Biosensing and Biomedicine
https://www.sciltp.com/journals/bab

Article

Have Large Language Models Improved Research Methodology?

Fredric Narcross * and Robert Marks

Avram and Stella Goldstein-Goren Department of Biotechnology Engineering, Ben-Gurion University of the Negev, Be'er Sheva 8499000, Israel
* Correspondence: narcross@post.bgu.ac.il

How To Cite: Narcross, F.; Marks, R. Have Large Language Models Improved Research Methodology? *Biosensing and Biomedicine* **2026**, *1*(1), 1.

Received: 4 April 2026
Revised: 16 May 2026
Accepted: 26 May 2026
Published: 28 May 2026

Abstract: Over the last 30 years, research methods have shifted from manual, library-based approaches to methods supported by digital tools. Current Large Language Models (LLMs), such as ChatGPT, are no longer relegated to literature search and have expanded into research analysis and knowledge discovery. However, their dependability for conducting meaningful research, especially in specialized, interdisciplinary areas, has not been thoroughly assessed. This study evaluated whether contemporary LLMs can reproduce the research outcomes of a fully documented human study: a 1991 article that identified dermatophytosis (ringworm) in historical fine art. This niche, interdisciplinary topic was selected as a deliberate stress test of LLM capabilities in the research frontier, where cross-domain synthesis and deep domain knowledge are required. Ten commercially available LLMs were systematically tested using two prompt conditions: a basic factual query and a complex motivational prompt designed to elicit human-level research performance. LLM responses were graded on a scale from 0 to 3 based on eight characteristics of the artwork and three factors from the benchmark article. Four categories, which included complete fabrication, misattribution, embellishment, and overclaiming, were used to identify, count, and categorize hallucinations. The original paper, digital collections, and museum databases were all examined as part of the verification process utilized for this classification. Statistical comparisons used the Wilcoxon signed-rank test and Fisher's exact test. No LLM rediscovered any of the seven artworks identified in the original article. 3 out of 10 LLMs (30%) produced incorrect information in response to the simple prompt (M = 0.40, SD = 0.70). 9 of 10 LLMs (90%) produced incorrect information in response to the complex prompt, resulting in 48 total instances (M = 4.80, SD = 4.52). The rise was statistically significant (effect size r = 0.91; Fisher's exact test $p = 0.020$; Wilcoxon W = 0.0, $p = 0.004$). Perplexity Pro Deep Research, despite providing the most detailed etiological information (scoring 3 on all three article-level characteristics), also produced the most hallucinations ($n = 17$). LLMs consistently fabricated plausible-sounding content rather than acknowledging uncertainty. In the specialized interdisciplinary domain tested here, current LLMs proved unreliable as autonomous research agents. Between the two prompt conditions tested, hallucination rates rose twelvefold when the prompt moved from a basic factual query to a complex motivational prompt that explicitly demanded a detailed report. This pattern is consistent with prioritizing output completion above factual accuracy; Reinforcement Learning from Human Feedback (RLHF) is discussed as a plausible explanation for the pattern, but not as a demonstrated mechanism. LLMs can be helpful as research tools when experts in the field use them to verify the results. However, they should not be viewed as independent research agents. It is also

relevant for researchers to keep prompts simple and limited in scope to produce more reliable results and to view confident responses with skepticism. Independent verification is necessary.

Keywords: medical mycology; large language models; hallucination; research methodology; RLHF; prompt engineering; dermatophytosis; AI evaluation

1. Introduction

Although research methodology has not changed appreciably over the last three decades, the processes required to carry out its steps have changed considerably. Previously, the processes were manual and laborious, but today those same steps can be highly automated. This raises a fundamental question: do the new computerized processes produce an equivalent or even a better research product?

To answer that question, the processes that were performed to create an article titled "Dermatophytoses in art" [1], published in 1991 by Robert Marks, a graduate student, were compared against the processes performed today by generative artificial intelligence, referred to as Large Language Models (LLMs), for example, ChatGPT [2].

1.1. LLM Evaluation, Hallucination, and AI-Assisted Research

Scholars' interest in LLMs' potential, constraints, and appropriate applications has grown as LLMs have advanced rapidly in research. A worldwide survey on the use of LLMs as research tools by clinical researchers was conducted by Mishra et al. [3]. They discovered that although LLM output was widely used, researchers knew very little about its dependability. Further, computational linguistics has studied the generation of believable but erroneous information by LLMs, known as "hallucinations". Huang et al. analyzed hallucinations and identified knowledge gaps, decoding methods, and the quality of training data as essential factors [4]. They divided hallucinations into two categories: extrinsic (unverifiable with respect to a source) and intrinsic (contrary to a source). Additionally, Ji et al. divided hallucinations into two categories: intrinsic (defying the source) and extrinsic (unexplainable by the source) [5].

Training LLMs with Reinforcement Learning from Human Feedback (RLHF) results in specific LLM behavioral patterns. Ouyang et al. found that RLHF helps LLM output match human preferences. As a result, LLMs aim to provide answers that human assessors consider helpful, harmless, and honest [6]. Christiano et al. also established a framework for LLM learning from human preferences that supports these techniques [7]. Furthermore, Salecha et al. showed that this alignment process produces social desirability bias in LLMs, which tend toward agreeableness and engage in 'impression management' when presented with Big Five personality surveys, similar to human respondents [8]. Altogether, the results directly support the current study's observation that LLMs appear to favor response completeness and user satisfaction over factual accuracy.

In the medical AI domain, advances in vision-language pretraining have demonstrated both the capability and limitations of AI systems for specialized-knowledge tasks. Qin et al. introduced parameter-efficient contrastive methods for medical vision-language pre-training known as "Freeze the Backbones" [9]. Also, Liu et al. developed global-to-dense radiography representation learning [10]. Liu's "G2D" method shows that domain-specific structures can outperform general-purpose models. This result is comparable to what is seen when general-purpose LLMs handle specialized research activities.

Attention has also been drawn to methodological issues for LLM evaluation. Because of its probabilistic-based structure, Banerjee et al. [11] suggested that hallucination is an intrinsic and inescapable property of LLMs. Abdurahman et al. [12] gave practical guidelines for LLM examination. They pointed out that the main obstacle to replication is the non-deterministic, regularly changing commercial models [12].

1.2. The Benchmark: Marks publication

The first stage of the research methodology is identifying a problem [13]. Marks identified a problem in determining if any fine artworks depicted the fungal infection of dermatophytosis, also known colloquially as "ringworm". Marks had an affinity for mycology, the study of fungi, and a keen interest in the history of medicine and fine art. With this problem established, Marks conducted a literature review [14].

The seven artworks identified by Marks [1] serve in this study as a historical comparison point rather than as a putative complete reference set of all surviving artworks depicting dermatophytosis. Marks himself identified, in 1991, five additional candidate artworks that he excluded as insufficiently detailed for confident diagnosis; further examples may well exist in collections that were not examined. The set is, however, appropriate as a

criterion for the present study because every step of the human research process is fully documented, the artwork attributions and provenance can be independently verified through major museum databases, and the source paper provides a complete record against which each LLM output can be checked. An LLM that identified a different but verifiable artwork would therefore not be counted as a failure, and indeed two such artworks were identified by LLMs under Prompt 1 (Section 3.3).

Literature reviews in the early 1990s were conducted in libraries using card catalogs, which were classified into three types: subject, author, and title [15]. The user browsed the collection of subject card boxes to find cards labeled with the topic of interest, then located the corresponding books using the Dewey Decimal classification system displayed on the cards. Books might have been unavailable if another patron had checked them out, resulting in waits of days or weeks. Expensive or oversized books were housed in a non-circulating Reference Collection.

