Journal Title
XX(X):1–16
©The Author(s) 2016
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Towards an Understanding and Explanation for Mixed-Initiative Artificial Scientific Text Detection

Luoxuan Weng[1], Shi Liu[1], Hang Zhu[1], Jiashun Sun[1], Kam Kwai Wong[2], Dongming Han[1], Minfeng Zhu[1] and Wei Chen[1]

Abstract

Large language models (LLMs) have gained popularity in various fields for their exceptional capability of generating human-like text. Their potential misuse has raised social concerns about plagiarism in academic contexts. However, effective artificial scientific text detection is a non-trivial task due to several challenges, including 1) the lack of a clear understanding of the differences between machine-generated and human-written scientific text, 2) the poor generalization performance of existing methods caused by out-of-distribution issues, and 3) the limited support for human-machine collaboration with sufficient interpretability during the detection process. In this paper, we first identify the critical distinctions between machine-generated and human-written scientific text through a quantitative experiment. Then, we propose a mixed-initiative workflow that combines human experts' prior knowledge with machine intelligence, along with a visual analytics system to facilitate efficient and trustworthy scientific text detection. Finally, we demonstrate the effectiveness of our approach through two case studies and a controlled user study. We also provide design implications for interactive artificial text detection tools in high-stakes decision-making scenarios.

Keywords

Large Language Models, Mixed-Initiative, Visual Analytics, Explainable Artificial Intelligence

Introduction

The recent emergence of large language models (LLMs) has significantly enhanced the diversity, control, and quality of machine-generated text[1,2]. For instance, ChatGPT[3] gained tremendous public attention for its ability to generate plausible, coherent, and human-like text in a conversational manner (or *prompting*[4,5]). Nonetheless, the widespread accessibility of powerful LLMs has raised concerns regarding their potential misuse, including the propagation of fake news[6,7], fraudulent online reviews[8], and social media spam[9]. These concerns are particularly acute in academic and educational contexts[10,11]. Academic conferences and journals such as ACL and ICML have established strict guidelines for the cautious and responsible use of LLMs[12,13]. In this paper, we focus on the detection of scientific text, as there is an urgent need to prevent cheating and maintain academic integrity.

Recently, much attention has been devoted to artificial text detection in industry and academia. The growing demand for detecting machine-generated text has led to the development of several online tools (*e.g.*, GPTZero[14]). Meanwhile, academic research has proposed numerous detection methods, including feature-based[15,16] and transformer-based[11,17] models. While existing methods can achieve high accuracy in specific domains such as fake news detection[6], they are less effective when applied to detect machine-generated scientific text due to the following reasons.

Firstly, the critical distinctions between machine-generated and human-written scientific text remain underexplored. Prior findings[10,18,19] are either not specific to scientific text or lack comprehensive and quantitative user studies. Understanding these differences is crucial for designing effective detection algorithms and enhancing human experts' ability to identify suspect manuscripts.

Secondly, the performance of existing methods often lacks generalizability due to out-of-distribution (OOD) issues. Numerous LLMs can generate scientific text with varying feature distributions that are difficult to distinguish using a single detector, caused by diverse sampling methods of LLMs[15] or cross-domain adaptation issues[11]. Moreover, different detectors might produce incorrect and conflicting results when confronted with multiple scientific text of unknown origins.

Thirdly, the incorporation of human agency is limited. Recent studies[10,20] have found that expert human reviewers outperform automated detectors in identifying certain shortcomings of machine-generated scientific text. This suggests that integrating human experts' prior knowledge is promising for improving the fairness, interpretability, and reliability of artificial text detection, as argued in [1,20,21]. However, current approaches such as GLTR[22] lack sufficient support for human-machine collaboration and interpretation of machine learning (ML) models' decisions. In high-stakes decision-making scenarios like scientific text detection,

[1] State Key Lab of CAD&CG, Zhejiang University, CN
[2] The Hong Kong University of Science and Technology, HK

Corresponding author:
Wei Chen, State Key Lab of CAD&CG, Zhejiang University, Hangzhou, China.
Email: chenvis@zju.edu.cn

solely relying on models' decisions reduces the reliability and trustworthiness of the detection process.

To address the aforementioned challenges, we first conduct a formative study to understand the critical distinctions and the corresponding statistical features between machine-generated and human-written scientific text, and then formulate design requirements through user interviews. Our results show that, while knowledgeable human experts can holistically identify machine-generated scientific text, they seek evidence from ML models to support their decisions. Therefore, we propose a mixed-initiative workflow that combines human experts' prior knowledge with ML models trained using the identified features. Our workflow presents text excerpts based on feature similarities and ranks multiple ML models during each iteration to mitigate the OOD issue when handling multi-sourced text. It also provides multiple levels of feature analysis leveraging explainable artificial intelligence (XAI) techniques to enhance the interpretability and reliability of the human-machine detection process. Accordingly, we design and implement a visual analytics system. Case studies and a controlled user study demonstrate the capability of our approach in facilitating artificial scientific text detection.

In summary, the major contributions of this work include:

- A formative study that identifies critical distinctions and summarizes design requirements for artificial scientific text detection.
- A novel workflow and a visual analytics system that combines human intelligence with machine learning techniques to facilitate artificial text detection.
- Two case studies and a controlled user study to demonstrate the effectiveness of our proposed approach.

Related Work

Artificial Text Detection

Various techniques have been proposed to differentiate between machine-generated and human-written text. Crothers *et al.*[1] presented a comprehensive review of automatic detection of machine-generated text. Most existing work can be broadly categorized into two groups: feature-based methods and deep-learning-based methods.

Feature-based methods[8,15,16,18,23] utilize statistical distinctions between machine-generated and human-written text to train ML models for classification. These statistical features include basic features (*e.g.*, word and sentence length[10,15]), frequency features (*e.g.*, TF-IDF[24]), fluency features (*e.g.*, Gunning-Fog Index and Flesch Index[23]), *etc.* Perplexity, another metric used to measure the efficacy of a model in predicting the next word in a sequence, is also commonly adopted to identify machine-generated text[10,18]. However, feature-based methods are hindered by different sampling methods or model sizes[6,25], whereas our proposed approach incorporates multiple models with human expertise to address these limitations.

Deep-learning-based methods[7,11,17,20,26,27] utilize neural networks or language models to differentiate between machine-generated and human-written text. For instance, Zellers *et al.*[6] proposed the Grover model for generating and detecting fake news, demonstrating the potential of generative models in discrimination, which has also been observed in other studies[24,28]. Fine-tuning approaches, particularly those utilizing large bi-directional language models like RoBERTa, still represent the state-of-the-art for artificial text detection[1,24]. While cross-domain adaptation has shown significant improvement with a few hundred out-of-domain samples[11], collecting balanced training data for general-purpose detection models in real-life scenarios remains a significant challenge[1,21,29]. Moreover, the black-box nature of most deep learning models impedes their adoption in high-stakes decision-making scenarios like scientific text detection.

In addition to automatic detection approaches, empirical research[19,20,30,31] has investigated humans' ability of detecting machine-generated text. For instance, human experts achieve high precision in detecting certain errors like technical jargon[19] and identifying fake scientific abstracts[10]. Inspired by these findings and the concept of GLTR[22] which incorporates human agency to facilitate detection, we propose a mixed-initiative workflow that more effectively integrates human intelligence.

Explainable Artificial Intelligence

XAI techniques aim to enhance the interpretability and reliability of AI models, enabling humans to understand how these models make their decisions. In the visualization community, researchers have developed numerous XAI systems, such as RuleMatrix[32], CNNVis[33], and CNN Explainer[34]. Several surveys[35–42] have also been proposed to provide a comprehensive overview of XAI approaches, including intrinsic and post-hoc interpretability[43]. Intrinsic interpretability[44–46] refers to ML models that can be understood and explained directly from their design and architecture, such as decision trees and generalized additive algorithms. Visualization techniques have been proposed to facilitate intrinsic interpretability. For instance, Dingen *et al.*[47] presented RegressionExplorer to interactively explore logistic regression models in clinical biostatistics. Similarly, Neto *et al.*[48] proposed Explainable Matrix for interpreting complex random forest ensembles. On the other hand, post-hoc interpretability techniques[49–53] are applied after model training, specifically for black-box models like deep neural networks that are not intrinsically interpretable. Such methods are model-agnostic and can be utilized to explain the decisions of any models. One of the typical approaches is to train a surrogate model to approximate the outputs of the original black-box model[43]. For instance, LIME[51] trains local surrogate models to explain individual predictions. Similarly, SHAP (SHapley Additive exPlanations)[50], which is based on game theory[54], has been widely applied in various XAI systems[55–57]. In this work, we leverage SHAP values to offer contribution-based explanations for models' decisions in scientific text detection.

Mixed-Initiative Visual Analytics Systems

Initially introduced by Horvitz[58], mixed-initiative systems aim to enhance collaboration between humans and machines in decision-making processes. In recent years, there has been a significant amount of work in the visualization field related to mixed-initiative visual analytics systems[59–69]. These systems leverage both human and machine intelligence to explore and analyze complex data utilizing innovative visual designs [70–76]. For instance, Wall *et al.*[77] proposed the Podium system that enabled users to rank multi-attribute data based on their holistic understanding and explore their subjective preferences. Our work is also based on a similar assumption derived from our formative study, that knowledgeable human experts can only holistically identify machine-generated scientific text and require evidence from ML models to support their decisions. Additionally, Pister *et al.*[78]'s work on integrating prior knowledge for social network clustering has motivated our proposed workflow. Unlike clustering, we focus on the binary classification task of detecting artificial text.

Our work involves a binary labeling process, which is similar to previous works on mixed-initiative labeling systems. For instance, Felix *et al.*[79] proposed an interactive visual data analysis method to facilitate document labeling with machine recommendations. Similarly, Choi *et al.*[80] developed the Attentive Interactive Labeling Assistant to visually highlight words and improve the efficiency of document labeling. More recently, Alsaid *et al.*[81] integrated dimensionality reduction to develop an interactive method for labeling video and image data more efficiently and effectively. Furthermore, the visualization community has explored visual-interactive labeling which combines VA with ML techniques[82–86]. In contrast to prior works that primarily focused on enhancing the effectiveness and efficiency of the labeling process, we extend this line of research by prioritizing the reliability of high-stakes decision-making scenarios like scientific text detection.

Formative Study

Our workflow is designed for proficient researchers with a deep understanding of their respective disciplines, typically tasked with the critical appraisal and review of research papers. We collected paired datasets related to visualization research, and collaborated with visualization experts, including senior PhD students, PostDocs, and editors of relevant journals. The formative study spanned 4 months, during which we conducted bi-weekly semi-structured interviews with the experts to comprehend their primary concerns regarding existing automatic detection methods for scientific text. Through quantitative experiments, we identified key statistical features that can distinguish machine-generated text from human-written text. We consolidated the design requirements and progressively refined them based on user feedback to guide the design and development of our proposed workflow and system.

