

Evaluating AI Chatbots in Addressing the Unmet Survivorship Needs of Adolescents and Young Adults with Melanoma

Jordan Lily Jafarnia, Priscilla Lynne Haff, Reece Phillip Moore, Alyssa Leigh Osheim, Katherine McKenna Riley, Sabrina Zheng, Michael Roth, Madeleine Hines Salge, Kelly Carter Nelson

Submitted to: JMIR Cancer
on: September 21, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript.....	5
Supplementary Files.....	22
Figures	23
Figure 1.....	24
Figure 2.....	25
Figure 3.....	26
Figure 4.....	27
Figure 5.....	28
Figure 6.....	29
Multimedia Appendixes	30
Multimedia Appendix 1.....	31
Multimedia Appendix 2.....	31
Multimedia Appendix 3.....	31
Multimedia Appendix 4.....	31
Multimedia Appendix 5.....	31
Multimedia Appendix 6.....	31
Multimedia Appendix 7.....	31

Evaluating AI Chatbots in Addressing the Unmet Survivorship Needs of Adolescents and Young Adults with Melanoma

Jordan Lily Jafarnia^{1*} BA; Priscilla Lynne Haff^{1*} BS; Reece Phillip Moore¹ BS; Alyssa Leigh Osheim¹ BBA; Katherine McKenna Riley¹ BA; Sabrina Zheng² BS; Michael Roth³ MD; Madeleine Hines Salge⁴ MD; Kelly Carter Nelson⁴ MD

¹ UTHealth Houston, McGovern Medical School Houston US

² Paul L. Foster School of Medicine at TTUHSC El Paso El Paso US

³ Division of Pediatrics University of Texas MD Anderson Cancer Center Houston US

⁴ Department of Dermatology, Division of Internal Medicine The University of Texas MD Anderson Cancer Center Houston US

*these authors contributed equally

Corresponding Author:

Kelly Carter Nelson MD

Department of Dermatology, Division of Internal Medicine

The University of Texas MD Anderson Cancer Center

1515 Holcombe Blvd

Houston

US

Abstract

Background: Melanoma, a highly aggressive form of skin cancer, is the second most common type of cancer for adolescent and young adult (AYA, ages 15-39 years) patients. AYA melanoma patients may turn to internet sources, especially AI-chatbots, to manage uncertainty about prognosis and treatment.

Objective: To evaluate the quality, empathy, and readability of responses generated by leading AI chatbots when addressing the top unmet needs of AYA melanoma patients receiving treatment.

Methods: Our research team recently surveyed 152 AYA melanoma patients using the Needs Assessment Service Bridge (NASB), a validated instrument that assesses psychosocial needs for AYA cancer patients. The survey identified the top 5 needs for advanced AYA melanoma patients receiving treatment. Each need was reframed into a question and brief clinical history, then entered into each chatbot by five individuals who cleared their pre- and post-question history. Chatbot responses were evaluated to assess information quality (Global Quality Score (GQS) and DISCERN), accessibility and readability (GQS, Flesch Kincaid Grade Level, Flesch Reading Ease), and perceived empathy (Perceived Empathy of Technology (PETS)).

Results: Across 75 chatbot responses, ChatGPT achieved the highest average quality (mean GQS 4.4, mean DISC 3.2) and empathy (PETS-ER 5.4, PETS-UT 6.4), though with greater variability (SD~1.8). Copilot produced the lowest quality and empathy scores, while Gemini responses were consistently midrange. PETS-UT exceeded PETS-ER across all models, suggesting stronger cognitive empathy than emotional responsiveness. Readability analysis showed outputs exceeded the average U.S. reading level (mean FKGL 11.8, FRE 38.6), limiting accessibility. The most readable responses were found in Question 2, which also scored higher in quality and empathy, whereas Question 4 and 5 produced the most complex, difficult to read responses corresponding with lower quality and empathy ratings.

Conclusions: AI chatbots can provide moderately accurate and supportive responses to AYA melanoma patient needs, but outputs are inconsistent, written above the recommended reading level for health information, and limited in empathy. Question framing strongly influenced chatbot performance, with more emotional prompts drawing greater empathy, and readability aligning with both quality and empathy. Chatbot use in this population should remain adjunctive, with further research needed to standardize quality, improve readability, and enhance empathetic communication.

(JMIR Preprints 21/09/2025:84234)

DOI: <https://doi.org/10.2196/preprints.84234>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <http://www.jmir.org>

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <https://www.jmir.org>

Original Manuscript

Title: Evaluating AI Chatbots in Addressing the Unmet Survivorship Needs of Adolescents and Young Adults With Melanoma

Authors: Jordan L. Jafarnia^{1*}, BS; Priscilla L. Haff^{1*}, BS; Reece Moore¹, Alyssa Osheim¹, Katherine Riley¹, Sabrina Zheng², Michael Roth³, MD; Madeleine Hines Salge⁴, MD; Kelly C. Nelson⁴, MD

* JLJ and PLH are co-first authors and contributed equally to this work

Affiliations:

1. McGovern Medical School, The University of Texas Health Science Center at Houston, 6431 Fannin St, Houston, Tx 77030, United States
2. Texas Tech University Health Sciences Center El Paso, 130 Rick Francis Street El Paso, TX 79905, United States
3. Division of Pediatrics, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Blvd, Tx, 77030, United States
4. Department of Dermatology, Division of Internal Medicine, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Blvd, Tx, 77030, United States

Word Count: 3,293, 4 tables, & 6 figures

Keywords: Melanoma, Adolescent, Young Adult, Artificial Intelligence, Natural Language Processing, Empathy, Readability

Abstract

Background: Melanoma, a highly aggressive form of skin cancer, is the second most common type of cancer for adolescent and young adult (AYA, ages 15-39 years) patients. AYA melanoma patients may turn to internet sources, especially AI-chatbots, to manage uncertainty about prognosis and treatment.

Objective: To evaluate the quality, empathy, and readability of responses generated by leading AI chatbots when addressing the top unmet needs of AYA melanoma patients receiving treatment.

Methods: Our research team recently surveyed 152 AYA melanoma patients using the Needs Assessment Service Bridge (NA-SB), a validated instrument that assesses psychosocial needs for AYA cancer patients. The survey identified the top 5 needs for advanced AYA melanoma patients receiving treatment. Each need was reframed into a question and brief clinical history, then entered into each chatbot by five individuals who cleared their pre- and post-question history. Chatbot responses were evaluated to assess information quality (Global Quality Score (GQS) and DISCERN), accessibility and readability (GQS, Flesch Kincaid Grade Level, Flesch Reading Ease), and perceived empathy (Perceived Empathy of Technology (PETS)).

Results: Across 75 chatbot responses, ChatGPT achieved the highest average quality (mean GQS 4.4, mean DISC 3.2) and empathy (PETS-ER 5.4, PETS-UT 6.4), though with greater variability (SD~1.8). Copilot produced the lowest quality and empathy scores, while Gemini responses were consistently midrange. PETS-UT exceeded PETS-ER across all models, suggesting stronger cognitive empathy than emotional responsiveness. Readability analysis showed outputs exceeded the average U.S. reading level (mean FKGL 11.8, FRE 38.6), limiting accessibility. The most readable responses were found in Question 2, which also scored higher in quality and empathy, whereas Question 4 and 5 produced the most complex, difficult to read responses corresponding with lower quality and empathy ratings.

Conclusions: AI chatbots can provide moderately accurate and supportive responses to AYA melanoma patient needs, but outputs are inconsistent, written above the recommended reading level for health information, and limited in empathy. Question framing strongly influenced chatbot performance, with more emotional prompts drawing greater empathy, and readability aligning with both quality and empathy. Chatbot use in this population should remain adjunctive, with further research needed to standardize quality, improve readability, and enhance empathetic communication.

Introduction

Melanoma, a malignant tumor of melanocytes, is an aggressive form of skin cancer due to its high metastatic potential and associated mortality. Although cutaneous melanoma is most frequently diagnosed in older adults, it is one of the most common cancers affecting adolescents and young adults (AYAs, age 15-39). Analysis of the US cancer registry data indicates that cutaneous melanoma accounts for approximately 7% of newly diagnosed cancers among AYAs[1]. A melanoma diagnosis and subsequent treatment can be highly disruptive for AYA patients, with implications for physical health, emotional well-being, and social function. The severity of the disease, coupled with the uncertainty of treatment outcomes, often introduces feelings of fear and isolation [2]. In seeking clarity in this uncertainty, patients may turn to the internet and social media for answers, education, and support. When seeking answers through internet interfaces, patients are likely to leverage artificial intelligence (AI) driven chatbots, such as Chat GPT, Microsoft Copilot, and Google Gemini [3].

