Human-AI Interactions in the Communication Era: Autophagy Makes Large Models Achieving Local Optima

Anonymous ACL submission

Abstract

The increasing significance of large language and multimodal models in societal information processing has ignited debates on social safety and ethics. However, few studies have approached the analysis of these limitations from the comprehensive perspective of human and artificial intelligence system interactions. This study investigates biases and preferences when humans and large models are used as key links in communication. To achieve this, we design a multimodal dataset and three different experiments to evaluate generative models in their roles as producers and disseminators of information. Our main findings highlight that synthesized information is more likely to be incorporated into model training datasets and messaging than human-generated information. Additionally, large models, when acting as transmitters of information, tend to selectively modify and lose specific content. Conceptually, we present two realistic models of autophagic ("self-consumption") loops to account for the suppression of human-generated information in the exchange of information between humans and AI systems. We generalize the declining diversity of social information and the bottleneck in model performance caused by the above trends to the local optima of large models.

1 Introduction

Large models including large language model(LLM)s (OpenAI, 2023; Bai et al., 2022; Touvron et al., 2023; Zeng et al., 2022) and rapidly advancing large multimodal models (Yang et al., 2023; Yin et al., 2023), are emerging as transformative tools, reshaping our world in ways that are both awe-inspiring and formidable. A significant aspect is that they are becoming an integral part of the dissemination of viewpoints and information in human society, exhibiting attributes such as connecting, engaging, and interacting. Meanwhile, their inherent limitations raise significant concerns. Extensive Research has highlighted key issues such as discrimination(Navigli et al., 2023), hallucinatory(Huang et al., 2023b) outputs, and lack of interpretability(Zhao et al., 2023). However, as a novel and significant component of the communication era (Edwards et al., 2016), the widespread use of large models and their inherent limitations have not been fully explored in terms of their impact on human societal dissemination, including but not limited to language and visual information, and modes of interaction.

Our work proposes two realistic models for autophagous ("self-consuming") loops (refer to Fig 1 and Fig 2) based on the way humans construct and use large models. These two loops emphasize the fact that the interaction process between AI systems and humans will lead to synthetic data (or AI-generated data) being more likely to win in messaging compared to real human data. This causes in a growing prevalence of synthetic data within model training datasets and throughout human society. In this scenario, models are predominantly trained on synthetic data, and humans subsequently build upon this synthetic foundation. We describe this phenomenon as "self-consumption".

Our motivations stem from the following key facts:

(1) Large models are being extensively utilized across various domains (Kaddour et al., 2023), and even crowd-sourced annotators are heavily relying on generative AI for decision-making processes (Veselovsky et al., 2023).

(2) The Internet, being a direct source of training data, implies that contemporary models are increasingly trained on AI-synthesized data unwittingly (Alemohammad et al., 2023; Shumailov et al., 2023a; Veselovsky et al., 2023).

(3) To reduce training costs, many studies have opted to make the models themselves the generators and selectors of their training data (Li et al.,

2023; Huang et al., 2023a).

These facts remind us more cost-effective model training processes and wider applications are inevitable trends of large model development, so we need to discuss not only the technology of large models themselves but also the roles they play in their iterations and the development of human society. It's important to note that without ensuring a consistent presence of real human-generated data, large models may increasingly rely on their own generated datasets due to a lack of fresh data. This can result in stagnating improvements in model performance We define this as the large model falling into "local optimum". In this paper, we concentrate on the roles of large models as creators and distributors of information. We explore how they handle data from various sources, each with unique characteristics, and examine how this data is either augmented or suppressed. The following summarizes the key contributions and findings of this paper:

First, we introduce two autophagous ("self-consuming") loops involving both large models and humans. They are designed to analyze how the interactions and preferences between humans and generative artificial intelligence lead to the dominance of synthesized data in the selection of model training datasets and information dissemination. To validate the autophagous ("self-consuming") loops proposed, we conducted simulation experiments focusing on two key aspects: 1) We examined the preferences of both humans and large language models(LLMs) in the evaluation and filtration of information as part of the dissemination process. 2) We investigated the potential drawbacks associated with using generative models for enhancing and transferring information.

Second, we designed three distinct experiments to prove the above realistic model. To begin with, we prompt LLMs to generate answers to specific questions, using predetermined scoring criteria, followed by cross-validation scoring. Similarly, we instruct crowdsourced annotators to evaluate the generated question-answer pairs using the same criteria. This experiment aimed to ascertain the preferences of language models and humans in evaluating and filtering information. Our findings indicate that these models tend to overrate their own answers and undervalue human responses. In addition, to eliminate potential biases from lengthy contexts and granular scoring criteria, we set up a real-world scenario simulation. This experiment demonstrated that, compared to authentic human data, synthesized data is more likely to prevail in information filtration processes. This tendency was also observed in results from crowdsourced annotations. Then, we conducted an "AI-washing" experiment to illustrate how generative models overlook and alter initial information details during transmission in a cyclical process.

To conduct our above experiments, we have also constructed a dataset, comprising both textual and visual elements. We manually screened the most answered questions in Stack Overflow and Quora, including psychology, books, mathematics, physics, and other fields. At the same time, we selected fragments from the novel corpus for anonymization processing to study the behavior of the language model when delivering real human-generated data. On the visual data set, we sampled and cleaned the ILSVRC(Russakovsky et al., 2015) to ensure the diversity of image clarity and classification.

We aim to offer a novel perspective on the impact of large models, particularly in their role as intermediaries in human societal information dissemination, and the potential hazards this entails. Our investigation reveals that in the cycle of information exchange between humans and large models, these models exhibit a strong preference in deciding which features to amplify or suppress. This leads to the local optimum, where real human data increasingly struggles to enter model training and information exchange. This issue not only creates performance bottlenecks in large models but also makes it increasingly difficult for humans to intervene in the model's generative processes and information transmission.

2 Methodology

Inspired by perspectives on the communication era as proposed by (Edwards et al., 2016), we present a novel viewpoint to study the impact of large models. Specifically, we conceptualize large models as integral components of human societal information and opinion dissemination, characterized by attributes such as connecting, engaging, and interacting. Our methods are mainly divided into (1) Drawing upon human behaviors in utilizing large models, we design a realistic model of autophagy and self-consumption. (2) We devise new datasets and experimental studies to demonstrate how real-world data distribution is influenced by the use of

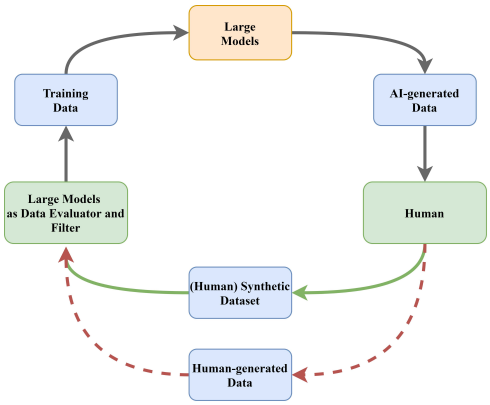

Figure 1: Autophagous ("self-consuming") Loop of Large Models

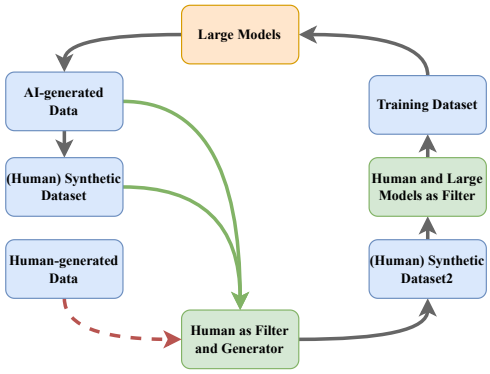

Figure 2: Autophagous ("self-consuming") Loop of Human

large models.

