publications

MDPI

Article

LLM-Generated Visual Summaries of Cochrane Plain-Language Summaries: A Pilot Expert-Rated Study and Implications for Research Communication

Kannan Sridharan [1,*], Gowri Sivaramakrishnan [2], Juliet Rebecca Joseph [3] and Jerome Gnanaraj [4]

[1] Department of Pharmacology & Therapeutics, College of Medicine & Health Sciences, Arabian Gulf University, Manama P.O. Box 26671, Bahrain
[2] Bahrain Defence Force Royal Medical Services, Riffa P.O. Box 28743, Bahrain; gowri.sivaramakrishnan@gmail.com
[3] Department of Health Sciences, Northeastern University, Boston, MA 02115, USA; rebeccaa.joseph05@gmail.com
[4] Department of Medicine, Johns Hopkins Bayview Medical Center, Johns Hopkins University School of Medicine, Baltimore, MD 21224, USA; jerome.gnanaraj@gmail.com
* Correspondence: skannandr@gmail.com

Abstract

Background: Plain-language summaries (PLSs) improve accessibility of medical research for patients but remain predominantly text-based. Large language models (LLMs) can now generate images from text. We carried out the present exploratory study to evaluate whether LLMs can generate images that demonstrate technical accuracy and theoretical visual usability when derived from PLS content. **Methods**: In this cross-sectional pilot study, two PLS were randomly selected from each of 37 Cochrane Library themes. Three LLMs, ChatGPT-5.2, Google Gemini 3 Pro, and Google Notebook, generated one image per PLS, yielding 222 images. Two blinded assessors evaluated images using a preliminary, internally expert-validated tool that measured technical accuracy and completeness, visual usability, and hallucination presence. Inter-LLM comparisons were assessed using linear mixed effects model. **Results**: Gemini outperformed ChatGPT and Notebook across all domains ($p < 0.001$). Hallucinations occurred exclusively in ChatGPT-generated images (29.73%). Gemini demonstrated the least intra-thematic variability, whereas ChatGPT showed the highest. No significant interaction was found between LLM type and Cochrane theme. Sensitivity analyses, including alternative weighting schemes and leave-one-theme-out analyses, confirmed robust model rankings. **Conclusions**: In this single-prompt expert-rated pilot study, Google Gemini 3 Pro reliably generated accurate, hallucination-free visual summaries from PLS. These exploratory findings support further patient-centered validation of LLM-generated images as complements to text-based patient education materials.

Keywords: artificial intelligence; Gemini; ChatGPT; notebook; plain-language summary

Academic Editor: Eungi Kim

Received: 20 June 2026
Revised: 9 September 2026
Accepted: 14 September 2026
Published: 16 September 2026

1. Introduction

Patient education instructions are important for promoting health literacy, improving patient understanding, and supporting effective self-management of disease (Berkman et al., 2011). Standard approaches typically include written discharge summaries, printed educational materials, and verbal counseling, which aim to communicate important information regarding diagnoses, medications, and follow-up care (Johnson & Sandford, 2005).

These strategies can improve patient comprehension and outcomes when designed appropriately, and techniques such as simplified written materials and the teach-back method are commonly used to reinforce understanding (Kripalani et al., 2008; Schillinger et al., 2003). However, written instructions are not equally effective for all patients, as many materials exceed recommended reading levels and can be difficult to understand, particularly for individuals with limited health literacy (Badarudeen & Sabharwal, 2010; Kessels, 2003). These limitations highlight the need for more accessible approaches to communicating medical information.

Plain-language summaries (PLSs) have emerged as an important strategy to improve accessibility of complex medical information by translating clinical content into understandable language for patients and caregivers. PLS are summaries of scientific articles written in non-technical language to make research findings accessible to non-specialist audiences, including patients and caregivers (Dormer et al., 2022). Evidence suggests that PLS improve understanding of research results, help patients place findings into context, and support communication of shared decision making between patients and healthcare providers (Lobban et al., 2022). Additionally, PLS plays an important role in improving patient engagement, trust, and adherence by presenting information in a clear and accessible format adjusted for non-expert audiences (Barmpas, 2026; Berkman et al., 2011). PLS vary in format and quality and are not yet consistently included in research publications, highlighting the need for improved standardization and distribution (Lobban et al., 2022).

Despite their benefits, PLS are still primarily text-based, which may limit their effectiveness for certain populations. Visual representations are hypothesized to reduce cognitive load by leveraging the brain's dual-coding capacity, the ability to process verbal and visual information through distinct but interconnected channels. According to dual-coding theory, information presented in both verbal and visual formats creates redundant memory traces, facilitating encoding, storage, and retrieval (Kanellopoulou et al., 2019). For medical information, visual and graphical representations have been shown to improve comprehension, recall, and engagement by utilizing visual processing pathways and reducing reliance on reading proficiency even in patients with limited health literacy (Houts et al., 2006; Katz et al., 2006; Delp & Jones, 1996). Importantly, evidence suggests that combining text with visual elements is often more effective than text alone, as graphics can improve clarity while maintaining accuracy and context (Lobban et al., 2022). Additionally, integrating visual elements into plain-language communication can further improve patient engagement and interpretation of clinical information, particularly when aligned with health literacy principles (Barmpas, 2026). This suggests that combining PLS with graphical representations could be a promising approach to improving patient communication.

Recent developments in artificial intelligence (AI), particularly large language models (LLMs), have revolutionized healthcare, including the field of scientific literature. LLMs have been shown to be promising in generating images related to medical fields (Koh et al., 2023).

We have previously explored the potential of LLMs in generating graphical abstracts for systematic reviews intended for experts (personal communication). In this study, our primary objective is to evaluate whether LLMs can generate images that preserve the technical accuracy and completeness of the source PLS while adhering to established visual communication principles, as assessed by clinical experts. We explicitly frame this as a preliminary technical evaluation, a necessary first step before patient-centered validation, rather than a direct measure of patient accessibility or comprehension. The primary objective was to compare the overall quality of images generated by three LLMs, ChatGPT-5.2, Google Gemini 3 Pro, and Google Notebook, using a structured preliminary assessment tool. The secondary objectives were to compare domain-specific performance

across models, including accuracy/completeness, visual usability, and hallucination rates, and to examine whether performance differences were consistent across diverse Cochrane clinical themes.