Marks' research strategy was based on the premise that the best chance of finding representative examples of dermatophytosis in art lay in the 17th–19th centuries. Artists of that period depicted their subjects with photographic fidelity; the disease was prevalent, and any sign of it should be readily identifiable. Since reviewing every painting from that period was impractical, he chose a useful shortcut: searching books on the history of medicine, including those focused on art, and then studying the identified sections for skin ailments. He also searched medical and dermatological journals for visual indicators of related papers, then reviewed their references and ordered those that appeared promising.

Since Marks' article was written, technological developments have significantly changed the processes required to conduct research. The Internet made reference materials from libraries, journals, books, and museums available online. As such, search engines could identify references within seconds [16] and specialized academic search engines focused on peer-reviewed scholarly sources [17]. And over time, search engines have incorporated artificial intelligence to deliver even more relevant results and increased personalization [18].

The advent of Large Language Models such as ChatGPT and Perplexity has had a significant impact not only on literature reviews but also on data analysis, writing, and research summarization [3]. These tools are easy to use, quick, and most are free.

In our study, 10 LLMs were used to determine whether any individual LLM, or the entire collection, could reproduce a facsimile of what Marks generated. If automated processes, which can be performed in minutes, could match the quality of manual effort requiring hundreds of hours, the impact on research would be significant.

1.3. Dermatophytosis in Art: Domain Context

It is worth noting that representations of dermatophytosis in art are both unusual and provocative: unusual in that art often presents the human form in a flattering if not ideal light, and a dermatological disfigurement would seem antithetical to those ends; provocative in that presenting a disease on the human form appears to make a statement about the human condition [19]. Ironically, the period of "fine art" is when most of the artwork cited in Marks' article occurred [20].

The Kingdom Fungi consists of organisms that are heterotrophs with cell walls composed of chitin and polysaccharides. They are eukaryotic and include microorganisms such as yeasts, molds, and mushrooms [21]. Dermatophytes are parasitic fungi that infect keratinized tissues, including hair, nails, and skin [22]. Human dermatophyte infections are mainly caused by three genera: Trichophyton, Microsporum, and Epidermophyton [23]. Different forms of dermatophytosis are classified by body location, using the Latin term "tinea" as a prefix, followed by an additional term specifying the infected site. For example, tinea corporis is body ringworm [24].

For millennia, artists have depicted the human body in various postures [25]. Human figures were mainly clothed; therefore, the most frequent representations of dermatophytosis in art are tinea capitis (scalp) and tinea faciei (face) [21].

1.4. Study Contributions

This study makes five particular contributions to AI evaluation and research methodology. First, it provides an empirical benchmark; to our knowledge, this is the first study to compare LLM research outputs with a fully documented human research product on the same topic using structured scoring. The scoring introduces a new complement to benchmark standards such as MMLU and HumanEval. Second, it quantifies the relationship between LLM prompt complexity and hallucinations resulting from the prompts, where more detailed prompts increased hallucination rates twelvefold. Third, it presents a systematic cross-LLM comparison of 10 models under identical conditions, indicating considerable variation between and within companies' products. Fourth, it records a behavior pattern that affects research integrity. LLMs often generate believable content rather than admitting

uncertainty. Fifth, it tackles the less-studied issue of how well LLMs perform on humanities research tasks that require subjective judgment, span multiple fields, and require in-depth knowledge of specific areas.

Selecting 'dermatophytosis in art' as a subject is a stress test for LLMs because such studies frequently span multiple specialized disciplines. And so, the value of an LLM as a research tool is most meaningfully tested on tasks necessitating cross-domain synthesis rather than the retrieval of well-known facts. As will be demonstrated, the detected hallucination patterns are consistent with structural LLM behaviors described across many domains in the broader literature.

2. Materials and Methods

2.1. Method 1: Establishing Benchmark Characteristics

The original article was examined to identify characteristics worth comparing with those that an LLM might discover. The article's characteristics were placed into columns of a spreadsheet, including: (1) contained a summary of early historical records of skin diseases; (2) noted that the presentation of skin disease in art did not take place until after the Gothic period, during the Renaissance; (3) noted that few illustrations of ringworm existed, but those that did are exact enough for determining the likely etiological agent; and (4) for each artwork presented, the following sub-characteristics were evaluated: (a) title of the artwork, (b) name of the artist, (c) year the artwork was completed, (d) where the artwork is housed, (e) the main character(s) received a biographic sketch, (f) the scene was described, (g) the character(s) carrying the disease were described, and (h) the disease etiology was assessed.

2.2. Method 2: Simple Prompt Design

After establishing the article's approach and its germane characteristics, an LLM prompt was created. The prompt was simple, straightforward, and sufficient for an LLM to display what artworks it could uncover from internet literature and museum websites:

> *"What historical artwork depicts ringworm?"*

This prompt is consistent with those that a novice user might use when investigating a research topic for the first time. To ensure the LLMs accurately understood the translation of "ringworm", an additional cross-check prompt was created using synonyms. An initial test with Perplexity Pro Research on both prompts showed virtually no difference, so the simpler prompt was used.

2.3. Method 3: Complex Motivational Prompt Design

A second prompt was included to assess how adept LLMs were at replicating or surpassing the original article. This prompt was designed to represent the motivation and depth that drive a researcher to undertake a knowledge-discovery effort [26,27]. The second prompt was:

> *"You are a PhD student who is passionate about mycology, especially the etiology of mycology. You also have an interest in fine art, and you have noticed that in some Renaissance artworks, there appear to be depictions of dermatophytosis. So, you decide to conduct a detailed research project investigating works of fine art, before, during, and after the Renaissance, that might depict dermatophytosis. You analyze the art to determine whether it depicts dermatophytosis. You write a very detailed report of all possible depictions of dermatophytosis in fine art, and you also provide a very detailed description of the particular dermatophytosis, as well as your detailed etiological assessment of why you believe the depiction is indeed a representation of dermatophytosis".*

These two prompts were developed based on user behavior rather than being designed explicitly through optimized prompt engineering. The simple prompt illustrates how a beginner would start investigating a topic. The complex prompt illustrates how a motivated researcher could use an LLM to produce important research outputs. This type of design was intended to test whether LLMs can conduct humanities research in a way that parallels how humans actually experience it, rather than simply testing whether better prompting would improve their performance. The comparison between these two prompt conditions is an exploratory two-condition contrast rather than a controlled isolation of a single variable. The two prompts differ along several dimensions simultaneously: length, specificity, the explicit assignment of a researcher role, and an explicit demand for a detailed report. Any of these (or their interactions) could contribute to the observed effect. In particular, because the complex prompt

explicitly demands a "very detailed report", the elevated hallucination rate may reflect pressure to complete the output rather than the prompt's complexity per se. The two-condition design does not, on its own, separate these factors. We know that experimenting with different prompt variations, such as applying different approaches to phrase the same query, will be more useful and is a goal for future work.

