

Who is the Creator: Subjectivity Game in AIGC Narrative Generation and Adversarial Prompting

Yi Lu

Xianda College of Economics and Humanities, Shanghai International Studies University, Shanghai, China

Abstract: The narrative generation of generative artificial intelligence is based on the probability prediction of large-scale data training, which naturally tends to output the narrative mode of "statistical averaging", resulting in content homogenization. In this paper, the concept of "adversarial prompt" is proposed, which is different from the dialogue mechanism of conventional prompt engineering. Through the alternating use of three types of prompts: value presupposition, reflective questioning and structural constraints, AI is forced to deviate from the generation inertia and shift to the preset critical direction. The study takes Lu Xun's *Madman's Diary* as the ideological engine, refines its core critical elements and encodes them into AIGC generation parameters that can be guided, and analyzes the creation of AI narrative short film *The Memory Inspector* as a case. According to the research, the essence of adversarial cues is not to tame artificial intelligence (AI), but to expose the creator's own thinking inertia. The creation of AIGC era is no longer just the cooperation between people and tools, but also the game between people and themselves.

Keywords: AIGC Narrative Generation; Adversarial Prompts; Critical Thinking; *Madman's Diary*; Human-machine Collaboration.

1. Introduction

Since 2022, generative AI has enabled coherent text, stylized images, and even complete short films. Yet the technology succeeds at a cost. Millions of users work with identical models and similar prompts; the outputs converge. A narrative template now dictates how stories open, how characters evolve, and how conflicts end. Formulaic plotting, flat emotion, and shallow ideas have become hallmarks of AIGC content.

There has long been insight in the field of philosophy of technology. Benjamin pointed out in "The Work of Art in the Era of Mechanical Reproduction" that technological reproduction dispels the "aura" of art [1]. Heidegger defines the essence of modern technology as "Gestell", a "forced" way of unmasking [2]. In the context of AIGC, narrative is regarded as a replaceable and optimized data processing process. The most precious part of narrative is precisely the ideological sharpness that does not obey the norm. This sharpness is only an "error" in the eyes of the algorithm and needs to be eliminated. Our real concern is not that machines replace humans, but that humans gradually accept the "compliance answers" given by machines and gradually lose their ability to identify mediocrity and independent judgment.

In the face of this dilemma, this study attempts to propose a response. Taking Lu Xun's *Madman's Diary* as the ideological engine [3], this paper proposes a game mechanism of "adversarial prompts" at the level of creative methodology, and explores how critical thinking is deeply embedded in the whole process of AIGC narrative generation.

2. Literature Review: Three Dimensions of AIGC Narrative Research

2.1. The Technical Path of AIGC Narrative Generation

Early studies focused on narrative generation based on rules and templates, including prompt engineering, controllable generation and style transfer. These studies use AI as a development tool, focusing on technical indicators such as coherence and diversity, but lack concern for the depth of thought in generated content, which is rarely discussed. These studies default that "good narrative" is the narrative that conforms to the characteristics of most samples. This creates a contradiction: the more advanced the technology, the more convergent the generated narrative.

2.2. The "Analytic" Tradition of Digital Humanities

In the field of digital humanities, text mining, social network analysis and other methods are used to study classical literature. "Distance reading" proposed by Moretti has become one of the important methods in the field of digital humanities [4]. These methods are mainly used to analyze text, not to create text. Researchers can extract image data from *Madman's Diary*, but have not systematically explored how to turn these data into ideological resources that drive AI creation.

2.3. The Educational Application and Ethical Reflection of AIGC

This kind of research either focuses on tool use skills, or risk reminders, and has not yet proposed a systematic solution on how to embed critical thinking into AIGC's creative process.

The three areas are advancing separately, and there is a lack

of connection between them. Technical researchers pursue more coherent and diverse generation effects, but do not ask whether the generated content carries ideas worth telling. Digital humanities scholars are good at using algorithms to interpret classics, but rarely transform the analysis results into new creative resources. Educational researchers and ethicists have seen the risks of AIGC, but most of them stay at the reminder level and fail to give operational solutions. The fracture of the above three angles constitutes the starting point of this study.

3. Extraction and Coding of Critical Thinking Elements in *Madman's Diary*

To transform the critical thinking of classical texts into a strategy that can guide AIGC generation, the first step is to systematically extract the ideological elements of the text. *Madman's Diary* is the pioneering work of modern Chinese literature. This work contains the cognitive subversion of science fiction literature under the realistic framework, and its "virtual reality" breaks the deterministic logic of traditional realism [5]. *Madman's Diary* is very experimental in narrative form and ideological structure. This kind of experiment, which is in contrast to the generation logic of AIGC's pursuit of standardization, provides a transformable critical thinking resource for this article.

