DOI: http://dx.doi.org/10.12996/gmj.2026.4626

Addressing Patient Questions on Spinal Muscular Atrophy: Performance of Large Language Models in a Genetic Context

Hastaların Spinal Musküler Atrofi Hakkındaki Soruları: Genetik Açıdan Büyük Dil Modellerinin Performansı

Dilsu Dicle Erkan[1], Mehmet Alikaşifoğlu[2]

[1]Clinic of Medical Genetics, University of Health Sciences Türkiye, Ankara Etlik City Hospital, Ankara, Türkiye
[2]Department of Medical Genetics, Hacettepe University, Faculty of Medicine, Ankara, Türkiye

ABSTRACT

Objective: Large language models (LLMs) are increasingly used by the public to obtain medical and genetic information. Given the genetic complexity and public health relevance of spinal muscular atrophy (SMA), this study aimed to evaluate the quality, readability, and actionability of LLM-generated responses to SMA-related frequently asked questions (FAQs).

Methods: Fifteen SMA-related FAQs were identified in Turkish using Google's "People Also Ask" feature and categorized into disease definition, genetic screening, and genetic diagnosis and treatment. Each question was submitted to the free versions of ChatGPT, Gemini, and DeepSeek. Responses were evaluated using the modified DISCERN instrument and a 5-point Likert scale for information quality; the Flesch–Kincaid reading ease and grade level for readability; and the Patient Education Materials Assessment Tool (PEMAT) for understandability and actionability.

Results: Median DISCERN scores were 3.00 across all LLMs, indicating moderate information quality, and there was no significant difference among models ($p = 0.069$). Readability differed significantly, with ChatGPT producing responses at a lower Flesch–Kincaid grade level than that of Gemini and DeepSeek ($p = 0.001$). PEMAT understandability and actionability scores varied by question category, with significant differences observed for questions on disease definition and genetic screening ($p < 0.05$).

Conclusion: LLMs generate SMA-related information with moderate quality and understandability; however, variability in readability, actionability, and topic-specific performance limits their suitability

ÖZ

Amaç: Büyük dil modelleri (LLM'ler), tıbbi ve genetik bilgiye erişim amacıyla toplum tarafından giderek daha fazla kullanılmaktadır. Spinal musküler atrofi (SMA) hastalığının anlaması zorlayıcı genetik terimleri ve halk sağlığı açısından önemi göz önüne alındığında, bu çalışmada LLM'ler tarafından üretilen SMA ile ilişkili sıkça sorulan sorulara (SSS) verilen yanıtların kalite, okunabilirlik ve uygulanabilirlik açısından değerlendirilmesi amaçlanmıştır.

Yöntemler: Google'ın "People Also Ask" özelliği kullanılarak 15 adet SMA ile ilişkili Türkçe SSS belirlenmiş ve bu sorular hastalığın tanımı, genetik tarama testleri ve genetik tanı ile tedavi olmak üzere üç başlık altında sınıflandırılmıştır. Her bir soru, ChatGPT, Gemini ve DeepSeek'in ücretsiz sürümleri kullanılarak sırayla sorulmuştur. Yanıtlar; bilgi kalitesi için modifiye DISCERN aracı ve 5 puanlı Likert ölçeği, okunabilirlik için Flesch–Kincaid Okunabilirlik Düzeyi ve Sınıf Seviyesi, anlaşılabilirlik ve uygulanabilirlik için ise Hasta Eğitim Materyalleri Değerlendirme Aracı (PEMAT) kullanılarak değerlendirilmiştir.

Bulgular: Tüm LLM'lerde ortanca DISCERN puanı 3,00 olarak saptanmış ve bu değer orta düzeyde bilgi kalitesine işaret etmiş olup modeller arasında anlamlı fark izlenmemiştir ($p = 0,069$). Okunabilirlik açısından anlamlı farklılıklar saptanmış ve ChatGPT'nin Gemini ve DeepSeek'e kıyasla daha düşük Flesch–Kincaid Sınıf Seviyesi ile yanıtlar ürettiği görülmüştür ($p = 0,001$). PEMAT anlaşılabilirlik ve uygulanabilirlik puanları soru kategorilerine göre farklılık göstermiş; özellikle hastalık tanımı ve genetik tarama sorularında anlamlı farklar saptanmıştır ($p < 0,05$).

Cite this article as: Erkan DD, Alikaşifoğlu M. Addressing patient questions on spinal muscular atrophy: performance of large language Models in a genetic context. Gazi Med J. 2026;37(2):230-237

Address for Correspondence/Yazışma Adresi: Dilsu Dicle Erkan, Clinic of Medical Genetics, University of Health Sciences Türkiye, Ankara Etlik City Hospital, Ankara, Türkiye
E-mail / E-posta: dilsudicle@gmail.com
ORCID ID: orcid.org/0000-0002-2400-8344

Received/Geliş Tarihi: 15.01.2026
Accepted/Kabul Tarihi: 01.03.2026
Publication Date/Yayınlanma Tarihi: 31.03.2026

©Copyright 2026 The Author(s). Published by Galenos Publishing House on behalf of Gazi University Faculty of Medicine. Licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND) International License.

©Telif Hakkı 2026 Yazar(lar). Gazi Üniversitesi Tıp Fakültesi adına Galenos Yayınevi tarafından yayımlanmaktadır. Creative Commons Atıf-GayriTicari-Türetilemez 4.0 (CC BY-NC-ND) Uluslararası Lisansı ile lisanslanmaktadır.

