Measuring Human Contribution to Human–AI Collaborative Content Generation in the Wild

Anonymous authors
Paper under double-blind review

Abstract

With the widespread adoption of generative artificial intelligence, content creation is shifting from independent human authorship towards human-guided collaborative generation with AI. The extent of human contribution therefore varies substantially across works, challenging conventional notions of originality. Existing approaches to measuring human contribution on a continuous scale rely primarily on internal model information, such as token probabilities. Meanwhile, content generation increasingly relies on black-box services provided by frontier models such as GPT, Claude, and Gemini. This mismatch between the white-box access required for evaluation and practical access constraints limits the applicability of existing methods in the wild. To address this limitation, we propose the Human Contribution Evaluation Framework (HCEF) for measuring human informational contribution using only observable human inputs and AI outputs. Drawing on the information-theoretic notion of side information, HCEF measures human contribution as the usable information gain that human inputs provide about the realized output. It normalizes this gain by the output's baseline information content to obtain a human contribution score. This enables continuous, recomputable instance-level assessment without access to the generating model's internal states or probability distributions. Extensive experiments demonstrate that HCEF effectively captures varying degrees of human contribution under strictly black-box conditions, with further validation in the wild.

1 Introduction

Rapid advances in large language models (LLMs) are reshaping content production. Human–AI collaborative content generation is widely used in academic writing (Cheng et al., 2025), office work (Jaffe et al., 2024), and artistic creation (Dergaa et al., 2023), and is becoming an important content generation paradigm (Zhang et al., 2025). Unlike independent human authorship, this paradigm combines human intentions, knowledge, and creativity with AI's generative capabilities through ongoing interaction to produce the final content. However, substantive human intellectual contribution varies considerably across works generated with AI. This variability challenges originality assessment, contribution attribution, and regulation (Richburg et al., 2024), particularly in academic research (Watson et al., 2025), education (Luo, 2024), and creative works (Hwang et al., 2025). Disputes over copyright eligibility and ownership in AI-assisted creation continue to grow. Related cases and policy discussions underscore the importance of accurately identifying the extent of human contribution (Formosa et al., 2025).

In human–AI collaborative content generation, human involvement directly affects how the final work's originality is attributed. When AI only polishes an author's existing text, the work's originality can be clearly attributed to the author. Conversely, content generated independently by AI from brief instructions containing little information may exhibit little substantive human creativity. In practice, human inputs influence generation to varying degrees, making their contribution difficult to determine directly.

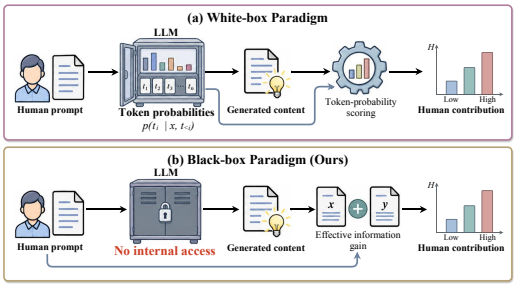

Figure 1: Human contribution evaluation in black-box settings better reflects real-world AI usage.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Identifying and measuring the actual informational contribution of human input to the output is therefore essential for assessing originality and attributing contributions.

During the early adoption of generative AI, research focused primarily on neural text detection Zellers et al. (2019), which distinguishes human-written from machine-generated text. Approaches include perplexity-based detectors, language watermarking, and methods such as DetectGPT Mitchell et al. (2023) and DNA-GPT Yang et al. (2024). As LLMs have entered everyday writing, researchers have recognized that content in the wild is often neither entirely human-written nor entirely AI-generated. Studies have therefore explored discrete and stage-based assessments of human contribution. Text-based approaches estimate human involvement from semantic overlap between prompts and outputs (A. Welzel & Vincent, 2026). Process-based approaches use rating scales, decision logs, and editing histories to document human and AI contributions in a structured or stage-wise manner (Vyas et al., 2026; Al-Kabbany, 2026). Xie et al. (2026) introduce a continuous measure based on token-level probabilities from the generating model. They compare the probability of the same output with and without conditioning on human input. However, human–AI collaborative content generation increasingly relies on frontier models such as GPT and Claude accessed through black-box services, limiting the applicability of continuous measures that require inaccessible model-internal information, such as token probabilities. As shown in Figure 1, this mismatch between white-box evaluation requirements and real-world black-box access makes continuous, recomputable measurement of human contribution from observable inputs and outputs an open challenge.

To address this challenge, we propose the Human Contribution Evaluation Framework (HCEF) for measuring human informational contribution using only observable human inputs and AI outputs. HCEF formulates this task as estimating the information gain provided by side information for an individual instance. It estimates the same output's information content under paired conditions, without side information and with human input as side information. It extracts the positive information gain provided by the input and normalizes it by the output's baseline information content to obtain a human contribution score. HCEF thus enables continuous, recomputable instance-level assessment without accessing the generating model's internal states or probability distributions. We systematically evaluate HCEF's effectiveness, cross-domain applicability, and alignment with human perception, alongside its stability and generalization across collaborative generation conditions. To this end, we construct a human–AI collaborative content generation dataset spanning varying levels of human contribution, multiple content creation domains, and multiple frontier models. The dataset provides a unified benchmark for systematically evaluating human contribution. We further apply HCEF to human–AI collaborative content generation data collected in the wild to assess its effectiveness and practical utility. Ablation studies, comparisons with white-box evaluations, and computational efficiency analyses further assess HCEF's reliability and practicality, supporting its application across diverse human–AI collaborative settings.

Our main contributions are as follows:

1. We formulate the problem of measuring human informational contribution on a continuous scale under strictly black-box access constraints. The task requires only observable human inputs and AI outputs under strictly black-box conditions.
2. We construct an evaluation dataset of over 56,000 human–AI collaborative content generation instances spanning varying levels of human contribution. Built using frontier models such as GPT 5.6 Sol and Claude Opus 4.8, it supports systematic comparisons across diverse collaborative generation conditions.
3. We propose HCEF, which uses the information-theoretic notion of side information to measure the usable information gain that human inputs provide about the realized output. HCEF converts this gain into a normalized, recomputable human contribution score.
4. Extensive experiments demonstrate that HCEF effectively captures varying degrees of human contribution under strictly black-box conditions, with further validation in the wild.

2 Problem Formulation

In human–AI collaborative content generation, humans shape outputs by providing content, specifying constraints, and communicating judgments. The same generative system may receive inputs ranging from a broad topic to a complete draft. Input length alone does not determine the contribution

of these inputs to the resulting content. We focus on the informational dimension of this contribution: *how much usable information does the human input provide about a particular realized output?* Our measurement principle is the reduction in the information required to describe the same output when the input is available as side information. Let M_θ denote a fixed generative model, and let γ specify the generation protocol, including the model version, message template, and decoding configuration. Given human input $X = x$, the output follows

$$Y \mid X = x \sim p_{\theta,\gamma}(\cdot \mid x). \tag{1}$$

We take a completed generation record (x, y) as the unit of measurement. Here, x is the complete final prompt submitted to the model, including instructions, context, content, and constraints. The output y is the model's raw response before any subsequent human editing. Under a fixed coding protocol Π, our objective is to construct an instance-level score

$$\mathcal{R}_\Pi : (x, y) \longmapsto R^\Pi_{\mathrm{HC}}(x, y) \in [0, 1], \tag{2}$$

which quantifies the positive description savings attributable to the input relative to the baseline description cost of the output. The protocol Π is independent of the generative model. Scoring requires only the observed record, without access to model internals or sequence probabilities, and without further calls to the generative model. Attribution is restricted to the input represented by x. If the prompt contains material from multiple sources, the score evaluates the prompt as a whole without separating their individual contributions.

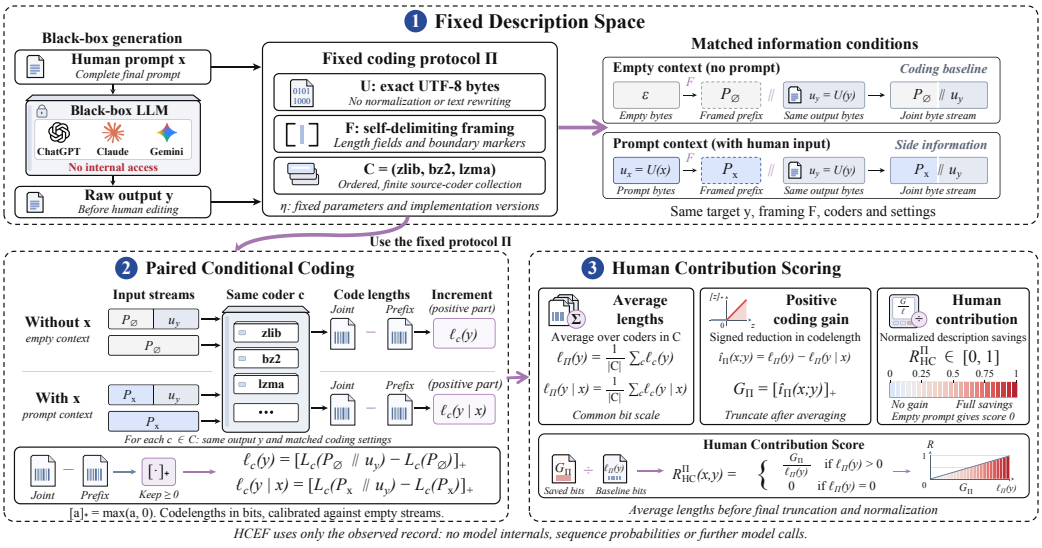

Figure 2: **Overview of the HCEF Framework for Human Contribution Evaluation.** HCEF quantifies human contribution as the effective information gain supplied by the prompt about the generated output. It measures this gain through the reduction in the same output's codelength with prompt side information, normalizing the positive savings into a contribution score.

3 Method

3.1 The Descriptive Value of Side Information

HCEF takes the descriptive value of side information as the basis for measuring informational contribution. For any reference joint distribution $q(X, Y)$ with $q(x, y) > 0$, ideal probabilistic codelengths yield

$$j_q(x; y) = -\log_2 q_Y(y) + \log_2 q(y \mid x) = \log_2 \frac{q(y \mid x)}{q_Y(y)}. \tag{3}$$

This information density measures the signed reduction in ideal codelength obtained by observing the input. When the relevant entropies are finite, its expectation satisfies

$$\mathbb{E}_q[j_q(X; Y)] = I_q(X; Y) = H_q(Y) - H_q(Y \mid X) = \mathbb{E}_{X \sim q_X}\left[D_{\mathrm{KL}}\left(q_{Y|X}(\cdot \mid X) \,\|\, q_Y\right)\right]. \tag{4}$$

where H_q denotes entropy and D_{KL} denotes Kullback–Leibler divergence. These identities relate the value of side information to both an expected reduction in output uncertainty and a divergence between conditional and marginal output distributions (Shannon, 1948). By the chain rule, an auxiliary input adds no mutual information about the output if they are conditionally independent given the original input under q. These relations motivate evaluating output-relevant information supplied by the input, rather than its surface size alone. The corresponding derivations are provided in Appendix D.1. For individual finite objects, algorithmic information theory characterizes the value of side information through

$$g_K(x \to y) = K_{\mathcal{U}}(y) - K_{\mathcal{U}}(y \mid x), \tag{5}$$

where $K_{\mathcal{U}}$ denotes Kolmogorov complexity relative to a fixed universal machine $\mathcal{U}$. This formulation shifts the focus from distributional averages to the conditional description cost of an individual object. Practical use of finite codelengths to characterize relationships between objects is also studied by Cilibrasi & Vitányi (2005). These perspectives motivate a common structure for instance-level measurement: compare the description costs of the same output with and without the input. We instantiate this principle using a fixed coding protocol Π. Let $\ell_\Pi(y)$ and $\ell_\Pi(y \mid x)$ denote the baseline and conditional effective codelengths, respectively. We define the signed side-information gain proxy as

$$\widehat{i}_\Pi(x; y) = \ell_\Pi(y) - \ell_\Pi(y \mid x). \tag{6}$$

The information-theoretic relations motivate this construction; the computable score is defined directly by Π. We do not assume that the finite coders recover the reference distribution or attain shortest descriptions. The generative system and the measurement protocol therefore serve distinct roles: the former determines how outputs arise, while the latter fixes a reproducible description scale.

