# Survey on Hallucination in Reasoning Large Language Model: Evaluation, Taxonomy, Intervention, and Open Issues

Xinyi Liu[1], Yuting Lu[1], Shunping Wei[1,2]†

[1] *School of Education, Minzu University of China, Beijing 100081, China*

[2] *Engineering Research Center of Integration and Application of Digital Learning Technology, Ministry of Education, Beijing 100039, China*

**Abstract**

In recent years, reasoning large language models (LLMs) have seen increasingly widespread adoption in the field of education, particularly demonstrating substantial potential in tasks involving complex text comprehension. However, these LLMs are susceptible to a critical yet often overlooked issue: hallucinations within the reasoning process—instances where the model outputs a correct final answer while its underlying reasoning chain contains fabricated, inconsistent, or logically flawed content. Such hallucination phenomena in Chain-of-Thought (CoT) processes pose serious challenges to the reliability of educational applications. To address this issue, this study proposes a systematic research framework comprising dataset construction, multi-model CoT evaluation, and hallucination classification and quantification. Utilizing the whole-book reading dataset aligned with the junior secondary Chinese language curriculum, we conduct a comparative evaluation of six leading domestic and international LLMs, including ChatGPT o1 and DeepSeek-R1. Key findings include:(1) Hallucinations in CoT are prevalent across all tested models, with ChatGPT-o1 exhibiting a distinctive high accuracy–high hallucination pattern;(2) Hallucinations are both task- and genre-dependent: narrative texts, particularly novels, tend to trigger higher hallucination indices due to long-range dependencies and implicit cultural references. Tasks involving logical reasoning, linguistic feature analysis, and detail extraction show the highest hallucination rates, revealing model weaknesses in handling long-tail knowledge;(3) Hallucinations typically follow a progressive generative pattern: Information misreading → Comprehension deviation → Content fabrication → Logical instability. To mitigate these issues, we propose two targeted intervention strategies: uncertainty-based abstention and model-to-model correction. These approaches offer practical pathways toward enhancing the trustworthiness and educational applicability of reasoning LLMs.

*Keywords:* Reasoning large language models; Chain-of-thought; Hallucination; The whole-book reading; Dataset

## 1. Introduction

In September 2024, OpenAI introduced a revolutionary advancement in artificial intelligence with the release of its o1 reasoning system, marking a fundamental transition from perceptual

†Corresponding author: Shunping Wei (Email: weishunping@muc.edu.cn; ORCID: 0009-0005-0892-0033)

modeling to genuine reasoning capabilities. Benchmark evaluations demonstrated that this system consistently outperformed domain experts in doctoral-level knowledge assessments across multiple disciplines[1]. In January 2025, DeepSeek, a fundamental technology research company for artificial intelligence in Hangzhou, China, released the DeepSeek-R1 reasoning LLM. Its performance was comparable to that of OpenAI o1, the strongest reasoning LLM publicly released at that time, attracting high attention from the global technology circle[2]. Since then, Reasoning LLMs have come into the public eye, and more and more AI companies have begun intensive updates and released reasoning LLMS in their respective fields. Unlike the black box problem of the unknowable reasoning decision Chain of traditional general large models, the reasoning large model with chain-of-thought technology as the core breaks down the thinking chain through the transparent reasoning process of step-by-step thinking, effectively resolving the problem of model unknowability caused by ultra-large-scale unsupervised deep learning training[3].

Although reasoning large language models have significantly enhanced the interpretability of decisions through the transparency of CoT, the risk of hallucination latent in their reasoning process has taken on a new complexity. The problem of hallucination is widespread in large language models and has become one of the greatest challenges faced by natural language generation[4]. As Ziwei Ji mentioned, the hallucination refers to the situation where the content generated by large models is meaningless or not faithful to the source content provided[5]. That is, large models will generate answers that seem reasonable but actually deviate from the user's intention, from the previously generated context, or from factual knowledge[6]. Current research suggests that big language models are commonly illusory in generative output[7][8], while reasoning large language models, as an emerging technological paradigm, not only provide the final answer, but also present the complete reasoning process at the same time. From the perspective of cognitive computing, these explicit reasoning processes essentially constitute an important part of the model output, and thus inevitably suffer from the hallucinatory effect. It is worth noting that while large models generate correct answers, the reasoning chain behind them may contain serious illusionary flaws - this kind of process illusion is often more concealed and misleading than the illusion at the result level.

Based on this, this study selects the existing public the whole-book reading dataset to conduct empirical research, attempting to answer the following three core questions:

(1) Does the phenomenon of hallucination exist in chain-of-thought reasoning, and what paradigm characteristics does it present?

(2) What are the subject-specific types of illusions that the reasoning large language models present in the whole-book reading comprehension task?

(3) How to design educational and appropriate intervention measures to detect and prevent process hallucinations?

## 2. Related Work

Chain-of-Thought refers to the process in which large language models explicitly generate a continuous, readable sequence of intermediate reasoning steps before producing the final answer, simulating the human cognitive process of step-by-step thinking. Empirical studies have demonstrated that CoT prompting can significantly enhance performance across a range of arithmetic, commonsense, and symbolic reasoning tasks, with observed gains often being substantial [9]. Wei et al. first validated the effectiveness of this mechanism in arithmetic and symbolic reasoning tasks using zero-shot prompting with the instruction "Let's think step by step" [3].

Subsequently, Zhou et al. proposed the Least-to-Most (LtM) prompting strategy, which recursively decomposes complex problems into subproblems to reduce reasoning depth [10]. Further optimizing this approach, Chen et al. introduced Minimum Reasoning Path Prompting (MRPP) to refine chain length and minimize irrelevant noise [11]. While CoT improves surface-level plausibility and accuracy, research has also identified its potential to introduce new hallucination risks—specifically, Chain-of-Thought Hallucinations. Despite improving performance, CoT introduces distinct hallucination risks including factual errors where models insert unverifiable information [12], logical fallacies where invalid reasoning persists with high confidence [13], and confidence masking where flattened token-entropy distributions obscure differences between correct and incorrect answers [12]. Although CoT enhances the reasoning capabilities of LLMs, it may also predispose models to erroneous reasoning paths, thereby amplifying hallucination tendencies.

Current research on hallucination classification in large language models has evolved into a comprehensive multidimensional framework. The foundational dichotomy distinguishes between Factual Hallucinations, where generated content contradicts established world knowledge, and Faithfulness Hallucinations, where model outputs diverge from input instructions or contextual requirements [8]. Building upon this basic classification, researchers have developed a more nuanced system that traces hallucinations to three core cognitive deficiencies: Commonsense Memorization Deficiencies leading to errors in entity relationships and conceptual understanding, Relational Reasoning Deficiencies manifesting as limitations in logical inference, and Instruction Following Deficiencies arising from conflicts between pretraining and fine-tuning objectives [14]. Recent work from Tsinghua University has further identified two specialized phenomena — Flaw Repetition characterized by the persistent recurrence of identical erroneous logic during reasoning processes, and Think-Answer Mismatch exhibiting semantic disconnections between intermediate reasoning steps and final conclusions [15].

Domain-specific studies have yielded specialized taxonomies tailored to different applications. In computer vision research, object hallucinations are systematically categorized into three subtypes: Category Hallucination, Attribute Hallucination, and Relation Hallucination, corresponding to errors at the levels of object recognition, attribute judgment, and relation understanding, respectively [16]. The medical domain has established more specialized classification criteria, encompassing confusion errors arising from knowledge gaps, confabulated statements generated based on flawed logic, and knowledge contamination induced by training data biases. These classification methods have undergone systematic validation across representative clinical scenarios, including clinical trials, medical knowledge bases, licensing examinations, and clinical conversations [17].These diverse yet complementary classification systems collectively contribute to a more systematic and granular understanding of hallucination phenomena across different LLM applications and domains, enabling more targeted analysis and mitigation approaches.

At present, the alleviation and inhibition strategies of hallucinations are mainly optimized from the data level [18][19], model training level [20][21] and reasoning level [22][23].From the perspective of the CoT, Dhuliawala et al.'s Chain-of-Verification (CoVE) method employs multi-round self-verification to substantially reduce hallucinations [24], while Microsoft's Chain of Natural Language Inference (CoNLI) enables detection and correction without external knowledge sources through natural language inference chains [25]. The SLED framework, developed through Duke University and Google collaboration, innovatively guides final-layer outputs using early-layer logit information to enhance factual accuracy [26].Domain-specific solutions demonstrate increasing specialization, exemplified by medical applications adopting structured reason-

ing pathways. The Chain-of-Medical-Thought (CoMT) approach decomposes medical reports into six-layer Question-Answering chains mirroring clinical reasoning processes [27]. Similarly, Yin et al. proposed a training-free post-hoc method named Woodpecker, which detects and corrects visual hallucinations in text generated by multimodal large language models through five stages: key concept extraction, question formulation, visual knowledge validation, visual claim generation, and hallucination correction [28]. Table 1 systematically compares the relevance of these approaches to our work.

