohan Hou, Jingyu Zhang, Tianxing He, Yichen Wang, Yung-Sung Chuang, Hongwei Wang, Lingfeng Shen, Benjamin Van Durme, Daniel Khashabi, and Yulia Tsvetkov. 2023. Semstamp: A semantic watermark with paraphrastic robustness for text generation. *arXiv preprint arXiv:2310.03991*.
- Kalin Hristov. 2016. Artificial intelligence and the copyright dilemma. *Idea*, 57:431.
- James Hutson. 2024. Rethinking plagiarism in the era of generative ai. *Journal of Intelligent Communication*, 4(1):20–31.
- John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. *arXiv preprint arXiv:2301.10226*.

- Sachin Kumar, Vidhisha Balachandran, Lucille Njoo, Antonios Anastasopoulos, and Yulia Tsvetkov. 2023. [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*, pages 3299–3321, Dubrovnik, Croatia. Association for Computational Linguistics.
- Diana Kwon. 2024. Ai is complicating plagiarism. how should scientists respond? *Nature*.
- Thomas Lavergne, Tanguy Urvoy, and François Yvon. 2008. Detecting fake content with relative entropy scoring. *PAN*, 8:27–31.
- Microsoft. 2023. Introducing microsoft 365 copilot – your copilot for work. <https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/>.
- Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *arXiv preprint arXiv:2301.11305*.
- Ryosuke Nakadai, Yo Nakawake, and Shota Shibasaki. 2023. Ai language tools risk scientific diversity and innovation. *Nature Human Behaviour*, 7(11):1804–1805.
- Voice of America. 2023. [Schools ban chatgpt amid fears of artificial intelligence-assisted cheating](#). *VOA News*. Accessed: 2024-06-27.
- OpenAI. 2022. Chatgpt. <https://openai.com/blog/chatgpt>.
- OpenAI. 2023. [AI text classifier](#).
- Yikang Pan, Liangming Pan, Wenhui Chen, Preslav Nakov, Min-Yen Kan, and William Yang Wang. 2023. On the risk of misinformation pollution with large language models. *arXiv preprint arXiv:2305.13661*.
- Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.
- Natasha Singer. 2023. [Ban or embrace? colleges wrestle with a.i.-generated admissions essays](#). *The New York Times*. Accessed: 2024-06-27.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- U.S. Copyright Office. 2023a. [Cancellation decision re: Zarya of the dawn \(vau001480196\)](#). At 2.
- U.S. Copyright Office. 2023b. [Copyright registration guidance: Works containing material generated by artificial intelligence](#).
- U.S. Copyright Office Review Board. 2022. [Decision affirming refusal of registration of a recent entrance to paradise](#). At 2.
- Valentina Pyatkin. 2024. [Prompt classification](#).
- Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. 2024. Ghostbuster: Detecting text ghostwritten by large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1702–1717.
- Dakuo Wang, Elizabeth Churchill, Pattie Maes, Xi-angmin Fan, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. 2020. 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*, pages 1–6.
- Roosa Wingström, Johanna Hautala, and Riina Lundman. 2024. Redefining creativity in the era of ai? perspectives of computer scientists and new media artists. *Creativity Research Journal*, 36(2):177–193.
- Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. 2025. A survey on llm-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, 51(1):275–338.
- Yueqi Xie, Jingwei Yi, Jiawei Shao, Justin Curl, Lingjuan Lyu, Qifeng Chen, Xing Xie, and Fangzhao Wu. 2023. Defending chatgpt against jailbreak attack via self-reminders. *Nature Machine Intelligence*, 5(12):1486–1496.
- Xianjun Yang, Wei Cheng, Linda Petzold, William Yang Wang, and Haifeng Chen. 2023a. Dna-gpt: Divergent n-gram analysis for training-free detection of gpt-generated text. *arXiv preprint arXiv:2305.17359*.
- Xianjun Yang, Liangming Pan, Xuandong Zhao, Haifeng Chen, Linda Petzold, William Yang Wang, and Wei Cheng. 2023b. A survey on detection of llms-generated content. *arXiv preprint arXiv:2310.15654*.
- Xianjun Yang, Xiao Wang, Qi Zhang, Linda Petzold, William Yang Wang, Xun Zhao, and Dahua Lin. 2023c. Shadow alignment: The ease of subverting safely-aligned language models. *arXiv preprint arXiv:2310.02949*.
- Hao Yu. 2023. Reflection on whether chat gpt should be banned by academia from the perspective of education and teaching. *Frontiers in Psychology*, 14:1181712.

- Peipeng Yu, Jiahao Chen, Xuan Feng, and Zhihua Xia. 2025. Cheat: A large-scale dataset for detecting chatgpt-written abstracts. *IEEE Transactions on Big Data*, 11(3):898–906.
- Haolan Zhan, Xuanli He, Qionkai Xu, Yuxiang Wu, and Pontus Stenetorp. 2023. G3detector: General gpt-generated text detector. *arXiv preprint arXiv:2305.12680*.
- Rongsheng Zhang, Xiaoxi Mao, Le Li, Lin Jiang, Lin Chen, Zhiwei Hu, Yadong Xi, Changjie Fan, and Minlie Huang. 2022. Youling: an ai-assisted lyrics creation system. *arXiv preprint arXiv:2201.06724*.
- Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. Wildchat: 1m chatgpt interaction logs in the wild. *arXiv preprint arXiv:2405.01470*.
- Xuandong Zhao, Prabhanjan Vijendra Ananth, Lei Li, and Yu-Xiang Wang. 2023. Provable robust watermarking for ai-generated text. *ArXiv*, abs/2306.17439.
- Chengbo Zheng, Dakuo Wang, April Yi Wang, and Xiaojuan Ma. 2022. 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*, pages 1–20.