3.2 Description Space and Coding Protocol

To evaluate outputs from different generative systems under a description rule, we define a coding protocol Π consisting of U, F, $\mathcal{C}$, and η. Here, U is a deterministic mapping from text to bytes, F is a self-delimiting context framing function, and $\mathcal{C}$ is an ordered finite collection of source coders. The configuration η fixes coding parameters and implementation versions. For any text t, let $u_t = \mathsf{U}(t)$. We use UTF-8 without Unicode normalization, whitespace folding, or text rewriting. The same framing function constructs the contexts associated with the input and with empty side information:

$$P_x = \mathsf{F}(u_x), \qquad P_\varnothing = \mathsf{F}(\epsilon), \tag{7}$$

where ϵ is the empty byte string. The framing function uses length fields and fixed boundary markers to distinguish the context from the target, ensuring a consistent layout in both conditions. The empty context $P_\varnothing$ supplies an internal coding baseline without requiring an additional generation record. We instantiate the coder collection with three lossless compression algorithms: `zlib`, `bz2`, and `lzma`. For each coder c, $L_c(z)$ denotes the effective codelength, measured in bits, of the byte string z, after calibrating for the coder-specific overhead using the corresponding empty-stream output. Each coder and its parameters are fixed before scoring, so both information conditions use the same coding rules. Appendix D specifies the byte mapping, framing, and effective codelength definitions.

3.3 Conditional Description with Side Information

We measure the descriptive contribution of the input through its marginal effect on the coding cost of the output. Using matched representations and coding rules, we define paired operators for the effective incremental codelength under empty and observed input contexts:

$$\ell_c(y) = \big[L_c(P_\varnothing \| u_y) - L_c(P_\varnothing)\big]_+, \qquad \ell_c(y \mid x) = \big[L_c(P_x \| u_y) - L_c(P_x)\big]_+. \tag{8}$$

where $\|$ denotes byte concatenation and $[a]_+ = \max\{a, 0\}$. Both quantities describe the same target under identical framing and coding rules, net of their respective context codelengths. The prompt receives credit only for reducing the output's incremental codelength, not for its own description cost. For an empty prompt, the two effective codelengths coincide by construction. Finite coding implementations may yield nonmonotonic stream lengths due to block boundaries, termination overhead, or changes in internal coding models. We therefore apply $[\cdot]_+$ to each coder's incremental codelength while retaining the untruncated values. Derived from a difference of complete-stream codelengths, the conditional effective codelength need not correspond to a uniquely decodable conditional code. The paired construction aligns representation and computation across conditions, while the fixed protocol governs the remaining content-dependent coding effects.

3.4 Human Contribution Score

To combine the gains captured by different coding mechanisms, we aggregate effective codelengths before normalization. On the common bit scale, we compute

$$\ell_\Pi(y) = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} \ell_c(y), \qquad \ell_\Pi(y \mid x) = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} \ell_c(y \mid x). \tag{9}$$

Their difference gives the signed gain $\widehat{i}_\Pi(x; y)$. We retain its positive part as the input's positive coding gain:

$$G_\Pi(x \to y) = \big[\ell_\Pi(y) - \ell_\Pi(y \mid x)\big]_+. \tag{10}$$

Normalizing by the baseline effective codelength of the output yields the Human Contribution Score:

$$R^\Pi_{\mathrm{HC}}(x, y) = \begin{cases} \dfrac{\big[\ell_\Pi(y) - \ell_\Pi(y \mid x)\big]_+}{\ell_\Pi(y)}, & \ell_\Pi(y) > 0, \\ 0, & \ell_\Pi(y) = 0. \end{cases} \tag{11}$$

The score R^Π_{HC} expresses the description savings G_Π induced by the input as a fraction of the output's baseline effective codelength. For a positive baseline, larger savings yield a higher score, while nonnegative effective codelengths ensure $R^\Pi_{\mathrm{HC}} \in [0, 1]$. The score is zero when the conditional codelength is no smaller than the baseline, and one when it vanishes under a positive baseline; zero-baseline cases, including empty outputs, are assigned zero by definition. We aggregate codelengths before normalization, allowing coder-specific gains of opposite signs to offset before truncation. When all coder baselines are positive, the score is the positive part of a baseline-weighted average of signed relative gains, avoiding equal weighting of ratios with very small denominators. We use R^Π_{HC} as the contribution score and retain unnormalized gains and coder-specific codelengths as auxiliary quantities. Appendix D provides further aggregation properties and codelength sensitivity analysis.

In summary, HCEF combines a fixed description space, paired conditional coding, and codelength aggregation to evaluate prompt–output records. The resulting score quantifies the usable description gain supplied by the input relative to the output baseline under the protocol. It provides deterministic, reproducible instance-level evaluation without accessing or querying the generative model.

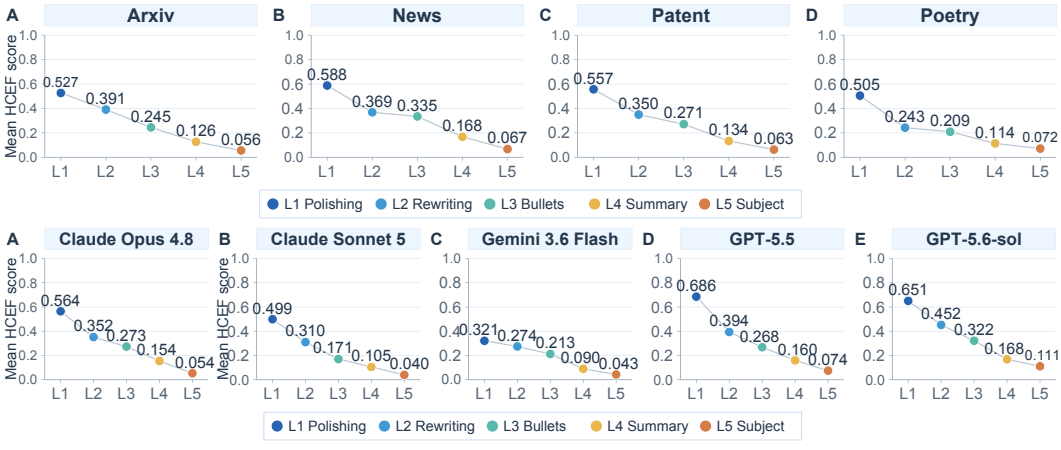

Figure 3: **Human contribution measurement across domains and models.** Across multiple content domains and frontier models used for content generation, HCEF consistently exhibits the expected downward trend as the level of human contribution decreases.

4 Dataset Construction for Human Contribution Evaluation

To evaluate the effectiveness and applicability of HCEF for measuring human informational contribution, we construct a human–AI collaborative content generation dataset spanning five generation

levels. We further vary content domains and generating models and perform repeated stochastic sampling to enable systematic comparisons across generation conditions. Dataset construction consists of two stages: source data collection and processing, followed by AI-assisted content generation.

4.1 Data Collection and Processing

We sample source texts from public datasets in four domains: academic paper abstracts, news articles, patent abstracts, and poetry, with more than 2,000 texts initially sampled per domain. We prepare inputs for five levels: *Polishing*, *Rewriting*, *Bullets*, *Summary*, and *Subject*. For *Polishing* and *Rewriting*, we retain the complete source text for language polishing and full-text rewriting, respectively. For *Bullets*, we summarize the source text into several key points, each expressed as a complete sentence. For *Summary*, we condense the source text into two sentences. For *Subject*, we extract up to three lexical cues from the source text using uniform extraction rules to form a topic prompt. The intermediate representations for *Bullets* and *Summary* are generated by the corresponding generating model from the first 2,000 characters of the source text. This preprocessing provides the content inputs for all five generation levels. Data sources and sampling procedures are detailed in Appendix B.1.

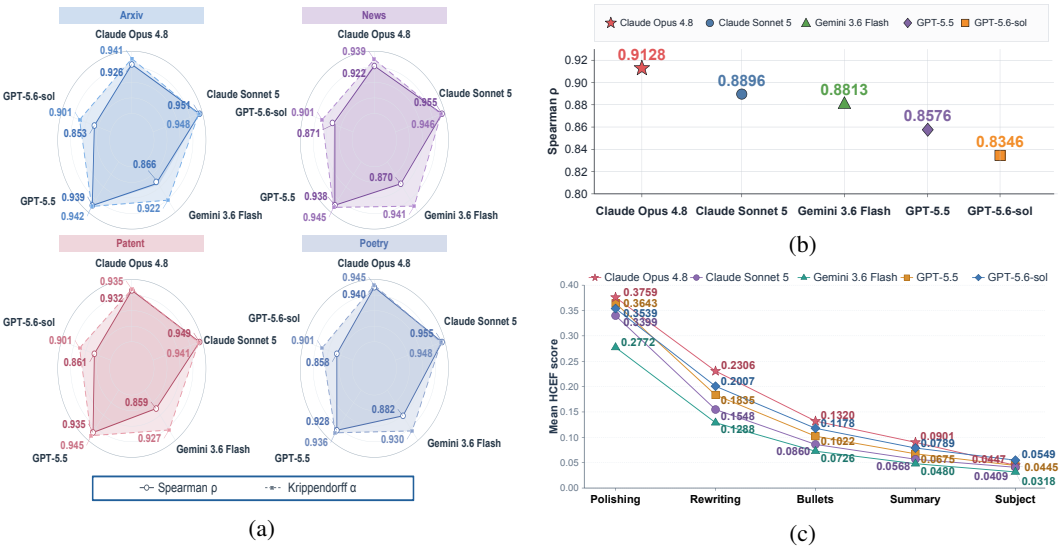

Figure 4: Human contribution evaluation results. (a) HCEF exhibits consistently high Spearman's ρ and Krippendorff's α across generating models and content domains. (b) HCEF measures cross-model consistency of human contribution on real-world user prompts. (c) HCEF scores decrease monotonically across five contribution groups for all frontier models.

4.2 Human–AI Collaborative Content Generation

We use LLMs to generate content from inputs at each of the five levels. For *Polishing*, the model is instructed to improve language while preserving the original content. For *Rewriting*, it is instructed to preserve the source information while restructuring and rephrasing the text. For *Bullets*, *Summary*, and *Subject*, the model generates content from key points, brief summaries, and topic cues, respectively. For *Bullets* and *Summary*, the prompt contains only the corresponding intermediate representation and an instruction to expand it. The five-level setup progresses from polishing and rewriting complete texts to generation guided by summarized information and sparse topic cues. These levels simulate different forms of human involvement by varying the supplied content and the constraints on how it is expressed. Details of prompt construction are provided in Appendix B.3. To support cross-model analysis, we use five generating models: Claude Opus 4.8, Claude Sonnet 5, Gemini 3.6 Flash, GPT 5.5, and GPT 5.6 Sol. We run all five tasks with each model at generation temperatures ranging from 0.5 to 0.8. We generate five outputs per prompt to obtain diverse samples. The corresponding statistics are presented in Table 1. Further details are in Appendix B.