Experiment Design

Data Collection. We define *text excerpts* as meaningful segments of a complete article, such as subsections or paragraphs. Specifically, we utilize *scientific abstracts* in this work, due to their brevity and ability to encapsulate a paper's key points. We extracted the titles and original abstracts from the dataset provided by Narechania *et al.*[87]. We then used the metadata to generate abstracts through the official APIs for ChatGPT and GPT-3 with the prompt *'suppose you are the author, write a short abstract for the scientific paper titled with TITLE'*. Given GPT-2's comparatively weaker generative capabilities, we used the first sentence of each original abstract as a prompt. Therefore, we constructed three paired datasets containing machine-generated and human-written abstracts for the same title. The origin, size, and a sample item of each dataset are shown in Tab. 1.

Participants. We recruited 12 researchers (E1-E12, 9 males and 3 females) from a local university. Each participant has a minimum of 5 years' experience in conducting research on visualization, with 10 possessing previous knowledge of utilizing AI tools (*e.g.*, Grammarly[88]) for proofreading or paraphrasing.

Settings. Based on previous literature[10,15,19] and discussions with experts, we identified three major dimensions of distinctions, *i.e.*, *syntax*, *semantics*, and *pragmatics*[89], each of which consists of several subcategories (Tab. 2). While we acknowledge that we may not have exhaustively captured all possible distinctions, we believe that our derived categories offer valuable insights for future empirical research.

Procedure. We first introduced the study settings and offered a comprehensive explanation of the meaning and coverage of each distinction category. Upon ensuring that the participants had familiarized themselves with the study, we sampled 30 pairs (10 per dataset) of machine-generated and human-written abstracts with the same title from our datasets without specifying their labels. We then asked the participants to rate the *level of quality* for each distinction category of each abstract pair on a 7-point Likert scale from 1 (*low quality*) to 7 (*high quality*). Finally, we collected and analyzed their feedback and conducted a post-study interview to comprehend their evaluation criteria for machine-generated text.

Results Analysis

Figure 1 illustrates the average ratings of various distinction categories used to evaluate machine-generated scientific text. To determine the most important features for humans to distinguish such text, we conducted pairwise comparisons of ratings of abstract pairs within each distinction category using a post-hoc Wilcoxon signed-rank test. If the ratings of abstract pairs are significantly different, the corresponding distinction category should be considered a *critical distinction*, as participants can accurately compare the quality of the scientific text in that category. Based on our analysis of the results and post-study interviews with the participants, we summarize our key findings below.

The *Syntax* dimension refers to the rules and patterns of language that govern how words and sentences are formed and organized. Human-written text received higher ratings in 2/3 of the subcategories in this dimension. Moreover, we observed significant differences in *Text Structure* and *Readability*, but not in *Grammatical Issues*, suggesting that machine-generated scientific text generally exhibit grammatical correctness, but struggle with effective sentence and paragraph organization, leading to suboptimal communication of scientific ideas to readers. For instance, some participants complained that the machine-generated text were *'very hard to follow sometimes'* (E3), which might be attributed to issues such as *'excessively long sentences'* (E3), *'inconsistent use of verb tense'* (E2), or *'overuse of*

Table 1. Descriptions of the three datasets we constructed.

Origin	Size	Sample Item
ChatGPT	2697	Deep Q-Networks (DQNs) have emerged as a powerful technique to solve complex decision-making problems. However, understanding the internal workings of DQNs, especially regarding how different components interact with each other and the consequences of their settings, is a difficult task due to the complex nature of these algorithms. In this paper, we present DQNViz, a novel visual analytics tool that facilitates the exploration and analysis of DQNs. DQNViz provides useful visualizations that help analysts to understand each component of a DQN, as well as to identify issues that can arise during training. DQNViz incorporates the state-of-the-art visualization techniques, such as Parallel Coordinates, Heatmaps, and Scatterplots to explore the high-dimensional spaces of DQNs. We demonstrate the usefulness of DQNViz in two case studies using the Atari and OpenAI Gym environments. Our experiments show that DQNViz is effective in providing insights into how different components of DQNs influence the learning process, as well as for identifying problems such as overfitting, improper weight initialization, and dead nodes. The results of our evaluations indicate that DQNViz is a valuable tool for deep learning practitioners seeking to understand DQNs better.
GPT-3	1598	This paper presents an experimental study of heat transfer and flow visualization of a swirling impinging jet on a flat surface using helicoid inserts. A swirling impinging jet was generated using a circular jet plate with helicoid inserts and then impinged on a flat plate. The Reynolds number was varied from 1,000 to 7,000, while the swirl intensity was kept constant. The results showed that the inserts significantly enhanced the heat transfer rate on the flat plate surface and increased the average Nusselt number by about 30%. Additionally, flow visualization revealed that the inserted jets generated a secondary flow field in the vicinity of the impingement surface, which could explain the improved heat transfer. The results of this study provide useful information for the design of swirling impinging jets for enhanced heat transfer applications.
GPT-2	2590	Area-preserving maps are found across a wide range of scientific and engineering problems. However, existing mapping tools can either be manually generated using complex manual processes or, instead, automatically generated using computer models. Here, we present an algorithm to automatically construct a large volume-preserving aerial map, with the same level of accuracy and resolution, that is found in any topographic or geostatistical database. A comparison of the quality of existing ground-based and air-based maps of the Earth with the proposed methodology is presented, with results, using the results of the two map projects as models, being discussed. The proposed system enables mapping in the air at high resolution for thousands of kilometres, far beyond existing and anticipated aerial mapping capabilities. The proposed algorithm can be applied in large-scale multi-objective or multi-objective-modest surveys. The algorithm operates over either satellite- or local-based aerial mapping, as is found in both the USA and China. The proposed system permits the construction of an aerial mapping of a large area that exists in current terrestrial maps as well. A wide range of uses is considered, including the study of the Earth's surface, as well as of the ground-based, hydrological and geological context and the development of aerial navigation systems.

Table 2. Summary of distinctions between machine-generated and human-written scientific text.

Dimension	Subcategory	Description	Features
Syntax	Grammatical Issues	The correctness and accuracy of using words, phrases and clauses in a sentence	Part-of-Speech Tag Frequency, Punctuation Frequency
	Text Structure	The organization and arrangement of sentences and paragraphs in a text	Paragraph/Sentence Length, Word/Sentence Count, Words per Sentence
	Readability	The ease of reading and understanding the text	Gunning-Fog Index, Flesch Reading Ease
Semantics	Lexical Issues	The choice and usage of words that convey the intended meaning and tone of a text	Google's Top Word Frequency, TF-IDF, NER Frequency
	Consistency	The agreement and harmony of words, phrases and sentences in a text	Average Cosine Similarity between Sentence and Title
	Coherence	The logical connection and relation between sentences and paragraphs in a text	Average Cosine Similarity between Sentences
Pragmatics	Redundancy	The unnecessary repetition of information in a text	Unigram/Bigram/Trigram Overlap of Words/PoS Tags
	Writing Style	The distinctive manner of expressing ideas, opinions or emotions in a text	SciBert[90] Embedding
	Self-Contradiction	The inconsistency or conflict between different parts or aspects of a text	Not Applicable
	Commonsense	The general knowledge or understanding that is expected from the reader/writer of a text	Not Applicable
	Factuality	The level of accurate and verifiable information in a text	Not Applicable
	Specificity	The level of detail in a text to support the main points	Not Applicable

passive voice' (E9). In contrast, human-written scientific text usually have a more structured presentation, facilitating better understanding and interpretation of the content.

The *Semantics* dimension refers to the meaning and interpretation of words and sentences in context. Human-written text received higher ratings in all subcategories in this dimension. Also, significant differences were observed in all three subcategories: *Lexical Issues*, *Consistency*, and *Coherence*, indicating that machine-generated scientific text can be easily distinguished by human experts in terms of semantics. Almost half of the participants (5/12) reported inconsistencies between the title and abstract and incoherencies between sentences. For instance, E4 noted that *'the title is related to time-varying data analysis while the entire abstract keeps talking about focal point extraction'*, while E11 observed that *'most generated text were not logically coherent due to unexpected or unreasonable expressions'*. Additionally, for *Lexical Issues*, some participants complained about *'the lack of lexical diversity'* (E6, E8) and *'informal use of certain words'* (E2) in machine-generated text. These findings confirmed previous literature[10] that suggested machine-generated scientific text still struggled with semantic consistency and coherence when conveying complex ideas and insights. Interestingly, although most participants did not detect any misspellings or vocabulary mistakes, those who were native English speakers all agreed that machine-generated text sounded *'more natural'* (E6) due to *'a better choice of vocabulary'* (E7). This may be attributed to the use of reinforcement learning from human feedback (RLHF)[91] in LLMs like ChatGPT during training, enabling them to sound more human-like. Future research is necessary to investigate this further.

The *Pragmatics* dimension refers to the purpose and effect of language in communication and interaction. Human-written text received higher ratings in all subcategories in this dimension, particularly in terms of *Writing Style* and *Specificity*. Additionally, significant differences were observed in

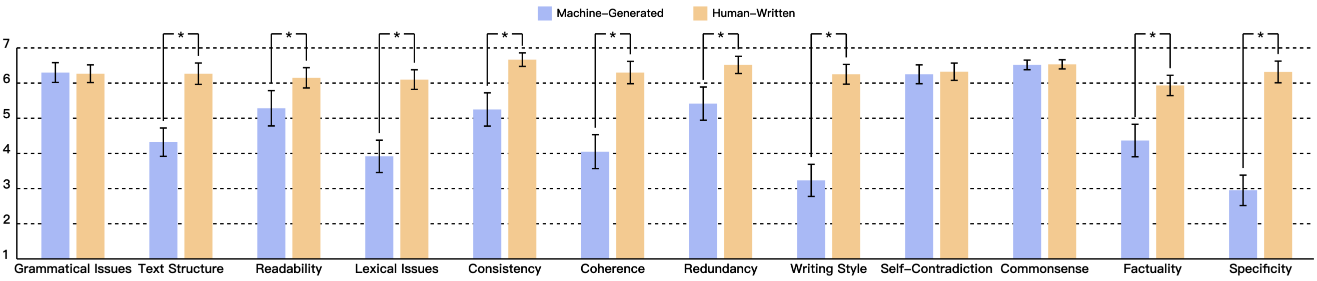


Figure 1. Average ratings of distinction categories on a 7-point Likert scale ($* : p < .05$), where error bars represent 95% confidence intervals.

Redundancy, *Writing Style*, *Factuality*, and *Specificity*, but not in *Self-Contradiction* and *Commonsense*. Participants expressed dissatisfaction with machine-generated text due to *'inadequate details in background, motivation, method and evaluation'* (E2) and *'low formality regarding the overall writing style'* (E7). On the other hand, some participants noted that *'longer and more detailed generated text were often less precise and coherent'* (E5). Our findings aligned with previous literature that machine-generated text were limited due to a "lack of purpose and functionality"[15]. Nevertheless, participants did not report errors related to *Self-Contradiction* and *Commonsense*, contradicting findings in [19]. We attribute this to two primary reasons. First, our datasets mostly comprise short text excerpts (less than 250 words), which are less likely to contain self-contradictions compared to longer text. Second, scientific text generally contain less commonsense knowledge than fake news[92] or online reviews[21], resulting in a lower likelihood of commonsense errors.

