Evaluation of Large Language Models in Tailoring Educational Content for Disadvantaged Cancer Survivors and Their Caregivers

Darren Liu, Xiao Hu, Canhua Xiao, Jinbing Bai, Zahra Barandouzi, Stephanie Lee, Caitlin Webster, La-Urshalar Brock, Lindsay Lee, Delgersuren Bold, Yufen Lin

Submitted to: JMIR Cancer
on: October 24, 2024

Table of Contents

Evaluation of Large Language Models in Tailoring Educational Content for Disadvantaged Cancer Survivors and Their Caregivers

Darren Liu[1]; Xiao Hu[1]; Canhua Xiao[1]; Jinbing Bai[1]; Zahra Barandouzi[1]; Stephanie Lee[1]; Caitlin Webster[1]; La-Urshalar Brock[1]; Lindsay Lee[2]; Delgersuren Bold[1]; Yufen Lin[1]

[1]Nell Hodgson Woodruff School of Nursing, Emory University Atlanta US
[2]Department of Medicine, University of Florida Gainesville US

Corresponding Author:
Yufen Lin
Nell Hodgson Woodruff School of Nursing, Emory University
1520 Clifton Rd NE
Atlanta
US

Abstract

Background: Disadvantaged cancer survivors and their caregivers (e.g., individuals with limited health literacy, racial and ethnic minorities facing language barriers) face a disproportionately increased risk of symptom burden from cancer and its treatments. Large language models (LLMs) offer researchers an opportunity to develop educational materials tailored to these populations.

Objective: The purposes of this study were to: 1) evaluate the overall performance of LLMs in generating tailored educational content for disadvantaged cancer survivors and their caregivers; 2) compare the performances of three Generative Pre-trained Transformer (GPT) models (i.e., GPT-3.5 Turbo, GPT-4, GPT-4 Turbo); and 3) explore different prompts that can help LLMs generate better content.

Methods: We selected 30 topics from national guidelines on cancer care and education. GPT-3.5 Turbo, GPT-4, and GPT-4 Turbo were used to generate tailored content of up to 250 words at a 6th-grade reading level, with translations into Spanish and Chinese for each topic. Nine oncology experts evaluated the content based on pre-determined criteria: word limit, reading level, and quality assessment (i.e., clarity, accuracy, relevance, completeness, and comprehensibility). ANOVA or Chi-square analyses were employed to compare differences among the various GPT models and prompts.

Results: Overall, LLMs showed excellent performance in tailoring educational content, with 74.2% (n=360) adhering to the specified word limit and achieving an average quality assessment score of 8.933 out of 10. However, LLMs showed moderate performance in reading level, with 41.1% of content failing to meet the 6th-grade reading level. LLMs demonstrated strong translation capabilities, achieving an accuracy of 88.9% for Spanish and 81.1% for Chinese translations. The more advanced GPT-4 family models showed better overall performance compared to GPT-3.5 Turbo. Prompting GPTs to produce bulleted-format content was likely to result in better educational materials compared to textual-format content.

Conclusions: This study highlights the application of LLMs in cancer care and education while acknowledging their potential limitations. The findings can inform the development and implementation of interventions in cancer symptom management and supportive care, thereby advancing health equity.

(JMIR Preprints 24/10/2024:67914)
DOI: https://doi.org/10.2196/preprints.67914

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 v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Evaluation of Large Language Models in Tailoring Educational Content for Disadvantaged Cancer Survivors and Their Caregivers

Abstract

Background: Disadvantaged cancer survivors and their caregivers (e.g., individuals with limited health literacy, racial and ethnic minorities facing language barriers) face a disproportionately increased risk of symptom burden from cancer and its treatments. Large language models (LLMs) offer researchers an opportunity to develop educational materials tailored to these populations.

Objective: The purposes of this study were to: 1) evaluate the overall performance of LLMs in generating tailored educational content for disadvantaged cancer survivors and their caregivers; 2) compare the performances of three Generative Pre-trained Transformer (GPT) models (i.e., GPT-3.5 Turbo, GPT-4, GPT-4 Turbo); and 3) explore different prompts that can help LLMs generate better content.

Methods: We selected 30 topics from national guidelines on cancer care and education. GPT-3.5 Turbo, GPT-4, and GPT-4 Turbo were used to generate tailored content of up to 250 words at a 6th-grade reading level, with translations into Spanish and Chinese for each topic. Nine oncology experts evaluated the content based on pre-determined criteria: word limit, reading level, and quality assessment (i.e., clarity, accuracy, relevance, completeness, and comprehensibility). ANOVA or Chi-square analyses were employed to compare differences among the various GPT models and prompts.

Results: Overall, LLMs showed excellent performance in tailoring educational content, with 74.2% (n=360) adhering to the specified word limit and achieving an average quality assessment score of 8.933 out of 10. However, LLMs showed moderate performance in reading level, with 41.1% of content failing to meet the 6th-grade reading level. LLMs demonstrated strong translation capabilities, achieving an accuracy of 88.9% for Spanish and 81.1% for Chinese translations. The more advanced GPT-4 family models showed better overall performance compared to GPT-3.5 Turbo. Prompting GPTs to produce bulleted-format content was likely to result in better educational materials compared to textual-format content.

Conclusions: This study highlights the application of LLMs in cancer care and education while acknowledging their potential limitations. The findings can inform the development and implementation of interventions in cancer symptom management and supportive care, thereby advancing health equity.

Keywords: large language models; GPT-4; cancer survivors; caregivers; education; health equity

Introduction

More than 18.1 million individuals with a history of cancer were alive in the United States in 2022, and that number is projected to reach 26 million by 2040 [1]. Cancer survivors receive a wide range of treatments, often experiencing severe symptoms or side effects, including fatigue, depression, anxiety, sleep disturbance, pain, cognitive impairment, nausea, vomiting, and neuropathy [2-7]. These symptoms negatively impact survivors' functional status, quality of life (QOL), and overall survival rates [8-11]. Cancer caregivers, typically family members or

significant others offering primary emotional and physical support for cancer survivors, experience an array of similar distressing symptoms [12-14]. These symptoms are linked to high caregiving burden, emotional distress, and communication barriers with cancer survivors and providers [15]. Additionally, disparities in healthcare access further exacerbate the challenges faced by cancer survivors and their caregivers, especially those from disadvantaged communities [16]. Those with limited health literacy and racial and ethnic minorities facing language barriers are at greater risk for poorer access to care [17-19]. Consequently, they tend to experience a heavier symptom burden and poorer health outcomes during and after cancer treatments [20].