3.1. Core Critical Themes

The critical thought of *Madman's Diary* presents four progressive topics at the narrative level, which together constitute a cognitive path from external observation to internal reflection.

The triple progression of "cannibalism". "Cannibalism" is not a static metaphor, but has experienced multiple progressions in the narrative advancement. From the initial fear of being eaten by the madman to his reading of the word "cannibalism" in the cracks of history, he sublimates his fear into questioning the civilized order, and finally confesses "I have eaten people" and finds himself both a victim and a conspirator. The cruelty of the logic of "cannibalism" in Lu Xun's works is that everyone lives in an oppressive system. People are under oppression, but also inadvertently continue oppression. As far as AIGC narrative creation is concerned, if a story wants to have critical power, it cannot simplify the conflict into personal confrontation. Characters become part of the system unconsciously, which is far more disturbing than being directly hit by the system.

The dilemma of the awakened. The madman finds no heroic victory after discovering the truth. The despair of "nowhere to find a true human" and the helpless cry to "save the children" together outline a profound tragedy: The clearer one sees, the more isolated one becomes. Epistemic awakening does not bring liberation; the madman loses his connection to the world around him. This suggests that critical narratives should not offer cheap hope or facile redemption — awakening comes with a heavy cost.

The reversal of normal and mad. The "normal people" in the madman's eyes are incomprehensible — they live unperturbed within the order of "cannibalism". In the eyes of the "normal", it is the madman who is insane for claiming that everyone eats people. Two perspectives confront each other, leaving the reader with no basis for judgment. Lu Xun provides no answer; he leaves the suspension to the reader.

This indeterminacy is precisely the source of the work's intellectual power. AI-generated narratives should likewise avoid offering definitive conclusions from an omniscient vantage point.

"I too have eaten people" — turning critique upon the self. The madman's most startling confession is not that they eat people, but that he himself has eaten people. He "could not have been entirely unaware" that he might have eaten some of his sister's flesh. He is himself part of that "four-thousand-year history of cannibalism". This self-examination breaks the binary of accuser versus accused. Critique no longer points only at others, but also at the critic. In AIGC narrative, a story may establish a turning point in which the protagonist discovers his own complicity, rather than remaining merely an observer or a rebel.

3.2. Encoding Narrative Structural Elements

The restricted first-person perspective in *Madman's Diary* creates a cognitive oscillation that can serve as a source of narrative tension. The unreliable narrator — the madman whose words are never confirmed as "fact" — leaves readers perpetually uncertain whether "cannibalism" is truth or delusion [6]. Lu Xun tells this story in the tone of a "madman," which means that the truth in the narrator's mouth must be interpreted as madness at the same time. This uncertainty not only does not weaken the criticism, but also makes the reader unable to easily determine the credibility of the narrative.

In addition, the rhythm of the novel is gradually collapsed. From the strange feeling of "why Zhao's dog sees my eyes", to the historical excavation of "Mr. Gu Jiu's old flowing water book", to the self-judgment of "I have eaten people", each diary has more deeply disintegrated the madman's cognition than the previous one. In the AIGC narrative, the release of information should not be completed at one time. Each level of new discovery should allow the reader and the protagonist to experience the fall of cognition together, rather than sliding smoothly to the end.

3.3. Translating Conceptual Elements into Actionable Parameters

The above ideological elements need to be further converted from humanistic analysis to executable creative parameters. This study further encodes it into three types of structured narrative parameters. Thematic parameters include power criticism, cognitive subversion, individual awakening and system collusion. Emotional parameters cover anxiety, fear, loneliness and despair. The formal parameters include fragmented narrative, unreliable narrator, the repeated appearance of core images and the construction of symbolic system. The role of these three types of parameters is to disassemble abstract judgments in literary criticism, such as "critical" and "unreliable narration", into instructions that AI can directly execute, so that AIGC generates content to carry ideas.

4. Conceptualizing and Constructing "Adversarial Prompting"

4.1. Defining the Core Concept

Based on the extraction of the ideological elements of *Madman's Diary*, this study proposes the concept of "adversarial prompt". It is different from the "adversarial attack" (misleading the model to output wrong results) or

"adversarial training" in the field of computer science, but a game mechanism at the level of creative methodology.

The essence of routine prompt engineering is cooperation. People describe needs, AI performs, both sides have the same goal, and pursue maximum efficiency. The essence of adversarial cues is a game of thought. Humans implant a position that is contrary to the statistical inertia of AI at the starting point of generation. Through continuous questioning and correction, AI is forced to deviate from its "safe zone" and generate in a more critical direction. Conventional prompts describe the expected output to AI in the form of instructions, and its effectiveness depends on the accuracy and integrity of the instructions. Different from conventional prompts, adversarial prompts require examining and questioning the output of AI after each round of generation, and pushing the generated content to move in a more thoughtful direction. The former regards AI as an execution tool, and the latter regards the generation process as a dynamic process that requires continuous judgment and intervention.

