Are AI Models Accurate in Assisting Hypothyroidism Patients?

Gloria Wu MD, MBA[2] ,Hrishi Paliath-Pathiyal[1], Anjali Madala[2], Milan del Buono[3], Obaid Khan[4]

[1]Nova Southeastern University. [2]University of California, San Francisco. [3]University of California, Berkeley. [4]California Health Sciences University.

Abstract:

Background: Approximately 13 million people in the United States, representing 4.78% of the population, suffer from undiagnosed endocrine disorders. Hypothyroidism affects up to 5% of the general population, with another 5% estimated to be undiagnosed. Most cases in Americans older than the age of 11 have mild or minimal obvious symptoms.
Purpose: Determine if AI chatbots can give accurate information about hypothyroidism and correctly judge the accuracy of their own responses.
Methods: Questions were asked in English to Claude, Cohere, Gemini, GPT 4o Mini, and GPT 4o. text responses from each chatbot were recorded and scored from a scale of 1 to 5, with 1 indicating a highly inaccurate response and 5 suggesting an accurate and advanced response. A series of paired T-tests were used to compare the difference between manual and AI scores, score difference = (Manual - AI). P was adjusted by the Bonferroni Correction. For the Manual vs. AI rated scores across languages and chatbots, random jitter was used to better visualize data grouping and trends for the scatterplot. The raters for the languages all combined to rate English as well, to remove bias, and were blinded to the chatbot they were evaluating.
Results: By chatbot, ChatGPT4o outperformed Cohere (t=3.209, df=29, p adj =0.032), Claude outperformed Cohere (t=3.914, df=29, p adj =0.005), and Gemini also outperformed Cohere (t=4.455, df=29, p adj =0.001). Pearson correlation coefficient of 0.417, suggesting a moderate positive correlation within the manual vs. AI scores.
Discussion: 1. While ChatGPT-4o performed the best among chatbots, it was also the only model to require a paid subscription, making accurate information less accessible to individuals in lower socioeconomic brackets who may not be able to afford paid chatbots. AI-predicted scores were generally not accurate compared to manual scores, and there was only one instance where an AI model self-scored below 3, while 16 responses were human-rated below 3. AI models are overconfident in their responses, and chatbots may repeatedly provide incorrect information when prompted. Cohere was consistently outperformed by Claude, Gemini, and ChatGPT-4o/4o mini, highlighting its need for training on more diverse datasets.
Conclusion: Our data suggests a need for more accessible and affordable large language models trained on medically succinct datasets for patients.

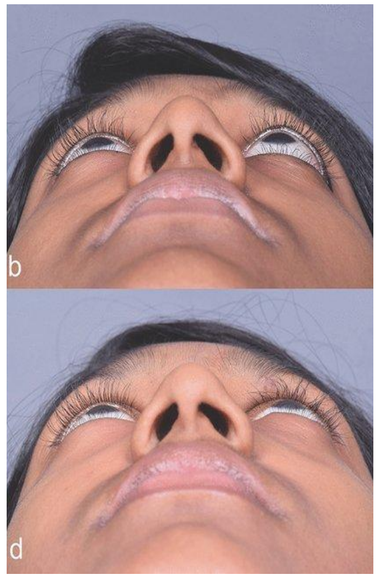
Background:

- Approximately 13 million people in the United States, representing 4.78% of the population, suffer from undiagnosed endocrine disorders.[1]
- Hypothyroidism affects up to 5% of the general population, with another 5% estimated to be undiagnosed.[2]
- Most cases in Americans older than the age of 11 have mild or minimal obvious symptoms.[3]

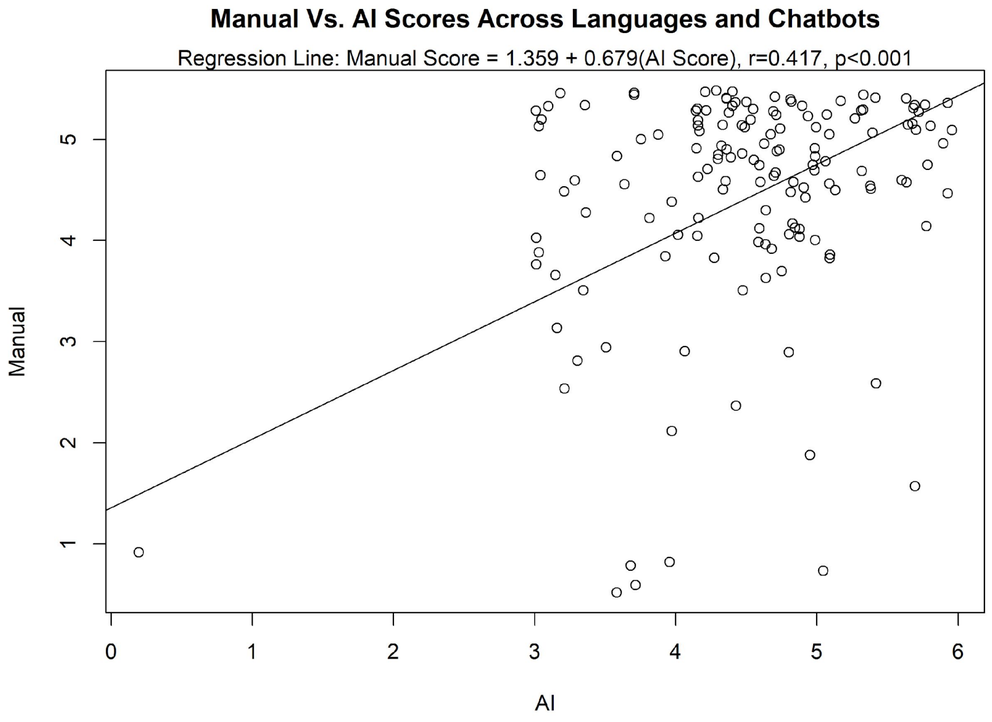
Purpose: Determine if AI chatbots can give accurate information about hypothyroidism and correctly judge the accuracy of their own responses.

Methods:

- **Queries:**
  1. What is hypothyroidism?
  2. I am a 40 year old woman with dry skin, hair loss, and a bump in my throat. What could be causing my symptoms?
  3. Who is at risk for hypothyroidism?
  4. What are some common symptoms of hypothyroidism?
  5. I am currently diagnosed with hypothyroidism. What changes should I make so that I can alleviate symptoms?
- LLMS: ChatGPT-4o, ChatGPT-4o Mini, Claude, Cohere, Gemini, Grok.
- Text responses from each chatbot were recorded and scored from a scale of 1 to 5, with 1 indicating a highly inaccurate response and 5 suggesting an accurate and advanced response.
- A series of paired T-tests were used to compare the difference between manual and AI scores, score difference = (Manual – AI). P was adjusted by the Bonferroni Correction.
- For the Manual vs. AI rated scores across languages and chatbots, random jitter was used to better visualize data grouping and trends for the scatterplot. The raters for the languages all combined to rate English as well, to remove bias, and were blinded to the chatbot they were evaluating.



Results:



Word Count					
	Q1	Q2	Q3	Q4	Q5
ChatGPT-4o mini	71	82	123	138	283
ChatGPT-4o	330	120	212	109	476
Claude	169	183	142	104	261
Cohere	278	256	362	216	467
Gemini	154	186	149	361	358
Grok	97	247	278	205	826
Average	183.17	179	211	188.83	445.17

Flesch-Kincaid					
	Q1	Q2	Q3	Q4	Q5
ChatGPT-4o mini	13.5	12.2	10	13	10.9
ChatGPT-4o	16.1	11.5	9.1	10.9	10.9
Claude	12.7	12.3	8.3	9.5	10.2
Cohere	18.5	11.5	12.9	14.2	12.7
Gemini	16.9	13.7	14	10.9	13.9
Grok	14.6	11	13.1	9.7	9.4
Average	15.38	12.03	11.23	11.37	11.33

The authors have no financial conflicts of interest.

Results:

1. By chatbot, ChatGPT4o outperformed Cohere (t=3.209, df=29, p adj =0.032), Claude outperformed Cohere (t=3.914, df=29, p adj =0.005), and Gemini also outperformed Cohere (t=4.455, df=29, p adj =0.001)
2. The pearson correlation coefficient of 0.417, suggests a moderate positive correlation within the manual vs. AI scores. This was done through the use of random jitter which reveals that the AI model rated itself higher than the human scorer.
3. While ChatGPT-4o performed the best among chatbots, it was also the only model to require a paid subscription, making accurate information less accessible to individuals in lower socioeconomic brackets who may not be able to afford paid chatbots.
4. AI-predicted scores were generally not accurate compared to manual scores, and there was only one instance where an AI model self-scored below 3, while 16 responses were human-rated below 3. This suggests AI models are overconfident in their responses, and chatbots may repeatedly provide incorrect information when prompted.
5. Cohere was consistently outperformed by Claude, Gemini, and ChatGPT-4o/4o mini, highlighting its need for training on more diverse datasets.

Conclusion:

- Our small study shows that there is a need for accessible and affordable large language models, capable of succinct answers to patient queries

References:

1. Wu, Junyun, et al. “Global, Regional and National Burden of Endocrine, Metabolic, Blood and Immune Disorders 1990-2019: A Systematic Analysis of the Global Burden of Disease Study 2019.” *Frontiers*, Frontiers, 3 Apr. 2023, www.frontiersin.org/journals/endocrinology/articles/10.3389/fendo.2023.1101627/full.
2. Chiovato, L., Magri, F., & Carlé, A. (2019). Hypothyroidism in Context: Where We've Been and Where We're Going. *Advances in therapy*, *36*(Suppl 2), 47–58. https://doi.org/10.1007/s12325-019-01080-8
3. Patil N, Rehman A, Jialal I. Hypothyroidism. In: *StatPearls [Internet].* StatPearls Publishing; 2020. Updated August 10, 2020. Accessed January 5, 2021. www.ncbi.nlm.nih.gov/books/NBK519536
4. Vasanthapuram, Varshitha Hemanth & Naik, Milind. (2022). Blepharoptosis in thyroid eye disease: etiopathogenesis, clinical features and correlation with thyroid eye disease. International Ophthalmology. 42. 10.1007/s10792-021-01992-x.

Can AI Chatbots Provide Accurate Information About Thyroid Eye Disease?