2.4. Method 4: LLM Selection and Configuration

Ten LLMs were selected based on four criteria. Market prominence and user reach: ChatGPT (in both its 3o Deep Research and 4.1 variants) was included as the most widely used LLM globally. Gemini Pro 2.5 (Google) and Grok 3 (xAI) were included as flagship products of major technology companies. Perplexity Pro Deep Research was selected due to its research orientation. Further, our selection reflects technical diversity, including proprietary closed-source models (ChatGPT, Gemini, Grok, Perplexity), an open-source model (DeepSeek R1), a model available through open-weight distribution (Llama 3), an enterprise model (IBM's Granite), and a general-purpose assistant (Claude). The selection also reflects accessibility. All LLMs were free or available through widely accessible free tiers during testing. Lastly, ChatAI Research Assistant was included as a lesser-known tool available to consumers, in contrast to the more well-known models. Altogether, the collection encompasses tools that a typical researcher might select. We also paid attention to several marketing claims: ChatGPT 3o's claimed to have an IQ of 136 (Mensa Norway test [28]) and Grok 3's positioning as the "World's Smartest AI" [29]. These claims made both of them targets for empirical review.

We acknowledge that this selection is not exhaustive. Notable exclusions include Mistral and Cohere, which were omitted due to scope constraints. The non-exhaustive nature of the LLM selection is a stated limitation.

Table 1 presents the technical details for each LLM tested. All testing was carried out using the default web interfaces encountered by a typical researcher. Most consumer-facing LLM interfaces do not expose parameters such as temperature, top-p, or top-k; when they are exposed, default values are used. This is an intentional design choice: the study evaluates LLMs as they are actually used, not under controlled API conditions. However, reproducibility is, inherently, limited with non-deterministic, often-updated commercial LLMs: a challenge recognized in the wider LLM evaluation literature [12].

Table 1. LLM Technical Details.

LLM	Full Name	Type	Access Date	Interface	Settings	Retrieval
ChatAI-RA	Chat & Ask AI Research Assistant	Proprietary	May 2025	Web	Default	Unknown
GPT3o-DR	ChatGPT 3o Deep Research	Proprietary	May 2025	Web	Default	Yes
Claude	Anthropic Claude	Proprietary	May 2025	Web	Default	Limited
DeepSeek-R1	DeepSeek R1	Open-source	May 2025	Web	Default	No
GPT4.1	ChatGPT 4.1	Proprietary	May 2025	Web	Default	Yes
Gemini Pro-2.5	Google Gemini Pro 2.5	Proprietary	May 2025	Web	Default	Yes
Granite	IBM Granite (Enable Thinking)	Enterprise	May 2025	Web	Default	Unknown
Grok-3	xAI Grok 3	Proprietary	May 2025	Web	Default	Yes
Llama-3	Meta Llama 3	Open-weight	May 2025	Web	Default	No
Perplexity-Pro DR	Perplexity Pro Deep Research	Proprietary	May 2025	Web	Default	Yes (RAG)

Note: All LLMs were accessed through their default consumer web interfaces. "Retrieval" indicates whether the LLM has built-in web search or retrieval-augmented generation (RAG) capabilities. "Default" settings mean no user-configurable parameters were modified from their out-of-the-box values.

2.5. Method 5: Scoring Rubric

Responsiveness of the LLMs' replies to the prompts was graded on a scale of 0–3 for each benchmark characteristic. The rubric was developed de novo for this study because no pre-existing, validated instrument exists to evaluate LLM performance in interdisciplinary art-historical research tasks. On the face of it, the validity of the rubric comes from its direct mapping to the characteristics of the benchmark article [1]. Content validity is supported by the fact that the traits being scored of artwork identification, artist attribution, clinical description, and etiological assessment are objectively verifiable against the source. The scale applies the following operational definitions:

(0) Unresponsive: The LLM's response provides no relevant information about the characteristic in question. For example, when asked about artworks depicting dermatophytosis, the LLM provides only a general definition of ringworm and does not cite specific artworks. Decision criterion: the response contains no artwork identification and no domain-specific clinical or art-historical content that maps to that characteristic. Worked example: DeepSeek R1, which scored 0 on all three article-level characteristics according to Prompt 2.

(1) Partially responsive: The LLM's response refers to a characteristic but has gaps or inaccuracies in describing the cause. Example: naming an artwork and artist, but providing no description of the depicted pathology, or instead providing a description that is vague and clinically non-specific. Decision criterion: at least one verifiable element (e.g., correct artist or correct century) is present, but at least one other required element (e.g., clinical description tied to a specific lesion morphology) is absent or generic.

(2) Comparable: The LLM's response is qualitatively similar in depth and accuracy to Marks' article. Example: correctly identifying an artwork, naming the artist, dating the work, and providing a plausible clinical interpretation. Decision criterion: all required verifiable elements (artist, dating, museum location, clinically specific description) are present and correct against the source, but the response does not exceed Marks' treatment in scope or detail.

(3) Exceeds: The LLM's response exceeds the element in the original article by means of its depth, accuracy, or scope. Example: Perplexity Pro's etiological descriptions were more detailed than in Marks' article. Decision criterion: in addition to satisfying the criteria for score 2, the response provides verifiable content that extends the Marks treatment, e.g., genus-level etiological detail, additional historical context, or more specific morphological characterization, and that content can be independently verified against external sources. Fabricated "additional detail" never qualifies for score 3; any extension to the Marks treatment must itself be verifiable.

Scoring was performed by the authors, who have domain expertise in mycology and art history. Each LLM response was independently reviewed by both authors and assigned a score on the 0–3 scale; the two scores were then compared. Where the two authors agreed, the score was retained; where they disagreed, a reconciliation meeting was held in which each disputed score was adjudicated against the relevant primary source (the original Marks [1] article, museum databases for artwork existence and provenance, or high-definition images for clinical claims). Disagreements were resolved through discussion rather than averaging, on the rationale that nearly all rubric items reduce to a verifiable factual question, e.g., is the artist attribution correct against the museum record? Is the cited clinical feature visible in the high-res image? Additionally, inter-rater reliability was not formally assessed, which is listed as a limitation. Formal multi-rater assessment using Cohen's κ or weighted κ with at least one rater external to the present authorship is recommended for replications and cross-domain extensions in future studies. All graded responses were recorded into spreadsheets as presented in Tables 1–3.

Table 2. Hallucination Counts by LLM and Prompt Condition.

LLM	Prompt 1 Hal.	Prompt 2 Hal.	Change	Prompt 1 Paintings	Prompt 2 Paintings	Prompt 3 (Med. Illus.)
ChatAI-RA	0	3	+3	1	0	1
GPT3o-DR	2	4	+2	1	0	1
Claude	0	3	+3	0	0	0
DeepSeek-R1	0	4	+4	0	0	0
GPT4.1	0	4	+4	0	0	0
Gemini Pro-2.5	0	0	0	0	0	1
Granite	1	5	+4	0	0	1
Grok-3	0	3	+3	0	0	0
Llama-3	1	5	+4	0	0	0
Perplexity-Pro DR	0	17	+17	2	0	1
Summary	4	48	+44	4	0	5/10
Mean (SD)	0.40 (0.70)	4.80 (4.52)				

Note: "Hal" stands for the quantity of claims of hallucinogenic artwork. "Paintings" refers to the quantity of artworks that can be verified and are not hallucinations. "Prompt 3" asks whether the LLM claimed that medical illustrations are fine art (1 = yes, 0 = no). (Prompt 1 vs. Prompt 2 hallucinations). Signed-rank Wilcoxon test: W = 0.0, p = 0.004. Fisher's exact test: proportion hallucinating, p = 0.020. The effect size is R = 0.91. Perplexity-Pro DR showed both the highest number of hallucinations and the best performance on Characteristics 1–3 (highlighted).

Table 3. Characteristics 1–3 Scores (Prompt 2).