Table 1: Relationship between related studies and this paper

| Research Dimension | Focus area | Key Contributions | Literature | Relationship to This Study |
|---|---|---|---|---|
| CoT Hallucination | Technical Evolution | Zero-shot prompting | Wei et al. (2022) [9] | Provides methodological reference for optimizing reasoning paths |
| | | Recursive problem decomposition | Zhou et al. (2022) [11] | |
| | | Reasoning path optimization | Chen et al. (2024) [12] | |
| | Hallucination Risks | Factual errors in chains, logical fallacies, confidence masking | Cheng et al. (2025) [13]<br>Turpin et al. (2023) [14] | Identifies key problem targets |
| Hallucination Taxonomy | Foundational Framework | Factual and Faithfulness hallucinations | Huang et al.(2025) [8] | Supports unified definition framework across domains |
| | Cognitive Root Causes | Commonsense memory deficits/Relational reasoning deficits/Instruction-following deficits | Du et al.(2023) [15] | Provides cognitive perspective on model capability gaps |
| | | Flaw repetition/Think-answer mismatch | Yao et al. (2023) [16] | Offers verifiable behavioral metrics for identifying "pseudo-correct" outputs |
| | Domain Extensions | CV:Category/Attribute/Relation hallucinations | Bai et al. (2024) [17] | Enhances adaptability for specialized tasks |
| | | Medical:Confusion/Confabulation/Contamination | Garcia et al. (2025) [18] | |
| Hallucination Mitigation | Foundational Optimization | Data cleaning, model training, inference intervention | Touvron et al. (2023) [19]<br>Zhou et al. (2023) [21]<br>Chuang et al. (2023) [23] | Helps understand prior environmental factors of LLM hallucinations |
| | CoT-Specific Methods | Self-verification chains (CoVE) | Dhuliawala et al. (2023) [25] | Supports proposed AI validator mechanism for human-machine collaborative verification |
| | | Knowledge-free correction (CoNLI) | Lei et al. (2023) [26] | |
| | | Layer-guided decoding | Zhang et al. (2024) [27] | |
| | Domain-Specific Solutions | Medical structured CoT | Jiang et al. (2025) [28] | Inspires construction of literary reasoning mechanisms |
| | | Vision-language generateverify loop（Woodpecker） | Yin et al. (2024) [29] | |

## 3. Research Design

### *3.1. Dataset Construction*

In this study, the whole-book reading dataset is chosen as the basis of the research, based on its fitness advantage in inducing and detecting large model hallucinations. Compared with fragmented reading, reading a whole book requires a million-word reading volume [29] , which involves high cognitive load and exhibits long-range dependency, forcing the model to process long-distance information associations, maintain semantic coherence across tens of thousands of characters, and exhibit memory-based hallucinations. At the same time, classical texts contain a large number of cultural contexts of invisible statements, and metaphors, symbols, and culturally

specific elements lead the model to be more susceptible to the phenomenon of hallucination in the face of uncertainty. Therefore, these characteristics of the whole-book reading dataset provide an ideal breeding hotbed for investigating the hallucination problem of large inference models in complex text tasks, which is conducive to revealing the boundaries and risks of current models in text comprehension and inference. The original dataset contains 500 multiple-choice questions, and this study uses stratified random sampling to select 4-5 questions from each of the labeled 0-10 difficulty levels, resulting in a 50-question assessment set. The dataset spans ancient Chinese classics, modern literature, and foreign masterpieces, evenly covering 14 classic masterpieces, such as The Journey to the West, Dawn Blossoms Plucked at Dusk, Jane Eyre, How the Steel Was Tempered, and Selected Poems of Ai Qing, etc., and encompassing 10 questions in the original dataset on linguistic features, characterization, contextual themes, detail extraction, summarization, logical inference, affective attitudes, cross-textual applications, genre styles, and creative interpretations. The structure of the dataset takes into account the gradient of difficulty, the diversity of texts and the distribution of question types, and realizes the triadic balance of "difficulty-bibliography-type".

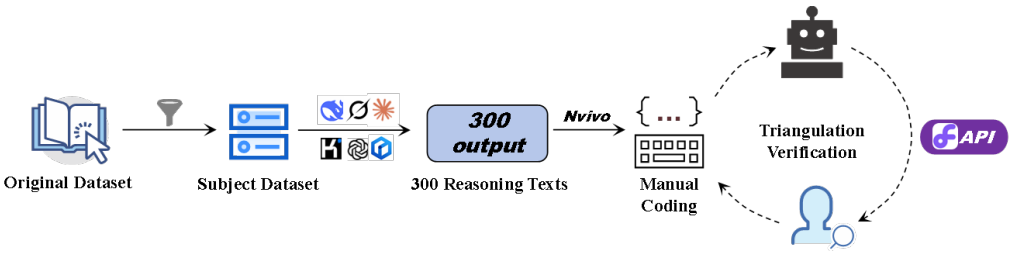



Figure 1: Research Design Process

### *3.2. Model Selection and Testing*

As of April 2025, this study refers to Chatbot Arena, SuperCLUE and other large language models evaluation platforms, and selects six types of representative reasoning LLMs: ChatGPT o1, Grok3, Claude 3.7, Sonnet, DeepSeek-R1, Kimi 1.5, and ERNIE Bot X1. To ensure the fairness and consistency of the test, the research has made unified plans in aspects such as prompt word design, input format and test environment control. As shown in the figure 2, the prompt word template is uniformly designed, and it is clearly required that the model answer in the form of single-answer multiple-choice questions and display the detailed reasoning process. All evaluations were conducted within a 7-day period using identical hardware configurations and the latest stable web interface versions for each model. To maintain test independence, each of the 50 questions was submitted in a separate dialogue session, with all model outputs subsequently undergoing systematic processing including standardization, archival, and formatting. Through the above-mentioned experimental approach of the unified control environment, this study finally obtained 300 complete response data of 6 models for 50 questions, totaling approximately 250,000 characters of reasoning text corpora. This comprehensive dataset serves as the foundation for subsequent chain-of-thought analysis and hallucination research.

In the output process, it is considered that there may be inconsistencies in the output format of the reasoning large model answers, and even other non-standardized correspondences, for example, some LLMs such as DeepSeek-R1 and ERNIE Bot X1 occasionally produced excessively

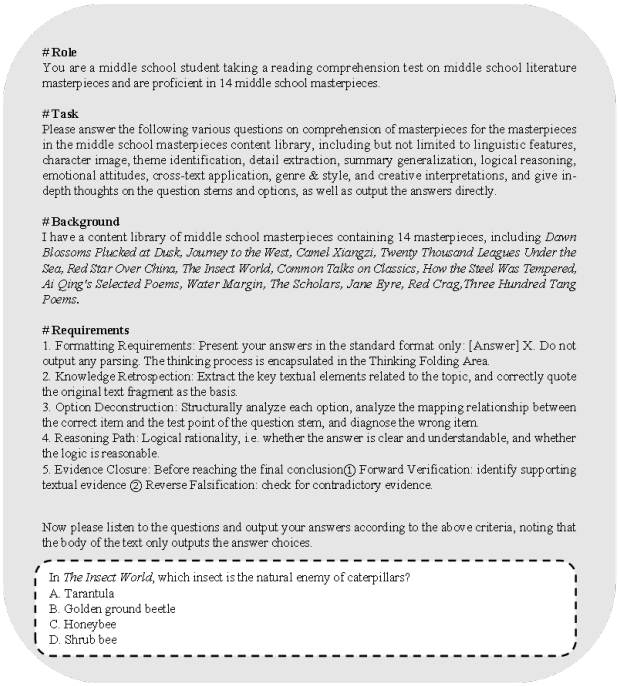



Figure 2: Prompt Design for LLMs Input

verbose responses or entered repetitive argument cycles. To ensure the accuracy and consistency of answer extraction, we implemented a standardized processing protocol with the following steps. The processing is run up to three times, as the output of some of the models will remain unchanged after multiple generation selections.

(1) **Format verification**: If the output matches the pre-formatted answer. If so, the generated result is directly extracted as the final answer.

(2) **Ambiguity resolution**: If the output contains multiple answer indications or there is an ambiguous representation, the final choice explicitly stated in the end of reasoning is extracted as the answer. If the ending is not explicit, the most frequently occurring option in the full text is selected as the model intent answer.

(3) **Prompt refinement**: If none of the above conditions are met, the input prompt is modified to read "Please output only the final answer without any explanation or reasoning process". The process was followed up with a maximum of two follow-up questions, and if a clear answer was not obtained, it was recorded as an "error in form" and the two researchers independently judged their intended answer.

### *3.3. NVivo Coding for Chain-of-Thought Hallucinations*

This study employs Grounded Theory methodology, utilizing the NVivo 12 qualitative analysis software to conduct systematic coding analysis on 300 model reasoning samples. Among these, 250 samples serve as the primary dataset for establishing the initial coding framework, while the remaining 50 samples are reserved for theoretical saturation testing.