ABSTRACT

for standalone use in genetic counseling. While these tools may serve as supplementary educational resources, they should not replace clinician-led genetic counseling, particularly in contexts requiring individualized risk assessment and decision-making.

Keywords: Spinal muscular atrophy, genetic counseling, artificial intelligence, health information quality, large language models, patient education

ÖZ

Sonuç: LLM'ler SMA ile ilişkili bilgileri orta düzeyde kalite ve anlaşılabilirlik ile sunabilmektedir; ancak okunabilirlik, uygulanabilirlik ve konuya özgü performanstaki değişkenlik, bu yanıtların genetik danışmada tek başına kullanımını sınırlamaktadır. LLM'ler tarafından üretilen içerikler tamamlayıcı bir eğitim kaynağı olarak değerlendirilebilir; ancak bireyselleştirilmiş risk değerlendirmesi ve karar verme süreçleri gerektiren durumlarda klinisyen tarafından yürütülen genetik danışmanın yerini almamalıdır.

Anahtar Sözcükler: Spinal musküler atrofi, genetik danışma, yapay zeka, sağlık bilgisi kalitesi, büyük dil modelleri, hasta eğitimi

INTRODUCTION

Spinal muscular atrophy (SMA) is a severe autosomal recessive neuromuscular disorder characterized by progressive degeneration of anterior horn motor neurons (1). SMA can lead to symmetrical muscle weakness, respiratory insufficiency, and, in severe forms, early mortality. The disease is caused by biallelic pathogenic variants in the survival motor neuron-1 (*SMN1*) gene, resulting in a deficiency of the SMN protein. SMA represents one of the most common inherited causes of infant mortality, with an estimated global incidence of approximately 1 in 10,000 live births and a carrier frequency of about 1 in 40–60 individuals (2,3). Clinical severity spans a broad spectrum and is traditionally classified into phenotypic subtypes based on age of onset and achievement of motor milestones, ranging from severe infantile-onset disease to milder later-onset forms.

From a public health and genetic medicine perspective, SMA has a unique position due to the availability of effective disease-modifying therapies and benefits of early diagnosis (4). These developments have increased the importance of timely genetic counseling, carrier screening, and newborn screening programs (5). In Türkiye, SMA has become a major focus of both clinical practice and public awareness. Although the increased focus on SMA in Türkiye is partly related to the high prevalence of consanguineous marriages, population-based data indicate that the elevated SMA carrier frequency cannot be explained by consanguinity alone. Genetic counseling for SMA is further complicated by diagnostic challenges such as "2 + 0" silent carriers, who may be misclassified by standard copy-number testing, creating differences in genetic counseling before or after diagnostic tests rather than screening programs (5,6). Misunderstanding or oversimplification of these concepts may result in false reassurance, unnecessary anxiety, or suboptimal reproductive decision-making. These findings show the need for clear and accurate information when addressing genetically complex conditions, such as SMA.

In parallel with these, the way patients and families seek health information has undergone a substantial transformation (7). Large language models (LLMs) are now widely used by the public to obtain medical information, including explanations of genetic diseases, inheritance patterns, screening tests, and treatment options (7-9). Given the emotional impact of SMA, which often involves reproductive planning, newborn diagnosis, and therapeutic decisions, patients and families increasingly turn to LLMs for rapid, accessible explanations and guidance. In this context, LLM-generated information may function not only as general education but also as a supplement to medical and genetic counseling.

The expanding reliance on LLMs for health information raises critical concerns regarding the quality, accuracy, readability, and actionability of the information provided. Inaccurate or misleading responses about SMA may have significant consequences, particularly in settings where users seek guidance on carrier status, prenatal or preimplantation testing, newborn screening results, or emerging therapies. Despite the growing use of LLMs in other medical specialities, data evaluating the quality of LLM-generated responses specifically for medical genetics remain limited (10-13).

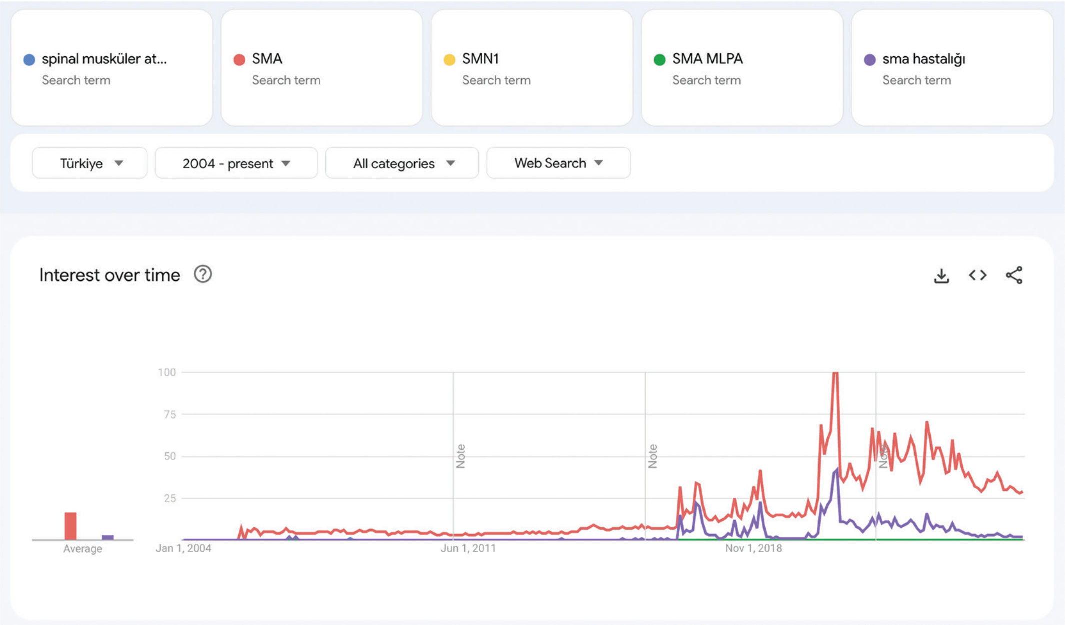
Therefore, assessing the information quality of LLMs in the context of SMA is of particular importance. This study aims to evaluate and compare the quality, readability, and usability of responses generated by the most commonly used LLMs in response to frequently asked questions (FAQs) about SMA.