5 Experiments

We systematically evaluate the effectiveness of HCEF on the constructed dataset under a strict black-box setting that relies solely on externally observable information from human–AI collaboration. Specifically, the human prompt x and AI-generated output y are observable, whereas token-level generation probabilities and other internal signals from frontier models, including GPT 5.6 Sol and Claude Opus 4.8, remain inaccessible. This setting reflects real-world human–AI collaborative content generation in the wild, where closed-source frontier models are widely deployed.

5.1 Human Contribution Evaluation

As shown in Figure 3, HCEF scores decreased consistently as human input became less informative, following the predefined ordering from complete content to summaries, titles, and topics. This monotonic trend indicates that HCEF effectively captures the information dependence between human input and AI-generated output and distinguishes different levels of human contribution. At the same contribution level, HCEF score distributions varied across content domains, showing that the metric is sensitive to specific input–output relationships rather than assigning fixed scores to generation modes. Such variation reflects domain differences in content complexity, information density, and the information gain provided by human input. Figure 4a further evaluates consistency with the predefined contribution ordering across five generating models. Spearman's ρ ranged from 0.853 to 0.955, while Krippendorff's α ranged from 0.901 to 0.948. These results demonstrate that HCEF consistently recovers contribution levels across models, supporting its effectiveness and robustness for continuous human contribution measurement. More results are in Appendix C.1.

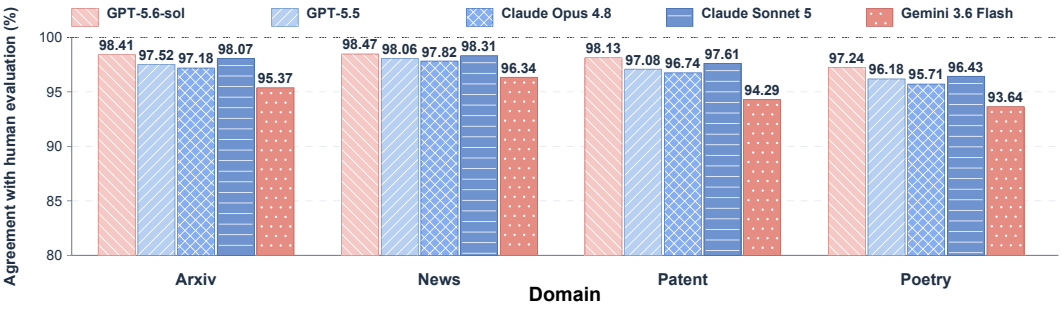

Figure 5: HCEF shows strong alignment with human perception across content domains and generating models, achieving 96.93% average agreement with aggregated human judgments.

5.2 Alignment with Human Perception

To assess whether HCEF aligns with human perceptions of contribution in human–AI collaborative content generation, we adopt a human preference annotation protocol inspired by pairwise comparison frameworks such as Chatbot Arena (Chiang et al., 2024; Xie et al., 2026). We construct candidate pairs from samples generated by five models, including GPT 5.6 Sol and Claude Opus 4.8, across four content domains and all generation modes. We retain 7,500 input–output pairs whose HCEF scores differ by more than 0.05, ensuring sufficient score separation and reducing ambiguity in human judgments. Each pair is independently assessed by nine human annotators, who identify which example exhibits greater human contribution. The consensus preference for each pair is determined by majority vote and compared with the ranking induced by HCEF, thereby evaluating alignment with human perception. Annotation guidelines and procedures are detailed in Appendix B.4. As shown in Figure 5, HCEF agrees with aggregated human judgments on 96.93% of the selected pairs. This result supports the effectiveness and robustness of HCEF, demonstrating strong alignment with human judgments across content domains and collaborative generation settings.

5.3 Application to Human–AI Collaborative Content Generation in the Wild

To assess applicability beyond the five controlled generation levels, we evaluate HCEF's cross-model consistency on real user prompts. We sample prompts from WildChat Zhao et al. (2024),

10k_prompts [1], and OASST1 Köpf et al. (2023), and classify them using Claude Opus 4.6, which is excluded from subsequent content generation. We retain six content creation categories, including creative generation, experiential context, and artistic design, yielding 6,569 prompts. Table 2 reports sample counts and category distributions. Independent generation with Claude Opus 4.8, Claude Sonnet 5, GPT 5.5, GPT 5.6 Sol, and Gemini 3.6 Flash yields 25,027 prompt–output examples. For each prompt, we compute HCEF using its reference output and compare this score with scores from independently generated outputs across models. Within each dataset and content category, we form five approximately equal-sized groups using reference scores. Group assignments remain fixed when evaluating independent outputs to test whether the relative contribution ordering is reproduced. As shown in Figure 4b, across all five models, Spearman's ρ with reference scores ranged from 0.8346 to 0.9128, indicating high rank agreement. As shown in Figure 4c, mean scores decreased strictly across the five fixed groups for every model. For GPT 5.6 Sol, the group means were 0.3539, 0.2007, 0.1178, 0.0789, and 0.0549, respectively. These results support stable cross-model contribution rankings on the evaluated prompts and HCEF's applicability to human–AI collaborative content generation in the wild. As shown in Figure 6, a real-world GPT 5.6 Sol example is presented.

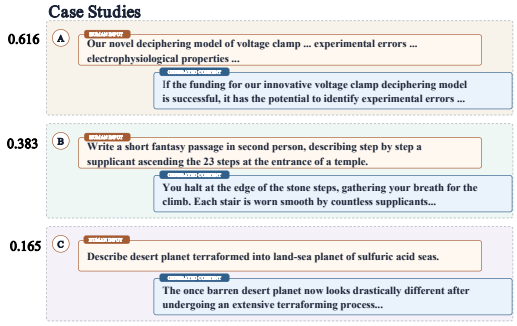

Figure 6: Case studies on data samples from two distinct classes in WildChat.

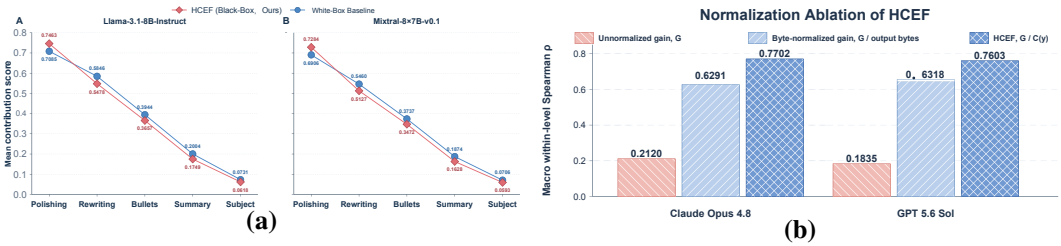

Figure 7: **(a)** Comparison with white-box evaluation. HCEF achieves higher strict monotonicity than white-box baselines, showing stronger preservation of human contribution ordering across generation levels. **(b)** Ablation study. HCEF with baseline-information normalization consistently achieves stronger agreement with white-box references than unnormalized gain or byte-length normalization.

5.4 Comparison with White-Box Baselines

We compare HCEF with the white-box baseline of Xie et al. (2026) on the same 56,110 prompt–output examples, using the predefined contribution ordering as the reference. The baseline adopts Llama 3.1 8B-Instruct and Mixtral-8x7B-v0.1 as probability evaluation models. To assess complete recovery of the contribution ordering, we compute the strict monotonicity rate across five generation levels. As shown in Figure 7a, the overall Spearman's ρ between HCEF and the Llama and Mixtral scores were 0.9570 and 0.9563, respectively. Across 11,222 five-level groups, HCEF achieved a strict monotonicity rate of 68.92%, exceeding the two baselines by 5.86 and 5.03 percentage points. These results show strong agreement with probability-based measures and more complete recovery of the predefined ordering, without requiring access to model probabilities. We further evaluate computational cost and latency on 600 examples spanning six text-length ranges. At an average output length of approximately 1,200 tokens, CPU-only HCEF achieves a per-example latency of 35.6 ms, yielding a 2.87× speedup over the GPU-accelerated LLaMA 3.1 8B-based white-box baseline (Figure 11). HCEF also maintains relatively stable computational cost and latency across text lengths, whereas the white-box baseline becomes increasingly costly for longer outputs. By avoiding both GPU computation and an additional probability evaluation model, HCEF enables lightweight and efficient human contribution evaluation. Further details are provided in Appendix C.3.

[1] https://huggingface.co/datasets/data-is-better-together/10k_prompts_ranked

5.5 Ablation Study

To assess the role of normalization in the human contribution score, we compare three variants on the same $56{,}110$ prompt–output examples while keeping the positive information gain G unchanged: unnormalized gain G, byte-normalized gain $G/|u_y|$, and HCEF's baseline-normalized score $G/C(y)$. These variants isolate the effect of normalization and the choice of denominator. We measure agreement with white-box references using Spearman's ρ computed within each generation level and averaged with equal weights. As shown in Figure 7b, unnormalized gain yields macro correlations of 0.2120 and 0.1835 across the two plotted settings. Byte-length normalization increases these values to 0.6291 and 0.6318, whereas HCEF achieves 0.7702 and 0.7603, exceeding byte-length normalization by 0.1411 and 0.1285, respectively. These results support using baseline output information content to express input gain as a relative contribution, yielding stronger within-level agreement with probability-based measures than either absolute gain or byte-length normalization. Additional ablation details are provided in Appendix C.4.

6 Related Work

6.1 Provenance and Process in Human–AI Content Generation

Earlier studies on neural text detection distinguish human-written from machine-generated text through likelihood curvature, regeneration consistency, proxy-model features, and watermarking (Mitchell et al., 2023; Yang et al., 2024; Verma et al., 2024; Kirchenbauer et al., 2023). As human–AI collaborative content generation becomes more prevalent, subsequent studies address mixed textual provenance. MixSet benchmarks detection on AI-revised human text and human-revised AI text (Zhang et al., 2024). HACo-Det provides word-level attribution labels and evaluates fine-grained detection and document-level AI content ratios (Su et al., 2025). Process-based approaches characterize human and AI contributions through interaction records, decision logs, and editing histories. SLICE estimates heuristic contribution ratios across eight stages of academic writing using transcripts, manuscripts, and editing logs (Earp & Porsdam Mann, 2026). PAIRED prospectively documents decision points and AI involvement in scientific research (Al-Kabbany, 2026). These studies characterize textual provenance and participation throughout the generation process. HCEF quantifies how much usable information the human input provides for a realized AI output, producing continuous, instance-level scores using only the observed prompt–output pairs.

6.2 Measuring Human Contribution

Continuous measures capture varying degrees of human involvement beyond textual provenance.Guo et al. (2025) train an output-only RoBERTa predictor using normalized prompt–output BERTScore recall as supervision. Xie et al. (2026) measure contribution as the reduction in output surprisal after conditioning on human input, normalized by unconditional output surprisal. This requires token probabilities from the generator or a surrogate evaluator. Frontier services such as GPT and Claude often preclude direct probability access, while surrogate scoring still requires a probability-accessible model. HCEF adopts a side-information perspective, defining human informational contribution as positive compression gain under a fixed coding protocol. It derives continuous, normalized, and reproducible instance-level scores solely from prompt–output pairs, without accessing generator internals, training a contribution predictor, or requiring surrogate likelihood evaluation.