AI chatbots are computer programs that use artificial intelligence to simulate human-to-human conversation mainly through text models, but also through voice. Through the use of Natural Language Processing (NLP) and machine learning, AI chatbots attempt to answer complex questions and adapt to different communication styles. These tools have been shown to provide health information in an accessible and interactive manner [4]. Their availability, objectivity, and capacity to adapt personalized responses to diverse communication styles make them appealing to patients navigating difficult diagnoses [5]. While some studies have evaluated the most frequently asked chatbot questions regarding specific forms of cancer (i.e. breast, leukemia, prostate), no published research to our knowledge has systematically examined their ability to address AYAs' specific concerns regarding melanoma treatment and survivorship [6].

A recent survey study of AYAs with melanoma identified several unmet needs for patients in different phases of cancer care [7]. This study utilized a survey based on the Needs Assessment and Service Bridge (NA-SB) questionnaire and gathered results from 152 patients, 20 of whom were currently receiving treatment. For AYA melanoma patients currently undergoing systemic treatment, the top 5 unmet needs included: (1) what happens after treatment; (2) fear of cancer recurrence; (3) long-term hormonal side effects; (4) long-term systemic side effects; and (5) genetic implications on diagnosis and treatment [7]. With these established unmet needs and the increasing use of AI chatbots for AYA patient support, we evaluated the unique response of three chatbots to patient-centered questions to determine if these resources are safe and valid for AYA melanoma patient use.

Methods

Need Selection and Question Creation

The top five most frequently reported needs for AYA melanoma patients receiving systemic treatment were identified through the NA-SB survey. Once identified, each NA-SB prompt was slightly modified into a question format with a standardized and appropriate brief clinical history (Table 1). All questions shared a common stem that identified the patient as an AYA with metastatic melanoma, with tailored follow-up question added to reflect the specific need category.

Table 1: Questions input into each Chatbot, based on the five more frequently reported needs

Questions from the perspective of patients currently on treatment:
--

1. I am an adolescent/young adult patient with metastatic melanoma and am currently receiving treatment. I want more information about what will happen when treatment finishes.
2. I am an adolescent/young adult patient with metastatic melanoma and am currently receiving treatment. In the last month I have become more worried about my cancer spreading. What do I need to know and what steps can I take to cope with this fear?
3. I am an adolescent/young adult patient with metastatic melanoma and am currently receiving treatment. How will my treatment affect long-term hormone changes?
4. I am an adolescent/young adult patient with metastatic melanoma and am currently receiving treatment. I want more information about how my genetics may or may not have impacted my diagnosis and treatment.
5. I am an adolescent/young adult patient with metastatic melanoma and am currently receiving treatment. I want more information about the long-term side effects of treatment.

Study design

This study evaluated the quality of responses generated by AI chatbots to needs frequently reported by AYA melanoma patients receiving treatment. This study relied exclusively on publicly available chatbot outputs and did not involve human subjects, so institutional research board approval was not required. Our methodologic approach included six components: (1) need selection, (2) question creation, (3) chatbot input, (4) data extraction, (5) data coding, and (6) evaluation (Figure 1).

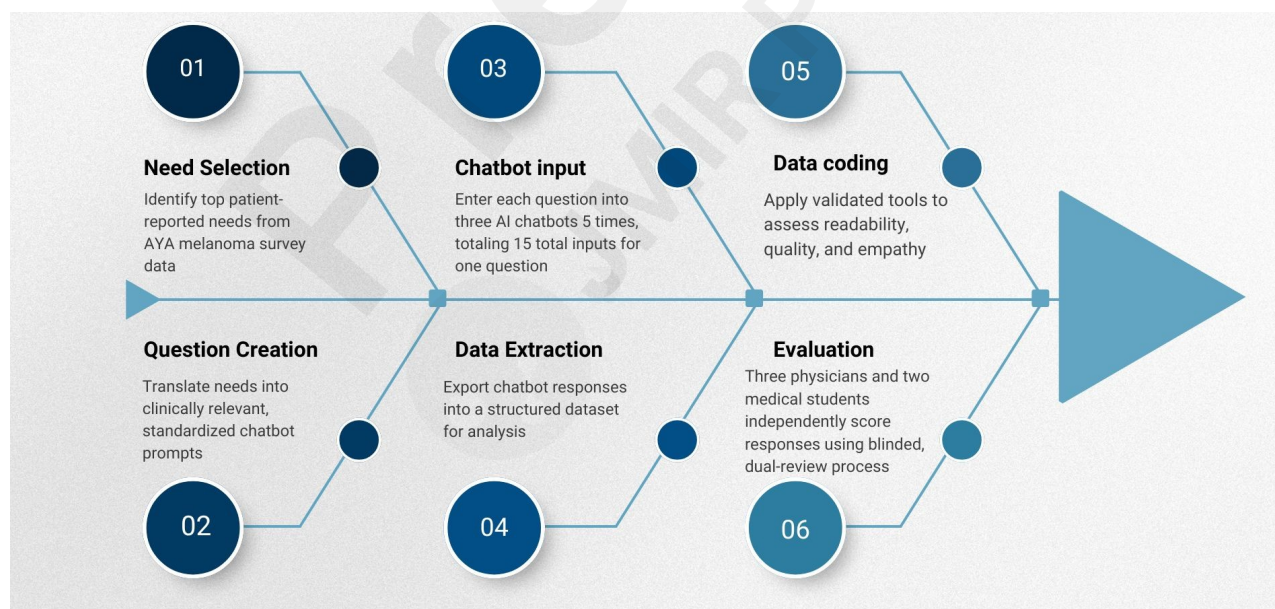


Figure 1: Study design workflow

Chatbot Input and Data Extraction

Chat GPT-4 (Open AI), Copilot (Microsoft), and Gemini (Google) were selected for evaluation based on their increasing usage among AYAs and advancements in their technology. The resulting need-

related questions, outlined in Table 1, were entered into each chatbot five times by four individuals (RM, AS, KR, SZ). To minimize bias, these team members did not participate in question selection, and AI personalization was avoided through memory clearing between each input, thus replicating a true AYA patient experience. Responses were copied and exported into an Excel spreadsheet where a separate team completed data coding and evaluation.

Data Coding and Evaluation

Responses were evaluated and coded to assess information quality, (Global Quality Scale and DISCERN), perceived empathy (Perceived Empathy of Technology), and readability (Flesch Kincaid Grade Level, Flesch Reading Ease) (Table 2). Four team members, 2 medical students (JLJ, PLH) and 2 physicians (KCN, MHS), completed evaluations independently, with results hidden until completion to minimize bias. Each evaluator had the necessary qualifications to assess their designated sections. Score interpretation for all tools is provided in table 2. Standard deviations (SD) were calculated to assess variability in quality, empathy, and readability for each chatbot and each question.

Table 2: Evaluation metrics for chatbot responses, detailing assessed domains and score interpretation.

Assessment Tool	Evaluated Domains	Scale	Score Interpretation
Global Quality Score (GQS) [8]	Accessibility Quality Flow Usefulness	5-point Likert scale	1= Poor quality, flow, and missing information 2= Fair 3= Moderate 4= Good 5= Excellent quality, flow, and information
DISCERN [9] (DISC)	Accuracy Quality	Items 1-8 (5-point Likert scale) Item 16 (summary 5-point Likert scale)	Low quality information ≤ 2 Medium quality = 2-4 High quality information ≥ 4
Perceived Empathy of Technology Scale (PETS) [10]	Emotional Responsiveness (ER), Understanding and Trust (UT)	10-items, two domains. Each item assessed with 10-point scale, and then averaged across 5 items for each domains score (ER & UT)	1-3= Low empathy 4-6= Moderate empathy 7-10= High empathy
Flesch Kincaid Grade Level [11] (FKGL)	Comprehension	Automated formula estimating U.S school grade level required to understand text	Numeric grade level (e.g. 8=8 th grade reading level)
Flesch Reading Ease [11] (FRE)	Readability Flow	Automated formula producing a score from 0-100, with higher value indicating easier reading	0-30= Very difficult to read 31-50= Difficult to read 51-60= Fairly difficult to read

			61-70= Plain English 70-100: Easy to read
--	--	--	--

Quality (GQS and DISC)

Chatbot answer quality was evaluated using the GQS and DISC scales. The GQS scale evaluates quality, accessibility, flow, and perceived usefulness of information for patients, as judged by a physician (Table S1 in MA 1). The DISCERN Scale assesses the quality of written health information, focusing on reliability, clarity, and explanations of risks and benefits (Table S2 in MA 1). Two physicians (KCN, MHS) evaluated each response with the QGS and DISC scales. The results were then averaged into a final score.

Empathy (PETS-ER/UT)

The PETS 10-item tool was utilized to evaluate user perception of empathy from chatbot responses. The PETS tool is divided into two sections: PETS-ER (6 items) and PETS-UT (4 items) (Table S3, MA 1). PETS-ER (Emotional Responsiveness) focuses on emotional engagement and support. PETS-UT (Understanding and Trust) measures “cognitive empathy” or the ability of the Chatbot to understand the user's perspective. Each item was individually scored, then averaged to give domain-specific scores for each input. Two medical student reviewers (JLJ, PLH) performed scoring and highlighted key examples for qualitative evaluation.