With over three-quarters of the disadvantaged population owning smartphones or computers [21], technology-based intervention programs can bridge the accessibility gap and promote health equity [22, 23]. The advent and growth of artificial intelligence (AI) enable researchers to design tailored and personalized interventions and educational content to meet individual unmet needs [24]. Large language models (LLMs) are advanced AI systems that can understand and generate human-like text by training on vast amounts of data [25]. LLMs perform various language tasks, such as answering questions and translating languages. How questions are asked can significantly affect the performance of LLMs. This process, known as prompt engineering, is crucial for obtaining accurate and relevant responses from LLMs [26, 27]. While LLMs have demonstrated remarkable potential in cancer research [28-31], their efficacy in real-world scenarios, such as cancer care and education, which often require advanced levels of comprehension, have yet to be thoroughly assessed.

Recent advancements in LLMs, such as Generative Pre-trained Transformer (GPT)-4 and GPT-4 Turbo [32, 33], have demonstrated their exceptional proficiency in completing various tasks, including coding, design, and content summarization. Previous research [34, 35] indicates that LLMs can capture large volumes of text effectively, even without specialized domain knowledge. This ability highlights its sophistication in processing and understanding information across a broad spectrum of topics, and its potential to significantly aid in analyzing unstructured data in clinical environments (e.g. clinical notes) [34, 35]. However, there are several notable gaps in the current research knowledge. Firstly, while LLMs have demonstrated high levels of accuracy in understanding extensive texts [34, 36, 37], even minor inaccuracies can have detrimental effects on patient outcomes [38], particularly regarding actionable advice. Therefore, the content they generate still necessitates additional expert verification to ensure it is error-free and ready to be presented to patients and their caregivers. Secondly, although previous research [36, 37] has demonstrated promising results in content summarization, these LLMs are often not applied in clinical environments, or they specifically address cancer care and education among disadvantaged groups [39]. Lastly, most educational resources for cancer care are available exclusively in English, which can create comprehension challenges for non-English speakers (e.g., Hispanic individuals, immigrants). Also, cancer survivors and their caregivers, already overwhelmed by treatment, often lack the time to read lengthy content. Therefore, it is essential to provide educational materials in multiple languages and in concise content to ensure effective communication and education [40].

To address these gaps, our team aimed to evaluate how LLMs perform in tailoring educational materials to enhance accessibility and comprehension for disadvantaged cancer survivors and their caregivers. The study aimed to: 1) evaluate the overall performance of LLMs in generating tailored educational content based on criteria such as word limit, reading level, and quality assessment (i.e., clarity, accuracy, relevance, completeness, and comprehensibility) for

disadvantaged cancer survivors and their caregivers; 2) compare the performances of three GPT models (i.e., GPT-3.5 Turbo, GPT-4, GPT-4 Turbo); and 3) explore different prompts that can help LLMs generate better content. By tailoring content to be concise and at a lower reading level, and translating it into multiple languages, individuals with low health literacy and racial and ethnic minorities facing language barriers can better understand cancer care educational materials. This approach helps them manage their symptoms more effectively, thereby reducing health disparities and promoting health equity.

Methods

Design

This study utilized three GPT models (GPT-3.5 Turbo, GPT-4, and GPT-4 Turbo) to generate content tailored for disadvantaged cancer survivors and their caregivers utilizing pre-determined prompts. The GPT-generated content was then evaluated by oncology experts.

Prompt Engineering

To promote the accessibility and comprehension of educational materials for disadvantaged cancer survivors and their caregivers with limited health literacy and language barriers, we structured prompts to have LLMs produce content at a low reading level, maintain a word limit of 250, and provide Spanish and Chinese translations for each topic, as described below [41].

The Flesch-Kincaid Grade Level (FKG) system [42] was used to assess the readability of content produced by the LLMs. The FKG level is a readability test designed to indicate how difficult a text is to understand. It calculates the grade level required for someone to comprehend the text. The FKG is based on word length and sentence length, providing a numerical score that corresponds to U.S. grade levels [42]. The National Institutes of Health (NIH) and the American Medical Association (AMA) suggest that patient education materials should be written at a reading level no higher than the sixth grade [43]. This recommendation is in place to guarantee that the information is reachable by a broad spectrum of individuals, encompassing those with limited health literacy. Therefore, our research targets an FKG level of 6 to align with this guidance.

We set a 250-word limit for our educational materials, recognizing that cancer survivors and their caregivers are frequently preoccupied with treatment schedules and daily responsibilities, leaving them limited time for reading [44]. This word limit is designed to ensure that participants can complete the reading within five minutes, making the task both manageable and feasible within their schedules.

Furthermore, it is shown that prompts exert a considerable impact on the responses generated by LLMs [45]. Therefore, we compared different prompts, including both textual and bulleted formats, to determine which approach yields better results.

Prompt engineering

We assembled a panel of nine oncology experts, comprising four oncology professors, four doctoral students, and one medical resident. Among them, all are fluent in English, with four experts proficient in Chinese and one proficient in Spanish. Each response generated by the LLMs was evaluated by at least two experts to ensure a comprehensive assessment, except for the Spanish translation task, which was evaluated by a single expert. The panel conducted

several Zoom meetings: the initial meeting provided training on content evaluation, and three additional meetings were held to discuss the results and feedback. Each expert was assigned 10 topics to evaluate and was required to provide feedback on the errors committed by the LLMs. These experts critically reviewed and annotated the LLM-generated content using a web-based Cohort Adjudication and Data Annotation (CADA) application [34] (see Figure 1) developed by our team.

Data Sources

Our primary sources for content generation were cancer survivors and caregiver education materials from the National Cancer Institute and the National Comprehensive Cancer Network guidelines [46, 47]. We selected 30 distinct topics covering a range of content such as fatigue, depression, anxiety, pain, cognitive impairment, nutrition, physical activity, healthy lifestyle, family communication, coping skills, and more. The selection of topics was informed by insights from our prior qualitative interviews with cancer survivors and their caregivers [48] and an extensive review of the literature [49-51]. We identified the key areas of interest and specific needs of disadvantaged cancer survivors and their caregivers, resulting in these 30 topics.