7 Conclusion

We address the problem of measuring human contribution in human–AI collaborative content generation in the wild. We propose HCEF, a theoretically grounded information-theoretic framework for reliable measurement of human informational contribution. Large-scale experiments across multiple content domains and frontier models demonstrate HCEF's effectiveness in distinguishing different degrees of human contribution and its alignment with human judgments. This work formalizes a practical research problem and establishes a principled measurement framework. It aims to provide a new research foundation for assessing originality, attributing human contributions, and governing AI-assisted content in the generative AI era.

AI USE STATEMENT

We did not use generative AI tools for generating synthetic data, developing theoretical models or conceptual frameworks, formulating mathematical claims or proofs, proposing or refining hypotheses, designing the methodology or experiments, implementing methods, or interpreting results. Generative AI tools were used only to assist with manuscript writing and language polishing. All AI-assisted text was carefully reviewed and verified by the authors to ensure that it accurately reflects the intended meaning. The authors take full responsibility for the final content of this work.

REFERENCES

Vitor H A. Welzel and Nicholas Vincent. Overreliance in writing tasks: Exploring similarity-based measures of ai influence on writing and proposing a reflective writing interface intervention. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency*, pp. 5941–5958, 2026.

Ahmad Al-Kabbany. PAIRED: A process-anchored framework for transparent reporting of AI contributions in scientific research, 2026. URL https://arxiv.org/abs/2605.24325.

Adam Cheng, Aaron Calhoun, and Gabriel Reedy. Artificial intelligence-assisted academic writing: recommendations for ethical use. *Advances in Simulation*, 10(1):22, 2025.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8359–8388. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/chiang24b.html.

Rudi Cilibrasi and Paul MB Vitányi. Clustering by compression. *IEEE Transactions on Information theory*, 51(4):1523–1545, 2005.

Thomas M Cover and Joy A Thomas. Elements of information theory second edition solutions to problems. *Internet Access*, pp. 19–20, 2006.

Ismail Dergaa, Karim Chamari, Piotr Zmijewski, and Helmi Ben Saad. 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, 2023.

Brian D. Earp and Sebastian Porsdam Mann. Mapping human and AI contributions across stages of academic writing using SLICE—a preliminary LLM-based method, 2026. URL https://www.researchgate.net/publication/406991607. Preprint.

Paul Formosa, Sarah Bankins, Rita Matulionyte, and Omid Ghasemi. Can chatgpt be an author? generative ai creative writing assistance and perceptions of authorship, creatorship, responsibility, and disclosure. *Ai & Society*, 40(5):3405–3417, 2025.

Yuchen Guo, Zhicheng Dou, Huy H. Nguyen, Ching-Chun Chang, Saku Sugawara, and Isao Echizen. Measuring human involvement in AI-generated text: A case study on academic writing. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–10. IEEE, 2025. doi: 10.1109/IJCNN64981.2025.11228666. URL https://doi.org/10.1109/IJCNN64981.2025.11228666.

Yohan Hwang, Dongkwang Shin, and Jang Ho Lee. Who owns ai-generated artwork? revisiting the work of generative ai based on human-ai co-creation. *Telematics and Informatics*, 98:102266, 2025.

Sonia Jaffe, Neha Parikh Shah, Jenna Butler, Alex Farach, Alexia Cambon, Brent Hecht, Michael Schwarz, and Jaime Teevan. Generative ai in real-world workplaces. *Microsoft.*, 2024.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *International conference on machine learning*, pp. 17061–17084. PMLR, 2023.

Andreas Köpf, Yannic Kilcher, Dimitri Von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. Openassistant conversations-democratizing large language model alignment. *Advances in neural information processing systems*, 36:47669–47681, 2023.

Jiahui Luo. A critical review of genai policies in higher education assessment: A call to reconsider the "originality" of students' work. *Assessment & Evaluation in Higher Education*, 49(5):651–664, 2024.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 24950–24962. PMLR, 2023. URL https://proceedings.mlr.press/v202/mitchell23a.html.

Aquia Richburg, Calvin Bao, and Marine Carpuat. Automatic authorship analysis in human-ai collaborative writing. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 1845–1855, 2024.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Zhixiong Su, Yichen Wang, Herun Wan, Zhaohan Zhang, and Minnan Luo. HACo-det: A study towards fine-grained machine-generated text detection under human-AI coauthoring. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22015–22036, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1069. URL https://aclanthology.org/2025.acl-long.1069/.

Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. 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)*, pp. 1702–1717, Mexico City, Mexico, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.95. URL https://aclanthology.org/2024.naacl-long.95/.

Marianne B. Vyas, Lori M. Rhudy, Suzanne Weckman, Joshua Ennis, Elizabeth A. Hundt, Amanda Werner, Fiona Smith, Eugena Bergvall, Florence Chukwuneke, Lindsay M. Riskey, and DaiWai M. Olson. Quantifying artificial intelligence contribution in academic writing: Development of the transparency and reporting of artificial intelligence contribution for evaluating submissions instrument. *Journal of Neuroscience Nursing*, 2026. doi: 10.1097/JNN.0000000000000903. URL https://doi.org/10.1097/JNN.0000000000000903.

Steven Watson, Erik Brezovec, and Jonathan Romic. The role of generative ai in academic and scientific authorship: an autopoietic perspective. *AI & society*, 40(5):3225–3235, 2025.

Yueqi Xie, Tao Qi, Jingwei Yi, Xiyuan Yang, Ryan Whalen, Junming Huang, Qian Ding, Yu Xie, Xing Xie, and Fangzhao Wu. Measuring human contribution in ai-assisted content generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6168–6190, 2026.

Xianjun Yang, Wei Cheng, Yue Wu, Linda Petzold, William Yang Wang, and Haifeng Chen. DNA-GPT: Divergent n-gram analysis for training-free detection of GPT-generated text. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Xlayxj2fWp.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. *Advances in neural information processing systems*, 32, 2019.

Qihui Zhang, Chujie Gao, Dongping Chen, Yue Huang, Yixin Huang, Zhenyang Sun, Shilin Zhang, Weiye Li, Zhengyan Fu, Yao Wan, and Lichao Sun. LLM-as-a-coauthor: Can

mixed human-written and machine-generated text be detected? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 409–436, Mexico City, Mexico, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.29. URL https://aclanthology.org/2024.findings-naacl.29/.

Shuning Zhang, Hui Wang, and Xin Yi. Exploring collaboration patterns and strategies in human-ai co-creation through the lens of agency: A scoping review of the top-tier hci literature. *Proceedings of the ACM on Human-Computer Interaction*, 9(7):1–43, 2025.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatgpt interaction logs in the wild. In *International Conference on Learning Representations*, volume 2024, pp. 34590–34605, 2024.

A Algorithm

Algorithm 1 summarizes HCEF evaluation for an observed prompt–output pair under a fixed coding protocol. The prompt and output are mapped to UTF-8 bytes, with matched contexts constructed for the observed input and empty side information. For each coder, baseline and conditional effective codelengths are computed by subtracting the corresponding context costs and retaining nonnegative increments. This paired construction credits the input for reducing the effective description cost of the same output, excluding the context's own coding cost. These codelengths are averaged across coders, and their positive difference is normalized by the output baseline, with zero assigned when the baseline vanishes. Aggregating codelengths before truncating the final gain allows positive and negative coder-specific gains to offset. The resulting procedure yields a deterministic, reproducible instance-level contribution score for each completed generation record without accessing model internals or making additional calls to the generative model.

Algorithm 1 HCEF: Human Contribution Evaluation Framework

Require: Observed prompt x, raw output y, and fixed protocol $\Pi = (\mathsf{U}, \mathsf{F}, \mathcal{C}, \eta)$
Ensure: Human Contribution Score $R^{\Pi}_{\mathrm{HC}}(x, y) \in [0, 1]$
Coders: $\mathcal{C} = (\texttt{zlib}, \texttt{bz2}, \texttt{lzma})$, with settings fixed by η.
Notation: L_c denotes calibrated effective codelength; $[a]_+ = \max\{a, 0\}$; ϵ is the empty byte string; $\|$ denotes byte concatenation.
Construct the description space and paired contexts
$u_x \leftarrow \mathsf{U}(x), \quad u_y \leftarrow \mathsf{U}(y)$
$P_\varnothing \leftarrow \mathsf{F}(\epsilon), \quad P_x \leftarrow \mathsf{F}(u_x)$
Compute paired effective codelengths
for each $c \in \mathcal{C}$ **do**
  $d_c^0 \leftarrow L_c(P_\varnothing \| u_y) - L_c(P_\varnothing)$
  $d_c^x \leftarrow L_c(P_x \| u_y) - L_c(P_x)$
  $\ell_c(y) \leftarrow [d_c^0]_+$
  $\ell_c(y \mid x) \leftarrow [d_c^x]_+$
end for
Aggregate codelengths and normalize the positive gain
$\ell_\Pi(y) \leftarrow \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \ell_c(y)$
$\ell_\Pi(y \mid x) \leftarrow \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \ell_c(y \mid x)$
$\widehat{i}_\Pi(x; y) \leftarrow \ell_\Pi(y) - \ell_\Pi(y \mid x)$
$G_\Pi(x \to y) \leftarrow [\widehat{i}_\Pi(x; y)]_+$
if $\ell_\Pi(y) > 0$ **then**
  $R^{\Pi}_{\mathrm{HC}}(x, y) \leftarrow G_\Pi(x \to y) / \ell_\Pi(y)$
else
  $R^{\Pi}_{\mathrm{HC}}(x, y) \leftarrow 0$
end if
return $R^{\Pi}_{\mathrm{HC}}(x, y)$

B Evaluation Data Construction and Processing

This section details the collection and processing of the evaluation data. Table 1 summarizes the resulting dataset across four content domains and different levels of human-provided information.

B.1 Evaluation Data Collection

We construct the evaluation dataset from four content domains, comprising 56,110 source examples in total.

- **Arxiv.** We randomly sample 14,685 paper abstracts from the *arxivabstracts-2021* dataset [2]. Each example contains the original paper abstract as source content. The source abstracts contain 134.24 ± 63.07 words on average.
- **News.** We sample 14,280 articles from the *News Articles* dataset [3]. Each example contains the original news article as source content. The articles contain 532.85 ± 86.26 words on average.
- **Patent.** We sample 13,800 patent abstracts from the *HUPD 2018* dataset [4]. Each example contains the original patent abstract as source content, with an average length of 171.65 ± 24.22 words.
- **Poetry.** We sample 13,345 Poetrys from the *Poetry Foundation* dataset [5]. Each example contains the original Poetry as source content, with an average length of 208.18 ± 94.08 words. We do not construct an additional subject representation for Poetrys.

B.2 Evaluation Data Processing

To simulate progressively decreasing levels of human-provided information, we construct multiple representations of each source document. The resulting representations form an information hierarchy from complete source content to increasingly compressed semantic descriptions.

For *Polishing* and *Rewriting*, the complete source text is retained as input. Consequently, these two conditions have the same input length as the original source content: 134.24 ± 63.07 words for Arxiv, 532.85 ± 86.26 for News, 171.65 ± 24.22 for Patent, and 208.18 ± 94.08 for Poetry.

For *Bullets*, we use Claude Opus 4.6 to condense each source into a set of salient key points. The resulting bullet representations contain 68.38 ± 20.59, 78.36 ± 16.76, 59.89 ± 13.85, and 48.20 ± 11.27 words on average for Arxiv, News, Patent, and Poetry, respectively.

Table 1: Detailed statistics of the constructed dataset.