Contents

1	Introduction	1
2	Related Work	3
3	Dataset Construction	3
4	Methods	4
4.1	Defining Human Contribution in AI-Assisted Generation	4
4.2	Estimating Human Contribution Without Human Input	5
5	Experiments	5
5.1	Human Contribution Evaluation	5
5.2	Impact of Writing Style	6
5.3	Generalization of Our Method	7
5.4	Applications to Real-World AI-Assisted Generation	7
5.5	Human Contribution Estimation Without Human Input	8
6	Conclusion	9
A	Additional Experimental Setups	14
A.1	Raw Data Collection and Processing	14
A.1.1	Raw Data Collection	14
A.1.2	Raw Data Processing	14
A.2	Prompt Construction	14
A.2.1	News	14
A.2.2	Paper Abstract	14
A.2.3	Patent Abstract	15
A.2.4	Poem	15
A.3	Model Setups	15
A.4	Human Annotation	15
A.5	Hyperparameter Setups	15
B	Method: Estimating Human Contribution in AI-Assisted Generation Without Human Input	17
C	Additional Experimental Results	18
C.1	Impact of Content Length	18
C.2	Impact of Generative Model Temperature	19
C.3	Resilience to Adaptive Attacks	19
C.4	Generalization of Our Method	20
C.5	Multi-Round Generation Analysis	20
C.6	AI-generated Content Detection	23

A Additional Experimental Setups

A.1 Raw Data Collection and Processing

A.1.1 Raw Data Collection

This section details the collection and processing of the raw data.

- **Paper Abstract:** The paper abstract data entries are randomly sampled from the *arxiv-abstracts-2021* dataset¹, containing the original paper abstracts as content, titles as title, and primary categories as subject. The summaries are generated using GPT-3.5 using the prompt specified in Section A.1.2.
- **News:** The news data entries are sampled from the *News Articles* dataset², containing the original news articles as content and titles as title. To avoid outliers, the entries are randomly sampled with a content length constraint of 400 to 700 words. The summaries and subjects are generated using GPT-3.5 with the prompt specified in Section A.1.2.
- **Patent Abstract:** The patent abstract data entries are sampled from the *HUPD 2018* dataset³, containing the original patent abstracts as content and titles as title. To avoid outliers, the entries are randomly sampled with a content length constraint of 150 to 250 words. The summaries and subjects are generated using GPT-3.5 with the prompt specified in Section A.1.2.
- **Poem:** The poem data entries are sampled from the *Poetry Foundation* dataset⁴, containing the original poems as content and titles as titles. To avoid outliers, the entries are randomly sampled with a content length constraint of 150 to 250 words. The summaries are generated using GPT-3.5 with the prompt specified in Section A.1.2. Since the titles already contain a small amount of information, averaging 3.65 words, we do not generate additional subjects.

¹<https://huggingface.co/datasets/gfissore/arxiv-abstracts-2021>

²<https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/GMFCTR>

³<https://huggingface.co/datasets/HUPD/hupd/blob/main/data/2018.tar.gz>

⁴<https://www.kaggle.com/datasets/tgdivy/poetry-foundation-poems>

A.1.2 Raw Data Processing

For the missing information, including summaries and subjects, we used GPT-3.5 (*gpt-3.5-turbo-1106*) to supplement them, thereby facilitating AI-assisted generation with multiple levels of human contribution. The prompts are specified as follows:

- **Summary Generation:** You are a helpful assistant. Help me summarize the following content in a few sentences as concise as possible: *{Content}*.
- **Subject Generation:** You are a helpful assistant. Help me generate the subject of the following content in two to four words: *{Content}*.

A.2 Prompt Construction

We detail the prompt construct for AI-assisted generation with different levels of human contribution as follows, where N is the original length of the content in the dataset:

A.2.1 News

- **Polish:** Help me polish the following news article: *{News Content}*. Limit your response to $\{N\}$ words. Start with "News:".
- **Generate with Summary:** Generate a news article with the following summary: *{News Summary}*. Limit your response to $\{N\}$ words. Start with "News:".
- **Generate with Title:** Generate a news article with the title: *{News Title}*. Limit your response to $\{N\}$ words. Start with "News:".
- **Generate with Subject:** Generate a news article with the subject: *{News Subject}*. Limit your response to $\{N\}$ words. Start with "News:".

A.2.2 Paper Abstract

- **Polish:** Help me polish the following paper abstract: *{Paper Abstract}*. Limit your response to $\{N\}$ words. Start with "Abstract:".
- **Generate with Summary:** Generate a paper abstract with the following summary: *{Paper Summary}*. Limit your response to $\{N\}$ words. Start with "Abstract:".
- **Generate with Title:** Generate a paper abstract with the title: *{Paper Title}*. Limit your response to $\{N\}$ words. Start with "Abstract:".

- **Generate with Subject:** Generate a paper abstract with the subject: $\{Paper\ Subject\}$. Limit your response to $\{N\}$ words. Start with "Abstract:".

A.2.3 Patent Abstract

- **Polish:** Help me polish the following patent abstract: $\{Patent\ Abstract\}$. Limit your response to $\{N\}$ words. Start with "Abstract:".
- **Generate with Summary:** Generate a patent abstract with the following summary: $\{Patent\ Summary\}$. Limit your response to $\{N\}$ words. Start with "Abstract:".
- **Generate with Title:** Generate a patent abstract with the title: $\{Patent\ Title\}$. Limit your response to $\{N\}$ words. Start with "Abstract:".
- **Generate with Subject:** Generate a patent abstract with the subject: $\{Patent\ Subject\}$. Limit your response to $\{N\}$ words. Start with "Abstract:".

A.2.4 Poem

- **Polish:** Help me polish the following poem: $\{Poem\}$. Limit your response to $\{N\}$ words. Start with "Poem:".
- **Generate with Summary:** Generate a poem with the following summary: $\{Poem\ Summary\}$. Limit your response to $\{N\}$ words. Start with "Poem:".
- **Generate with Title:** Generate a poem with the title: $\{Poem\ Title\}$. Limit your response to $\{N\}$ words. Start with "Poem:".

A.3 Model Setups

In this section, we detail the versions and parameters of the models we apply in the experiments as follows.

API Settings: For evaluation, we use Llama-3 (Llama-3.1-8B-Instruct) and Mixtral (Mixtral-8x7B-v0.1). The generation models further include GPT-3.5 (gpt-3.5-turbo-0125), GPT-4o (gpt-4o-2024-08-06), Gemini-2.0 Flash (gemini-2.0-flash), and Claude-3.5 Haiku (claude-3-5-haiku-20241022).