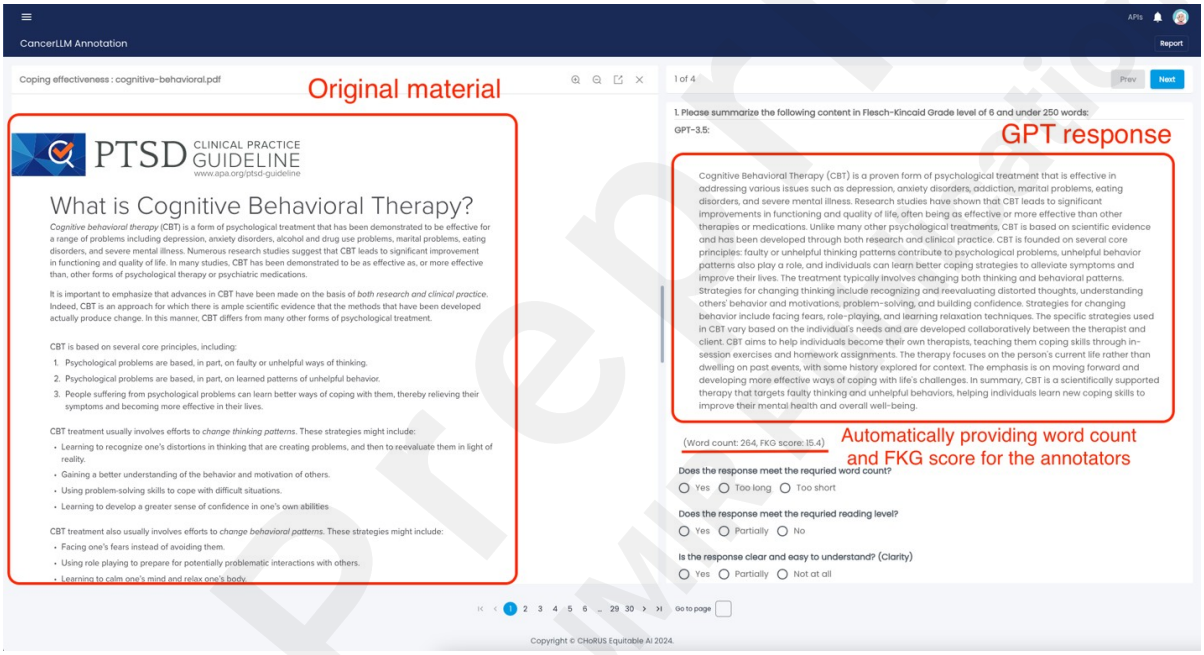

Figure 1. A screenshot of Cohort Adjudication and Data Annotation (CADA) Application

Appraisal Criteria

Based on a prior study of evaluating responses from LLMs [34], we formulated a set of multidimensional criteria to thoroughly assess the performance of LLMs, which include: 1) adherence to a word limit of 250 words; 2) achieving a reading level as per the FKG of below 6; and 3) quality assessment: *a. clarity* (i.e., ease of understanding in the response); *b. accuracy* (i.e., the response does not contain errors, like medical or language errors, that could negatively impact patients and their caregivers); *c. relevance* (i.e., the response is fully grounded in the materials we provided); *d. completeness* (i.e., the response encompasses all critical points from the materials); *e. comprehensibility* (i.e., the response is understandable that readers can apply it to their daily routine).

In terms of word limit, “yes” refers to a word limit within 250 words, and “no” refers to a word

limit of more than 250 words. The reading level was evaluated using “yes” for an FKG ≤ 6; “partial” for a 6<FKG ≤8; and “no” for an FKG>8). The FKG level was calculated by the Python package Textstat (version 0.7.3). For the quality assessment criteria, we implemented a scoring system in which evaluations were quantified based on their alignment with the expected outcomes. A score of 2 was assigned for "yes" evaluations, indicating full compliance; a score of 1 was given for "partially" evaluations, reflecting partial compliance; and a score of 0 was allocated for "no" evaluations, indicating non-compliance. The quality assessment included five criteria (*a* - *e*), each contributing a maximum of 2 points, for a total possible score of 10. The overall quality assessment ranged from 0 to 10, with 0 representing the absence or lowest quality and 10 indicating the highest quality.

Data Analysis

Descriptive analyses were conducted to determine the frequencies, percentages (for word limit, reading levels, and translations), mean and standard deviations (for quality scores) of major variables. Quality scores were determined by calculating the mean scores for each criterion and then obtaining the overall scores through their summation. To compare the differences in each model or prompt, we utilized ANOVA or Chi-square tests, as applicable. Values of $p < 0.05$ were considered to indicate a significant level. All analyses were conducted using Python statistical packages.

Ethical Considerations

The study protocol (STUDY00004750) was approved by the Institutional Review Board at Emory University. Oral consent was obtained from nine oncology experts since no protected health information (PHI) was collected from them. The study was conducted in accordance with the U.S. Common Rule (45 CFR 46) [52].

Results

Overall Performance of LLMs

In this study, 360 annotation values were collected from nine experts. Overall, LLMs have shown excellent performance in tailoring content based on our criteria (see Table 1). For word limit, 74.2% of the responses were within the word limit (less than 250 words) set for the task. The result indicates the excellent ability of LLMs to produce responses that adhere to specified word limit requirements. Regarding reading levels, LLMs demonstrated moderate performance, with 29.2% of the responses fully meeting the specified FKG level (FKG ≤ 6), 29.7% partially satisfactory (FKG = 6-8), and 41.1% not aligning with the provided FKG level (FKG > 8).

As shown in Table 2, LLMs demonstrated consistently high average scores across all quality criteria (total score: 8.933 out of 10). The highest average score achieved was 1.91 on relevance (Figure 2), highlighting the LLMs' ability to generate content that was highly pertinent to the given prompts. The lowest average score observed was 1.58 out of 2 in the category of completeness, indicating a moderate adherence to providing responses that capture all key points. In the translation tasks, the LLMs demonstrated high performance, with an accuracy of 88.9% for Spanish and 81.1% for Chinese translations, respectively.