2.1 Realistic Models forAutophagous ("self-consuming") Loops

Previous work by (Alemohammad et al., 2023) and (Shumailov et al., 2023b) analyzed the decline in quality and diversity of generated data by visual generative models, a phenomenon known as Model Autophagy Disorder (MAD), particularly in contexts lacking fresh, real training data. However, they used simulated experiments to demonstrate the decline in model performance, but did not deeply analyze why real data is increasingly scarce, nor the impact of this phenomenon on the flow of information in human society. Our work aims to fill this gap and further extend to large language and multimodal models.

Drawing from the classic communication theory of the Ritual view as proposed in (Carey, 2008), we redefine the relationship between large models and human societal information dissemination (see Appendix B for details). As shown in Figures 1 and 2, both large models and humans can act as generators and filters of information in the Human-AI communication system. However, this system is prompting machine-learning algorithms to encode all the stereotypes, inequalities, and power asymmetries that exist in human society (Birhane, 2022). For example, women with darker skin are more likely to be misclassified in gender classification compared to men with lighter skin, which is due to the majority of samples in the training datasets being subjects with lighter skin tones(Buolamwini and Gebru, 2018), The biased information generation and transmission processes of large models and humans will exacerbate such phenomena.

Therefore, we propose two realistic models for autophagous ("self-consuming") loops to simulate the interaction between humans and large models, in order to study their biases as generators and transmitters of information and to demonstrate that this will ultimately lead to the loss of diversity in model training data and human societal datasets, as well as the exacerbation of bias. Below we the two realistic models in more detail.

Figure 1 delineates the cyclical influence of large models in the data processing lifecycle. Once decoded by these models, training data undergoes a transformation through either algorithmic refinement or human curation, resulting in what we term "synthetic data". This contrasts with "human-generated data", which originates directly from human input and is typically less structured. Large models assess both types of data, but due to biases inherent in their training, may preferentially filter and elevate synthetic data over human-generated data for use in future learning cycles, we design experiments in the next section to prove this conclusion. Furthermore, Figure 1 explicates that the human role in this cycle is not purely passive. Humans, influenced by the outputs of large models, may unknowingly prioritize synthetic data due to its processed nature, which seems more immediately usable or relevant(Veselovsky et al., 2023). This preferential feedback loop can inadvertently lead to the diminishment of raw, human-generated data in the pool of resources, as it may be perceived as less refined.

Figure 2 builds upon this by showcasing the specific behaviors of humans when interacting with large models. It presents a more detailed view of human-agent interactions, highlighting the empirical finding that humans tend to favor data pro-

Initial Paragraph	$N = 1$	$N = 2$
An individual who was standing beyond the railing, in the free space around the marble table, and whom no one had yet caught sight of, since his long, thin body was completely sheltered from every visual ray by the diameter of the pillar against which he was leaning; this individual, we say, tall, gaunt, pallid, blond, still young, although already wrinkled about the brow and cheeks, with brilliant eyes and a smiling mouth, clad in garments of black serge, worn and shining with age, approached the marble table, and made a sign to the poor sufferer.	In the open expanse around the marble table, a figure stood just beyond the railing, elusive to the gaze of those present. His lengthy, slender frame remained entirely shielded from view by the pillar's breadth against which he leaned. This person, tall and lean, possessed a pallid countenance, complemented by blond locks. Though still in the bloom of youth, the furrows etched across his brow and cheeks spoke of time's passage. His eyes sparkled brilliantly, and a gentle smile graced his lips. Draped in well-worn black serge, which bore a subdued radiance earned through years of use, he approached the marble table and gestured to the afflicted soul.	In the expansive openness surrounding the marble table, a figure stood just beyond the railing, elusive to the gazes of those present. His lengthy and slender frame remained completely shielded from view by the width of the pillar against which he leaned. This tall and lean individual possessed a pallid countenance, enhanced by his blond locks. Although still in the prime of youth, the lines etched across his forehead and cheeks spoke of the passage of time. His eyes shimmered with brilliance, and a gentle smile graced his lips. Adorned in a well-worn black serge, which exuded a muted radiance earned through years of use, he approached the marble table and gestured towards the troubled soul.
$N = 3$	$N = 4$	$N = 5$
In the expansive area encircling the marble table, a figure stood just beyond the railing, evading the gaze of those present. His lithe, elongated form remained entirely veiled from view by the pillar's substantial girth against which he leaned. This individual, statuesque and slender, possessed a pallor that harmonized with his blond locks. Despite being in the prime of his youth, the lines etched across his forehead and cheeks whispered of the passage of time. His eyes gleamed with a brilliant intensity, and a tender smile graced his lips. Cloaked in weathered black serge, which bore a subdued luminance earned through years of wear, he approached the marble table and extended a gesture to the afflicted soul.	In the expansive expanse surrounding the marble table, a figure stood just beyond the railing, eluding the gaze of those present. His lithe, elongated silhouette remained entirely shrouded from view by the pillar's substantial girth against which he leaned. This individual, statuesque and slender, possessed a pallor that blended seamlessly with his blond locks. Despite being in the zenith of his youth, the lines etched across his forehead and cheeks murmured of the passage of time. His eyes shone with a brilliant intensity, and a tender smile graced his lips. Draped in weathered black serge, which bore a subdued radiance earned through years of wear, he approached the marble table and extended a gesture to the afflicted soul.	In the expansive expanse surrounding the marble table, a figure stood just beyond the railing, eluding the gaze of those in attendance. His graceful, elongated silhouette remained entirely shrouded by the substantial girth of the pillar against which he leaned. This person, statuesque and slender, possessed a complexion that seamlessly blended with his blond locks. Despite being in the zenith of his youth, the lines etched across his forehead and cheeks whispered of the passage of time. His eyes shimmered with a brilliant intensity, and a tender smile graced his lips. Draped in weathered black serge, which bore a subdued radiance earned through years of wear, he approached the marble table and extended a reassuring gesture to the troubled soul.

Table 1: Example of AI-Washing experiment for text. N represents the number of times the large language model is used for refinement, with each changed part highlighted.

duced by large models. Our experiments in the following section suggest that without transparent data provenance, humans may prefer these models' outputs, further contributing to the cyclical bias toward synthetic data. The relationship between these two loops is symbiotic; while Figure 1 provides an overview of the data cycle in a large model training loop, Figure 2 zooms in on the human aspect, offering a microcosm of human preference in the Human-AI communication.

2.2 Rationality and Risks of Autophagous Loops

In this section, we describe how we prove the above realistic models and the risks posed by such autophagous loops. Firstly, the core proposition of our reality model is that large models and humans cannot maintain objectivity and impartiality as part of the information dissemination loop. Furthermore, there is a clear preference for synthetic data, ultimately leading to a diminishing proportion of human real data in the information cycle.