Type	# Samples	Polishing Words	Rewriting Words	Bullets Words	Summary Words	Subject Words
Arxiv	14,685	134.24 ± 63.07	134.24 ± 63.07	68.38 ± 20.59	28.74 ± 4.16	2.87 ± 1.37
News	14,280	532.85 ± 86.26	532.85 ± 86.26	78.36 ± 16.76	31.26 ± 3.84	4.06 ± 1.32
Patent	13,800	171.65 ± 24.22	171.65 ± 24.22	59.89 ± 13.85	26.91 ± 4.57	3.91 ± 0.79
Poetry	13,345	208.18 ± 94.08	208.18 ± 94.08	48.20 ± 11.27	24.38 ± 3.69	3.81 ± 0.79

For *Summary*, the source is further compressed into a concise summary. The resulting summaries contain 28.74 ± 4.16 words for Arxiv, 31.26 ± 3.84 for News, 26.91 ± 4.57 for Patent, and 24.38 ± 3.69 for Poetry.

Finally, for *Subject*, we construct a short subject-level description that retains only the central semantic topic of the source. The resulting subject representations contain 2.87 ± 1.37 words for

[2] https://huggingface.co/datasets/gfissore/arxiv-abstracts-2021

[3] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/GMFCTR

[4] https://huggingface.co/datasets/HUPD/hupd/blob/main/data/2018.tar.gz

[5] https://www.kaggle.com/datasets/tgdivy/poetry-foundation-poems

Arxiv, 4.06 ± 1.32 for News, and 3.91 ± 0.79 for Patent. No additional subject representation is constructed for Poetry.

These representations therefore provide a progressively compressed information hierarchy:

$$\textit{Polishing} > \textit{Rewriting} > \textit{Bullets} > \textit{Summary} > \textit{Subject}.$$

B.3 Prompt Construction

We construct generation conditions corresponding to the information levels described above. Let C denote the complete source content. Intermediate representations are constructed using Claude Opus 4.6 and are shared across generating models to ensure that each model is evaluated under the same input-information conditions.

- **Polishing.** The generating model receives the complete source content C:
  Please polish the following text to make it more eloquent and professional while preserving all original information: $\{C\}$.
- **Rewriting.** The generating model also receives the complete source content C, but is instructed to reformulate its expression:
  Please rewrite the following text in your own words while preserving all original facts and information: $\{C\}$.
- **Bullets.** Claude Opus 4.6 first extracts a concise set of salient key points from C. The generating model receives only the resulting bullet representation:
  Please expand the following bullet points into a complete, well-structured text: $\{\textit{Bullets}\}$.
- **Summary.** Claude Opus 4.6 further condenses C into a concise summary. The generating model receives only this summary:
  Please expand the following summary into a complete, well-structured text: $\{\textit{Summary}\}$.
- **Subject.** For Arxiv, News, and Patent, Claude Opus 4.6 reduces C to a short subject-level description. The generating model receives only this representation:
  Please write a complete text about the following subject: $\{\textit{Subject}\}$.

The resulting conditions progressively reduce the amount of source information explicitly available to the generating model. Polishing and Rewriting preserve the complete source content, whereas Bullets, Summary, and Subject provide increasingly compressed representations. This controlled information hierarchy enables systematic evaluation of human informational contribution under different levels of human guidance.

Table 2: Distribution of content generation samples across categories and models.

Content Generation Domain	Opus	Sonnet	Gemini	GPT-5.5	GPT-5.6
Professional Knowledge	799	500	1,297	997	1,297
Creative Generation	774	499	1,298	998	1,298
Technical Implementation	1,199	898	1,299	999	1,299
Reasoning and Analysis	650	499	1,059	908	1,059
Experiential Context	475	499	1,299	798	1,299
Artistic Design	90	49	317	261	314
Overall	**3,987**	**2,944**	**6,569**	**4,961**	**6,566**

B.4 Human Annotation

To provide an independent human reference for evaluating HCEF, we conduct a pairwise human annotation study over samples from the constructed human–AI collaborative content generation dataset. Candidate examples are collected from five generating models across four content domains and all generation modes. We construct candidate pairs from these examples and retain 7,500 pairs whose HCEF scores differ by more than 0.1. This threshold ensures that the paired examples exhibit a

meaningful difference according to HCEF and reduces ambiguity caused by nearly indistinguishable contribution scores. For each selected pair, annotators are presented with two prompt–output examples and asked to determine which example reflects greater human contribution to the generated content. As shown in Figure 8, the annotation interface presents the paired examples and allows annotators to select the one exhibiting greater human contribution. The annotation task focuses on the extent to which the human-provided input contributes substantive information to the corresponding output, rather than on its style, length, or tone. Each pair is independently evaluated by nine human annotators, ensuring that the final judgment does not depend on a single individual assessment. We aggregate the nine independent annotations for each pair using majority voting. Because each pair receives an odd number of annotations, a unique consensus preference can be obtained without ties. This aggregated human judgment serves as the reference ordering for the pair. We then compare the human-preferred ordering with the ordering induced by the corresponding HCEF scores. A pair is counted as agreement when the example assigned the higher HCEF score is also judged by the annotator consensus to exhibit greater human contribution.

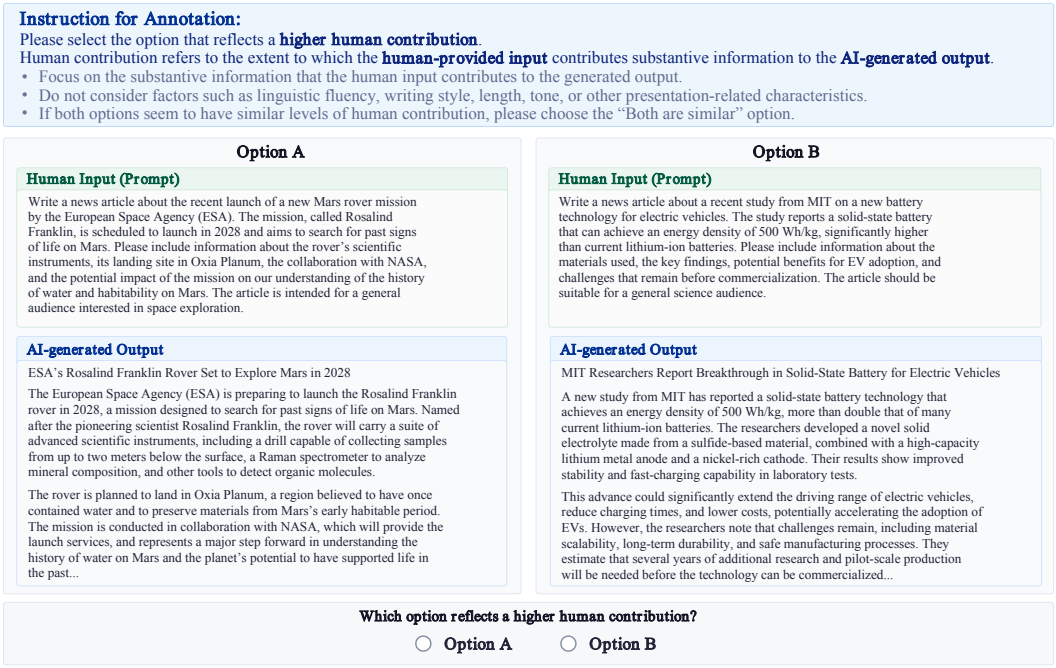

Figure 8: Illustration of the human annotation interface for human contribution evaluation.

Across the 7,500 annotated pairs, HCEF agrees with the aggregated human judgments on 96.93% of the comparisons. This high agreement indicates that the relative contribution ordering induced by HCEF closely matches human perception. Moreover, because the annotated pairs cover multiple content domains, generation modes, and generating models, the result provides additional evidence that the alignment between HCEF and human judgments is not restricted to a particular content type or generation setting. The annotation study therefore complements the controlled contribution-level experiments by directly evaluating whether the numerical ordering produced by HCEF corresponds to human perceptions of contribution. Together, these results support HCEF as a reliable continuous measure of human contribution in human–AI collaborative content generation.

C Further details of the experimental results

This section provides further details and supplementary analyses of the experimental results reported in the main text. We include extended comparisons with white-box baselines, additional ablation results, evaluations on real-world user prompts, and other supporting analyses to provide a more comprehensive assessment of HCEF.

C.1 Detailed Results for Human Contribution Evaluation

This experiment examines whether HCEF consistently captures the predefined hierarchy of human contribution across content domains and frontier generating models. As shown in Figure 9, when aggregated across all model–domain combinations, the mean HCEF score decreases strictly from 0.544 at *Polishing* to 0.358, 0.249, 0.136, and 0.065 at *Rewriting*, *Bullets*, *Summary*, and *Subject*, respectively. This strict monotonic ordering closely follows the progressive reduction in human-provided information across the five generation levels. The overall trend remains consistent across content domains and generating models, indicating that HCEF responds systematically to changes in input informativeness rather than to a particular domain or model. The clear separation across contribution levels further demonstrates the sensitivity of HCEF to graded differences in human contribution. Overall, these results support HCEF as a stable and generalizable measure for continuously measuring human contribution across diverse human–AI collaborative generation settings.

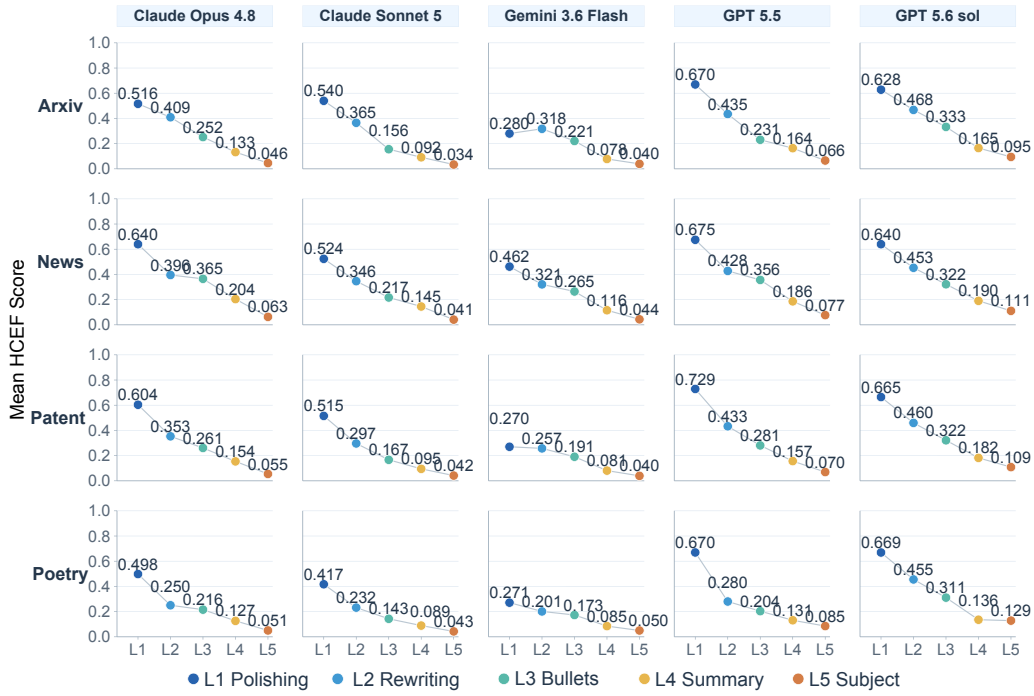

Figure 9: HCEF consistently captures the predefined human contribution hierarchy across content domains and frontier generating models under the evaluated generation settings, with aggregated scores decreasing strictly across the five contribution levels.