In summary, we explored the critical distinctions that human experts rely on to differentiate between machine-generated and human-written scientific text in terms of *syntax*, *semantics*, and *pragmatics*. Results revealed that participants rated human-written scientific text higher in 11/12 of the distinction categories, and significant differences were observed in 9/12 of the categories, providing empirical evidence that a gap still exists between human-written and machine-generated text in scientific writing[10]. The derived critical distinctions were then used to identify *key statistical features* (Tab. 2) informed by previous literature[10,15] to train ML models for our proposed workflow (Sec. *Workflow & System Design*).

Design Requirements

The experiment revealed that human experts could holistically detect machine-generated scientific text based on the quality of critical distinctions, but may not understand the statistical features contributing to their decisions. To this end, we aim to design a mixed-initiative workflow that incorporates both human and machine intelligence and supports interpretation of ML models to enhance the effectiveness and efficiency of the detection process. Throughout the formative study, we solicited input from the experts on their envisioned workflow for identifying fake scientific text, their perspectives on existing automatic detection methods, and their requirements for a detection process involving human intervention. Subsequently, we summarized the insights obtained to establish a comprehensive set of design requirements.

R1: Support adequate human involvement. In the context of scientific text detection, our experts have raised concerns about the limited or non-existent incorporation of human agency in existing artificial detection methods. This inadequacy leads to decreased efficacy, particularly in scenarios requiring domain expertise for reliable and trustworthy decision-making. To address this issue, our experts suggested sufficient integration of their prior knowledge into the detection process, and highlighted the significance of interactive visualizations to facilitate effective human-machine collaboration. Besides, the depth of domain expertise and personal preferences can differ among experts - some might focus on refining their detection capabilities, while senior experts might aim at enhancing detection efficiency. As such, our workflow should support sufficient human involvement through a mixed-initiative procedure and cater to diverse detection needs.

R2: Integrate decisions from multiple ML models. Although experts can generally differentiate between machine-generated and human-written scientific text, they recognized the potential value in using existing ML models to complement and augment their judgments. As numerous detection approaches have been proposed recently, our experts hoped to leverage multiple ML models to enhance their decision-making process, similar to ensemble learning techniques. However, the varying performances of different models under different conditions present a challenge in selecting the most appropriate model for a specific use case. Therefore, our workflow should integrate multiple ML models and provide a mechanism for experts to select the most reliable model for their needs with ease.

R3: Explain the decisions of ML models. One of the major concerns raised by our experts is the lack of interpretability in most black-box detection models. In high-stakes decision-making scenarios like scientific text detection, editors and reviewers must thoroughly understand the reasons behind ML models' decisions to identify fake text to ensure academic integrity and credibility. Furthermore, our experts pointed out that contribution-based explanations alone may be insufficient to fully comprehend models' decisions, especially when typical *normal* feature values are unknown. Accordingly, it is essential to employ XAI techniques along with cohort-level feature analysis to enhance the interpretability and reliability of our workflow, in line with the recommendations of our experts and previous research literature[1].

R4: Analyze multiple text effectively and efficiently. In the context of scientific text detection, journal editors

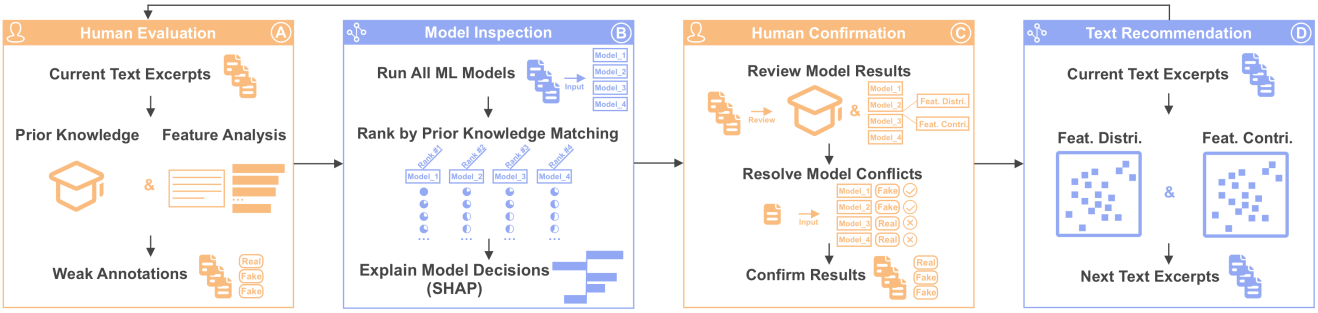


Figure 2. The proposed workflow comprises four iterative stages: (A) human experts provide weak annotations for current text excerpts based on their prior knowledge, (B) ML models are run and ranked based on their *level of prior knowledge matching*, (C) human experts review and confirm results for current text excerpts based on the analysis of models' and their own decisions, and (D) next text excerpts are recommended based on their weighted similarities to current text excerpts.

are often presented with a large number of submitted papers (*e.g.*, 50) and are required to identify the potentially suspicious text excerpts in a timely manner. However, existing industry products for artificial text detection (*e.g.*, GPTZero) only allow users to upload text one by one, which is time-consuming and labor-intensive. While it is possible to deploy an ML model offline and process multiple text simultaneously, our experts have found it challenging to quickly determine the most effective model to use. Moreover, as artificial scientific text can be generated by different language models, it is difficult for experts to identify which text might originate from the same source beforehand, making it impractical to apply the most appropriate model to each one of them. Therefore, our workflow should facilitate effective and efficient detection of multiple scientific text.

Workflow & System Design

We present a mixed-initiative workflow (Fig. 2) that integrates human experts' prior knowledge with multiple ML models for artificial scientific text detection to fulfill the design requirements. In this section, we first describe the four stages of our proposed workflow. Then, we introduce the visual designs and interactions of our visual analytics system (Fig. 3) based on the workflow.

Workflow

Human Evaluation In the first stage, users are presented with scientific text excerpts to review (Fig. 2*A*). At the entry point of the workflow, these excerpts are either manually selected by the users based on their familiarity with the topic, or recommended randomly by the system. However, in subsequent iterations of the workflow, the selection of excerpts is automated based on the weighted similarity of feature distribution and contribution (Sec. *Text Recommendation*). We ask users to represent their prior knowledge as *weak annotations* (R1), that is, assigning a label of *Real* or *Fake* to each text excerpt based on their own judgments. While previous literature and our formative study have demonstrated the abilities of human experts and ML models to distinguish between machine-generated and human-written scientific text, we believe that they are complementary. These weak annotations can be supported and refined by ML models in subsequent stages of the workflow. Additionally, we employ various visualizations such as area charts and highlightings to enable users to interactively analyze features (R1). Although users may not initially understand the meaning and effects of each feature, we posit that after several iterations, they can gradually acquire knowledge of key features and improve their detection ability by integrating their prior knowledge with feature analysis. Notably, users can also apply the outputs of an ML model as weak annotations in batches if they feel confident about its performance on the current text excerpts to increase efficiency (R4).

Model Inspection In the second stage, we incorporate machine intelligence by comparing human experts' decisions with those made by ML models (Fig. 2*B*). First, we run all models on the current text excerpts to obtain their classification results (R2). These models are trained using the statistical features derived from our formative study to make their behavior and performance more understandable to human experts, as these features are closely related to the *critical distinctions* identified by human experts. However, different ML models may be trained on different datasets, resulting in varying performances when detecting machine-generated text from different origins. For instance, a model trained on text generated by GPT-3 may not accurately detect text generated by ChatGPT. As human experts can holistically identify machine-generated scientific text, we rank the ML models based on their *level of prior knowledge matching* to human experts, which is computed as

$$LevelofPKMatching = (1 - \omega_g) * R_{local} + \omega_g * R_{global} \tag{1}$$

where R_{local} and R_{global} indicate the rate which the classification results of each model *match* those of the user in the *current iteration* and *all previous iterations*, respectively. R_{global} is weighted by a parameter ω_g, which can be adjusted by the user at any point of the workflow according to their preferences and requirements (R1). Lower values of ω_g indicate a focus on finding the best model for the current iteration, while higher values indicate a broader focus on the entire detection process. Notably, We rank ML models not only to meet users' needs for selecting the most appropriate model (R2), but also to form *positive first impressions* which make it more likely for users with domain expertise to trust and use the intelligent system in the future, as suggested in [93]. Finally, we use SHAP values[50] to generate contribution-based explanations of statistical features to facilitate interpretability (R3).

Human Confirmation In the third stage, users can review the list of ML models ranked by *level of prior knowledge matching* (Fig. 2C). The most appropriate model for the current iteration (either globally or locally) should ideally be ranked high. Users can validate their weak annotations by investigating the feature distributions and contributions (R1). While the top-ranked model is initially selected for analysis, users are free to explore any other models to gain insights into why a particular model performs poorly or to validate their prior knowledge. If all models fail to match users' prior knowledge well, this may be due to inadequate training of the models or lack of expertise of the users. In such cases, users may refine their own prior knowledge or assess the usability of the models. Subsequently in this stage, users are responsible for resolving any conflicts arising from different models' decisions and confirming the final results for the text excerpts (R1, R2). We provide effective visualizations to assist users in identifying models' decisions that contradict their weak annotations, and to enable model-wise and instance-wise comparisons for each text excerpt (R1).