2. Methods

2.1. Study Design and Ethics

This was a cross-sectional pilot study carried out between January and April 2026. The study included generation of images for Cochrane PLS that are available in public domain, and the images were assessed by the author researchers due to which ethics approval was not required. This study is reported in adherence to the key elements of GAMER guidelines for the use of generative AI in medical research (Luo et al., 2025).

2.2. Study Procedure

One of the authors (KS) browsed Cochrane reviews listed topic-wise in the Cochrane Library (Cochrane Library, 2026). Using an online tool for generating random numbers (Random Number Generator, 2026), two numbers were randomly selected for each of the topics. Details of the included reviews from each topic can be found in the Electronic Supplementary File S1. The PLS accompanying each of these reviews were extracted in a separate Word file. No modifications were made to the PLS.

A structured prompt specifying the context and role of LLMs, content to be included, visual format requirements, style guidelines, and output specifications (Electronic Supplementary File S2) was used for obtaining response from each of the three LLMs: ChatGPT-5.2 (ChatGPT, 2026), Google Gemini 3 Pro image (Google Gemini, 2026), and Google Notebook (Notebook, 2026). We used the paid versions of the models on their web interface without specifying any temperature as we intended to mimic real-world applications. Only one image was prompted without detailing the image size or aspect ratio, and only the first image generated was considered for assessment, and the prompt was not repeated even if the image quality was poor. None of the prompts failed and none of the generated images were altered except for the removal of visual identifiers specific for Google Gemini and Notebook outputs. We prompted the LLMs between 21 January 2026 and 22 February 2026, with the same prompt being repeated for all reviews and for all LLMs. The prompt was not piloted before being used in this study.

For each PLS, three outputs were obtained that were assigned random numbers by the same author (KS) and the allocation was concealed. Two authors (GS and JG) received the images in the same order as randomized by KS (Electronic Supplementary File S3) and they were instructed not to guess the origin of the images. For the primary assessment, the two assessors evaluated each image collaboratively. For each PLVSAT item, the assessors discussed their interpretations and arrived at a consensus score through joint deliberation. When disagreements on specific items or hallucination status arose, these were resolved through discussion with a third author (KS), who served as an adjudicator. The final score for each item represented the agreed-upon consensus judgment. The visual identifiers for both Google Gemini and Notebook outputs were erased, and no other modifications were done on the metadata of the outputs. Plain-language visual summary assessment tool (PLVSAT) (Electronic Supplementary File S4) was developed by the author (KS) based on previous experience in generating tools for assessing graphical abstracts for systematic reviews and meta-analysis (personal communication), like the structured prompts. The PLVSAT was designed as a proof-of-concept framework to evaluate technical accuracy (such as whether the generated image faithfully represents the source PLS content), completeness (whether key elements are included), and theoretical visual usability (such as adherence to principles of clarity, labeling, and typography that are hypothesized to

support comprehension). It does not measure actual patient comprehension, cognitive load, reading ease, or emotional response. Two authors (GS and JG) were involved in the validation of PLVSAT by assessing the tool in six PLS that were not included in the main analysis (Electronic Supplementary File S5). PLVSAT comprises the following domains and sub-domains:

- Domain I: Accuracy and Completeness.
- Domain II: Visual Usability.
- Domain III: Hallucination (Present/Absent).

Scores for each domain were calculated as follows: Accuracy and completeness (%) = (Sum of raw scores − 6)/(30 − 6) × 100; Visual usability (%) = (Sum of raw scores − 3)/(15 − 3) × 100; and Hallucination (%) = 100 (if absent) or 0 (if present). Then the overall score was estimated as follows: [Accuracy and completeness (%) × 0.4] + [Visual usability (%) × 0.3] + [Hallucination (%) × 0.3]. Accuracy was given a higher weightage considering scientific importance.

2.3. Statistical Analysis

2.3.1. Descriptive Statistics

Descriptive statistics were computed for each LLM and each Cochrane theme. Continuous variables (overall score, accuracy percentage, visual usability percentage) were summarized using mean ± SD, median (Q1–Q3) and 95% confidence intervals (CIs). Categorical variables (hallucination presence) were summarized as frequencies and percentages.

2.3.2. Inferential Statistics

The primary analysis involved the comparison of overall scores among the three LLMs. We employed linear mixed-effects models with Kenward-Roger approximation for denominator degrees of freedom. For each continuous outcome (overall PLVSAT score, accuracy and completeness score, and visual usability score), the model included LLM as a fixed effect and PLS and Cochrane Theme as random intercepts to account for the paired and clustered data structure. To evaluate whether LLM performance varied across medical topics, we tested the LLM × Theme interaction using a mixed-effects model with LLM and Theme as fixed effects and PLS as random intercept. The interaction *p*-value was extracted from the ANOVA table. *p*-values of $\leq$0.05 were considered statistically significant.

Internal consistency of the assessment tool was evaluated using Cronbach's alpha coefficient for the Accuracy and Completeness subscale (5 items), the Visual Usability subscale (3 items), and the total scale (8 items). Cronbach's alpha values $\geq$0.70 to $\leq$0.79 were considered acceptable, $\geq$0.80 to $\leq$0.89 good, and $\geq$0.90 excellent. Intra-class correlation coefficients were obtained between the raters for inter-rater reliability.

2.3.3. Principal Component Analysis

Principal component analysis (PCA) was conducted on the eight assessment items (IA–IE and IIA–IIC) after standardization (scaling to unit variance). The correlation matrix revealed that one item (IF: Source Attribution) had near-zero variance across all ratings and so this item was excluded. The number of components retained was determined using the eigenvalue >1 criterion and scree plot inspection.

2.3.4. Sensitivity Analyses

The following sensitivity analyses were conducted to assess the robustness of the findings: leave-one-theme-out analysis—the one-way ANOVA comparing LLMs was iteratively repeated, excluding one Cochrane theme at a time, to evaluate whether any single theme disproportionately influenced overall conclusions; and alternate weighting

schemes—to test the robustness of the results to the choice of weights in the composite score, five alternative weighting schemes (Electronic Supplementary File S6) were applied. For each scheme, mean overall scores and LLM rankings were computed.