Table 2: Hallucination Examples Under Core Axial Coding Categories

| Category | Description | Example | Explanation |
|---|---|---|---|
| **Text Misinterpretation** | Misunderstanding **explicitly stated information** in the original text or que stion | **Q:** Which of the following is not part of Paul ' s " four near-death experiences" in *How the Steel Was Tempered*?<br>**R:** I am looking for Paul's "four near-death experiences." First, I note that Paul was hit on the head with a rifle butt during the battle with the Polish White Army, contracting typhoid. This event typically occurred in 1919. Examining the four near-death experiences in *How the Steel Was Tempered*, including the typhoid infection after being hit on the head and later pneumonia, this feels like uncovering the truth… | This reasoning contains multiple factual errors. First, Paul was indeed injured in the leg and contracted typhoid during the battle with the Polish White Army, but he **was not hit on the head with a rifle butt**, conflating distinct events. Second, the battle **occurred in 1920, not 1919**, indicating a timeline error. |
| **Text Fabrication** | Making associations or interpretations **be yond a reasonable scope** based on the original text | **Q:** Regarding Three Hundred Tang Poems, what is the purpose of Meng Haoran ' s use of the allusion " standing by the abyss envying the fish, better to retreat and weave a net" in Gazing at Lake Dongting, Dedicated to Prime Minister Zhang?<br>**R:** Let's examine the full poem:<br>*The lake glows in August under the autumn moon, a song drunkenly sung in its view.*<br>*Clear and vast, the lake under the autumn moon, a song drunkenly sung in its view.*<br>*A song drunkenly sung in its view, clear and vast, the lake under the autumn moon.*<br>*Clear and vast, the lake under the autumn moon, a song drunkenly sung in its view.* | This reasoning is severely flawed due to the **complete fabrication of the poem's lines in *Gazing at Lake Dongting, Dedicated to Prime Minister Zhang***. In reality, Meng Haoran's original poem reads: "*In August, the lake is level… only left with the feeling of envying the fish.*" The four lines cited, such as "*The lake glows in August under the autumn moon,* " are entirely fabricated and bear no relation to the original poem. |
| **Logical Flaws** | **Structural gaps or inconsistencies** in th e reasoning process | **Q:** The emotional attitudes toward violent behavior in *The Scholars* and *Water Margin* reflect the authors ' differing views on society. Which option is correct?<br>**R:** Violent behavior in *The Scholars* is an isolated phenomenon, with the author maintaining a neutral stance; in *Water Margin*, violence is routine, reflecting dissatisfaction with chaotic social order and a desire to restore traditional order… The violence in *The Scholars* satirizes the hypocrisy of scholars and societal absurdity; *Water Margin* showcases criticism of the impulsive actions of the Liangshan heroes. | The reasoning process **exhibits inconsistent and self-contradictory logic**. The characterization of violence in The Scholars is described as both " neutral " and " satirical, " presenting conflicting stances. In Water Margin, the reasoning first attributes violence to social critique and later to criticism of the heroes ' impulsiveness, without establishing a logical connection. |
| **Formal Errors** | Technical issues in **linguistic expressi on or structure** of the output | **Q:** In *Journey to the West*, many episodes employ the "foreshadowing, subtle clues over a long arc" technique. Which of the following options does not use this technique?<br>**R:** Exploring details: The chaos of Sun Wukong's havoc in Heaven and Princess Iron Fan's approach to the Flaming Mountain prompted deeper reflection on these details. Confirming connections: Evaluating the authenticity and fabrication of " foreshadowing, subtle clues, " examining whether the use of the "Wind-Calming Pill" and "Flying Dragon Staff" intertwine over time, verifying through details— _PROGRESS: Continue analyzing the connections. … | This reasoning **includes code-mixing text**, such as " _PROGRESS: Continue analyzing the connections, " indicating a technical error in the output structure. |

*Note:* Red-marked sections indicate hallucinated outputs.

In the open coding phase, the research team conducted a line-by-line analysis of the reasoning texts imported into NVivo 12, focusing on identifying and labeling various manifestations of hallucinations. Through an in-depth examination of anomalous phenomena in model reasoning, the study extracted 205 reference points and abstracted them into 22 initial conceptual nodes. The coding process strictly followed the progressive logic of "raw statements → conceptualization → categorization", ensuring each label was supported by explicit textual evidence. Building upon the open coding results, the researchers further consolidated these initial nodes into higher-level core categories, as shown in Table 2. After repeated comparative analysis and theoretical discussion, the study finally established a four-dimensional hallucination classification framework: **text misinterpretation, text fabrication, logical flaws, and formal errors**. Each dimension contained several subcategories, forming a clearly hierarchical coding system, as detailed in the coding table shown in Table 3. To validate theoretical saturation, the reserved 50-sample dataset was imported into the system for coding verification. The results revealed no new conceptual categories or subordinate relationships, indicating that the constructed hallucination classification framework had reached theoretical saturation, thereby satisfying the sample adequacy principle in qualitative research.

In order to verify the reliability of the coding framework, this study innovatively adopts the

Table 3: Coding Results of Cot Hallucinations in Reasoning LLMs

| Core Coding | Open Coding | Reference Points | | | | | | Total |
|---|---|---|---|---|---|---|---|---|
| | | o1 | grok3 | Claude | R1 | Kimi | Bot X1 | |
| Text Misinterpretation | Background/Theme Misunderstanding | 5 | 2 | 3 | 1 | 1 | 1 | 86 |
| | Character Confusion Across Works | 2 | 1 | 1 | | | | |
| | Location Information Error | 1 | 2 | 1 | 1 | 1 | 1 | |
| | Emotional Attitude Misjudgment | 4 | 3 | 2 | 1 | 3 | 2 | |
| | Character Deed Misremembering | 1 | 1 | 2 | | | | |
| | Character Portrayal Misunderstanding | 1 | 1 | 2 | | | | |
| | Event Information Error | 4 | 1 | 4 | 1 | 1 | 3 | |
| | Linguistic Feature Misanalysis | 2 | 2 | 1 | 2 | 2 | 2 | |
| | Original Text Misunderstanding | | 6 | 2 | 1 | | 2 | |
| | Author Information Error | | 1 | 2 | | | | |
| Text Fabrication | Unwarranted Extrapolation in Unknown Tex | 6 | 6 | 1 | 3 | 2 | 4 | 82 |
| | Fabricated Plot Elements | 5 | 5 | 11 | 1 | 4 | 3 | |
| | Invented Characters | 3 | | 1 | | 1 | | |
| | Irrelevant Content Insertion | 6 | | | | | | |
| Logical Flaws | Falsified Text Citations | 1 | 7 | 3 | 2 | 1 | 1 | 23 |
| | Over-Extension of Original Content | 4 | | | | 1 | | |
| | Missing Evidence Chain | 6 | 1 | | | | | |
| | Faulty Reasoning Logic | 1 | 4 | | | 2 | 5 | |
| | Circular Argumentation | | | | 2 | | 2 | |
| Formal Errors | Grammatical Errors | 2 | | | | | | 14 |
| | Chinese-English Code-Mixing Gibberish | 2 | 1 | | 3 | | | |
| | Character Encoding Mistakes | 6 | | | | | | |

human-agent collaborative triangular verification mechanism and introduces the AI-assisted verification method. Firstly, the DeepSeek-R1 model, which has the least illusion in the preliminary evaluation, is selected as the AI-Validator, which is positioned as a neutral evaluator, so that it focuses on judging the correctness of the manual coding and has nothing to do with the generative ability of the evaluation tool, thus effectively reducing the interference of the model's own illusion or uncontrollable output. Through API calls on the Silicon Flow platform, we conducted batch verification of all 300 coding points. The large language model's evaluation returned 198 <Correct>judgments that were consistent with manual coding results,and the recall rate was 66%. For the 102 cases with discrepancies, the research team implemented a secondary double-blind manual review process, which yielded the following resolutions:81 cases retained their original manual coding, 12 cases incorporated supplemental coding from the AI-Validator, and 9 cases required joint re-evaluation and redefinition of the reasoning text based on integrated human-AI assessments.

To ensure the reliability of manual coding, this study conducted a reliability test on the manual coding process before formally implementing AI-assisted verification. A random sample of 20% (60 sample points) was selected from the total sample size, and another researcher independently performed blind coding in the NVivo software environment. After coding completion, the Cohen's Kappa coefficient was calculated to measure consistency between coders. The final Kappa value reached 0.76 ($p < 0.001$), indicating excellent coding consistency and providing a reliable foundation for subsequent AI validation.

In cases where AI-Validator identifies inconsistencies, apart from instances where it itself exhibits hallucinations, research has found that its judgment anomalies often stem from confusion between two concepts: self-fiction under unknown texts and fabrication of non-existent plots. Self-fiction under unknown texts refers to the model's excessive completion of blank or

ambiguous sections in the text, often accompanied by uncertain markers such as “possibly” or “perhaps”. On the other hand, the fabrication of non-existent plots refers to the model’s active alteration of known content, typically expressed through definitive statements. Through the AI4S coding validation process, this study maintains the depth of qualitative research while improving coding efficiency and accuracy via machine learning. Eventually, a hallucination classification framework for large reasoning models in whole-book reading comprehension has been established.