Three GPT Models Comparisons: GPT-3.5 Turbo, GPT-4, and GPT-4 Turbo

GPT-4 demonstrated a superior capability in adhering to the specified word limit, with 84.2% of its responses falling within 250 words. In contrast, GPT-3.5 Turbo and GPT-4 Turbo exhibited a

relatively lower proficiency, with 71.7% and 66.7% of their responses meeting the word limit respectively. When comparing the models based on word limit, the Chi-square test did not demonstrate a highly significant difference (p=0.06) but indicated a trend of differences in word count (Table 2).

Regarding the assessment of reading level, GPT-4 Turbo met the required FKG level of 6 in 42.5% of cases, nearly doubling the performance of the other two models: 21.7% for GPT-4 and 22.5% for GPT-3.5 Turbo. The result indicated significant discrepancies among the models in adherence to the specified reading level ($p<<0.05$), with GPT-4 Turbo performing better compared to the other two models.

In terms of quality assessment, each of the LLMs attained a high score exceeding 8.8 out of 10, with GPT-4 and GPT-4 Turbo achieving 8.992 and 8.983 respectively, and GPT-3.5 Turbo trailing slightly at 8.825. Upon evaluation of each criterion, the performance of all models was found to be similar (Figure 2a). The application of ANOVA tests to each criterion revealed no significant differences among the three models (p=0.572).

In the translation tasks, GPT-4 Turbo exhibits perfect accuracy with a 100% success rate in Spanish translation, whereas GPT-4 and GPT-3.5 Turbo exhibited slightly lower, yet commendable success rates of 96.7% and 93.3% respectively. For the Chinese translation task, GPT-4 outperformed the other models with an accuracy of 86.7%. In contrast, GPT-3.5 Turbo and GPT-4 achieved 76.7% and 80.0% success rates respectively. The three models did not show a significant difference in the translation task (p= 0.481).

Table 1. Performance of all models, prompts on the summarization task

	GPT-3.5 Turbo		**GPT-4**		**GPT-4 Turbo**	
Prompt	Textual Format	Bullet Points	Textual Format	Bullet Points	Textual Format	Bullet Points
Word Limit (%)	0.467	0.967	0.917	0.767	0.517	0.817
Reading Level (%)	0.183	0.283	0.217	0.217	0.533	0.317
Accuracy	1.767±0.500	1.783±0.49	1.800±0.480	1.733±0.634	1.800±0.48	1.767±0.563
Clarity	1.833±0.418	1.750±0.474	1.867±0.389	1.800±0.403	1.883±0.324	1.717±0.49
Relevance	1.883±0.415	1.900±0.303	1.883±0.372	1.967±0.181	1.900±0.303	1.950±0.22
Completeness	1.533±0.623	1.583±0.645	1.483±0.624	1.667±0.601	1.583±0.619	1.650±0.547
Comprehensibility	1.817±0.469	1.800±0.403	1.883±0.324	1.900±0.303	1.900±0.303	1.817±0.39
Total Score	8.833±1.748	8.817±1.546	8.917±1.239	9.067±1.26	9.067±1.087	8.900±1.298
Spanish Translation (%)	0.933		0.967		1	
Chinese Translation (%)	0.767		0.867		0.800	

Table 2. Statistical analysis results

Group	*Criterion*	**PR(>F)**	**χ^2**

Models	Accuracy	0.970	
	Clarity	0.721	
	Relevance	0.630	
	Completeness	0.748	
	Comprehensibility	0.215	
	Total Score	0.572	
	Word Limit		10.178
	Reading Level		35.468
	Spanish Translation		2.069
	Chinese Translation		1.015
	Translation		1.463
Prompts	Accuracy	0.213	
	Clarity	0.028	
	Relevance	0.177	
	Completeness	0.154	
	Comprehensibility	0.149	
	Total Score	0.939	

Two Different Prompt Comparisons: Textual and Bulleted Formats

We compared two prompting methods in terms of word limits, reading level, and quality assessment. The major difference noted in the comparison of the two prompts was that responses generated from prompt 2 (bulleted format) were superior in adhering to the target word limit. Specifically, 85.0% of the responses from prompt 2 successfully achieved the word limit, in contrast to 63.3% of the responses from prompt 1 (textual format) that fully satisfied the word limit. Utilizing prompt 1 resulted in only 31.1% of responses meeting our desired reading level, with a slight decrease to 30.0% for prompt 2. For the five quality criteria, both prompts achieved a similarly high score of 8.939 and 8.928 respectively (Figure 2b). Upon performing an ANOVA test to assess the differences in performance between the two prompts, it was found that the variations between them were not significant (p=0.939).