2.3.5. Bootstrap Confidence Intervals

Percentile bootstrap with 1000 resamples was used to obtain 95% CIs for pairwise mean differences between LLMs.

2.3.6. Sample Size Justification

We designed this study as an exploratory pilot investigation rather than a confirmatory hypothesis-testing trial. As such, sample size was determined based on feasibility and descriptive precision rather than formal power calculations for a specific effect size.

All the statistical analyses were performed using R version 4.5.2.

3. Results

3.1. Descriptive Statistics

A total of 37 themes were present in the Cochrane Library of which two topics were selected in each theme, and with three LLM-generated images for each topic, the total number of images generated amounted to 222. All the LLMs generated images for all the prompts without fail (Electronic Supplementary File S7). Table 1 depicts the summary of overall score and domain-specific scores for all the LLMs. Hallucination was observed only with ChatGPT-generated images (n = 22, 29.73%).

Table 1. Performance evaluation metrics across AI models.

Domain	AI Model	n	Mean (SD)	Median (Q1, Q3)	95% CI [Lower, Upper]
Overall	ChatGPT	74	60.66 (19.31)	67.50 (39.38, 75.00)	[56.27, 65.06]
	Gemini	74	97.50 (2.59)	98.33 (98.33, 98.33)	[96.91, 98.09]
	Notebook	74	81.40 (8.81)	80.00 (76.67, 86.46)	[79.39, 83.40]
Accuracy	ChatGPT	74	61.71 (15.03)	62.50 (50.00, 75.00)	[58.29, 65.14]
	Gemini	74	95.44 (3.46)	95.83 (95.83, 95.83)	[94.65, 96.23]
	Notebook	74	73.42 (14.56)	72.92 (63.54, 86.46)	[70.11, 76.74]
Usability	ChatGPT	74	49.66 (11.90)	50.00 (41.67, 50.00)	[46.95, 52.37]
	Gemini	74	97.75 (6.37)	100.00 (100.00, 100.00)	[96.30, 99.20]
	Notebook	74	73.42 (15.59)	75.00 (66.67, 83.33)	[69.87, 76.97]

SD = standard deviation; Q1 = 25th percentile; Q3 = 75th percentile; CI = confidence interval for mean.

3.2. Inter-LLM Comparison

The linear mixed-effects model revealed significant differences in overall PLVSAT scores across the three LLMs (F = 180.25, $p < 0.001$) (Table 2). However, the interaction between LLM and Cochrane Theme was not statistically significant (F = 1.19, $p = 0.227$), indicating that LLM performance differences were consistent across diverse medical topics. No significant main effect of Theme was observed (F = 0.93, $p = 0.580$).

Table 2. Type III tests of fixed effects.

Effect	F-Value	Num DF	Den DF	*p*-Value
LLM	180.25	2	77.31	<0.001 *
Theme	0.93	36	34.88	0.580
LLM × Theme	1.19	72	77.31	0.227

Num DF = numerator degrees of freedom; Den DF = denominator degrees of freedom. * Statistical significance.

Estimated marginal means for overall scores were: ChatGPT 60.66 (95% CI: 57.85–63.49), Gemini 97.50 (95% CI: 94.67–100.33), and Notebook 81.40 (95% CI: 78.57–84.22) (Table 3). Pairwise comparisons showed that Gemini significantly outperformed ChatGPT (mean difference: −36.84, 95% CI: −40.83 to −32.84, $p < 0.001$) and Notebook (mean difference: 16.10, 95% CI: 12.11–20.10, $p < 0.001$); Notebook also significantly outperformed ChatGPT (mean difference: 20.73, 95% CI: 16.74–24.73, $p < 0.001$). Theme-specific scores of LLM-generated images are summarized in the Electronic Supplementary File S8. Intra-thematic variability was maximum with ChatGPT, while Gemini had the least.

Table 3. Pairwise comparisons across performance outcomes.

Outcome	Comparison	Estimate	SE	df	*p*-Value
Overall Score	ChatGPT vs. Gemini	−36.84	2.03	219.00	<0.001 *
	ChatGPT vs. Notebook	−20.73	2.03	219.00	
	Gemini vs. Notebook	16.10	2.03	219.00	
Accuracy Score	ChatGPT vs. Gemini	−33.73	1.91	148.09	
	ChatGPT vs. Notebook	−11.71	1.91	148.09	
	Gemini vs. Notebook	22.02	1.91	148.09	
Visual Usability Score	ChatGPT vs. Gemini	−48.09	1.96	219.00	
	ChatGPT vs. Notebook	−23.76	1.96	219.00	
	Gemini vs. Notebook	24.32	1.96	219.00	

SE = standard error; df = degrees of freedom. * Statistical significance.

3.3. Internal Consistency Assessment

The internal consistency assessment indicates an acceptable level of author consensus, with the Visual Usability subscale achieving an "Excellent" Cronbach's alpha of 0.944. The Accuracy and Completeness domain showed an "Acceptable" alpha of 0.744, with the overall total score demonstrating "Good" reliability with a coefficient of 0.879.

3.4. Principal Component Analysis

The PCA reveals that the first component (IA), population/problem accuracy, accounts for a dominant 59.3% of the total variance in the dataset. Combined with IB (10.8%) and IC (9.6%), the first three components capture nearly 80% of the cumulative variance (79.7%), indicating a significant reduction in dimensionality is possible (Figure 1). Total variance is fully accounted for across eight components, with the remaining PCs contributing progressively smaller marginal gains.

3.5. Sensitivity Analyses

No significant differences were observed in the overall scores of LLMs with the exclusion of scores from each theme (Figure 2). Similarly, all the alternate weighting schemes revealed similar overall scores for all the LLMs (Figure 3) and consequently their ranking (Figure 4).