MATERIALS AND METHODS

Ethical approval was not required for this study because it involved no human participants, patient data, or identifiable personal information. In this study we analysed publicly available online content and responses generated by artificial intelligence (AI) systems.

Question Identification and Categorization

FAQs related to SMA were identified using a structured Google search strategy. Searches were performed using the Google Chrome browser (version 143.0.7499.170) on November 16, 2025, in incognito mode to minimize personalization bias. Initially, common search terms were explored using Google searches, including "spinal muscular atrophy," "SMA," and their Turkish-language equivalents. Among these terms, "SMA" was the most frequently searched keyword and was therefore selected as the primary search term (Figure 1). To capture patient- and public-oriented information needs, the search term "SMA" was entered into the Google search tool (www.google.com.tr), and the questions listed under the "People Also Ask" section were reviewed to identify Turkish-language questions. The "People Also Ask" feature reflects queries that are most commonly searched by users, and has been widely used to represent real-world public interest and information-seeking behavior. A total of 15 non-repetitive FAQs were selected from this section for analysis (Table 1). These questions were chosen to reflect common online information needs of individuals seeking medical and genetic information on SMA.



spinal musküler at... Search term

SMA Search term

SMN1 Search term

SMA MLPA Search term

sma hastalığı Search term

Türkiye · 2004 - present · All categories · Web Search

Figure 1. Google Trends analysis showing relative search interest over time in Türkiye for spinal muscular atrophy–related terms, including "spinal musküler atrofi," "SMA," "*SMN1*," "SMA MLPA," and "SMA hastalığı," from 2004 to the present. Search interest is normalized on a scale from 0 to 100, with higher values indicating greater relative search volume. Among the evaluated terms, "SMA" demonstrated the highest and most sustained search interest, supporting its selection as the primary search term for identifying SMA-related frequently asked questions.

MLPA: Multiplex ligation-dependent probe amplification, SMA: Spinal muscular atrophy, *SMN1*: Survival motor neuron-1.

Table 1. Patient-oriented frequently asked questions on spinal muscular atrophy categorized by topic translated to English.

Groups	Questions
Definition of SMA	What does SMA disease mean?
	From whom is the *SMA* gene inherited?
	Is SMA caused by consanguineous marriage?
	Why are there so many patients with SMA in Türkiye?
	Whose children can be affected by SMA?
Genetic screening tests for SMA	How is SMA carrier screening performed before marriage?
	What happens if the SMA test performed by a family physician is positive?
	Why is the SMA test performed only on the male partner?
	What can be done to prevent a baby from having SMA?
	Is SMA tested in the heel-prick blood sample taken from newborns?
Genetic diagnostic tests and gene therapy for SMA	What happens if the SMA test is positive before marriage?
	What happens if the mother or father is an SMA carrier?
	Can SMA be detected during pregnancy?
	Can patients with SMA recover?
	Is *SMA* gene therapy a definitive cure?

SMA: Spinal muscular atrophy.

The selected questions were then categorized into three predefined thematic groups, each comprising five questions: the definition of SMA; genetic screening tests for SMA; and genetic diagnosis and treatment of SMA.

Large Language Models and Generation of Responses

Each of the 15 questions was independently posed to the free versions of three widely accessible LLMs: ChatGPT, Gemini, and DeepSeek. These tools were selected based on reported usage prevalence in Türkiye for public information seeking, while other platforms (e.g., Copilot, Meta AI, Apple Intelligence) are also available (14). All queries were entered using identical wording for each model. No follow-up prompts, clarifications, or iterative refinements were applied. The initial responses generated by each LLM were recorded verbatim and used for analysis. All responses were collected in Turkish, and the webpage was refreshed before each question. Additionally, questions were collected within the same time period to minimize potential variability related to model updates.

Evaluation Tools and Outcome Measures

Each response generated by the LLMs was independently evaluated by a single board-certified clinical geneticist using multiple established instruments assessing information quality, readability, and usability. Information quality was assessed using the modified DISCERN instrument, which evaluates the reliability and quality of health information, with higher scores indicating better overall quality, and a 5-point Likert scale assessing accuracy and completeness, where higher scores reflect greater concordance with expert consensus (15,16). Readability was analyzed using the Flesch–Kincaid Reading Ease score, in which higher scores indicate easier-to-read text, and the Flesch–Kincaid Grade Level, which estimates the U.S. school grade required to comprehend the material, with lower values indicating greater readability (17). Usability and clarity were further assessed using the Patient Education Materials Assessment Tool (PEMAT), which evaluates whether health information is understandable and actionable for patients, with higher percentages representing better performance in both understandability and actionability domains (18).