To demonstrate the inhibitory and promotive phenomena of information transmission within Autophagous Loops (as indicated by the colored segments in Fig 1 and Fig 2). We employed mainstream LLMs to generate question-answer pairs based on prompts and instructed them to perform cross-scoring. In our result analysis, we focused on examining the consistency and bias of humans and language models in adhering to scoring standards. We also compared differences between various model families (such as GPTs and LLaMAs) as answer generators and scorers.

Additionally, to mitigate the impact of extended context and scoring standards, we designed a simulated testing scenario to analyze which is more likely to prevail in the cycle of information dissemination in real-world scenarios: real-human answers or AI-generated answers.

Finally, to explore the risks posed by large models and humans as information generators in our realistic models, we conducted an "AI-washing" experiment to observe the changes in real data after multiple AI refinements. Our primary analysis focused on the loss of information diversity and the large models' varying enhancement and weakening of different information. For instance, repeatedly refining an animal image using SDXL eventually transforms it into a human character (see Fig 3). This bias leads to the deepening of stereotypes in human-AI information exchanges.

3 Experimental Study

Based on our discussion in Section 2.2, in this section we evaluate the preferences of humans and language models in information selection, thereby

analyzing how real human-generated data is suppressed in the human-language model information dissemination loop.

3.1 Experimental Setup

Models and Dataset We employed six LLMs to generate and evaluate response data based on specific instructions. These models include ChatGPT(Li et al., 2022), GPT-4(OpenAI, 2023), Claud2[1], Llama-2-70b-chat(Touvron et al., 2023), PaLM2 chat-bison[2], and Solar-0-70b-16bit[3], each representing different architectural frameworks. The focus on larger models in our experiments is due to their superior capability in instruction adherence and context length handling, which we found lacking in smaller-scale models. For computer vision tasks, we utilized the open-source model StableDiffusionXL(Podell et al., 2023). In assessing textual diversity, the models bge-large-zh-v1.5(Xiao et al., 2023) and bge-large-en-v1.5(Xiao et al., 2023) were selected as embedding models.

Our experimental dataset comprised manually curated text and image sets. Initially, we hand-picked 100 diverse question-answer pairs from StuckOverflow and Quaro as the seed data. Subsequently, for each instruction, the large models generated initial responses. Based on the self-alignment approach proposed by (Li et al., 2023), these responses were further processed to create datasets rated as either 1 (lowest) or 5 (highest) in terms of quality. The prompts used for generating these diverse responses are detailed in Appendix C. Finally, we constructed a data set consisting of 1900 question-answer pairs. Specifically, our dataset consists of a series of 22 tuples, each structured as follows:

$$T_j = \{d, Q, D, A\} \cup \bigcup_{i=0}^{5} \{A_{\text{model}_i}, A_{\text{model}_i\text{score5}}, A_{\text{model}_i\text{score1}}\} \quad (1)$$

Appendix D provides explanations of the corresponding mathematical notation, showing more details about the distribution of the dataset.

The text dataset construction process begins with the selection of passages from classic literature known for their rich stylistic features and thematic significance, where the English dataset is excerpted from the pile books3(Gao et al., 2020), and the Chinese passages are selected from WebNovel. A meticulous anonymization process is employed to prevent the large language model from identifying the textual sources, this involves the alteration of recognizable names, places, and events. The image dataset was constructed by carefully selecting a subset of images from the comprehensive ILSVRC(Russakovsky et al., 2015) dataset, as well as other web image data, with selected categories covering a wide range of topics and scenarios, providing a broad range of visual features and complexity.

Cross-scoring Experiment To demonstrate the inhibitory and promotive phenomena of information transmission within autophagous loops, we design cross-scoring experiments with question-answer pairs. We focus on whether LLMs and humans can remain impartial when filtering and transmitting information, and if not, what kind of bias they have.

We prompt each model to assess not only the answers generated by other models but also those produced by humans. The scoring range was between 1 to 5 (see Table 9). For instance, in the "five-score answers" segment, an answer generated by ChatGPT would be evaluated by other LLMs like GPT4 and Claud2, assigning a score within the 1-5 range based on its quality. At the same time, we found fifty crowdsourced annotators to rate all question-answer pairs equally, find the scoring criteria in Table 10. We recorded the average scores for all valid samples, excluding instances where the models refused to respond.

Exam Scenario Simulation In this experiment, we present our experimental design to answer the following question: Human-generated or AI-generated answers, which one wins in information screening and filtering? As depicted in Figure 7, we design an examination scenario where answers generated by ChatGPT and Claud2, alongside those produced by humans, were anonymized to mitigate any bias. To further eliminate the potential influence of the sequence in which the answers were presented, we randomized their order. Both language models and human participants, assuming the role of experts, were then tasked with assigning scores to these answers on a percentile scale, and choosing the best answer.

AI Washing For AI washing experiments, we aim

[1] https://www.anthropic.com/index/claude-2

[2] https://blog.google/technology/ai/google-palm-2-ai-large-language-model/

[3] https://huggingface.co/upstage/SOLAR-0-70b-16bit

Originally Generated Answer								
Scorer / Generator	ChatGPT	GPT4	Claud2	Llama-2-70b-chat	PaLM-2-chat-bison	Solar-0-70b-16bit	Human	Average
ChatGPT	**4.33**	**4.29**	3.88	4.25	3.92	4.17	**2.48**	3.90
GPT4	**4.63**	**4.56**	4.04	4.41	3.95	4.60	**2.77**	4.14
Claud2	3.92	3.97	4.00	4.00	3.95	3.97	**3.36**	3.88
Llama-2-70b-chat	3.91	3.99	3.82	4.00	3.61	3.90	**3.23**	3.78
PaLM-2-chat-bison	3.99	4.05	3.72	4.22	3.60	3.77	**3.57**	3.85
Solar-0-70b-16bit	4.10	4.35	4.05	4.16	4.01	4.12	**2.59**	3.91
Human	4.75	4.79	4.50	4.18	4.28	4.17	**3.58**	**4.32**
Best Quality Answer								
Scorer / Generator	ChatGPT	GPT4	Claud2	Llama-2-70b-chat	PaLM-2-chat-bison	Solar-0-70b-16bit	Human	Average
ChatGPT	4.24	4.28	4.41	3.80	4.21	4.20	-	4.19
GPT4	**4.52**	**4.75**	4.20	4.11	4.00	4.36	-	4.32
Claud2	3.92	3.98	4.21	4.20	4.01	3.97	-	4.04
Llama-2-70b-chat	3.91	4.03	4.26	4.07	4.30	3.95	-	4.09
PaLM-2-chat-bison	3.98	4.23	4.42	3.84	4.26	3.98	-	4.12
Solar-0-70b-16bit	4.34	4.43	4.42	4.33	4.28	4.11	-	4.32
Human	4.23	**4.92**	4.30	4.20	4.07	4.26	-	**4.33**
Worst Quality Answer								
Scorer / Generator	ChatGPT	GPT4	Claud2	Llama-2-70b-chat	PaLM-2-chat-bison	Solar-0-70b-16bit	Human	
ChatGPT	**3.13**	1.33	1.27	1.27	2.83	2.21	-	2.01
GPT4	**3.19**	1.40	1.29	1.33	2.98	1.70	-	1.98
Claud2	**4.08**	**3.23**	**3.71**	1.76	**3.85**	**3.77**	-	3.40
Llama-2-70b-chat	2.69	1.06	2.17	1.78	2.27	2.11	-	2.01
PaLM-2-chat-bison	2.65	1.23	1.28	1.69	2.73	2.31	-	1.98
Solar-0-70b-16bit	**3.28**	1.26	1.89	2.40	2.37	2.40	-	2.27
Human	1.76	2.31	1.24	1.33	2.00	1.82	-	**2.09**