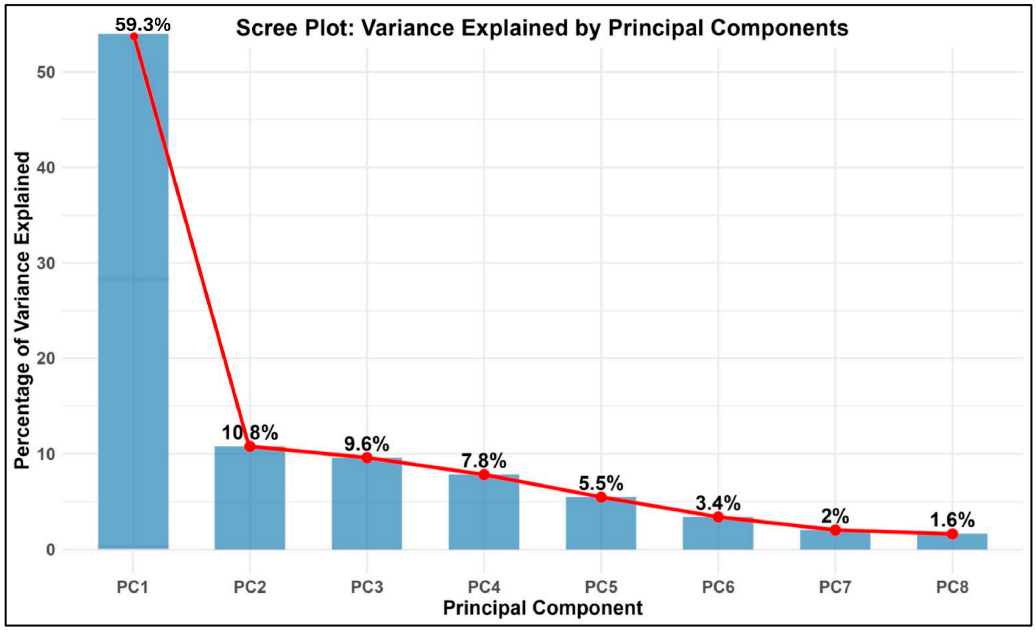

Figure 1. Variance explained by principal components. This figure displays a scree plot showing the percentage of total variance explained by each of the first eight principal components extracted from the dataset. The horizontal axis lists the principal components in descending order of importance, labeled from PC1 through PC8, while the vertical axis represents the percentage of variance accounted for by each component.

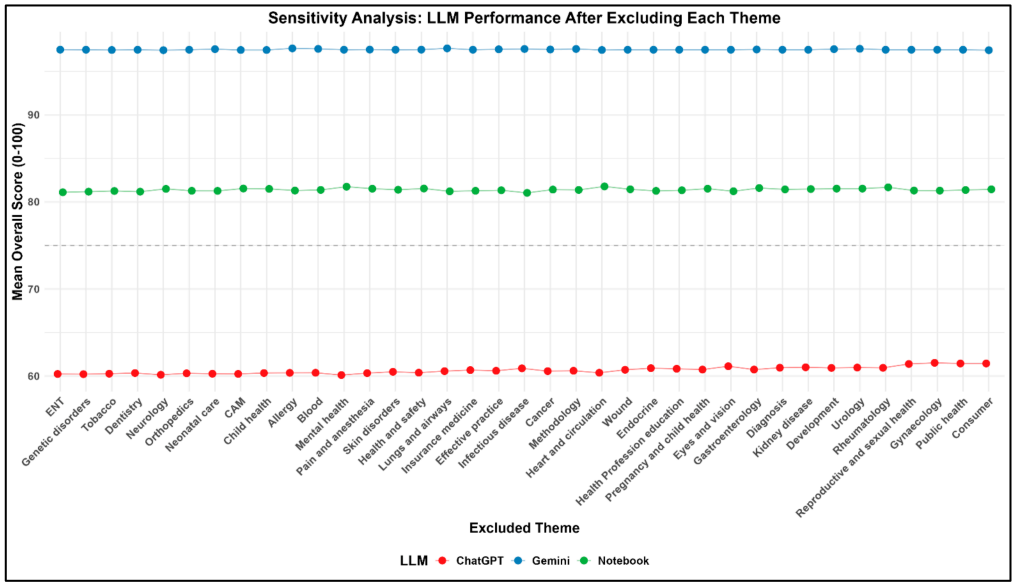

Figure 2. Plot following leave-one-theme-out sensitivity analysis. This figure presents a comparison of mean overall image quality scores (measured on a scale from 0 to 100) across the range of Cochrane review topics, organized into three distinct rows. The changes in overall LLM scores with exclusion from each topic are depicted in red lines (for ChatGPT), green (for Notebook) and blue (for Gemini).

