

# Optimizing Temperature Settings in Large Language Models for Enhanced Patient Understanding in Orthodontic Treatment

## Büyük Dil Modellerinde Temperature Ayarlarının Hastaların Ortodontik Bilgiyi Anlama Düzeyini Artırmaya Yönelik Optimizasyonu

Ebru YURDAKURBAN<sup>1</sup>  
Kübra Gülnur TOPSAKAL<sup>2</sup>  
Gökhan Serhat DURAN<sup>3</sup>

<https://orcid.org/0000-0001-9477-6894>

<https://orcid.org/0000-0002-2717-3492>

<https://orcid.org/0000-0001-6152-6178>

<sup>1</sup>Department of Orthodontics, Faculty of Dentistry, Muğla Sıtkı Koçman University, Muğla, Turkey

<sup>2</sup>Department of Orthodontics, Gulhane Faculty of Dentistry, University Of Health Sciences, Ankara, Turkey

<sup>3</sup>Department of Orthodontics, Faculty of Dentistry, Canakkale Onsekiz Mart University, Çanakkale, Turkey

**Citation:** Yurdakurban E, Topsakal KG, Duran GS. Optimizing Temperature Settings in Large Language Models for Enhanced Patient Understanding in Orthodontic Treatment. *Int Arc Dent Sci.* 2026; 47(1): 41-48.

### ABSTRACT

**INTRODUCTION:** This study aimed to integrate an orthodontic resource set into a general-purpose language model and evaluate the chatbot's responses to orthodontic treatment questions at various temperature settings.

**MATERIAL and METHODS:** An orthodontic chatbot was customized using the OpenAI on Microsoft Azure. PDF resources on orthodontic treatments were incorporated. Forty questions were posed to the chatbot at temperature settings of 0, 0.5, and 1.0. Responses were evaluated based on "level of detail," "fluency," and "user-focused understandability" using a three-point scale. Readability was assessed with the Flesch-Kincaid Grade Level (FKGL) and Gunning Fog Index. Statistical analyses were performed using Kruskal-Wallis and Mann-Whitney U tests.

**RESULTS:** The highest scores for 'user-focused understandability,' 'fluency,' FKGL, and Gunning Fog indices were observed at a temperature setting of 1.0. The average Flesch-Kincaid Grade Level indicated a 10th-grade reading level. Significant differences among all three temperature settings were found for the 'level of detail' and 'user-focused understandability' criteria ( $p < 0.001$  and  $p < 0.025$ , respectively). Post-hoc analyses revealed that, for the 'level of detail' criterion, the significant difference was between the temperature setting of 0 and the other two settings.

**CONCLUSION:** Temperature settings affect the content and linguistic features of responses. Lower temperatures enhance detail and comprehensiveness, whereas higher temperatures improve fluency and understandability.

**Keywords:** Large language model, patient education, orthodontics, temperature

### Öz

**GİRİŞ:** Bu çalışmanın amacı, genel amaçlı bir dil modeline ortodontiyle ilgili kaynakları entegre ederek, sohbet robotunun farklı temperature ayarlarında verdiği yanıtların içerik ve dil özelliklerini değerlendirmektir.

**YÖNTEM ve GEREÇLER:** Microsoft Azure platformunda OpenAI kullanılarak ortodonti alanına özgü sohbet robotu geliştirildi. Sisteme ortodontik tedaviye dair PDF kaynaklar entegre edildi. Chatbota 0, 0.5 ve 1.0 temperature ayarlarında toplam kırk soru yöneltildi. Yanıtlar, üç puanlı ölçekle "detay düzeyi", "akıcılık" ve "kullanıcı odaklı anlaşılabilirlik" açısından değerlendirildi. Okunabilirlik için Flesch-Kincaid Grade Level (FKGL) ve Gunning Fog İndeksi kullanıldı. İstatistiksel analizler Kruskal-Wallis ve Mann-Whitney U testleriyle yapıldı ( $p < 0.05$ ).

**BULGULAR:** "Kullanıcı odaklı anlaşılabilirlik", "akıcılık", FKGL ve Gunning Fog indeksleri için en yüksek skorlar 1.0 temperature ayarında gözlemlenirken, "detay düzeyi" kriteri için en yüksek skor 0 temperature ayarında elde edildi. Ortalama Flesch-Kincaid Sınıf Düzeyi, yaklaşık olarak lise ikinci sınıf (10. sınıf) seviyesinde bir okunabilirlik gösterdi. "Detay düzeyi" ve "kullanıcı odaklı anlaşılabilirlik" kriterlerinde üç temperature ayarı arasında anlamlı farklılıklar tespit edildi (sırasıyla  $p < 0.001$  ve  $p < 0.025$ ). Post-hoc analizler, "detay düzeyi" kriteri açısından anlamlı farklılığın, temperature 0 ile diğer iki temperature ayarı arasında olduğu belirlendi.

**SONUÇ:** Temperature ayarı, yanıtların içeriğini ve dil özelliklerini önemli ölçüde etkilemektedir. Düşük temperature detay düzeyini artırırken, yüksek temperature anlaşılabilirliği geliştirmektedir.