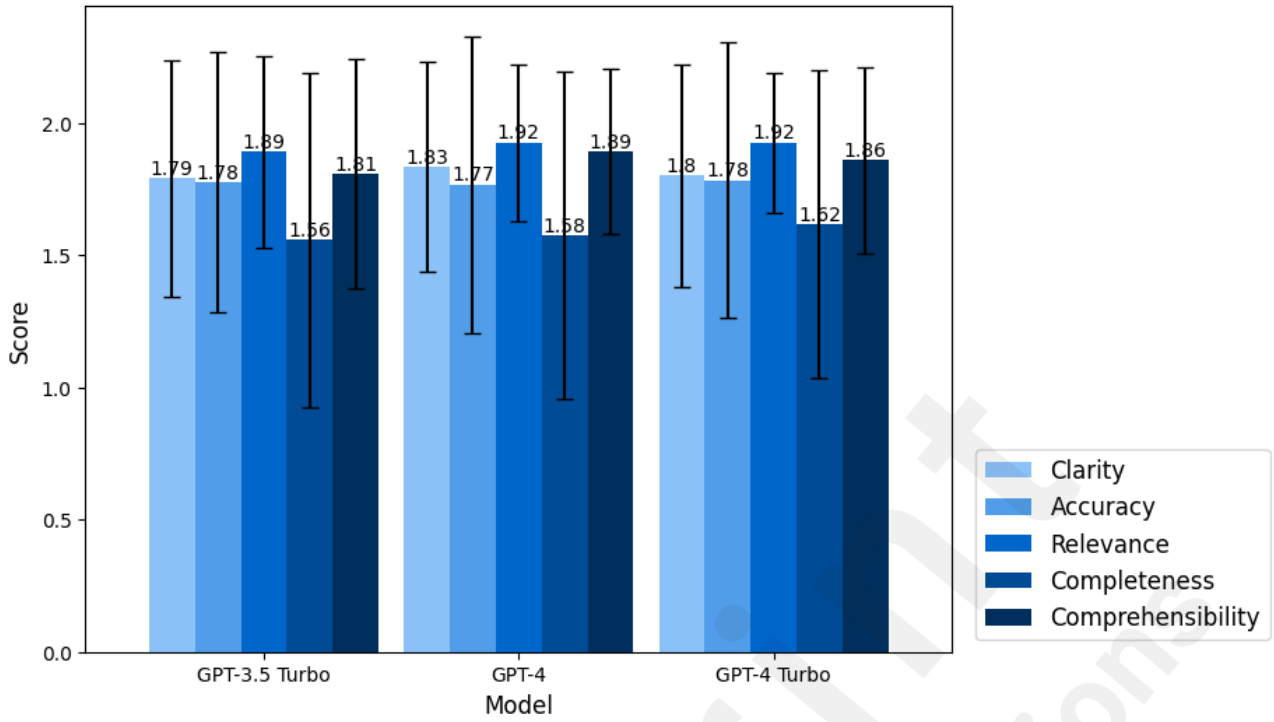

Figure 2a. Assessment scores on each criterion between different models

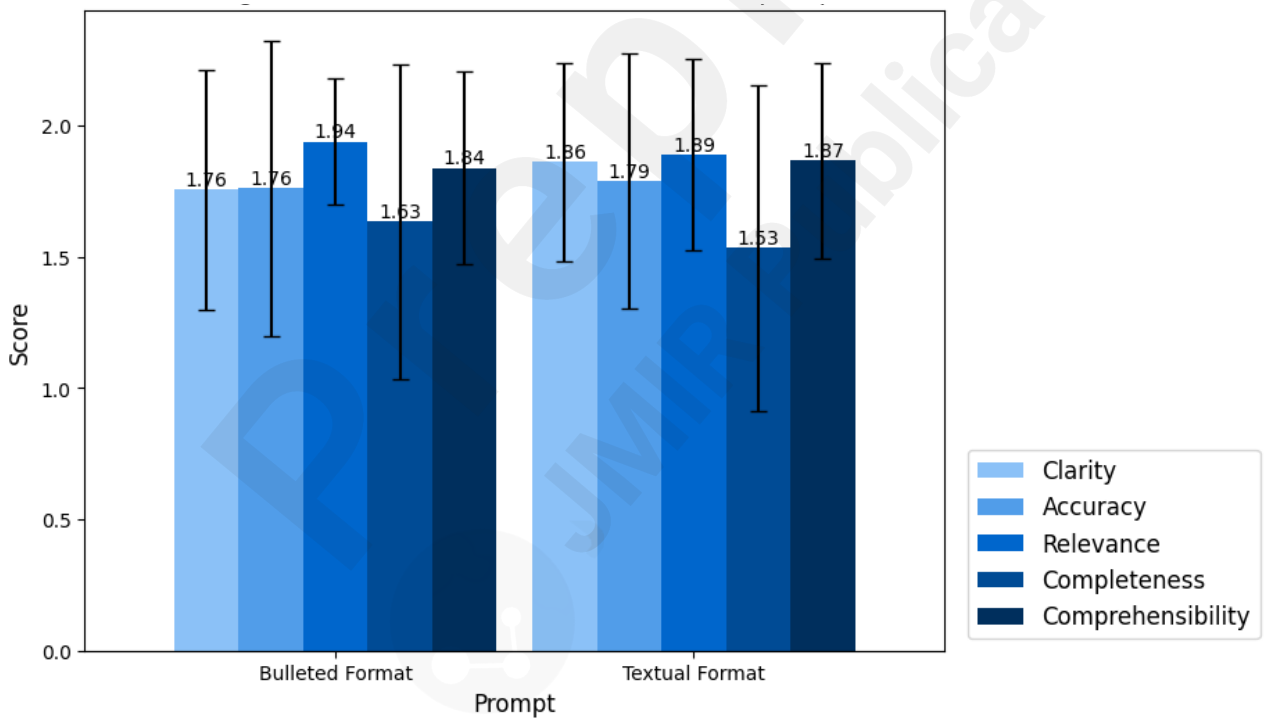

Figure 2b. Assessment scores on each criterion between different prompts

Errors Evaluation

The errors that LLMs committed were categorized into inaccurate scope, inaccurate definition, inaccurate expression, and meaningless points. Some examples are shown in Table 3.

A common error observed with LLMs is their tendency to integrate their own knowledge and interpretation rather than adhering strictly to the provided materials, such as an inaccurate scope. For instance, when the text specified "*to limit red meat*." in the *Nutrition* topic, GPT-3.5 Turbo inaccurately generalized this advice to *"limiting animal-based food."* This interpretation is not entirely correct, as animal-based food encompasses more than just red meat, including white meat such as chicken, which the original material did not intend to restrict.

Other observed errors involve inaccurate expressions. For instance, in *Sexual health issues in men with cancer* topic, the original content suggested, *"It is probably still important to maintain intimacy with a partner."* However, GPT-3.5 Turbo revised this to *"it is still important to maintain intimacy with a partner."* This alteration results in a tone that may seem judgmental, deviating from the original's more tentative stance.

Table 3. Error cases

Model	**Topic**	**Output**	**Error Type**	**Reason**
GPT-3.5 Turbo	Nutrition	“It advises limiting animal-based food, processed food, and alcohol consumption.”	Inaccurate scope	The chapter only mentions to limit red meat, not all animal-based foods (says it can make up half or less of diet).
GPT-3.5 Turbo	Sexual Health Issues in Men with Cancer	"It is still important to maintain intimacy with a partner."	Inaccurate expression	The tailored content sounds a little judgmental whereas the original document says, "probably still important" and is less assuming.
GPT-4	Mindfulness	“These practices involve focusing the mind on present sensations, such as breathing, a sound, or an image.”	Inaccurate definition	It seems to define meditation and mindfulness in one overarching definition, which only defines meditation. The model merged definitions of MBSR and MBCT together and did not include difference between types.
GPT-4 Turbo	Making a Difference	"Learning: Educating yourself about cancer can empower you to assist others. Resources are available online, by phone, and in print."	Meaningless point	The customized content falls short in terms of actionability. The purpose of tailoring content is to educate patients and caregivers, rather than expecting them to educate themselves.