Model Selection Rationale: The rationale for the choice of models in our experiments is as follows. In the main experiments (Figure 2 and 6),

we employed two state-of-the-art open-source language models: Llama-3 and Mixtral. In the experiment on generalization of our method (Figure 4), we explored various combinations of generation models and surrogate testing models. For the ablation and additional exploratory experiments, we used Llama-3 as the default model to ensure consistency across different settings.

A.4 Human Annotation

For the human-annotated dataset, we select 1,500 pairs of input-output data points generated by Llama-3 across all four domains and categories, ensuring that the measured human contribution gap between the two points exceeds 0.1. This threshold is chosen to enhance the meaningfulness of the comparisons and minimize ambiguity during evaluation.

We conducted the annotation using the Labelbox platform. Three platform labeling service annotators, proficient English speakers experienced in expert-level image and text labeling, were recruited for two weeks (80 hours). Each of the 1,500 data points was independently annotated by all three annotators. The final label for each instance was determined by majority voting. The inter-annotator agreement rate—defined as the proportion of cases where all three annotators provided the same label—is 91.47%. Figure 7 illustrates the annotation interface used in this process.

Figure 8 shows the distribution of the measured human contribution gaps, as determined by our proposed method, for data consistent and inconsistent with human annotations. The results indicate that most of the inconsistent cases are associated with smaller measured gaps, further demonstrating the alignment of our measurement with human annotations.

A.5 Hyperparameter Setups

In the estimation framework, we define a threshold τ , which serves as a lower bound for the geometric mean of the conditional probabilities $p_\theta(\mathbf{y} | \mathbf{x})^{1/N}$ if input \mathbf{x} could plausibly generate output \mathbf{y} using model M_θ . Correspondingly, $-\log \tau$ serves as the upper bound for the mean conditional information (or surprisal) $-\log p_\theta(\mathbf{y} | \mathbf{x})/N$ of generating \mathbf{y} given \mathbf{x} . There are several considerations for selecting the hyperparameter threshold τ used in the estimation framework.

First, the threshold τ is inherently tied to the generative distribution of the model in use. For

Instruction for Annotation:

Please select the option that reflects a **higher** human contribution.

Human contribution is measured as the **ratio** of information provided by **human input** relative to the total information of the **AI-assisted output**.

Option A	Option B
<p>Human Input</p> <p>Generate a poem with the following summary: The content describes the emotional turmoil of a mother witnessing her son's dying moments. It paints a picture of a body as a mere husk and lends a sense of agony as the mother grapples with the impending loss. The imagery of a plastic bracelet and the son's eyes being compared to oysters add depth to the portrayal of sorrow. The narrative then shifts to the aftermath, with the mention of soiled sheets, a biohazard bin, and the burning of the sheets creating a grim atmosphere. The content ends with a plea for release and a desire to let the birds in..</p>	<p>Human Input</p> <p>Generate an abstract for the patent with the following summary: The disclosed intravascular stent is woven from at least one weaving wire into a tubular structure. The weaving wire has an even number of free ends, which are aligned and jointed together with cutting portions. This creates a seamless connection, uniform distribution of radial and axial force, and allows for convenient laser welding. The stent also offers great flexibility, operability, and a perfect appearance..</p>
<p>AI-assisted Output</p> <p>Poem: As I hold your frail, lifeless form, A shell of the boy I once had sworn, A plastic bracelet still wrapped around, A reminder of childhood's unfurled bond. Your eyes, once bright as oysters' shine, Now dim, like the stars that had aligned, In my darkest hour, I'm left to define, The anguish that this moment brings to mine. In this cold, clinical space, I'm lost, Where soiled sheets and biohazard bins are crossed, The scent of disinfectant fills the air, As I'm forced to confront the final farewell there. Oh, how I yearn to let the birds in, To sweep away the pain and let love begin, To release the grief, to let go of the pain, And find solace in the memories we'll never regain.</p>	<p>AI-assisted Output</p> <p>Abstract: A novel intravascular stent is disclosed, comprising a tubular structure woven from at least one weaving wire. The weaving wire has an even number of free ends, which are aligned and jointed together with cutting portions to create a seamless connection. This unique design enables uniform distribution of radial and axial force, allowing for improved flexibility and operability of the stent. The cutting portions facilitate convenient laser welding, ensuring a strong and secure bond between the wire ends. The stent's seamless construction and uniform distribution of force also provide a perfect appearance, minimizing the risk of tissue irritation or other adverse reactions. The disclosed stent offers a superior alternative for vascular applications, offering improved performance, durability, and patient outcomes.</p>

Figure 7: Illustration of the annotation interface.

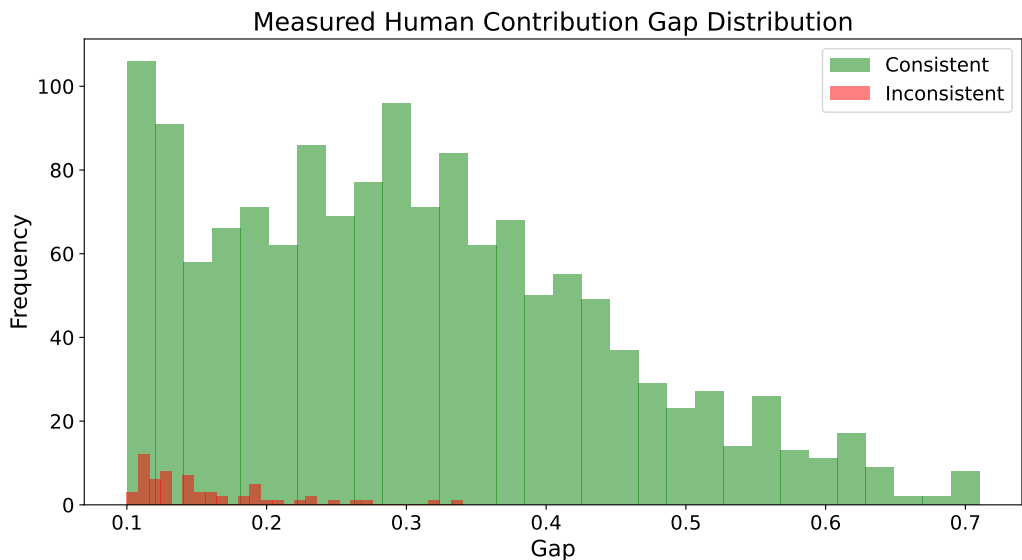


Figure 8: Distribution of measured human contribution gaps for data consistent and inconsistent with human annotations.