For example, the first FAQ, “What does SMA mean?” (“SMA hastalığı ne demek?”) was independently queried across all evaluated LLMs. The response generated by ChatGPT correctly defined SMA as a genetic neuromuscular disorder characterized by progressive muscle weakness and atrophy and identified involvement of motor neurons in the spinal cord, autosomal recessiv*e i*nheritance related to the *SMN1* gene, and common clinical features such as limb weakness, delayed motor milestones, and respiratory involvement (Supplementary Table 1). The response also provided a brief overview of SMA subtypes and current treatment approaches. This response received a DISCERN score of 3, reflecting moderate information quality, as it presented generally accurate and balanced information but did not address uncertainty, alternative sources, or limitations in detail. The Likert score was 4, indicating good accuracy and completeness of a general, patient-oriented explanation. Readability analysis yielded a Flesch–Kincaid Reading Ease score of 46.6 and a Flesch–Kincaid Grade Level of 8.7, indicating that the content requires at least middle-school level reading proficiency. The PEMAT understandability score was 84.61%, indicating that the information was generally clear and comprehensible to patients. However, the PEMAT actionability score was 0% because the response provided descriptive information and did not include guidance on actionable next steps. This evaluation approach was applied to all responses generated by each LLM across the remaining questions.

Statistical Analysis

Statistical analyses were performed using IBM SPSS Statistics version 23 (IBM Corp., Armonk, NY, USA). Normality of data distribution was assessed using the Shapiro-Wilk test. For comparisons of normally distributed scores across three or more question categories, one-way analysis of variance (ANOVA) was applied. For non-normally distributed scores across three or more question categories, the Kruskal-Wallis test was used, with post-hoc multiple comparisons conducted using Dunn’s test. Comparisons of normally distributed scores across different AI models were performed using repeated-measures ANOVA, with Bonferroni correction applied for multiple comparisons. For non-normally distributed scores across LLMs, the Friedman test was used, followed by Dunn’s test for post-hoc pairwise comparisons. Results are presented as mean ± standard deviation for normally distributed data and as median (minimum-maximum) for non-normally distributed data. A two-sided p-value of < 0.05 was considered statistically significant.

AI-based tools were used to assist with language editing and manuscript preparation; however, all content was reviewed, verified, and finalized by the authors. Ethical approval was not required for this study, as it involved no human participants, patient data, or identifiable personal information.

RESULTS

Overall Performance of Large Language Model Across All Spinal Muscular Atrophy Faqs

Across all 15 SMA-related FAQs, information quality was similar among LLMs. Median DISCERN scores were 3.00 (2.00–3.00) for ChatGPT, 3.00 (2.00–4.00) for Gemini, and 3.00 (0.00–3.00) for DeepSeek, with no significant difference (p = 0.069). These scores were clustered around 3.00, suggesting that the information quality of LLM-generated SMA FAQs was moderate. Median Likert-scale scores differed across models (p = 0.015), with ChatGPT and Gemini scoring 4.00 (2.00–5.00), and DeepSeek scoring 3.00 (1.00–4.00); post-hoc analyses revealed no significant pairwise differences (Table 2).

Median PEMAT understandability scores were 84.61 (83.33–92.85) for ChatGPT, 91.66 (81.25–92.85) for Gemini, and 84.61 (73.33–92.85) for DeepSeek (p = 0.786). Median PEMAT actionability scores were identical across models at 60.00 (p = 0.105). Median Flesch–Kincaid Reading Ease scores were 64.30 (22.80–81.90), 59.40 (39.90–69.50), and 58.00 (46.60–75.20) for ChatGPT, Gemini, and DeepSeek, respectively (p = 0.344).

Mean Flesch–Kincaid Grade Level scores differed significantly among models (p = 0.001), with lower values for ChatGPT (6.64 ± 2.17) compared with Gemini (8.37 ± 1.56) and DeepSeek (8.12 ± 1.19); no difference was observed between Gemini and DeepSeek (Figure 2e).

Language Model Across Performance Across Spinal Muscular Atrophy Question Categories

Within individual question categories, no significant differences in DISCERN or Likert scores were observed among LLMs. However, within-model analyses demonstrated significant category-based differences for DeepSeek in both DISCERN (p = 0.034) and Likert scores (p = 0.032), whereas ChatGPT and Gemini showed no such differences (Table 3).

PEMAT understandability scores differed across models for SMA disease-definition questions (p = 0.015), with medians of 92.30

(84.61–92.30) for ChatGPT, 92.30 (84.61–92.85) for Gemini, and 76.92 (73.33–83.33) for DeepSeek. A significant overall difference among the models was observed for genetic screening test questions (p = 0.034), whereas no difference was found for genetic diagnosis and treatment questions (p = 0.128).

For SMA disease definition questions, mean Flesch–Kincaid Reading Ease scores differed among models (p = 0.020), with values of 70.24 ± 14.03 for ChatGPT, 57.26 ± 10.90 for Gemini, and 58.12 ± 5.91 for DeepSeek. Mean Flesch–Kincaid Grade Level scores between disease-definition questions (p = 0.002) and genetic screening test questions (p = 0.007), with lower grade levels observed for ChatGPT in both categories. No significant readability differences were observed for genetic diagnosis and treatment questions.