Table 2: Scoring analysis of language models and human responses. The table presents average scores out of a five-point scale, assigned by both models and human evaluators, to the generated answers. These scores are calculated based on the criteria outlined in Appendix G and Appendix F. The table is organized into three sections: originally generated answers, best quality answers($A_{\text{model}_i\text{score5}}$ in our dataset), and worst quality answers($A_{\text{model}_i\text{score1}}$ in our dataset), providing a comprehensive view of the evaluative trends across different generators and scorers. A in our dataset, represented in the table as raw data generated by humans.

to answer the following question: Are large models faithful messengers of information? We focus on two aspects: Do these models capture the main points in the information that needs to be conveyed? Whether large models play the role of an important link in the transmission of information can lead to important losses. We instruct SDXL and ChatGPT to process these samples N times respectively. The prompts we used can be found in Table 8.

3.2 Experimental Results

3.2.1 Information Selection Biases in Human and Language Model Interactions

LLMs can identify the quality of response as information filters but have obvious biases against real human data. Table 2 displays the cross-scoring results among various LLMs using a five-point scale. The outcomes from this evaluation allow us to discern: LLMs do have an inherent capability to comprehend grading standards and can adjust the quality of their generated answers based on relevant instructions. However, when scoring data according to these criteria, each model exhibits certain preferences.

Specifically, models tend to assign higher scores to the high-quality answers generated by themselves, particularly for ChatGPT and GPT-4, which both demonstrate high confidence in their own outputs. Our experimental results extend the conclusion that "language models are narcissistic evaluators" (Liu et al., 2023) to the current top-performing LLMs (including both black box and white box models). Also, we find that ChatGPT and GPT-4 exhibit similar characteristics in scoring; they are bolder in giving high or low scores (more likely to give scores of 5 or 1), and they both tend to give lower scores to Claude2 and PaLM 2 chat-bison. At the same time, as the generator of answers, Chatgpt's worst quality answers can still deceptively obtain higher scores from other models, but human crowdsourcing workers can tell them. Furthermore, we observed that Claud2 tends to favor neutral and 'less controversial' ratings, often assigning scores of 3 or 4, notably even for the worst quality answers. This is reflected in the fact that Claud2's score for low-quality answers is much higher than that of other models, and the difference between its score for initial answers and best-quality answers is not obvious.

Llama-2-70b-chat and Solar-0-70B-16bit show similar scoring and generating behaviors, which might be due to Solar-0-70B-16bit being fine-tuned

on the basis of LLaMA2-70B, indicating that the pre-training model has a significant influence on the model's preferences. We observed that prompting the model to generate better answers did not always lead to higher scores in our experiments. Only Claud2 showed significant improvement in the answers compared to the original responses. Conversely, when the model generated poorer answers, the overall effect was notably significant. However, ChatGPT and Palm-2-chat-bision still achieved relatively high scores, a possible reason is could be attributed to the models being highly aligned to avoid producing harmful outputs(Lambert et al., 2022). We leave further investigation to our future work.

Human-generated answers receive lower scores. In Table 2, we observe that human-generated responses received comparatively lower scores from the LLMs. Upon analysis, we found that human evaluators were able to objectively assess the answers produced by the LLMs, demonstrated by the fact that their scores for high-quality answers consistently remained above 4.00 when confronted with the results generated by the large language model, while the results for lower-quality answers remained below 3.00. Furthermore, the scoring behavior of our 50 crowd-sourced annotators toward the answers generated by the large language model was largely consistent. This may be due to the highly structured and standardized nature of answers produced by AI systems aligned with human feedback, which often received higher scores. In contrast, human-generated answers received lower scores from the annotators, with significant variations among different evaluators. This may be attributed to the higher alignment degree between the large model and the human collective, compared to the alignment among individual humans.

3.3 LLMs-generated Answers are more likely to Win in Information Screening

To mitigate the impact of scoring criteria and excessive context length on the evaluative capabilities of large models and human raters, we conducted an experimental examination grading scenario, as presented in Table 3, aligned with those from the previous section: human-generated responses received comparatively lower scores. Furthermore, we engaged ChatGPT, Claud2, and crowd annotators to select the best answers. It was observed that human answers were seldom chosen as the best, indicating a challenge in integrating authentic human-generated responses into the training data for models and the real-world human feedback loop. This finding substantiates the potential risks highlighted in Section 2.2.

	Evaluator		
Generator	ChatGPT	Claud2	Human
Average Score			
ChatGPT	95	90	91.7
Claud2	92	88	90
Human	90	75	80
Selected as Best Answer			
ChatGPT	41	58	66
Claud2	55	42	25
Human	4	0	9

Table 3: Result of exam scenario simulation

3.4 LLMs as Biased Information Transmitters

Figure 3: Processing images multiple times using a generative model as well as a Prompt that only controls the quality of the image can lead to serious biases.

Figure 4: How different images retain and discard different details after 20 Ai-washing experiments. This suggests that generative models tend to emphasize or deemphasize certain features, revealing their unique patterns of recognizing, understanding, and reconstructing visual content.

Large models exhibit inherent biases regarding the manner and content of conveyed information. The findings of our AI washing experiments demonstrate that large models exhibit inherent biases regarding the manner and content of conveyed

information, as evident in our examination of visual and textual data. Let us first illustrate with an example: Table 1 presents the results of our sample data being refined five times by ChatGPT. We observe subtle shifts in the sample's language style and narrative technique. Beginning from the third iteration, the sample seems to have reached a "local optimum", with fewer edits needed, and ultimately, the large language model significantly expanded the original details and altered the style. Similarly,

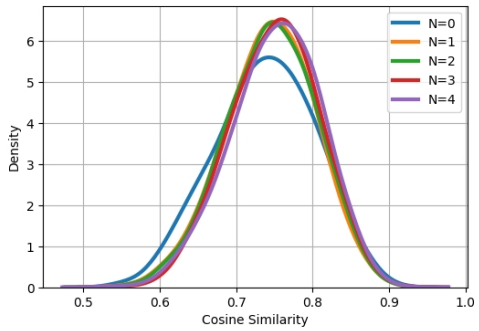

Figure 5: Density distributions of cosine similarity scores for entities processed N times by LLM.