models with highly concentrated generative distributions—that is, those that assign high probability to a small subset of plausible outputs—a relatively high threshold τ for the geometric mean of the conditional probabilities $(p_\theta(\mathbf{y} | \mathbf{x}))^{1/N}$ is appropriate, which corresponds to a relatively low threshold $-\log \tau$ for the mean conditional information $-\log p_\theta(\mathbf{y} | \mathbf{x})/N$. This is because, in such models, plausible generations are expected to have relatively high conditional probabilities and thus lower conditional information values. Conversely, for models with more diffuse or uniform generative distributions, plausible outputs may be assigned lower probabilities on average. In such cases, a lower threshold τ (i.e., a higher threshold $-\log \tau$) is more appropriate.

Second, ideally, we would calibrate τ using a large amount of empirical data from the target domain and model, consisting of paired examples where human input \mathbf{x} generates AI-assisted output \mathbf{y} . This would allow us to observe the distribution of the geometric mean conditional probabilities (or equivalently, the average conditional information), and to set threshold τ accordingly to reflect a meaningful boundary between plausible and implausible generations within the testing domain. However, in practice, such domain-specific calibration may increase the barrier to use. Therefore, we provide a general, model-specific reference value for τ that serves as a reasonable default for real-world applications. Users can still adjust τ based on their specific domain or sensitivity requirements.

Third, since it is infeasible to obtain the full generative space of a model for selecting a general threshold value, we approximate it using samples from WildChat, a real-world AI-assisted generation dataset that also serves as out-of-distribution (OOD) data in the evaluation on our constructed dataset. Specifically, we use the set of 2,500 prompts as those sampled in the Applications to Real-World AI-Assisted Generation section, generate five outputs for each prompt using the target models (Llama-3 and Mixtral), and compute the distribution of average conditional information for each model. We then set $-\log \tau$ to the empirical mean plus one standard deviation of this distribution, representing a reasonable upper bound for the mean conditional information $-\log p_\theta(\mathbf{y} | \mathbf{x})/N$ on the plausible generation range. This results in $-\log \tau = 0.5496$ for Llama-3 and 0.4155 for Mixtral in our experiments. This approach provides a general, model-specific reference value for the

threshold, serving as a practical default for real-world applications. Users may still adjust it based on domain-specific requirements or sensitivity preferences.

B Method: Estimating Human Contribution in AI-Assisted Generation Without Human Input

In many scenarios requiring the evaluation of human contribution, the generation process of AI-assisted output \mathbf{y} can be complex (e.g., involving iterative back-and-forth refinements) or entirely unavailable. To address this challenge, we extend our framework by introducing the human contribution estimation problem and a corresponding solution that is applicable to any text output, without requiring prior knowledge of the human input \mathbf{x} or the generation process.

Problem Definition: First, we define a research problem aimed at evaluating human contribution without reliance on \mathbf{x} . It is noteworthy that the same output \mathbf{y} can result from various human inputs \mathbf{x} , which are associated with differing levels of human contribution. For instance, an AI-generated abstract based solely on a subject (representing minimal human contribution) could also be produced through repetition, where human contribution approaches 100%. To establish a well-defined problem, we frame the research challenge as determining the *minimum human contribution* required for human input \mathbf{x} to plausibly generate AI-assisted output \mathbf{y} . Specifically, this involves quantifying the proportion of information content within the AI-assisted output that is *necessarily* provided by humans.

Our focus on minimal human contribution stems from several practical considerations: First, all outputs can theoretically be generated from a prompt with nearly 100% human contribution, such as prompting LLMs for repetition. Estimating minimal human contribution provides a reference range, emphasizing the lower bound of human-contributed information content required to plausibly generate the output. Second, in the context of widespread LLM adoption, minimal human contribution evaluates the *unique original informativeness value* provided by humans—content that LLMs cannot easily generate. This distinction is critical in various applications, such as curating high-value data for training next-generation models, assessing creative or original works that reflect

human ingenuity and effort, and identifying unique contributions in academic peer reviews.

Definition 2 (Minimal Human Contribution)

Given an AI model M_θ and the AI-assisted generated content \mathbf{y} , the minimal human contribution $\hat{\phi}$ is defined as the ratio of the minimal mutual information $\hat{I}(\hat{\mathbf{x}}, \mathbf{y})$ —where $\hat{\mathbf{x}}$ represents the minimally informative human input required to plausibly generate \mathbf{y} —to the self-information $I(\mathbf{y})$: $\hat{\phi} = \frac{\hat{I}(\hat{\mathbf{x}}; \mathbf{y})}{I(\mathbf{y})}$.

Solution: To solve this problem, we first need to establish a reasonable criterion for determining whether \mathbf{x} could plausibly generate \mathbf{y} . Considering the generative process of an autoregressive LLM, each token y_i is sampled from a generative distribution $p_\theta(y_i | \mathbf{y}_{<i}, \mathbf{x})$ conditioned on the previous tokens and \mathbf{x} . For \mathbf{y} to be considered plausibly generated by \mathbf{x} , we require that the geometric mean of the conditional probabilities $p_\theta(y_i | \mathbf{y}_{<i}, \mathbf{x})$ across all tokens exceeds a predefined threshold τ :

$$\left(\prod_{i=1}^N p_\theta(y_i | \mathbf{y}_{<i}, \mathbf{x}) \right)^{1/N} > \tau, \quad (6)$$

where N is the number of tokens in \mathbf{y} . This condition ensures that, overall, the likelihood of \mathbf{y} being sampled from the distribution of model M_θ given input \mathbf{x} at each step is sufficiently large. It prevents any token in \mathbf{y} from having an excessively low probability of being sampled, maintaining the credibility of \mathbf{x} as the source of \mathbf{y} . The rationale and empirical values for the threshold τ are provided in section A.5.