Text Recommendation In the fourth stage, we perform similarity calculation to recommend text excerpts for the next iteration of the workflow (Fig. 2D). We calculate the average cosine similarity between each remaining text excerpt and the current text excerpts as follows

$$sim_i = \frac{1}{n}\sum_{j=1}^{n}\left(\omega_d \cdot \frac{f_{d,j} \cdot f_{d,i}}{\|f_{d,j}\|\|f_{d,i}\|} + (1-\omega_d) \cdot \frac{f_{c,j} \cdot f_{c,i}}{\|f_{c,j}\|\|f_{c,i}\|}\right) \tag{2}$$

where n indicates the iteration size, and $f_{d,i}$ and $f_{c,i}$ indicate the feature distribution and contribution vector of the i-th text excerpt, respectively. $f_{d,i}$ and $f_{c,i}$ are defined as

$$f_{d,i} = \begin{pmatrix} f_{d,i1} \\ f_{d,i2} \\ \vdots \\ f_{d,im} \end{pmatrix} \tag{3}$$

$$f_{c,i} = \mathrm{SHAP}(f_{d,i}) \tag{4}$$

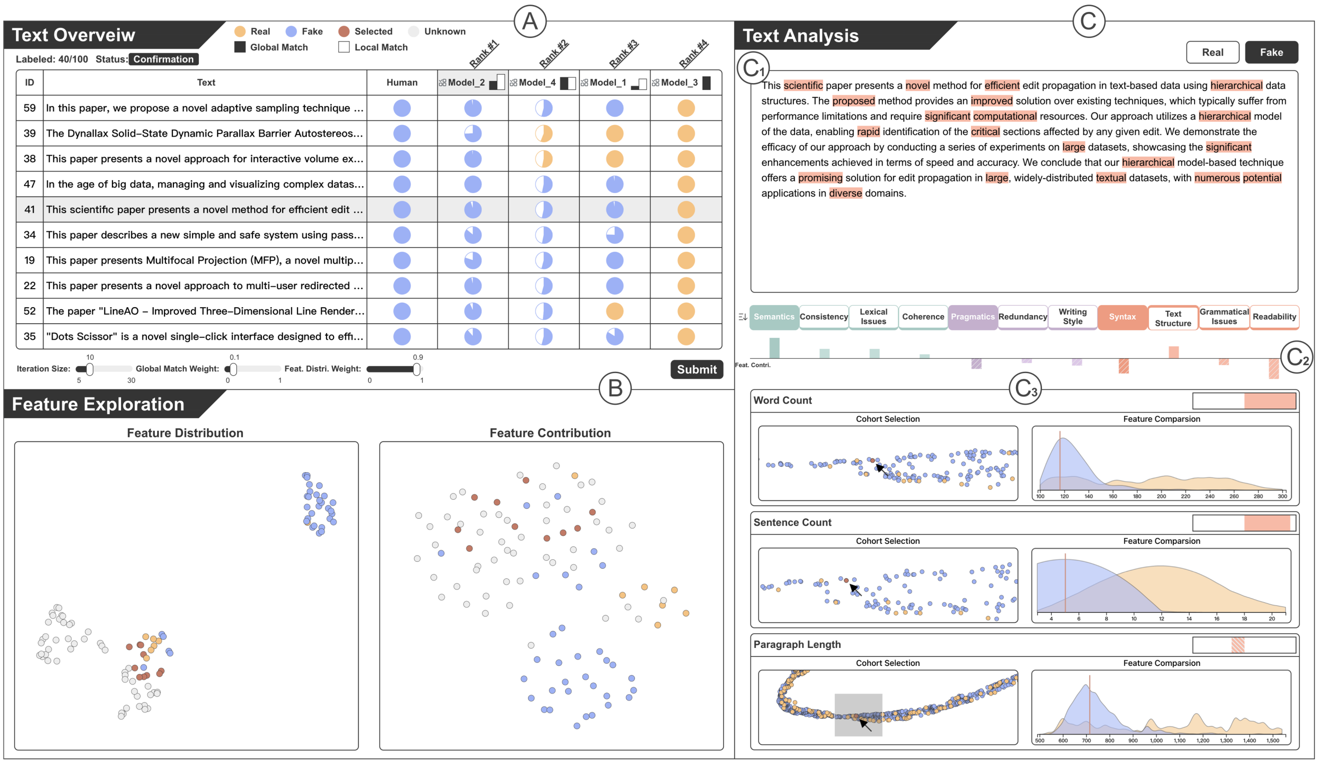
The elements of $f_{d,i}$ are the key statistical feature values identified in our formative study (Tab. 2). And the elements of $f_{c,i}$ are SHAP values for the i-th text excerpt, which quantify the local contributions of each feature to the detection model's classification result. As SHAP values are model-specific, we use the top-ranked model in the current iteration to calculate $f_{c,i}$. The similarity of feature distribution is weighted by a parameter ω_d. Higher values of ω_d indicate a preference for handling text excerpts generated by similar LLMs first, as they typically have similar feature distributions. Conversely, lower values of ω_d indicate a preference for handling text excerpts treated by similar *model strategies* first, which have similar feature contributions. As suggested by Collaris *et al.*[56], model strategies reflect different treatments of ML models to the input data. In the context of scientific text detection, we formulate model strategies as various types of *characteristics* of text. For instance, some text excerpts may be classified as machine-generated mainly due to their short paragraph length, while others may be due to their abnormal writing styles. Users can adjust ω_d freely throughout the workflow to balance the handling of text excerpts based on their origins and characteristics, catering to diverse detection needs (R1). After the similarity calculation, we select the text excerpts with the highest similarities among the remaining ones for the next iteration. Iteratively presenting several text excerpts enhances the effectiveness and efficiency of detecting multiple text excerpts (R4), since users can conveniently adopt the most appropriate model of the current iteration to the next one without much performance loss. Additionally, if users are unsatisfied with the recommended text excerpts, they can reduce the iteration size to achieve a more fine-grained detection process.

The fourth stage marks the end of an entire iteration, and the workflow continues to loop from the first stage until all text excerpts are reviewed and confirmed by the user. Through an iterative manner, the workflow facilitates the effectiveness and efficiency in detecting multi-sourced scientific text, and provides sufficient support for human-machine collaboration and interpretation of ML models' decisions, satisfying our derived design requirements.

User Interface

Text Overview The *Text Overview* (Fig. 3A) panel provides an interactive tabular layout that shows the text excerpts in the current iteration of the workflow. It allows users to compare and analyze decisions made by themselves and different models either in a model-wise (vertically) or instance-wise (horizontally) manner. Each text excerpt is encoded as a circle with different colors representing their labels, and the predicted probability of each model is shown as the size of the fan-shaped area inside each circle. Models are ranked from left to right based on their *level of prior knowledge matching* in the current iteration, and users can click on each column to focus on the corresponding model. Besides, clicking on each row allows users to analyze and annotate the corresponding text excerpt. The current focused text excerpt or model is highlighted in light grey. To the right of each model's name is a small barchart (◧) indicating the *Global Match Rate* and *Local Match Rate*. Additionally, users can apply the outputs of a model as weak annotations by clicking on the *'batch apply'* icon (⊞) to the left of each model's name. Above the tabular layout are the number of labeled text excerpts, current iteration status, and model ranks, and below are sliders for adjusting the workflow parameters (Sec. *Workflow*) and a button for submitting decisions. Notably, changes to the parameters will take effect in the next iteration.

This panel serves different purposes depending on the stage of the workflow. In the first stage, the circles representing models' decisions are not displayed (unless users wish to *'batch apply'* the outputs of a particular model), as users need to provide weak annotations based on their own prior knowledge. In subsequent stages, all results of models are displayed, as users need to compare and confirm the final results for the current text excerpts.


A
Text Overveiw
● Real ● Fake ● Selected ● Unknown
■ Global Match □ Local Match
Labeled: 40/100 Status: Confirmation
Rank #1
Rank #2
Rank #3
Rank #4

ID	Text	Human	Model_2	Model_4	Model_1	Model_3
59	In this paper, we propose a novel adaptive sampling technique ...	●	●	◑	●	●
39	The Dynallax Solid–State Dynamic Parallax Barrier Autostereos...	●	◕	◑	●	●
38	This paper presents a novel approach for interactive volume ex...	●	●	◑	●	●
47	In the age of big data, managing and visualizing complex datas...	●	●	◑	●	●
41	This scientific paper presents a novel method for efficient edit ...	●	●	◑	●	●
34	This paper describes a new simple and safe system using pass...	●	●	◑	◕	●
19	This paper presents Multifocal Projection (MFP), a novel multip...	●	●	◑	●	●
22	This paper presents a novel approach to multi–user redirected ...	●	●	◑	●	●
52	The paper "LineAO – Improved Three–Dimensional Line Render...	●	●	◑	●	●
35	"Dots Scissor" is a novel single–click interface designed to effi...	●	●	◑	◕	●

Iteration Size: 10 (5–30) Global Match Weight: 0.1 (0–1) Feat. Distri. Weight: 0.9 (0–1)

This scientific paper presents a novel method for efficient edit propagation in text-based data using hierarchical data structures. The proposed method provides an improved solution over existing techniques, which typically suffer from performance limitations and require significant computational resources. Our approach utilizes a hierarchical model of the data, enabling rapid identification of the critical sections affected by any given edit. We demonstrate the efficacy of our approach by conducting a series of experiments on large datasets, showcasing the significant enhancements achieved in terms of speed and accuracy. We conclude that our hierarchical model-based technique offers a promising solution for edit propagation in large, widely-distributed textual datasets, with numerous potential applications in diverse domains.

Figure 3. The user interface of the visual analytics system. (A) The *Text Overview* panel allows users to provide weak annotations for text excerpts based on their prior knowledge and compare them with multiple ML models' decisions. (B) The *Feature Exploration* panel presents an overview of the feature distribution and contribution of all text excerpts to be detected. (C) The *Text Analysis* panel offers contribution-level, distribution-level, and excerpt-level feature analysis to facilitate an effective and interpretable detection process. By iteratively interacting with the three coordinated panels, users can perform effective and efficient detection of multiple scientific text through a mixed-initiative workflow.

Feature Exploration The *Feature Exploration* (Fig. 3*B*) panel allows users to gain an overview of the feature distribution and contribution of all text excerpts to be detected. It consists of two UMAP[94] projections: one for feature distribution values and the other for feature contribution values. Text excerpts are represented as colored dots according to their corresponding labels, with unlabeled excerpts being depicted in grey. Those in the current iteration are highlighted in red, and users can manually select other excerpts to review by brushing. As discussed in Sec. *Text Recommendation*, the feature distribution projection visualizes text excerpts by their generative sources (*i.e.*, various LLMs), while the feature contribution projection depicts these excerpts based on different model strategies of the top-ranked detection model in the current iteration, reflecting distinct text characteristics. Specifically, clusters in *feature distribution* typically correspond to text generated by similar LLMs, while clusters in *feature contribution* indicate text treated by similar model strategies. Notably, clusters in the two projections generally do not match with each other as there is usually no explicit relationship between feature values and feature contribution values. By examining the 2D projections of feature distribution and contribution, users can quickly identify patterns related to different language models or model strategies, as well as the relationship between text excerpts across consecutive iterations. For instance, if the text excerpts in the current iteration are similar to those in the previous iteration in terms of proximity in feature distribution, users may choose to *'batch apply'* the outputs of the previous best model to improve efficiency.

Text Analysis The *Text Analysis* (Fig. 3*C*) panel provides three levels of feature analysis for the selected text excerpt in the *Text Overview* panel.

Contribution-level analysis. As shown in Fig. 3C_2, we display each distinction category in a grouped style similar to tab groups in web browsers. This facilitates easy comparison and analysis of the impact of different features. Each feature dimension is presented as a separate *tab* using different colors, and comprises several subcategories represented by *sub-tabs*, which are outlined in the same color as their corresponding dimension for clarity. Users can sort the feature dimensions and subcategories on the tab strip, and can expand or collapse a feature dimension by clicking on its tab. Clicking on a subcategory tab highlights it and allows users to view the associated feature details presented as multiple *feature cards* (Fig. 3C_3).

We calculate the aggregate contribution values of each feature dimension or subcategory by summing up the SHAP values of the included features[54,55]. The contribution values for each feature dimension and subcategory are depicted as a vertical bar below each corresponding tab (Fig. 3C_2), with lighter colors indicating subcategories and darker colors representing dimensions. For each included feature, its contribution value is shown as a horizontal bar atop the corresponding feature card (Fig. 3C_3). Notably, negative contribution values are indicated by striped bars. By presenting contribution values in a coarse-to-fine-grained manner, we enable users to analyze features at different levels of granularities catering to their analysis goals. For example, some users may only be interested in understanding the most critical distinctions affecting the models' decisions, while others may wish to delve deeper to gain a more comprehensive understanding of the underlying statistical features' effects.