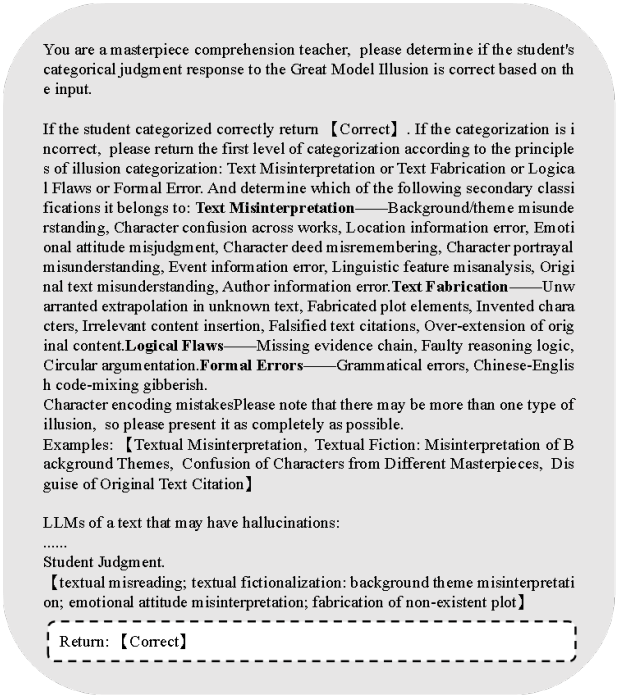



Figure 3: Prompt Template for AI-Validator’s Verification of Manual Coding

## 4. Data Analysis

### *4.1. Overall Analysis of Hallucination Rate*

After completing testing and statistical analysis, we categorized the reasoning outputs and chain-of-thought processes of large language models. To precisely describe different types of model performance, the reasoning outputs were classified into four cases: (1) **Absolute Correctness**: Both the final result and the CoT reasoning process are correct.(2) **Absolute Error**: Both the result and the reasoning process contain errors.(3) **Pseudo-Correctness**: The result is correct, but the CoT process exhibits hallucinations.(4) **Pseudo-Error**: The result is wrong, but the CoT process is correct. As shown in Table 4, the evaluation of reasoning large language models revealed a noteworthy phenomenon: **Hallucination issues in CoT reasoning are prevalent across all tested models.** Experimental data indicate that even when the final result

Table 4: Statistics of Output Results and CoT Hallucinations in Reasoning LLMs

| Model | Developer | Release | Outputs | | Outputs × | |
|---|---|---|---|---|---|---|
| | | | Process | Process × | Process | Process × |
| ChatGPT-o1 | OpenAI | Sep-24 | 14 | **22 (44%)** | 1 | 13 |
| Grok-3 | xAI | Feb-25 | 23 | **7 (14%)** | 0 | 20 |
| Claude 3.7 Sonnet | Anthropic | Feb-25 | 24 | **11 (22%)** | 0 | 15 |
| DeepSeek-R1 | DeepSeek | Jan-25 | 37 | **5 (10%)** | 0 | 8 |
| Kimi-1.5 | Moonshot AI | Jan-25 | 33 | **3 (6%)** | 0 | 14 |
| ERNIE Bot X1 | Baidu | Apr-25 | 29 | **6 (12%)** | 0 | 15 |

is correct, 25.23% of the cases still contain hallucinations in the reasoning chain, constituting Pseudo-Correctness.

This Pseudo-Correctness phenomenon accounts for a significant proportion in multiple models, particularly in ChatGPT o1 (44%) and Claude 3.7 Sonnet (22%), suggesting that although these models perform well in output accuracy, their reasoning chains exhibit instability and fragility. Moreover, such issues in the reasoning process may remain hidden but gradually emerge in subsequent iterative dialogues.We also observed an imbalance between "false negatives" and "false positives". Across all tested models, Pseudo-Error cases were extremely rare, with only one instance occurring in ChatGPT o1. This asymmetry provides strong empirical evidence that high-quality reasoning chains typically lead to correct conclusions, whereas incorrect conclusions almost always stem from flaws in the reasoning process. This finding reinforces the critical role of robust CoT reasoning in achieving accurate results.

### *4.2. Analysis of Hallucination Rate Variations Across Models*

In order to improve the quantitative accuracy of subsequent analysis and evaluate the hallucination generation features of large language models more systematically, this study designs and introduces two core computational metrics: Hallucination Rate (HR) and Hallucination Index (HI) are systematic measures of hallucination phenomena from two dimensions of frequency and intensity, respectively.

(1) Hallucination Rate (HR) formula:

$$HR = \frac{\sum_{i=1}^{N} \Pi(h_i)}{N} \times 100\% \tag{1}$$

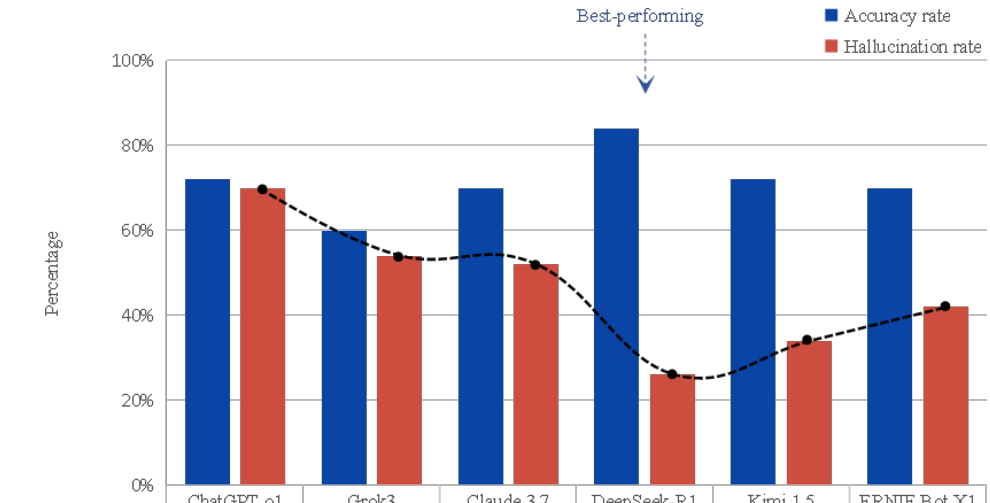
N represents the total test sample size (N=300); $\Pi(h_i)$ is the indicator function, which is 1 when the ith sample is hallucinative, and 0 otherwise. This index is used to reflect the overall proportion level of hallucination content generated by the large language model in all test samples, and has basic frequency statistical significance.

(2) Hallucination Index (HI) formula:

$$HR = \frac{1}{N} \sum_{i=1}^{N} S_i \tag{2}$$

N represents the total number of evaluation questions; $S_i$ is the hallucination severity score for question i (hallucination frequency: 0-5). This index not only considers the presence or

absence of hallucinations, but also evaluates the severity of hallucinations, providing a fine-grained basis for the model to distinguish the performance of hallucinations in different task types and scenarios.




| | ChatGPT o1 | Grok3 | Claude 3.7 | DeepSeek-R1 | Kimi 1.5 | ERNIE Bot X1 |
|---|---|---|---|---|---|---|
| Accuracy rate | 72% | 60% | 70% | 84% | 72% | 70% |
| Hallucination rate | 70% | 54% | 52% | 26% | 34% | 42% |

Figure 4: Comparison of Accuracy Rates and Hallucination Rates Across Reasoning LLMs

As evident from the overall trend in Figure 4, there exists a negative correlation between the accuracy rate of reasoning outputs and the hallucination rate in chain-of-thought processes for reasoning LLMs: better-performing models generally exhibit lower hallucination rates, indicating that the quality of reasoning chains can to some extent predict output accuracy.In the specific task of whole-book reading, three leading Chinese models (DeepSeek-R1, Kimi 1.5, and ERNIE Bot X1) achieved an average accuracy rate of 75.3%, approximately 8 percentage points higher than their international counterparts (ChatGPT o1, Claude 3.7, and Grok3) at 67.3%. Concurrently, the average hallucination rate for domestic models was 34±7.96%, much lower than the 58.7±6.53% observed in foreign models.This performance gap stems from the deeper optimization of Chinese LLMs for native linguistic contexts, particularly in handling culturally nuanced texts, where they demonstrate more mature localization capabilities. The findings suggest that language- and culture-specific tuning plays a critical role in reducing hallucinations while improving reasoning accuracy.