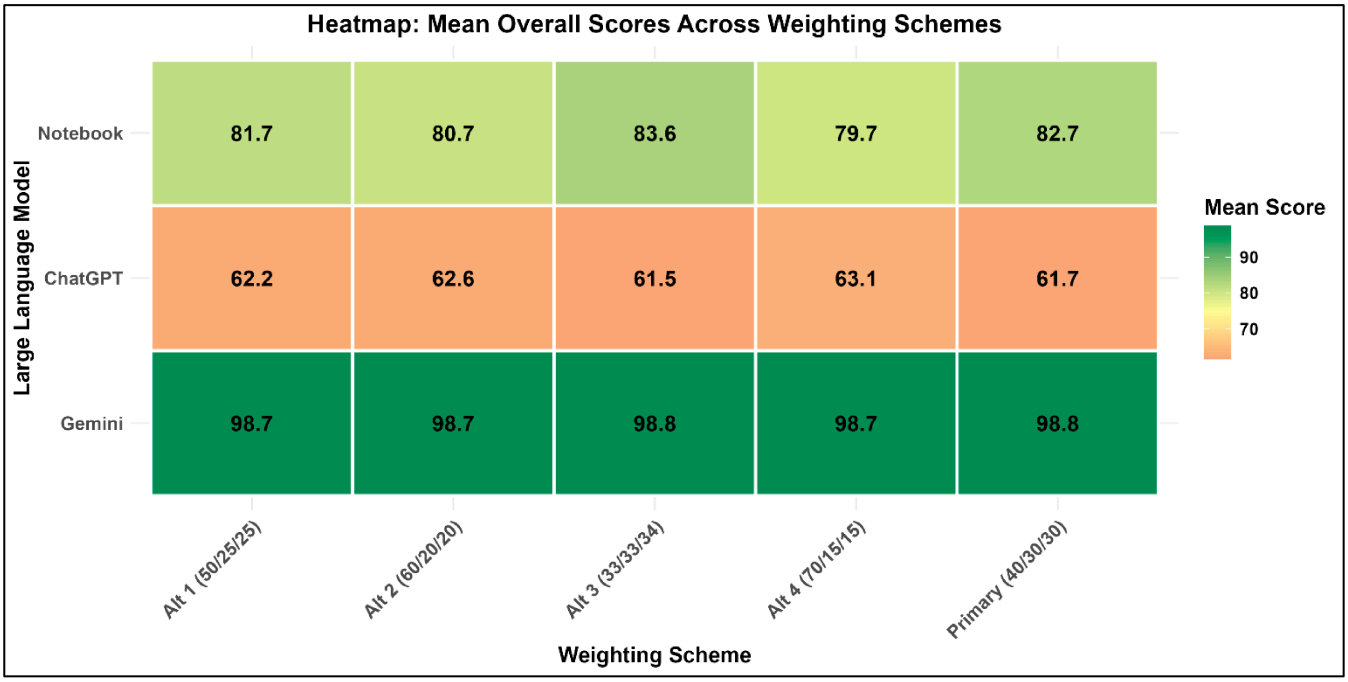

Figure 3. Sensitivity analysis for alternate weighting schemes. This figure presents a heatmap of mean overall image quality scores (measured on a scale from 0 to 100) across alternate weighting schemes. The weighting schemes are arranged horizontally and the LLMs vertically.

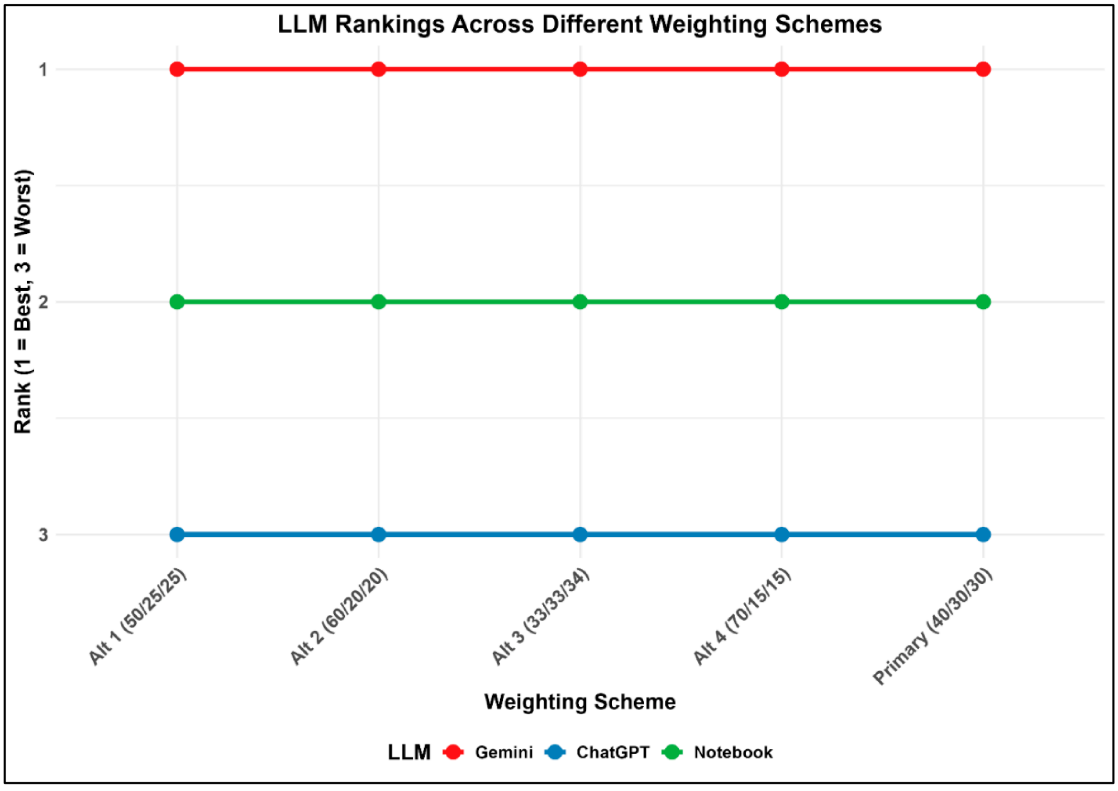

Figure 4. LLM rankings with alternate weighting schemes. This figure presents a comparison of ranking of LLMs (measured in a scale from 1 to 3) across alternate weighting schemes. The weighting schemes are arranged horizontally and the rank of LLMs in each scheme vertically.

4. Discussion

4.1. Statement of Key Findings

In this single-prompt expert-rated pilot exploratory study comparing three LLMs on their ability to generate accurate, complete, and visually usable images from Cochrane PLS, several key findings emerged. First, Google Gemini 3 Pro consistently outperformed both ChatGPT and Notebook across all the domains, achieving the highest overall scores with minimal variability and no hallucination. However, we emphasize that these are preliminary findings reflecting expert assessment of technical fidelity and theoretical usability, not demonstrated patient comprehension or accessibility. The performance differences among LLMs were consistent across all the Cochrane themes, with no significant interaction between model type and clinical topic. Collectively, these results suggest that LLMs can generate visual summaries from plain-language health content, with Gemini demonstrating the most reliable and high-quality performance in this context.

4.2. Comparison with Existing Literature

The observed performance differences between LLMs likely reflect underlying fundamental architectural distinctions among the models (Islam & Ahmed, 2024). Gemini's advanced multimodal framework and ability to integrate well-defined prompts may explain its superior accuracy and completeness. Notebook's intermediate performance can be attributed to its source-grounded, retrieval-augmented generation (RAG) approach, which restricts outputs to provided texts and likely accounts for the absence of hallucinations (Albrecht-Crane, 2025). In contrast, ChatGPT's higher hallucination rate raises important safety concerns for clinical use, as even minor inaccuracies in patient-directed materials may lead to misunderstanding, poor decision making, or diminished trust in health information (McCarthy et al., 2012). The significant intra-thematic variability observed with ChatGPT further undermines its reliability in real-world applications where consistency is critical (Patel & Lam, 2023). Notably, no significant interaction was found between LLM type and Cochrane theme. This finding potentially suggests that performance differences are attributable to the models themselves rather than to topic complexity or theme-specific variation.