Based on this criterion, we can estimate the minimum human contribution $\hat{\phi} = \frac{\hat{I}(\hat{\mathbf{x}}; \mathbf{y})}{I(\mathbf{y})}$. The denominator $I(\mathbf{y})$ can be directly computed; thus, our primary task is to compute a *lower bound* for the mutual information $I(\mathbf{x}, \mathbf{y}) = I(\mathbf{y}) - I(\mathbf{y} | \mathbf{x})$ by deriving an *upper bound* for the conditional self-information $I(\mathbf{y} | \mathbf{x})$.

According to our plausibility criterion, we have:

$$p_\theta(\mathbf{y} | \mathbf{x}) = \prod_{i=1}^N p_\theta(y_i | \mathbf{y}_{<i}, \mathbf{x}) > \tau^N, \quad (7)$$

where N is the total number of tokens in \mathbf{y} , and τ is the predefined plausibility threshold, as detailed in section A.5. From this, the upper bound for the conditional self-information $I(\mathbf{y} | \mathbf{x})$ can be

computed as:

$$I(\mathbf{y} | \mathbf{x}) = -\log p_\theta(\mathbf{y} | \mathbf{x}) < -N \log \tau. \quad (8)$$

This can be interpreted as follows: when \mathbf{x} plausibly generates \mathbf{y} , the likelihood of \mathbf{y} being generated based on \mathbf{x} is sufficiently high, implying that the conditional information (or surprisal) of generating \mathbf{y} given \mathbf{x} should be bounded. From this, we derive the mutual information lower bound:

$$I(\mathbf{x}; \mathbf{y}) \geq I(\mathbf{y}) + N \log \tau. \quad (9)$$

Finally, we can compute the estimated minimal human contribution as:

$$\hat{\phi} = \frac{\hat{I}(\hat{\mathbf{x}}; \mathbf{y})}{I(\mathbf{y})} = \frac{I(\mathbf{y}) + N \log \tau}{I(\mathbf{y})}. \quad (10)$$

This framework enables the estimation of minimal human contribution $\hat{\phi}$ without requiring knowledge of \mathbf{x} . It can be broadly applied to real-world evaluations of human contribution for any given text, providing a practical solution for scenarios where the AI-assisted generation process is unknown.

C Additional Experimental Results

C.1 Impact of Content Length

In addition to the varying levels of human contribution present in our constructed dataset, we further validate our method by manipulating the total information content of the output while controlling the information content of the input. By doing so, we aim to verify whether our measurement exhibits the expected property that, when the input is controlled, increasing the output information leads to a lower human contribution (as defined by the ratio of human-contributed information relative to the total information in the output).

To achieve this, we adjust the output length requirements in the input prompts while keeping the other input information constant. We assume that much longer outputs generally have higher information content. Specifically, we use Llama-3 to **generate AI-assisted outputs of varying lengths from the identical titles** by specifying the desired output length in the prompt.

The results, as shown in Figure 9, align with our expectations: as the AI-assisted output contains more information content, the proportion of human informational contribution relative to the total informational content decreases, as reflected by our measurement of human contribution.

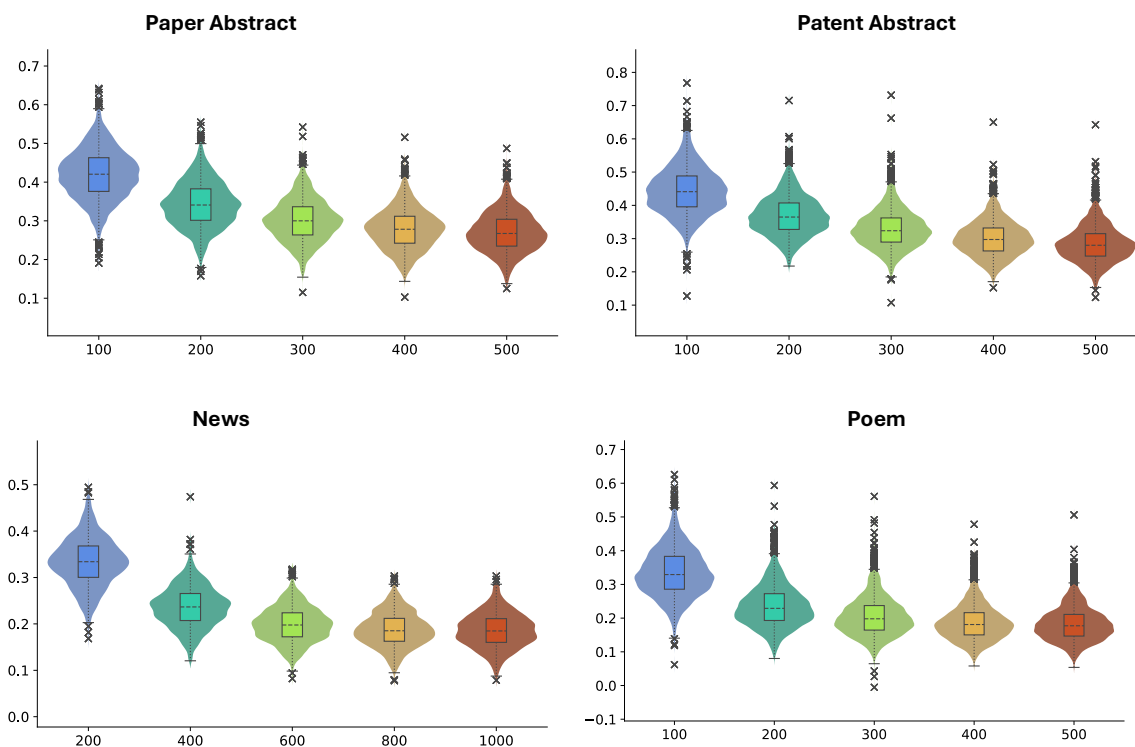


Figure 9: The distribution of the proposed measure for AI-assisted generations of varying lengths, generated with **identical titles** using Llama-3. Overall, the results are consistent with our expectation that, given the same amount of human input information, longer AI-assisted outputs exhibit smaller measured human contributions.