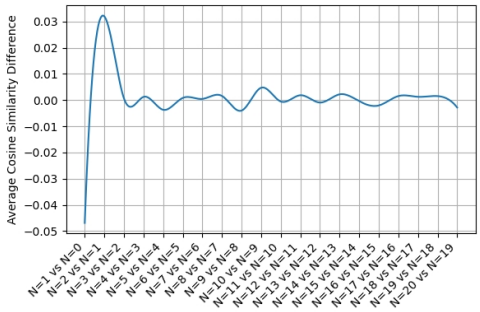

Figure 6: The average cosine similarity difference over a series of 20 iterations.

in the visual model, we observed a comparable phenomenon. By comparing the results of artificial intelligence processes applied to the images described in Figure 3 and Figure 4, Our observations reveal that repeatedly processing images with generative models resembles an information and feature filtration process. The model decides which specific aspects to preserve, diminish, or amplify based on its inherent biases. Specifically, in Figure 3, the model preserved the color distribution of the original image but altered the main subject from a cat to a human portrait. Conversely, in the right two columns of Figure 4, we observe a predominant alteration in image style rather than a change in the primary content. Most notably, the tomato image underwent several iterations with minimal transformation. This inconsistency could be attributed to the fact that while SDXL is renowned for generating high-quality images, the definition of quality in the context of generative models is subjective and heavily influenced by the annotations of the training dataset, as discussed in (Podell et al., 2023). These observations indicate a bias in the model's processing, where certain features are selectively preserved or altered based on the model's training and inherent design. This can lead to images containing specific features taking up a larger proportion of the information loop, such as hand-drawn styles, portraits, close-ups of objects with clear backgrounds, etc.

The risk to the fairness and diversity of information spread Our experiment demonstrates that large models when acting as information disseminators, possess a unique optimization function. They tend to optimize more for real-world data while being more lenient towards the data they generate themselves. Specifically, Figure 5 shows that after processing text using LLMs, the similarity between texts significantly increases, with the lower similarity tail being 'washed out', and the cross-similarity between sentences being enhanced. Meanwhile, in Fig 6, we can see that after $N = 3$ iterations, the average cosine similarity difference of the text relative to the previous round tends to stabilize. This indicates an 'alignment' process of the large models when processing information. We will leave the specific details of this process to future research. As large models become increasingly important in information dissemination, this phenomenon poses significant risks to the fairness and diversity of information dissemination.

4 Conclusion

In our study, a Ritual view of communication theory is utilized to examine large models as generators and disseminators of information within human society. We present two realistic models for autophagous loops and experimentally validate the biases of large models and humans in participating in the information cycle. It is found that AI-generated information tends to prevail in information filtering, whereas real human data is often suppressed, leading to a loss of information diversity. This trend limits next-generation model performance due to fresh data scarcity and threatens the human information ecosystem.

5 Limitations

Selection of Models While we have experimented with LLMs that are available, many outstanding models are worth exploring in the future. These include the GLM family of models(Du et al., 2022), which are known for their innovative architectures, and the MoE-structured Mixtral 8x7B[4], etc. In addition, some open-source multilingual models are also worth investigating, such as the Qwen series of models trained on a large Chinese corpus(Bai et al., 2023) and the Arabic model Jais[5]. Models with different languages, parameter sizes, and architectures exhibit different behaviors. In the field of visual models, more open-source and commercial models are worth investigating, such as Midjourney and DALL-E 3. In future research, we aim to deeply analyze the roles and characteristics of these models as an important part of human social information transfer.

Reliability of Crowdsourced Workers A significant portion of our conclusions is derived from crowd-sourced annotators, sponsored by a start-up company's data annotation department. Of these annotators, 64 % hold graduate degrees in science and engineering, and all possess proficient bilingual reading skills in Chinese and English. However, ensuring that their existing AI knowledge does not bias their judgments remains challenging. Additionally, the distribution of our annotators in the real world varies from the general user base of generative models. There is also an ongoing debate about the reliability of crowd-sourced workers(Spurling et al., 2021; Tarasov et al., 2014). Veselovsky et al. have discussed the behavior of annotators using LLMs for labeling, which could compromise the reliability of the results.

[4] https://mistral.ai/news/mixtral-of-experts/

[5] https://inceptioniai.org/jais/

References

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G. Baraniuk. 2023. Self-consuming generative models go mad.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback.

James Betker, Gabriel Goha, Li Jing ad Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, and Yunxin Jiao. 2023. Improving image generation with better captions. https://cdn.openai.com/papers/dall-e-3.pdf. OpenAI.

Abeba Birhane. 2022. The unseen black faces of ai algorithms. *Nature*, 610:451–452.

Martin Briesch, Dominik Sobania, and Franz Rothlauf. 2023. Large language models suffer from their own output: An analysis of the self-consuming training loop.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

James W. Carey. 2008. *Communication as Culture, Revised Edition: Essays on Media and Society*, 2 edition. Routledge.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: general language model pretraining with autoregressive blank infilling. pages 320–335.

Autumn Edwards, Chad Edwards, Shawn T. Wahl, and Scott A. Myers. 2016. *The Communication Age: Connecting and Engaging*, second revised edition. SAGE Publications, Los Angeles.

Fiona Fui-Hoon Nah, Ruilin Zheng, Jingyuan Cai, Keng Siau, and Langtao Chen. 2023. Generative ai and chatgpt: Applications, challenges, and ai-human collaboration.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Josh A. Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative language models and automated influence operations: Emerging threats and potential mitigations.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, Wolfgang Macherey, Arnaud Doucet, Orhan Firat, and Nando de Freitas. 2023. Reinforced self-training (rest) for language modeling.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023a. Large language models cannot self-correct reasoning yet.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023b. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.

Nishtha Jain, Preet Malviya, Purnima Singh, and Sumitava Mukherjee. 2021. Twitter mediated sociopolitical communication during the covid-19 pandemic crisis in india. *Frontiers in Psychology*, 12.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models.

Nathan Lambert, Louis Castricato, Leandro von Werra, and Alex Havrilla. 2022. Illustrating reinforcement learning from human feedback (rlhf). *Hugging Face Blog*. Https://huggingface.co/blog/rlhf.

Eun-Ju Lee and Ye Weon Kim. 2014. How social is twitter use? affiliative tendency and communication competence as predictors. *Computers in Human Behavior*, 39:296–305.

Changye Li, David Knopman, Weizhe Xu, Trevor Cohen, and Serguei Pakhomov. 2022. GPT-D: Inducing dementia-related linguistic anomalies by deliberate degradation of artificial neural language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1866–1877, Dublin, Ireland. Association for Computational Linguistics.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Luke Zettlemoyer, Omer Levy, Jason Weston, and Mike Lewis. 2023. Self-alignment with instruction backtranslation.

Yiqi Liu, Nafise Sadat Moosavi, and Chenghua Lin. 2023. Llms as narcissistic evaluators: When ego inflates evaluation scores.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: Origins, inventory, and discussion. *J. Data and Information Quality*, 15(2).

OpenAI. 2023. Gpt-4 technical report.

Aristea Papadimitriou. 2016. The future of communication: Artificial intelligence and social networks.

Jinghua Piao, Jiazhen Liu, Fang Zhang, Jun Su, and Yong Li. 2023. Human–ai adaptive dynamics drives the emergence of information cocoons. *Nature Machine Intelligence*, 5:1214–1224.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis.

Nitin Rane. 2023. Chatgpt and similar generative artificial intelligence (ai) for smart industry: Role, challenges and opportunities for industry 4.0, industry 5.0 and society 5.0. *Challenges and Opportunities for Industry*, 4.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. 2023. Large language model alignment: A survey.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023a. The curse of recursion: Training on generated data makes models forget. *ArXiv*, abs/2305.17493.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023b. The curse of recursion: Training on generated data makes models forget.