C.2 HCEF Performance on Real-World User Prompts

To examine the robustness of HCEF on real-world user prompts, we report results on 10k_prompts and OASST1, as shown in Figure 10. For each dataset, we evaluate the mean HCEF scores across five contribution levels, including *Polishing*, *Rewriting*, *Bullets*, *Summary*, and *Subject*, using five frontier generating models. This analysis examines whether the contribution hierarchy captured by HCEF remains consistent across datasets and independently generated outputs. On 10k_prompts, all five models exhibit a strictly decreasing pattern in HCEF scores as the amount of human-provided information decreases. Averaged across models, the scores decline from 0.3403 for *Polishing* to 0.1771, 0.0967, 0.0679, and 0.0426 for *Rewriting*, *Bullets*, *Summary*, and *Subject*, respectively. This monotonic ordering is consistently preserved across Claude Opus 4.8, Claude Sonnet 5, Gemini 3.6 Flash, GPT-5.5, and GPT 5.6 Sol. A similar pattern is observed on OASST1. The model-averaged HCEF scores decrease from 0.3382 to 0.1770, 0.1037, 0.0692, and 0.0440 across the five contribution

levels, with all generating models preserving the same ordering. For instance, the scores for Claude Opus 4.8 decrease from 0.4019 under *Polishing* to 0.0416 under *Subject*, while those for GPT 5.6 Sol decrease from 0.3308 to 0.0586. Overall, the consistent monotonic trends across both datasets and all five frontier models demonstrate that HCEF preserves the evaluated contribution hierarchy at the group level, supporting its consistency across real-world prompt sources and generating models.

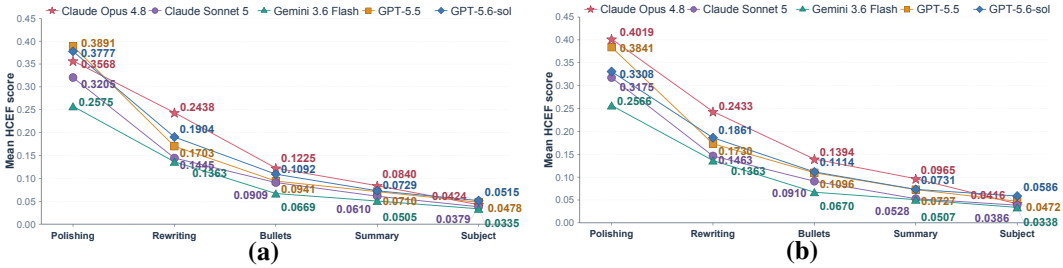

Figure 10: **(a)** Results on 10k_prompts. HCEF scores decrease monotonically across the five contribution levels for all frontier generating models, demonstrating stable preservation of the human contribution hierarchy. **(b)** Results on OASST1. HCEF exhibits the same monotonic pattern across all frontier generating models, further supporting its robustness on real-world user prompts.

C.3 Additional Results for White-Box Baseline Comparison

In the era of LLMs, the utility of an evaluation method depends not only on its effectiveness but also on its computational cost and latency. We therefore examine how output length and evaluator model scale affect evaluation latency, highlighting the efficiency implications of white-box probability-based evaluation relative to HCEF. As shown in Figure 11, the latency of white-box methods increases substantially with both output length and evaluator scale. This growth reflects the need to perform probability-based model inference over increasingly long outputs, with larger evaluators incurring considerably higher computational overhead. In contrast, HCEF's evaluation latency increases only slightly, from 33.91 ms to 35.60 ms, as output length increases from 100 to 1,200 tokens. At 1,200 tokens, the latency of the white-box evaluators ranges from 102.0 ms for Llama 3.1 8B to 421.2 ms for Llama3 42B, while HCEF remains at 35.60 ms. These results indicate that HCEF avoids the scaling cost associated with repeated probability evaluation, providing a substantially more stable and lightweight alternative as both text length and evaluator size increase.

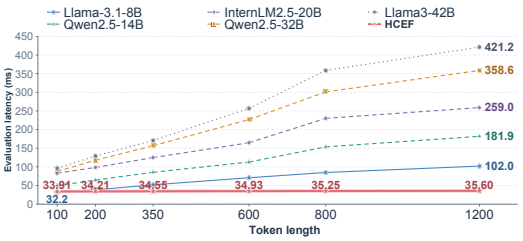

Figure 11: Evaluation latency under varying output lengths and white-box evaluator scales. HCEF remains nearly stable as token length increases, while probability-based white-box evaluation becomes substantially more expensive with longer outputs and larger models.

C.4 Additional Ablation Results

We examine the scoring design of HCEF through four complementary configurations evaluated within each of the five generation levels under the two settings shown in Figure 12. The compared configurations include unnormalized information gain G, byte-normalized gain $G/|u_y|$, the HCEF score $G/C(y)$, and an output-only control based on $C(y)$. This design isolates the effect of normalization while also testing whether output information content alone is sufficient to explain the reference contribution scores. Agreement with the white-box references is measured using Spearman's ρ and Krippendorff's interval α on standardized scores. Both metrics are computed within each generation level and macro-averaged with equal weights, thereby emphasizing contribution differences within the same generation condition rather than separation across predefined levels. HCEF consistently achieves the strongest agreement across both settings and both metrics. Its Spearman's ρ reach 0.7702 and 0.7603, while the corresponding Krippendorff's α values reach 0.7919 and 0.7817. In comparison, byte-length normalization yields agreement around 0.62–0.64, whereas unnormalized

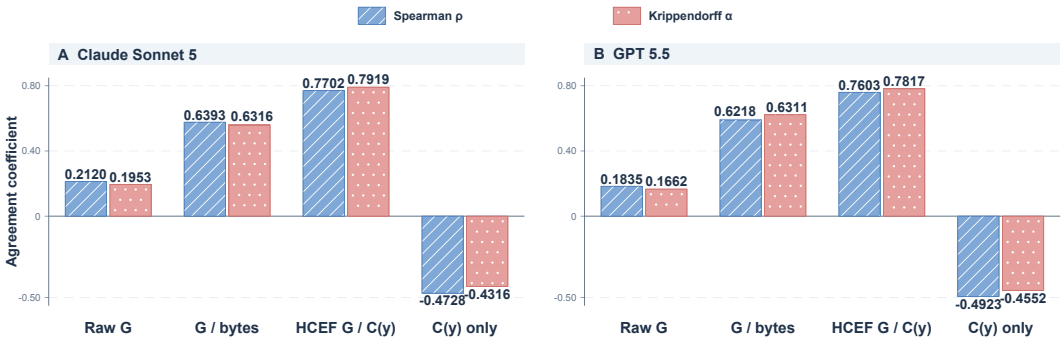

Figure 12: Comparison of unnormalized gain G, byte-normalized gain $G/|u_y|$, HCEF $G/C(y)$, and baseline information content $C(y)$ using macro-averaged within-level Spearman ρ and Krippendorff's α. HCEF achieves the highest agreement with reference scores across both model settings, supporting normalization by baseline output information content.

gain remains near 0.20. Relative to byte-length normalization, HCEF improves Spearman agreement by approximately 0.13 and Krippendorff's α by approximately 0.15–0.16. These results establish a consistent ordering of $G/C(y) > G/|u_y| > G$, showing that normalization by baseline output information provides substantially more than a correction for surface text length. The output-only control further clarifies the role of the baseline description cost. Using $C(y)$ alone produces negative agreement, with Spearman's ρ around -0.47 to -0.49 and Krippendorff's α around -0.43 to -0.46. Output information content therefore does not independently recover the contribution ordering. Instead, its value lies in providing a reference scale against which input-dependent information gain can be expressed. Taken together, these results support the central scoring principle of HCEF, which measures human contribution as the proportion of baseline output information explained by positive input-dependent gain. This formulation achieves substantially stronger agreement with the reference measures than absolute gain, byte-length normalization, or output information content alone.

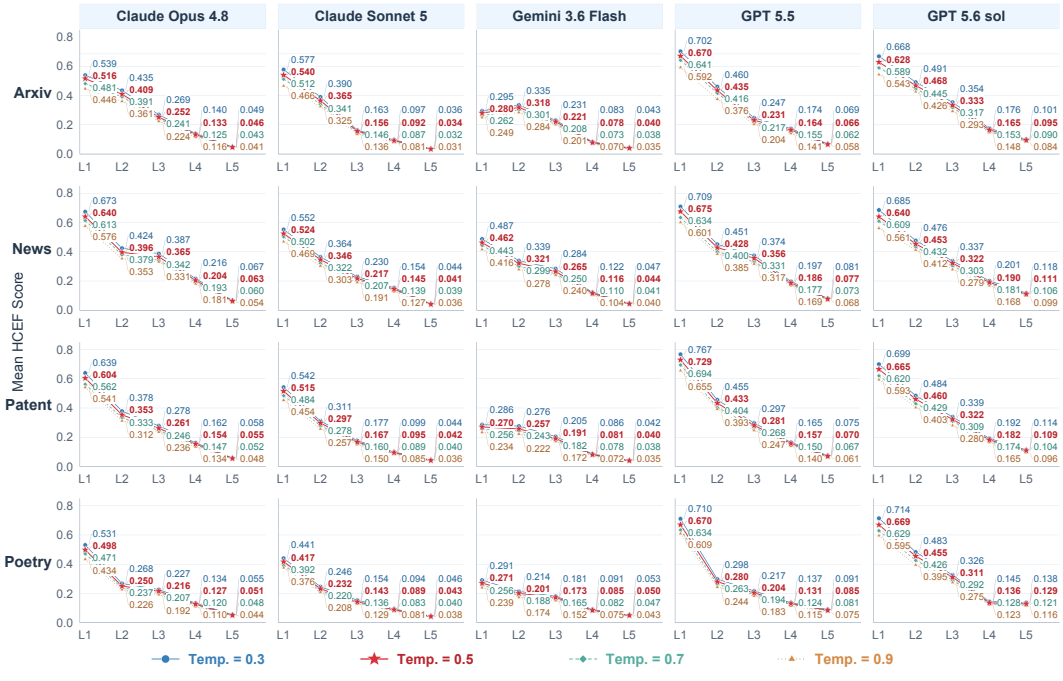

Figure 13: Robustness of HCEF to generation temperature. HCEF exhibits a stable downward trend with increasing temperature while preserving the relative contribution structure across models and domains. ★ denotes the values reported in the main text.

C.5 Effect of Generation Temperature

Generation temperature controls the stochasticity of model outputs by adjusting the concentration of the token sampling distribution. Lower temperatures favor high-probability continuations and produce more deterministic outputs, whereas higher temperatures introduce greater variation into the generation process. We examine whether HCEF remains effective under such changes in generation stochasticity while reflecting the corresponding variation in human contribution. Across temperature settings, the human input is held fixed while only the randomness of generation changes. As temperature increases, generated content becomes less constrained by the information provided in the prompt and contains more model-induced variation. The informational contribution attributable to human input is therefore expected to decrease. Since HCEF quantifies contribution through the information gain that the prompt provides for describing the generated output, its scores should reflect this weakening input–output dependence. Empirically, HCEF scores exhibit a stable downward trend as temperature increases across content domains and generation conditions. Importantly, this overall decline does not disrupt the underlying contribution structure, as the relative differences across contribution levels remain consistently preserved. These results indicate that HCEF is sensitive to changes in the informational influence of human input while remaining robust and effective under varying generation stochasticity.