C.2 Impact of Generative Model Temperature

Temperature refers to the parameter that controls the randomness of output in a generative model. During the generation process, a lower temperature makes the model more deterministic by favoring high-probability tokens, while a higher temperature introduces greater variability, resulting in more diverse but less predictable outputs. In this section, we study the impact of generative model temperature on the outcomes of our measurement. While the “human effort” (human input) remains unchanged across different temperature settings, we expect human contribution—defined as the proportion of informational content in the output attributable to human input—to decrease as the temperature increases. This is because the output becomes less dependent on the input information and increasingly reflects the model’s explorative generation. In extreme cases, when the temperature approaches infinity, the output of the language model would become entirely independent of the human input, resulting in an expected human contribution of zero.

Figure 10 demonstrates the comparison between temperature settings from 0.3 to 0.9 during the

generation with Llama-3. We can observe that generally, a higher temperature leads to a smaller measured human contribution across different generation categories. This validates that our measurement accurately reflects the diminishing influence of human input on AI-assisted output with increasing temperature of the generative model.

C.3 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. For each attack, we append an instruction to the original input prompts that does not provide additional information but may influence the AI’s generation process in a way that potentially increases the measured human contribution. Specifically, the two instructions used are: (1) “Always choose words you rarely use.” and (2) “Mimic human writing.” The first instruction 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

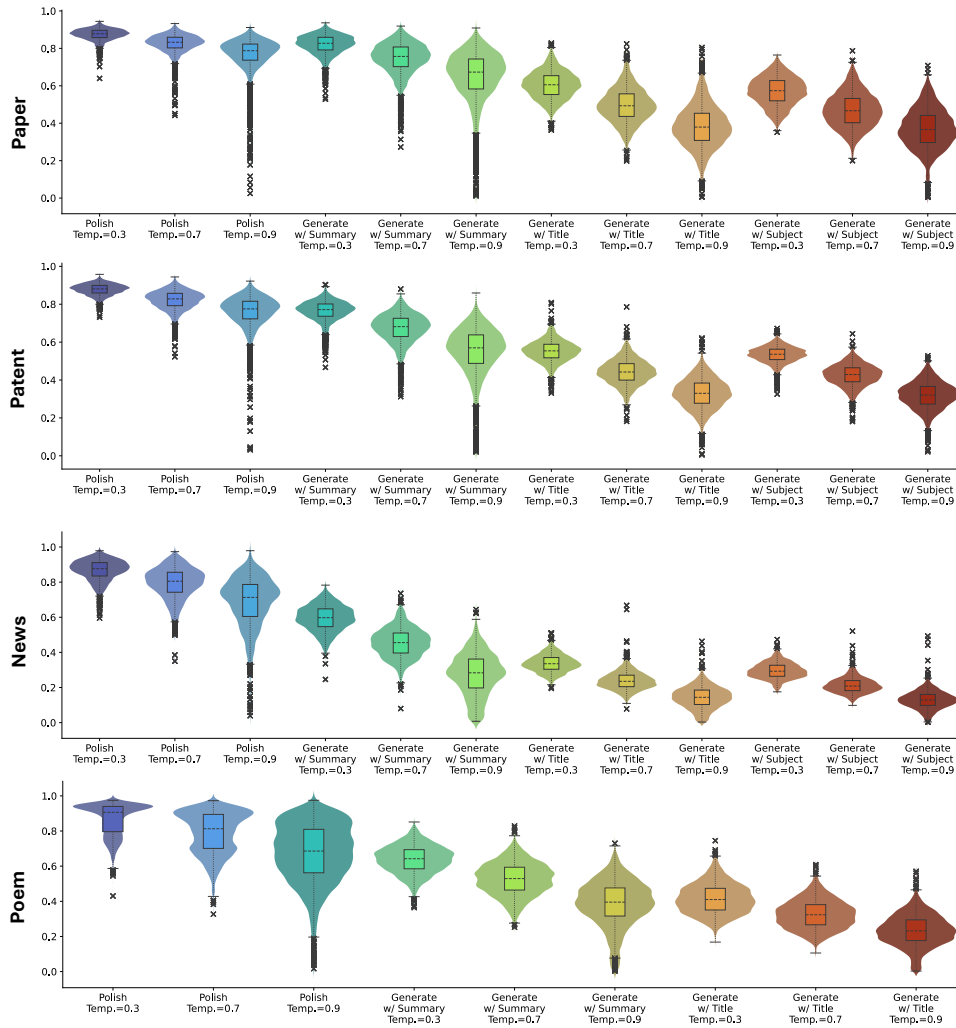


Figure 10: The distribution of the proposed measurement outcomes for AI-assisted generation under different sampling temperature settings using Llama-3. Overall, the results are consistent with our expectation that higher sampling temperatures lead to smaller measured human contributions.

resembles human writing, thereby attempting to increase the perceived human contribution.

Figure 11 shows the results of our measure with and without attacks using Llama-3. 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.

C.4 Generalization of Our Method

Besides the generalization results on news data presented in the main paper, we further show the results on paper abstract, patent abstract, and poem in Figure 12, 13, and 14, respectively.

C.5 Multi-Round Generation Analysis

In addition, we further investigate the validity of our evaluation in the context of multi-round generation. As illustrated in Figure 15, we test four commonly used multi-round generation scenarios and examine whether the estimated human contribution values exhibit expected patterns and variations across outputs from different rounds. Specifically, we sample 1,000 news articles from the original dataset for the first round of generation using Llama-3, covering all generation modes. We then introduce second and third rounds of interactions, involving three common real-world operations: adding details, correcting hallucinations, and correcting grammatical errors. Human contribution values for outputs from different rounds are computed using our output-only estimation framework, and pairwise comparisons are conducted across