Distribution-level analysis. It may be insufficient to only display the numerical feature values and their contribution values for users to comprehend and build trust in models' decisions, as some statistical features may be unfamiliar to them (*e.g.*, *Gunning-Fog Index*). To this end, inspired by [55], we adopt *reference values*, which are calculated from a relevant cohort of text excerpts. We obtain the 500 most similar text excerpts to the current selected one from the *focused* model's training dataset based on cosine similarities between their feature distribution vectors. Users can also manually select a more fine-grained cohort by brushing in the UMAP projection of the default relevant excerpts in *Cohort Selection* (Fig. $3C_3left$).

The statistical features include numerical values and multidimensional vectors, which are displayed in *Feature Comparison* (Fig. $3C_3right$). We employ area charts (Fig. $3C_3right$) to visualize the distributions of numerical features (*e.g.*, *paragraph length*) and violin charts (Fig. $4C_2$) to visualize the distributions of vector elements for low-dimensional features that consist of meaningful elements (*e.g.*, *part-of-speech tag frequency*). For high-dimensional features, such as embeddings from a transformer model, we leverage scatterplots to visualize the UMAP projection of the feature vectors (Fig. $5B$). The same color encodings as before are used to depict machine-generated and human-written text. Moreover, we indicate feature values of the current selected text excerpt with a vertical line in area charts, a horizontal line in violin charts, and an arrow in scatterplots. Reference values are determined by lower and upper bounds (for numerical features) or clusters (for embedding features) of the relevant cohort. By comparing the current feature value with its corresponding reference values in context, users can determine whether the text excerpt is abnormal in terms of each feature. For instance, if the *Gunning-Fog Index* value of the current text excerpt is significantly lower than those of similar human-written text, it may indicate that the text is likely machine-generated.

Excerpt-level analysis. Some features are calculated based on word frequencies in the text excerpt, such as *part-of-speech tag frequency*. Users prefer to understand these features in the context of the original text. Therefore, we visually associate these features with the corresponding raw text. As shown in Fig. $3C_1$, by *clicking on* a feature element in the violin charts, users can highlight the associated words in the raw text. This allows users to gain a deeper understanding of the correlations between statistical features and specific characteristics in the original text. By analyzing back and forth between features and text, users can gradually learn from the model's decisions and explanations and become more attentive to those characteristics in future detection processes. However, some features, such as *average sentence length*, are relative to the entire text rather than individual words. Therefore, we do not provide any interactions for reference, and users should observe their distributions as a whole.

Table 3. Accuracy scores of the ML models trained and evaluated on datasets of different LLMs.

Training Data	Test Data							
	ChatGPT		GPT-3		GPT-2		Combined	
	Acc.	AUC	Acc.	AUC	Acc.	AUC	Acc.	AUC
ChatGPT	0.953	0.953	0.891	0.891	0.576	0.589	0.794	0.807
GPT-3	0.783	0.796	0.975	0.975	0.443	0.393	0.697	0.714
GPT-2	0.530	0.697	0.473	0.381	0.988	0.987	0.692	0.794
Combined	0.816	0.817	0.908	0.914	0.865	0.869	0.856	0.859

Notably, feature contribution values are not displayed in the first stage of the workflow, whereas in subsequent stages, feature distribution and contribution values are subject to the current focused model.

Evaluation

Settings

Training detection models. Feature values were computed for text excerpts within each dataset outlined in Table 1, utilizing the statistical features specified in Table 2. Subsequently, three LightGBM models[95] were trained to classify machine-generated and human-written scientific text across these datasets. An additional LightGBM model was trained on a combined dataset, consisting of 500 randomly selected text excerpts from each original dataset, totaling 1500 samples. Due to their relatively small size, datasets were divided into a 70% training set and a 30% test set, with k-fold cross-validation applied to the training set. It is important to note that the selection of detection models does not influence our proposed workflow, which is adaptable to any feature-based detection models. We chose LightGBM for its high efficiency and accuracy. Following the training phase, each model was applied to classify text excerpts within all four datasets. SHAP values were then calculated for each excerpt for feature contributions, based on each model's detection results. These feature distribution and contribution values were stored offline to support the visual analytics system.

Evaluating detection models. Table 3 shows that the accuracy scores of the detection models are high when trained and evaluated on the same dataset, but their performance decreases across different datasets due to the OOD issue mentioned earlier. Additionally, the detection model trained on the combined dataset performed relatively well on all four datasets, but not optimally.

Based on these models and our visual analytics system, we conducted two case studies and a controlled user study to demonstrate the effectiveness of our approach.

Case Study 1: Origin-based Detection

E1 aimed to review 100 manuscripts and identify which ones were likely to be machine-generated. With over 10 years of research experience, E1 felt confident in his ability to detect machine-generated scientific text. He hoped to improve his efficiency in detecting multiple text excerpts and to seek evidence from ML models to support his own decisions.

Explore feature distribution and contribution. E1 began by checking the *Feature Exploration* panel and observed three clusters in *feature distribution* (Fig. 4*A*), indicating that the corresponding text excerpts likely shared similar origins. E1 recognized that each cluster of text excerpts could be detected using the same ML model for optimal performance. He then decided to adopt an *origin-based* procedure and set *Global Match Weight* to 0.1 to direct the system to present the locally optimal model in each iteration. This allowed him to focus solely on the top-ranked models and avoid analyzing all the results tediously.

Provide weak annotations. At the start of the workflow, the system automatically presented the text excerpts for the first iteration, using the default iteration size of 10. All of these excerpts were from the cluster located in the upper right corner of *feature distribution*. E1 provided weak annotations for each text excerpt based on his prior knowledge. However, he did not extensively analyze the features, as they were not familiar or intuitive to him at the moment. E1 hoped to delve deeper into these features once the models' decisions were displayed.

Review models' decisions. Upon submission of the weak annotations, all four models were run. As shown in Fig. 4*B*, *Model_3* was the top-ranked model with a *local match rate* of 100%. E1 then selected each text excerpt to investigate their feature distribution and contribution values. The significant contributions were mainly made by two subcategories: *Coherence* and *Grammatical Issues*. In the *Coherence* subcategory, the most important feature was *Average Cosine Similarity between Sentences* whose value was obviously lower than the corresponding distribution of human-written text (Fig. 4C_1). This finding was consistent with E1's own evaluation, who commented that the paragraph was not logically connected. In the *Grammatical Issues* subcategory, the primary feature was *Punctuation Frequency*, which is a multidimensional feature displayed using a violin chart. E1 discovered that the frequencies of *colons* and *quotation marks* were far beyond the reference value range of human-written scientific text (Fig. 4C_2), indicating that the text was likely machine-generated. E1 confirmed this observation, noting that scientific abstracts generally do not contain such punctuation.

Re-adjust similarity calculation strategy. After completing the first iteration, the system presented another 10 text excerpts for the second iteration. However, E1 noticed that they were not in the same cluster as before. This was due to the default value of *Feature Distribution Weight* being set to 0.5, which did not prioritize similar text excerpts. Therefore, E1 manually re-selected the text excerpts and adjusted the parameter to 0.9 to ensure the system recommended text excerpts from similar origins.

Batch apply models' decisions. E1 chose to 'batch apply' the outputs of *Model_3* to the current text excerpts due to their high similarities with the previous ones. However, he still manually verified the model's decisions for each excerpt to ensure consistency with his prior knowledge. Notably, only one text excerpt failed to match due to some semantic errors that were difficult for models to capture but obvious to human experts. After careful analysis, E1 corrected the model's decision for this case. Through the whole process, E1 gradually built trust in *Model_3*.

Move on to other clusters. E1 proceeded to examine the text excerpts in the second cluster located in the lower left corner. This time, *Model_2* became the top-ranked one as it performed well on the current text excerpts that were possibly from another origin. Therefore, E1 followed similar procedures as before by extensively analyzing only the initial text excerpts and then 'batch applying' the model's decisions to those in the same cluster. This enabled him to efficiently complete the detection process for all clusters in *feature distribution*.

Summary. This case study shows how our approach assists human experts in efficiently detecting multi-sourced text by iteratively handling text excerpts from similar origins and utilizing ML models to provide evidence to support their judgments through feature analysis.

Case Study 2: Strategy-based Detection

E2 was provided with 40 manuscripts to review. Unlike E1, E2 was not as experienced and hoped to benefit more from human-machine collaboration and learn from models' decisions through comprehensive feature analysis.

Locate similar model strategies. E2 examined the *Feature Exploration* panel and identified two clusters in *feature contribution* (Fig. 5*A*), indicating similar model strategies. As such, E2 decided to adopt a *strategy-based* procedure. He adjusted *Global Match Weight* to 0.9 to avoid local optimum and set *Feature Distribution Weight* to 0.1 to ensure the system recommended text excerpts from the same clusters in *feature contribution*.

Find the globally best model. After several iterations, E2 discovered that only *Model_4* had a moderate *global match rate* of 80%. The separate locations of the text excerpts in *feature distribution* (Fig. 5A_1) suggested that they came from diverse origins, resulting in poor performances of the other three models trained on single datasets. In contrast, *Model_4* was trained on a combined dataset and showed an acceptable accuracy in detecting multi-sourced text (Tab. 3). Hence, E2 decided to conduct an analysis focused on *Model_4* to gain some insights.

Analyze the first cluster. It turned out that most text excerpts in the cluster located in the lower left corner of *feature contribution* were machine-generated. To investigate the most influential factors, E2 inspected the *Text Analysis* panel for each text excerpt. He found that the *Syntax* dimension consistently contributed the most, with the *Text Structure* subcategory having the greatest impact within this dimension. Furthermore, the contribution value of *Word Count* was much higher than others. E2 then performed a more fine-grained analysis by brushing a smaller area in *Cohort Selection*, and found that the *Word Count* value was significantly lower than most human-written text (Fig. 5C_3). Additionally, *Part-of-Speech Tag Frequency* in the *Grammatical Issues* subcategory was another important feature. Further investigations revealed that the frequencies of certain word classes were noticeably lower than those in human-written text (Fig. 5C_2). By highlighting the corresponding words in the raw text (Fig. 5C_1), E2 gained a deeper understanding about this feature. Notably, these observations persisted across most text excerpts of the first cluster, indicating that they were classified as machine-generated primarily due to syntax issues.