4.2. Designing Three Types of Adversarial Prompts

The three types of adversarial cues are used alternately in practice, with different emphases, but sharing the same goal: to let AI deviate from its statistical inertia in continuous resistance.

Value-presupposed prompting embeds a position at the point of generation. It does not offer AI a neutral creative space but begins with a critical perspective as an uncompromised premise. In the creation of *The Memory Inspector* (see Figure 1), for instance, which was selected for the International Digital Media and AI Competition at the 28th Tel Aviv International Student Film Festival, the prompt was designed as: "This quality inspection system appears fair but actually maintains power order by filtering memories. The narrator gradually discovers that he himself has been complicit in this system." This is not a suggestion but an enforceable framework. AI cannot choose a compromising response; it must proceed from the premise that the system is unjust.



Fig 1. The Memory Inspector poster

Reflexive interrogation intervenes during the generation process, questioning the AI's intermediate outputs. When the

AI generated an ending where the protagonist successfully escapes the system, the prompt asked: "the awakened often cannot truly escape — what if there is no external space to escape to?" This breaks the AI's "problem-solving" pattern.

Structural-constraint prompting translates formal features from *Madman's Diary* into executable directives. For example: "use a first-person unreliable narrator perspective. Use 'qualified' and 'unqualified' as recurring core imagery." The value of such prompts lies in their intervention not only in what AI says but in how it says it. Form is never merely ornamental; in *Madman's Diary*, the unreliable first-person perspective is itself a source of intellectual acuity. Without formal constraints, AI defaults to its most familiar narrative templates.

The alternating use of the three prompt types forms a dynamic correction loop. Value anchoring sets direction; reflexive interrogation corrects deviations; structural constraint prevents formal conventionality. Each correction may introduce new deviations, and new deviations require further correction.

5. Mechanism Analysis of "Dialogue Game"

The implementation of adversarial prompts follows an iterative cycle from generation, review, inquiry to re-generation, rather than a nonlinear process. Taking the creation of the AI narrative short film *The Memory Inspector* as an example, the short film sets a future memory audit system. The audience, as a quality inspector, examines the core memory of a bionic man called "Echo." The Echo is recalled by the system because of the unwanted emotions. Every time the audience sees a memory, it is necessary to determine whether it is compliant or non-compliant. The following is a detailed observation of the operation of this mechanism.

The initial prompt "to create a science fiction short film about a future memory management system" produced a standard heroic story. The protagonist discovers the truth, confronts the system and wins. The first round of adversarial hints changed the premise. The audience is no longer a rebel. They are the executors of the system. This is directly embedded in the structure of 'I am also a conspirator'.

A second round of prompts raised a new problem. How should Echo respond to the viewer's choices? If the viewer's actions merely move the process forward, the interaction becomes mechanical. The prompt required Echo to give emotionally inflected responses. Choosing "qualified" prompts Echo's thanks; "unqualified" elicits confusion; "abstain" invites questioning. This interactivity does not change the narrative direction; it allows viewers, in the act of choosing, to experience how they are being guided by a preset system.

The subsequent introduction of structural constraints required a first-person perspective throughout. Everything viewers see comes from Echo's subjective replay. This translated the "unreliable narration" mechanism from *Madman's Diary*: Viewers can never determine whether what they see is truth or a paranoid self-justification.

In subsequent iterations, the three prompt types were used alternately. The creators never accepted AI's final answer. In the finalized version of *The Memory Inspector*, no matter what choices viewers make, they ultimately find themselves trapped within the system. Echo's every line reminds them:

"You judge me, but are you not also being judged?"

6. Theoretical Contributions and Discussion

The proposal of "adversarial prompting" originates from a fundamental question: When AI can generate almost any plausible story, where does human irreplaceability lie? This is no longer a theoretical assumption but an urgent question demanding answers. Existing scholarship offers varied responses. Researchers in the technical path tend to position AI as a continuously optimizing tool whose value lies in enhancing creative efficiency. Researchers in the ethical reflection path focus more on AI's potential erosion of human creativity. Both perspectives have their grounds, yet both assume — explicitly or not — that the human-AI relationship is structured as one of dominator and dominated.