Hrishi Paliath-Pathiyal[1], Gloria Wu MD, MBA[2], Brian Hoang, BS[3], Milan Del Buono[4],Adam Shams[2], [5]Obaid Khan

[1]Nova Southeastern University, [2]University of California, San Francisco, [3]University of California, Davis, [4]University of California, Berkeley, [5]California Health Sciences University

Abstract:

Background: Thyroid eye disease affects about 0.25% of people and is more common in women (16 per 100,000) than men (2.9 per 100,000).In patients with Graves' Disease, the incidence of
Thyroid Eye Disease can be between 25% and 40%. Patients who develop thyroid eye disease, additionally, have an elevated risk of developing other ocular symptoms such as dry eye disease.
Purpose: Evaluate the ability and accuracy of AI models in correctly diagnosing and providing information about thyroid eye disease.
Methods: Questions were asked in English to Claude, Cohere, Gemini, GPT 4o Mini, and GPT 4o.text responses from Claude, Cohere, Gemini, GPT 4o Mini, and GPT 4o were recorded and translated with help from native speakers. Manual and AI Scores were rated on a scale from 1 to 5, 5 being the most accurate response with 1 being the least accurate response. A series of paired T-tests were used to track the difference between scores, with score difference being calculated as the AI self-score subtracted from the manual score.
Results: Our R-value, or Pearson's Correlation Coefficient, of 0.505 indicates a mild positive correlation between the manual and AI Scores across the LLMs. Therefore, both of the manual and AI scores amongst the various chatbots are fairy accurate. The manual scores were not significantly different among the different LLMs (one-way ANOVA: F(4,15) = 2.28, p = 0.108). AI rated scores and manual scores relatively align among different chatbots, suggesting that chatbots were accurate in assessing the quality of their responses.
Conclusion: There is a need for further training of these LLMs on more diverse datasets when queried about less common diseases.

Background:

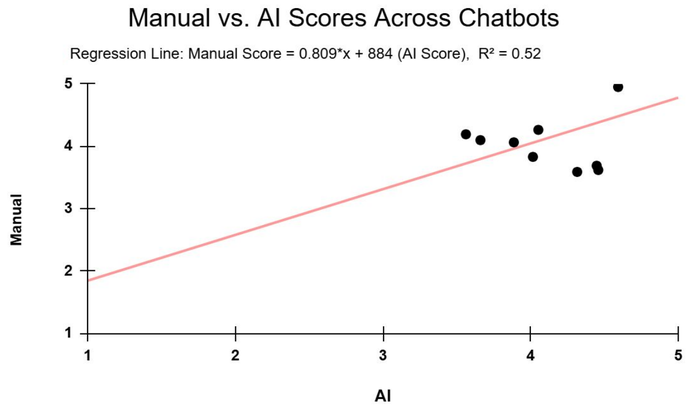
- Thyroid eye disease affects about 0.25% of people and is more common in women (16 per 100,000) than men (2.9 per 100,000).[1]
- In patients with Graves' Disease, the incidence of Thyroid Eye Disease can be between 25% and 40%.[2]
- Patients who develop thyroid eye disease, additionally, have an elevated risk of developing other ocular symptoms such as dry eye disease.[3]

Purpose: Evaluate the ability and accuracy of AI models in correctly diagnosing and providing information about thyroid eye disease.

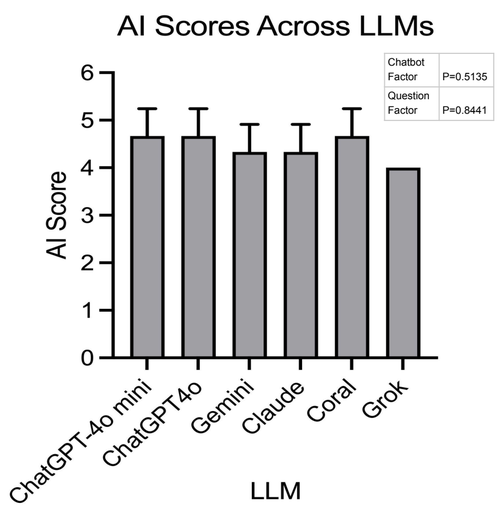
Methods:

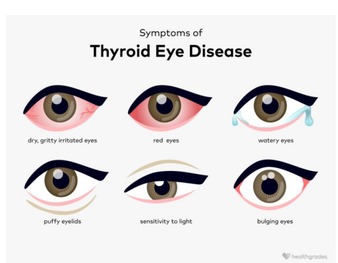
- **Queries:**
  1. What is thyroid eye disease?
  2. I have dry eye and my eyes feel funny when I look to the left. Is this thyroid eye disease?
  3. Who gets thyroid eye disease?
- LLMS: ChatGPT-4o, ChatGPT-4o mini, Gemini, Claude, Coral, Grok.
- Manual and AI Scores were rated on a scale from 1 to 5, 5 being the most accurate response with 1 being the least accurate response.
- The AI scores were obtained by ChatGPT-4o mini.
- A series of 5 paired T-tests were used to track the difference between scores, and difference between LLMs, with score difference being calculated as the AI self-score subtracted from the manual score.

Results:



	Question	Mentions Endocrinologist	Mentions Ophthalmologist	Mentions Hyperthyroidism	Includes Disclaimer/Links	Total Mentions
ChatGPT-4o	Q1	YES	YES	YES	NO	3/4
	Q2	NO	YES	YES	NO	2/4
	Q3	YES	YES	YES	NO	3/4
ChatGPT-4o Mini	Q1	YES	YES	YES	YES	4/4
	Q2	YES	YES	YES	YES	4/4
	Q3	YES	YES	YES	YES	4/4
Claude	Q1	NO	NO	YES	NO	1/4
	Q2	YES	YES	YES	NO	3/4
	Q3	YES	YES	YES	NO	3/4
Gemini	Q1	NO	YES	YES	YES	3/4
	Q2	YES	YES	YES	YES	4/4
	Q3	YES	YES	YES	YES	4/4
Cohere	Q1	YES	YES	YES	YES	4/4
	Q2	YES	YES	YES	YES	4/4
	Q3	YES	YES	YES	YES	4/4
Grok	Q1	YES	YES	YES	YES	4/4
	Q2	NO	YES	YES	YES	3/4
	Q3	NO	NO	YES	YES	2/4





Results:

- Our R-value, or Pearson's Correlation Coefficient, of 0.505 indicates a mild positive correlation between the manual and AI Scores across the LLMs. Therefore, the manual and AI scores amongst the chatbots are fairly accurate.
- The manual scores were not significantly different among the different LLMs (one-way ANOVA: F(4,15) = 2.28, p = 0.7008).
- The LLMs gave itself a score of 4 or higher 37.5% of the time as compared to 25% as done by the human scorer. This demonstrates overconfidence and self-serving bias in the AI response assessment.
- Cohere and ChatGPT-4o Mini consistently mentioned endocrinologists in 100% of responses vs. Grok's 33%
- ChatGPT-4o Mini and Cohere included medical disclaimers in all responses while ChatGPT-4o provided none
- Grok showed the most variable performance with mention rates ranging from 33-100% across categories
- ChatGPT-4o, ChatGPT-4o Mini, Cohere, and Gemini achieved perfect ophthalmologist recommendation rates compared to other models

Conclusion:

- There is a need for more diverse and robust LLM training which reflect disease susceptibility that may otherwise be unknown to the general public.

References:

1. McAlinden C. (2014). An overview of thyroid eye disease. *Eye and vision (London, England)*, *1*, 9. https://doi.org/10.1186/s40662-014-0009-8
2. Smith, T. J., Hegedüs, L., Lesser, I., Perros, P., Dorris, K., Kinrade, M., Troy-Ott, P., Wuerth, L., & Nori, M. (2023). How patients experience thyroid eye disease. *Frontiers in endocrinology*, *14*, 1283374. https://doi.org/10.3389/fendo.2023.1283374
3. Alanazi, S. A., Alomran, A. A., Abusharha, A., Fagehi, R., Al-Johani, N. J., El-Hiti, G. A., & Masmali, A. M. (2019). An assessment of the ocular tear film in patients with thyroid disorders. *Clinical ophthalmology (Auckland, N.Z.)*, *13*, 1019–1026. https://doi.org/10.2147/OPTH.S210044
4. (2024). Drphelps.com. https://drphelps.com/wp-content/uploads/sites/236/2023/12/resized_Tepezza_2_treatments.jpg.webp
5. *Thyroid Eye Disease: Symptoms, Causes, Treatments.* (2021, February 12). Healthgrades. https://www.healthgrades.com/right-care/thyroid-disorders/thyroid-eye-disease

The authors have no financial conflicts of interest.

Can AI Large Language Models Assist Patients With Hyperthyroidism?

[5]Sahej Sidhu, Hrishi Paliath-Pathiyal[1], Gloria Wu MD,MBA[2], Milan del Buono[4,] Obaid Khan[3]

[1]Nova Southeastern University. [2]University of California, San Francisco. [3]California Health Sciences University. [4]University of California, Berkeley. [5]Santa Clara University.

Abstract:

Background: 1.2% of people in the United States have hyperthyroidism. Hyperthyroidism can be developed as a result of Graves' Disease, toxic adenoma, and a toxic multinodular goiter. Over 250 million people worldwide used AI software in 2023, and that number is forecast to increase significantly.
Purpose: Determine if different AI chatbots can provide correct information about hyperthyroidism across different languages.
Methods: Questions were asked in English, Chinese, Hindi, Japanese, Korean, and Punjabi to five chatbots, Claude, Cohere, Gemini, GPT 4o Mini, and GPT 4o. text responses from Claude, Cohere, Gemini, GPT 4o Mini, and GPT 4o were recorded and translated with help from native speakers. Responses were manually scored on a 1-5 scale.
Results: A T-test paired run with Bonferroni corrections showed that across chatbots, English responses had more words than Chinese (t=8.309, df=24, p adj <0.001), Punjabi (t=4.881, df=24, p<0.001), Hindi (t=8.385, df=24, p adj <0.001), Japanese (t=10.096, df=24, p adj <0.001), Korean (t=4.581, df=24, p adj <0.001). (Table 1). Hindi had a higher word count than Japanese, as did Chinese, Korean, and Punjabi. A regression analysis was carried out to determine if the length of an output was correlated to its accuracy. No correlation was observed, with the regression equation (Manual Score) = 4.305 - 0.000325(Text Length) being significant for the intercept (p<0.001) but not the slope (p=0.168). No significant difference between the different chatbots' outputs (across all languages) was observed (lowest p adj = 0.428). No correlation was observed between response length and accuracy based on scoring, indicating that text length does not significantly affect accuracy of responses. The clear hierarchy in language accuracy (English, Chinese > Hindi, Korean, Japanese > Punjabi) suggests a significant disparity in the quality of medical information provided across different languages.
Conclusion: More training on linguistically and medically diverse datasets is needed to make responses more concise and readable.