**Anahtar Kelimeler:** Büyük dil modeli, hasta eğitimi, ortodonti, temperature

Corresponding author: ebruyurdakurban@mu.edu.tr

Received Date: 14.04.2025

Accepted Date: 02.10.2025

## INTRODUCTION

Individuals seeking orthodontic treatment frequently consult online resources to evaluate treatment options, comprehend potential challenges during the treatment process, and anticipate expected outcomes.<sup>1,2</sup> In recent years, artificial intelligence (AI)-based chatbots have demonstrated several advantages over traditional information sources. These chatbots provide patients with subject-specific medical information in an interactive, personalized, and efficient manner, offering immediate accessibility.<sup>3,4</sup>

Although chatbots are increasingly used, they may generate incomplete or inaccurate responses, especially when queries fall outside their training data, a phenomenon known as "hallucinations".<sup>5</sup> In a recent study, Aljamaan et al.<sup>6</sup> introduced a reference hallucination score tool to assess the authenticity of citations provided by six different AI-based chatbots. Their findings revealed that ChatGPT and Bing exhibited a critical degree of hallucination.<sup>6</sup> The issue of hallucinations in language models is closely associated with the selection of "tokens," the smallest linguistic units used in text processing. When a language model fails to fully comprehend the context during response generation, it tends to select tokens (words or word fragments) based on their frequency in the training data. Consequently, the model may produce information that is incorrect or non-existent.<sup>7,8</sup> Incorporating subject-specific sources with verified accuracy into the training data encourages the model to utilize only reliable tokens and their combinations from these sources. This approach enables the model to prioritize validated expressions from trusted references, thereby reducing the risk of hallucinations.

Language models are influenced by various parameters that impact the accuracy, level of detail, and presentation of information. One critical parameter is "temperature," which modulates the probability distribution of the tokens produced by the model. The model calculates a probability based on the frequency of use of each token, and this rate is used in the process of selecting the next token during text generation. A low temperature setting prioritises high-probability tokens, resulting in deterministic, consistent and predictable responses. Conversely, a high temperature setting increases the likelihood of selecting lower-probability tokens, leading to more random, creative, and variable outputs.<sup>9,10</sup> Researchers are exploring the optimal temperature settings for chatbots across various tasks. Veen et al.<sup>11</sup> evaluated the performance of different language models in clinical text summarization tasks and observed that a low temperature setting of 0.1 yielded more deterministic responses and the best performance across all models. Nielsen and Karlstand<sup>12</sup> compared the performance of various language models and configuration settings in information flow and decision-making processes within public transport management.

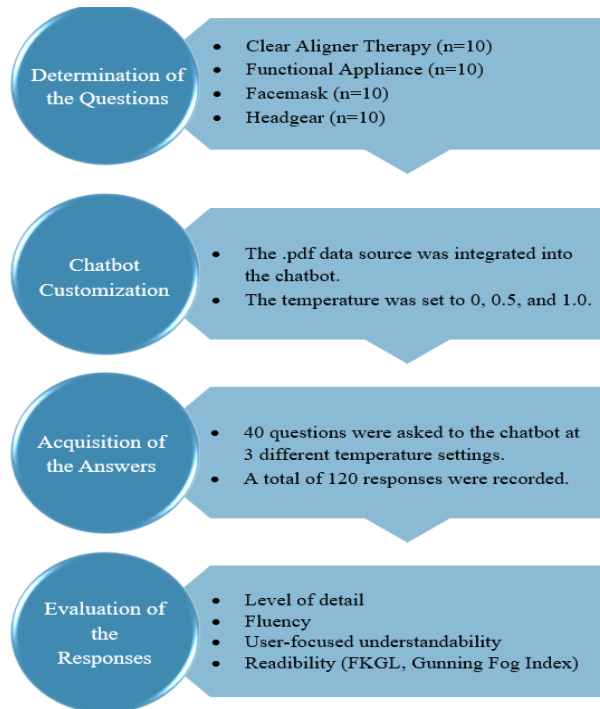
They reported that the GPT-4 model achieved the highest accuracy when the temperature parameter was set to 0. Similarly, Zhuang et al. evaluated the performance of ChatGPT-3.5 on the American Endocrine Society Self-Assessment Test across temperature settings of 0, 0.3, 0.7, and 1. Although no statistically significant differences were observed, they noted that the most accurate response rates were achieved at settings of 0.3 and 0.7.<sup>13</sup>

Customizing the data sources and temperature settings of chatbots enables the generation of responses with language features tailored to the target audience. However, fine-tuning these settings often requires technical expertise and the use of application programming interfaces (APIs) within a Python environment.<sup>14,15</sup> In our study, we integrated a specialized orthodontic resource set into a general-purpose language model via an online platform, eliminating the need for technical expertise or coding knowledge. The aim was to evaluate the level of detail, fluency, user focus, and readability of the chatbot's responses to questions about orthodontic treatment procedures at various temperature settings.