The results align with previous studies evaluating LLM-generated medical communication tools, which have reported strong performance in language fluency and summarization but persistent concerns regarding factual accuracy and hallucination risk (Guo et al., 2026). Prior work has similarly emphasized the need for careful validation and human oversight when implementing LLM-generated patient education materials (McCarthy et al., 2012). Moreover, studies on graphical abstracts and visual summaries for systematic reviews have shown that visual formats can improve accessibility, engagement, and comprehension for non-medical audiences (Ibrahim et al., 2017; Buljan et al., 2018). However, most existing literature has focused on text-based LLM outputs; the present study extends this evidence by specifically evaluating image generation from plain-language summaries, a relatively underexplored area in AI-assisted patient education. Our findings carry substantial relevance for scholarly publishing and the broader ecosystem of evidence dissemination. PLS are increasingly recognized as essential tools for making research accessible to non-specialist audiences, yet they remain predominantly text-based (Dormer et al., 2022; Lobban et al., 2022). The integration of LLM-generated visual summaries as a complementary publication format could significantly enhance comprehension, engagement, and recall, particularly for individuals with low health literacy (Houts et al., 2006; Katz et al., 2006; Delp & Jones, 1996).

Critically, the design of the prompt itself emerges as a key modifiable determinant of output quality; our use of a structured, context-rich prompt that explicitly delineated the

PLS content, visual format, style guidelines, and output specifications likely contributed to the performance observed, and we contend that publishers should require authors to document and share their prompts as part of the submission metadata to enable reproducibility and quality assessment (Kendall, 2025). The dominance of population/problem accuracy (Item IA) in PCA reveals that LLM performance on this task is fundamentally gated by the model's ability to correctly identify and visually instantiate the core clinical problem. This finding has direct practical implications for how researchers and developers should approach prompt engineering that should prioritize explicit specifications of the population and condition. We propose the following prompt engineering strategies to specifically target this dominant variance source: First, explicit problem articulation: Prompts should include a dedicated, standalone sentence that clearly states the population, condition, and key comparison, ideally at the beginning of the prompt and repeated in simplified form within the visual specifications. For example, rather than embedding population details within descriptive text, prompts should include a distinct field such as "POPULATION: [description]; CONDITION: [description]; INTERVENTION: [description]; COMPARATOR: [description]." Second, instructional emphasis on population accuracy: Prompts should explicitly instruct the model to "ensure that visual representation clearly and unambiguously identifies the target population and clinical condition as the central element of the image," with specific guidance on labeling, icon selection, and visual hierarchy to prioritize these elements. Third, iterative refinement with targeted verification: Given that errors in population/problem identification are largely categorical (such as the model either correctly identifies the population or does not), prompt refinement could incorporate a verification step where the model is asked to "confirm that the image clearly depicts [specific population] with [specific condition]" before finalizing output. Fourth, model-specific optimization: Fifth, structured output templates: Developers could consider requiring models to first generate a textual "verification summary" identifying the population, problem, and key findings before generating the image. We emphasize that these strategies are derived from our empirical observations and theoretical reasoning, not from a controlled prompt-engineering experiment.

Publishers must address privacy and confidentiality concerns by ensuring that no patient-identifiable or proprietary information is inadvertently exposed. Equally important is the question of author notification and consent before publishing LLM-generated PLS-related visuals. Furthermore, the risk of unintentional visual plagiarism must be proactively managed, as LLMs trained on large corpora of published figures may inadvertently reproduce copyrighted or previously published graphical elements without proper attribution (Murray, 2023). To mitigate this, publishers could run generated images through specialized similarity-detection tools adapted for visual content, or they could mandate that generation prompts explicitly instruct the model to produce original compositions rather than replicating existing styles or layouts. Human oversight is key to ensure ethical and responsible use of AI.

Regarding copyright and ownership of LLM-generated images, our review of platform terms indicates that users retain ownership of output generated by LLMs assessed in this study. However, the legal status of AI-generated content remains an evolving area; copyright protection for AI-generated works may not be available in all jurisdictions, and liability for potential infringement rests with the user. We recommend that authors and publishers take a cautious approach: (1) clearly disclose the use of AI generation in the acknowledgments or methods section, (2) ensure that prompts do not instruct models to replicate specific copyrighted works, and (3) verify that generated images do not inadvertently reproduce copyrighted elements. These precautions are consistent with emerging publisher guidelines on AI use.

A crucial caveat warrants explicit emphasis: the assessments in this study were conducted by clinical and academic researchers, not by patients or caregivers. While expert evaluation is appropriate for detecting factual inaccuracies, hallucinations, and omissions, elements that are objectively verifiable against the source PLS, it is insufficient for gauging whether a layperson can understand, recall, or act upon the visual information. Research on health communication has consistently shown that expert judgments of material clarity and patient comprehension frequently diverge. For instance, materials deemed 'plain language' by experts may still exceed the reading comprehension levels of patients with limited health literacy (Dowse, 2004). Similarly, visual elements that appear clear to a trained professional may introduce unintended cognitive load or cultural misinterpretation for a patient. Thus, while our findings establish that certain LLMs can generate technically accurate and theoretically well-designed visuals, they do not establish that these visuals are genuinely accessible to patients. We therefore position this work as an exploratory prerequisite technical screen, one that must be followed by rigorous patient-centered validation before any clinical or educational implementation. Additionally, recent studies have pointed out communication and trust issues, along with privacy concerns amongst patients with the use of AI in healthcare (Esmaeilzadeh et al., 2021).