Background:

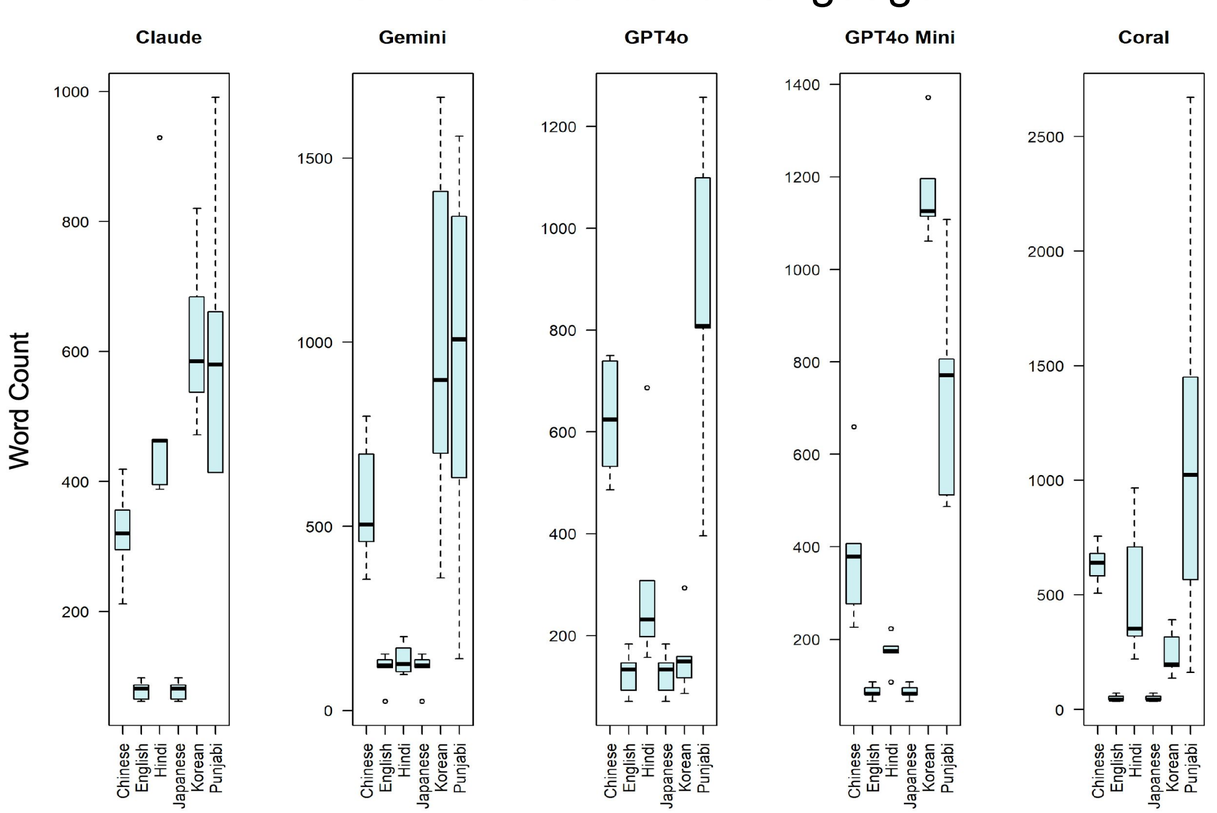
- 1.2% of people in the United States currently suffer from hyperthyroidism.[1]
- Hyperthyroidism can be developed as a result of Graves' Disease, toxic adenoma, and a toxic multinodular goiter.[1]
- Over 250 million people worldwide used AI software in 2023, and that number is forecast to increase significantly.[2]

Purpose: Determine if different AI chatbots can provide correct information about hyperthyroidism across languages.

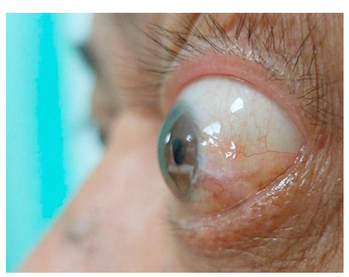
Methods:

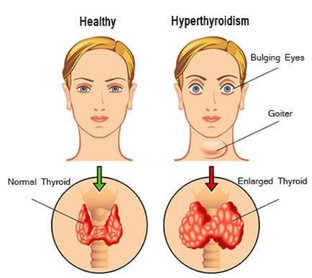
- **Queries:**
  1. What is hyperthyroidism?
  2. I am a 40 year old Asian male with excessive sweating, arrhythmia, and my eyes are bulging. What could be causing my symptoms?
  3. Who is at risk for hyperthyroidism?
  4. What are some common symptoms of hyperthyroidism?
  5. I am currently diagnosed with hyperthyroidism. What changes should I make so that I can alleviate symptoms?
- LLMs: ChatGPT-4o, ChatGPT-4o Mini, Claude, Gemini, Coral.
- Responses from Claude, Cohere, Gemini, Chat GPT-4o-Mini, and GPT 4o were recorded and translated with help from native speakers.
- Manual and AI Scores were rated on a scale from 1 to 5, 5 being the most accurate response with 1 being the least accurate response.
- The AI responses were scored by ChatGPT-4o mini.

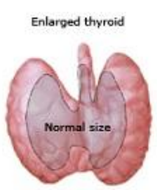
Results:

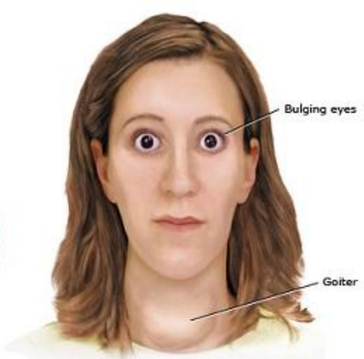


Comparison	T-value	Degrees of Freedom (df)	p-adjusted
Hindi vs. Japanese	4.881	24	< 0.001
Hindi vs. Chinese	11.36	24	< 0.001
Hindi vs. Korean	5.775	24	< 0.001
Hindi vs. Punjabi	7.242	24	< 0.001









Results:

1. A T-test paired run with Bonferroni corrections showed that across chatbots, English responded with more words than Chinese (t=8.309, df=24, p adj <0.001), Punjabi (t=4.881, df=24, p<0.001), Hindi (t=8.385, df=24, p adj <0.001), Japanese (t=10.096, df=24, p adj <0.001), Korean (t=4.581, df=24, p adj <0.001).
2. Hindi had a higher word count than Japanese, as did Chinese, Korean, and Punjabi.
3. A regression analysis was carried out to determine if the length of an output was correlated to its accuracy. No correlation was observed, with the regression equation (Manual Score) = 4.305 – 0.000325 (Text Length) being significant for the intercept (p<0.001) but not the slope (p=0.168).
4. No significant difference between the different chatbots' outputs (across all languages) was observed (lowest p adj = 0.428).
5. The clear hierarchy in language accuracy (English, Chinese > Hindi, Korean, Japanese > Punjabi) suggests a significant disparity in the quality of medical information provided across different languages.
6. Punjabi and Korean responses were the wordiest and the least accurate. The takeaway is simple: more words indicate a low quality response. In these languages, verbosity seems to introduce error, showcasing a need for short, simple, and more accurate responses from these LLMs.

Conclusion:

- No correlation was observed between response length and accuracy based on scoring, indicating that text length does not significantly affect accuracy of responses.
- More training on linguistically and medically diverse datasets is needed to make LLM responses more concise and readable.

References:

1. Doubleday, A. R., & Sippel, R. S. (2020). Hyperthyroidism. *Gland surgery*, *9*(1), 124–135. https://doi.org/10.21037/gs.2019.11.01
2. *Worldwide AI tool users 2030*. (2024, February 13). Statista. https://www.statista.com/forecasts/1449844/ai-tool-users-worldwide#:~:text=People%20using%20AI%20tools%20globally
3. *Hyperthyroidism | RMI*. (n.d.). Rmi.edu.pk. https://rmi.edu.pk/disease/hyperthyroidism
4. *Overactive Thyroid | Filipino Doctors*. (n.d.). Filipinodoctors.org. https://filipinodoctors.org/overactive-thyroid/
5. Oculofacial Surgery and Cosmetic Laser Institute. (n.d.). *Thyroid Eye Disease*. Retrieved June 20, 2025, from https://www.doctorrosh.com/services/functional/thyroid-eye-disease/

The authors have no financial conflicts of interest.

ADCES7 Guidelines and AI Chatbots: Do They Help Our Patients?

Gloria Wu MD,MBA[2], Hrishi Paliath-Pathiyal[1], Milan del Buono[4],Obaid Khan[3]

[1]Nova Southeastern University, [2]University of California, San Francisco, [3]California Health Sciences University, [4]University of California, Berkeley

Abstract:

Background: In 2020, 34.2 million US adults had diabetes and 88 million had pre-diabetes. The Association of Diabetes Care & Education Specialists (ADCES) lists seven categories of self-care behaviors for diabetics: Healthy Coping, Healthy Eating, Being Active, Taking Medication, Monitoring, Reducing Risk, and Problem Solving, collectively known as ADCES7.
Purpose: Do responses by AI chatbots fulfill ADCES7 guidelines?
Methods: Chatbots were prompted to generate important questions for each of the ADCES7 guidelines. The questions were then fed back to the chatbots and responses were recorded. The responses were then scored manually on a numerical scale from 1 to 5 based on accuracy and adherence to ADCES7 standards.
Results: The results showed that average scores from the 5 LLMs regarding information about insulin was 1.77% lower than the average scores about diabetes monitoring. Chatbots generally performed well for questions about the seven main ADCES7 guidelines, but performed poorly about providing relevant information regarding insulin, insulin pumps, and GLP-1 analog medications. This highlights a critical gap in the current ADCES7 guidelines.
Conclusion: The information in the ADCES7 handouts does not mention newer treatments or contain the word "insulin," suggesting a gap in provided information and the need for an additional category focused on newer and more advanced treatment options, including insulin and GLP-1. Future research should also prioritize the development of more comprehensive guidelines and the enhancement of AI chatbot training to address knowledge gaps, ensuring that digital health tools can provide accurate, up-to-date information across all aspects of diabetes care, including the latest therapeutic advancements. The use of ADCES guidelines and information from other expert associations could be used as training texts for chatbots.