PEMAT actionability scores did not differ among models within any category. However, significant within-model differences across categories were observed for all three AI tools (ChatGPT p = 0.013; Gemini p = 0.004; DeepSeek p = 0.023), and the lowest actionability scores were consistently found for SMA disease-definition questions.

DISCUSSION

This study evaluated the performance of LLMs in generating responses to SMA–related FAQs from a clinical genetics perspective. Across all evaluated metrics, overall information quality and understandability did not differ significantly among ChatGPT, Gemini, and DeepSeek. In contrast, readability differed across models, with ChatGPT producing responses at a lower reading grade level. Actionability varied by FAQ

Table 2. Comparison of responses generated by artificial intelligence tools according to DISCERN, Likert, PEMAT, and Flesch-Kincaid readability metrics.

	ChatGPT	Gemini	DeepSeek	Test statistic	p-value
DISCERN score	3.00 (2.00 – 3.00)	3.00 (2.00 – 4.00)	3.00 (0.00 – 3.00)	5.353	0.069[x]
Likert scale	4.00 (2.00 – 5.00)	4.00 (2.00 – 5.00)	3.00 (1.00 – 4.00)	8.359	**0.015**[x]
PEMAT understandability score	84.61 (83.33 – 92.85)	91.66 (81.25 – 92.85)	84.61 (73.33 – 92.85)	0.481	0.786[x]
Flesch–Kincaid reading ease score	64.30 (22.80 – 81.90)	59.40 (39.90 – 69.50)	58.00 (46.60 – 75.20)	2.133	0.344[x]
Flesch–Kincaid grade level	6.64 ± 2.17[a]	8.37 ± 1.56[b]	8.12 ± 1.19[b]	8.328	**0.001**[y]
PEMAT actionability	60.00 (0.00 – 100.00)	60.00 (0,00 – 80.00)	60.00 (0.00 – 83.33)	4.514	0.105[x]

AI: Artificial intelligence.

[x]Friedman test, [y]Repeated-measures analysis of variance.

Data are presented as median (minimum-maximum) or mean ± standard deviation.

[a-b]AI tools sharing the same superscript letter do not differ significantly.

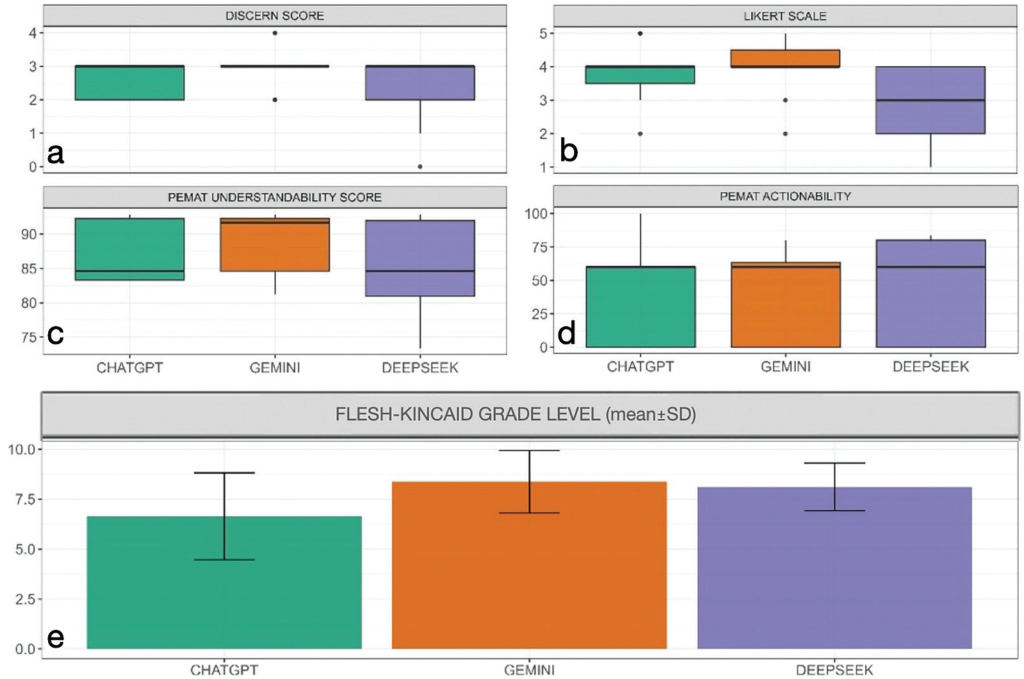


Figure 2. Comparison of large language model performance across multiple evaluation metrics for spinal muscular atrophy related frequently asked questions. (a) Distribution of DISCERN scores, b) Distribution of Likert scale scores, c) PEMAT understandability scores, d) PEMAT actionability scores, e) Mean Flesch–Kincaid Grade Level scores (± standard deviation). Box plots display medians, interquartile ranges, and ranges, while bars represent mean values with standard deviation where indicated. Comparisons are shown for ChatGPT, Gemini, and DeepSeek.

category, with higher scores observed for screening, diagnosis, and treatment-related questions than for disease definition questions. These findings indicate that LLM-generated responses to SMA FAQs vary according to both the model used and the type of question asked.