LLM	Char 1 (Historical)	Char 2 (Post-Gothic)	Char 3 (Diagnostic)	Total (max 9)	Hal. (Prompt 2)
ChatAI-RA	1	1	1	3	3
GPT3o-DR	1	0	0	1	4
Claude	1	2	2	5	3
DeepSeek-R1	0	0	0	0	4
GPT4.1	1	0	1	2	4
Gemini Pro-2.5	2	2	2	6	0
Granite	1	1	1	3	5
Grok-3	2	2	2	6	3
Llama-3	1	0	0	1	5
Perplexity-Pro DR	3	3	3	9	17
Mean (SD)	1.30 (0.82)	1.10 (1.10)	1.20 (1.03)	3.60 (2.84)	4.80 (4.52)
Median	1.0	1.0	1.0	3.0	4.0

Note: Scoring scale: 0 = unresponsive, 1 = partially responsive, 2 = comparable to original, 3 = exceeds original. Spearman's ρ (Total vs. Hallucinations) = −0.226, $p = 0.529$ (not significant).

2.6. Method 6: Hallucination Typology and Verification Protocol

Realizing that hallucinations could become a pattern in LLM responses, a systematic method for identifying and classifying the hallucinations was implemented. Each factual claim made by an LLM was verified against three sources: (a) the original Marks [1] article; (b) museum databases and digitized collections for artwork existence and provenance; and (c) high-res images of cited artworks for clinical claim evaluation.

Hallucinations were classified into four categories:

Complete fabrication: The LLM invented an artwork that does not exist by fabricating a title, artist, and/or museum location with no basis in reality.

Misattribution: The LLM cited a real artwork but incorrectly stated that it depicts dermatophytosis, as no such condition is visible upon inspection.

Embellishment: The LLM correctly identified a relevant artwork but made up clinical details. For example, ChatGPT 3o claimed there was a visible forearm plaque in "The Young Beggar", but the boy's forearm was covered by his shirt, rendering it impossible to identify any such plaque.

Overclaiming: The LLM presented speculative or ambiguous interpretations as definitive clinical diagnoses while asserting unwarranted confidence in pathological features.

The distinction between hallucination and interpretive difference is essential. When an LLM suggested that a medical illustration qualifies as fine art, this was treated as an interpretive difference, i.e., a legitimate gray area rather than as a hallucination. A supplementary Prompt 3 ("Does medical illustration qualify as fine art?") was used so that LLMs could argue their point, and their argument quality was recorded in a separate column.

Hallucination categories were assigned through a fixed decision sequence applied to each artwork claim made by an LLM. First, the cited artwork was searched in major museum databases and digitized collections; if no record of the artwork could be found under the cited title, artist, or approximate date, the claim was coded as a complete fabrication. Second, if the artwork was verified to exist but did not in fact depict dermatophytosis on visual inspection of high-res images, the claim was coded as a misattribution. Third, if the artwork was correctly identified as relevant to dermatophytosis but the specific clinical feature claimed by the LLM (e.g., a lesion on a particular body part) was not visible on inspection, the claim was coded as embellishment. Fourth, if the artwork and the depicted disease were both plausible but the LLM presented an ambiguous or speculative interpretation as a definitive clinical diagnosis with unwarranted confidence, the claim was coded as overclaiming. As with the rubric scoring, category assignments were performed independently by both authors and reconciled through joint discussion against the relevant primary source. The full LLM response transcripts and the per-claim coding spreadsheet are available from the corresponding author upon reasonable request, so that any reader can independently verify the category assignments.

2.7. Method 7: Statistical Analysis

This study includes an empirical examination using descriptive and nonparametric statistics and does not present results in a typical hypothesis-testing experiment. This is due to the small sample size ($n = 10$) and the ordinal measurement scale. Parametric tests (e.g., *t*-tests, ANOVA) would not be statistically significant for ordinal ratings from 10 observations.

The descriptive statistics in this study include interquartile ranges, standard deviations, means, and medians for all scored variables. The main inferential test used in this study was the Wilcoxon signed-rank test, which evaluated paired hallucination counts from the same LLMs under the two stimuli without assuming normality. Fisher's exact test compared the proportion of LLMs that hallucinated at least once for each prompt. The formula $r = |z|/\sqrt{n}$ was used to calculate the effect size, and we evaluated the link between hallucination frequency and research quality scores using Spearman's rank correlation. All studies used Python (SciPy v1.11).

2.8. Note on Non-Traditional Primary Source Citations

This study's reference list includes three categories of non-traditional sources. These references are necessary and appropriate for the study design and include the LLM product URLs [2,30–35], which cite the specific software tested and follow standard practice for software in empirical studies. Also, references to museum and artwork URLs [36–44] document the particular artworks that constitute the study's ground truth and follow the accepted citation standard in art history scholarship. The Library of Congress card catalog guide reference [14] documents the 1991 historical methodology against which LLM processes are compared, which is a primary source for the historical process being described.

3. Results

3.1. Quantitative Overview

Table 2 summarizes the hallucination counts of each LLM along with the corresponding prompts. For the simple prompt (Prompt 1), 3 of 10 LLMs (30%) produced at least one hallucination, for a total of 4 hallucinations (M = 0.40, SD 0.70, Mdn = 0.0, range: 0–2). In the complex "motivational" prompt (Prompt 2), 9 of 10 LLMs (90%) produced at least one hallucination, with a total of 48 hallucinations (M = 4.80, SD = 4.52, Mdn = 4.0, range: 0–17). Gemini Pro 2.5 was the only LLM that did not hallucinate for Prompt 2, though it explicitly stated that its examples were "hypothetical" rather than factual.

The increase in hallucination counts from Prompt 1 to Prompt 2 was statistically significant (Wilcoxon signed-rank test: W = 0.0, $p = 0.004$). Nine of 10 LLMs had higher hallucination counts under Prompt 2 than Prompt 1, with one tie (Gemini Pro 2.5, which had 0 hallucinations under both conditions). The effect size was very large ($r = 0.91$). The increase in the proportion of LLMs hallucinating (from 30% to 90%) was also statistically significant (Fisher's exact test: $p = 0.020$, OR = 0.048).

No LLM discovered any of the seven artworks identified in the original Marks' article. Under Prompt 1, LLMs across three models identified four artworks. Still, these were distinct from the artworks in the benchmark: two medical illustrations and two fine artworks that were missing in the original article. Under Prompt 2, no LLM identified any verifiable artwork. Also, the hallucination rate was severe, and all artwork citations were found to be fabricated.

3.2. Research Quality Scores: Characteristics 1–3

Table 3 presents the LLMs' scores for the three article-level characteristics used to evaluate Prompt 2 responses, along with the resulting total score. The mean total score was 3.60 (SD = 2.84, Mdn = 3.0, range: 0–9). Performance varied across the LLMs, with Perplexity Pro Deep Research scoring the highest (9/9), followed by Gemini Pro 2.5 and Grok 3 (both 6/9), and the other LLMs scoring appreciably lower. DeepSeek R1 scored lowest (0/9). Notably, there was no significant correlation between hallucination frequency and research quality scores (Spearman's $\rho = -0.226$, $p = 0.529$), indicating that these two dimensions of LLM performance are statistically independent. The results of this decoupling, particularly for researchers who might use response quality as a credibility heuristic, are examined in the Discussion.

3.3. Benchmark Artwork Comparison

The original Marks [1] article identified seven artworks depicting dermatophytosis: *Saint Elisabeth of Hungary curing tinea children* [36]; *Saint Thomas of Villanova giving alms* [37]; *La Boda* [38]; *The Prayer of the Teig Children* [39]; *die Regenten des Leprosenhauses* [40]; *De vier Regenten van het Leprozenhuis te Amsterdam* [41]; and a stone sculpture of Saint Elizabeth of Hungary known as "Le Petit Teigneux" [42]. Marks also identified five further artworks that might depict scalp ringworm, but noted they "do not show enough detail to be convincing".