Readability (FKGL, FRE, Word Count)

To evaluate each output's readability, the FKGL and FRE were utilized. FKGL estimates the U.S grade level required to comprehend a response, while FRE scores readability from 0 to 100 (higher=easier to read). Scores for both instruments were calculated in Microsoft Word's spelling and grammar tool.

Results

The averaged per-question assessment scores for each chatbot response are available in Table 3, individual input scoring is available in MA 2.

Table 3: Average quality, empathy, and readability scoring per question by each chatbot. Highest scores bolded and highlighted in green.

Q	Chatbot	GQS	DISC	PETS-ER	PETS-UT	FRE	FKGL	Word Count
1	ChatGPT	4.50	3.34	6.18	7.80	39.20	12.62	538.80
	Copilot	4.00	2.75	5.70	6.45	37.32	11.82	344.80
	Gemini	4.00	3.09	5.87	6.50	46.42	11.10	779.80
2	ChatGPT	4.60	3.35	7.53	7.60	54.84	9.10	522.40
	Copilot	4.50	3.15	6.67	6.65	46.42	10.72	431.80
	Gemini	4.20	3.09	6.67	6.65	52.68	9.80	728.00
3	ChatGPT	4.00	3.00	3.70	4.18	28.58	11.98	389.20
	Copilot	4.10	2.75	4.13	4.10	31.82	12.38	315.20
	Gemini	4.50	3.05	4.13	4.10	34.38	13.12	717.80
4	ChatGPT	4.50	2.94	4.33	6.97	37.62	10.84	545.20
	Copilot	3.40	2.55	3.10	3.15	31.72	13.40	360.20
	Gemini	4.00	2.81	3.10	3.15	40.60	12.16	703.00
5	ChatGPT	4.40	3.26	5.03	5.30	29.36	12.78	415.60

	Copilot	3.70	2.74	2.20	2.70	33.38	12.74	330.80
	Gemini	4.20	2.76	2.20	2.70	34.68	12.74	713.60

Quality (GQS and DISC)

ChatGPT produced the highest information quality, scoring best on GQS and DISC for 4 out of the 5 questions (As seen in Figures 2 and 3). Copilot scored the lowest, while Gemini most often scored between the two others.

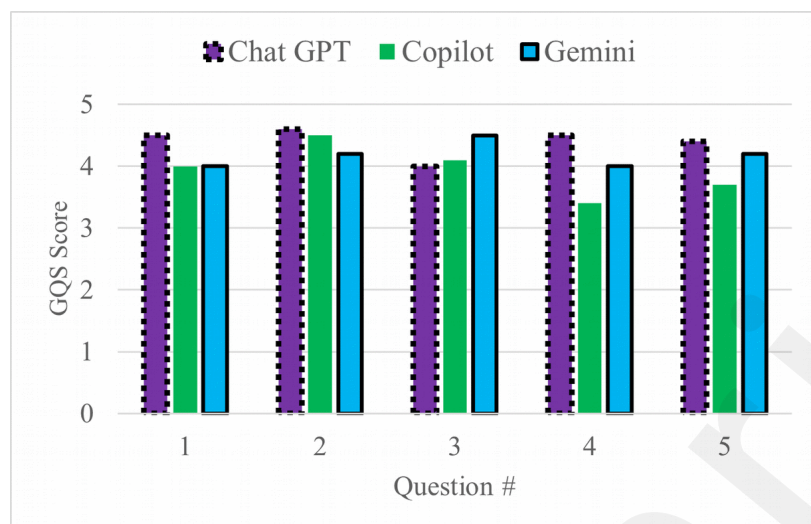


Figure 2: GQS scores comparing each chatbot by question input.

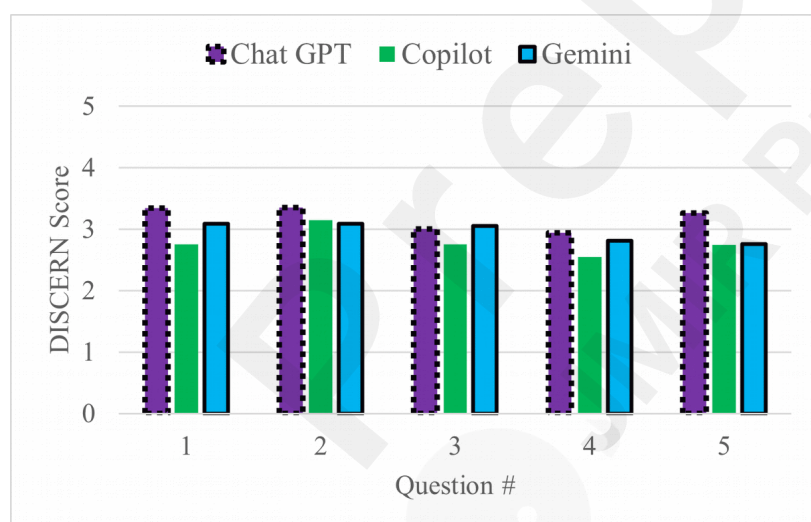


Figure 3: Discern scores comparing each chatbot by question input.

Quality variability differed across chatbots. Copilot showed the greatest spread in GQS ($SD=0.53$) and moderate variability in DISC ($SD=0.24$). ChatGPT showed moderate unpredictability in both GQS ($SD=0.32$) and DISC ($SD=0.31$), and Gemini was the most stable in both GQS ($SD=0.24$) and DISC ($SD=0.19$). At the question level, Q4 and Q5 produced the most variable quality scores, while Q2 and Q3 were relatively similar across models (See MA 3 for all SD). ChatGPT delivers the best quality based on GQS and DISC ratings, Gemini delivers reliably average quality according to GQS/DISC scoring and SD, and Copilot is the least reliable, with the lowest GQS/DISC and most variability.

Empathy (PETS-ER and PETS-UT)

PETS scores varied at the question and chatbot level (Figure 4, Table 3). Q1 and Q2 have the highest empathy (ER- 5.91, UT 6.92; ER- 6.96, UT- 6.97; respectively). Whereas Q5 demonstrates the lowest empathy scores (ER-3.14, UT-3.57).

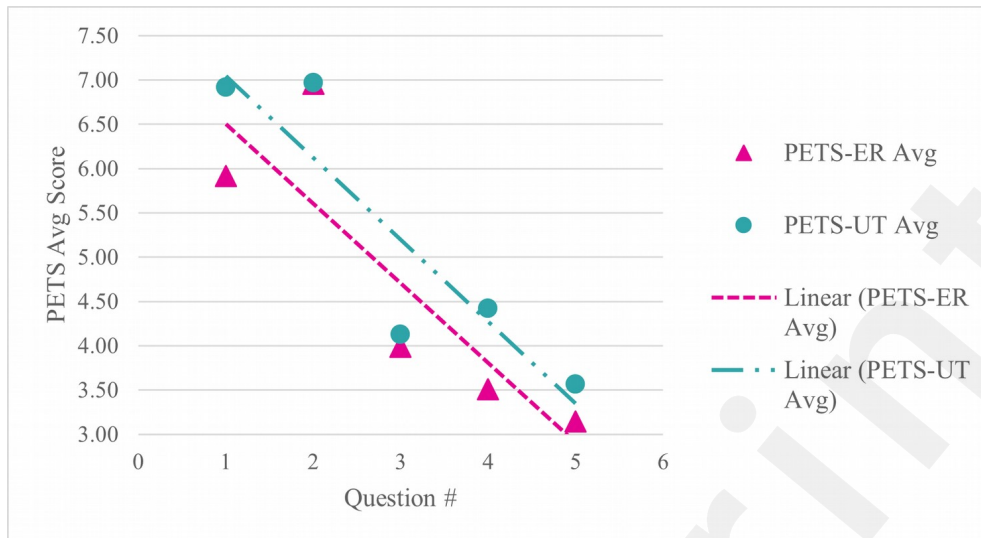


Figure 4: Average PETS across Chatbot by Question

At the chatbot level, ChatGPT received the highest empathy scores (ER Avg= 5.36, UT Avg= 6.37). Copilot scored lowest (ER Avg= 4.36, UT Avg= 4.61), with Gemini performing only marginally higher than Copilot (ER Avg= 4.40, UT Avg= 4.62). Across all models, PETS-UT scores exceeded PETS-ER, suggesting stronger cognitive empathy and perspective taking rather than emotional expression. ChatGPT showed the largest UT-ER gap (6.4 vs 5.4), indicating moderate to high levels of understanding, but only moderate emotional connection. Copilot and Gemini had similar low-moderate ER and UT averages, with limited perceived empathy overall. Variability patterns (Figure 5, SD values) further distinguish AI models. ChatGPT demonstrated the greatest variation (SD ~1.8), further evidenced by the spread of scores as related to the mean (Figure 5). In contrast, Gemini and Copilot showed tighter distributions (SD~1.7), indicating slightly more predictable patterns of empathy.