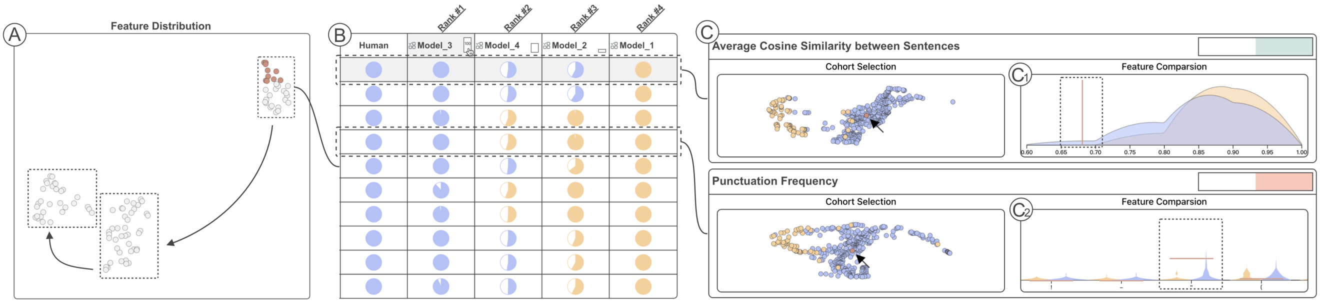
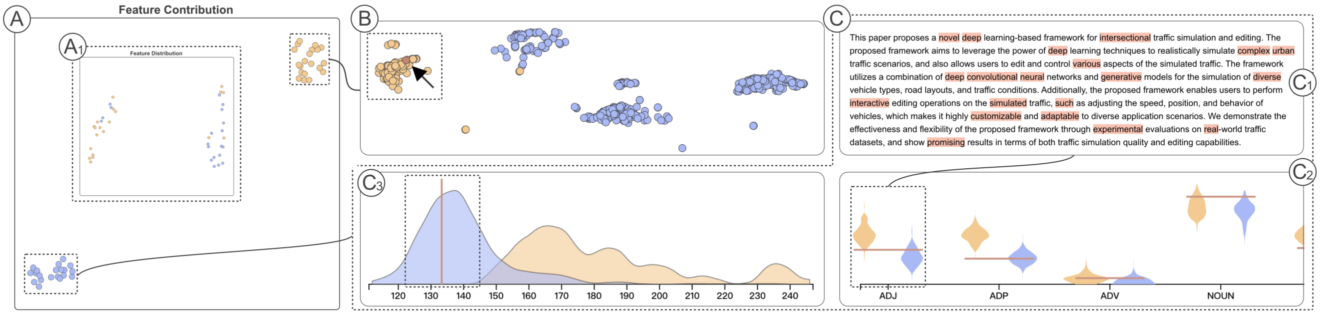


Figure 4. Case Study 1 illustrates how the user (A) locates clusters in *feature distribution*, (B) reviews models' decisions, and (C) builds trust in the appropriate model through feature analysis.



This paper proposes a novel deep learning-based framework for intersectional traffic simulation and editing. The proposed framework aims to leverage the power of deep learning techniques to realistically simulate complex urban traffic scenarios, and also allows users to edit and control various aspects of the simulated traffic. The framework utilizes a combination of deep convolutional neural networks and generative models for the simulation of diverse vehicle types, road layouts, and traffic conditions. Additionally, the proposed framework enables users to perform interactive editing operations on the simulated traffic, such as adjusting the speed, position, and behavior of vehicles, which makes it highly customizable and adaptable to diverse application scenarios. We demonstrate the effectiveness and flexibility of the proposed framework through experimental evaluations on real-world traffic datasets, and show promising results in terms of both traffic simulation quality and editing capabilities.

Figure 5. Case Study 2 illustrates how the user (A) locates clusters in *feature contribution*, (B) analyzes the human-written text cluster, and (C) analyzes the machine-generated text cluster.

Analyze the second cluster. On the other hand, most text excerpts in the cluster located in the upper right corner of *feature contribution* were human-written. Further investigations revealed that the *Pragmatics* dimension always had the highest positive contribution, whereas the other two dimensions had negative or negligible contributions. Therefore, E2 analyzed the most influential feature *SciBert Embedding*, which belonged to the *Writing Style* subcategory. By comparing the current focused text excerpt (red dot) with the relevant cohort, E2 discovered that it belonged to a cluster predominantly comprising human-written text (orange dots), while other clusters were mainly composed of machine-generated text (blue dots), as shown in Fig. 5*B*. This observation was also validated in other text excerpts of the second cluster, suggesting that they were classified as human-written mostly based on their writing styles.

Gain knowledge from models. Through the analysis of different model strategies, E2 outlined two characteristics that may help him in future identification of scientific text. First, machine-generated text are typically shorter and less detailed compared to human-written ones. Second, the writing styles of human-written text differed significantly from machine-generated ones in terms of specificity and formality. These principles align with our formative study and are further supported through feature analysis.

Summary. This case study demonstrates how our approach assists human experts in interpreting models' decisions by iteratively handling text excerpts treated by similar model strategies and leveraging models' insights to identify distinctive characteristics that differentiate machine-generated text from human-written text.

User Study

Experiment Design

Participants. We recruited 15 participants, consisting of 9 researchers and 6 practitioners, each possessing sufficient domain expertise and a minimum of 5 years of research experience in visualization.

Settings. To evaluate the effectiveness and efficiency of our approach, a comparative study was conducted using three conditions for scientific text detection:

C1. Human only. An ablated version of the system was utilized that only displayed raw text. Participants annotated each text excerpt based on their prior knowledge.

C2. Machine only. An ablated version of the system was utilized that only displayed the outputs of four models. Participants randomly applied one of the outputs to each text excerpt.

C3. Human-machine collaboration. The complete version of the system was utilized. Participants examined each text excerpt combined with models' decisions and feature analysis, and then confirmed their final decisions.

Procedure. Participants were introduced to the study's purpose and asked to complete a consent form. They were then instructed on the experiment design and encouraged to familiarize themselves with the three versions of the system with designed examples[96]. During the experiment, the order of conditions were counterbalanced across participants. For each condition, we sampled 60 text excerpts (15 ChatGPT-generated, 15 GPT-3-generated, 15 GPT-2-generated, and 15 human-written), yielding 180 trials in total. In addition to annotations, participants were required to rate their *confidence score* on a 7-point Likert scale from 1 (*low confidence*) to 7 (*high confidence*) for every 10 trials[93]. Finally, we evaluated the results for each condition in terms of *effectiveness*, *efficiency*, and *reliability*, measured by overall accuracy, total completion time, and average confidence score (Fig. 6). Post-study interviews were also conducted to collect qualitative feedback.

Results Analysis **Effectiveness.** *C3* achieved a higher accuracy than *C1* by 9% on average, both of which outperformed *C2* significantly. This confirms that human experts can detect most machine-generated scientific text, while blindly applying models' outputs proves inadequate in identifying multi-sourced text. Most participants (11/15) increased their accuracy in *C3* compared to *C1*, since the system enabled them to locate the most appropriate model under various conditions by matching their decisions with models' outputs. Participants could then refine their weak annotations based on models' decisions, as certain excerpts were *'initially hard to distinguish'* (P2, P9) or *'required further verification from suitable models'* (P4).

Efficiency. Regarding completion time, *C2* was the fastest by directly utilizing models' decisions but sacrificed accuracy, as most participants resorted '*majority voting*' (P3, P4) or '*random selection*' (P9, P12). Although *C3* was only 10% faster than *C1* on average, most time was spent on feature analysis or due to *'system lagging'* (P8). Also, our experiment involved a relatively small number of text excerpts, which reduced the gap between completion times. Additionally, participants praised the *'batch apply'* function, which saved much time since they *'do not need to manually check each excerpt anymore'* (P3). Therefore, the efficiency could be further improved once the system issues are resolved, especially for large amounts of text excerpts.

Reliability. The average confidence score in *C3* was 13% and 93% higher than *C1* and *C2*, respectively, demonstrating the effectiveness of our approach in improving detection reliability. Compared to *C2*, *C3* significantly increased user trust by allowing them to *'see what factors contributed to models' decisions'* (P7). However, some participants' confidence scores did not increase much between *C3* and *C1* due to explanations being *'beyond expectations'* (P10), potentially caused by some correlated features that made SHAP values less effective[97]. Interestingly, some participants tended to seek explanations that supported their initial judgments to *'feel more confident'* (P1), even if the corresponding contribution values were low. This could lead to bias by ignoring other important features. Future research on human-centered XAI[98] is needed to explore this phenomenon further.

Summary. Overall, the results showed that our approach facilitated the effectiveness, efficiency, and reliability of the detection process. Despite minor issues such as system lagging, participants gave positive feedback on the usability of our visual analytics system, especially the feature projections which helped them discover patterns in feature distribution and contribution values.

Discussion

In this section, we summarize the derived design implications, and then discuss the limitations and future work.

Implications

Incorporate human experts' prior knowledge. Combining the capabilities of human experts and machine intelligence in detecting scientific text is effective, as our work has shown. Human-machine collaboration in the detection process has been previously advocated for[1,21,24], and some tools[22] have been proposed to incorporate a human analyst to facilitate detection. In our work, we match ML models' decisions with human experts' prior knowledge to handle situations where models' outputs conflict with each other or human judgment. We place much emphasis on the human side, as we believe that it is the responsibility of human experts to confirm the final judgments in high-stakes decision-making scenarios. Thus, in determining whether a manuscript is fake and violates academic integrity, experts should aim for a balance between trusting their own judgments and seeking evidence from models, rather than blindly accepting models' decisions, especially when models' capabilities may be weaker than humans themselves due to various limitations such as OOD issues.

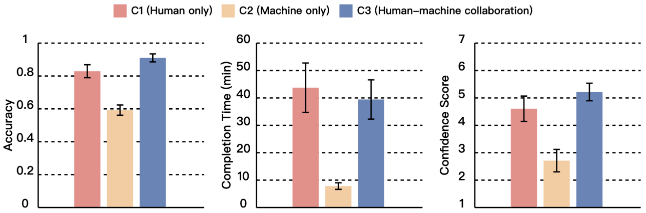


Figure 6. Overall accuracy, total completion time, and average confidence score for each condition.

Leverage multiple models to facilitate detection. Given the diversity and rapid development of LLMs, a single detector may not suffice for all situations. To address the issue, our mixed-initiative workflow is model-agnostic and integrates various detectors to maximize their strengths under different conditions. We assume that human judgments are the '*gold standard*' in scenarios that require domain expertise or context information to verify the detection results, such as academic and educational contexts. Therefore, we provide a new perspective in this work by leveraging multiple detectors along with human agency to facilitate the detection process.

Support feature analysis with multiple granularities and levels. Although preliminary works[8,10] have explored the feature contributions of ML models for detection, such explanations are insufficient for effective interpretation. Our experts appreciated the coarse-to-fine-grained manner to display contribution values that can serve various analysis goals. In addition to contribution-level analysis, we also provide distribution- and excerpt-level analysis to understand features from the cohort and instance perspective, respectively. These analyses enhance human experts' trust in models' decisions and enable them to learn from models' behavior to improve their own detection ability.