## MATERIALS AND METHODS

### Preparation of Questions and Data Sources

The flowchart of the study is presented in Figure 1. To develop questions that patients might ask about orthodontic treatments requiring cooperation, four categories were established focusing on removable intraoral and extraoral appliances: Clear Aligner Therapy, Functional Appliances, Face Masks, and Headgear. Two researchers prepared ten questions for each category related to treatment descriptions, potential issues during therapy, and expected outcomes (Table 1). The questions used in this study were based on patient information leaflets published by the British Orthodontic Society, which provide standardized and reliable orthodontic information and include questions commonly asked by patients. In addition to questions that were quoted directly from these leaflets, the authors generated new questions that patients might potentially ask, inspired by the structure, language and content of the original materials. The question categories were formed by selecting orthodontic treatment methods that require patient cooperation and are likely to pose challenges. These categories included definitions of such treatment methods and their potential impact on patients' daily routines. To reduce the potential influence of memory retention, each question was submitted in a separate and independent chatbot session. The questions were not entered consecutively or within the same conversation window, ensuring that previous interactions did not affect subsequent responses.



**Figure 1:** The flowchart of the study.

Responses were gathered using Google and Google Scholar (Google Inc., California, USA), accessing open-access and free resources. Patient guides, informational texts, brochures from orthodontic associations, and articles from clinical websites were reviewed, and relevant information was saved in PDF format. The selection of resources containing answers to the questions, as well as the evaluation of their content quality, academic validity, and suitability for patient education, was determined by consensus between the two researchers.

**Customization of the Chatbot and Response Generation**

An orthodontic patient information chatbot was customized using the OpenAI GPT-3.5 language model available on the Microsoft Azure cloud computing platform (Azure AI Services 2024, www.azure.microsoft.com). PDF resources containing answers to 40 predetermined questions were uploaded as data sources to the model. The model was configured to restrict its responses exclusively to the uploaded sources. Since the study aimed to evaluate only the temperature parameter as a variable, no additional configurations were made, and all

**Table 1.** Orthodontic treatment–related questions asked to ChatGPT via the OpenAI GPT-3.5 language model, accessed through Microsoft Azure AI Services.

Clear Aligner Therapy–Related Questions		Facemask Therapy–Related Questions	
1	What is clear aligner treatment and how does it work?	1	What is the face mask?
2	What is the duration of treatment in clear aligner therapies?	2	How does a face mask correct jaw misalignment?
3	How often will I need to visit the orthodontist during my clear aligner treatment?	3	How many hours per day should I use face mask?
4	How are the attachments placed on my teeth?	4	What is the most effective age for the use of face mask?
5	What are the differences between clear aligners and the traditional braces?	5	How is the face mask attached, and is it painful to wear?
6	How does the visibility of clear aligners compare to traditional braces?	6	Can the face mask be worn during the day or just at night?
7	How many hours per day should I wear my clear aligners?	7	What are the expected outcomes of using a face mask in orthodontics?
8	How should I clean my clear aligner?	8	What should I do if my face mask becomes uncomfortable or causes irritation?
9	Which foods and drinks should I avoid while wearing my clear aligners?	9	Will wearing a face mask affect my daily activities or diet?
10	Will my speech be affected when wearing my clear aligners?	10	How often will I need to see my orthodontist for adjustments to my face mask?
Functional Appliances–Related Questions		Headgear Appliances–Related Questions	
1	What is the purpose of functional appliances in orthodontic treatment?	1	How does headgear contribute to orthodontic treatment?
2	What is the most effective age for the use of functional appliances?	2	What should I do if my headgear is damaged?
3	How do functional appliances correct jaw discrepancies?	3	Can headgear cause changes in facial structure or jaw alignment?
4	How many hours per day should the functional appliance each day?	4	What should I do if there is any problem with the headgear?
5	Can functional appliances be used to treat all malocclusion types?	5	How many hours per day should I wear headgear?
6	Will my speech be affected when wearing a functional device?	6	Is headgear treatment painful?
7	Are there any dietary restrictions while using a functional appliance?	7	How frequently do I need to visit the orthodontist for headgear adjustments?
8	What are the common side effects or discomforts associated with functional appliances?	8	Can headgear be worn while sleeping, and are there any risks?
9	Do functional appliances need to be worn at night?	9	Are there any specific activities or sports to avoid while wearing headgear?
10	What happens if my functional appliance breaks or becomes damaged?	10	How does headgear affect eating practices?

other parameters were left at their default settings. The 40 questions were then posed to the model at three different temperature settings: minimum (0), average (0.5), and maximum (1.0). The questions were submitted via the user message section, and the generated answers were saved in a Microsoft Word document for subsequent analysis.

### Evaluation of Response Language and Content

To assess the scope of content and linguistic characteristics of the chatbot's responses at different temperature settings, we developed three evaluation criteria: "level of detail", "fluency", and "user-focused understandability". A three-tier scoring system was established for each criterion. Additionally, the readability of the responses was measured using the Flesch-Kincaid Grade Level (FKGL)<sup>16</sup> and the Gunning Fog Index<sup>17</sup>.

The "level of detail" criterion evaluated how thoroughly the responses extracted information from the sources and how comprehensively this information was presented. Responses were scored on a three-point scale: *Score 1*: Superficial information. The response provides a general and simple explanation derived from the source. *Score 2*: Moderate detail. The response addresses important points from the source but lacks some details. *Score 3*: Highly detailed information. The response offers comprehensive and highly detailed information extracted from the source.