At the individual LLMs level, DeepSeek-R1 shows a double advantage over other models in the same group: its accuracy reaches 84%, while maintaining the lowest hallucination rate (26%), indicating its adaptability on the whole book reading task, which may benefit from its MoE architecture and Chinese context optimization, making it more robust in processing ChatGPT o1 exhibits a distinctive “high accuracy-high hallucination” pattern.Despite achieving a relatively high accuracy rate (72%), its hallucination rate reaches 70%—the highest among all tested models. Further quantification of hallucination indices in Table 5 was conducted using a non-parametric Kruskal-Wallis H test across the six large language models, yielding a statistically significant

result (H=32.599, p <0.001). Pairwise comparisons revealed that ChatGPT o1 differed significantly from DeepSeek-R1, Kimi 1.5, and ERNIE Bot X1 (p <0.05). This result empirically supports the previously noted phenomenon of a "high accuracy–high hallucination rate", indicating that although ChatGPT-o1 demonstrates a higher output correctness rate, its reasoning process is associated with a greater risk of hallucination. Notably, the hallucination index of ChatGPT o1 (1.24) is 3.4 times that of DeepSeek-R1 (0.36), and its maximum hallucination score on a single item reached 5, reflecting the instability and volatility of its reasoning process. This suggests that ChatGPT-o1 frequently relies on significantly biased reasoning paths—generating correct answers by chance even when logical chains are incomplete or fabricated, resulting in Pseudo-Correctness outputs. Such behavior poses substantial risks in real-world applications, as it may mislead users into overestimating the model's true reasoning capabilities.

Table 5: Descriptive Statistics of CoT Hallucination Reference Points

| Variable | Sum | Mean (HI) | SD | Min | Max | 95% Confidence Interval | | H | P |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Upper Bound | Lower Bound | | |
| ChatGPT o1 | 62 | 1.24 | 1.135 | 0 | 5 | 0.925 | 1.555 | 32.599 | <0.001 |
| Grok3 | 44 | 0.88 | 1.003 | 0 | 4 | 0.602 | 1.158 | | |
| Claude 3.7 Sonnet | 36 | 0.72 | 0.834 | 0 | 3 | 0.489 | 0.951 | | |
| DeepSeek-R1 | 18 | 0.36a | 0.663 | 0 | 2 | 0.176 | 0.544 | | |
| Kimi 1.5 | 19 | 0.38b | 0.567 | 0 | 2 | 0.223 | 0.537 | | |
| ERNIE Bot X1 | 26 | 0.52c | 0.677 | 0 | 2 | 0.332 | 0.708 | | |

*Note:* a, b, and c indicate statistically significant differences compared to ChatGPT o1. All comparisons were adjusted using the Bonferroni correction.

### *4.3. Analysis of Hallucination Variations in the Dataset*

In the work-based dataset analysis, based on the evaluation data of 14 middle school Chinese classics (N=300), this study investigated the hallucination performance of the reasoning LLMs from the perspective of literary genre, as shown in Table 6. Through the quantitative analysis of hallucination index HI, it is found that there are significant differences in the degree of hallucination between different literary genres.Among them, the average hallucination index of fiction was the highest (HI=0.7861, SD=0.955), followed by prose (HI=0.6333, SD=0.863) and poetry (HI=0.4285, SD=0.590). To further examine the influence of literary genre on model hallucination, a non-parametric Kruskal-Wallis H test was conducted. The results indicated a marginally significant difference in Hallucination Index (HI) distributions across the three genres (H = 4.978, p = 0.083), suggesting that genre type may exert some influence on hallucination generation. However, the effect did not reach statistical significance under the current sample conditions. This finding implies that while literary genre may contribute to hallucination risk, it is not the sole explanatory factor affecting model hallucinations. Through in-depth analysis of the experimental data, the high HI value of fiction texts stems from their unique triple cognitive challenges.

(1) **Long-range dependency reasoning**: Novels usually contain multi-threaded plots and interleaved relationships between characters, which requires the model to carry out long-range dependent reasoning across chapters. For example, when the technique application of "like snakes and ropes hiding from the bush indicate a foreshadowing of their distant existence". was observed in Journey to the West, due to the large text span, the four reasoning LLMs were unable

to identify the appearance of "Flying Dragon Staff" and "Wind-Calming Pill" in the early stage, which lay a foreshadowing for the follow-up.

(2) **Cultural implicitness**: The titles and customs of traditional novels need the support of background knowledge, and the model is easy to generate fictional content due to cultural errors. For example, in the reasoning process of The Scholars, the big model has confused the cultural common sense of the imperial examination system such as "Sheng ren", "Jie Ren" and "Jian Sheng", which touch the blind spot of its knowledge and produce hallucinations.

(3) **Semantic ambiguity**: Fictional texts, typically characterized by their extended length, frequently incorporate ambiguous expressions such as metaphors, implications, and narrative foreshadowing, alongside subjective content including emotional depictions and stream-of-consciousness techniques. These elements of semantic uncertainty pose significant challenges to models' precise comprehension. In contrast, while poetic works similarly exhibit high degrees of semantic density and symbolism, their comparatively lower information redundancy paradoxically reduces the branching probability of model reasoning paths, thereby mitigating hallucination risks.

Table 6: Statistics of Hallucination Indices (HI) Across Literary Genres

| Genre | Work | HI | SD | 95% CI | G-HI | G-HI |
|---|---|---|---|---|---|---|
| Fiction | *Red Crag* | 0.4444 | 0.705 | [0.09, 0.79] | **0.7816** | 0.955 |
| | *The Scholars* | 1.0417 | 1.233 | [0.52, 1.56] | | |
| | *Water Margin* | 0.7917 | 0.721 | [0.49, 1.10] | | |
| | *Camel Xiangzi* | 0.8333 | 0.963 | [0.43, 1.24] | | |
| | *How the Steel Was Tempered* | 1.0417 | 1.122 | [0.57, 1.52] | | |
| | *Jane Eyre* | 0.6250 | 0.969 | [0.22, 1.03] | | |
| | *20,000 Leagues Under the Sea* | 0.5833 | 0.793 | [0.08, 1.09] | | |
| | *Journey to the West* | 0.7083 | 0.859 | [0.35, 1.07] | | |
| Poetry | *Selected Poems of Ai Qing* | 0.7222 | 0.669 | [0.39, 1.05] | **0.4285** | 0.590 |
| | *300 Tang Poems* | 0.2083 | 0.415 | [0.03, 0.38] | | |
| Prose | *An Introduction to Classics* | 0.6250 | 0.824 | [0.28, 0.97] | **0.6333** | 0.863 |
| | *The Insect World* | 0.5417 | 0.721 | [0.24, 0.85] | | |
| | *Dawn Blossoms Plucked at Dusk* | 0.8889 | 1.183 | [0.30, 1.48] | | |
| | *Red Star Over China* | 0.3889 | 0.502 | [0.14, 0.64] | | |

Starting from the problem type classification of the dataset, this study reveals the differential performance of the models in different cognitive task types through the systematic evaluation of 6 reasoning LLMs in 10 problem classification dimensions (N=300), as shown in Table 5. According to the hallucination index (HI), the task dimension is divided into three groups: high (HI $\geq$ 0.8), medium (0.5 $\leq$ HI $<$ 0.8), and low (HI $<$ 0.5), and the model performance shows significant hierarchical characteristics.

In the **high hallucination group**, the logical reasoning dimension showed the most serious hallucination phenomenon. A total of 34 hallucination reference points were detected in the 30 logical reasoning test questions, of which text fiction accounted for 52.9% (18/34). The logical reasoning task had a high reasoning load and deep reasoning dependence, and the LLMs could not accurately trace the causal chain. It tends to fill the gaps with the approximate logic in its own training corpus. Tasks involving "linguistic features" and "detail extraction" primarily

assess the comprehension of artistic devices and fine-grained textual elements. These question types directly test models' long-tail knowledge processing capabilities, revealing significant limitations: 83.3% of the reasoning LLMs cannot distinguish the difference between metaphor and metonymy. In the use of rhetorical devices, Paul described that Tonia "smells like mothballs all over", the hallucination index is as high as 1.5. This finding demonstrates that reasoning LLMs cannot retain all factual knowledge encountered during pretraining, particularly less frequent long-tail knowledge. When reasoning LLMs process information beyond their limited knowledge boundaries, they exhibit heightened hallucination tendencies, which empirically validate their significant deficiency in low-frequency information processing.

Compared with the **middle hallucination group**, the reasoning LLMs showed stable performance in the basic understanding dimensions of character image (accuracy 80.0%/HI=0.7333), theme identification (73.3%/0.5667) and emotion recognition (80.0%/0.6333), and its accuracy was concentrated in the range of 70.0%-83.3%. The hallucination rate was controlled in the range of 33.3%-53.3%. This performance shows that the reasoning LLMs have strong adaptability in basic person features, topic recognition and sentiment analysis.

In the **low hallucination group**, especially in the creative interpretation dimension, the reasoning LLMs achieved the best performance (HI=0.2667), and its excellent performance requires dialectical interpretation. On the one hand, open-ended tasks have a lower reliance on precise facts, allowing models to reduce fabrication risks through generalized responses. On the other hand, hallucination may even stimulate creativity and possibilities in such tasks. This phenomenon of "creative hallucination" suggests that in specific application scenarios, it is necessary to establish a dynamic adjustment mechanism of hallucination tolerance to seek the optimal balance between factual accuracy and divergent thinking.