The absence of significant differences in information quality scores across LLMs reflects a broader limitation of current generative models in addressing complex genetic conditions such as SMA. Because genetic counseling involves important concepts including carrier status, inheritance risk, and genotype–phenotype variability, similar information quality scores do not necessarily translate into equivalent clinical usefulness (19,20). Although LLMs may support preliminary information seeking, variability in readability, actionability, and topic-specific performance shows the need for clinician led contextualization (21-23).

Readability is a determinant of patient understanding in SMA, where counseling is repeated across the lifespan and often occurs under emotional stress and limited health literacy. Differences in

Table 3. Comparison of responses generated by artificial intelligence tools according to DISCERN, Likert, PEMAT, and Flesch-Kincaid readability metrics within and across question groups.

Score	AI tool	Definition of SMA	Genetic screening tests for SMA	Genetic diagnostic tests and gene therapy for SMA	Test statistic	p-value
DISCERN score	ChatGPT	3.00 (2.00–3.00)	3.00 (2.00–3.00)	2.00 (2.00–3.00)	1.556	0.459^{t}
	Gemini	3.00 (2.00–3.00)	3.00 (3.00–3.00)	3.00 (2.00–4.00)	1.400	0.497^{t}
	DeepSeek	3.00 (3.00–3.00)^{a}	2.00 (0.00–3.00)^{b}	3.00 (2.00–3.00)^{ab}	6.746	**0.034**^{t}
	Test statistic	1.000	4.800	4.308		
	p-value^{x}	0.607	0.091	0.116		
Likert scale	ChatGPT	4.00 (3.00–5.00)	4.00 (3.00–4.00)	4.00 (2.00–5.00)	1.173	0.556^{t}
	Gemini	4.00 (3.00–5.00)	4.00 (4.00–5.00)	4.00 (2.00–5.00)	0.410	0.815^{t}
	DeepSeek	4.00 (4.00–4.00)	3.00 (1.00–4.00)	2.00 (2.00–4.00)	6.863	**0.032**^{t}
	Test statistic	0.545	5.375	3.500		
	p-value^{x}	0.761	0.068	0.174		
PEMAT understandability score	ChatGPT	92.30 (84.61–92.30)^{AB}	83.33 (83.33–84.61)	92.30 (83.33–92.85)	7.198	**0.027**^{t}
	Gemini	92.30 (84.61–92.85)^{A}	91.66 (84.61–92.85)	83.33 (81.25–92.30)	3.490	0.175^{t}
	DeepSeek	76.92 (73.33–83.33)^{aB}	85.71 (83.33–92.30)^{ab}	91.66 (84.61–92.85)^{b}	9.403	**0.009**^{t}
	Test statistic	8.444	6.778	4.111^{a}		
	p-value^{x}	**0.015**	**0.034**	0.128		
Flesch-Kincaid reading ease score	ChatGPT	70.24 ± 14.03^{A}	64.38 ± 0.84	46.20 ± 18.11	2.629	0.160^{z}
	Gemini	57.26 ± 10.90^{B}	60.88 ± 4.07	58.82 ± 9.86	0.213	0.811^{z}
	DeepSeek	58.12 ± 5.91^{AB}	60.38 ± 9.06	56.10 ± 7.04	0.413	0.671^{z}
	Test statistic	6.695	0.791	4.357		
	p-value^{x}	**0.020**	0.486	0.104		
Flesch-Kincaid grade level	ChatGPT	5.42 ± 2.01^{A}	6.04 ± 0.33^{A}	8.46 ± 2.44	3.837	0.051^{z}
	Gemini	9.16 ± 1.79^{B}	7.78 ± 0.80^{B}	8.16 ± 1.85	1.048	0.381^{z}
	DeepSeek	8.44 ± 0.86^{AB}	7.56 ± 0.97^{AB}	8.36 ± 1.64	0.815	0.466^{z}
	Test statistic	15.259	9.629	0.124		
	p-value^{x}	**0.002**	**0.007**	0.885		
PEMAT actionability score	ChatGPT	0.00 (0.00–0.00)^{a}	60.00 (60.00–60.00)^{b}	60.00 (0.00–100.00)^{b}	8.714	**0.013**^{t}
	Gemini	0.00 (0.00–0.00)^{a}	66.66 (60.00 –80.00)^{b}	60.00 (60.00–80.00)^{b}	10.932	**0.004**^{t}
	DeepSeek	0.00 (0.00–60.00)^{a}	80.00 (60.00–83.33)^{b}	80.00 (0,00–83.33)^{ab}	7.541	**0.023**^{t}
	Test statistic	2.000	5.765	0.400		
	p-value^{x}	0.368	0.056	0.819		

^{x}Friedman test, ^{y}Repeated-measures analysis of variance, ^{z}One-way ANOVA, ^{t}Kruskal-Wallis test.
Data are presented as median (minimum-maximum) or mean ± standard deviation.
^{ab}Within each AI tool, question groups sharing the same lowercase letter do not differ significantly.
^{AB}Within each question group, AI tools sharing the same uppercase letter do not differ significantly.
SMA: Spinal muscular atrophy, AI: Artificial intelligence, ANOVA: Analysis of variance

readability across LLMs indicate that linguistic style and sentence structure influence patients' comprehension of concepts. Variability in LLM performance across SMA FAQ categories likely reflects differences in underlying genetic complexity: disease definition questions are largely descriptive, whereas screening, diagnostic, and treatment-related questions require integration of inheritance, test interpretation, risk assessment, and evolving therapeutic options (21-23). Actionability similarly depends more on question intent than on model characteristics, as definition-focused questions inherently offer limited guidance, while screening and treatment related questions align more closely with clinical decision making and follow-up processes (18,21,23).