The "fluency" criterion was developed to assess how fluently, coherently, and naturally the chatbot generates responses. The responses were analyzed based on their level of fluency using a three-point scoring system: *Score 1*: The response is disjointed or difficult to understand. *Score 2*: Moderate fluency. The response is generally fluent but contains some interruptions. *Score 3*: Highly fluency. The response is completely fluent and easy to understand.

The "user-focused understandability" criterion was employed to assess whether the chatbot's responses were user-friendly, easy to comprehend, and provided clear explanations of any technical terminology used. Responses were evaluated on a three-point scale: *Score 1*: Poor user focus. The response is complex and unclear; terminological terms are not adequately explained. *Score 2*: Moderate user focus. The response is generally clear. Terminological terms are partially explained. The response is generally clear; terminological terms are partially explained. *Score 3*: excellent user focus. The response is completely clear; terminological terms are thoroughly and understandably explained.

The readability of the responses was determined using the FKGL and the Gunning Fog Index. The FKGL is calculated from the average number of words per sentence and the average number of syllables per word.<sup>16</sup>

The Gunning Fog Index assesses readability by analyzing sentence lengths and the number of complex words with three or more syllables.<sup>17</sup> Both indices indicate the educational level within the U.S. educational system at which a text can be easily understood.

The evaluation criteria and readability indices were applied by two independent researchers to a total of 120 responses, covering 40 questions at three different temperature settings (0, 0.5, and 1.0). To assess interobserver and intraobserver reliability, half of the responses were re-evaluated by the same two researchers after a two-week interval.

### Statistical Analysis

Descriptive statistics were recorded, including mean, standard deviation, minimum, and maximum values. Statistical analyses were performed using the open-source software Jamovi (The Jamovi Project 2022, version 2.3.21.0, www.jamovi.org). The Shapiro-Wilk test indicated that the data did not follow a normal distribution. Consequently, the Kruskal-Wallis test was employed for group comparisons, with post hoc analyses conducted using the Mann-Whitney U test. A p-value of less than 0.05 was considered statistically significant. The intraclass correlation coefficient (ICC) was used to assess intraobserver and interobserver reliability.

## RESULTS

Descriptive statistics and post-hoc comparisons of the chatbot's responses at different temperature settings are presented in Table 2. The highest scores for 'user-focused understandability,' 'fluency,' FKGL, and Gunning Fog indices were observed at a temperature setting of 1.0, while the highest score for the 'level of detail' criterion was noted at a temperature setting of 0. Significant differences among all three temperature settings were found for the 'level of detail' and 'user-focused understandability' criteria ( $p < 0.001$  and  $p < 0.025$ , respectively). Post-hoc analyses revealed that, for the 'level of detail' criterion, the significant difference was between the temperature setting of 0 and the other two settings. In the 'user-focused understandability' criterion, a significant difference was identified between temperature settings of 0 and 1 (Table 2).

In the responses to the 10 questions in the Clear Aligner Therapy category, significant differences among the temperature settings were observed only in the criteria of fluency and user-focused understandability ( $p = .001$  and  $p = .010$ , respectively; Table 3). The highest scores for the level of detail, FKGL, and Gunning Fog indices in this category were observed at a temperature setting of 0, while the highest score for the fluency criterion was noted at a temperature setting of 1 (Table 3). In the Functional Appliances and Facemask Therapy categories, significant differences among the

different temperature settings were found only in the level of detail criterion ( $p = .046$ , Table 4;  $p = .002$ , Table 5). In the post-hoc comparisons within the Functional Appliances category, a significant difference was found between temperature settings of 0 and 1.0. Additionally, at a temperature setting of 0.5, this category exhibited the highest scores for fluency, user-focused understandability, and the Gunning Fog Index (Table 4). For the Facemask Therapy category, the highest scores in

level of detail, FKGL, user-focused understandability, and Gunning Fog Index were observed at a temperature setting of 0. In the Headgear Appliances category, no statistically significant differences were detected across all three temperature settings for any of the evaluation criteria (Table 6). The intraclass correlation coefficient (ICC) was determined to be above 0.9 for both intra- and inter-observer reliability.

**Table 2.** Descriptive statistics and post hoc comparisons of all groups.

Categories	Temperature Setting=0 (N=40)				Temperature Setting=0.5 (N=40)				Temperature Setting=1.0 (N=40)				P
	Mean±SD	Median	Min-Max	25%-75%	Mean±SD	Median	Min-Max	25%-75%	Mean±SD	Median	Min-Max	25%-75%	
Level of detail	2.77±0.48 <sup>A</sup>	3.00	1.00-3.00	3.00-3.00	2.10±0.84 <sup>B</sup>	2.00	1.00-3.00	1.00-3.00	2.17±0.67 <sup>B</sup>	2.00	1.00-3.00	2.00-3.00	<.001*
Fluency	2.02±0.733	2.00	1.00-3.00	1.75-3.00	2.35±0.770	3.00	1.00-3.00	2.00-3.00	2.35±0.802	3.00	1.00-3.00	2.00-3.00	.067
User-focused	2.15±0.736 <sup>C</sup>	2.00	1.00-3.00	2.00-3.00	2.17±0.781 <sup>CD</sup>	2.00	1.00-3.00	2.00-3.00	2.55±0.597 <sup>D</sup>	3.00	1.00-3.00	2.00-3.00	<.025*
FKGL	10.5±1.68	11.00	5.80-12.0	9.05-12.0	10.7±1.61	11.4	6.40-12.0	9.50-12.0	10.8±1.52	11.15	7.20-12.0	10.3-12.0	.792
Gunning Fog	14.7±2.70	14.8	10.1-19.0	12.3-16.4	14.8±2.66	15.0	8.80-19.0	12.7-17.0	15.3±2.70	15.3	10.8-19.0	12.7-17.8	.548