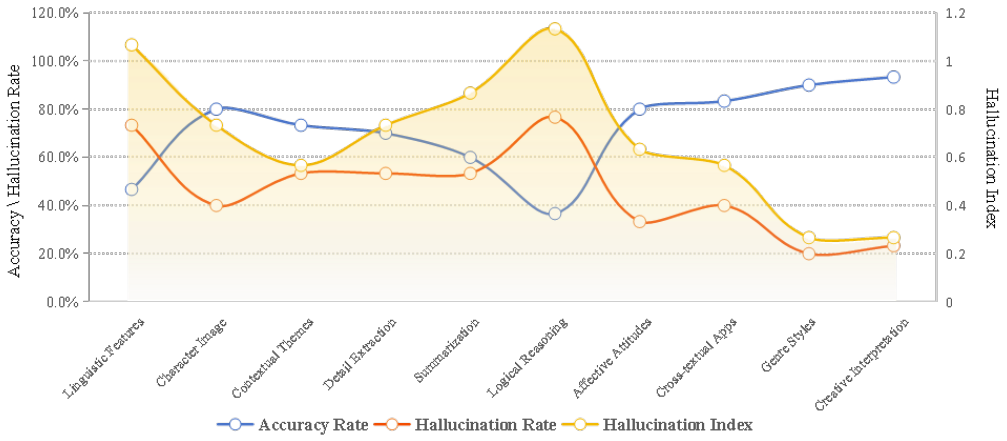



Figure 5: Comparison of the type of problem with the index of hallucinations

In summary, interpreted from a cognitive science perspective, reasoning LLMs performance has an inverted U-shaped relationship with the cognitive load of the task-basic comprehension tasks (e.g., character image) and highly open-ended tasks (e.g., creative interpretation) perform better, whereas medium-complexity tasks (e.g., detail extraction), which require precise logical reasoning, perform the worst. The three-tier classification system constructed in this study provides differentiated optimization pathways: for high-hallucination dimensions, it is neces-

sary to enhance symbolic reasoning modules and long-tail knowledge pretraining; for medium-hallucination dimensions, the knowledge retrieval mechanism should be optimized; and for low-hallucination dimensions, appropriate creative space can be retained. Similarly, differences in literary genres substantially impact the completeness and accuracy of large language models' reasoning chains, particularly when processing literary texts with dense interweaving of narrative and logic, where the risk of LLMs hallucinations increases significantly. Future research should further refine the task-text adaptation mechanism, integrating textual features with reasoning LLMs modeling capabilities to develop pretraining objectives embedded with cultural knowledge graphs and design improved attention mechanisms for long-range dependencies. This will enhance the controllability and interpretability of reasoning LLMs in complex contextual scenarios.

### *4.4. Analysis of Differences in Hallucination Coding*

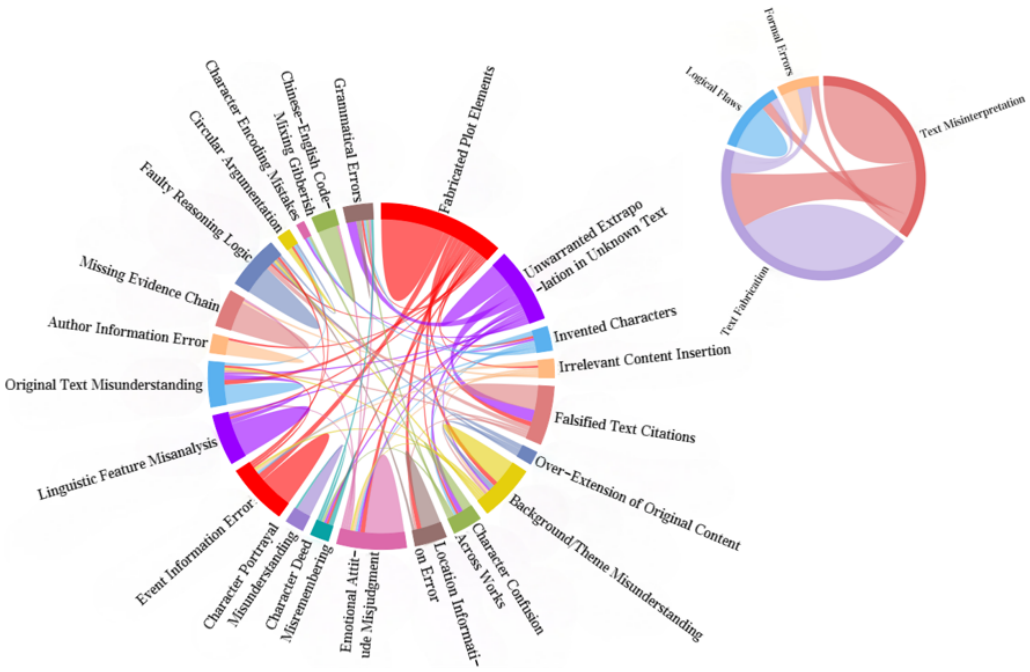



Figure 6: Co-occurrence Relationships Between Core and Open Coding Categories

This study, based on co-occurrence analysis of keywords from core coding and open coding of each output, reveals the distribution characteristics and internal correlations of the primary hallucination types in reasoning large language models. As shown in Figure 6, text misinterpretation and text fabrication constitute the main components in the principal axis coding. Among the 86 reference points for text misinterpretation, multidimensional features are observed, such as confusion in spatiotemporal dimensions, semantic comprehension biases, and misplacement of character relationships, with these misinterpretation phenomena significantly correlated with working memory load. The 82 cases of textual fiction show two typical patterns, one is "compensation for lack of knowledge" (68.2%), which is manifested in the fabrication of non-existent

plots and characters, and the other one is "over-generation out of control" (31.8%), which includes the appearance of completely irrelevant content and the faking of quotations from the original text, including the appearance of completely unrelated content and disguised quotations from the original text.

The axial coding analysis reveals a critical association between text fabrication and text misinterpretation, with 33 co-occurrences, indicating these hallucination types frequently intertwine in model reasoning processes. The intrinsic logic of their high co-occurrence suggests that fabrication often originates from misinterpretation. When LLMs misinterpret the original content, background information, or character traits, it often results in a phenomenon of error propagation. Faced with knowledge gaps, the LLMs tend to engage in creative filling, subsequently constructing seemingly plausible but fundamentally inaccurate content based on this initial misunderstanding. In this process, misinterpretation frequently serves as the contextual trigger for factual fabrication. Errors such as misjudging the thematic background, misconstruing character traits, or misaligning emotional tone impair the LLMs' ability to generate responses grounded in true semantic understanding. As a result, the LLMs are more likely to draw upon the most similar fragments in their pretraining corpus, producing content that appears coherent but is, in fact, fabricated. This mechanism of hallucinatory creativity is particularly pronounced in reasoning-intensive tasks. A detailed analysis of hallucination type co-occurrence reveals the following:

• **Text misinterpretation** co-occurs with text fabrication 33 times, with logical flaws 5 times, and with formal error 8 times. Among these, text misinterpretation often acts as the origin of hallucinations—once it occurs, it tends to trigger more complex hallucinations.

• **Text fabrication** typically emerges as the representational form of hallucination and is often induced by prior misinterpretation, thus forming the "final stage" of hallucinatory output.

• **Logical Flaws** manifest as structural flaws in the reasoning chain, such as breaks in logical progression or causal leaps.

• **Formal errors**, by contrast, are mostly "self-co-occurring", functioning more as technical noise—byproducts of unstable output quality—characterized by irregular expression or formatting errors.

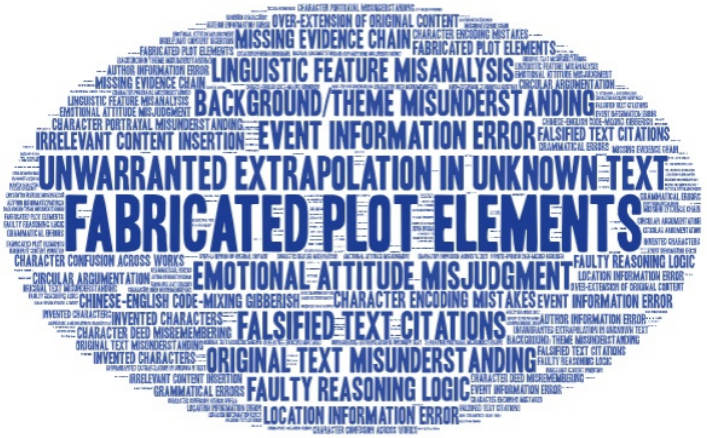



Figure 7: Frequency Distribution of Open Coding Categories

These patterns indicate that "text misinterpretation" is a key mediating variable in the hallu-

cination chain. It can both disrupt the logic of reasoning and serve as the root cause of factual fabrication, thereby playing a central role in the formation of complex hallucinations.