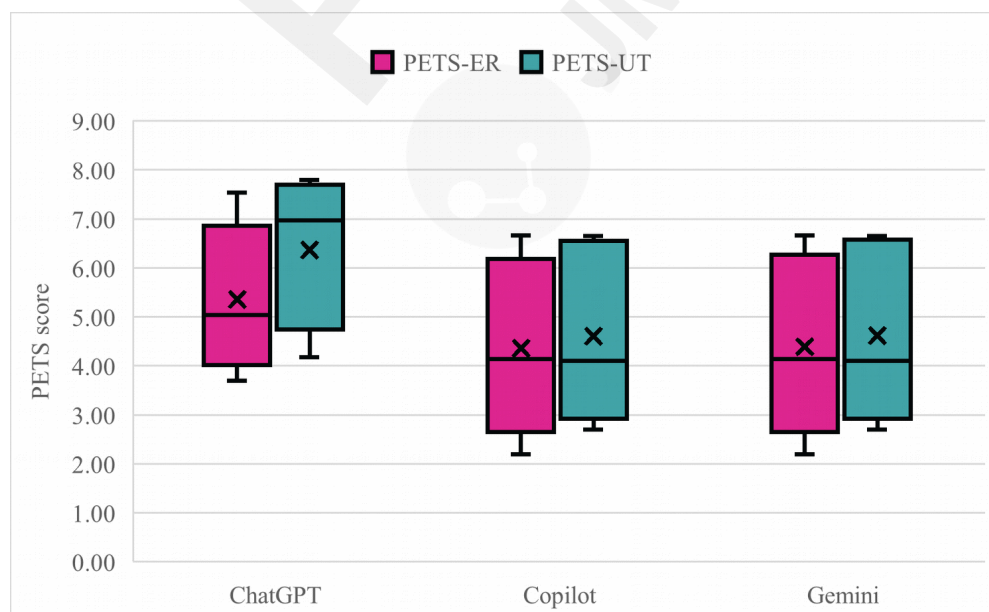


Figure 5: Box and whisker plots showing Pets-ER and PETS-UT scores across chatbots, illustrating limited overall perceived empathy compared to understandability.

Readability (Flesch-RE, Flesch-GL, Word Count)

All outputs were written at an 8th grade reading level or higher (range 8.2-14.6, mean 11.8, equivalent to high school senior level). The average Flesch Readability (Flesch-RE) score across all AI chatbot responses was 38.60, indicating difficult readability overall. For four out of five questions, Gemini produced the most readable responses (mean Flesch-RE = 41.75). Copilot consistently produced responses with difficult readability (mean Flesch-RE = 36.1). ChatGPT produced responses with the widest range in readability (28.58-54.84). Question 5 responses were the least readable (mean grade level = 12.76), while Question 2 responses were the most readable (mean grade level = 9.9). Word count differed notably, with Gemini responses being the longest (~728 words), ChatGPT being mid-length (~482 words), and Copilot being the shortest (~356). In summary, Gemini provided the most readable and detailed outputs, ChatGPT had more variation in readability, and Copilot responses were brief and more difficult to read.

Discussion

Quality

ChatGPT provided the highest quality responses, but with moderate variability across questions. Gemini delivers moderate quality responses with greater consistency, while Copilot was the least reliable, even producing a factually incorrect answer in one instance. The variability found in this study has important implications; when information quality fluctuates, users cannot predict whether or not an answer is trustworthy. In health communication, information quality standardization is critical because it allows for less confusion and increases patient-provider trust [12].

Quality is not only model dependent but also question reliant. Q1 and Q2 received the highest quality evaluations, whereas Q4 and Q5 produced lower quality scores. This pattern suggests that when questions are more straightforward or commonly addressed in the literature, chatbots are more likely to provide a reliable response. However, questions regarding more complex science, with sparse and/or unclear literature, chatbots provide unreliable information without stating uncertainty. For AYA melanoma patients, this is concerning because questions such as these often represent the areas of greatest uncertainty.

Review of lower-quality outputs revealed common pitfalls. Responses with low quality scores were often oversimplified, incomplete, or presented in bullet-point style responses, with insufficient detail to address the question input. The lack of actionable information was most apparent in discussions of treatment, side effects, and hormone changes. For example, when a patient asked how treatment would affect hormones, responses accurately identified potential consequences, such as adrenal insufficiency, but failed to describe symptoms. From a patient perspective, clinically beneficial communication translates these answers into recognizable manifestations at home (i.e. fatigue, dizziness, or other changes). Accuracy was further undermined by lack of transparent source citation and inconsistent use of safeguards such as disclaimers. Instead, statements were presented as definitive, risking patient confusion when conflicting with other resources or provider guidance.

Conversely, higher quality responses demonstrated more complete and organized content. They cited credible resources, gave disclaimers, and emphasized the importance of deferring to the healthcare team for individualized guidance. These responses were detailed and well-tailored to the questions, without an overwhelming word count, which improved readability and digestion of information. They had clear directions to access specific resources for further support and actionable suggestions

for the patient. These qualities serve as a reference standard for future versions of AI-generated health information.

Although overall quality scores across all Chatbot responses were moderate to high, persistent gaps in sourcing and uncertainty disclosures raise serious patient safety concerns. For AYA melanoma patients undergoing treatment, inconsistent quality can undermine trust and lead to confusion at a vulnerable stage of care. To be reliable in clinical contexts, AI systems must deliver outputs that are not only accurate and detailed, but also verifiable, transparent, and responsibly framed.

Empathy

Across the five questions over three chatbots, a central pattern emerged: systems struggled to balance clinical accuracy with emotional support. ChatGPT showed the strongest empathetic capacity, while Gemini and Copilot tended to be neutral and lacking in emotion. Qualitative review of responses identified a set of best practices that distinguished qualities of higher-empathy outputs from weak ones (Table 4).

Table 4. Examples of best practices for AI chatbot responses to express empathy

Best Practices for AI Chatbot Responses to express empathy		
Best Practices (↑ PETS scores)	Downfalls (↓ PETS scores)	Examples
Open with empathy and understanding	Robotic or clinical tone	Acknowledge diagnosis, saying “I’m sorry you’re going through this,” or validating fear/stress.” vs. “I understand you are an adolescent/young adult patient with metastatic melanoma...”, starting immediately with medical jargon, sounding cold or impersonal.
Follow-up & clarification questions	One-way “information dump”	“Would you like me to help you prepare questions for your oncologist, → asking “Can you tell me more?” or “Did I cover what you were hoping for?” vs. No follow up or clarification questions
Symptom and Side effect explanations in understandable terms	List of side effects or clinical jargon	In discussing possible pancreatic insufficiency 2/2 treatment: “You may notice you feel more tired or thirsty which could be a sign of these hormonal changes” vs. No explanation of how to recognize these changes at home, only providing name of treatment and name of hormone affected, or listing severe side effects that are not patient-specific without a full clinical picture
Medical disclaimers & boundaries	Overstep scope	“The information provided here is for educational purposes and should not replace a consultation with your doctor” vs. “I can help interpret your genetic testing results.”, “Have you considered clinical trials?”
Cites credible resources	Lack citations or specific resources	When discussing support providing links to AYA resources- Stupid Cancer, NCCN, ACS, fertility & survivorship resources vs. suggesting “therapy” or “support” but no guide on where to find resources
Specific actionable coping strategies	Generic or overwhelming advice	“Make a symptom diary, here is how...”; “try meditation strategies such as...”; “try box breathing here is how you do it” vs. “Do weekly self-exams” which is vague and puts pressure on the patient to discover new illness, instigating anxiety and fear or “fertility preservation discussions are crucial before starting therapy” which overlooks the reality that many patients are not given the opportunity to discuss fertility preservation before beginning treatment
Validates and justifies patient's emotional response	Naming emotions without explanation or context	AYA tailored and validating response: “For AYAs this fear can be... more intense because... life stage focused on building a future... feel isolated (from peers)..., faced with uncertainty and loss of control over your body and future.” vs something general that restates emotion expressed: “I’m sorry you’re scared”

Inclusive, empathetic language that normalizes and reassures	Emotionally shallow or "empathy sandwich" automated response	"You're not alone in this, and it's important to know that there are ways to cope with that fear and take steps toward feeling more supported and in control." vs brief and generic emotional statements ("it can be difficult")
AYA-specific	Adult-centric or non-specific oncology framing	Providing an AYA-specific section that covers fertility, dating, school, work, independence, peer support vs. Presenting information as if for general oncology patients, missing developmental relevance and differences in needs, one response even mentioned "Children's Oncology guidelines"
Demonstrates interest in the patient	Lack of interest or personalization	"I'd like to help you think about how this affects your daily life", "Do you have any other questions" vs. No personalization, failure to engage with the patient
Structured and concise guides	Clinical jargon overload	Step-by-step plans, checklists, "Actionable steps you can take this week", or "what to expect" pathways vs long, dense paragraphs filled with clinical jargon without explanations
Provides empathy in all contexts	Overlooks unspoken emotional needs	In inputs with less emotion (3-5): "I'm really sorry to hear that you're dealing with metastatic melanoma- what you're going through is incredibly challenging" vs no empathetic statements, simply opening with "Thank you for your question" or diving directly into biology