N: Sample size, SD: Standard Deviation, Min: Minimum, Max: Maximum, 25%: 25 quartile values, 75%: 75 quartile values. FKGL: Flesch-Kincaid Grade Level \*There is a statistically significant difference at  $p < .05$ .

**Table 3.** Descriptive statistics and post hoc comparisons of the ‘Clear Aligner Therapy’ section.

Clear Aligner Therapy	Temperature Setting=0 (N=10)				Temperature Setting=0.5 (N=10)				Temperature Setting=1.0 (N=10)				P
	Mean±SD	Median	Min-Max	25%-75%	Mean±SD	Median	Min-Max	25%-75%	Mean±SD	Median	Min-Max	25%-75%	
Level of detail	2.70±0.483	3.00	2.00-3.00	2.25-3.00	2.30±0.823	2.50	1.00-3.00	2.00-3.00	2.50±0.707	3.00	1.00-3.00	2.00-3.00	.523
Fluency	1.60±0.516 <sup>A</sup>	2.00	1.00-2.00	1.00-2.00	2.50±0.527 <sup>B</sup>	2.50	2.00-3.00	2.00-3.00	2.70±0.675 <sup>B</sup>	3.00	1.00-3.00	3.00-3.00	.001*
User-focused	1.80±0.789 <sup>C</sup>	2.00	1.00-3.00	1.00-2.00	2.20±0.632 <sup>CD</sup>	2.00	1.00-3.00	2.00-2.75	2.80±0.422 <sup>D</sup>	3.00	2.00-3.00	3.00-3.00	.010*
FKGL	9.99±2.09	10.8	5.80-12.0	9.33-11.4	10.1±2.19	11.1	6.40-12.0	8.93-11.9	10.1±1.85	10.6	7.20-12.0	8.97-11.7	.978
Gunning Fog	14.1±2.76	14.2	10.1-19.0	12.5-16.0	13.8±2.70	14.3	8.80-17.3	12.1-15.3	14.1±3.15	13.2	10.8-19.0	11.5-16.8	.977

N: Sample size, SD: Standard Deviation, Min: Minimum, Max: Maximum, 25%: 25 quartile values, 75%: 75 quartile values. FKGL: Flesch-Kincaid Grade Level \*There is a statistically significant difference at  $p < .05$ .

**Table 4.** Descriptive statistics and post hoc comparisons of the ‘Functional Appliances’ section.

Functional Appliances	Temperature Setting=0 (N=10)				Temperature Setting=0.5 (N=10)				Temperature Setting=1.0 (N=10)				P
	Mean±SD	Median	Min-Max	25%-75%	Mean±SD	Median	Min-Max	25%-75%	Mean±SD	Median	Min-Max	25%-75%	
Level of detail	2.60±0.699 <sup>A</sup>	3.00	1.00-3.00	2.25-3.00	2.00±0.816 <sup>AB</sup>	2.00	1.00-3.00	1.25-2.75	1.80±0.632 <sup>B</sup>	2.00	1.00-3.00	1.25-2.00	.046*
Fluency	2.20±0.919	2.50	1.00-3.00	1.25-3.00	2.20±0.789	2.00	1.00-3.00	2.00-3.00	2.00±0.667	2.00	1.00-3.00	2.00-2.00	.752
User-focused	2.10±0.568	2.00	1.00-3.00	2.00-2.00	2.50±0.707	3.00	1.00-3.00	2.00-3.00	2.50±0.527	2.50	2.00-3.00	2.00-3.00	.213
FKGL	10.9±1.70	12.0	7.60-12.0	9.75-12.0	11.3±1.49	12.0	8.0-12.0	12.0-12.0	11.4±1.11	12.0	8.60-12.0	11.1-12.0	.690
Gunning Fog	15.4±2.6	16.0	11.3-18.4	13.3-17.6	16.9±2.65	17.9	11.6-19.0	16.6-18.8	17.3±2.30	18.5	12.3-19.0	15.8-19.0	.118

N: Sample size, SD: Standard Deviation, Min: Minimum, Max: Maximum, 25%: 25 quartile values, 75%: 75 quartile values. FKGL: Flesch-Kincaid Grade Level \*There is a statistically significant difference at  $p < .05$ .