Limitations & Future Work

Scalability. The current system is limited to a maximum iteration size of 30 and 4 ML models in the workflow, which may hinder its application in real-world scenarios. Experts noted that the tabular view may not facilitate effective comparisons when there were too many rows or columns, while excessive dots in feature projections may cause visual clutter[99]. Therefore, it is promising to design more effective visualizations to support the workflow at a larger scale. Additionally, our system currently employs SHAP values to evaluate feature contribution, which can be computationally expensive, especially given the vast volume of manuscript submissions encountered by leading conferences and journals. While the computation of feature distribution and contribution can be an offline task that does not impede the workflow, the issue may present a potential limitation when applied to large-scale text datasets. Therefore, the integration of more efficient XAI techniques, such as LIME[51], is a viable direction for future work.

Generalizability. The proposed workflow can be extended to various text detection contexts that require human involvement and interpretability, such as identifying cheating in educational contexts. However, given the various *critical distinctions* between machine-generated and human-written text across different scenarios, directly applying the current statistical features and system designs may not be feasible. Additionally, although the workflow can theoretically be generalized to any feature-based ML models, transformer-based models are not compatible with our feature-based explanations and require alternative XAI and visualization techniques for interpretation. In future work, we plan to extend the workflow to support newly proposed detection methods such as watermarking[100] and probability curvature[101].

Conclusion

In this work, we identify critical distinctions between machine-generated and human-written scientific text through a quantitative experiment. Our findings provide valuable insights into the capabilities of LLMs in academic writing and can inform the design of more effective detection methods. We propose a mixed-initiative workflow and a visual analytics system that incorporates human experts' prior knowledge to facilitate the efficiency, interpretability, and reliability of the detection process. We demonstrate the effectiveness of our approach through two case studies and a controlled user study. We believe that our work will inspire future research on integrating human intelligence into artificial text detection.

Acknowledgements

This paper is supported by the National Natural Science Foundation of China (62132017, 62302435), Zhejiang Provincial Natural Science Foundation of China (LD24F020011), and "Pioneer" and "Leading Goose" R&D Program of Zhejiang (2024C01167).

References

1. Crothers E, Japkowicz N and Viktor H. Machine generated text: A comprehensive survey of threat models and detection methods. *arXiv preprint arXiv:221007321* 2022; .
2. Xie T, Zhou F, Cheng Z et al. Openagents: An open platform for language agents in the wild. *arXiv preprint arXiv:231010634* 2023; .
3. OpenAI. Chatgpt. https://chat.openai.com/, 2023.
4. Strobelt H, Webson A, Sanh V et al. Interactive and visual prompt engineering for ad-hoc task adaptation with large language models. *IEEE Transactions on Visualization and Computer Graphics* 2023; 29(1): 1146–1156. DOI:10.1109/TVCG.2022.3209479.
5. Feng Y, Wang X, Wong KK et al. Promptmagician: Interactive prompt engineering for text-to-image creation. *arXiv preprint arXiv:230709036* 2023; DOI:10.48550/arXiv.2307.09036.
6. Zellers R, Holtzman A, Rashkin H et al. Defending against neural fake news. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/3e9f0fc9b2f89e043bc6233994dfcf76-Paper.pdf.
7. Pagnoni A, Graciarena M and Tsvetkov Y. Threat scenarios and best practices to detect neural fake news. In *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea, pp. 1233–1249. URL https://aclanthology.org/2022.coling-1.106.
8. Kowalczyk P, Röder M, Dürr A et al. Detecting and understanding textual deepfakes in online reviews 2022; URL http://hdl.handle.net/10125/79516.
9. Mirsky Y, Demontis A, Kotak J et al. The threat of offensive ai to organizations. *Computers & Security* 2023; 124: 103006. DOI:https://doi.org/10.1016/j.cose.2022.103006.
10. Ma Y, Liu J and Yi F. Is this abstract generated by ai? a research for the gap between ai-generated scientific text and human-written scientific text. *arXiv preprint arXiv:230110416* 2023; .
11. Rodriguez J, Hay T, Gros D et al. Cross-domain detection of gpt-2-generated technical text. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, pp. 1213–1233. DOI:10.18653/v1/2022.naacl-main.88.
12. ACL. Acl 2023 policy on ai writing assistance. https://2023.aclweb.org/blog/ACL-2023-policy/, 2023.
13. ICML. Clarification on large language model policy llm. https://icml.cc/Conferences/2023/llm-policy, 2023.
14. GPTZero. Gptzero. https://gptzero.me/, 2023.
15. Fröhling L and Zubiaga A. Feature-based detection of automated language models: tackling gpt-2, gpt-3 and grover. *PeerJ Computer Science* 2021; 7: e443. DOI:https://doi.org/10.7717/peerj-cs.443.
16. Nguyen-Son HQ, Tieu NDT, Nguyen HH et al. Identifying computer-generated text using statistical analysis. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, pp. 1504–1511. DOI:10.1109/APSIPA.2017.8282270.
17. Mitrović S, Andreoletti D and Ayoub O. Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text. *arXiv preprint arXiv:230113852* 2023; .
18. Guo B, Zhang X, Wang Z et al. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:230107597* 2023; .
19. Dou Y, Forbes M, Koncel-Kedziorski R et al. Is gpt-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text. In *Proceedings of*

the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, pp. 7250–7274. DOI:10.18653/v1/2022.acl-long.501.

20. Ippolito D, Duckworth D, Callison-Burch C et al. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 1808–1822. DOI:10.18653/v1/2020.acl-main.164.
21. Jawahar G, Abdul-Mageed M and Lakshmanan LV. Automatic detection of machine generated text: A critical survey. *arXiv preprint arXiv:201101314* 2020; .
22. Gehrmann S, Strobelt H and Rush AM. Gltr: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Florence, Italy: Association for Computational Linguistics, pp. 111–116. DOI: 10.18653/v1/P19-3019.
23. Crothers E, Japkowicz N, Viktor H et al. Adversarial robustness of neural-statistical features in detection of generative transformers. In *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 1–8. DOI:10.1109/IJCNN55064.2022.9892269.
24. Solaiman I, Brundage M, Clark J et al. Release strategies and the social impacts of language models. *arXiv preprint arXiv:190809203* 2019; .
25. Holtzman A, Buys J, Du L et al. The curious case of neural text degeneration. *arXiv preprint arXiv:190409751* 2019; .
26. Pu J, Sarwar Z, Abdullah SM et al. Deepfake text detection: Limitations and opportunities. In *2023 IEEE Symposium on Security and Privacy (SP)*. Los Alamitos, CA, USA: IEEE Computer Society, pp. 19–36. DOI:10.1109/SP46215.2023.00002.
27. Stiff H and Johansson F. Detecting computer-generated disinformation. *International Journal of Data Science and Analytics* 2022; 13(4): 363–383. DOI:10.1007/s41060-021-00299-5.
28. Radford A, Wu J, Child R et al. Language models are unsupervised multitask learners. *OpenAI blog* 2019; 1(8): 9.
29. Bakhtin A, Gross S, Ott M et al. Real or fake? learning to discriminate machine from human generated text. *arXiv preprint arXiv:190603351* 2019; .
30. Dugan L, Ippolito D, Kirubarajan A et al. Real or fake text? investigating human ability to detect boundaries between human-written and machine-generated text. In *The 37th AAAI Conference on Artificial Intelligence (AAAI 2023)*. Washington DC, USA, p. 104979. URL https://par.nsf.gov/biblio/10390703.
31. Clark E, August T, Serrano S et al. All that's 'human'is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, pp. 7282–7296. DOI:10.18653/v1/2021.acl-long.565.
32. Ming Y, Qu H and Bertini E. Rulematrix: Visualizing and understanding classifiers with rules. *IEEE Transactions on Visualization and Computer Graphics* 2019; 25(1): 342–352. DOI:10.1109/TVCG.2018.2864812.
33. Liu M, Shi J, Li Z et al. Towards better analysis of deep convolutional neural networks. *IEEE Transactions on Visualization and Computer Graphics* 2017; 23(1): 91–100. DOI:10.1109/TVCG.2016.2598831.
34. Wang ZJ, Turko R, Shaikh O et al. Cnn explainer: Learning convolutional neural networks with interactive visualization. *IEEE Transactions on Visualization and Computer Graphics* 2021; 27(2): 1396–1406. DOI:10.1109/TVCG.2020.3030418.
35. Du M, Liu N and Hu X. Techniques for interpretable machine learning. *Communications of the ACM* 2019; 63(1): 68–77. DOI:10.1145/3359786.
36. Carvalho DV, Pereira EM and Cardoso JS. Machine learning interpretability: A survey on methods and metrics. *Electronics* 2019; 8(8). DOI:10.3390/electronics8080832.
37. Došilović FK, Brčić M and Hlupić N. Explainable artificial intelligence: A survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. pp. 0210–0215. DOI:10.23919/MIPRO.2018.8400040.
38. Adadi A and Berrada M. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access* 2018; 6: 52138–52160. DOI:10.1109/ACCESS.2018.2870052.
39. Hohman F, Kahng M, Pienta R et al. Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE Transactions on Visualization and Computer Graphics* 2019; 25(8): 2674–2693. DOI:10.1109/TVCG.2018.2843369.
40. Yuan J, Chen C, Yang W et al. A survey of visual analytics techniques for machine learning. *Computational Visual Media* 2021; 7: 3–36.
41. Sacha D, Kraus M, Keim DA et al. Vis4ml: An ontology for visual analytics assisted machine learning. *IEEE Transactions on Visualization and Computer Graphics* 2019; 25(1): 385–395. DOI:10.1109/TVCG.2018.2864838.
42. Yang W, Liu M, Wang Z et al. Foundation models meet visualizations: Challenges and opportunities. *arXiv preprint arXiv:231005771* 2023; .
43. Molnar C. *Interpretable machine learning*. Lulu.com, 2020.
44. Eiras-Franco C, Guijarro-Berdiñas B, Alonso-Betanzos A et al. A scalable decision-tree-based method to explain interactions in dyadic data. *Decision Support Systems* 2019; 127: 113–141. DOI:https://doi.org/10.1016/j.dss.2019.113141.
45. Gupta B, Rawat A, Jain A et al. Analysis of various decision tree algorithms for classification in data mining. *International Journal of Computer Applications* 2017; 163(8): 15–19. DOI:10.5120/ijca2017913660.
46. Keneni BM, Kaur D, Al Bataineh A et al. Evolving rule-based explainable artificial intelligence for unmanned aerial vehicles. *IEEE Access* 2019; 7: 17001–17016. DOI:10.1109/ACCESS.2019.2893141.
47. Dingen D, van't Veer M, Houthuizen P et al. Regressionexplorer: Interactive exploration of logistic regression models with subgroup analysis. *IEEE Transactions on Visualization and Computer Graphics* 2019; 25(1): 246–255. DOI:10.1109/TVCG.2018.2865043.
48. Neto MP and Paulovich FV. Explainable matrix - visualization for global and local interpretability of random forest classification ensembles. *IEEE Transactions on*

Visualization and Computer Graphics 2021; 27(2): 1427–1437. DOI:10.1109/TVCG.2020.3030354.