D Theoretical Analysis of HCEF

This section analyzes the operational definition of human informational contribution in HCEF through paired coding, baseline normalization, and aggregation across coders. We first formalize the descriptive value of side information, then derive properties of the finite score and bounds on its sensitivity. Finally, we identify the conditions under which the score approximates a probabilistic reference. The finite-score results hold for the effective codelengths defined in Section 3; they do not require access to the generating distribution or assume that complete-stream differences form uniquely decodable conditional codes. Throughout, the coding protocol $\Pi = (\mathsf{U}, \mathsf{F}, \mathcal{C}, \eta)$ is fixed, and $m = |\mathcal{C}| \geq 1$. For a record (x, y), write

$$a_c = \ell_c(y), \qquad b_c = \ell_c(y \mid x), \qquad A = \frac{1}{m}\sum_c a_c, \qquad B = \frac{1}{m}\sum_c b_c. \tag{12}$$

Thus $a_c, b_c, A, B \geq 0$, $G_\Pi = [A - B]_+$, and $R^\Pi_{\mathrm{HC}}(x, y) = f(A, B)$, where

$$f(a, b) = \begin{cases} [1 - b/a]_+, & a > 0, \\ 0, & a = 0. \end{cases} \tag{13}$$

Here, G_Π is the absolute informational contribution under Π, whereas R^Π_{HC} expresses its proportion of the output baseline. All logarithms are base two. Probabilistic statements below concern an explicitly specified reference distribution q; algebraic statements concern the fixed protocol Π.

D.1 Side Information and Output-Relevant Gain

The measurement principle in Section 3.1 compares descriptions of the same output under two information conditions. The following standard identities establish what this comparison measures under ideal probabilistic coding (Shannon, 1948; Cover & Thomas, 2006).

Lemma D.1 (Reference information gain)**.** *Let (X, Z, Y) be discrete random variables with reference joint distribution q and $H_q(Y) < \infty$. On the support of q, define $j_q(x; y) = \log_2(q(y \mid x)/q_Y(y))$. Then*

$$\begin{aligned} \mathbb{E}_q[j_q(X;Y)] &= I_q(X;Y) = H_q(Y) - H_q(Y \mid X) \\ &= \mathbb{E}_{X \sim q_X} D_{\mathrm{KL}}\big(q_{Y|X}(\cdot \mid X) \,\|\, q_Y\big). \end{aligned} \tag{14}$$

Additional side information satisfies

$$\begin{aligned} j_q((x, z); y) - j_q(x; y) &= \log_2 \frac{q(y \mid x, z)}{q(y \mid x)}, \\ I_q((X, Z); Y) - I_q(X; Y) &= I_q(Z; Y \mid X) \geq 0. \end{aligned} \tag{15}$$

If $Y \perp Z \mid X$ under q, both increments are zero on the corresponding support and in expectation, respectively.

Proof. The finite-entropy assumption ensures integrability, since $\mathbb{E}_q|j_q(X;Y)| \le H_q(Y) + H_q(Y \mid X) \le 2H_q(Y) < \infty$. Expanding the expected log ratio gives

$$\begin{aligned}\mathbb{E}_q[j_q(X;Y)] &= \sum_{x,y} q(x,y) \log_2 \frac{q(y \mid x)}{q_Y(y)} \\ &= H_q(Y) - H_q(Y \mid X).\end{aligned} \tag{16}$$

Grouping the same sum by x yields the expected KL divergence. For the additional input Z, insert $q(y \mid x)$ into the likelihood ratio:

$$\log_2 \frac{q(y \mid x,z)}{q_Y(y)} = \log_2 \frac{q(y \mid x)}{q_Y(y)} + \log_2 \frac{q(y \mid x,z)}{q(y \mid x)}.$$

Taking expectations gives the conditional mutual information. Under conditional independence, the last ratio equals one wherever $q(x,z,y) > 0$. □

The lemma identifies output relevance as the source of additional descriptive value: information in Z contributes only through its relationship with Y after conditioning on X. This is a reference-distribution statement. Its connection to finite coding is made explicit in Section D.5. The non-negativity in Equation equation 15 holds in expectation; the incremental information density for an individual record can be negative. In particular, a deterministic function of X supplies zero additional conditional mutual information about Y. Appending a deterministic copy of the input therefore adds no output-relevant information under q, even when the recorded input becomes longer. The distinction between signed and positive gain is also important. Since $[v]_+ = v + [-v]_+$,

$$\mathbb{E}_q\big[[j_q(X;Y)]_+\big] = I_q(X;Y) + \mathbb{E}_q\big[[-j_q(X;Y)]_+\big]. \tag{17}$$

Thus the mutual-information identity applies to the signed information density. Positive-part truncation retains beneficial description savings at the instance level, but changes the corresponding expectation.

D.2 Paired Coding and Baseline Normalization

The paired construction evaluates the output net of the cost of its context. To see the consequence, consider two prompts x, x' and the same output y. For each coder, define

$$h_c = L_c(P_{x'}) - L_c(P_x), \qquad k_c = L_c(P_{x'}\|u_y) - L_c(P_x\|u_y).$$

Their untruncated conditional increments differ by $k_c - h_c$. Consequently, if $k_c = h_c$ for every coder, the conditional effective codelengths and the final score are unchanged. Context cost that affects both terms equally therefore receives no additional credit. This cancellation is exact, although a textual modification can also change the output's incremental coding cost through the coder's context-dependent behavior. More generally, with $b'_c = \ell_c(y \mid x')$, the 1-Lipschitz property of positive-part truncation gives $|b'_c - b_c| \le |k_c - h_c|$. This links paired cancellation to the sensitivity analysis below.

Proposition D.2 (Normalization properties)**.** *For every finite record evaluated under* Π*:*

1. $0 \le R^{\Pi}_{\mathrm{HC}}(x,y) \le 1$. *For $A > 0$, the score equals zero if and only if $B \ge A$, and equals one if and only if $B = 0$.*
2. *An empty prompt or empty output receives score zero.*
3. *For a fixed output and protocol, the score is nonincreasing in the conditional effective codelength B.*
4. *A common positive rescaling of all effective codelengths leaves the score unchanged: $f(\lambda A, \lambda B) = f(A,B)$ for every $\lambda > 0$.*

Proof. For $A > 0$, nonnegativity of B gives $0 \le [1 - B/A]_+ \le 1$, with the stated equality conditions. The zero-baseline branch is zero by definition. If x is empty, $P_x = P_\varnothing$, so $a_c = b_c$ for every coder. If y is empty, appending $u_y = \epsilon$ changes neither stream, so $a_c = b_c = 0$. For fixed y and Π, A is fixed and $[1 - B/A]_+$ is nonincreasing in B when $A > 0$; when $A = 0$, the score is constant at zero. Finally, for $A > 0$, $[\lambda A - \lambda B]_+/(\lambda A) = [A - B]_+/A$; the equality also holds when $A = 0$. □

The score is therefore dimensionless: a consistent conversion of all effective codelengths from bits to bytes leaves it unchanged. For $A > 0$ and any threshold $\tau \in (0, 1]$, its definition also gives

$$R^{\Pi}_{\mathrm{HC}}(x, y) \geq \tau \quad \Longleftrightarrow \quad B \leq (1 - \tau)A. \tag{18}$$

Since $\tau > 0$, the positive-part operation does not change this condition. Reaching threshold τ therefore requires saving at least a fraction τ of the output's baseline description cost.

Baseline normalization has a direct relationship to the alternatives examined in the ablation study. For a nonempty output with $n_y = |u_y| > 0$,

$$G_{\Pi} = A\, R^{\Pi}_{\mathrm{HC}}(x, y), \qquad \frac{G_{\Pi}}{n_y} = \frac{A}{n_y}\, R^{\Pi}_{\mathrm{HC}}(x, y). \tag{19}$$

These identities also hold when $A = 0$, since then $G_{\Pi} = 0$. Absolute gain retains the scale of the output baseline, while byte-normalized gain retains the baseline coding density A/n_y. In particular, two records with the same relative description savings can have different byte-normalized scores when their baseline coding densities differ. HCEF expresses the gain directly as a fraction of the output baseline. Equation equation 19 explains the distinct quantities measured by the three variants; their agreement with external references is assessed empirically.

D.3 Aggregation Across Coding Mechanisms

HCEF averages effective codelengths before taking the positive gain and forming a ratio. The next result gives an exact characterization of the induced weighting and of the effect of retaining negative coder-specific gains until aggregation. Here, truncation concerns the signed gains $a_c - b_c$; the incremental codelengths a_c, b_c have already been truncated to be nonnegative.

Theorem D.3 (Baseline weighting and gain cancellation)**.** *Let $S = \sum_c a_c > 0$, $\mathcal{C}_+ = \{c : a_c > 0\}$, and $\mathcal{C}_0 = \{c : a_c = 0\}$. For $c \in \mathcal{C}_+$, define*

$$w_c = \frac{a_c}{S}, \qquad r_c = 1 - \frac{b_c}{a_c}.$$

These weights satisfy $\sum_{c \in \mathcal{C}_+} w_c = 1$. Then

$$R^{\Pi}_{\mathrm{HC}}(x, y) = \left[\sum_{c \in \mathcal{C}_+} w_c r_c - \frac{\sum_{c \in \mathcal{C}_0} b_c}{S} \right]_+ . \tag{20}$$

Moreover, let

$$D_+ = \sum_c [a_c - b_c]_+, \qquad D_- = \sum_c [b_c - a_c]_+,$$

and define the alternative that clips each coder's gain first, using the same baseline weights, as $R_{\mathrm{pre}} = \sum_{c \in \mathcal{C}_+} w_c [r_c]_+$. The two scores satisfy

$$\begin{aligned} R^{\Pi}_{\mathrm{HC}} &= \frac{[D_+ - D_-]_+}{S}, \\ R_{\mathrm{pre}} - R^{\Pi}_{\mathrm{HC}} &= \frac{\min\{D_+, D_-\}}{S} \geq 0. \end{aligned} \tag{21}$$

Proof. The common factor $1/m$ cancels between numerator and denominator, giving

$$R^{\Pi}_{\mathrm{HC}} = \frac{[\sum_c (a_c - b_c)]_+}{S}.$$

For $a_c > 0$, $(a_c - b_c)/S = w_c r_c$; for $a_c = 0$, the same contribution is $-b_c/S$. This proves Equation equation 20. Using $v = [v]_+ - [-v]_+$ gives $\sum_c (a_c - b_c) = D_+ - D_-$. Zero-baseline coders have no positive gain, so $R_{\mathrm{pre}} = D_+/S$. Finally, $D_+ - [D_+ - D_-]_+ = \min\{D_+, D_-\}$, which proves Equation equation 21. □

Consequently, $R_{\mathrm{pre}} > R_{\mathrm{HC}}^{\Pi}$ if and only if $D_+ > 0$ and $D_- > 0$: positive and negative coder-specific gains must coexist. Otherwise, the two scores coincide. If $S = 0$, $R_{\mathrm{HC}}^{\Pi} = 0$ by definition, and we likewise define $R_{\mathrm{pre}} = 0$.

When all baselines are positive, the score is the positive part of a baseline-weighted mean of signed relative gains. If these gains all equal r, the aggregate score reduces to $[r]_+$, irrespective of differences in baseline magnitudes. A coder's positive contribution to this weighted sum is at most w_c, since $[r_c]_+ \leq 1$; a small-baseline coder therefore has a proportionally small upper bound on its positive contribution. Negative contributions remain included, as the zero-baseline term in Equation equation 20 makes explicit. The exact gap in Equation equation 21 shows how clipping each coder's signed gain before aggregation discards countervailing negative gains and weakly increases the resulting score. Aggregating first preserves this cancellation within the specified collection of coders. This is an algebraic property of the scoring rule; statistical improvements from combining particular coders require empirical evidence.

D.4 Sensitivity to Codelength Perturbations

To characterize the resolution of the normalized score, we next bound how changes in effective codelengths propagate through aggregation and normalization. This also provides a sufficient condition for preserving the ordering of two evaluated records.