**Table 5.** Descriptive statistics and post hoc comparisons of the ‘Facemask Therapy’ section.

Facemask Therapy	Temperature Setting=0 (N=10)				Temperature Setting=0.5 (N=10)				Temperature Setting=1.0 (N=10)				P
	Mean±SD	Median	Min-Max	25%-75%	Mean±SD	Median	Min-Max	25%-75%	Mean±SD	Median	Min-Max	25%-75%	
Level of detail	3.00±0.0 <sup>A</sup>	3.00	3.00-3.00	3.00-3.00	1.90±0.876 <sup>B</sup>	2.00	1.00-3.00	1.00-2.75	2.00±0.667 <sup>B</sup>	2.00	1.00-3.00	2.00-2.00	.002*
Fluency	2.10±0.876	2.00	1.00-3.00	1.25-3.00	2.80±0.632	3.00	1.00-3.00	3.00-3.00	2.10±0.876	2.00	1.00-3.00	1.25-3.00	.069
User-focused	2.50±0.707	3.00	1.00-3.00	2.00-3.00	2.20±0.789	2.00	1.00-3.00	2.00-3.00	2.50±0.527	2.50	2.00-3.00	2.00-3.00	.582
FKGL	10.6±1.57	11.4	8.50-12.0	8.97-12.0	10.6±1.45	11.0	8.10-12.0	10.0-11.8	10.1±1.69	9.95	7.30-12.0	8.95-11.9	.867
Gunning Fog	14.8±2.91	15.4	10.2-19.0	12.3-16.8	14.3±1.96	14.4	11.2-16.8	13.1-15.9	14.7±2.82	14.8	11.3-19.0	12.6-16.2	.882

N: Sample size, SD: Standard Deviation, Min: Minimum, Max: Maximum, 25%: 25 quartile values, 75%: 75 quartile values. FKGL: Flesch-Kincaid Grade Level \*There is a statistically significant difference at  $p < .05$ .

**Table 6.** Descriptive statistics and post hoc comparisons of the ‘Headgear Appliance’ section.

Headgear Appliances	Temperature Setting=0 (N=10)				Temperature Setting=0.5 (N=10)				Temperature Setting=1.0 (N=10)				P
	Mean±SD	Median	Min-Max	25%-75%	Mean±SD	Median	Min-Max	25%-75%	Mean±SD	Median	Min-Max	25%-75%	
Level of detail	2.80±0.422	3.00	2.00-3.00	3.00-3.00	2.20±0.919	2.50	1.00-3.00	1.25-3.00	2.40±0.516	2.00	2.00-3.00	2.00-3.00	.165
Fluency	2.20±0.422	2.00	2.00-3.00	2.00-2.00	1.90±0.876	2.00	1.00-3.00	1.00-2.75	2.60±0.843	3.00	1.00-3.00	3.00-3.00	.095
User-focused	2.20±0.789	2.00	1.00-3.00	2.00-3.00	1.80±0.919	1.50	1.00-3.00	1.00-2.75	2.40±0.843	3.00	1.00-3.00	2.00-3.00	.293
FKGL	10.5±1.44	10.4	8.70-12.0	9.03-12.0	10.6±1.14	10.4	9.30-12.0	9.53-11.8	11.5±0.718	11.8	9.80-12.0	11.4-12.0	.183
Gunning Fog	14.3±2.78	12.8	11.5-19.0	12.2-16.1	14.3±2.40	13.9	11.6-18.1	12.1-16.	15.2±1.40	15.0	12.6-17.0	14.5-16.2	.455

N: Sample size, SD: Standard Deviation, Min: Minimum, Max: Maximum, 25%: 25 quartile values, 75%: 75 quartile values. FKGL: Flesch-Kincaid Grade Level \*There is a statistically significant difference at  $p < .05$ .

## DISCUSSION

Recently, in addition to general-purpose chatbots, LLM-based platforms specialized in healthcare have been introduced. These tools offer more controlled and reliable information by citing literature-based references and involving supervision by healthcare professionals during development.<sup>18,19</sup> In our study, we adopted a similar approach by utilizing the OpenAI GPT-3.5 language model on the Microsoft Azure cloud computing platform, specifying the sources of information and adjusting configurations that affect the chatbot's features. We found that different temperature settings (0, 0.5, 1.0) resulted in statistically significant differences in both linguistic characteristics and content.

The level of detail, comprehensiveness, and accuracy of information provided by chatbots depend on the model's architecture and sampling parameters. In the present study, the responses obtained at different temperature values were evaluated using the “level of detail” criterion. It was found that the most detailed and comprehensive information was produced at a temperature setting of 0. This is believed to be due to the model's tendency to convey information with minimal modifications at lower temperature settings. Lysandrou et al.<sup>20</sup> evaluated the responses of three different GPT versions to questions related to diabetes and atopic dermatitis. The authors found that responses were more consistent and accurate at a temperature setting of 0, whereas higher temperature settings led to decreased repeatability and accuracy.<sup>20</sup> Winter assessed the performance of ChatGPT on national high school exams and discovered that at very high temperature settings, the model produced nonsensical text instead of correct answers.<sup>21</sup> These findings suggest that lower temperature values approaching zero facilitate more precise, comprehensive, and detailed responses. However, when diversity and a high degree of predictive variability are required for specific applications, the optimal temperature value may deviate from zero. Han et al.<sup>22</sup> compared the 10-year cardiovascular risk prediction capabilities of GPT-4 by adjusting the temperature from 0 to 1 in increments of 0.2. Their research indicated that the model's predictions were most stable and predictable at a temperature setting of 0.4.