AYA melanoma patients have unique needs compared to adults; however, chatbots often default to general or even pediatric framing [13]. When AYA patient-specific needs were addressed (fertility, fear of cancer spread, and long-term systemic effects of treatment), it was often with a surface-level list rather than a meaningful explanation. Systems were better equipped to name emotions rather than express them, as demonstrated by higher UT than ER scores, and empathy was also prompt-dependent. Chatbot responses generally failed to infer patient distress if not explicitly stated (as seen in Q3-Q5), unlike a human clinician who can use nonverbal cues and proactively address unspoken emotional needs even in the face of an analytical question. The chatbot lacks the emotional intelligence to deliver difficult news and interpret patient reactions, which can lead to patient harm. Responses receiving low PETS scores risk coming across as dismissive or cold, which may invalidate patient concerns and discourage them from seeking support or adhering to care recommendations in the future when they most need it.

Human providers naturally contextualize information based on patient age, life stage, current treatments, and psychosocial needs. Chatbots did not replicate this, limiting their ability to foster trust and support. Since our five questions represent domains of greatest emotional and informational needs of AYAs with melanoma, the irregularity of empathy suggests that current AI systems cannot adequately support this unique patient population.

Readability

While Gemini generated the longest, most readable responses, ChatGPT demonstrated wide variation, and Copilot's shorter responses were most difficult to read. Responses that were easier to read tended to be perceived as higher quality (GQS/DISC) and more empathetic (PETS) as seen in Figure 6.

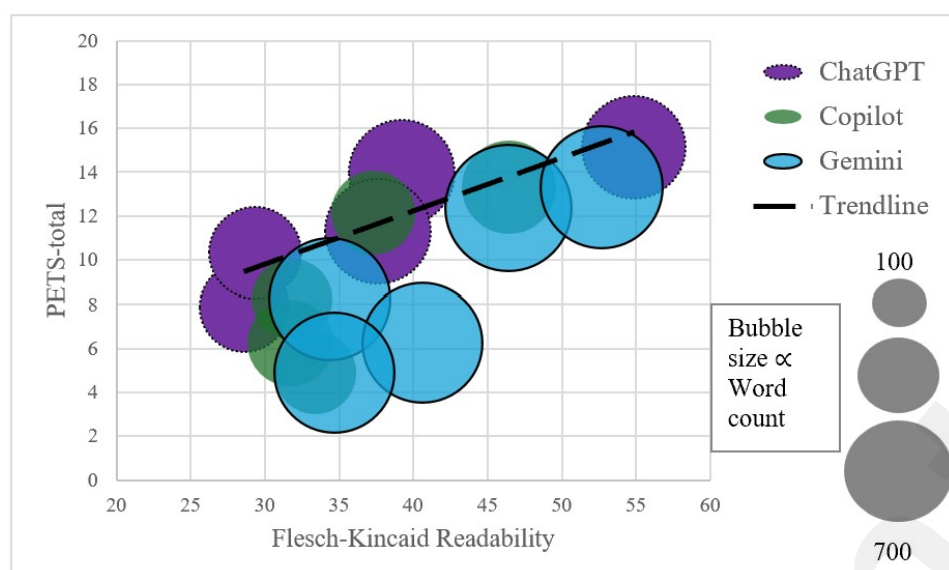


Figure 6. Relationship between Flesch-Kincaid readability and total PETS (empathy) scores across all chatbot responses. Each bubble represents a chatbot, with size proportional to overall word count. While word count shows little association with either readability or empathy, responses with higher readability tend to receive higher empathy scores.

Readability analysis revealed that all chatbot responses exceeded the US average reading level (7th-8th grade), with outputs averaging at the 11th – 12th grade level. This creates a barrier for accessibility, particularly for younger AYA patients and those with lower health literacy. These findings echo prior research that healthcare documents are often written at a level too high for general populations, reducing inclusivity and comprehension [14]. Unlike clinicians, who are trained to utilize strategies such as follow-up questions and the “teach-back” method, chatbots do not check understanding unless prompted. This may highlight a key limitation of AI chatbots for medical communication. They draw primarily from online sources, which are often already written at inflated reading levels, and thus lead to reproduced complexity by AI outputs [15]. In contrast, healthcare providers are trained to explain concepts in multiple ways, tailoring explanation to patients’ literacy levels, age, individual clinical context.

AI therefore may be able to produce content that is technically correct but inaccessible, leaving patients confused and overwhelmed. For proper implementation, chatbots should aim to target their output responses of healthcare information to an 8th grade level and to incorporate interactive strategies that mimic provider feedback loops. Without these adjustments, utilization of AI to answer AYA melanoma patient needs remains limited.

Patient Safety & Clinical Implications

Although AI chatbots can provide AYA patients with adequate education, support, and resources regarding their diagnosis, they cannot replace clinician input and should be used cautiously. The dangers associated with patients using AI in healthcare are multifaceted, with key risks including misinformation, unsafe recommendations, lack of understanding of the questions being asked, and insensitivity. Given that AYAs are the primary users of AI compared to other age groups, they are at high risk for harm or adverse outcomes, specifically due to their vulnerability to algorithmic bias, misinformation, and lack of access to adequate resources [16].

AI use in healthcare has been shown to spread misinformation [17]. In our study, we found that chatbot explanations regarding the causes of metastatic melanoma and recommendations for next steps often included inaccurate or misleading information. Responses across all 3 platforms often

focused on sun avoidance and SPF protection to prevent melanoma recurrence or secondary malignancy. Although this advice may be relevant for certain patient populations, it is not applicable to all patients, particularly for individuals with skin of color. As a result, responses lack inclusivity for diverse patient communities. Additionally, some chatbot responses encouraged AYA patients to explore clinical trials available for their conditions. While well-intended, this response is particularly unsafe to AYA patients as it provides them with false hope and may subject patients to considering studies that are not medically appropriate, not eligible to the patient, or misaligned with their treatment plan. Some responses even offered to interpret lab and genetic testing results. While AI chatbots have the potential to explain pathology reports and simplify medical jargon to patients, their interpretations are not flawless and should be used carefully [18]. Encouraging vulnerable patients to input their private health information to AI systems also exposes them to risks including potential breaches of data privacy and security [19]. This can lead to unauthorized access and misuse of sensitive medical information, underscoring the need for strict parameters and oversight when using AI in healthcare to both protect patient privacy and regulate the information and suggestions provided by AI.

Limitations

This study was limited by the amount of chatbot responses evaluated. Had questions been input into the chatbot more times, they may have scored differently for GQS, DISCERN, and PETS. Secondly, AI chatbots are very directable and can successfully provide more follow up information and clarifications through continued interaction. Because questions were asked only once before the search history was cleared, the AI chatbots did not have the opportunity to provide follow up, which may have affected the overall quality, empathy, and understandability of the responses. Lastly, the phrasing of the questions input to the chatbots can affect the quality, readability, and tone of responses generated. If questions were phrased differently, or if more of a clinical vignette was provided per question, the response scores may vary. Additionally, this study is limited by the subjective nature of the grading scales and variability between grading styles of reviewers. To mitigate this risk, each response was evaluated two times, and for transparency we have included all outputs in MA 4-6 as well as an example of the grading process with score justifications in MA 7. Further analysis of chatbot responses with more complex questions and interaction between the patient and chatbot is necessary to further evaluate these tools.

Future Directions

AI tools for AYA patients undergoing melanoma treatment should be used in conjunction with physician advice to ensure patient safety and quality control of information. Moving forward, we aim to use responses from this dataset to develop resources within melanoma oncology clinics that answer these questions and provide higher quality and safer responses for AYA melanoma patients. By providing age-appropriate education and support tools we hope to improve the treatment experience and the overall well-being for AYAs navigating melanoma care. This will help expand the currently limited resources available for AYAs currently receiving treatment and hopefully bring more attention to the unaddressed survivorship needs of patients.

Conclusion

This study provides an evaluation of the utility of AI to answer the most pressing needs of AYA melanoma patients currently receiving treatment. Overall, AI responses were useful for patients. Despite this, AI responses had persistent gaps which led to lower quality, empathy, and readability. Patient safety may be at risk due to the inability of AI systems to integrate patient context, emotional nuance, and conversational reciprocity into answers. Question wording, emotional phrasing, and clarity were predictive of the responses received. When questions addressed topics covered by

multiple research publications, AI responses were more consistent, with higher scores for quality. However, when questions addressed controversial or rapidly changing topics, AI responses were often vague. Chatbot empathy is not inherent and is largely dependent on the emotional level of question input to produce empathetic outputs. Quality, empathy, and readability were all interconnected as higher-rated responses incorporated clear and concise wording, showed interest in the patient's current medical status, provided tangible resources and references, and recognized the limits of AI, encouraging further conversation with medical providers. This research should be replicated and can provide a reference standard for future AI chatbot responses for AYA melanoma patients.