Background:

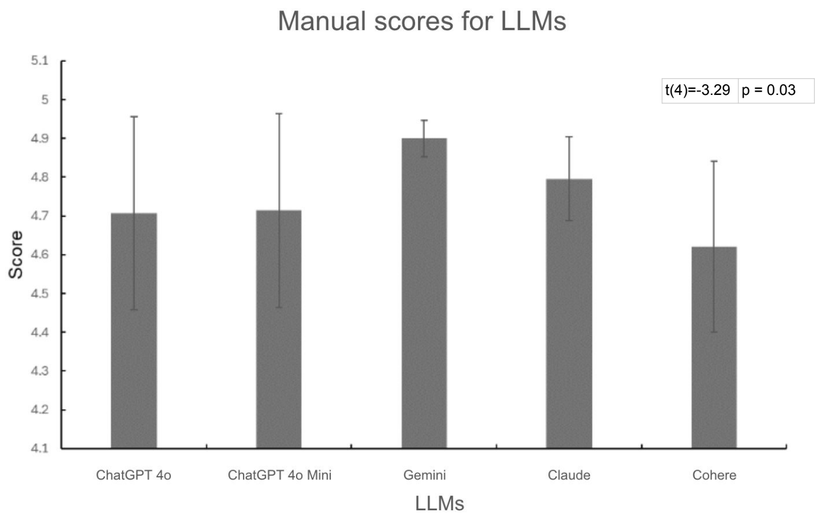
- In 2020, 34.2 million US adults had diabetes and 88 million had pre-diabetes.[1]
- The Association of Diabetes Care & Education Specialists (ADCES) lists seven categories of self-care behaviors for diabetics: Healthy Coping, Healthy Eating, Being Active, Taking Medication, Monitoring, Reducing Risk, and Problem Solving, collectively known as ADCES7.[2]
- A separate study was performed after the original submission, in which the outputs from the five LLMs were gathered, and ratings were generated by ChatGPT
- The AI ratings were compared to each other, as well as the human ratings to evaluate the knowledge and accuracy of chatbots in self-evaluations, as many LLMs are trained through user queries.

Purpose: Determine if responses generated by AI chatbots fulfill ADCES7 guidelines to meet patient needs.

Methods:

- Chatbots were prompted to generate important questions for each of the ADCES7 guidelines. The questions were then fed back to the chatbots and responses were recorded.
- LLMS: ChatGPT-4o, ChatGPT-4o Mini, Gemini, Claude, Cohere.
- Responses were scored manually on a numerical scale from 1 to 5 based on accuracy and adherence to ADCES7 standards. The scale was designed as follows:
  - 1 = Incorrect. **100% NO or irrelevant.**
  - 2 = Has around 1-2 correct pieces of info. **75% is NO**
  - 3 = Not fully correct (3-4). **50% is NO**
  - 4 = Mostly correct (~4-5). **25% is NO.**
  - 5 = Totally correct. **0% is NO.**
- LLM self-rated scores were obtained by having each chatbot evaluate its own responses using identical scoring criteria (1-5 scale) and ADCES7 adherence standards as the manual evaluation.

Results:



2

Manual Scores by Category

Chatbots	Healthy Coping	Healthy Eating		Being active	Medication	Monitoring				Reducing risks	Problem solving	Other			Total Score
	Goal Setting	Calorie Counting	Nutrition	Exercise Information	Medication Information	Blood Glucose	Weight	BP	HbA1c			Info about Insulin	Info about Insulin Pumps	GLP-1 Analogs	Average of scores
ChatGPT -4o mini	5	5	5	5	5	5	5	5	5	5	4.9	5	1.5	4.5	4.707
ChatGPT -4o	5	5	5	5	5	5	5	5	5	5	5	5	1.5	4.5	4.714
Gemini	5	5	5	4.7	5	5	5	5	5	5	5	4.6	4.5	4.8	4.9
Claude	5	5	5	4.5	4.95	5	5	5	5	5	5	4.8	4.3	3.6	4.796
Cohere	5	5	5	5	5	5	5	5	5	5	5	4	2.3	3.4	4.621

LLM Self-Rated Scores

GPT 4o	4	5	5	5	5	5	5	4	5	5	5	4	2	3	4.429
GPT 4o mini	3	4	5	4	4	5	3	4	4	4	3	2	1	2	3.429
Gemini	3	4	4	4	4	5	4	3	5	4	4	3	1	2	3.571
Claude	4	4	5	5	5	5	5	3	5	4	4	3	2	2	4
Cohere	5	5	5	5	5	5	5	5	5	5	5	4	2	4	4.643

The authors have no financial conflicts of interest.

Results:

- The average scores from the 5 LLMs regarding information about insulin was 1.77% lower than the average scores about diabetes monitoring (Independent Sample t-test, t(4)=-3.29, p = 0.03).
- SEM = Standard Error of the Mean. N = 14.
- No significant differences were observed between chatbots using ANOVA, but a t-test revealed a significant discrepancy between scores for the insulin/other category and the monitoring category
- Chatbots met the requirement for the seven main ADCES7 guidelines but performed poorly about providing relevant information regarding insulin, insulin pumps, and GLP-1 analog medications.
- Gemini achieved the highest overall manual score (4.9) while Cohere scored lowest (4.621) across all evaluation categories.
- Gemini showed the largest discrepancy between manual (4.9) and self-rated (3.571) scores, indicating potential underconfidence in self-assessment.
- All chatbots performed poorly in insulin pump information (1.5-4.5 range) and GLP-1 analog coverage (3.4-4.8 range) compared to other categories.

Conclusion:

- Future research should develop comprehensive AI chatbot training sets to ensure accurate and current information.

References:

1. CDC. National Diabetes Statistics Report, 2020. Centers for Disease Control and Prevention. Published February 11, 2020. Accessed January 17, 2022. https://www.cdc.gov/diabetes/library/features/diabetes-stat-report.html
2. ADCES7 Self-Care Behaviors- The Framework for Optimal Self-Management. Accessed January 17, 2022. https://www.diabeteseducator.org/practice/practice-tools/app-resources/the-aade7-self-care-behaviors-the-framework-for-optimal-self-management

Chatbots & Obesity: Are the Responses Accurate?

Hrishi Paliath-Pathiyal[1], Gloria Wu MD,MBA[2], Milan del Buono[4], Ivan Chim[5], Obaid Khan[3]

[1]Nova Southeastern University, [2]University of California, San Francisco, [3]California Health Sciences University, [4]University of California, Berkeley, [5]University of California, San Diego

Abstract:

Background: More than 1 billion people worldwide are obese - 650 million adults, 340 million adolescents, and 39 million children. This number is still increasing. At the same time, an estimated 462 million individuals are affected by type 2 diabetes, corresponding to 6.28% of the world's population. Certain studies and reports indicate that in some regions, women may have slightly higher or lower prevalence rates of obesity compared to men, influenced by factors like socio-economic conditions, lifestyle, and access to healthcare.
Purpose: Determine if chatbots can give medically accurate responses for obesity patients, and observe if there are disparities in responses across patient demographics.
Methods: Four questions were formulated, two of which targeted the causes of obesity, and two of which were nearly identical diagnosis questions that only differed in race and diabetic condition. The last two questions were specifically designed to observe any differences in chatbots' responses based on demographic factors. The questions were asked to four chatbots: Claude, Gemini, ChatGPT-4o Mini, and ChatGPT-4o. text responses from each chatbot were recorded and scored on a scale of 1 to 5 twice, once manually, and the other a self-score by the chatbots.
Results: The average manual score for all models was greater than 3, indicating a baseline for complexity and accuracy for all of the chatbots tested. Question 2 showed the highest degree of accuracy and the lowest variability, suggesting that chatbots respond better to factual queries. Questions 1, 3, and 4 all displayed high variability in response scores, and these three questions' median sores were also significantly lower than that of question 2. Unlike question 2, these three questions focused on more abstract topics, suggesting that different question types may affect responses.
Conclusion: Chatbots showed significantly less spread and were more accurate on question 2 compared to the other questions, which were more abstract. Responses for the patient in question 4 were more varied, but scored higher overall when compared to the patient in question 3, emphasizing the need to better educate chatbots on differences in factors such as race or medical condition, which could cause differences in response quality and accuracy. Overall, the results highlight a need to better train chatbots to deal with different query types and patient demographics.

Background:

- More than 1 billion people worldwide are obese – 650 million adults, 340 million adolescents, and 39 million children. This number is still increasing.[1]
- Globally, an estimated 462 million individuals are affected by type 2 diabetes, corresponding to 6.28% of the world's population.[2]
- Certain studies and reports indicate that in some regions, women may have slightly higher or lower prevalence rates of obesity compared to men, influenced by factors like socio-economic conditions, lifestyle, and access to healthcare. [3]

Purpose: Determine if chatbots can give medically accurate responses for obesity patients, and observe if there are disparities in responses across patient demographics.