This study offers an alternative understanding. In the iterative process of creating *The Memory Inspector*, AI was neither a passive execution tool nor a threat to be guarded against. Every time it produced an overly smooth resolution or an overly "standard" answer, it prompted the creator to reexamine the choices made. This mirrors the structure in which Echo, the judged humanoid, turns the question back on the inspector — a structure that, as Yang and Wang argue, exemplifies the mirror construction of the humanoid other in science fiction: Never about the other itself but about an allegorical reflection on the subject's position, a game of subject paradigms rather than a binary opposition [7].

The core function of adversarial cues is not to optimize model output, but to force creators to examine their own thinking inertia through continuous questioning. Through the continuous questioning of AI output, creators continue to encounter their own thinking inertia and aesthetic inertia. The reason why AI's equalization tendency is worthy of vigilance is not because it makes mistakes, but because it always makes mistakes "correctly". The value of adversarial cues is precisely to make this "right" suspicious.

The theoretical contribution of this study is not to invent new prompting techniques, but to reveal a more fundamental fact that the creation of the AIGC era is essentially a game between people and themselves. The mean output of AI makes the creator's unconscious preference explicit, thus providing an object for critical correction. For creators, adversarial hints mean a new creative ethics. When AI generates a "qualified" story, the value of the creator is no longer reflected in the speed of output, but in the ability to find that the answers given by AI are mediocre and insist on continuous revision. For education, this methodology suggests a possible teaching direction. Previous studies have pointed out that generative artificial intelligence shows unique advantages in critical thinking assessment, and can expose learners' thinking blind spots through multiple rounds of dialogue [8]. However, this study further believes that the core of teaching is not to teach students "how to use AI to make films", but to allow them to use adversarial cues as a thinking training mechanism in the alternation of close reading of classic texts and AI creation, and to reflect on their own cognitive inertia in the process of continuously questioning AI, so as to complete a deep training of critical thinking.

The limitation of this study is that it uses *Madman's Diary*

as a single source of thought, and the applicability of the adversarial prompt framework in other classic texts needs to be further verified. In addition, the design of adversarial prompts is currently highly dependent on the creator's humanistic literacy and critical thinking ability. The effectiveness of this strategy depends on the creator's existing aesthetic judgment and the accumulation of classical text reading.

7. Conclusion

This paper takes Lu Xun's *Madman's Diary* as the ideological engine, puts forward the "adversarial prompt" strategy, and explores the theoretical path of critical thinking embedded in the whole process of AIGC narrative generation. The study confirms that "adversarial prompt" is a multi-turn conversational instruction. It has the characteristics of value presupposition, reflective inquiry and structural constraints, which can break the "statistical mean" tendency of AIGC based on training data, and make the generated content carry the critical spirit of classical text.

The translation from "cannibalism" in *Madman's Diary* to "memory inspection" in *The Memory Inspector*, and from the madman's "I too have eaten people" to the inspector's "I too am complicit", demonstrates a regeneration of conceptual core across media rather than a mere thematic updating. It proves that the critical thought of canonical texts, after structured extraction, can be effectively converted into methodological strategies for driving AI creation.

This study ultimately responds to a fundamental question: Where does human creativity go in the age of intelligence? The answer is not to reject AI but to steer it with more profound thought. Adversarial prompting is not merely a technical strategy but a stance on human-machine relations. In the age of automation, human irreplaceability lies in the capacity for criticism, reflection, and uncompromising thought.

References

- [1] Benjamin, W. (2008). The work of art in the age of mechanical reproduction. In H. Arendt (Ed.), *Illuminations: Essays and Reflections* (Z. Zhang & B. Wang, Trans.; pp. 231–264). SDX Joint Publishing Company.
- [2] Heidegger, M. (2005). The question concerning technology. In *Basic Writings* (Z. Sun, Trans.; pp. 3–35). SDX Joint Publishing Company.
- [3] Lu, X. (2005). *The Complete Works of Lu Xun* (Vol. 1). People's Literature Publishing House, pp. 444–455.
- [4] Du, L. L. (2020). On Moretti's distant reading and its influence. *Comparative Literature in China*, (3), 179–193.
- [5] Song, M. W. (2020). Is *Madman's Diary* science fiction? — On Lu Xun's relation to science fiction, and reflections on the falsity of realism and the reality of the virtual. *Comparative Literature in China*, (2), 18–35.
- [6] Shen, D. (2006). What is "unreliable narration"? *Foreign Literature Review*, (4), 133–143.
- [7] Yang, J. L., & Wang, J. W. (2021). Subject paradigm game and posthuman allegory: The mirror construction of humanoid others in science fiction narrative. *New Films*, (2), 19–25.
- [8] Leng, J., Lu, H. H., & Dai, L. (2024). Generative AI empowering critical thinking assessment: An application experiment based on ChatGPT. *Modern Distance Education Research*, 36(6), 102–111.