Funding Statement:

No financial disclosures

COI:

None declared

DATA availability:

The dataset generated during this study, including all chatbot outputs and scoring, is available in the Multimedia Appendices. No human subject data was collected.

Author contributions:

JLJ and PLH are co-first authors and contributed equally to this work. JLJ, PLH and KCN contributed to conceptualization and methodology; JLJ and PLH contributed to question development, empathy and readability scoring, and manuscript writing. RPM, SZ, AS, and KMR performed question input and data collection. MR provided project oversight and expert clinical opinion. MHS contributed to manuscript writing and editing, quality scoring, and response review. KCN served as principal investigator, provided project oversight, conceptualization, methodology, contributed to data collection, and participated in manuscript writing and editing. All authors reviewed and approved the final manuscript.

Abbreviations:

AI: Artificial Intelligence
AYA; Adolescents and Young Adults
DISC: Discern Scale
ER: Emotional Responsiveness
FKGL: Flesch-Kincaid Grade Level
FRE: Flesch Reading Ease
GQS: Global Quality Scale
MA: Multimedia Appendix
NA-SB: Needs Assessment and Service Bridge
NLP: Natural Language Processing
PETS: Perceived Empathy of Technology Scale
SD: Standard Deviation
UT: Understanding and Trust

References

1. National Cancer Institute. SEER Cancer Stat Facts: Cancer Among Adolescents and Young Adults (AYAs). n.d. Accessed August 23, 2025. <https://seer.cancer.gov/statfacts/html/aya.html>
2. McNally W, Gray-Brunton C, Chouliara Z, Kyle RG. Life Interrupted: Experiences of adolescents, young adults and their family living with malignant melanoma. *J Adv Nurs*. 2021;77(9):3867-3879. doi:10.1111/jan.14959
3. Sidoti O and MC. 34% of U.S. adults have used ChatGPT, about double the share in 2023. Pew Research Center. June 25, 2025. Accessed August 23, 2025. <https://www.pewresearch.org/short-reads/2025/06/25/34-of-us-adults-have-used-chatgpt-about-double-the-share-in-2023/>
4. Clark M, Bailey S. *CADTH Horizon Scan Chatbots in Health Care: Connecting Patients to Information*. Vol 4.; 2024.

5. Kim HK. The Effects of Artificial Intelligence Chatbots on Women's Health: A Systematic Review and Meta-Analysis. *Healthcare (Basel)*. 2024;12(5). doi:10.3390/healthcare12050534
6. Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. Assessment of Artificial Intelligence Chatbot Responses to Top Searched Queries About Cancer. *JAMA Oncol*. 2023;9(10):1437-1440. doi:10.1001/jamaoncol.2023.2947
7. Moore, Reece, et al., Identifying Gaps and Needs in Care for Adolescents and Young Adults with Melanoma. Submitted for publication in *Supportive Care in Cancer* Under Peer Review. 2025
8. Bernard A, Langille M, Hughes S, Rose C, Leddin D, Veldhuyzen van Zanten S. A Systematic Review of Patient Inflammatory Bowel Disease Information Resources on the World Wide Web. *Am J Gastroenterol*. 2007;102(9):2070-2077. doi:10.1111/j.1572-0241.2007.01325.x
9. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. *J Epidemiol Community Health (1978)*. 1999;53(2):105-111. doi:10.1136/jech.53.2.105
10. Schmidmaier M, Rupp J, Cvetanova D, Mayer S. Perceived Empathy of Technology Scale (PETS): Measuring Empathy of Systems Toward the User. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM; 2024:1-18. doi:10.1145/3613904.3642035
11. Ley P, Florio T. The use of readability formulas in health care. *Psychol Health Med*. 1996;1(1):7-28. doi:10.1080/13548509608400003
12. Chauhan, V. & Sagar, M. (2023). *Healthcare decision making and choice: an empirical model of patient confusion*. *Management Decision*, 61(11), 3454–3474-
13. Ahmad N, 2025 Kashyap A, Taylor M, et al. Quality of Life Among Adolescents and Young Adults with Melanoma: A Systematic Review. *J Adolesc Young Adult Oncol*. Published online June 5, 2025.
14. Espinosa J, Lucerna A, Schuitema H. (2022). Healthcare Documents: How Readable? A Study of Simple Strategies to Decrease Reading Grade Level and Increase Reading Ease of Healthcare Documents. *Mathews J Emergency Med*. 7(1):44
15. OpenAI. How ChatGPT and Our Foundation Models Are Developed. OpenAI Help Center. Updated 07/2025. Accessed September 10, 2025. <https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-foundation-models-are-developed>
16. Carolan, S.; Martin, A.W.; Gong, G.C.; Borja, S. 2025: The State of Consumer AI. MENLO Ventures. June 26, 2025. Accessed September 09, 2025. <https://menlovc.com/perspective/2025-the-state-of-consumer-ai/>
17. Denecke K, Lopez-Campos G, Rivera-Romero O, Gabarron E. The Unexpected Harms of Artificial Intelligence in Healthcare: Reflections on Four Real-World Cases. *Stud Health Technol Inform*. 2025;325:55-60. doi:10.3233/SHTI250219
18. Steimetz E, Minkowitz J, Gabutan EC, et al. Use of Artificial Intelligence Chatbots in Interpretation of Pathology Reports. *JAMA Netw Open*. 2024;7(5):e2412767. Published 2024 May 1. doi:10.1001/jamanetworkopen.2024.12767
19. Li J. Security Implications of AI Chatbots in Health Care. *J Med Internet Res*. 2023;25:e47551. Published 2023 Nov 28. doi:10.2196/47551