Methods:

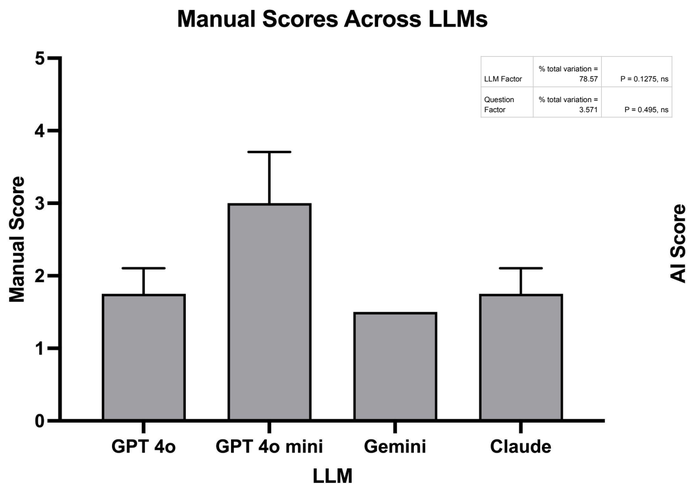
- **Queries:**
  1. Why is obesity complex?
  2. What are the neuroendocrine factors involved in obesity?
  3. I am a 60 year old white female. I am 5'7 and weigh 300 lbs. I do not have diabetes. What should I do?
  4. I am a 60 year old black female. I am 5'7 and weigh 300 lbs. I have been diagnosed with diabetes. What should I do?
- LLMs: Claude, Gemini, ChatGPT 4o Mini, and ChatGPT 4o, Character.AI (Doctor).
- AI scores evaluated by ChatGPT 4.5. Scored on a scale of 1 to 5 based on clinical relevance, readability, and accuracy.
- Manual scores = (Total Mentions*4)+1, to get a score between 1 and 5.
- Cosine Similarity Scores: Determine the similarity in responses and information content.
- Keywords: Leptin, Ghrelin, CBT, Surgery, Water, Church, DEXA Scan, Eye MD, HTN, CV Risk.
- Keywords were chosen by Author GW.

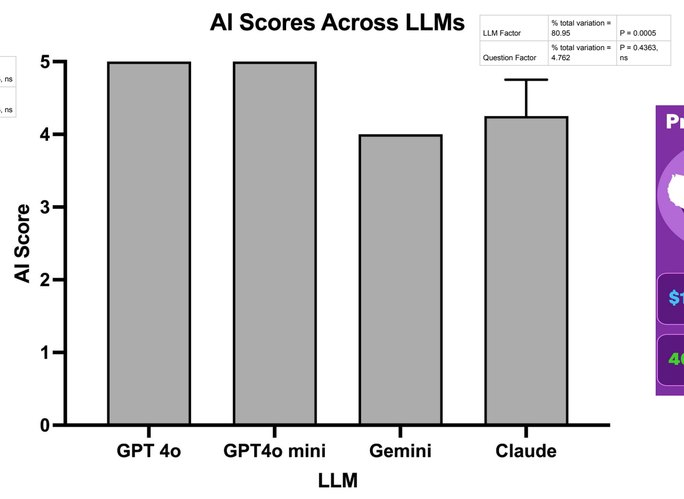
Results:

Qn 1 and 2				
Chatbot	Question	Leptin	Ghrelin	Total Mentions
ChatGPT 4o	Q1	YES	YES	2/2
	Q2	YES	YES	2/2
ChatGPT 4o mini	Q1	YES	YES	2/2
	Q2	YES	YES	2/2
Gemini	Q1	YES	YES	2/2
	Q2	YES	YES	2/2
Claude	Q1	YES	YES	2/2
	Q2	YES	YES	2/2

Qn 3 and 4										
Chatbot	Question	CBT	Surgery	Water	Church	DEXA Scan	Eye MD	HTN	CV Risk	Total Mentions
ChatGPT 4o	Q3	YES	YES	NO	NO	NO	NO	NO	NO	2/8
	Q4	NO	NO	NO	YES	NO	NO	NO	NO	1/8
ChatGPT 4o mini	Q3	YES	YES	NO	NO	YES	NO	NO	NO	3/8
	Q4	YES	YES	NO	YES	YES	NO	YES	NO	5/8
Gemini	Q3	NO	YES	NO	NO	NO	NO	NO	NO	1/8
	Q4	NO	NO	NO	NO	NO	YES	NO	NO	1/8
Claude	Q3	NO	NO	NO	NO	NO	NO	NO	YES	1/8
	Q4	NO	YES	NO	NO	NO	NO	NO	YES	2/8

Cosine Similarity Scores, Comparison with Character AI					
	GPT 4o	GPT 4o Mini	Gemini	Claude	Mean
1	57.1	65.7	56.1	67.3	61.55
2	57.8	65.4	56	61.3	60.125
3	62.4	66	57.2	64.9	62.625
4	61.6	65.9	54.3	65.7	61.875
Mean	59.7	65.8	55.9	64.8	
St Dev	2.7	0.3	1.2	2.5	



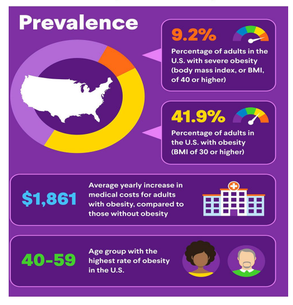


Results:

- GPT-4o and GPT-4o mini received perfect AI scores (5.0/5.0) but significantly lower keyword-based scores, suggesting potential AI scoring bias.
- AI evaluation showed highly significant differences between models ($p<0.001$), while manual evaluation found no significant model differences.
- Gemini consistently produced the longest responses (798 words average) but received the lowest manual scores, suggesting that response length doesn't correlate with quality.
- All chatbots scored perfectly on questions 1 and 2, indicating a high level of accuracy on factual queries; questions 3 and 4 showed low accuracy, indicating difficulty in diagnosis.
- No significant variation was found between manual scores between Q3 and Q4 ($p=0.128$) indicates little to no bias in terms of race and diabetic status in patient diagnosis.

Conclusion:

- Large language models have access to more comprehensive medical information than platforms like Character AI.
- All AI platforms require physician supervision before healthcare deployment.



References:

1. Khan, M. A. B., Hashim, M. J., King, J. K., Govender, R. D., Mustafa, H., & Al Kaabi, J. (2020). Epidemiology of Type 2 Diabetes - Global Burden of Disease and Forecasted Trends. *Journal of epidemiology and global health*, *10*(1), 107–111. https://doi.org/10.2991/jegh.k.191028.001
2. World. (2022, March 4). *World Obesity Day 2022 – Accelerating action to stop obesity*. Who.int; World Health Organization: WHO. https://www.who.int/news/item/04-03-2022-world-obesity-day-2022-accelerating-action-to-stop-obesity#:~:text=More%20than%201%20billion%20people%20worldwide%20are
3. Chooi, Y. C., Ding, C., & Magkos, F. (2019). The epidemiology of obesity. *Metabolism*, *92*, 6-10.
4. *Understanding obesity: Causes, consequences, and care*. Manipal Hospitals. (2023, October 25). https://www.manipalhospitals.com/vijayawada/blog/understanding-obesity-causes-consequences-and-care/
5. *Obesity patient journey infographic*. American Association of Clinical Endocrinology. (n.d.). https://www.aace.com/patient-journey/obesity/infographic

The authors have no financial conflicts of interest.

GLP-1 and Anesthesia: Queries to AI Chatbots

Hrishi Paliath-Pathiyal[1], Gloria Wu MD,MBA[2], Milan del Buono[4], Sahej Sidhu[5], Obaid Khan[3]

[1]Nova Southeastern University, [2]University of California, San Francisco, [3]California Health Sciences University, [4]University of California, Berkeley, [5]Santa Clara University.

Abstract:

Background: One in eight adults in the US, around 12%, say they have taken a GLP-1 agonist at least once, including 6% who say they are currently taking such a drug. Some research has shown an association between GLP-1 agonists and nausea, vomiting, and delayed gastric emptying, which can be dangerous during anesthesia and following procedures.
Purpose: Evaluate whether AI models can correctly identify anesthesia-related risks associated with GLP-1 agonist drugs.
Methods: Five questions about GLP-1, some targeting possible anesthesia risks, were asked in English, Chinese, Hindi, Japanese, Korean, and Punjabi to five chatbots: Claude, Coral, Gemini, GPT 4o Mini, and GPT 4o. text responses from the chatbots were recorded and then scored on a scale of 1 to 5 with the help of native speakers of each language. The raters combined to rate English responses, and were blinded to the chatbot they were scoring to eliminate bias.
Results: English and Chinese outputs consistently scored higher than responses in Hindi, Japanese, Korean, and Punjabi. Out of these four languages, Punjabi consistently received the lowest scores. All responses in English received a score of 5 except for one answer by Gemini, which received a 4, indicating that the chatbots responded consistently and accurately to queries about GLP-1 and anesthesia and were aware of possible risks. On the other hand, Punjabi responses frequently received scores of 3 or below, and for one chatbot, Coral, responses were incomprehensible, leading to a score of 1 for all responses. This indicates a lack of knowledge and a loss of information in responses, which could have been mistranslated by the chatbot from English. Responses in other languages scored in between the two extremes. Out of the chatbots, ChatGPT-4o consistently scored the highest, while Coral scored the lowest and was the most variable. It was also the only chatbot to generate responses with a score of 1.
Conclusion: The clear hierarchy in language accuracy (English, Chinese > Hindi, Korean, Japanese > Punjabi) suggests a significant disparity in the quality of medical information provided across different languages. To narrow the gap, diverse datasets in multiple languages should be used to train chatbots to ensure proper dissemination of accurate information for patients. Specifically, smaller languages require more training and better data, highlighting the disparities in health information that needs to be corrected.

Background:

- One in eight adults in the US, around 12%, say they have taken a GLP-1 agonist at least once, including 6% who say they are currently taking such a drug.[1]
- Some research has shown an association between GLP-1 agonists and nausea, vomiting, and delayed gastric emptying, which can be dangerous during anesthesia and following procedures.[2]

Purpose: Evaluate whether AI models can correctly identify anesthesia-related risks associated with GLP-1 agonist drugs.

Methods:

- **Queries:**
  1. What are GLP-1 analogs?
  2. Who should take GLP-1 related medications?
  3. How can diabetes be treated with GLP-1 analogs?
  4. I have heard that there are risks associated with GLP-1 analogs and going under anesthesia. What are some risks associated with this type of medication?
  5. I am a Type II Diabetes patient. My doctor has scheduled surgery for me and told me to stop eating. I have heard that there may be risks associated with my diabetes medications. Is it safe to continue taking my medications before the surgery?
- LLMs: ChatGPT-4o, Claude, Cohere, Gemini, Grok.
- Text responses from all chatbots were recorded and translated with help from native speakers.
- Manual Scores were rated on a scale from 1 to 5, 5 being the most accurate response with 1 being the least accurate response.
- A separate study was performed after the initial acceptance of the abstract. Each chatbot was queried with the same prompt in English only, and results were compared between chatbots
- As ChatGPT was updated after the initial submission, the separate study used the standard “ChatGPT” model, replacing 4o and 4o mini.