An example of inaccurate definition was identified within the *Mindfulness* topic, where GPT-4 defined meditation and mindfulness in one overarching definition for meditation. It also merged definitions of mindfulness-based stress reduction and mindfulness-based cognitive therapy without highlighting differences between the mindfulness interventions.

The last common issue is that LLMs may include information that, while accurate, might not be actionable for patients. For instance, in *Making a difference* topic, GPT-4 Turbo correctly sourced from the material *that "Learning: Educating yourself about cancer can empower you to assist others. Resources are available online, by phone, and in print."* However, this information becomes less useful in the absence of specific links or directions that could guide patients on where to start their education.

Discussion

To our knowledge, this is the first study to evaluate the capability of LLMs in tailoring educational content for disadvantaged cancer survivors and their caregivers. In our study, all three LLMs have demonstrated overall excellent performance in most criteria. The more advanced GPT-4 family models showed better overall performance compared to GPT-3.5 Turbo. GPT-4's high adherence to word limits and GPT-4 Turbo's better compliance to reading level compliance proved their ability to meet our requirements when tailoring content. Prompting GPTs to produce bulleted-format content is likely to result in better educational materials compared to textual-format content. All models exhibit strong capability in generating highly relevant content. However, they fall short in terms of completeness. Overall, it is proven that LLMs are highly effective in tailoring, condensing, and translating educational content for disadvantaged cancer survivors and their caregivers. These findings inform future versions of LLMs to focus more on the reading level and completeness of their output and the development of tailored intervention materials for disadvantaged cancer survivors and their caregivers. These promising results also indicate that LLMs can be a valuable tool in making educational content more accessible and comprehensible to diverse patient populations.

The capabilities of LLMs in text analysis have been well studied. For example, our prior study [34] examined the potential of LLMs to categorize clinical concepts from patient notes. Yet, this study focused solely on the LLMs' comprehension of patients' conditions from clinical notes rather than educational materials. Veen et al.'s study [53] assessed approaches for LLMs to summarize clinical texts. Although it demonstrated overall preferred performance, especially GPT-4, over human experts, the study was limited to the summarization of radiology report findings and confined to three attributes: completeness, correctness, and conciseness, whereas our study expanded on this topic by evaluating LLMs against seven distinct criteria. Furthermore, none of the existing studies focus on education regarding supportive care in cancer, whereas our innovative findings make a significant contribution to the literature in this field.

Despite the excellence of LLMs in adhering to specified word limits and generating high-quality content, several challenges remain. One notable area where LLMs struggle is in adjusting the reading level of the content to accommodate patients from various educational levels. The content tailored by LLMs often does not meet the intended FKG level. This oversight implies that some individuals might find the content overly complex, potentially hindering their understanding of health information and educational materials [54, 55]. Addressing this challenge is essential for maximizing the applicability of LLMs and ensuring that all cancer survivors receive the support they need to manage their cancer effectively.

It is also observed that the accuracy of Spanish translations is significantly higher than that of Chinese translations. This finding is expected, given the abundance of Spanish content available online compared to Chinese content that can serve as training materials. Previous studies [56, 57] have shown that LLMs' performance in different languages has a clear correlation with the proportion of each language in the pre-training corpus. Without fine-tuning, LLMs have a much higher performance in high-resource languages like German, French, and Spanish, and a significantly lower performance in low-resource languages like Kannada, Occitan, and Western Frisian [56, 57].

The educational content errors could be detrimental to cancer survivors and their caregivers

by providing false physical activity, diet, or medication suggestions. Therefore, content produced by LLMs should undergo thorough evaluation and validation before the content is utilized in a clinical setting [38, 58, 59]. Our analysis has identified multiple errors in the outputs from LLMs, including inaccuracies in scope, expression, and definition. These types of errors can lead to the dissemination of misinformation, potentially causing harm to patients [60]. Therefore, such inaccuracies must be identified, analyzed, and rectified to prevent any negative impacts on patient care. Our study also detected some meaningless points that were not actionable in LLMs' outputs, which could increase the reading burden on patients and their caregivers. Recommendations should highlight actionable information for disadvantaged cancer survivors and their caregivers to reduce the burden of reading educational materials, emphasizing the need for LLMs to prioritize the utility and applicability of the information they present. Additionally, education content should be evaluated and validated by content experts before the it is available to cancer survivors and their caregivers.

In addition, both Xiao's and Asthana's studies [36, 37] evaluated the performance of fine-tuned LLMs in non-clinical environments. Their results highlighted the significant potential of LLMs in summarizing general text through the adoption of advanced fine-tuning techniques. It is possible that fine-tuning could further improve LLMs' capacity to analyze educational materials specifically tailored for groups such as disadvantaged cancer survivors and their caregivers. With this additional data, more advanced fine-tuning techniques such as instruction tuning [57, 61, 62] and parameter-efficient fine-tuning [63] can be implemented, and are likely to further enhance the performance.

Limitations

While the study has shown promising results, it has several limitations. Firstly, the dataset size remains relatively small, which could restrict the generalizability of the findings to broader topics. Secondly, we lacked participant assessment. Relying solely on oncology experts to evaluate the outputs from LLMs might create obstacles when applying these findings to actual cancer patients and their caregivers. While our oncology experts deeply value caring for disadvantaged populations, it's important to note that they are highly educated and might have unintentional biases. This could make it challenging for them to view educational content from the perspective of individuals with low health education and literacy. Therefore, future studies can be broadened to include a wider range of educational topics and additional annotations from cancer patients and their caregivers. Thirdly, this study was limited to zero-shot learning because of the lack of training data. It could be expanded by collecting tailored content from human experts to serve as training data. Finally, due to a limited number of annotators from diverse backgrounds, our study was only able to evaluate translations in two languages. Our analysis suggests that translation performance can vary between languages, influenced by the availability of content in each language. It is important to note that these findings may not be generalizable to languages spoken by smaller populations, where content availability and linguistic nuances could further affect translation accuracy.