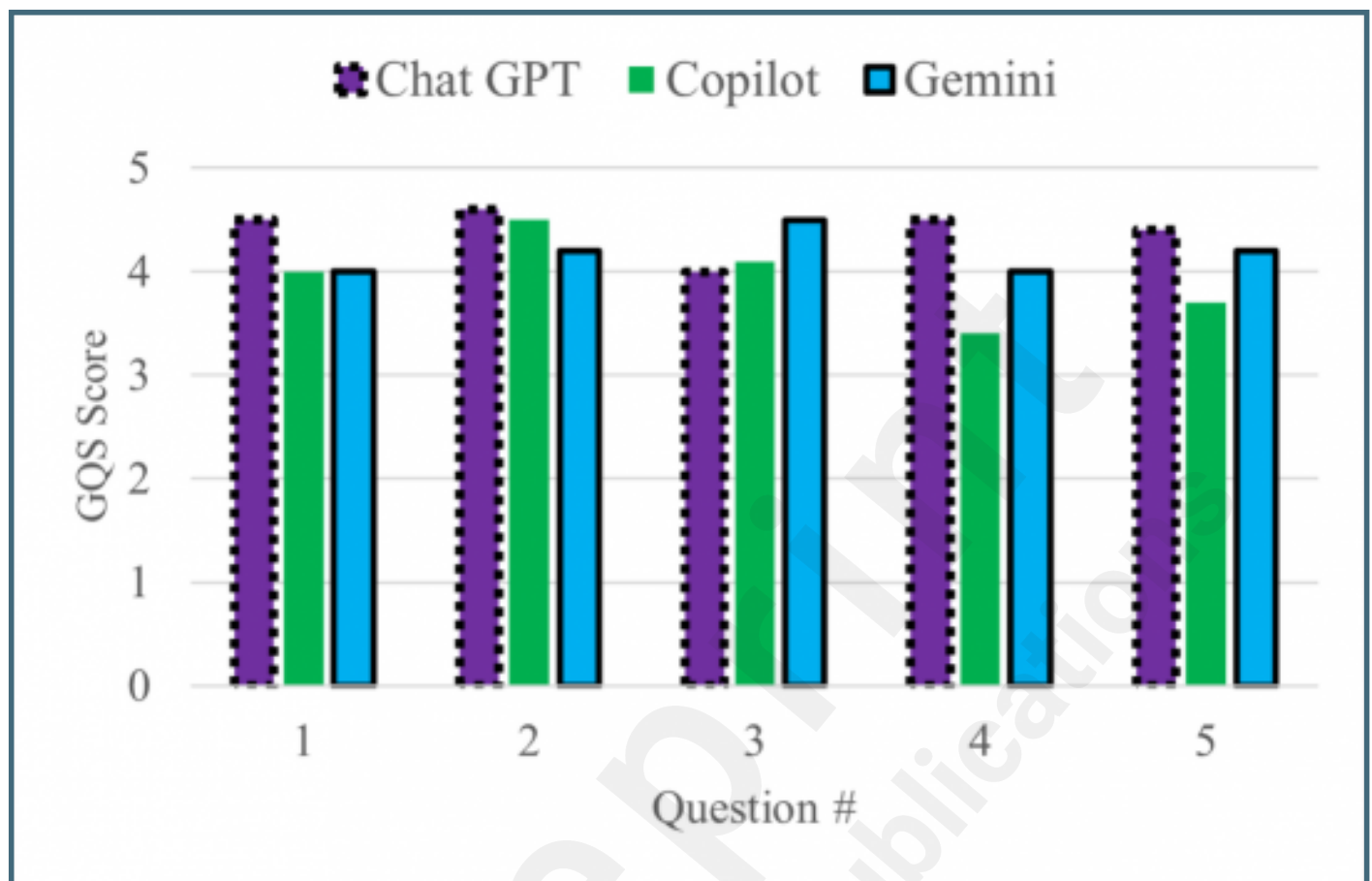
Supplementary Files

Figures

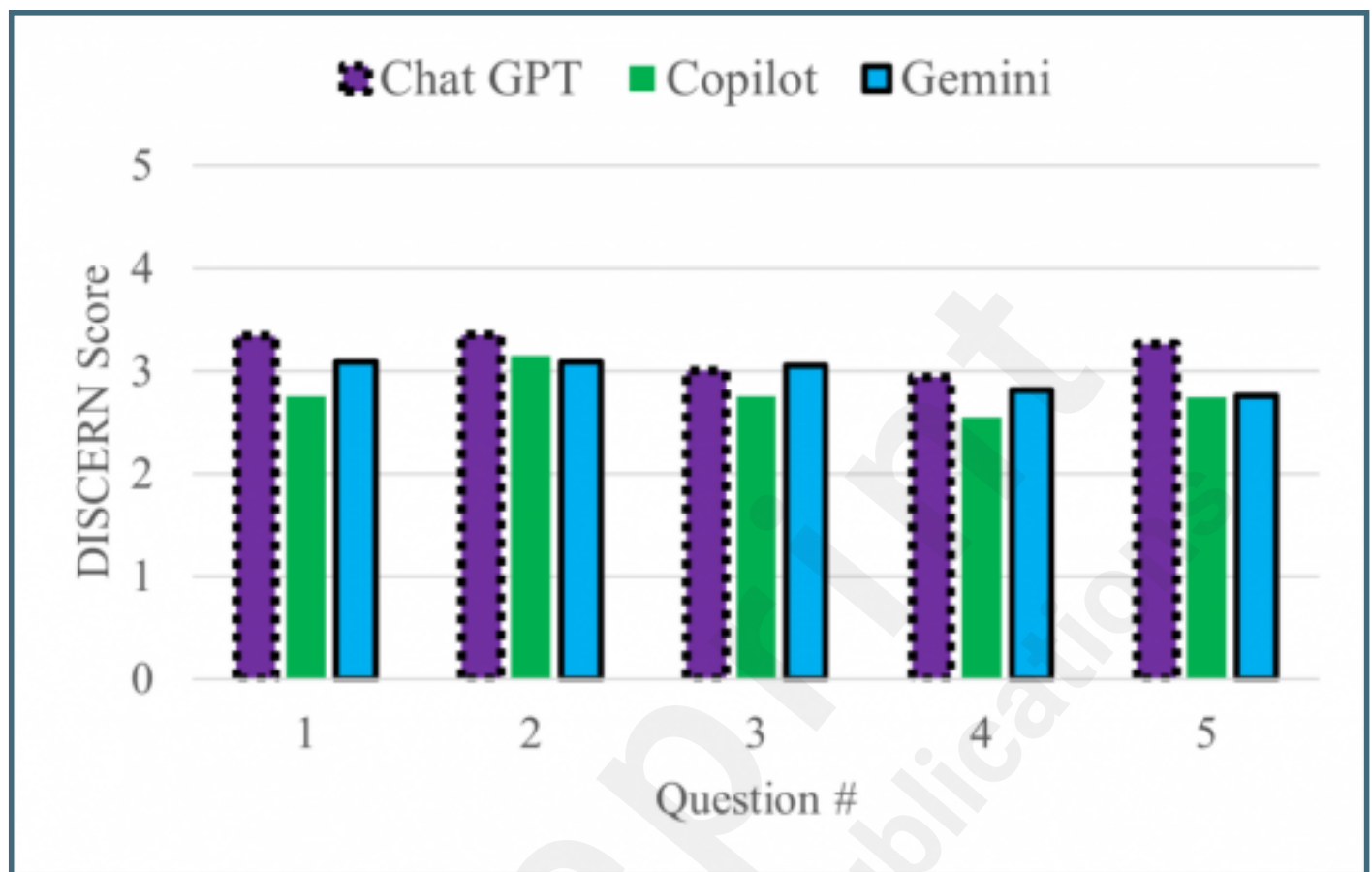
Study design workflow.



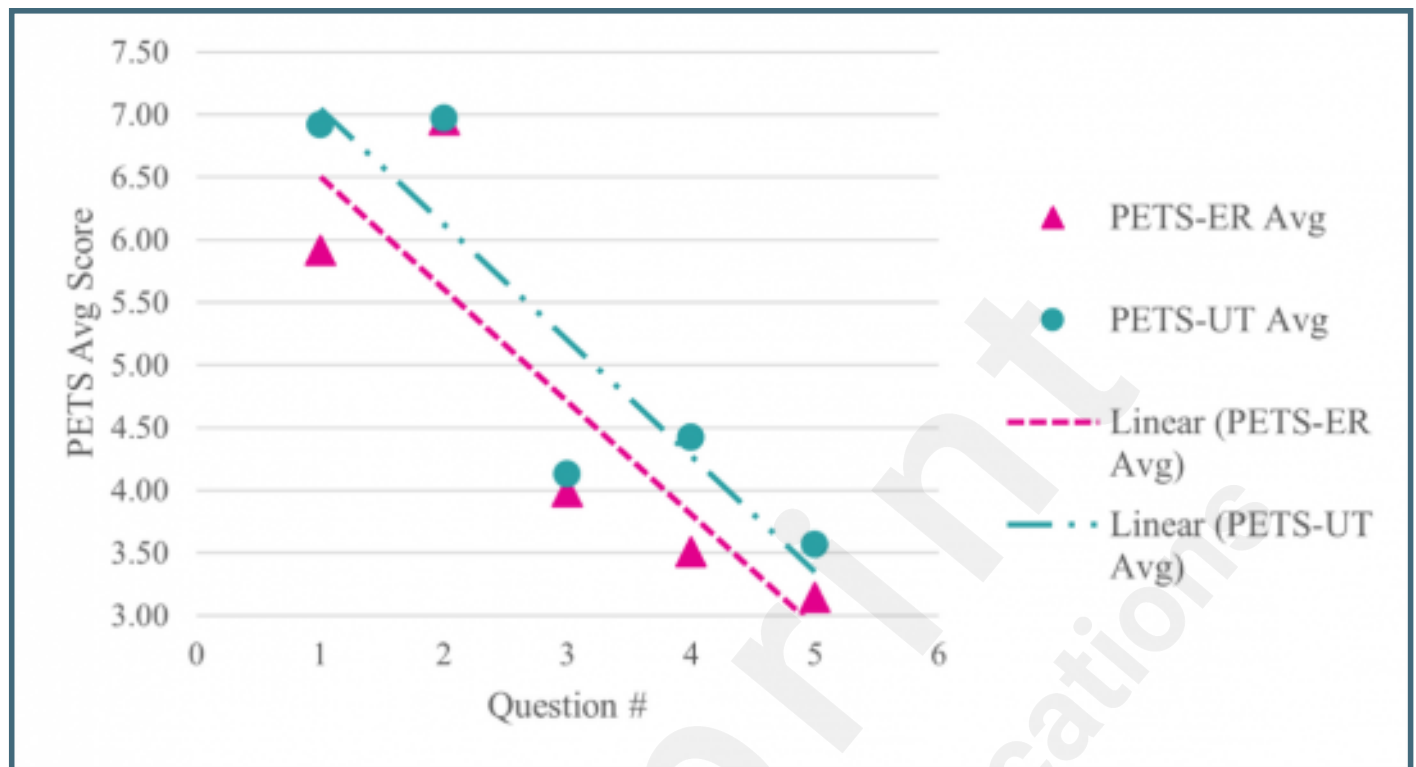
GQS scores comparing each chatbot by question input.