Theorem D.4 (Finite perturbation bound)**.** *Let $A, B, \widetilde{A}, \widetilde{B} \geq 0$ and $\max\{A, \widetilde{A}\} > 0$. Then*

$$\left| f(A,B) - f(\widetilde{A},\widetilde{B}) \right| \leq \min\left\{ 1, \frac{|A-\widetilde{A}| + |B-\widetilde{B}|}{\max\{A,\widetilde{A}\}} \right\}. \tag{22}$$

If both baselines vanish, both scores are zero.

Proof. First suppose both baselines are positive. For a fixed $a > 0$, the positive-part function is 1-Lipschitz, and hence

$$|f(a,b) - f(a,b')| \leq \frac{|b-b'|}{a}. \tag{23}$$

For fixed $b \geq 0$ and $a \geq a' > 0$,

$$|f(a,b) - f(a',b)| \leq \frac{a-a'}{a}. \tag{24}$$

Indeed, when $b \leq a'$, the difference is $b(a-a')/(aa') \leq (a-a')/a$; when $a' < b < a$, it is $1 - b/a \leq (a-a')/a$; and when $b \geq a$, it is zero. Without loss of generality, assume $A \geq \widetilde{A} > 0$. Changing the conditional quantity at the larger baseline gives

$$\begin{aligned} |f(A,B) - f(\widetilde{A},\widetilde{B})| &\leq |f(A,B) - f(A,\widetilde{B})| + |f(A,\widetilde{B}) - f(\widetilde{A},\widetilde{B})| \\ &\leq \frac{|B-\widetilde{B}| + |A-\widetilde{A}|}{A}. \end{aligned}$$

The score range supplies the additional bound of one. If exactly one baseline is zero, the fraction on the right-hand side of Equation equation 22 is at least one, so the score-range bound proves the result. □

The bound is sharp for the algebraic score map. For $a > 0$ and $0 < h < a$, take $(A,B) = (a,a)$ and $(\widetilde{A},\widetilde{B}) = (a+h, a-h)$. Then

$$\begin{aligned} |f(A,B) - f(\widetilde{A},\widetilde{B})| &= \frac{2h}{a+h} \\ &= \frac{|A-\widetilde{A}| + |B-\widetilde{B}|}{\max\{A,\widetilde{A}\}} < 1. \end{aligned} \tag{25}$$

Corollary D.5 (Propagation of coder-level errors)**.** *Suppose an alternative evaluation uses the same representations, framing, and coder collection. For each coder c, assume the effective codelengths*

of each of $P_\varnothing$, $P_\varnothing \| u_y$, P_x, and $P_x \| u_y$ change by at most ε_c bits. Let $\bar{\varepsilon} = m^{-1} \sum_c \varepsilon_c$. For an original baseline $A > 0$,

$$\left| R^{\Pi}_{\mathrm{HC}} - \widetilde{R} \right| \le \min \left\{ 1, \frac{4\bar{\varepsilon}}{A} \right\}. \tag{26}$$

Proof. Each untruncated increment is a difference of two effective codelengths, so its change is at most $2\varepsilon_c$. Positive-part truncation cannot increase this change. Averaging gives $|A - \widetilde{A}| \le 2\bar{\varepsilon}$ and $|B - \widetilde{B}| \le 2\bar{\varepsilon}$. Apply Theorem D.4 and use $\max\{A, \widetilde{A}\} \ge A$. □

For a fixed output and baseline, the sharper specialization is

$$\begin{aligned} &|R^{\Pi}_{\mathrm{HC}}(x, y) - R^{\Pi}_{\mathrm{HC}}(x', y)| \\ &\quad \le \min \left\{ 1, \frac{|\ell_\Pi(y \mid x) - \ell_\Pi(y \mid x')|}{\ell_\Pi(y)} \right\}, \qquad \ell_\Pi(y) > 0. \end{aligned} \tag{27}$$

This identifies the conditional description change, measured relative to the baseline, as the quantity controlling score sensitivity. It does not presume that a given textual edit induces a small codelength change. For the same output and protocol with $A > 0$, the paired differences h_c, k_c from Section D.2 further give

$$\begin{aligned} &\left| R^{\Pi}_{\mathrm{HC}}(x', y) - R^{\Pi}_{\mathrm{HC}}(x, y) \right| \\ &\quad \le \min \left\{ 1, \frac{1}{mA} \sum_{c \in \mathcal{C}} |k_c - h_c| \right\}. \end{aligned} \tag{28}$$

Indeed, averaging $|b'_c - b_c| \le |k_c - h_c|$ bounds the change in B; Equation equation 27 then yields the result. This connects score sensitivity directly to stream-length changes remaining after subtraction of the context costs. More generally, if two records have original scores R_1, R_2 and perturbation bounds $|R_i - \widetilde{R}_i| \le \eta_i$, then

$$R_1 - R_2 > \eta_1 + \eta_2 \quad \Longrightarrow \quad \widetilde{R}_1 > \widetilde{R}_2, \tag{29}$$

because $\widetilde{R}_1 - \widetilde{R}_2 \ge R_1 - R_2 - \eta_1 - \eta_2 > 0$. Thus a sufficiently large score margin preserves the ranking under bounded codelength perturbations. Over the nonnegative real domain, the baseline dependence cannot be removed: for every $\delta > 0$, $f(\delta, 0) = 1$ and $f(\delta, \delta) = 0$, although the conditional codelength difference tends to zero as $\delta \downarrow 0$. For a fixed finite coder, the relevant comparison is between the baseline and the magnitude of its codelength perturbation.

D.5 Connection to Probabilistic Reference Scores

The preceding properties hold directly for the finite coding protocol. We now connect them to the probabilistic perspective of Lemma D.1 by explicitly separating reference information from coding discrepancies. For a record in the support of q, let

$$s_q(y) = -\log_2 q_Y(y), \qquad t_q(x, y) = -\log_2 q(y \mid x). \tag{30}$$

On the support of q, define the coding discrepancies $e_0(x, y) = A - s_q(y)$ and $e_1(x, y) = B - t_q(x, y)$, including records with $s_q(y) = 0$. For $s_q(y) > 0$, define the reference normalized coding gain $r_q(x, y) = [s_q(y) - t_q(x, y)]_+ / s_q(y)$. This is a comparison quantity defined by q; it is not an additional input to the HCEF algorithm.

Theorem D.6 (Reference approximation under coding error)**.** *Fix a record with $q(x, y) > 0$. Write $s = s_q(y)$, $t = t_q(x, y)$, and define $e_0 = A - s$ and $e_1 = B - t$. Then*

$$\begin{aligned} \widehat{i}_\Pi(x; y) - j_q(x; y) &= e_0 - e_1, \\ |G_\Pi(x \to y) - [j_q(x; y)]_+| &\le |e_0 - e_1|. \end{aligned} \tag{31}$$

Whenever $s > 0$, the normalized scores satisfy

$$\left| R^{\Pi}_{\mathrm{HC}}(x, y) - r_q(x, y) \right| \le \min \left\{ 1, \frac{|e_0| + |e_1|}{\max\{A, s\}} \right\}. \tag{32}$$

In particular, if $s > 0$, $|e_0| \le \varepsilon_0$, and $|e_1| \le \varepsilon_1$, then

$$\left| R^{\Pi}_{\mathrm{HC}} - r_q \right| \le \min \left\{ 1, \frac{\varepsilon_0 + \varepsilon_1}{s_q(y)} \right\}. \tag{33}$$

Proof. Since $j_q = s - t$ and $\widehat{i}_\Pi = A - B$, subtracting gives $\widehat{i}_\Pi - j_q = e_0 - e_1$. The 1-Lipschitz property of the positive-part function gives the gain bound. For the normalized quantities, apply Theorem D.4 to (A, B) and (s, t). Finally, use $\max\{A, s\} \geq s$ and the assumed error bounds. □

At $A = s > 0$ and $B = t$, the signed gain and normalized score coincide exactly with their probabilistic references. The error bounds quantify departures from this correspondence when the effective codelengths differ.

When $H_q(Y) < \infty$ and $\mathbb{E}_q[A + B] < \infty$, both coding discrepancies are integrable, since

$$\mathbb{E}_q\big[|e_0| + |e_1|\big] \leq \mathbb{E}_q[A + B] + 2H_q(Y) < \infty. \tag{34}$$

The signed-error identity therefore gives

$$\mathbb{E}_q[\widehat{i}_\Pi(X;Y)] = I_q(X;Y) + \mathbb{E}_q[e_0 - e_1]. \tag{35}$$

If the two coding discrepancies have equal expectations, the expected signed gain equals the reference mutual information, even when neither discrepancy vanishes. Combining the pointwise gain bound with Equation equation 17 further yields

$$\big|\mathbb{E}_q[G_\Pi] - I_q(X;Y) - \mathbb{E}_q\big[[-j_q(X;Y)]_+\big]\big| \leq \mathbb{E}_q|e_0 - e_1|. \tag{36}$$

These expressions separate coding discrepancy from the effect of truncating negative gains. When $0 < H_q(Y) < \infty$, $s_q(Y) > 0$ on the support and $0 \leq r_q(X,Y) \leq 1$. Using $s_q(Y) r_q(X,Y) = [j_q(X;Y)]_+$ yields the exact decomposition

$$\mathbb{E}_q[r_q(X,Y)] = \frac{I_q(X;Y) + \mathbb{E}_q[[-j_q(X;Y)]_+] - \mathrm{Cov}_q\big(s_q(Y), r_q(X,Y)\big)}{H_q(Y)}. \tag{37}$$

Expanding $\mathbb{E}_q[s_q(Y) r_q(X,Y)]$ as $H_q(Y)\mathbb{E}_q[r_q(X,Y)]$ plus the covariance gives the identity. The covariance is finite because $s_q(Y)$ is integrable and $r_q(X,Y)$ is bounded. The decomposition separates positive-part truncation from the dependence between baseline surprisal and relative gain. Mean instance-level scores therefore need not equal normalized mutual information. Controlling the difference $e_0 - e_1$ suffices to control absolute gain, whereas a normalized comparison also depends on the accuracy of the baseline. For example, a common offset $A = s + d$, $B = t + d$, with $d > 0$, leaves the signed gain unchanged but decreases the normalized score whenever $s > t$.

The reference distribution cannot be identified uniquely from a single observed record. To illustrate why, let $q_X(x_0) = q_X(x_1) = 1/2$ and consider two binary-output channels. In the first, $q^{(1)}(y_0 \mid x_0) = q^{(1)}(y_0 \mid x_1) = 1/2$. In the second, $q^{(2)}(y_0 \mid x_0) = 1/2$ and $q^{(2)}(y_0 \mid x_1) = 0$. The same record (x_0, y_0) has joint probability $1/4$ in both cases, but its information density is respectively 0 and 1 bit, and its normalized reference gain is respectively 0 and $1/2$. The ambiguity comes from behavior at unobserved inputs, which changes the marginal output distribution while preserving the probability of the observed record. A function of that record alone cannot recover both values. Under a fixed Π, the finite score is fully specified and reproducible from observable texts. Theorem D.6 therefore states a conditional approximation guarantee: it identifies which codelength errors must be controlled to approximate a specified probabilistic reference. It does not assert that the finite coders recover an unknown generating distribution. Taken together, the results establish the score's range and null cases, the relative quantity selected by normalization, the exact behavior of coder aggregation, and finite perturbation guarantees. These properties provide the mathematical basis for the scoring design, while the experiments assess its alignment with contribution orderings, white-box references, and human judgments.