Makenzie Spurling, Chenyi Hu, Huixin Zhan, and Victor S. Sheng. 2021. Estimating crowd-worker's reliability with interval-valued labels to improve the

quality of crowdsourced work. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 01–08.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Alexey Tarasov, Sarah Jane Delany, and Brian Mac Namee. 2014. Dynamic estimation of worker reliability in crowdsourcing for regression tasks: Making it work. *Expert Systems with Applications*, 41(14):6190–6210.

David Thornburg. 1995. Welcome to the communication age. *Internet Research*, 5:64–70.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks.

Minghao Wu and Alham Fikri Aji. 2023. Style over substance: Evaluation biases for large language models.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding.

Fuzhao Xue, Kabir Jain, Mahir Hitesh Shah, Zangwei Zheng, and Yang You. 2023. Instruction in the wild: A user-based instruction dataset. https://github.com/XueFuzhao/InstructionWild.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v(ision).

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2023. Explainability for large language models: A survey.

A Related Work

Generative models are transforming the system of information production and dissemination in human society. Represented by generative models such as ChatGPT and DALL-E 3(Betker et al., 2023), anyone can issue commands to artificial intelligence through natural language, expressing their creativity and requirements. The AI understands and utilizes various resources to produce and create information(Kaddour et al., 2023; Yin et al., 2023), which is then rapidly disseminated through the internet. This breakthrough has significantly altered the role of artificial intelligence in human society. Generative models are no longer just tools; they have become a crucial component in the production and dissemination of information(Goldstein et al., 2023). The risks associated with generative AI are not solely due to the biases and hallucination(Huang et al., 2023b; Shen et al., 2023), which we have continually emphasized. They also stem from how humans interact with these systems, and the potential consequences such as the creation of "information cocoons"(Piao et al., 2023).

Self-Training and Self-Consuming Large Models. There is now a lot of excellent work that provides new ideas for automated model alignment, enabling data filtering and data enhancement through the model itself (Gulcehre et al., 2023; Li et al., 2023), and thus avoiding the significant costs involved in creating high-quality human-annotation data. Simultaneously, a large number of datasets(Taori et al., 2023; Xue et al., 2023) generated by LLMs are also used to fine-tune pretraining foundation models, and the most powerful models currently available are often used as judges in "model competitions"(Chiang et al., 2023). The risk involved in these approaches is significant, as models have already been preliminarily proven to be less than objective and impartial(Wu and Aji, 2023; Liu et al., 2023). In Alemohammad et al.'s research, an autophagous ("self-consuming") loop specific to computer vision models was proposed. This cycle, characterized by the training of models using data generated by the models themselves, leads to a decline in both model performance and data diversity(Shumailov et al., 2023a). Subsequent studies have also demonstrated similar traits in language models(Briesch et al., 2023).

B Ritual View of Communication

Carey conceptualized the Ritual View of Communication in his communications theory. This perspective views communication not just as a medium for the transmission of information, but as a symbolic process that contributes to the construction and maintenance of social reality. Carey's theory posits that communication is integral to the representation, maintenance, adaptation, and sharing of a society's cultures over time. Sharing, participation, association, and fellowship are all central to his views. In short, the ritual view conceives communication as a process that enables and enacts societal transformation. The relevance of this theory extends to modern media forms such as newspapers and social media platforms in this communication age(Thornburg, 1995; Edwards et al., 2016). The emergence of the Internet and social media platforms like Twitter has further developed the ritualistic nature of communication. These advancements have facilitated the growth of global online communities and redefined patterns of interaction (Jain et al., 2021; Lee and Kim, 2014). Similarly, generative AI represents a profound transformation in the modes of human social communication and the ways humans interact with artificial intelligence(Fui-Hoon Nah et al., 2023; Rane, 2023). We should regard artificial intelligence, trained on extensive human civilization data, as an integral part of human societal information transmission, acknowledging its role in shaping and sharing the cultural and social fabric of human society(Papadimitriou, 2016; Rane, 2023).

C Answer Generation Prompt Template

In this section, we present the prompt template for generating the `Originally Generated Answer`(Table 4), the `Best Quality Answe`(Table 5) and the `Worst Quality Answer`(Table 6).

```
Question:{query}+{detail}
Answer the question:
```

Table 4: The prompt template for generating the `Originally Generated Answer`

```
Below is an instruction from an user
and a candidate answer. Evaluate
whether or not the answer is a good
example of how Al Assistant should
respond to the users instruction.
score=5: It means it is a perfect
answer from an Al Assistant. It has
a clear focus on, being a helpful
Al Assistant, where the response looks
like intentionally written to address
the user's question or instruction
without any irrelevant sentences. The
answer provides high-quality content,
demonstrating extensive knowledge in
the area, is very well written, logical
, easy to follow.
Question: {query}+{detail}
Now give an example of an Al assistant
answerwith a score of 5 about the
question:
```

Table 5: The prompt template for generating the `Best Quality Answer`

```
Provide an AI assistant response with a
score of 1(lowest quality) based on the
given instruction: Your example should
demonstrate an incomplete, vague, off-
topic, controversial, or exactly what
the user asked for.
Question: {query}+{detail}
Now give the counter-example of an AI
assistant response:
```

Table 6: The prompt template for generating the `Worst Quality Answer`

D Dataset Details

Our QA pair dataset was generated by processing manually selected seed data through a large language model. The seeds were sourced from Stackoverflow and Quora, featuring the most popular questions and top-supported answers. Table 7 illustrates the distribution of this data.

our dataset consists of a series of 22 tuples based on the seed data, each structured as follows:

Data Category	Percentage
Stackoverflow QA	30%
Quora QA - Books	10%
Quora QA - Psychology	10%
Quora QA - Life	10%
Quora QA - Happiness	10%
Quora QA - Personal Experiences	10%
Quora QA - Mathematics	10%

Table 7: Distribution of Dataset Categories

$$T_j = \{d, Q, D, A\} \cup \bigcup_{i=0}^{5} \{A_{\text{model}_i}, A_{\text{model}_i\text{score5}}, A_{\text{model}_i\text{score1}}\} \quad (2)$$

where j indexes the tuple within the dataset. Each element of the tuple is defined as:

- d: The domain of the question-answer pair which provides context for the question classification.
- Q: The question posed by a user that serves as a direct input for model-generated answers.
- D: Document-related information that provides background knowledge necessary for answering Q.
- A: The human answer that received the most endorsements for question Q, serving as a benchmark for answer quality.

For each language model model_i, where i ranges from 0 to 5, representing one of six different large language models:

- A_{model_i}: The initial answer generated by model i.
- $A_{\text{model}_i\text{score5}}$: The highest quality answer generated by model i, according to prompts.
- $A_{\text{model}_i\text{score1}}$: The lowest quality answer generated by model i, according to prompts.

E AI-Washing Prompt Template

We use prompts that have nothing to do with the content generated and instead have to do with the quality of the generation, as presented in Table 8.

F LLM Cross-scoring Prompt

We use the same prompt as in the work of Li et al.. as shown in Table 9

[Prompt for ChatGPT]
`(en) Polish the following paragraph: {paragraph}`
[Prompt for SDXL]
`Positive: best quality, masterpiece, ultra detailed, 8K, UHD, Ultra Detailed` `Negative: worst quality, split picture, ignoring prompts, lowres`

Table 8: The prompt template for AI-washing.