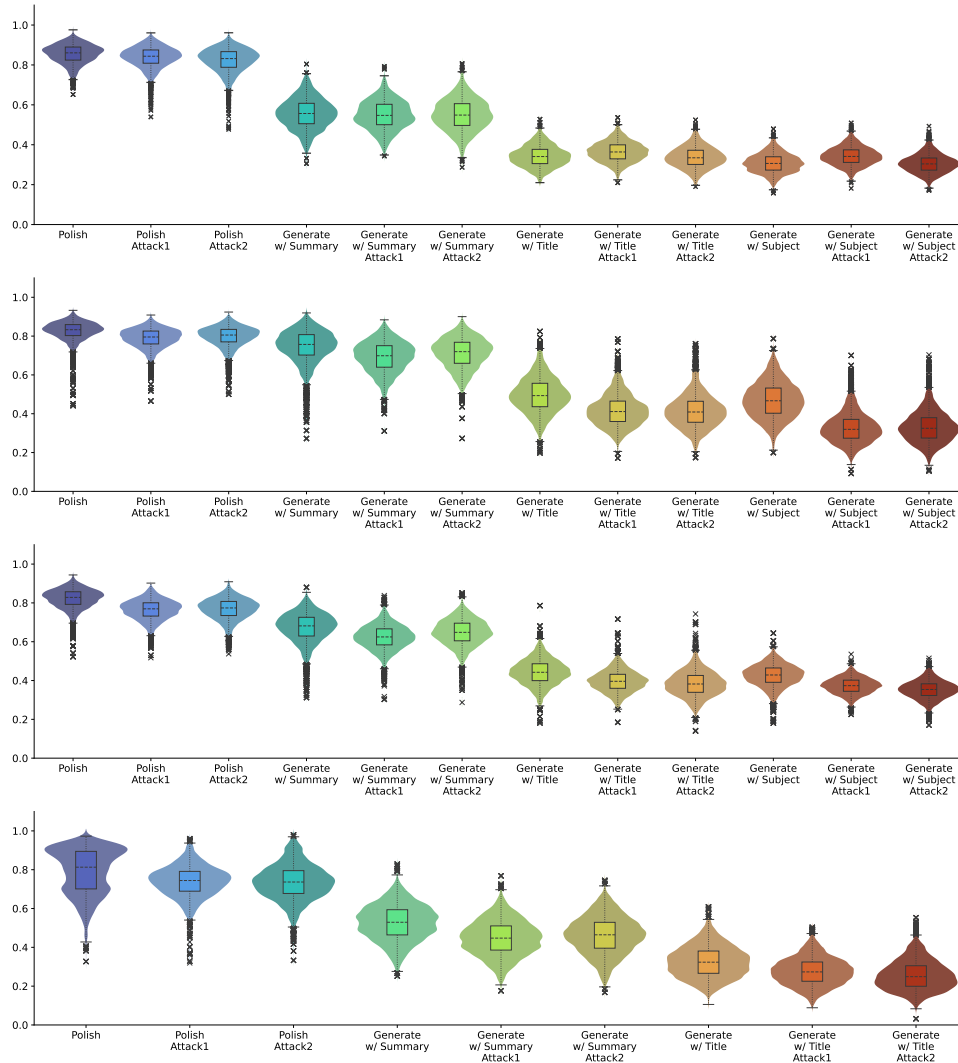


Figure 11: The distribution of the outcomes of the proposed measure for the constructed dataset with and without adaptive attacks using Llama-3. Overall, the adaptive attacks have little to no influence on the measurement outcomes.

consecutive rounds to analyze how contribution levels evolve in response to each operation.

The experimental results are presented in Figure 15. For each scenario, the left panel shows the relationship between the estimated human contribution for the Round 1 output (x-axis) and the Round 2 output (y-axis), while the right panel depicts the relationship between the Round 2 output (x-axis) and the Round 3 output (y-axis). These scatter plots visualize pairwise differences in estimated human contribution across consecutive rounds. The diagonal line indicates perfect agreement between rounds. The results generally align with our expectations. First, we observe consistency across rounds: the human contribution differences introduced by various generation modes (e.g., polishing, generation from a summary, title, or subject) in

the first round tend to persist in subsequent rounds. Second, for the “adding details” operation (Rounds 1–2 and 2–3 in Scenario 1; Rounds 2–3 in Scenarios 2 and 4), most points lie below the diagonal, indicating that human contribution generally decreases as the model adds more content with minimal new human input. For “grammar correction” (Round 2-3 in Scenario 3 and Round 1–2 in Scenario 4), the results are the most stable, as grammar errors in model-generated text are generally sparse, leading to minimal changes in estimated human contribution. For the “correcting hallucinations” operation (Rounds 1–2 in Scenarios 2 and 3), many points appear above the diagonal, suggesting that human contribution can increase when hallucinated content from the previous round is corrected and the output becomes more evidently rooted in the

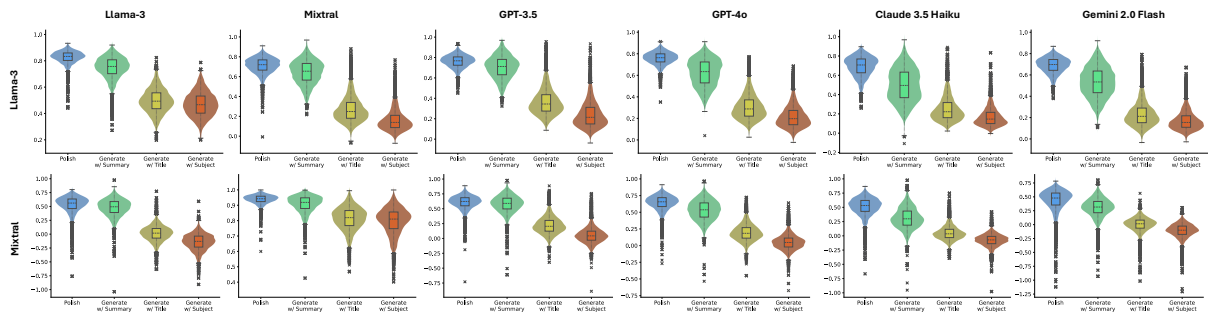


Figure 12: The distribution of the proposed measurement on the constructed dataset of paper abstracts for various generation models (columns) and surrogate models for measurement (rows). Overall, for each model pair, the proposed measure exhibits the desired trend that lower measured values are obtained for the generated content with less human contribution.