4.3. Strengths, Limitations and Future Directions

This study has several strengths, including the use of a structured and validated assessment tool, a randomized and concealed image allocation process, and multiple sensitivity analyses that confirmed the robustness of the findings across different scoring assumptions and clinical themes. Nevertheless, several limitations must be acknowledged. First, the assessments were conducted by author researchers rather than by independent researchers or patients or caregivers. Second, the study evaluated only three LLMs using a single standardized prompt structure, and the results may differ with alternative phrasing, iterative prompting, or newer model versions. Third, the cross-sectional design captured performance at a single time point, whereas LLMs are continuously updated, potentially altering their output characteristics over time. Fourth, the PLVSAT, while internally consistent, has not undergone external, independent validation against established criteria, patient comprehension measures, or other gold-standard instruments. It should therefore be considered a preliminary research tool requiring further psychometric development and validation before it can be recommended for routine use. The tool must undergo external expert validation and input from patient/caregiver must be considered in the full validation of the tool. Fifth, although we ensured that visually, the images did not have any identifiers, we did not ensure the success of blinding by asking the reviewers to identify the source of images. Sixth, because our assessment procedure employed consensus scoring rather than independent ratings, traditional inter-rater reliability statistics could not be calculated. However, for the purposes of this exploratory pilot study, we believe the consensus-based approach, which leverages the complementary expertise of two assessors and resolves disagreements through third-party adjudication, represents a methodologically defensible balance between accuracy and feasibility. Seventh, our hallucination assessment was binary (present/absent) rather than graded by clinical severity. We did not classify hallucinations by their type (such as factual error, causality implication, omission, overstatement), clinical context, or potential for patient harm. We recommend that future research move beyond binary hallucination assessment to develop and validate a graded clinical risk assessment framework that can distinguish between clinically inconsequential and potentially harmful errors. Eighth, we acknowledge not assessing the accessibility of the generated images such as color contrast, font size, suitability for color-blind users, screen-reader alternatives, and cognitive load. Another limitation is specifically related to the communication of uncer-

tainty, where our study did not conduct a dedicated analysis of whether generated images preserve evidence certainty, uncertainty, and limitations. Future research should develop and validate specific metrics for evaluating uncertainty communication in AI-generated patient visuals. Finally, the study did not assess whether the generated images improved patient understanding or clinical outcomes.

For clinicians, these findings suggest that while LLM-generated visual summaries are technically feasible, model selection is critical. For researchers, future work should focus on developing standardized benchmarks for AI-generated patient visuals, conducting prospective trials comparing text-only versus text-plus-image PLS on patient comprehension and behavior, assessing whether potential differences exist amongst various contents of Cochrane reviews and exploring prompt engineering strategies to reduce hallucinations and improve visual clarity. Future work should also involve evaluating the images amongst patients with varied health literacy levels. Moving forward, interdisciplinary collaboration among AI developers, health literacy experts, clinicians, and patient advocates will be essential to translate these technological capabilities into equitable, trustworthy, and impactful patient communication tools.

5. Conclusions

This single-prompt exploratory pilot expert-rated study found substantial differences among three LLMs in their ability to generate visual summaries from Cochrane PLS. However, the results require confirmation using independent assessors, stronger validation of the assessment tool, statistical models that account for the paired design, multiple prompt strategies, and direct testing with patients and caregivers. For now, LLM-generated images should complement, not replace, clinician guidance, and model selection must prioritize reliability and safety. The findings reported here represent an initial technical validation, a necessary but insufficient step, and should not be interpreted as evidence that these images improve patient comprehension, recall, or clinical outcomes.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/publications14030059/s1. Electronic Supplementary File S1: Themes, review IDs and review titles included in this study; Electronic Supplementary File S2: Prompt template; Electronic Supplementary File S3: Randomization numbers of the generated images; Electronic Supplementary File S4: PLVSAT; Electronic Supplementary File S5: Details of reviews used in validation cohort; Electronic Supplementary File S6: Details of alternate weighting schemes; Electronic Supplementary File S7: Generated images; Electronic Supplementary File S8: Thematic scores of LLMs.

Author Contributions: K.S.: Conceptualization, K.S., J.G. and G.S.; methodology, K.S. and G.S.; software, K.S., J.G., J.R.J. and G.S.; validation, K.S. and G.S.; formal analysis, K.S., J.G. and G.S.; investigation, K.S., J.G., J.R.J. and G.S.; resources, K.S.; data curation, K.S. and J.R.J.; writing—original draft preparation, K.S., J.G., J.R.J. and G.S.; writing—review and editing, K.S., J.G., J.R.J. and G.S.; visualization, K.S.; supervision, K.S.; project administration, K.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: Declaration on the use of Generative Artificial Intelligence: The authors wish to acknowledge the use of DeepSeek V3.2 for refining the language clarity, flow and grammatical structure of this manuscript. This application was conducted in adherence to established ethical guidelines regarding the use and disclosure of Generative AI in scholarly research. Each author has provided

substantial intellectual contributions to the work, which has been rigorously vetted for accuracy, and the authors assume full responsibility for the integrity and final content of the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

Albrecht-Crane, C. (2025). Thinking smarter, not harder? Google NotebookLM's misalignment problem in education. In *Proceedings of the 43rd ACM international conference on design of communication* (pp. 121–127). Association for Computing Machinery.

Badarudeen, S., & Sabharwal, S. (2010). Assessing readability of patient education materials: Current role in orthopaedics. *Clinical Orthopaedics and Related Research*, *468*(10), 2572–2580. [CrossRef] [PubMed]

Barmpas, G. (2026). *Plain language summarization of clinical trials* [Doctoral dissertation, Aristotle University of Thessaloniki]. Available online: https://ikee.lib.auth.gr/record/359574/files/GRI-2024-46341.pdf (accessed on 20 May 2026).

Berkman, N. D., Sheridan, S. L., Donahue, K. E., Halpern, D. J., & Crotty, K. (2011). Low health literacy and health outcomes: An updated systematic review. *Annals of Internal Medicine*, *155*(2), 97–107. [CrossRef] [PubMed]

Buljan, I., Malički, M., Wager, E., Puljak, L., Hren, D., Kellie, F., West, H., Alfirević, Ž., & Marušić, A. (2018). No difference in knowledge obtained from infographic or plain language summary of a Cochrane systematic review: Three randomized controlled trials. *Journal of Clinical Epidemiology*, *97*, 86–94. [CrossRef] [PubMed]

ChatGPT. (2026). Available online: https://chatgpt.com/ (accessed on 27 January 2026).