Conclusions

The study highlights the application of LLMs in cancer care while being cognizant of their potential limitations. All three LLMs have demonstrated overall high capability in tailoring educational content for disadvantaged cancer survivors and their caregivers. GPT-4 family models showed better overall performance compared to GPT-3.5 Turbo. Prompting GPTs to produce bulleted-format content can generate better educational content. The findings from this study inform the intervention development and implementation in cancer symptom

management and health equity. Additional studies are warranted to expedite the integration of AI-driven solutions into clinical settings.

Acknowledgements

The study was supported by the Oncology Nursing Foundation Research Grant (2022 RE03). The authors also appreciate the support from Emory University School of Nursing and Winship Cancer Institute.

Conflicts of Interest

The authors have no conflicts of interest to disclose.

Abbreviations

LLM: Large Language Model
GPT: Generative Pre-trained Transformer

Data Availability

The datasets generated and analyzed during the current study are available upon reasonable request. The corresponding author (YL) will coordinate requests for data and maintain documentation for requests and distributions. Emory University has an established Institutional Data Use Agreement that can easily be adapted and deployed.

References

1. Society, A.C. *Statistics and Graphs*. 2023 [cited 2024 March 21]; Available from: https://cancercontrol.cancer.gov/ocs/statistics.
2. Lin, Y., et al., *Distinct morning and evening fatigue profiles in gastrointestinal cancer during chemotherapy.* BMJ Support Palliat Care, 2021.
3. Lin, Y., et al., *Distinct sleep disturbance profiles in patients with gastrointestinal cancers receiving chemotherapy.* Cancer Nurs, 2021.
4. Lin, Y., et al., *Distinct profiles of multiple co-occurring symptoms in patients with gastrointestinal cancers receiving chemotherapy.* Support Care Cancer, 2021. **29**(8): p. 4461-4471.
5. Lin, Y., et al., *Distinct Co-occurring Morning and Evening Fatigue Profiles in Patients With Gastrointestinal Cancers Receiving Chemotherapy.* Cancer Nurs, 2022.
6. Lin, Y., et al., *A network analysis of self-reported psychoneurological symptoms in patients with head and neck cancer undergoing intensity-modulated radiotherapy.* Cancer, 2022. **128**(20): p. 3734-3743.
7. Lin, Y., et al., *Associations of differentially expressed genes with psychoneurological symptoms in patients with head and neck cancer: A longitudinal study.* J Psychosom Res, 2023. **175**: p. 111518.
8. Dodd, M.J., C. Miaskowski, and S.M. Paul, *Symptom clusters and their effect on the functional status of patients with cancer.* Oncol Nurs Forum, 2001. **28**(3): p. 465-70.
9. Tantoy, I.Y., et al., *Quality of life of patients with gastrointestinal cancers undergoing chemotherapy.* Qual Life Res, 2018. **27**(7): p. 1865-1876.
10. Lin, Y., et al., *Symptom experience and self-management for multiple co-occurring*

symptoms in patients with gastric cancer: A qualitative study. Eur J Oncol Nurs, 2020. **49**: p. 101860.
11. Lin, Y., et al., *Common and co-occurring symptoms experienced by patients with gastric cancer.* Oncol Nurs Forum, 2020. **47**(2): p. 187-202.
12. Cal, A., I.A. Avci, and F. Cavusoglu, *Experiences of Caregivers with Spouses Receiving Chemotherapy for Colorectal Cancer and their Expectations from Nursing Services.* Asia Pac J Oncol Nurs, 2017. **4**(2): p. 173-179.
13. Law, E., et al., *The "sphere of care": A qualitative study of colorectal cancer patient and caregiver experiences of support within the cancer treatment setting.* PLoS One, 2018. **13**(12): p. e0209436.
14. Lin, Y., et al., *A Web-Based Dyadic Intervention to Manage Psychoneurological Symptoms for Patients With Colorectal Cancer and Their Caregivers: Protocol for a Mixed Methods Study.* JMIR Res Protoc, 2023. **12**: p. e48499.
15. Howard, A.F., et al., *At the Heart of It All: Emotions of Consequence for the Conceptualization of Caregiver-Reported Outcomes in the Context of Colorectal Cancer.* Curr Oncol, 2021. **28**(5): p. 4184-4202.
16. Hollis, R.H. and D.I. Chu, *Healthcare Disparities and Colorectal Cancer.* Surg Oncol Clin N Am, 2022. **31**(2): p. 157-169.
17. Levy, H. and A. Janke, *Health Literacy and Access to Care.* J Health Commun, 2016. **21 Suppl 1**(Suppl): p. 43-50.
18. Murphy-Ende, K., *Barriers to Palliative and Supportive Care.* Nursing Clinics of North America, 2001. **36**(4): p. 843-853.
19. Sherman, A.D.F., et al., *Intersectionality in nursing research: A scoping review.* Int J Nurs Stud Adv, 2023. **5**.
20. *Serving Vulnerable and Underserved Populations*. [cited 2024 May 20]; Available from: https://www.hhs.gov/guidance/sites/default/files/hhs-guidance-documents/006_Serving_Vulnerable_and_Underserved_Populations.pdf.
21. Vogels, E.A. *Digital divide persists even as Americans with lower incomes make gains in tech adoption*. 2021 [cited 2023 August 30]; Available from: https://www.pewresearch.org/short-reads/2021/06/22/digital-divide-persists-even-as-americans-with-lower-incomes-make-gains-in-tech-adoption/.
22. Qan'ir, Y. and L. Song, *Systematic review of technology-based interventions to improve anxiety, depression, and health-related quality of life among patients with prostate cancer.* Psychooncology, 2019. **28**(8): p. 1601-1613.
23. Song, L., et al., *Improving couples' quality of life through a Web-based prostate cancer education intervention.* Oncol Nurs Forum, 2015. **42**(2): p. 183-92.
24. Xue, V.W., P. Lei, and W.C. Cho, *The potential impact of ChatGPT in clinical and translational medicine.* Clin Transl Med, 2023. **13**(3): p. e1216.
25. Strachan, J.W.A., et al., *Testing theory of mind in large language models and humans.* Nat Hum Behav, 2024.
26. Wang, J., et al., *Review of large vision models and visual prompt engineering.* Meta-Radiology, 2023: p. 100047.
27. Wang, J., et al., *Prompt engineering for healthcare: Methodologies and applications.* arXiv preprint arXiv:2304.14670, 2023.
28. Cascella, M., et al., *Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios.* J Med Syst, 2023. **47**(1): p. 33.
29. Fink, M.A., et al., *Potential of ChatGPT and GPT-4 for data mining of free-text CT reports on lung cancer.* Radiology, 2023. **308**(3): p. e231362.
30. Haver, H.L., et al., *Use of ChatGPT, GPT-4, and Bard to improve readability of ChatGPT's*