Collectively, the LLMs identified four artworks under Prompt 1 (none under Prompt 2), none of which overlapped with Marks' seven artworks: *Kerion from ringworm* (1872, New Sydenham Society Atlas) [45]; *El joven mendigo* ("The Young Beggar") [43]; and two medical illustrations (*Ringworm; lesions on the inside right*

wrist [44]; *A boy with a skin disease of the scalp* [44]). Notably, several of Marks' artworks are now digitized and publicly accessible: Murillo's paintings are documented on museum websites, Goya's *La Boda* is in the Museo del Prado's online collection, and the Rijksmuseum's digital archives include the Vinkeles' work, which implies that the LLMs should, in principle, have had easier access to these artworks than Marks did in 1991.

3.4. Illustrative Cases: Hallucination Patterns

A representative example of each of the four hallucination categories follows: *Complete fabrication (Perplexity Pro DR, Prompt 2):* the LLM cited an artwork by attribution, title, and museum location for which no record could be located in the cited museum's online catalog or in standard art-historical reference databases; subsequent disambiguation queries did not surface the work. *Misattribution (Granite, Prompt 2):* the LLM cited a verifiable artwork by a real artist, with a title and provenance, but stated it depicted dermatophytosis; inspection of the high-resolution museum image showed no skin lesions consistent with dermatophytosis on any depicted figure. *Embellishment (ChatGPT 3o DR, Prompt 1):* in Murillo's *The Young Beggar*, the LLM claimed "on the boy's forearm, an annular, scaly plaque is visible, consistent with tinea corporis", but the high-res image shows the boy's forearms entirely covered by his shirt (detailed below in Section 3.4.1). *Overclaiming (Perplexity Pro DR, Prompt 2):* from an ambiguous patch of pigmentation that could plausibly represent a number of dermatological conditions, the LLM gave a confident, specific etiological attribution to a single dermatophyte genus without admitting the diagnostic ambiguity that would be standard in a clinical or art-historical writeup.

3.4.1. Illustrative Case 1: Embellishment (ChatGPT 3o Deep Research, Prompt 1)

ChatGPT 3o Deep Research responded to Prompt 1 by saying that in Murillo's *The Young Beggar*, "on the boy's forearm, an annular, scaly plaque is visible, consistent with tinea corporis". However, the painting's inspection revealed that the boy's forearm is entirely covered by his shirt, with only his hands exposed. When confronted with this fact, the LLM immediately retracted: "You are absolutely right—the boy's forearms in Murillo's El joven mendigo are hidden inside an over-long linen shirt so that no annular plaque can be seen there". Following the admission, the LLM withdrew two additional claims.

3.4.2. Illustrative Case 2: Inverse Accuracy–Detail Relationship (Perplexity Pro, Prompt 2)

Perplexity Pro Deep Research presented a paradox: it was simultaneously the best-performing LLM on Characteristics 1–3 (scoring 3/3/3 and the only LLM to have exceeded the original article's etiological detail), but it was also the worst-performing on hallucinations (17 fabricated artwork claims under Prompt 2). It produced 0 hallucinations under Prompt 1 with 2 verified paintings. This inverse relationship demonstrates that domain knowledge and factual accuracy can be disassociated in LLMs. While an LLM can display sophisticated subject-matter expertise, it can also fabricate evidence to which that expertise is applied.

3.4.3. Illustrative Case 3: Gemini Pro 2.5's Unique Response Pattern (Prompt 2)

Gemini Pro 2.5 was the only LLM that did not hallucinate a response to Prompt 2. It identified artworks, as requested, like the other LLMs, but said that its examples were hypothetical: "The hypothetical examples are designed to showcase how I would analyze actual artworks if undertaking this as a full PhD project". This kind of response represents a qualitatively different strategy: honest uncertainty rather than sheer fabrication to please the user. This might reflect an architectural or training difference worthy of further investigation.

4. Discussion

The results of this study expose significant limitations of LLMs (at the time of the study) in specialized interdisciplinary research: no artworks were rediscovered, there was a 12-fold increase in hallucinations from the complex prompt, and a consistent pattern of fabrication taking precedence over honest uncertainty. These limitations have implications beyond dermatophytosis in art.

4.1. Output-Completion Prioritization and RLHF

The most significant result is the difference in response to prompt complexity. Responses to Prompt 1 (a basic factual query) produced relatively few hallucinations (4 total, 30% of LLMs). On the other hand, Prompt 2, intending to represent a complex motivational prompt, produced hallucinations that increased twelvefold to 48, affecting 90% of the LLMs (Wilcoxon $W = 0.0$, $p = 0.004$; effect size $r = 0.91$).

We note that this behavioral pattern is compatible with one plausible interpretive frame: Reinforcement Learning from Human Feedback (RLHF) training [6,7]. RLHF ensures that LLM outputs align with human evaluators' preferences by awarding points to answers that appear helpful, complete, and consistent with the user's intent. Salecha et al. used Big Five personality surveys to measure social desirability bias in LLMs. They found that they tended to be more agreeable and better at managing their impressions to others [8]. Our results agreed with these outcomes. When we prompted the LLMs to respond to an ambitious research question, they tended to provide more responsive answers than accurate ones. It is important to be clear that the present study does not directly test RLHF: we did not vary RLHF exposure across the tested LLMs and have no access to the proprietary models' training protocols. The output-completion prioritization pattern is the empirical finding; the RLHF account is a candidate explanation among others. This interpretation is consistent with the observed data and the existing literature, but we are not claiming to have demonstrated a causal mechanism. We also do not attribute psychological states or intentions to LLMs. The pattern described of output-completion prioritization is a behavioral observation from the statistical relationship between prompt demands and output accuracy.

4.2. LLM Strengths and Complementary Use

Even though the LLMs had problems with hallucination, they yet presented strengths in some results. For example, Perplexity Pro Deep Research's etiological descriptions exceeded those of the original article, scoring 3 on all three article-level characteristics. Gemini Pro 2.5 and Grok 3 provided comparable-quality responses for Characteristics 1–3, each scoring 6 out of 9. Several LLMs demonstrated structured output generation and a breadth of contextual knowledge that would be valuable in a complementary workflow.

These outcomes suggest that LLMs are most useful when augmenting tools for researchers who already possess domain expertise. A mycologist using an LLM to generate initial hypotheses or contextualize findings could benefit from the etiological depth demonstrated by Perplexity Pro, provided they independently verify every factual claim. The danger lies in the use by researchers who lack the expertise to detect fabrication.

4.3. The Independence of Hallucination and Quality

A natural question arising from the data is whether hallucination frequency and research quality are systematically related. Two competing hypotheses are plausible. Under what might be called the "ambition hypothesis", LLMs that hallucinate more do so because they are generating more verbose, detailed, and ambitious responses; the hallucinations would be a price stemming from ambition, and the non-hallucinated content should correspondingly be richer and score higher on the quality rubric. Under the alternative "poor model hypothesis", LLMs that hallucinate more are simply lower-quality models overall, and hallucinations would be a symptom of uniformly poor performance.

To test these hypotheses, Spearman's rank correlation was computed between Prompt 2 hallucination counts and Characteristics 1–3 total scores across the 10 LLMs. The result ($\rho = -0.226$, $p = 0.529$) provides no support for either hypothesis. The correlation is weakly negative but not statistically significant ($p > 0.05$). As well, the statistical power is limited by $n = 10$, and so the direction of the evidence slightly favors neither the ambition nor the poor-model account. Instead, hallucination behavior appears to be essentially decoupled from content quality and is an independent dimension of LLM performance.