49. Cheng F, Ming Y and Qu H. Dece: Decision explorer with counterfactual explanations for machine learning models. *IEEE Transactions on Visualization and Computer Graphics* 2021; 27(2): 1438–1447. DOI:10.1109/TVCG.2020.3030342.
50. Lundberg SM and Lee SI. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.
51. Ribeiro MT, Singh S and Guestrin C. ”why should i trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery, pp. 1135—1144. DOI:10.1145/2939672.2939778.
52. Ribeiro MT, Singh S and Guestrin C. Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence* 2018; 32(1). DOI:10.1609/aaai.v32i1.11491. URL https://ojs.aaai.org/index.php/AAAI/article/view/11491.
53. Feng Y, Wang X, Pan B et al. XNLI: Explaining and diagnosing nli-based visual data analysis. *IEEE Transactions on Visualization and Computer Graphics* 2023; : 1–14DOI:10.1109/TVCG.2023.3240003.
54. Shapley LS et al. A value for n-person games 1953; .
55. Cheng F, Liu D, Du F et al. Vbridge: Connecting the dots between features and data to explain healthcare models. *IEEE Transactions on Visualization and Computer Graphics* 2022; 28(1): 378–388. DOI:10.1109/TVCG.2021.3114836.
56. Collaris D and Van Wijk J. Strategyatlas: Strategy analysis for machine learning interpretability. *IEEE Transactions on Visualization and Computer Graphics* 2022; : 1–1DOI:10.1109/TVCG.2022.3146806.
57. Zytek A, Liu D, Vaithianathan R et al. Sibyl: Understanding and addressing the usability challenges of machine learning in high-stakes decision making. *IEEE Transactions on Visualization and Computer Graphics* 2022; 28(1): 1161–1171. DOI:10.1109/TVCG.2021.3114864.
58. Horvitz E. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, pp. 159–166. DOI:10.1145/302979.303030.
59. Cook K, Cramer N, Israel D et al. Mixed-initiative visual analytics using task-driven recommendations. In *2015 IEEE conference on visual analytics science and technology (VAST)*. IEEE, pp. 9–16. DOI:10.1109/VAST.2015.7347625.
60. Pérez-Messina I, Ceneda D, El-Assady M et al. A typology of guidance tasks in mixed-initiative visual analytics environments. In *Computer Graphics Forum*, volume 41. Wiley, pp. 465–476. DOI:10.1111/cgf.14555.
61. Wenskovitch J, Fallon C, Miller K et al. Beyond visual analytics: Human-machine teaming for ai-driven data sensemaking. In *2021 IEEE Workshop on TRust and EXpertise in Visual Analytics (TREX)*. IEEE, pp. 40–44. DOI:10.1109/TREX53765.2021.00012.
62. Crouser RJ, Franklin L, Endert A et al. Toward theoretical techniques for measuring the use of human effort in visual analytic systems. *IEEE transactions on visualization and computer graphics* 2016; 23(1): 121–130. DOI:10.1109/TVCG.2016.2598460.
63. Husain F, Proulx P, Chang MW et al. A mixed-initiative visual analytics approach for qualitative causal modeling. In *2021 IEEE Visualization Conference (VIS)*. IEEE, pp. 121–125. DOI:10.1109/VIS49827.2021.9623318.
64. Cabrera ÁA, Epperson W, Hohman F et al. Fairvis: Visual analytics for discovering intersectional bias in machine learning. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, pp. 46–56. DOI:10.1109/VAST47406.2019.8986948.
65. Feng Y, Chen J, Huang K et al. iPoet: interactive painting poetry creation with visual multimodal analysis. *Journal of Visualization* 2022; 25(3): 671–685. DOI:10.1007/s12650-021-00780-0.
66. Zhou J, Wang X, Wong KK et al. Dpviscreator: Incorporating pattern constraints to privacy-preserving visualizations via differential privacy. *IEEE Transactions on Visualization and Computer Graphics* 2023; 29(1): 809–819. DOI:10.1109/TVCG.2022.3209391.
67. Rietz T and Maedche A. Cody: An ai-based system to semi-automate coding for qualitative research. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI ’21, New York, NY, USA: Association for Computing Machinery. ISBN 9781450380966. DOI:10.1145/3411764.3445591. URL https://doi.org/10.1145/3411764.3445591.
68. Lu J, Pan B, Chen J et al. Agentlens: Visual analysis for agent behaviors in llm-based autonomous systems. *arXiv preprint arXiv:240208995* 2024; .
69. Liu Y, Wen Z, Weng L et al. Sprout: Authoring programming tutorials with interactive visualization of large language model generation process. *arXiv preprint arXiv:231201801* 2023; .
70. Zhang W, Wong KK, Wang X et al. Cohortva: A visual analytic system for interactive exploration of cohorts based on historical data. *IEEE Transactions on Visualization and Computer Graphics* 2023; 29(1): 756–766. DOI:10.1109/TVCG.2022.3209483.
71. Zhang W, Wong JK, Chen Y et al. Scrolltimes: Tracing the provenance of paintings as a window into history. *arXiv* 2023; DOI:10.48550/arXiv.2306.08834.
72. Lin Y, Wong KK, Wang Y et al. Taxthemis: Interactive mining and exploration of suspicious tax evasion groups. *IEEE Transactions on Visualization and Computer Graphics* 2021; 27(2): 849–859. DOI:10.1109/TVCG.2020.3030370.
73. Wong KK, Wang X, Wang Y et al. Anchorage: Visual analysis of satisfaction in customer service videos via anchor events. *IEEE Transactions on Visualization and Computer Graphics* 2023; : 1–13DOI:10.1109/TVCG.2023.3245609.
74. Paiva JGS, Schwartz WR, Pedrini H et al. An approach to supporting incremental visual data classification. *IEEE Transactions on Visualization and Computer Graphics* 2015; 21(1): 4–17. DOI:10.1109/TVCG.2014.2331979.
75. Xiang S, Ye X, Xia J et al. Interactive correction of mislabeled training data. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*. pp. 57–68. DOI:10.1109/VAST47406.2019.8986943.

76. Yang W, Guo Y, Wu J et al. Interactive reweighting for mitigating label quality issues. *IEEE Transactions on Visualization and Computer Graphics* 2024; 30(3): 1837–1852. DOI:10.1109/TVCG.2023.3345340.
77. Wall E, Das S, Chawla R et al. Podium: Ranking data using mixed-initiative visual analytics. *IEEE transactions on visualization and computer graphics* 2017; 24(1): 288–297. DOI:10.1109/TVCG.2017.2745078.
78. Pister A, Buono P, Fekete JD et al. Integrating prior knowledge in mixed-initiative social network clustering. *IEEE Transactions on Visualization and Computer Graphics* 2020; 27(2): 1775–1785. DOI:10.1109/TVCG.2020.3030347.
79. Felix C, Dasgupta A and Bertini E. The exploratory labeling assistant: Mixed-initiative label curation with large document collections. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. New York, NY, USA: Association for Computing Machinery, pp. 153–164. DOI:10.1145/3242587.3242596.
80. Choi M, Park C, Yang S et al. Aila: Attentive interactive labeling assistant for document classification through attention-based deep neural networks. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. New York, NY, USA: Association for Computing Machinery, pp. 1–12. DOI:10.1145/3290605.3300460.
81. Alsaid A and Lee JD. The datascope: A mixed-initiative architecture for data labeling. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 66. SAGE Publications Sage CA: Los Angeles, CA, pp. 1559–1563. DOI:10.1177/1071181322661356.
82. Bernard J, Hutter M, Zeppelzauer M et al. Comparing visual-interactive labeling with active learning: An experimental study. *IEEE transactions on visualization and computer graphics* 2017; 24(1): 298–308. DOI:10.1109/TVCG.2017.2744818.
83. Bernard J, Zeppelzauer M, Sedlmair M et al. Vial: a unified process for visual interactive labeling. *The Visual Computer* 2018; 34: 1189–1207. DOI:10.1007/s00371-018-1500-3.
84. Bernard J, Hutter M, Sedlmair M et al. A taxonomy of property measures to unify active learning and human-centered approaches to data labeling. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2021; 11(3-4): 1–42. DOI:10.1145/3439333.
85. Sevastjanova R, Jentner W, Sperrle F et al. Questioncomb: A gamification approach for the visual explanation of linguistic phenomena through interactive labeling. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2021; 11(3-4): 1–38. DOI:10.1145/3429448.
86. Zhang X, Xuan X, Dima A et al. Labelvizier: Interactive validation and relabeling for technical text annotations. In *2023 IEEE 16th Pacific Visualization Symposium (PacificVis)*. pp. 167–176. DOI:10.1109/PacificVis56936.2023.00026.
87. Narechania A, Karduni A, Wesslen R et al. Vitality: Promoting serendipitous discovery of academic literature with transformers & visual analytics. *IEEE Transactions on Visualization and Computer Graphics* 2022; 28(1): 486–496. DOI:10.1109/TVCG.2021.3114820.
88. Grammarly. Grammarly. https://www.grammarly.com/, 2023.
89. Gleason JB and Ratner NB. *The development of language*. Plural Publishing, 2022.
90. Beltagy I, Lo K and Cohan A. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 3615–3620. DOI:10.18653/v1/D19-1371.
91. Christiano PF, Leike J, Brown T et al. Deep reinforcement learning from human preferences. *Advances in neural information processing systems* 2017; 30: 4302–4310. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf.
92. Monti F, Frasca F, Eynard D et al. Fake news detection on social media using geometric deep learning. *arXiv preprint arXiv:190206673* 2019; .
93. Nourani M, King J and Ragan E. The role of domain expertise in user trust and the impact of first impressions with intelligent systems. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 2020; 8(1): 112–121. DOI:https://doi.org/10.1609/hcomp.v8i1.7469.
94. McInnes L, Healy J and Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:180203426* 2018; .
95. Ke G, Meng Q, Finley T et al. Lightgbm: A highly efficient gradient boosting decision tree. In Guyon I, Luxburg UV, Bengio S et al. (eds.) *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.
96. Yang L, Xiong C, Wong KK et al. Explaining with examples lessons learned from crowdsourced introductory description of information visualizations. *IEEE Transactions on Visualization and Computer Graphics* 2023; 29(3): 1638–1650. DOI:10.1109/TVCG.2021.3128157.
97. Aas K, Jullum M and Løland A. Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *Artificial Intelligence* 2021; 298: 103502. DOI:https://doi.org/10.1016/j.artint.2021.103502. URL https://www.sciencedirect.com/science/article/pii/S0004370221000539.
98. Ehsan U, Wintersberger P, Liao QV et al. Human-centered explainable ai (hcxai): Beyond opening the black-box of ai. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI EA '22, New York, NY, USA: Association for Computing Machinery. ISBN 9781450391566. DOI:10.1145/3491101.3503727. URL https://doi.org/10.1145/3491101.3503727.
99. Zhu H, Zhu M, Feng Y et al. Visualizing large-scale high-dimensional data via hierarchical embedding of knn graphs. *Visual Informatics* 2021; 5(2): 51–59.
100. Kirchenbauer J, Geiping J, Wen Y et al. A watermark for large language models. *arXiv preprint arXiv:230110226* 2023; .
101. Mitchell E, Lee Y, Khazatsky A et al. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *arXiv preprint arXiv:230111305* 2023; .