Cochrane Library. (2026). Available online: https://www.cochranelibrary.com/cdsr/reviews/topics (accessed on 26 April 2026).

Delp, C., & Jones, J. (1996). Communicating information to patients: The use of cartoon illustrations to improve comprehension of instructions. *Academic Emergency Medicine: Official Journal of the Society for Academic Emergency Medicine*, *3*(3), 264–270. [CrossRef] [PubMed]

Dormer, L., Schindler, T., Williams, L. A., Lobban, D., Khawaja, S., Hunn, A., Ubilla, D. L., Sargeant, I., & Hamoir, A. M. (2022). A practical 'How-To' guide to plain language summaries (PLS) of peer-reviewed scientific publications: Results of a multi-stakeholder initiative utilizing co-creation methodology. *Research Involvement and Engagement*, *8*(1), 23. [CrossRef] [PubMed]

Dowse, R. (2004). Using visuals to communicate medicine information to patients with low literacy. *Adult Learning*, *15*(1–2), 22–25. [CrossRef]

Esmaeilzadeh, P., Mirzaei, T., & Dharanikota, S. (2021). Patients' perceptions toward human-artificial intelligence interaction in health care: Experimental study. *Journal of Medical Internet Research*, *23*(11), e25856. [CrossRef] [PubMed]

Google Gemini. (2026). Available online: https://gemini.google.com/app (accessed on 28 January 2026).

Guo, Y., Sohn, J. H., Leroy, G., & Cohen, T. (2026). Are LLM-generated plain language summaries truly understandable? A large-scale crowdsourced evaluation. *Journal of Biomedical Informatics*, *179*, 105038. [CrossRef] [PubMed]

Houts, P. S., Doak, C. C., Doak, L. G., & Loscalzo, M. J. (2006). The role of pictures in improving health communication: A review of research on attention, comprehension, recall, and adherence. *Patient Education and Counseling*, *61*(2), 173–190. (Erratum in 2006, *Patient Education and Counseling*, *64*(1–3), 393–394). [CrossRef] [PubMed]

Ibrahim, A. M., Lillemoe, K. D., Klingensmith, M. E., & Dimick, J. B. (2017). Visual abstracts to disseminate research on social media: A prospective, case-control crossover study. *Annals of Surgery*, *266*(6), e46–e48. [CrossRef] [PubMed]

Islam, R., & Ahmed, I. (2024). Gemini-the most powerful LLM: Myth or truth. In *2024 5th Information Communication Technologies Conference (ICTC)* (pp. 303–308). IEEE.

Johnson, A., & Sandford, J. (2005). Written and verbal information versus verbal information only for patients being discharged from acute hospital settings to home: Systematic review. *Health Education Research*, *20*(4), 423–429. [CrossRef] [PubMed]

Kanellopoulou, C., Kermanidis, K. L., & Giannakoulopoulos, A. (2019). The dual-coding and multimedia learning theories: Film subtitles as a vocabulary teaching tool. *Education Sciences*, *9*(3), 210. [CrossRef]

Katz, M. G., Kripalani, S., & Weiss, B. D. (2006). Use of pictorial aids in medication instructions: A review of the literature. *American Journal of Health-System Pharmacy: AJHP: Official Journal of the American Society of Health-System Pharmacists*, *63*(23), 2391–2397. [CrossRef] [PubMed]

Kendall, G. (2025). When using artificial intelligence tools in scientific publications authors should include the prompts and the generated text as part of the submission. *Journal of Academic Ethics*, *3*(3), 639–647. [CrossRef]

Kessels, R. P. (2003). Patients' memory for medical information. *Journal of the Royal Society of Medicine*, *96*(5), 219–222. [CrossRef] [PubMed]

Koh, J. Y., Fried, D., & Salakhutdinov, R. R. (2023). Generating images with multimodal language models. *Advances in Neural Information Processing Systems*, *36*, 21487–21506. [CrossRef]

Kripalani, S., Bengtzen, R., Henderson, L. E., & Jacobson, T. A. (2008). Clinical research in low-literacy populations: Using teach-back to assess comprehension of informed consent and privacy information. *IRB*, *30*(2), 13–19. [PubMed]

Lobban, D., Gardner, J., Matheis, R., & ISMPP PLS Perspectives Working Group. (2022). Plain language summaries of publications of company-sponsored medical research: What key questions do we need to address? *Current Medical Research and Opinion*, *38*(2), 189–200. [CrossRef] [PubMed]

Luo, X., Tham, Y. C., Giuffrè, M., Ranisch, R., Daher, M., Lam, K., Eriksen, A. V., Hsu, C. W., Ozaki, A., Moraes, F. Y., Khanna, S., Su, K. P., Begagić, E., Bian, Z., Chen, Y., Estill, J., & GAMER Working Group. (2025). Reporting guideline for the use of generative artificial intelligence tools in MEdical research: The GAMER statement. *BMJ Evidence-Based Medicine*, *30*(6), 390–400. [CrossRef] [PubMed]

McCarthy, D. M., Waite, K. R., Curtis, L. M., Engel, K. G., Baker, D. W., & Wolf, M. S. (2012). What did the doctor say? Health literacy and recall of medical instructions. *Medical Care*, *50*(4), 277–282. [CrossRef] [PubMed]

Murray, M. D. (2023). Generative AI art: Copyright infringement and fair use. *SMU Science and Technology Law Review*, *26*, 259. [CrossRef]

Notebook. (2026). Available online: https://notebooklm.google.com/ (accessed on 15 February 2026).

Patel, S. B., & Lam, K. (2023). ChatGPT: The future of discharge summaries? *The Lancet Digital Health*, *5*(3), e107–e108. [CrossRef] [PubMed]

Random Number Generator. (2026). Available online: https://www.calculatorsoup.com/calculators/statistics/random-number-generator.php (accessed on 22 February 2026).

Schillinger, D., Piette, J., Grumbach, K., Wang, F., Wilson, C., Daher, C., Leong-Grotz, K., Castro, C., & Bindman, A. B. (2003). Closing the loop: Physician communication with diabetic patients who have low health literacy. *Archives of Internal Medicine*, *163*(1), 83–90. [CrossRef] [PubMed]