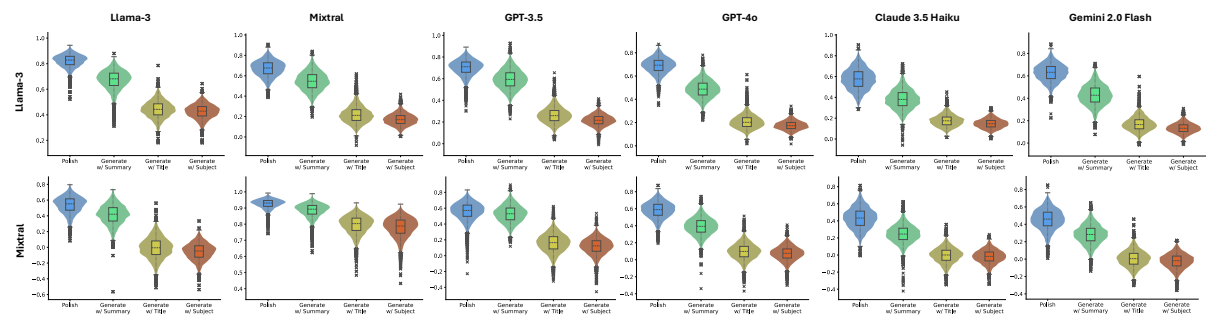


Figure 13: The distribution of the proposed measurement on the constructed dataset of patent abstracts for various generation models (columns) and surrogate models for measurement (rows). Overall, for each model pair, the proposed measure exhibits the desired trend that lower measured values are obtained for the generated content with less human contribution.

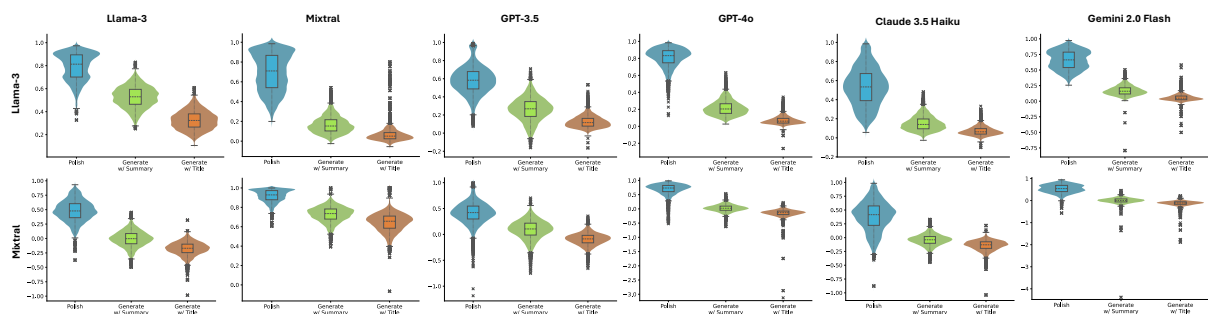


Figure 14: The distribution of the proposed measurement on the constructed dataset of poems for various generation models (columns) and surrogate models for measurement (rows). Overall, for each model pair, the proposed measure exhibits the desired trend that lower measured values are obtained for the generated content with less human contribution.

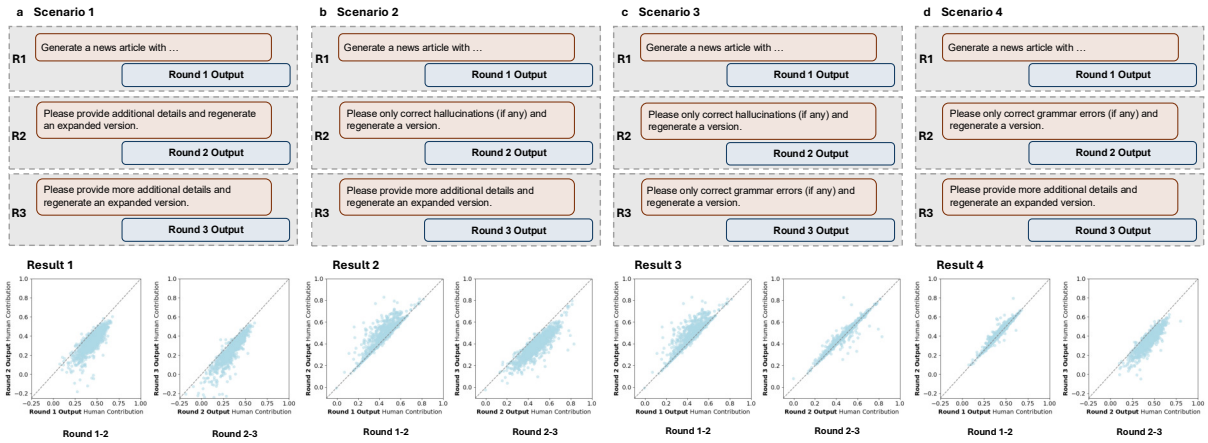


Figure 15: Illustration of the four multi-round generation scenarios and corresponding evaluated results. The scatter plots (Results 1–4) visualize pairwise comparisons of estimated human contribution between rounds. For each subplot, the x-axis and y-axis represent estimated human contribution values for two consecutive rounds. The diagonal line indicates perfect agreement.

	CHEAT (Yu et al., 2025)	HC3-English (Guo et al., 2023)	HC3-Chinese (Guo et al., 2023)	Ghostbuster (Verma et al., 2024)
AUC	0.9593	0.9931	0.9710	0.9857

Table 2: ROC–AUC performance of the proposed estimation framework on AI-generated content detection benchmarks.

human-provided information. If no hallucinations are present, the contribution remains stable.

C.6 AI-generated Content Detection

Our estimation methods proposed in section 4.2 can also be applied to AI-generated content detection tasks. To evaluate this capability, we conduct experiments on four detection benchmarks spanning multiple domains, including academic abstracts, student essays, creative fiction, news articles, Reddit posts, Wikipedia entries, as well as medical and financial texts. The evaluation covers two languages, English and Chinese.

For each dataset, we randomly sample 1,000 human-written texts and 1,000 AI-generated texts. Using only the text, we compute the human-contribution estimation score and assess its discriminative performance using ROC–AUC.

The results (Table 2) show consistently high AUC values across datasets, indicating that the output-only estimation framework can effectively distinguish AI-generated content from human-written content. This further demonstrates that the proposed output-only estimation metric captures meaningful signals related to the degree of human involvement.