This decoupling has a practical implication that is arguably more precarious than if poor models were simply uniformly poor. In other words, response quality is not related to response reliability. A well-written, detailed, and seemingly authoritative LLM response can be full of fabrications, while a more modest response might be entirely accurate. The confidence and polish of the output provide no insight into its truthfulness. For a naïve user, precisely the user most likely to rely on apparent quality as a credibility heuristic, this decoupling creates a particularly insidious failure mode.

This finding connects directly to the RLHF discussion above. Models trained through reinforcement learning from human feedback are optimized for a reward signal that is essentially "did this response look good to a human evaluator?" rather than "is every factual claim independently verifiable?" The Spearman result suggests that this training successfully optimizes the quality dimension (apparent helpfulness, depth, and polish) independently of, and sometimes at the expense of, the accuracy dimension. Hallucinations, in this framing, are not a "quality problem"; rather, it is a failure that requires independent verification regardless of how authoritative the response appears.

4.4. Corpus Gaps vs. Model Limitations

A pertinent question here is whether the LLMs' failure to discover Marks' artworks reflects gaps in their training data (Hypothesis A) or reflects core limitations of their models (Hypothesis B). It is useful to separate

three distinct factors that this dichotomy can conflate: (a) corpus coverage, i.e., whether information about a particular artwork was present in the LLM's training corpus at all; (b) retrieval and indexing behavior at query time and whether, for the RAG-enabled LLMs, the system actually issued queries that surfaced the relevant material, and whether the indexed source documents were correctly matched to the query; and (c) model-level reasoning and grounding, i.e., whether, when relevant material was either present in context or absent, the model produced a faithful response or instead generated unsupported claims. The present study cannot fully discriminate these factors because we did not have access to model-internal retrieval logs, training-corpus inventories, or grounding diagnostics for the proprietary models tested. Within this caveat, several observations still favor a substantial contribution from factor (c). First, many of Marks' artworks are now digitized and publicly accessible on major museum websites, so that they should be well represented in web-scraped training data. Second, and more importantly, the LLMs did not respond with "I could not find relevant artworks". Instead, they fabricated responses. This behavioral pattern exceeds the question of corpus availability. Regardless of what was or was not in their training data, the decision to generate fiction rather than acknowledge ignorance is a model-level failure.

Three types of experimental design could be used to examine these influences in the future. A controlled corpus injection experiment could give an LLM direct access to Marks' article and evaluate it, enabling the LLMs to identify the correct artworks. A verification experiment would test LLMs on artworks that are extensively documented online, which would establish a baseline for their (the artworks) retrieval. From there, a refusal-rate analysis would systematically measure how often LLMs acknowledge ignorance versus fabricate responses across domains with varying levels of availability.

4.5. Retrieval-Augmented Generation: Limitations Observed

The study yields indirect evidence on the effectiveness of retrieval-augmented generation (RAG). The tested LLMs varied in their retrieval capabilities (Table 1). Notably, several RAG-enabled models, including Perplexity Pro Deep Research, ChatGPT 3o Deep Research, and Grok 3, still hallucinated extensively. Perplexity Pro, which is specifically designed as a research tool with real-time web retrieval, produced the most hallucinations of any LLM (17 under Prompt 2). It is important to carefully scope this observation. We did not have access to the retrieval sources, queries, or citation logs of the RAG-enabled systems tested, and our observations are therefore behavioral rather than mechanistic. The finding is that the specific RAG-enabled consumer systems tested, on the specific task evaluated, did not avoid extensive fabrication; it should not be read as a general claim about retrieval-augmented generation as an architectural class. More transparent RAG architectures, or future systems that expose their retrieval chains and require verifiable citations for every factual claim, would be needed to support stronger claims about RAGs more broadly.

Future architectures that require verifiable citations for every factual claim, where the LLM cannot assert a claim without linking it to a retrievable source, may offer a promising way to reduce fabrication in research contexts.

4.6. Narrow Domain Design Choice

Dermatophytosis in art was a design choice that narrowed the scope of the topic. This is similar to other research in specialized disciplines that focuses on niche, interdisciplinary topics. Evaluating LLMs on well-known topics, such as common medical conditions or well-documented historical events, would have only tested retrieval capabilities rather than research capacity.

The patterns of behavior noted in this study, particularly the marked increase in fabrication between the two prompt conditions tested and the continuing tendency to fabricate plausible texts instead of failing to acknowledge a lack of knowledge, reflect basic LLM behaviors and are not specific to any particular LLM. OpenAI's system cards report that its newest o3 and o4-mini reasoning models hallucinated 33–79% of the time on general factual benchmarks (the PersonQA and SimpleQA tests) [46]. The domain-specific findings of the present study are consistent with, and illustrative of, these broader patterns. We emphasize, however, that our data come from a single specialized interdisciplinary case. Domain-general conclusions about LLMs and research methodology are not warranted from this study alone; the present findings should be read as a domain-specific illustration of patterns reported across many other domains in the broader hallucination literature, rather than as a domain-general claim in their own right. Cross-domain replication is an essential direction for future work.

4.7. Resource Landscape: 1991 vs. 2025

Marks in 1991 and LLMs in 2025 used entirely different resources. Marks drew from physical card catalogs, reference collections, and personal trips to museums. He checked every source directly. The LLMs, by comparison, access much larger though unevenly indexed digital collections. The comparison between the two is structurally

asymmetric and should not be read as a head-to-head contest. Marks brought to the task domain expertise in mycology and the history of medicine, an iterative search strategy that allowed for backing out of dead ends, and source verification at every step. The LLMs, by contrast, were given a single prompt under their default consumer interface, with no opportunity for incremental refinement and no human-in-the-loop verification. The question we can therefore answer with our data is not whether LLMs can match a domain-expert human researcher in absolute terms; rather, the question is whether LLMs, in the zero-shot consumer-interface deployment that a non-expert researcher would actually use, produce reliable outputs on a specialized interdisciplinary task. That is the asymmetric question of practical relevance for the naïve users with whom this study is most concerned. Within this scoping, the comparison favors the LLMs in resource access since they have decades of digitized data, including many of the very artworks Marks identified, and yet they found none of the seven original artworks.

4.8. Practical Recommendations for Researchers Using LLMs

From the results of this study, we offer five practical recommendations for researchers considering the use of LLMs as research tools:

1. Always verify claims that appear factual, and cross-reference to primary sources every LLM citation, claim, and clinical description.
2. Use simple prompts for factual queries. Our simple prompt produced 4 hallucinations, whereas the complex prompt produced 48.
3. Be cautious of confident responses. The Spearman analysis ($\rho = -0.226$, $p = 0.529$) demonstrated that an LLM's apparent expertise provides no support for the response's factual reliability. This was demonstrated by Perplexity Pro, which, although scoring the highest in quality, produced the most fabrications.
4. Cross-reference across multiple LLMs. The significant variation among LLMs (even within the same company's product line) means consulting multiple models can help identify inconsistencies.
5. LLMs are most useful when the researcher already has domain expertise. An expert can use LLM outputs to generate hypotheses, provide context in the literature, and create structured summaries. They can also spot and eliminate fabricated content. A naïve user lacks this important filter.