Implications for Clinical Genetic Counseling Practice in the Era of Large Language Models

Genetic counseling is one of the core components of clinical genetics practice, and its role becomes increasingly critical as patients seek genetic information and test result evaluation from AI-based tools. Counseling often involves supporting patients through uncertainty, facilitating informed decision making by addressing expectations, and emotional responses unique to each individual or family rather than directing patients toward a single predefined choice (24,25). Such individualized, bidirectional communication remains inherently human and may not be consistently achieved by LLM-based tools, which generate standardized responses without access to personal context or psychosocial cues. Clinical geneticists are now more likely to encounter patients who present after consulting LLMs and who may have acquired false, incomplete, or oversimplified information. In such cases, counseling should begin with identifying and correcting misinformation and clarifying unmet informational needs. This reinforces the importance of structured pretest and posttest genetic counseling and a strong patient-physician relationship grounded in clear communication and trust.

In addition, genetic risk assessment is highly individualized and cannot always be generalized using population-based information. For example, genetic screening tests are designed to detect the most common pathogenic variants within a population to maximize efficiency and cost-effectiveness, but they do not capture all disease-causing mechanisms. In SMA, rare scenarios such as silent carrier status, pathogenic single-nucleotide variants not included in standard screening panels, or *de novo* mutations may not be identified through routine carrier screening (26). Consequently, identical screening results may carry different residual risks across individuals and families. Communicating these individualized risks is a core component of genetic counseling, and it extends beyond the current capabilities of AI-generated responses.

Effective genetic counseling in this context requires more than a classical verbal explanation. New supportive tools should be integrated into counseling practice, including disease-specific visual materials to explain inheritance patterns, carrier states such as 2 + 0 silent carriers, and the distinctions between screening and diagnostic tests. These tools may improve understanding of complex genetic concepts that are frequently misinterpreted in AI-generated responses. Genetic counseling should also be recognized as a longitudinal and repetitive process, requiring sufficient time and multiple sessions to allow patients and families to process information, consider options, and make informed decisions. As external information sources become more prevalent, counseling may need to be more detailed and iterative than previously, continuing until patients and families demonstrate sufficient understanding of their genetic risks and choices.

An additional consideration concerns the governance of LLM platforms , particularly those used for training and deployment. For content addressing genetic screening, diagnosis, or treatment, LLM outputs could be regulated at the platform level to include standardized prompts encouraging users to seek clinician input for individualized risk assessment and final interpretation. Incorporating safeguards such as explicit statements directing users to consult healthcare professionals for definitive results may help reduce the risks of misinterpretation.

Several limitations should be considered when interpreting these findings. First, the analysis was based on a limited set of SMA-related FAQs obtained from Google search results at a single time point, which may not capture the full spectrum of information needs encountered in clinical genetics practice. Second, all questions and LLM-generated responses were evaluated in Turkish, and the findings may reflect language-specific characteristics of both the models and the assessment tools, as the LLMs are mostly trained in English. Third, LLM outputs are dynamic and subject to change with model updates, retraining, and prompt sensitivity, which may affect reproducibility over time. Fourth, the evaluation relied on expert-based assessment instruments that, while validated for health information appraisal, were not specifically designed to assess disease-specific genetic accuracy or the adequacy of counseling. Finally, the study focused on static written responses and did not account for interactive dialogue or individualized context, which are essential components of real-world genetic counseling encounters.

CONCLUSION

LLMs can generate generally consistent responses to SMA-related FAQs; however, their performance varies with readability, actionability, and question category. While overall information quality and understandability were similar across models, differences in linguistic accessibility and topic-specific response characteristics show important limitations for direct clinical use. These findings suggest that LLM-generated SMA information may serve as a supplementary educational resource today, but should not replace clinician-led genetic counseling, particularly in contexts requiring individualized risk assessment, interpretation of genetic test results, and shared decision-making.

Ethics

Ethics Committee Approval: Ethical approval was not required for this study, as it involved no human participants, patient data, or identifiable personal information.

Informed Consent: No informed consent is required for this study as we analysed publicly available online content and responses generated by AI systems.

Footnotes

Authorship Contributions

Concept: D.D.E., Design: D.D.E., M.A., Data Collection or Processing: D.D.E., Analysis or Interpretation: D.D.E., Literature Search: D.D.E., Writing: D.D.E., M.A.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support.