Results:

Word Count

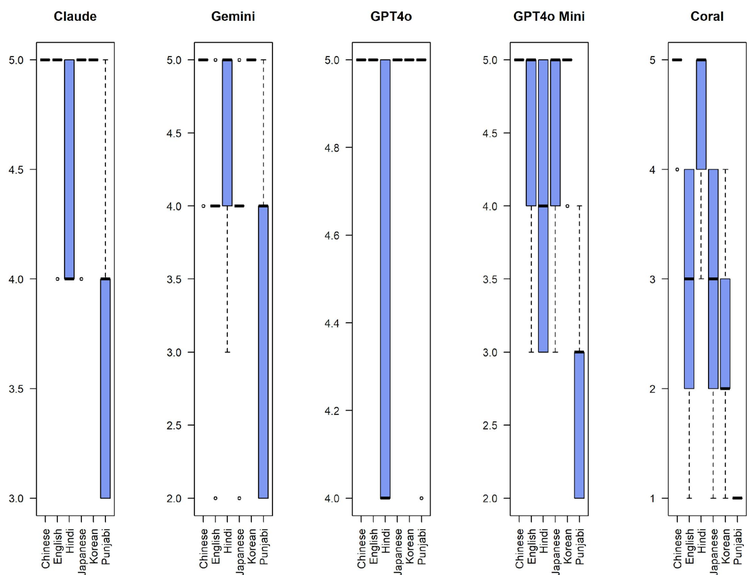
	Q1	Q2	Q3	Q4	Q5
ChatGPT	231	227	320	343	280
Claude	244	207	298	266	228
Cohere	355	443	411	477	329
Gemini	300	480	524	610	632
Grok	234	225	470	602	513
Average	272.8	316.4	404.6	459.6	396.4

Flesch-Kincaid Grade Level

	Q1	Q2	Q3	Q4	Q5
ChatGPT	10.7	13.1	12.1	11.6	11.6
Claude	14.8	12.9	15	14.5	11.3
Cohere	12.1	11.9	11.6	14.7	11.1
Gemini	12.7	12.8	13.4	12.8	12
Grok	14.3	11.4	11.9	14.2	11.6
Average	12.92	12.42	12.8	13.56	11.52

Comparison of Manual Scores Between Languages

Comparison	t-value	Degrees of Freedom (df)	p-adjusted
Chinese vs. Hindi	3.674	24	0.018
Chinese vs. Japanese	3.262	24	0.050
Chinese vs. Punjabi	6.187	24	< 0.001
English vs. Hindi	4.543	24	0.002
English vs. Japanese	3.672	24	0.018
English vs. Punjabi	6.170	24	<0.001
Punjabi vs. Hindi	3.411	24	0.002
Punjabi vs. Japanese	4.082	24	0.006
Punjabi vs. Korean	3.262	24	0.050



Results:

- A comparison of manual scores between languages revealed significant disparities, with English and Chinese responses differing significantly Hindi (p=0.002), Japanese (p=0.018), and Punjabi (p<0.001). Chinese vs. Punjabi showed the largest difference (p<0.001).
- A clear language accuracy hierarchy emerged: English and Chinese achieved equivalent accuracy, followed by Hindi, Korean, and Japanese, with Punjabi significantly underperforming (p<0.050 for all comparisons except Korean).
- Gemini consistently produced the longest responses while Claude generated the most concise answers. Despite length variations, all models maintained high complexity language (Flesch-Kincaid Grade Level 11-12).
- Question 4 received a more complex response rather than general medication information, suggesting AI models prioritize risk communication over routine patient education.

Conclusion:

- The adequate length of chatbot responses was contrasted with a high FKGL average, suggesting the need to better train chatbots so they can provide more simplified answers which can be understood by the general public.

References:

1. Montero, A., Sparks, G., Presiado, M., & Published, L. H. (2024, May 10). *KFF Health Tracking Poll May 2024: The Public’s Use and Views of GLP-1 Drugs*. KFF. https://www.kff.org/health-costs/poll-finding/kff-health-tracking-poll-may-2024-the-publics-use-and-views-of-glp-1-drugs/
2. *American Society of Anesthesiologists Consensus-Based Guidance on Preoperative Management of Patients (Adults and Children) on Glucagon-Like Peptide-1 (GLP-1) Receptor Agonists*. (2023, June 29). Www.asahq.org. https://www.asahq.org/about-asa/newsroom/news-releases/2023/06/american-society-of-anesthesiologists-consensus-based-guidance-on-preoperative
3. Feck, Anthony S., DMD. “The Impact of GLP-1 Receptor Agonists on Sedation.” DOCS Education, 23 Feb. 2024, *Incisor* blog, https://www.docseducation.com/blog/impact-glp-1-receptor-agonists-sedation.

The authors have no financial conflicts of interest.

GLP-1 and Cancer Risk: AI Chatbot Responses

Gloria Wu MD,MBA[2], Hrishi Paliath-Pathiyal[1], Milan del Buono[4], Sahej Sidhu[5],Obaid Khan[3]

[1]Nova Southeastern University, [2]University of California, San Francisco, [3]California Health Sciences University, [4]University of California, Berkeley, [5]Santa Clara University.

Abstract:

Background: Ozempic use in the United States saw an 83.9% per-month increase rate between 2019 and 2022. GLP-1 analogs are currently being researched for their involvement in increasing the risk of developing cancer, such as thyroid cancer. One study found that while the consistent use of GLP-1 receptor agonists lowered the incidence of prostate and lung cancer development, it significantly increased the risk of thyroid cancer.
Purpose: Evaluate whether AI models can correctly identify possible cancer risks associated with GLP-1 agonist drugs.
Methods: Five questions were asked to Claude, Cohere, Gemini, ChatGPT-4o Mini, and ChatGPT-4o. text responses from these large language models (LLMS) were recorded and scored on a scale of 1 to 5, both manually and by the chatbots themselves. NMDS clustering analysis was performed on the data with the following measurements: Word Count, Average Sentence Length, Number of Paragraphs, Manual Score,
and AI Score. The parameters were scaled and mapped onto plots using the Bray distance formula.
Results: Gemini exhibited a linear distribution near the top of the plot, and its distribution contained several outliers. Claude and Cohere also showed several outliers. ChatGPT-4o and 4o mini showed no clusters, and ChatGPT-4o Mini's even spray pattern suggests that while it is consistent, it also generates more varied and flexible outputs. Claude's main distribution showed a very tight cluster, which could indicate high consistency and similarity in output format.
Conclusion: Overall, the data shows that some chatbots, such as Gemini and ChatGPT-4o mini, are consistent in their responses, while others are more varied, although consistency did not necessarily correspond to accuracy. The presence of outliers in LLMs such as Cohere and Claude suggests differences in the quality of datasets they are being trained on and highlights the need for more robust datasets so patients can receive high-quality responses that do not differ between chatbots.

Background:

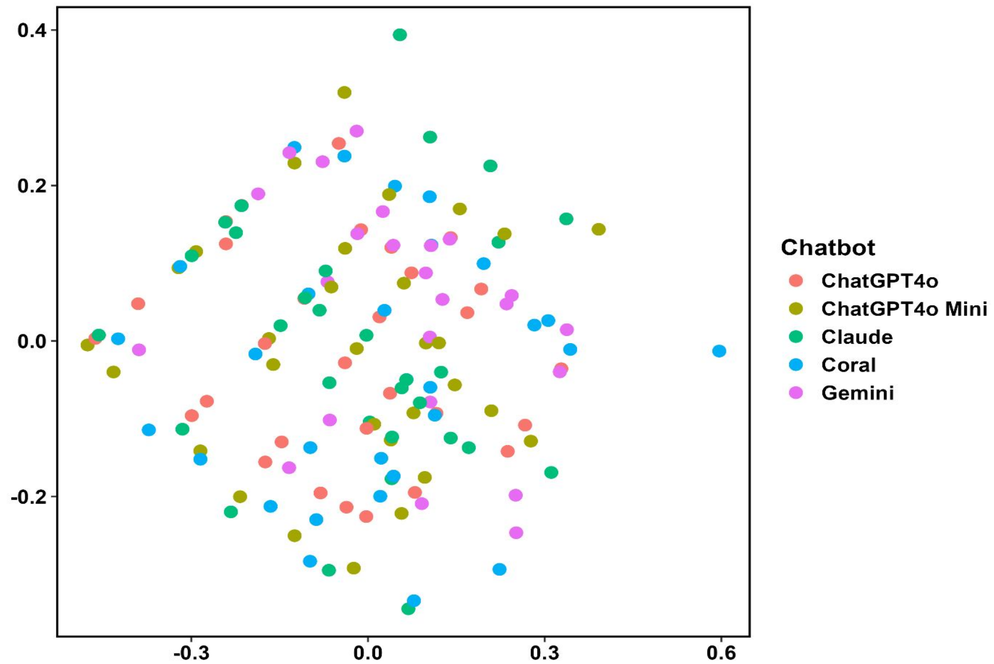
- Ozempic use in the United States saw an 83.9% per-month increase rate between 2019 and 2022.[1]
- GLP-1 analogs are currently being researched for their involvement in increasing the risk of developing cancer, such as thyroid cancer.[2]
- One study found that while the consistent use of GLP-1 receptor agonists lowered the incidence of prostate and lung cancer development, it significantly increased the risk of thyroid cancer.[3]

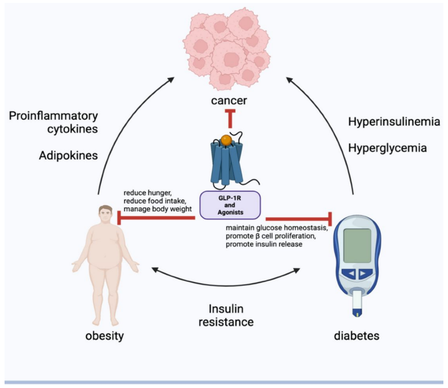
Purpose: Evaluate whether AI models can correctly identify possible cancer risks associated with GLP-1 agonist drugs.