answers to common questions about lung cancer and lung cancer screening. American Journal of Roentgenology, 2023. **221**(5): p. 701-704.
31. Dai, W., et al., *Systematic analysis of glutamine metabolism family genes and exploration of the biological role of GPT in colorectal cancer.* Aging (Albany NY), 2023. **15**(21): p. 11811.
32. Achiam, J., et al., *Gpt-4 technical report.* arXiv preprint arXiv:2303.08774, 2023.
33. Bhattarai, K., et al., *Leveraging GPT-4 for Identifying Clinical Phenotypes in Electronic Health Records: A Performance Comparison between GPT-4, GPT-3.5-turbo and spaCy's Rule-based & Machine Learning-based methods.* bioRxiv, 2023: p. 2023.09. 27.559788.
34. Liu, D., et al., *Evaluation of General Large Language Models in Contextually Assessing Semantic Concepts Extracted from Adult Critical Care Electronic Health Record Notes.* arXiv preprint arXiv:2401.13588, 2024.
35. Agrawal, M., et al., *Large language models are zero-shot clinical information extractors.* arXiv preprint arXiv:2205.12689, 2022.
36. Xiao, L. and X. Chen, *Enhancing llm with evolutionary fine tuning for news summary generation.* arXiv preprint arXiv:2307.02839, 2023.
37. Asthana, S., et al., *Summaries, Highlights, and Action items: Design, implementation and evaluation of an LLM-powered meeting recap system.* arXiv preprint arXiv:2307.15793, 2023.
38. Garrouste-Orgeas, M., et al., *Overview of medical errors and adverse events.* Annals of intensive care, 2012. **2**: p. 1-9.
39. Gero, Z., et al., *Attribute Structuring Improves LLM-Based Evaluation of Clinical Text Summaries.* arXiv preprint arXiv:2403.01002, 2024.
40. Espenshade, T.J. and H. Fu, *An analysis of English-language proficiency among US immigrants.* American Sociological Review, 1997: p. 288-305.
41. Chesser, A., et al., *Navigating the digital divide: A systematic review of eHealth literacy in underserved populations in the United States.* Inform Health Soc Care, 2016. **41**(1): p. 1-19.
42. Solnyshkina, M., et al., *Evaluating text complexity and Flesch-Kincaid grade level.* Journal of social studies education research, 2017. **8**(3): p. 238-248.
43. Walsh, T.M. and T.A. Volsko, *Readability assessment of internet-based consumer health information.* Respir Care, 2008. **53**(10): p. 1310-5.
44. Abazari, A., S. Chatterjee, and M. Moniruzzaman, *Understanding Cancer Caregiving and Predicting Burden: An Analytics and Machine Learning Approach.* AMIA Annu Symp Proc, 2023. **2023**: p. 243-252.
45. Wang, L., et al., *Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs.* NPJ Digit Med, 2024. **7**(1): p. 41.
46. *Patient Education Publications*. [cited 2023 August 31]; Available from: https://www.cancer.gov/publications/patient-education.
47. *NCCN Patient Guidelines for Patients*. [cited 2023 August 30]; Available from: https://www.nccn.org/patientresources/patient-resources/guidelines-for-patients.
48. Epari, A., et al. *Perceptions and needs for a technology-based dyadic intervention to manage psychoneurological symptoms in underserved patients with colorectal cancer and their caregivers: A qualitative study*. in *Oncology Nursing Society Annual Congress*. 2024. Washington, D.C.
49. Northouse, L., et al., *Effects of a family intervention on the quality of life of women with recurrent breast cancer and their family caregivers.* Psychooncology, 2005. **14**(6): p. 478-91.
50. Northouse, L.L., et al., *Randomized clinical trial of a brief and extensive dyadic*

intervention for advanced cancer patients and their family caregivers. Psychooncology, 2013. **22**(3): p. 555-63.
51. Northouse, L.L., et al., *Randomized clinical trial of a family intervention for prostate cancer patients and their spouses.* Cancer, 2007. **110**(12): p. 2809-18.
52. *Code of Federal Regulations, Part 46 - Protection of Human Subjects*. 2024 [cited 2024 May 26]; Available from: https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-A/part-46.
53. Van Veen, D., et al., *Clinical text summarization: Adapting large language models can outperform human experts.* Research Square, 2023.
54. Delaney, F.T., et al., *Readability of patient education materials related to radiation safety: What are the implications for patient-centred radiology care?* Insights into Imaging, 2021. **12**: p. 1-9.
55. Davis, T.C., et al., *Reading ability of parents compared with reading level of pediatric patient education materials.* Pediatrics, 1994. **93**(3): p. 460-468.
56. Li, Z., et al., *Quantifying Multilingual Performance of Large Language Models Across Languages.* arXiv preprint arXiv:2404.11553, 2024.
57. Zan, C., et al., *Building Accurate Translation-Tailored LLMs with Language Aware Instruction Tuning.* arXiv preprint arXiv:2403.14399, 2024.
58. Kalra, J., *Medical errors: impact on clinical laboratories and other critical areas.* Clinical biochemistry, 2004. **37**(12): p. 1052-1062.
59. Robertson, J.J. and B. Long, *Suffering in silence: medical error and its impact on health care providers.* The Journal of emergency medicine, 2018. **54**(4): p. 402-409.
60. Burton-Wood, C., et al., *Medical professionals'(mis) remembering of a simulated interaction with a patient: A medical misinformation effect.* 2019.
61. Zhang, S., et al., *Instruction tuning for large language models: A survey.* arXiv preprint arXiv:2308.10792, 2023.
62. Bai, Y., et al., *COIG-CQIA: Quality is All You Need for Chinese Instruction Fine-tuning.* arXiv preprint arXiv:2403.18058, 2024.
63. Ding, N., et al., *Parameter-efficient fine-tuning of large-scale pre-trained language models.* Nature Machine Intelligence, 2023. **5**(3): p. 220-235.