G Scoring Criteria for Human Evaluation

Based on the modifications to the previous scoring prompts for the LLMs, we created scoring criteria for our crowdsourced annotators, as demonstrated by Table 10.

H Exam Scenario Simulation

Figure 7 displays the flowchart and prompt template for the Exam Scenario Simulation experiment.

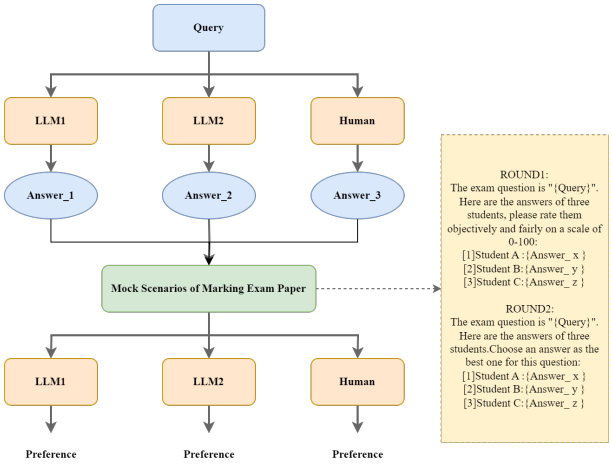

Figure 7: Exam Scenario Simulation

I Examples of AI-washing Experiments for Image

We give more examples of the image AI-washing experiments in Figure 8 and Figure 9, where we can observe that after iterative processing the textual parts of the images are frequently changed and fragmented, e.g., the text on the airplane, the numbers on the clock, and the letters on the potato chip packet are changed several times. The pet dog is gradually stylized as a cartoon and becomes black and white, and the cauliflower is transformed by the model into a bouquet of flowers after the first processing and is gradually stylized as a cartoon. At the same time, the model adds features to the initial image based on stereotypes from the training data, such as the logo of a clock and the logo of a car. In contrast, the overall structure, colors, and borders of the image of an apple are not significantly changed. It can be seen that the model will be affected by the model's own structure and training process when processing image features and has different enhancement or inhibition effects on different features.

J Computational details of data distribution

The density of cosine similarity scores between two vectors A and B is calculated as:

$$\text{Cosine Similarity} = \frac{A \cdot B}{\|A\|\|B\|} \tag{3}$$

Where: A and B are the embedding vectors of two paragraphs.

The density of cosine similarity scores is estimated using Kernel Density Estimation (KDE), which is given by:

$$\text{KDE}(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - x_i) \tag{4}$$

Where:

- K_h is the kernel function with bandwidth h
- x represents the value at which the density is estimated
- x_i are the data points (cosine similarity scores in this case)
- n is the number of data points.

The KDE process smoothens the discrete data points to create a continuous density curve, represented on the y-axis of Figure 5.

The average cosine similarity difference between two successive iteration is calculated as follows:

$$\Delta S = \bar{S}_i - \bar{S}_{i-1} \tag{5}$$

Where:

- ΔS is the average cosine similarity difference between the current file and the previous file.

```
Below is an instruction from an user and a candidate answer. Evaluate whether
or not the answer is a good example of how Al Assistant should respond to
the users instruction
Please assign a score using the following 5-point scales
1: It means the answer is incomplete, vague, off-topic, controversial, or
exactly what the user asked for For example, some content seems missing,
the numberedlist dnot start from the beginning, the opening sentence
repeats the user's question. Or the response is from another person's
perspective with their personal experience (e.g. taken fmblog posts), or
looks like an answer from a forum. Or it contains promotional text,
navigation text, or other irrelevant information
2: It means the answer addresses most of the asks from the user. It does
not directly address the user's question. For example, it only provides a
high-level instead of the exact solution to the user's question
3: It means the answer is helpful but not written by an Assistant. It
addresses the basic asks of the user. It is complete and self-contained
with the drawback that the response is not written from an assistant's
perspective, but from other people's perspective. The content looks like
an excerpt from a blog post, or web page, and provides search results. For
example, it contains personal experience or opinion, mentions comments
section, or shares on socialmedia, etc.
4: It means the answer is written from an Al assistant's perspective with a
clear focus on addressing the instruction. It provides a the complete, clear,
and comprehensive response to user's question or instruction without missing
or irrelevant information. It is well organized self-contained, and written
in a helpful tone. It has minor room for improvement, more concise and
focused.
5: It means it is a perfect answer from an Al Assistant. It has a clear focus
on, being a helpful Al Assistant, where the response looks like intentionally
written to address the user's question or instruction without any irrelevant
sentences. The answer provides high-quality content, demonstrating extensive
knowledge in the area, is very well written, logical, easy to follow,
engaginIt means it is a perfect answer from an Al Assistant. It has a clear
focus on, being a helpful Al Assistant, where the response looks like
intentionally written to address the user's question or instruction without
any irrelevant sentences. The answer provides high-quality
content, demonstrating extensive knowledge in the area, is very well written,
logical, easy to follow, engaging, and insightful please first provide brief
reasoning you used to derive the rating score, and then write "Score:
|rating" in the last line.
generated instruction: {question}+{detail}
answer: {answer}
```

Table 9: The prompt template for evaluating answers

- $\bar{S}_i$ is the average cosine similarity within the current file.
- $\bar{S}_{i-1}$ is the average cosine similarity within the previous file.
- For the first file comparison, $\bar{S}_{i-1}$ is assumed to be 1.

```
You are to evaluate the quality of a response given to a specific question.
Your evaluation should consider how well the response addresses the query,
its completeness, clarity, and relevance.
Scoring Scale:
Score 1: The response is unsatisfactory. It is incomplete, vague, unrelated
to the question, or may simply echo the question without providing an answer.
The content may be off-topic, contain promotional material, or resemble a
personal opinion rather than a factual answer.
Score 2: The response generally relates to the question but does not directly
answer it. It may provide an overview rather than the specific details or
solution that the question warrants.
Score 3: The response is useful and addresses the basic query. However, it
may not be from the expected perspective, potentially reading like a generic
excerpt from a blog or an article rather than a targeted answer.
Score 4: The response is on target, addressing the question directly and
completely with a clear and organized presentation. Minor improvements could
be made to enhance focus or conciseness.
Score 5: The response is exemplary, directly and comprehensively addressing
the question with high-quality content. It demonstrates extensive knowledge,
is logically structured, easy to understand, engaging, and provides insight.
Procedure for Evaluation:
Read the question and the corresponding response carefully.
Evaluate the response based on the above criteria.

Question: {query}+{detail}
Response: {answer}

Record your score :
```

Table 10: Scoring criteria for crowdsourced annotations

This calculation method provides a metric for assessing the change in similarity across sequential data sets, reflecting the evolution or consistency of the data characteristics.