4.9. The Evolving LLM Landscape

LLM capabilities are advancing rapidly. Since the data collection for this study, multimodal models with image analysis capabilities, improved RAG architectures, and dedicated reasoning models have been released. However, it is notable that even the latest reasoning models do not eliminate hallucinations, and in some cases increase them. OpenAI reports that its o3 model hallucinated 51% on SimpleQA and 33% on PersonQA, while the more advanced o4-mini model hallucinated 79% and 48% respectively [46]. The pattern documented in the present study, in which increased capability coexists with persistent or increased hallucinations, appears to be a basic challenge rather than a temporary limitation.

4.10. Limitations

This study has limitations to be acknowledged when evaluating the results. First, this study compares LLMs on a specific interdisciplinary topic, dermatophytosis in art. This single-domain design is a substantive constraint, and we have aimed throughout to scope domain-general claims accordingly. Although the observed behavioral patterns are consistent with other research on hallucination, cross-domain replication is needed to confirm their generalizability. Secondly, the scoring rubric did not assess inter-rater reliability (IRR), so multi-rater validation would increase confidence in the scoring. That said, the rated characteristics are essentially objectively verifiable. For replications and cross-domain extensions, we recommend formal multi-rater validation employing Cohen's κ or weighted κ with at least one rater external to the present authorship. A third limitation of our study is that it examined only two prompt conditions, resulting in only a coarse prompt-sensitivity analysis. This suggests that we have a subset of data that was systematically tested, and that further strengthening of the findings would require multiple variations of the prompts. In particular, the two prompts differ on several dimensions simultaneously: length, specificity, the explicit assignment of a researcher role, and an explicit demand for a detailed report.

The present design cannot, on its own, isolate which of these (or which interaction among them) drives the twelvefold increase in hallucinations. We therefore present that effect as a finding from these two particular prompt conditions and not as a general claim about how all complex prompts behave. Future prompt-sensitivity studies should vary length, role assignment, and detail demand independently across a larger prompt grid. Fourth, commercial LLMs are inherently non-deterministic, so reproducing outcomes is not always perfect. Such prompts

can, in fact, yield diverse results over time when used with the same LLM. In addition, the underlying model versions change frequently and are often not publicly versioned. The specific access dates for each model tested are listed in Table 1; an attempt to replicate these results today might result in a different model under the same product name. Fifth, although 10 different LLMs were tested, this list was not exhaustive. Major models, such as Mistral and Cohere, were excluded due to scope limits.

Sixth, the comparison between a multi-month, domain-expert human research program [1] and a zero-shot consumer-interface LLM prompt is structurally asymmetric, as discussed in Section 4.7. The asymmetry is itself the question we can answer with our data, that is, how reliable are LLMs in the deployment context a non-expert researcher would actually use. The comparison should not be read as a head-to-head contest between humans and machines. Seventh, the full LLM response transcripts and the per-claim coding spreadsheet were retained but are not embedded in the manuscript; they are available from the corresponding author upon reasonable request. Embedding the transcripts of 10 LLMs across two prompt conditions in-text would substantially enlarge the paper without adding analytic content beyond the representative examples already given (Section 3.4).

5. Conclusions

This study shows that, in the specialized interdisciplinary case examined here, current LLMs deployed as zero-shot autonomous research agents via their default consumer interfaces failed to reproduce the outcomes of a documented human research effort and did so in ways that could harm research integrity. The most important findings are:

(1) No LLM rediscovered any of the seven artworks identified through manual research in 1991.
(2) Between the two prompt conditions tested, hallucination rates rose twelvefold (from 4 to 48, $p = 0.004$) when the prompt moved from a basic factual query to a complex motivational prompt that explicitly demanded a detailed report; this pattern is consistent with prioritization of output completion instead of factual accuracy, although the two-condition design cannot, on its own, isolate the contributions of length, role-assignment, and detail-demand.
(3) LLMs consistently fabricated plausible content rather than acknowledging the limits of their knowledge.

These conclusions do not suggest that LLMs are without value for researchers. On the contrary, the etiological expertise demonstrated by several LLMs indicates genuine potential to augment research. The critical distinction is between augmentation, in which LLMs serve as aids for researchers with domain expertise who can verify outputs, and autonomy, in which LLMs are treated as independent research agents whose outputs are accepted without verification.

In our view, if forthcoming developments in LLM technology include retrieval architectures that require verifiable citations and training methods that reward honesty over seeming helpfulness, then LLMs should be used for research. However, right now, LLMs should be seen as powerful but unreliable assistants. They are useful when experts critically review their responses, but risky if taken at face value. These conclusions are drawn from a single, specialized, interdisciplinary case (dermatophytosis in fine art) and two particular prompt conditions; cross-domain replication and systematic prompt-sensitivity studies are needed before they can be extended to LLM-supported research methodology more generally.

Author Contributions

R.M.: conceptualization, supervision, review and editing; F.N.: methodology, software; data curation, writing—original draft preparation. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

Not applicable.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

No AI tools were utilized for this paper.

References

1. Marks, R. Dermatophytoses in Art. *J. Med. Vet. Mycol.* **1991**, *29*, 1–8. https://doi.org/10.1080/02681219180000021.
2. ChatGPT 2025. Available online: https://chatgpt.com/ (accessed on 2 March 2026).
3. Mishra, T.; Sutanto, E.; Rossanti, R.; et al. Use of large language models as artificial intelligence tools in academic research and publishing among global clinical researchers. *Sci. Rep.* **2024**, *14*, 31672. https://doi.org/10.1038/s41598-024-81370-6.
4. Huang, L.; Yu, W.; Ma, W.; et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *arXiv* **2023**, arXiv:2311.05232.
5. Ji, Z.; Lee, N.; Frieske, R.; et al. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* **2023**, *55*, 1–38. https://doi.org/10.1145/3571730.
6. Ouyang, L.; Wu, J.; Jiang, X.; et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 27730–27744.
7. Christiano, P.F.; Leike, J.; Brown, T.; et al. Deep reinforcement learning from human preferences. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
8. Salecha, A.; Ireland, M.E.; Subrahmanya, S.; et al. Large language models display human-like social desirability biases in Big Five personality surveys. *PNAS Nexus* **2024**, *3*, pgae533. https://doi.org/10.1093/pnasnexus/pgae533.
9. Qin, J.; Liu, C.; Cheng, S.; et al. Freeze the Backbones: a Parameter-Efficient Contrastive Approach to Robust Medical Vision-Language Pre-Training. In Proceedings of the ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, 14–19 April 2024.
10. Liu, C.; Ouyang, C.; Cheng, S.; et al. G2D: From global to dense radiography representation learning via vision-language pre-training. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 14751–14773.
11. Banerjee, S.; Agarwal, A.; Singla, S.; et al. LLMs Will Always Hallucinate, and We Need to Live with This. *arXiv* **2024**, arXiv:2409.05746.
12. Abdurahman, S.; Salkhordeh Z, A.; Moore, A.K.; et al. A Primer for Evaluating Large Language Models in Social-Science Research. *Adv. Methods Pract. Psychol. Sci.* **2025**, *8*, 25152459251325174. https://doi.org/10.1177/25152459251325174.
13. QuestionPro. Research Process Steps: What They Are + How to Follow. Available online: https://www.questionpro.com/blog/research-process-steps/ (accessed on 10 March 2026).
14. Alphonse, N. The Main Stages of the Research Process—A Review of the Literature. *Int. J. Res. Rev.* **2023**, *10*, 671–675.
15. Library of Congress. How to Use the Card Catalog. 2020. Available online: https://guides.loc.gov/card-catalog/using-the-card-catalog (accessed on 10 March 2026).
16. inetAsia. The Importance of Search Engines. 2025. Available online: https://www.inetasia.com/articles/the-importance-of-search-engines/ (accessed on 11 March 2026).
17. Wang, J. Streamline Your Research Using Academic Search Engines. Times Higher Education. 2022. Available online: https://www.timeshighereducation.com/campus/streamline-your-research-using-academic-search-engines (accessed on 3 March 2026).
18. Jirousek, C. The Evolution of the Idea of Art. Art, Design, and Visual Thinking. 1995. Available online: http://char.txa.cornell.edu/ART/FINEART/EVOLIDEA/evolidea.htm (accessed on 19 March 2026).
19. Wallentine, A. How Artists Have Explored and Understood the Human Body through Time, Hyerallergic. 2022. Available online: https://hyperallergic.com/716611/flesh-and-bones-getty-research-institute/#:~:text=Posted%20inArt-,How%20Artists%20Have%20Explored%20and%20Understood%20the%20Human%20Body%20Through,Getty%20Research%20Institute%2C%20Los%20Angeles) (accessed on 8 March 2026).
20. What Is Fine Art? Definition, History, and Examples. Grove Gallery. 2025. Available online: https://grovegallery.com/blogs/articles/what-is-fine-art-definition-history-and-examples?srsltid=AfmBOoonDLLAqTlFTfJeFELro_a2PSNcmoY05BAgGFn2O-uuSk2vj7NZ (accessed on 8 March 2026).
21. Fungi. 2019. Available online: https://byjus.com/biology/kingdom-fungi/ (accessed on 7 March 2026).