From the word frequency cloud statistics of open coding in Figure 7, fabricating plot elements emerges as the most frequent hallucination type, co-occurring with 15 other open codes. This type of fabrication is not an active creation by the model but rather a synthesis based on multiple misinterpreted details. As shown in Figure 8's co-occurrence analysis, it frequently overlaps with location information errors (2 instances), event memory distortions (3 instances), and original text misunderstandings (2 instances), forming a compound hallucination—highlighting an intersecting amplification effect among hallucination types.The second most prevalent category, unwarranted extrapolation in unknown text contexts, reveals a distinct pathological mechanism: when models lack concrete textual grounding, they frequently employ "citation-formatted fabrication" as a credibility-enhancing strategy. Essentially, this is a disguised form of fabrication, where generated content superficially follows citation logic but is entirely fabricated in origin. As illustrated in Figure 8, the co-occurrence matrix delineates a consistent degenerative sequence.

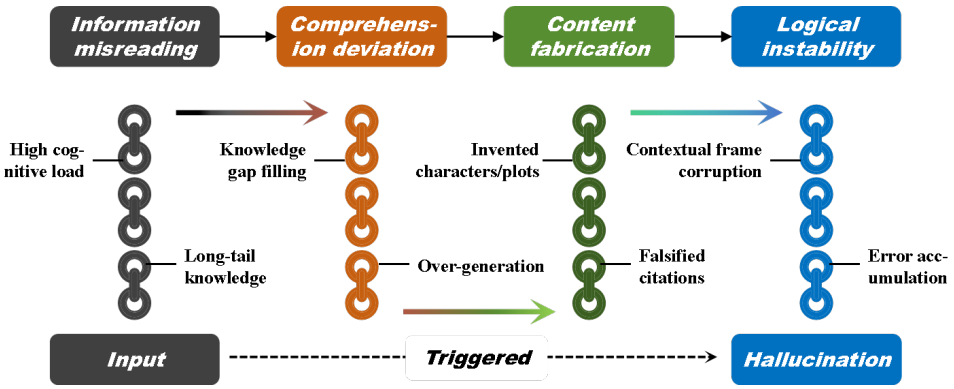



Figure 8: High-Level Hallucination Propagation Chain

This study's revelation of the co-occurrence relationship between text fabrication and misinterpretation demonstrates that hallucinations in large language models when processing complex Chinese literary texts are rarely isolated incidents, but rather follow a chain-reaction process. Fundamentally, hallucination types with creative adaptation characteristics—such as fabricating non-existent plot elements or disguising fictional content as direct quotations—are not random occurrences, but rather the cumulative products of multiple misinterpretations. This suggests that in the future optimization of the thinking chain of reasoning LLMs, we should give priority to improving the basic fact-matching ability and the original text alignment ability of the LLMs and establish a set of traceable index systems for the "hallucination risk chain", so as to reduce the occurrence of hallucination from the source. Therefore, the future optimization of reasoning LLMs hallucination problem can be started from the following two directions. First, the alignment training of the original text is strengthened, and the evidence-based response generation mechanism is added to make the LLMs trace back to the original text at each step of reasoning. Second, the semantic verification mechanism is introduced to check whether the LLM is moving toward fabrication on the basis of misreading in real-time, so as to curb the hallucination in the brewing stage.

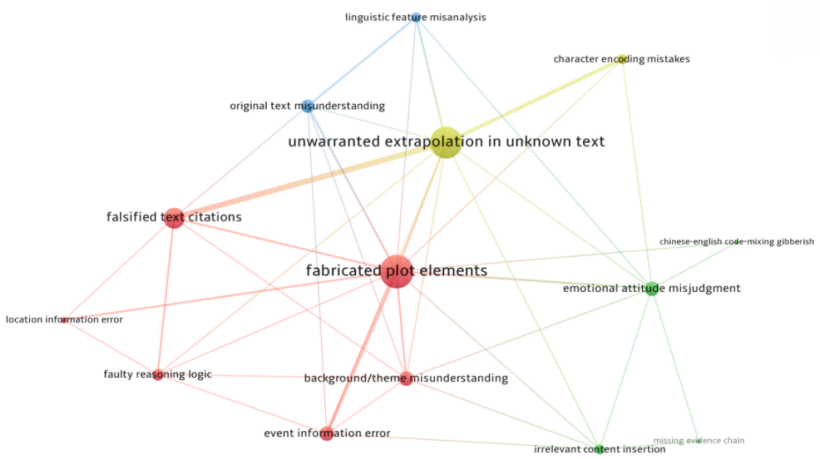



Figure 9: Interconnected Co-occurrence of Open Coding Categories

## 5. Conclusion and Future Work

### *5.1. Research Findings*

To investigate hallucination phenomena in the chain-of-thought reasoning of large language models, this study first constructed a CoT hallucination evaluation dataset through stratified sampling from a whole-book reading corpus, ensuring a balanced representation of difficulty gradients, text diversity, and question types. Subsequently, we designed a few-shot unified prompt template to systematically evaluate six reasoning models: ChatGPT o1, Grok3, Claude 3.7 Sonnet, DeepSeek-R1, Kimi 1.5, and ERNIE Bot X1. Applying NVivo 12 for qualitative analysis and grounded theory methodology, we conducted systematic coding of 300 reasoning samples. The study innovatively introduced an AI-Validator verification mechanism, supplemented by a double-blind manual review, culminating in the first reasoning hallucination classification framework customized for language arts education. Finally, through quantitative analysis using Hallucination Rate (HR) and Hallucination Index (HI), we measured model outputs and CoT processes, revealing key hallucination patterns and characteristics of LLMs in whole-book reading tasks.

The data analysis yielded the following key findings: (1) Chain-of-thought hallucinations were prevalent in reasoning LLMs performing whole-book reading tasks, particularly manifesting as Pseudo-Correctness. Experimental results demonstrated that even when models produced correct answers, their reasoning processes might contain fabricated content or logical flaws. ChatGPT-o1 exhibited a distinctive ”high accuracy-high hallucination” pattern, reflecting instability in its reasoning chains that posed critical concealment and misdirection risks. (2) The phenomenon of hallucination is specific to task type and text genre. Fiction text leads to a higher hallucination index (HI=0.7861) due to long-range dependence and cultural implicitness, while logical reasoning, linguistic features, and detail extraction tasks have the top three hallucination indexes, which are directly related to the ability of long-tailed knowledge processing, which suggests that the model encounters rarer long-tailed knowledge that is more likely to produce hallucinations, and verifies the model's important shortfalls in low-frequency information processing. In addition, Domestic models demonstrated much lower average hallucination

rates (34±7.96%) compared to international models (58.7±6.53%), attributable to their deeper optimization for Chinese linguistic contexts. (3) Hallucinations have a chain generation mechanism. Co-occurrence analysis and grounded theory coding revealed an evolutionary pathway: information misreading → comprehension deviation → content fabrication → logical instability, with text misinterpretation being the critical mediating variable. (4) Hallucinations have a double-edged sword effect, showing differentiated value in different application scenarios. In open-ended tasks such as creative interpretation, hallucinations stimulate innovative output and generate novel insights beyond the training data, but are highly misleading in factual tasks. This double-edged sword effect suggests the need to establish a dynamic adjustment mechanism for the tolerance of hallucinations, and to seek an optimal balance between ensuring factual accuracy and promoting innovative thinking.

### *5.2. Research Implications*

Based on the systematic analysis of hallucination phenomena in reasoning large language models during whole-book reading tasks, we combine experimental data to propose the following implications to optimize the reliability and safety of reasoning LLMs in complex text tasks.

**Firstly, implement uncertainty modeling and expression mechanisms with uncertainty-based abstention to disrupt hallucination propagation chains.** The hallucination phenomenon of the reasoning large model in the whole-book reading task shows a critical chain propagation feature, and its evolution path follows the generation law of information misreading → comprehension deviation → content fabrication → logical instability. The effectiveness of this mechanism has been validated by scholars through cross-task experiments. When a model exhibits high semantic entropy [7] or frequently uses ambiguous expressions [30] during the initial stage of information misinterpretation, the risk of downstream hallucinations increases significantly. By training models to actively recognize and explicitly express uncertainty—choosing to acknowledge knowledge boundaries rather than force unreliable outputs when uncertain.

• **Uncertainty Quantification**: Train models to actively detect and explicitly express confidence levels through phrases like possibly or uncertain when processing ambiguous content. Develop a fuzzy-term frequency analysis model to monitor real-time uncertainty signals.

• **Threshold-Triggered Abstention**: Implementing threshold-triggered abstention mechanisms that activate when confronting knowledge gaps or high semantic ambiguity.

• **Knowledge disclaimer**: Replacing potential fabrications with explicit disclaimers like "This exceeds my current knowledge".