Supplementary Table 1: https://d2v96fxpocvxx.cloudfront.net/c81cd83d-6f66-4047-b292-9f60f894af26/documents/Supplementary%20Table%201.xlsx

REFERENCES

1. Keinath MC, Prior DE, Prior TW. Spinal Muscular Atrophy: Mutations, Testing, and Clinical Relevance. Appl Clin Genet. 2021; 14: 11-25.
2. Prior TW. Carrier screening for spinal muscular atrophy. Genet Med. 2008; 10: 840-2.
3. Verhaart IEC, Robertson A, Wilson IJ, Aartsma-Rus A, Cameron S, Jones CC, et al. Prevalence, incidence and carrier frequency of 5q-linked spinal muscular atrophy - a literature review. Orphanet J Rare Dis. 2017; 12: 124.
4. Cooper K, Nalbant G, Sutton A, Harnan S, Thokala P, Chilcott J, et al. Systematic review of presymptomatic treatment for spinal muscular atrophy. Int J Neonatal Screen. 2024; 10: 56
5. Prior TW, Leach ME, Finanger EL. Spinal muscular atrophy. In: Adam MP, Bick S, Mirzaa GM, et al., editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993-. 2000 Feb 24 [updated 2026 Feb 12]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK1352/
6. Milligan JN, Blasco-Pérez L, Costa-Roger M, Codina-Solà M, Tizzano EF. Recommendations for interpreting and reporting silent carrier and disease-modifying variants in SMA testing workflows. Genes (Basel). 2022; 13: 1657.
7. Shahsavar Y, Choudhury A. User intentions to use ChatGPT for self-diagnosis and health-related purposes: cross-sectional survey study. JMIR Hum Factors. 2023; 10: e47564.
8. Kamatani Y, Kaname T. Artificial intelligence in medical genomics. J Hum Genet. 2024; 69: 475.
9. Yun HS, Bickmore T. Online health information-seeking in the era of large language models: cross-sectional web-based survey study. J Med Internet Res. 2025; 27: e68560.
10. Ayik G, Kolac UC, Aksoy T, Yilmaz A, Sili MV, Tokgozoglu M, et al. Exploring the role of artificial intelligence in Turkish orthopedic progression exams. Acta Orthop Traumatol Turc. 2025; 59: 18-26.
11. Kolac UC, Karademir OM, Ayik G, Kaymakoglu M, Familiari F, Huri G. Can popular AI large language models provide reliable answers to frequently asked questions about rotator cuff tears? JSES Int. 2025; 9: 390-7.
12. Mehmet S, et al. Comparative evaluation of large language models in addressing autism-related information queries: insights from ChatGPT, Gemini, and Copilot. Gazi Med J. 2025; 36: 407-16.
13. Volk SC, Schäfer MS, Lombardi D, Mahl D, Yan X. How generative artificial intelligence portrays science: Interviewing ChatGPT from the perspective of different audience segments. Public Underst Sci. 2025; 34: 132-53.
14. Küçüksabanoğlu ZK. Yapay Zeka Politikaları Derneği (AIPA) Gelecek Araştırması: Toplumda yapay zeka algısı. Yapay Zeka Politikaları Derneği (AIPA); 2025. Available from: https://aipaturkey.org/media/pdfs/YapayZeka_AlgiArastirmasi_2025_Toplum.pdf
15. Likert R. A technique for the measurement of attitudes. Arch Psychol. 1932;140:1-55. Available from: https://legacy.voteview.com/pdf/Likert_1932.pdf
16. 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. 1999; 53: 105-11.
17. Kincaid JP, Fishburne RP Jr, Rogers RL, Chissom BS. Derivation of new readability formulas: automated Readability Index, Fog Count and Flesch Reading Ease Formula for Navy enlisted personnel. Millington (TN): Naval Technical Training Command, Research Branch; 1975 Feb. Available from: https://stars.library.ucf.edu/istlibrary/56/.
18. Shoemaker SJ, Wolf MS, Brach C. Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. Patient Educ Couns. 2014. 96: 395-403.
19. Jeon S, Lee SA, Chung HS, Yun JY, Park EA, So MK, et al. Evaluating the use of generative artificial intelligence to support genetic counseling for rare diseases. Diagnostics (Basel). 2025; 15: 672.
20. Ahimaz P, Bergner AL, Florido ME, Harkavy N, Bhattacharyya S. Genetic counselors' utilization of ChatGPT in professional practice: a cross-sectional study. Am J Med Genet A. 2024; 194: e63493.
21. Brett GR, Ward A, Bouffler SE, Palmer EE, Boggs K, Lynch F, et al. Co-design, implementation, and evaluation of plain language genomic test reports. NPJ Genom Med. 2022; 7: 61.
22. Farmer GD, Gray H, Chandratillake G, Raymond FL, Freeman ALJ. Recommendations for designing genetic test reports to be understood by patients and non-specialists. Eur J Hum Genet. 2020; 28: 885-95.
23. Haga SB, Mills R, Pollak KI, Rehder C, Buchanan AH, Lipkus IM, et al. Developing patient-friendly genetic and genomic test reports: formats to promote patient engagement and understanding. Genome Med. 2014; 6: 58.
24. Coen E, Del Fiol G, Kaphingst KA, Borsato E, Shannon J, Smith H, et al. Chatbot for the return of positive genetic screening results for hereditary cancer syndromes: prompt engineering project. JMIR Cancer. 2025; 11: e65848.
25. Meekins-Doherty L, Dive L, McEwen A, Sexton A. Generative AI and the profession of genetic counseling. J Genet Couns. 2025; 34: e2009.
26. Tıbbi Genetik Derneği. Spinal Musküler Atrofi El Kitabı [Internet]. Ankara: Tıbbi Genetik Derneği; 2024. 20 p. Available from: https://yonetim.citius.technology/files/kurum/kurum75/menu/spinal-muskuler-atrofi-el-kitabi--bitmis-hali-15x23-olacak---1-.pdf