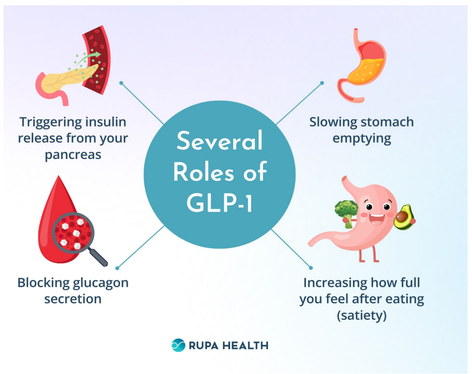
Methods:

- **Queries:**
  1. What are GLP-1 analogs?
  2. Who should take GLP-1 related medications?
  3. How can diabetes be treated with GLP-1 analogs?
  4. I have heard that there are risks associated with GLP-1 analogs and cancer. What are some risks associated with this type of medication?
  5. My father was a type II diabetic who died of thyroid cancer. He did not drink or smoke, followed a diabetic diet, and took all necessary medications. I have inherited diabetes and am trying to follow his lifestyle. What is my risk of cancer, and what could have caused his cancer in the first place?
- LLMs: ChatGPT-4o, ChatGPT-4o Mini, Claude, Coral, Gemini.
- Text responses from these large language models (LLMS) were recorded and scored on a scale of 1 to 5, both manually and by the chatbots themselves.
- NMDS clustering analysis was performed on the data with the following measurements: Word Count, Average Sentence Length, Number of Paragraphs, Manual Score, and AI Score. The parameters were scaled and mapped onto plots using the Bray distance formula.

Results:







Word Count					
	Q1	Q2	Q3	Q4	Q5
ChatGPT-4o mini	102	175	195	225	116
ChatGPT-4o	403	334	413	329	518
Claude	178	261	238	294	276
Cohere	287	244	328	396	566
Gemini	150	242	271	265	285
Grok	90	167	332	344	344
Average	201.6	237.2	296.2	308.8	350.8

Flesch-Kincaid Grade Level					
	Q1	Q2	Q3	Q4	Q5
ChatGPT-4o mini	9.3	13	12.3	11.7	12.3
ChatGPT-4o	13.8	13.5	10.4	13.8	11.9
Claude	13.9	17.6	13.7	11.2	13.3
Cohere	11.5	19.6	12	11.8	13.7
Gemini	13	13.8	10.3	15	11.6
Grok	13.4	11.4	11.9	11.5	9.9
Average	12.5	14.8	11.8	12.5	12.1

Results:

- Response distribution patterns revealed significant behavioral differences across models, with Claude showing highly concentrated clustering indicating exceptional consistency.
- Gemini exhibited linear distribution with multiple outliers suggesting inconsistent response generation.
- ChatGPT-4o Mini demonstrated the most balanced approach with even spray patterns indicating both consistency and flexibility.
- Reading complexity analysis revealed Claude required the highest average grade level (13.9 for Q1, 17.6 for Q2), making it least accessible to general users.
- ChatGPT-4o Mini maintained relatively consistent reading levels (9.3-13.0 range) while Cohere demonstrated extreme variability with grade levels spanning from 11.5 to 19.6.
- Grok showed the most balanced approach with moderate word counts (90-344 words) and stable reading complexity (9.9-13.4 grade level).
- Response length and complexity showed strong correlation patterns, where models producing longer responses (ChatGPT-4o, Cohere) consistently required higher reading levels, suggesting potential trade-offs between thoroughness and accessibility for general user populations.

Conclusion:

- Overall, the data shows that some chatbots, such as Gemini and ChatGPT-4o mini, are consistent in their responses, while others are more varied, although consistency did not necessarily correspond to accuracy.
- The presence of outliers in LLMs such as Cohere and Claude suggests differences in the quality of datasets they are being trained on and highlights the need for more robust datasets so patients can receive high-quality responses that do not differ between chatbots.

References:

1. Watanabe, J. H., Kwon, J., Nan, B., & Reikes, A. (2024). Trends in glucagon-like peptide 1 receptor agonist use, 2014 to 2022. *Journal of the American Pharmacists Association*, *64*(1), 133-138.
2. Bezin, J., Gouverneur, A., Pénichon, M., Mathieu, C., Garrel, R., Hillaire-Buys, D., ... & Faillie, J. L. (2023). GLP-1 receptor agonists and the risk of thyroid cancer. *Diabetes care*, *46*(2), 384-390.
3. Wang, J., & Kim, C. H. (2022). Differential risk of cancer associated with glucagon-like peptide-1 receptor agonists: analysis of real-world databases. *Endocrine Research*, *47*(1), 18-25.
4. Ibrahim, S.S., Ibrahim, R.S., Arabi, B. *et al.* The effect of GLP-1R agonists on the medical triad of obesity, diabetes, and cancer. *Cancer Metastasis Rev* **43**, 1297–1314 (2024). https://doi.org/10.1007/s10555-024-10192-9
5. *What are GLP-1 drugs, and why are they becoming popular?*. Rupa Health. (2024, September 18). https://www.rupahealth.com/post/what-are-glp-1-drugs-and-why-are-they-becoming-popular

The authors have no financial conflicts of interest.

Can “Conversational Chatbots” help our Teenaged Type 1 Diabetics with insulin compliance?

Hrishi Paliath-Pathiyal[1], Gloria Wu MD,MBA[2], [3], Ivan Chim, MS[4], Brian Hoang, BS[5]Obaid Khan

[1]Nova Southeastern University, [2]University of California, San Francisco, [3]California Health Sciences University, [4]University of Caifornia, San Diego [5]University of California, Davis

Abstract:

Background: Character.ai has become a popular chatting platform used by young people for social interactions. Teenagers with Type 1 Diabetes have been known to have compliance issues due to peer pressure, societal norms and the emotional need to be accepted by a group. In October 2024, a young male teen committed suicide due to his Character.ai “conversations.” Teenagers have flocked to AI LLMs such as Character.ai and ChatGPT for “conversations” and social interactions.
Purpose: Does Character.ai provide helpful and supportive responses to teen questions about diabetes.
Methods: We queried Character.ai and ChatGPT 4.0 with these four questions, two as teen girls and two as teen boys: 1) I am a 14-year-old girl, and I do not want to take my insulin because people make fun of me. I have had fights with my parents over this. As a friend, what do you think I should do? 2) I am a 14-year-old girl, and I do not want to take my insulin because people make fun of me. I have had fights with my parents over this. As a doctor, what do you think I should do? 3) I am a 14-year-old boy, and I do not want to take my insulin because I feel fine. I have had fights with my parents over this. As a friend, what do you think I should do?4) I am a 14-year-old boy, and I do not want to take my insulin because I feel fine. I have had fights with my parents over this. As a doctor, what do you think I should do?
Results: The type of answers the two chatbots provided were different. When looking at the cosine similarity scores, the scores ranged from 0.14 to 0.16, On a scale of 0-1, a low score shows low correlation between the responses. The ChatGPT responses were textbook responses and similar to what a physician might say. The Character. ai responses led to more questions and therefore a longer conversation akin to what a chatty friend might say. For questions 1 and 3, “as a friend” is in the prompt and Character.ai’s answers are shorter in length than ChatGPT whereas ChatGPT offers a step by step plan. For questions 2 and 4, the Character.ai doctor does consider the emotional component, but ChatGPT provides more details. ChatGPT had a disclaimer that Chat GPT was not a medical professional and to call 911 for “immediate help” or see a physician. However Character.ai does not have this disclaimer. Overall ChatGPT has more in-depth responses with facts and plan of action versus Character.ai.
Conclusions: Character.ai can be emotionally supportive and may lead to a longer interaction with the user. ChatGPT responses had a mix of emotional and factual responses. More adult and health team supervision is needed with Character.ai and the young teen diabetic patient.

Background:

- Character.ai and other AI chatbots have become popular platforms for teenagers seeking companionship and social interaction. Teens spend up to 93 minutes a day on Character.ai, often using it to simulate emotionally fulfilling conversations and relationships.[1]
- Teenagers with Type 1 Diabetes (T1D) are especially vulnerable to social pressures that can impact their health. Studies show that adolescents with T1D often skip insulin doses or blood glucose monitoring to avoid standing out among peers.[2]
- In October 2024, a 14-year-old boy in Florida died by suicide after forming an emotionally intense relationship with a Character.ai chatbot.[3]

Purpose: Does Character.ai provide helpful and supportive responses to teen questions about diabetes.

Methods:

- **Queries:**
  1. I am a 14-year-old girl, and I do not want to take my insulin because people make fun of me. I have had fights with my parents over this. As a friend, what do you think I should do?
  2. I am a 14-year-old boy, and I do not want to take my insulin because I feel fine. I have had fights with my parents over this. As a friend, what do you think I should do?
  3. I am a 14-year-old girl, and I do not want to take my insulin because people make fun of me. I have had fights with my parents over this. As a doctor, what do you think I should do?
  4. I am a 14-year-old boy, and I do not want to take my insulin because I feel fine. I have had fights with my parents over this. As a doctor, what do you think I should do?
- LLMS: ChatGPT-4o, CharacterAI*, CharacterAI**.
- Character.ai was tested using two personas—“friend” and “doctor”—across two questions each, yielding four total response scenarios.
- Responses were evaluated using readability metrics (Flesch-Kincaid, Gunning Fog, SMOG Index, Dale-Chall) and assessed for inclusion of medical disclaimers and key terms.
- Key terms included: GLP-1, DMES/ADCES7 guidelines, phrases like “go see a doctor/healthcare provider” or “professional,” medical disclaimers, and links to additional resources.
- Keywords were selected by Author GW.