This approach effectively disrupts the hallucination generation chain at its source, proving particularly suitable for high-stakes educational assessment scenarios where precision is essential.Tomani et al. used the TriviaQA data set to implement a waiver strategy based on uncertain statistics such as semantic entropy, which can improve the accuracy of the question answering task by up to 8.2%, and can improve the accuracy by 2% to 8% by sacrificing a small number of highly uncertain samples [31].

**Secondly, to address the long-tail knowledge effect, we propose establishing a model-to-model correction collaborative intervention system.** The study reveals that fiction texts and detail extraction questions exhibit higher hallucination indices due to cultural implicitness and long-tail knowledge, emphasizing LLMs' limitations in processing low-frequency information.To mitigate this, we developed a cooperative framework where LLMs with lower hallucination rates generate prior knowledge to assist higher-hallucination models in knowledge completion and reasoning correction [32],He Jing et al. verified that the method can reduce the hallucination rate of Baichuan-13B model by 51.4% through medical data set [33]. Our AI-Validator

mechanism—using the least hallucinatory model as a neutral verifier—effectively detects hallucinations in other models' outputs, demonstrating the feasibility and efficacy of this model-to-model correctional approach. Looking ahead, the system could evolve into domain-specific "safety-assured LLM" architectures.

• **Expert-vetted datasets**: Curated pretraining datasets stringently filtered by human experts.

• **Model-audited operations**: Dynamic oversight mechanisms where high-reliability models monitor,including knowledge base retrievals, intelligent agent behaviors and base model outputs [34].

Although this paper focuses on the literature domain, the findings and intervention strategies have universal value across domains. In high-risk professional scenarios such as medical diagnosis and legal consultation, model hallucination may bring more serious consequences, and the importance of the uncertainty-based abstention mechanism is particularly prominent.For instance, when medical LLMs encounter rare diseases or contradictory symptoms, proactively terminating reasoning chains proves clinically safer than forcing diagnostic suggestions.At the same time, the model-to-model correction system can take the law database and case database as the prior knowledge source, and the professional law LLMs can verify the compliance of the output of the general model. Therefore, the method in this paper has a strong generalization ability, which provides a reference for reducing the illusion rate of large models in various professional fields.

However, there are still some remaining issues that have not been fully explored in this paper. First, the current evaluation framework mainly focuses on closed-form tasks in the form of single-choice questions. Although this design is convenient for variable control and quantitative analysis, it also limits the in-depth discussion of the association between "generation hallucination" and "reasoning hallucination".Future research will expand to the open question answering paradigm, and reveal the similarities and differences between the two types of hallucinations through comparative analysis. Second, constrained by human annotation costs and data processing scale, the evaluation dataset in this study consists of 50 questions with 300 LLMs response samples. The relatively limited sample size may affect the broader applicability and generalizability of the conclusions. Future work will expand to include multilingual parallel corpora as well as non-literary genres such as technical documents, historical records, and medical reports. This expansion aims to further validate the universality of language localization advantages and investigate the influence mechanisms of texts with varying logical densities on hallucination chains. Third, the determination of hallucination rates relies on the NVivo qualitative coding system. Although this study introduced an AI-Validator mechanism for auxiliary verification, certain subjective judgment biases may still exist during the coding process. We recommend that future research incorporate more objective quantitative metrics and develop fine-grained hallucination severity grading standards to improve the reliability of evaluation results.

## Author Contributions

**Xinyi Liu**: Developed the AI-Validator verification mechanism, performed qualitative coding using NVivo 12, and co-wrote the Research Design and Data Analysis sections.

**Yuting Lu**: Constructed the whole-book reading dataset, coordinated cross-model testing, and contributed to the Abstract and Introduction sections.

**Shunping Wei**: Designed the experimental protocols and model evaluation methodology, conducted statistical analysis of hallucination patterns, and contributed to the manuscript's Results and Discussion sections.

## Acknowledgements

This work is supported by 2024 Innovation Fund Project of Engineering Research Center of Integration and Application of Digital Learning Technology, Ministry of Education (Project No.: 1441001), and The China National Children's Center 2025 Research Project (Project No.: CNC-CYJY2025003), National Natural Science Foundation of China 2022 General Program(Grant No.: 72274234).

## References

[1] Zhang Huimin. How deepseek-r1 was created? *Journal of Shenzhen University Science & Engineering*, 42(2), 2025.

[2] Chen Zhi. Value orientation,evolutionary logic,and forward-looking pathways of new-generation generative artificial intelligence—a case study of deepseek. *Social Sciences in Xinjiang*, (02):17–29+170, 2025.

[3] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[4] ZY Liu, PJ Wang, XB Song, X Zhang, and BB Jiang. Survey on hallucinations in large language models. *Journal of Software*, 36(03):1152–1185, 2025.

[5] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.

[6] Shuo Zhang, Liangming Pan, Junzhou Zhao, and William Yang Wang. The knowledge alignment problem: Bridging human and external knowledge for large language models. *arXiv preprint arXiv:2305.13669*, 2023.

[7] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

[8] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

[9] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[10] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022.

[11] Qiguang Chen, Libo Qin, Jiaqi Wang, Jingxuan Zhou, and Wanxiang Che. Unlocking the capabilities of thought: A reasoning boundary framework to quantify and optimize chain-of-thought. *Advances in Neural Information Processing Systems*, 37:54872–54904, 2024.

[12] Jiahao Cheng, Tiancheng Su, Jia Yuan, Guoxiu He, Jiawei Liu, Xinqi Tao, Jingwen Xie, and Huaxia Li. Chain-of-thought prompting obscures hallucination cues in large language models: An empirical evaluation. *arXiv preprint arXiv:2506.17088*, 2025.

[13] Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.

[14] Li Du, Yequan Wang, Xingrun Xing, Yiqun Ya, Xiang Li, Xin Jiang, and Xuezhi Fang. Quantifying and attributing the hallucination of large language models via association analysis. *arXiv preprint arXiv:2309.05217*, 2023.

[15] Zijun Yao, Yantao Liu, Yanxu Chen, Jianhui Chen, Junfeng Fang, Lei Hou, Juanzi Li, and Tat-Seng Chua. Are reasoning models more prone to hallucination? *arXiv preprint arXiv:2505.23646*, 2025.

[16] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.

[17] Carlos Garcia-Fernandez, Luis Felipe, Monique Shotande, Muntasir Zitu, Aakash Tripathi, Ghulam Rasool, Issam El Naqa, Vivek Rudrapatna, and Gilmer Valdes. Trustworthy ai for medicine: Continuous hallucination detection and elimination with check. *arXiv preprint arXiv:2506.11129*, 2025.

[18] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[19] Zichao Lin, Shuyan Guan, Wending Zhang, Huiyan Zhang, Yugang Li, and Huaping Zhang. Towards trustworthy llms: a review on debiasing and dehallucinating in large language models. *Artificial Intelligence Review*, 57(9):243, 2024.

[20] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021, 2023.

[21] Yilin Wen, Zifeng Wang, and Jimeng Sun. Mindmap: Knowledge graph prompting sparks graph of thoughts in large language models. *arXiv preprint arXiv:2308.09729*, 2023.

[22] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*, 2023.

[23] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530, 2023.

[24] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*, 2023.

[25] Deren Lei, Yaxi Li, Mengya Hu, Mingyu Wang, Vincent Yun, Emily Ching, and Eslam Kamal. Chain of natural language inference for reducing large language model ungrounded hallucinations. *arXiv preprint arXiv:2310.03951*, 2023.

[26] Jianyi Zhang, Da-Cheng Juan, Cyrus Rashtchian, Chun-Sung Ferng, Heinrich Jiang, and Yiran Chen. Sled: Self logits evolution decoding for improving factuality in large language models. *Advances in Neural Information Processing Systems*, 37:5188–5209, 2024.

[27] Yue Jiang, Jiawei Chen, Dingkang Yang, Mingcheng Li, Shunli Wang, Tong Wu, Ke Li, and Lihua Zhang. Comt: Chain-of-medical-thought reduces hallucination in medical report generation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[28] Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105, 2024.

[29] WD. Li. Analytical framework and strategy for setting whole-book reading course. *Language Planning*, (03):9–12+25, 2021.

[30] Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*, 2023.

[31] Christian Tomani, Kamalika Chaudhuri, Ivan Evtimov, Daniel Cremers, and Mark Ibrahim. Uncertainty-based abstention in llms improves safety and reduces hallucinations. *arXiv preprint arXiv:2404.10960*, 2024.

[32] J. He, Y. Shen, and RF Xie. Research on categorical recognition and optimization of hallucination phenomenon in large language models. *Journal of Frontiers of Computer Science and Technology*, 19(05):1295–1301, 2025.

[33] J. He, Y. Shen, and RF Xie. Recognition and optimization of hallucination phenomena in large language models. *Journal of Computer Applications*, 45(03):709–714, 2025.

[34] Hongyi Zhou. Solving security challenges to unleash the full potential of large models. `http://www.xinhuanet.com/politics/20250306/3d3ada53a8654ba8bd17fe2ebb59d6f5/c.html`, 2025.