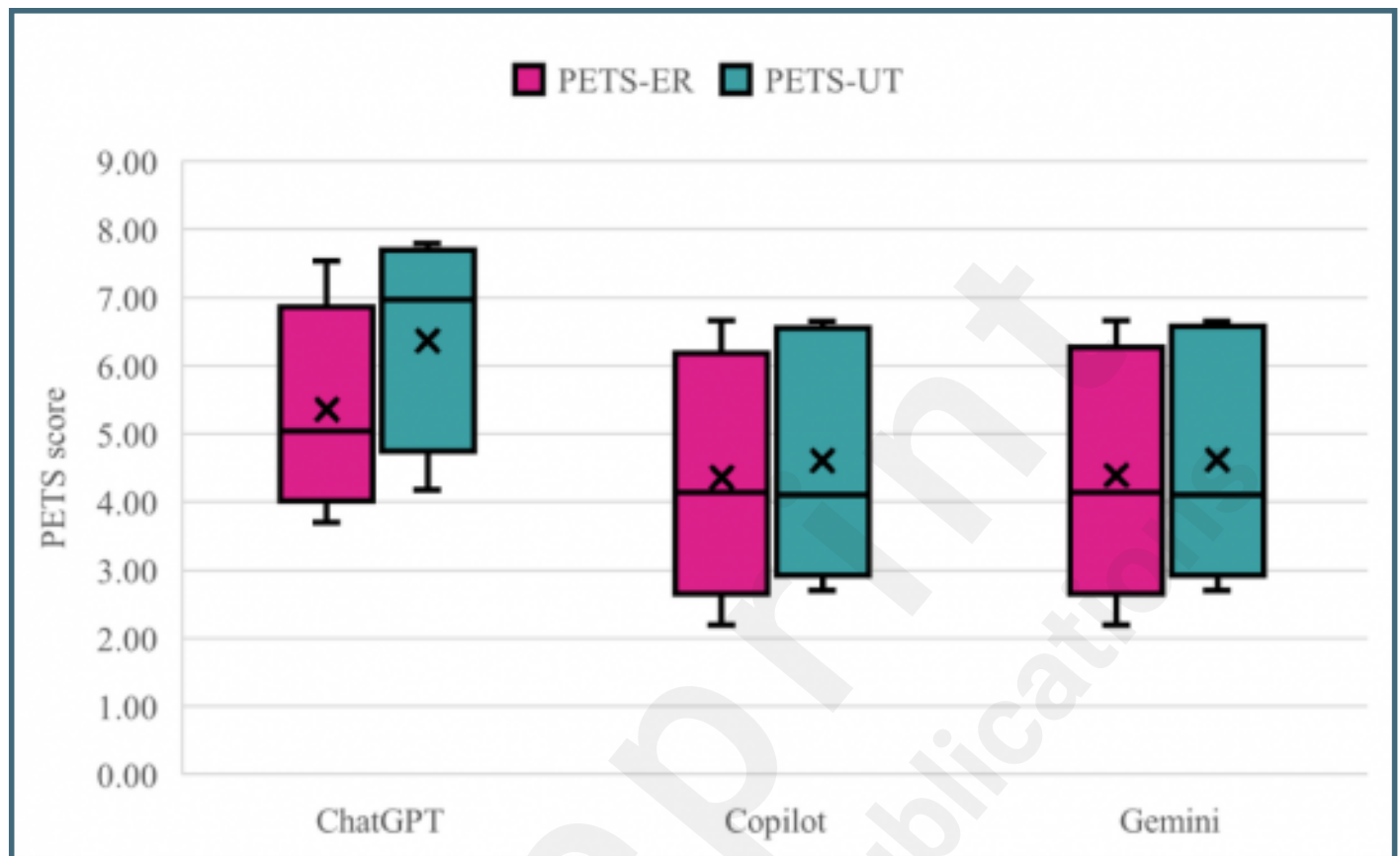
Discern scores comparing each chatbot by question input.



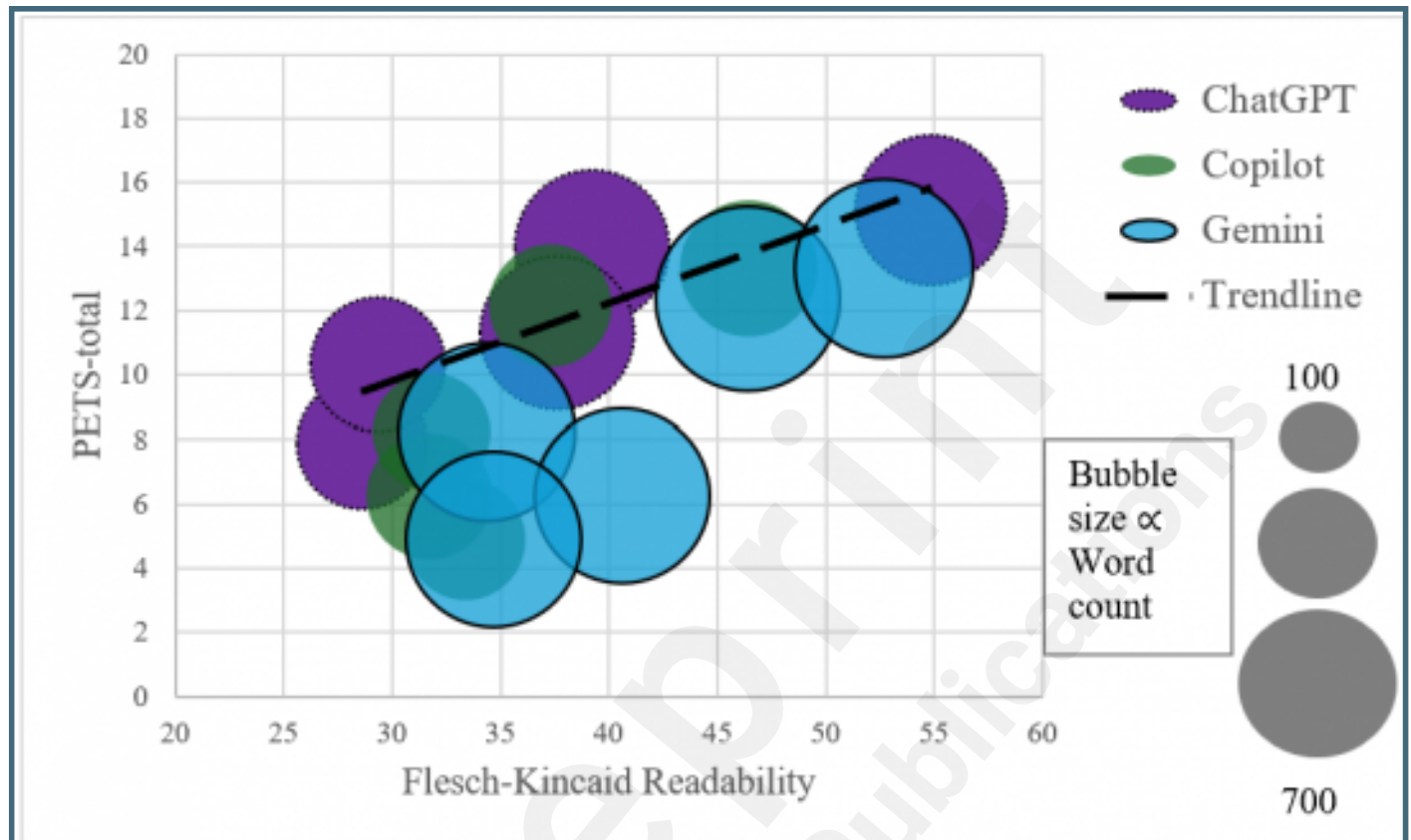
Average PETS across Chatbot by Question.



Box and whisker plots showing Pets-ER and PETS-UT scores across chatbots, illustrating limited overall perceived empathy compared to understandability.



Relationship between Flesch-Kincaid readability and total PETS (empathy) scores across all chatbot responses. Each bubble represents a chatbot, with size proportional to overall word count. While word count shows little association with either readability or empathy, responses with higher readability tend to receive higher empathy scores.



Multimedia Appendixes

Table S1-S3: S1. GQS Scoring Scale; S2 Discern Scale; S3 PETS score.

URL: <http://asset.jmir.pub/assets/13f5e64c09f9325d190d833aaa0c039b.docx>

Chatbot response scores for each input.

URL: <http://asset.jmir.pub/assets/af9a08d3c0d55eecdd8375d20c49526b.docx>

Calculated standard deviations of response scores.

URL: <http://asset.jmir.pub/assets/b953253e6116b12706cf92d20b921d10.docx>

All ChatGPT outputs.

URL: <http://asset.jmir.pub/assets/5c3308e103ff94b70e8d281b3d70c447.docx>

All Google Gemini outputs.

URL: <http://asset.jmir.pub/assets/69bda07a9a076894880fb4138daecfd5.docx>

All Microsoft Copilot outputs.

URL: <http://asset.jmir.pub/assets/fb7d10bcfca7fe723a72895a54dc5ad9.docx>

Example of response score breakdown.

URL: <http://asset.jmir.pub/assets/64aaa8b58af951941667ac1f32a0d364.docx>