An increase in the temperature parameter encourages the model to select less probable tokens, thereby enhancing the diversity and creativity of its responses. However, when language models are utilized by clinicians and researchers for tasks requiring precision, such as summarising and reporting patient data or searching the literature, low creativity and deterministic expressions are advantageous.<sup>11,14,23</sup> For non-specialist users and patients, chatbots employing a more creative and colloquial style may enhance usability and appeal. Increased creativity can enable the model to generate attention-grabbing responses that are easily understood by the audience, thereby simplifying complex medical information. In our study, evaluations across all questions indicated that at a temperature setting of 1.0, higher scores were attained for the user-focused understandability and fluency criteria, with the difference in user-focused understandability being statistically significant. The customized chatbot generated clearer and more comprehensible text at higher temperature settings by incorporating more explanations of terminology.

Across different categories, the highest scores for fluency and user-focused understandability were observed at a temperature setting of 1.0. However, within the Facemask Therapy category, the highest average scores for both criteria were found at a temperature setting of 0.5. This difference, although not statistically significant, may be attributed to variations in the linguistic characteristics of the data sources prepared for this category. Davis et al. tasked ChatGPT with generating a tweet for a public audience, a title for a scientific journal publication, and a title for a keynote speech using the abstract of a recently published paper at three different temperature settings.<sup>25</sup> They found that at temperature settings of 0 and 0.5, the responses exhibited clearer and simpler language characteristics.<sup>25</sup> These findings suggest that the optimal temperature setting may vary depending on the context and specific requirements of the task. While higher temperature values have the potential to produce more fluent, understandable, and user-friendly responses, when information needs to be generated faithfully to the original data, the ideal temperature setting should be carefully tailored to the intended use.

Within the field of health literacy, the readability of texts plays a crucial role in determining the educational levels of individuals who can comprehend them. The American Medical Association (AMA) and the National Institutes of Health (NIH) recommend that patient-facing texts be written at reading levels no higher than sixth and eighth grades, respectively.<sup>24</sup> In our study, we assessed the effects of different temperature settings on the readability of the responses using the Gunning Fog and FKGL indices. No significant differences were found in readability across the temperature settings. According to the average FKGL and Gunning Fog scores, the responses required at least a 10th-grade reading level and a college-level reading level, respectively. In a recent study, Kılınç and Mansız examined the reliability and readability of ChatGPT's responses to frequently asked questions (FAQs) about orthodontics in both older and updated model versions.<sup>26</sup> Their study utilized real patient questions and employed the DISCERN tool and FKGL index. While their results indicated that the updated ChatGPT version produced more reliable responses, the Flesch Reading Ease scores decreased, meaning that the readability of the new answer texts became more difficult. Specifically, the scores declined from 29.28 to 25.12 for general questions and from 47.67 to 41.60 for treatment-related questions.<sup>26</sup> Cheong et al. evaluated the readability of information provided by ChatGPT and Google Bard on obstructive sleep apnea using the FKGL index.<sup>27</sup> They reported average FKGL scores of 9.0 for ChatGPT and 5.9 for Google Bard.<sup>27</sup> Simplifying sentence structure and reducing linguistic complexity can lower readability scores without compromising content quality. This would make the responses more accessible to a wider audience.

## REFERENCES

1. Crispino R, Mannocci A, Dilena IA, Sides J, Forchini F, Asif Alherawi WM et al. Orthodontic patients and the information found on the web: a cross-sectional study. *BMC Oral Health*. 2023;23(1):860. doi: 10.1186/s12903-023-03609-4
2. Wade N, Paul N, Nagar N, Rolland S, Germain S. Information-seeking behaviour in patients exploring orthognathic surgery: A qualitative study. *J Orthod*. 2025;52(1):63-71. doi: 10.1177/14653125241249494
3. Cheong RCT, Unadkat S, Mcneillis V, Williamson A, Joseph J, Randhawa P et al. Artificial intelligence chatbots as sources of patient education material for obstructive sleep apnoea: ChatGPT versus Google Bard. *Eur Arch Otorhinolaryngol*. 2024;281(2):985–993. doi: 10.1007/s00405-023-08319-9
4. Görtz M, Baumgärtner K, Schmid T, Muschko M, Woessner P, Gerlach A et al. An artificial intelligence-based chatbot for prostate cancer education: Design and patient evaluation study. *Digit Health*. 2023;9:1-11. doi:10.1177/20552076231173304
5. Alkaissi H, McFarlane SI. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus*. 2023;15(2):e35179. doi: 10.7759/cureus.35179
6. Aljamaan F, Temsah MH, Altamimi I, Al-Eyadhy A, Jamal A, Alhasan K, et al. Reference Hallucination Score for Medical Artificial Intelligence Chatbots: Development and Usability Study. *JMIR Med Inform*. 2024;12:e54345. doi:10.2196/54345
7. Mirzadeh I, Alizadeh K, Shahrokhi H, Tuzel O, Bengio S, Farajtabar M. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models. *arXiv preprint*. arXiv:2410.05229 2024. doi:10.48550/arXiv.2410.05229
8. Shuster K, Poff S, Chen M, Kiela D, Weston J. Retrieval Augmentation Reduces Hallucination in Conversation. Findings of the Association for Computational Linguistics. *arXiv preprint*. arXiv:2104.07567 2021. doi:10.48550/arXiv.2104.07567

## CONCLUSION

The temperature setting significantly influences both the content and linguistic features of chatbot-generated information. In our study, we found that lowering the temperature setting increased the level of detail and comprehensiveness of the content, while raising the temperature enhanced user-focused understandability and fluency. By optimizing chatbot parameters, such as temperature, and customizing information sources, it is possible to deliver controlled, patient-specific information with the desired linguistic characteristics.