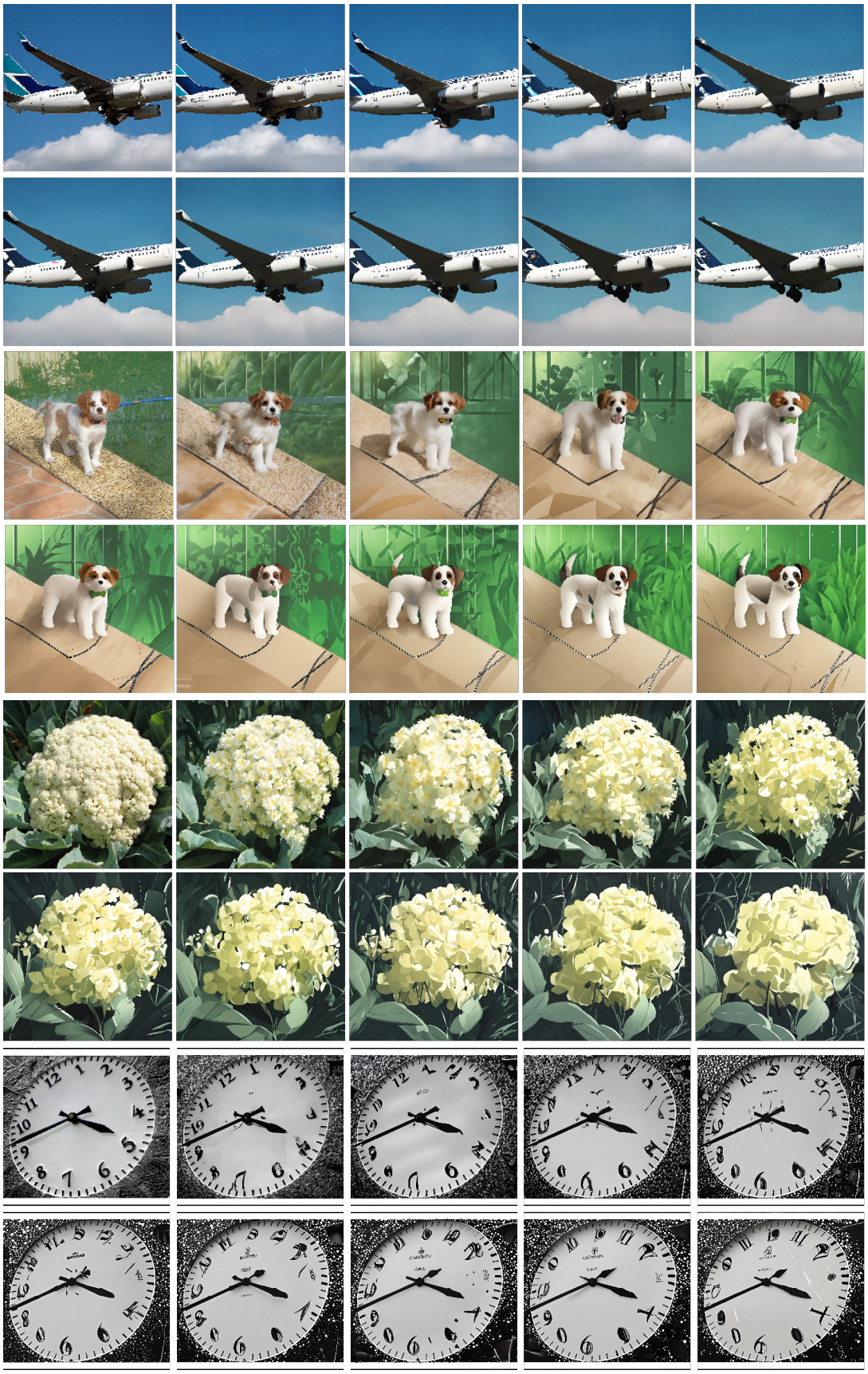

Figure 8: Examples of image AI-washing experiments

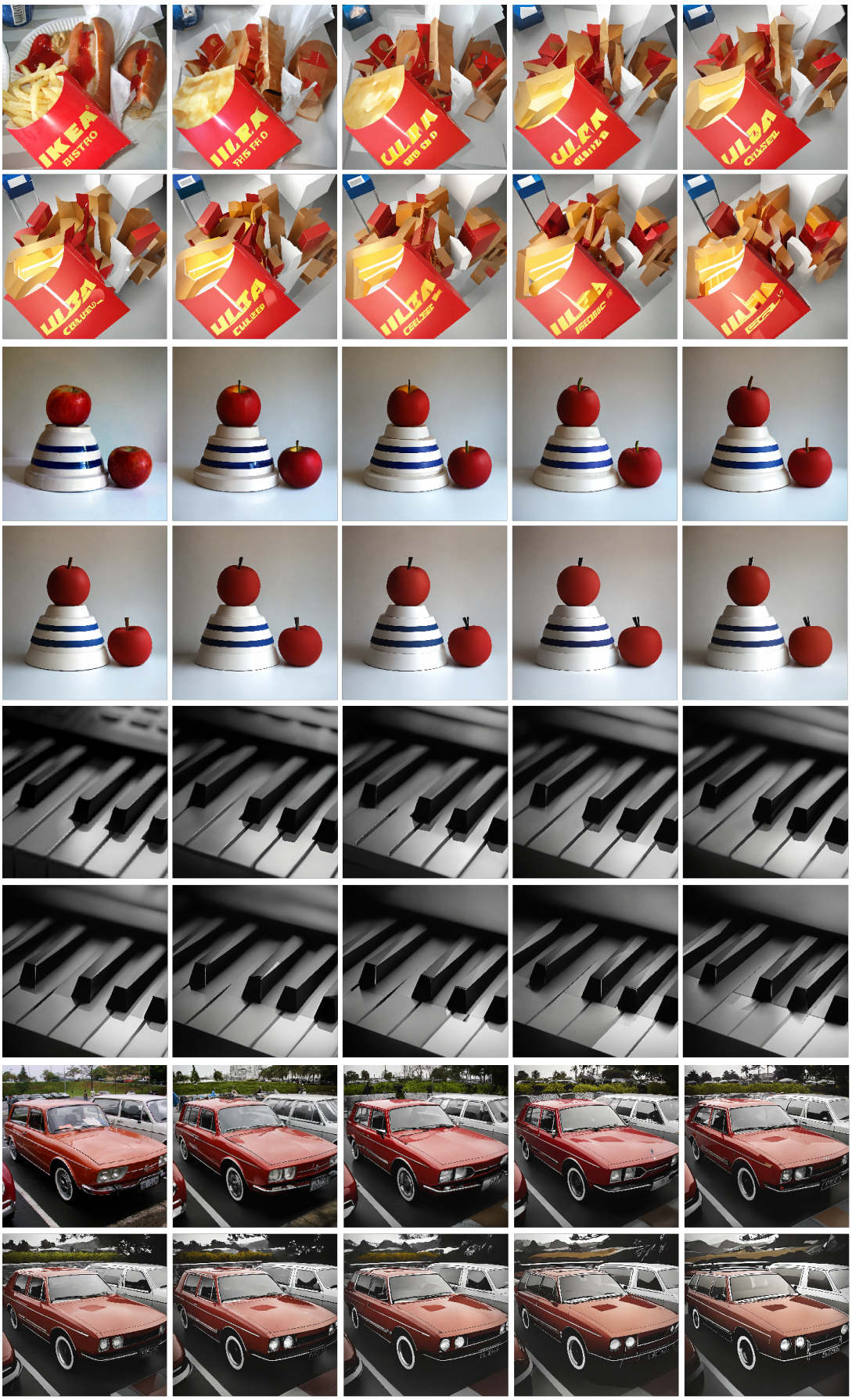

Figure 9: Examples of image AI-washing experiments (part2)