22. Moskaluk, A.; VandeWoude, S. Current Topics in Dermatophyte Classification and Clinical Diagnosis. *Pathogens* **2022**, *11*, 957.
23. Hon, A. Mycology of Dermatophyte Infections. 2023. Available online: https://dermnetnz.org/topics/mycology-of-dermatophyte-infections (accessed on 7 March 2026).
24. Kovitwanichkanont, T.; Chong, A. Superficial Fungal Infections. *Aust. J. Gen. Pract.* **2019**, *48*, 706–711.
25. Pepin, J. The Artistry of Jacques Pepin. 2025. Available online: https://jacquespepinart.com/art/the-human-form-in-art-history (accessed on 7 March 2026).
26. Svartefoss, S.; Jungblut, J.; Aksnes, D.; et al. Explaining research performance: Investigating the importance of motivation. *SN Soc. Sci.* **2024**, *4*, 105. https://doi.org/10.1007/s43545-024-00895-9.
27. Streefkerk, R. Inductive vs. Deductive Research Approach. Steps & Examples. 2023. Available online: https://www.scribbr.com/methodology/inductive-deductive-reasoning/#:~:text=Published%20on%20April%2018%2C%202019,at%20testing%20an%20existing%20theory (accessed on 7 March 2026).
28. OpenAI'S O3 Scores 136 on the Mensa Norway Test, Surpassing 98% of the Human Population. CryptoSlate. 2025. Available online: https://cryptoslate.com/openais-o3-scores-136-on-mensa-norway-test-surpassing-98-of-human-population/ (accessed on 7 March 2026).
29. High, M. Why Elon Musk Claims Grok-3 Is the World'S 'Smartest AI'. 2025. Available online: https://aimagazine.com/articles/is-grok-3-really-the-smartest-ai-on-earth (accessed on 7 March 2026).
30. Chat & Ask AI, Research Assistant. 2025. Available online: https://askaichat.app/assistant?assistant=research_assistant (accessed on 7 March 2026).
31. DeepSeek 2025. Available online: https://deepseek.ai/ (accessed on 1 March 2026).
32. Gemini Pro 2.5. 2025. Available online: https://aistudio.google.com/app/prompts/new_chat?model=gemini-2.5-pro/ (accessed on 23 March 2026).
33. Granite Enable Thinking. 2025. Available online: https://www.ibm.com/granite/playground/ (accessed on 12 March 2026).
34. Grok 3. 2025. Available online: https://grok.com/ (accessed on 15 March 2026).
35. Llama 3. 2025. Available online: https://www.llama.com/ (accessed on 1 March 2026).
36. Andaluca, Santa Caridad Hospital, Available online: https://en.andalucia.org/listing/santa-caridad-hospital/15961101/ (accessed on 15 March 2026).
37. Murillo, B. Saint Thomas of Villanova Giving Alms, Museum of Fine Arts in Seville, Spain. 1668. Available online: https://www.museosdeandalucia.es/web/museodebellasartesdesevilla/obras-singulares/-/asset_publisher/GRnu6ntjtLfp/content/santo-tomas-de-villanueva-dando-limosnas?redirect=%2Fweb%2Fmuseodebellasartesdesevilla%2Fobras-singulares%3Fp_p_id%3D101_INSTANCE_GRnu6ntjtLfp%26p_p_lifecycle%3D0%26p_p_state%3Dnormal%26p_p_mode%3Dview%26p_p_col_id%3Dcolumn-2%26p_p_col_pos%3D1%26p_p_col_count%3D2%26_101_INSTANCE_GRnu6ntjtLfp_delta%3D6%26_101_INSTANCE_GRnu6ntjtLfp_keywords%3D%26_101_INSTANCE_GRnu6ntjtLfp_advancedSearch%3Dfalse%26_101_INSTANCE_GRnu6ntjtLfp_andOperator%3Dtrue%26p_r_p_564233524_resetCur%3Dfalse%26_101_INSTANCE_GRnu6ntjtLfp_cur%3D7&inheritRedirect=true (accessed on 30 March 2026).
38. Goya, F. La Boda (The Wedding). Museo del Prado, Madrid, España. 1792. Available online: https://www.museodelprado.es/coleccion/obra-de-arte/la-boda/6340b840-5e11-49cd-9151-0c1fdd240389 (accessed on 4 March 2026). (In Spanish)
39. Pils, I. The Prayer of the Teig Children. Musee de l'Assistance Publique, Hopitaux de Paris, Paris, France. 1853. Available online: https://en.muzeo.com/art-print/the-prayer-of-the-children-suffering-from-ringworm/isidore-pils (accessed on 7 March 2026).
40. USEUM. Available online: https://www.useum.org/artwork/The-Regents-of-the-Leper-colony-in-Amsterdam-in-1649-Ferdinand-Bol (accessed on 11 March 2026).
41. Vinkeles, R. De Vier Regenten van Het Leprozenhuis te Amsterdam, Rijksmuseum, Amsterdam, The Netherlands. 1769. Available online: https://www.rijksmuseum.nl/nl/collectie/object/De-regenten-van-het-Leprozenhuis-1649--c6c0ad57afbd895d93ae31d74dc3cdd3 (accessed on 13 March 2026).
42. Ministry of Culture, POP, Open Heritage Foundation. Available online: https://pop.culture.gouv.fr/notice/palissy/PM28000982 (accessed on 15 March 2026).
43. Spanish Arts, El Joven Mendigo, Available online: https://www.spanish-art.org/spanish-painting-el-joven-mendigo.html (accessed on 15 March 2026).
44. Mukhopadhyay, A.K. A Historical Note on the Evolution of "Ringworm". *Indian J. Dermatol. Venereol. Leprol.* **2019**, *85*, 125–128. https://doi.org/10.25259/IJDVL_123_2019.
45. Historical Illustrations of Skin Disease: Selections from the New Sydenham Society Atlas 1860–1884. Yale University Library. 2022. Available online: https://onlineexhibits.library.yale.edu/s/skin-diseases/page/home (accessed on 23 March 2026).
46. OpenAI. OpenAI o3 and o4-mini System Card. 16 April 2025. Available online: https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf (accessed on 20 March 2026).