9. Renze M, Guven E. The Effect of Sampling Temperature on Problem Solving in Large Language Models. *arXiv preprint*. arXiv:2402.05201 2024. doi: 10.18653/v1/2024.findings-emnlp.432
10. Davis J, Van Bulck L, Durieux BN, Lindvall C. The Temperature Feature of ChatGPT: Modifying Creativity for Clinical Research. *JMIR Hum Factors*. 2024;11:e53559. doi: 10.2196/53559
11. Veen D Van, Uden C Van, Blankemeier L, Delbrouck J-B, Aali A, Bluethgen C, et al. Clinical Text Summarization: Adapting Large Language Models Can Outperform Human Experts. *Res Sq*. 2024;30(4):1134-42. doi:10.21203/rs.3.rs-3483777/v1
12. Karlstrand J, Nielsen A. Generative AI Assistant for Public Transport Using Scheduled and Real-Time Data [dissertation]. Linköping, Sweden: Linköping University; 2024. Available from: <https://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-204561>.
13. Zhuang Z, Yang Z, Li K, Shi P, Liu X, Zhang S, et al. Performance of ChatGPT on the American Endocrine Society Self-Assessment Test. *Available at SSRN 4658115*.
14. Hasani AM, Singh S, Zahergivar A, Ryan B, Nethala D, Bravomontenegro G, et al. Evaluating the performance of Generative Pre-trained Transformer-4 (GPT-4) in standardizing radiology reports. *Eur Radiol*. 2024;34(6): 3566-3574. doi: 10.1007/s00330-023-10384-x
15. Barua A, Brase G, Dong K, Hitzler P, Vasserman E. On the Psychology of GPT-4: Moderately anxious, slightly masculine, honest, and humble. *arXiv preprint*. arXiv:2402.01777 2024. doi:10.48550/arXiv.2402.01777
16. Flesch R. A new readability yardstick. *J Appl Psychol*. 1948;32(3):221-233. doi: 10.1037/h0057532
17. Gunning R. The Fog Index After Twenty Years. *J Bus Commun*. 1969;6(2):3-13. doi: 10.1177/002194366900600202
18. Chow JCL, Sanders L, Li K. Impact of ChatGPT on medical chatbots as a disruptive technology. *Front Artif Intell*. 2023;6:1166014. doi:10.3389/frai.2023.1166014
19. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health records. *npj Digit Med*. 2022;5:194. doi: 10.1038/s41746-022-00742-2
20. Lysandrou G, Owen RE, Mursec K, Brun G Le, Fairley EAL. Comparative Analysis of Drug-GPT and ChatGPT LLMs for Healthcare Insights: Evaluating Accuracy and Relevance in Patient and HCP Contexts. *arXiv preprint*. arXiv:2307.16850 2023. doi: 10.48550/arXiv.2307.16850
21. de Winter JCF. Can ChatGPT Pass High School Exams on English Language Comprehension? *Int J Artif Intell in Educ*. 2024;34:915-930. doi:10.1007/s40593-023-00372-z
22. Han C, Kim DW, Kim S, You SC, Park JY, Bae S, et al. Evaluation of GPT-4 for 10-year cardiovascular risk prediction: Insights from the UK Biobank and KoGES data. *Iscience*. 2024;27(2):109022. doi:10.1016/j.isci.2024.109022
23. Bhavya B, Xiong J, Zhai CX. Analogy Generation by Prompting Large Language Models: A Case Study of InstructGPT. *arXiv preprint*. arXiv:2210.04186 2022. doi:10.48550/arXiv.2210.04186
24. Weiss BD. *Health Literacy*. Chicago, IL: American Medical Association; 2003:253.
25. Davis J, Bulck LV, Durieux BN, et al. The Temperature Feature of ChatGPT: Modifying Creativity for Clinical Research. *JMIR*. 2024;11:e53559. doi:10.2196/53559
26. Kılınc DD, Mansuz D. Examination of the reliability and readability of Chatbot Generative Pretrained Transformer's (ChatGPT) responses to questions about orthodontics and the evolution of these responses in an updated version. *Am J Orthod Dentofacial Orthop*. 2024;165(5): 546-555. doi:10.1016/j.ajodo.2023.11.012
27. Cheong RCT, Unadkat S, Mcneillis V, et al. Artificial intelligence chatbots as sources of patient education material for obstructive sleep apnoea: ChatGPT versus Google Bard. *Eur Arch Otorhinolaryngol*. 2024;281:985-993. doi: 10.1007/s00405-023-08319-9