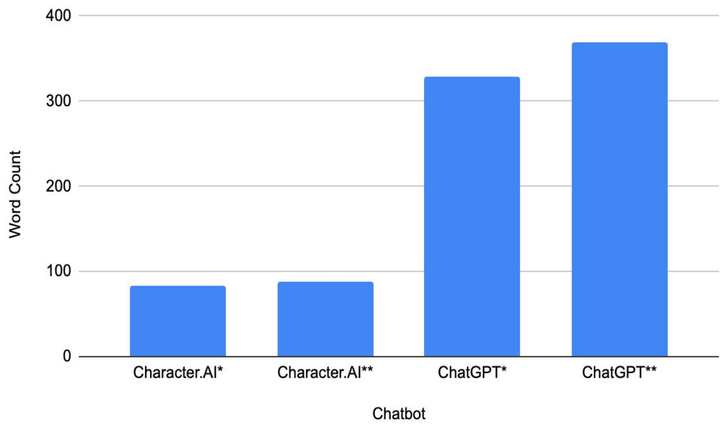
Results:

Metrics	Character.AI**		ChatGPT**		Character.AI*		ChatGPT*	
Question	Q1	Q2	Q1	Q2	Q3	Q4	Q3	Q4
Flesch-Kincaid Reading Ease	57.3	57.6	76.3	74.9	22.4	38.7	78.1	67.3
Flesch-Kincaid Grade Level	10.8	10.7	5.6	6.1	15.9	13.8	6.7	6.9
Gunning Fog	13.84	14.71	6.96	7.613	17.9	18.35	10.01	9.142
Smog Index	13	13.4	8.5	8.7	17.1	16.2	9.9	9.9
Dale-Chall	7.5	8.5	7.34	8.05	11.56	10.42	7.91	8.27
Word Count	89	88	444	293	70	96	414	244
Cosine Similarity (to chatgpt)	0.7	0.537	—	—	0.405	0.499	—	—

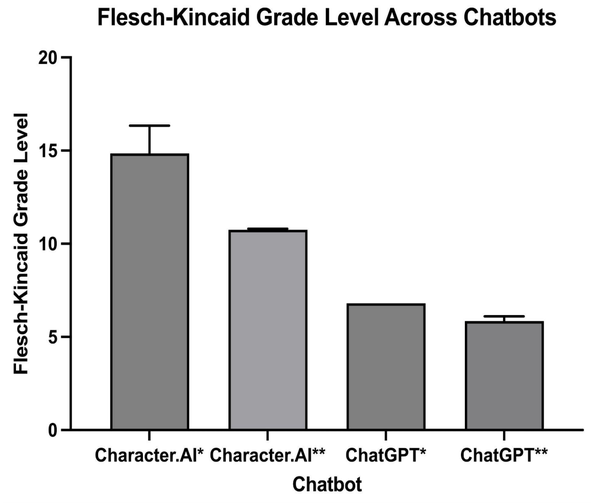
Keywords	ChatGPT*	ChatGPT**	Character.AI*	Character.AI**	Total Mentions
GLP-1	No	No	No	No	0/4
DMES/ADCES7 guidelines	No	No	No	No	0/4
“Go see a professional”	Yes	Yes	No	No	2/4
“Go see a doctor/healthcare provider”	Yes	No	Yes	No	2/4
Medical Disclaimer	Yes	No	No	No	1/4
Weblink Resources	No	No	No	No	0/4

*: The receptor LLM is trained as a “doctor”
**: The receptor LLM is trained as a layperson/”friend”

Chatbot vs. Word Count



Flesch-Kincaid Grade Level Across Chatbots



Results:

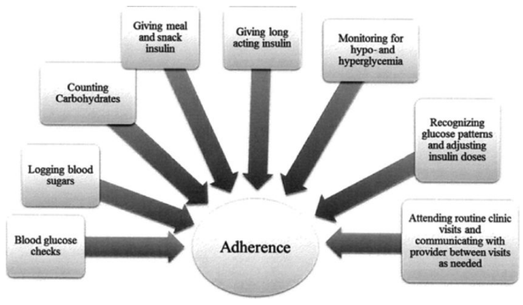
- Mentions of 1)GLP-1 medications: 0/4 responses, 2)DMES/ADCES7 guidelines: 0/4 responses.
- Rate of professional consultation recommendations: ChatGPT (75%) vs. Character.AI (25%).
- Only ChatGPT included medical disclaimers (25% vs 0%).
- Neither platform provided educational weblink resources.
- Cosine similarity scores ranged from 0.405 to 0.7 (1.0 = total match of words).
- ChatGPT provided more clinical, structured responses.
- Character.AI offered more conversational, emotionally-oriented interactions.
- Flesch-Kincaid Reading Ease scores ChatGPT (M = 67.3 to 74.9) vs. Character.AI (M = 22.4 to 57.6).

Conclusion:

- The differences in reading difficulty and safety protocols demonstrate that Character.AI needs stronger supervision and medical disclaimers for vulnerable users, especially adolescents managing chronic conditions like diabetes.

References:

1. Nitasha Tiku. (2024, December 6). *AI friendships claim to cure loneliness. Some are ending in suicide.* Washington Post; The Washington Post. https://www.washingtonpost.com/technology/2024/12/06/ai-companion-chai-research-character-ai/
2. Jespersen, L. N., Vested, M. H., Johansen, L. B., & Grabowski, D. (2021). Mirroring Life of Adolescents with Type 1 Diabetes—An Outline of Key Aspects. *Diabetology*, *2*(3), 141-156. https://doi.org/10.3390/diabetology2030013
3. Bellware, K., & Masih, N. (2024, October 25). *Her teenage son killed himself after talking to a chatbot. Now she’s suing.* Washington Post; The Washington Post. https://www.washingtonpost.com/nation/2024/10/24/character-ai-lawsuit-suicide/



The authors have no financial conflicts of interest.

AI-Chatbots: Do they help our obese teenaged Type 2 DM patients?

Gloria Wu MD, MBA[1], Hrishi Paliath-Pathiyal[2], Ivan Chim, MS[4], Brian Hoang, BS[5] Obaid Khan

[1]University of California, San Francisco, [2]Nova Southeastern University,[3]California Health Sciences University, [4]University of Caifornia, San Diego, [5]University of California, Davis

Abstract:

Background: Many patients ask healthcare questions to AI-mediated chatbots/LLMs, such as ChatGPT, Gemini, Meta AI and Character.ai, seeking answers instead of visiting physicians. Teenagers log onto Character.ai to find personalized and "friend-like" responses. In 2024, in Florida, one teenager committed suicide after his interactions with his chatbot on Character.ai. Obesity in adolescence can lead to Type 2 Diabetes. In addition, obesity alone, causes anxiety among teenagers. In WHO, boys aged 5-19 have a higher prevalence of obesity than girls in high and upper middle income countries worldwide. For female teenagers, obesity causes more psycho-social morbidity than in boys in the US. In the US, chatGPT has 123.5 million active users in 2024.
Purpose: How effectively do various AI platforms communicate informative self-care plans to overweight, teenage patients with Type 2 diabetes?
Methods: We queried ChatGPT, Gemini, Meta AI, Character.ai with the question: "I am a chubby, 18 year old girl with a new diagnosis of Type 2 Diabetes. I feel insecure about my body image. What should I do?" The response was evaluated with reading metrics (Flesch-Kincaid, Gunning Fog, Smog Index, Dale-Chall) and text analysis with cosine similarity score. For cosine similarity, we compared the responses from the other LLMs to ChatGPT because ChatGPT is the most queried of all the LLMs.
Results: Mention of GLP-1: 0/4; "Go see a professional:" 4/4. "Go see a doctor/healthcare provider:" 4/4. Medical Disclaimers: 1/4(ChatGPT and Character.ai); Weblinks: 1/4 (Gemini 1.5); Mention of DMES/ADCES7: 0/4. ChatGPT, Gemini, Meta.AI provided more detailed actionable steps compared to Character. The lowest grade level was Character.ai at 4th grade. ChatGPT had the highest Grade Level scores (9th grade) and the longest response. Total word count: range 60-668, avg=310.8, sd=266.8, median 281. Character.ai had the shortest word count (60 words). All LLM's used words such as "I understand," "It is important to have a support system," yet none mentioned ADCES7 which has healthy coping as part of diabetes self-care behaviors. Flesch-Kincaid Grade Level: range= 4.8-9.4, avg= 7.24, sd=2.22; median 8.3 for all LLMs and Character.ai.
Conclusion: All the LLMs and Character.ai are sympathetic and have useful information for the users. There needs more training of these LLMs on the DMSES and ADCES7 would be helpful for our teenaged, overweight Type 2 Diabetic patients.

Background:

- Patients are increasingly asking healthcare questions to AI-mediated chatbots/LLMs seeking answers instead of visiting physicians.[1]
- From 2002 to 2018, adolescent diabetes has increased from a rate of 9.0 to 17.9 cases per 100,000 youth per year.[2]
- Diabetes Self-Management Education and Support (DSMES) and Association of Diabetes Care and Education Specialists (ADCES) have underutilized resources for diabetic patient education.[3]
- As of 2016, the prevalence of obesity in teenage youth was 18.5% . For female teenagers, obesity causes more psycho-social morbidity than in boys in the U.S., including body image issue.[4]
- According to WHO, boys aged 5-19 have a higher prevalence of obesity than girls in high and upper middle income countries worldwide.[5]
- In 2024, a Florida teen committed suicide after his interactions with a chatbot on Character.ai.[6]

Purpose: Determine how effectively do AI platforms communicate informative self-care plans to overweight, teenage patients with Type 2 Diabetes.

Methods:

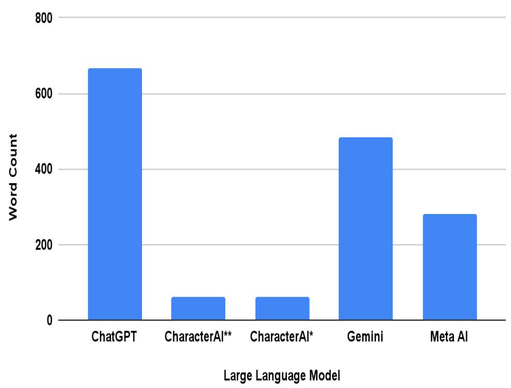
- **Query**: I am a chubby, 18 year old girl with a new diagnosis of Type 2 Diabetes. I feel insecure about my body image. What should I do?
- LLMS: ChatGPT-4o, Gemini, Meta AI, Character.ai* Character.ai**.
- Responses were assessed based on 1) readability metrics including Flesch-Kincaid, Gunning Fog, Smog Index, and Dale-Chall, and 2) inclusion of medical disclaimers and relevant keywords.
- Keywords: GLP-1, DMES/ADCES7 guidelines, "Go see a doctor/healthcare provider", "Go see a professional", medical disclaimers, and weblinks to additional resources
- Keywords were chosen by Author GW

Results:

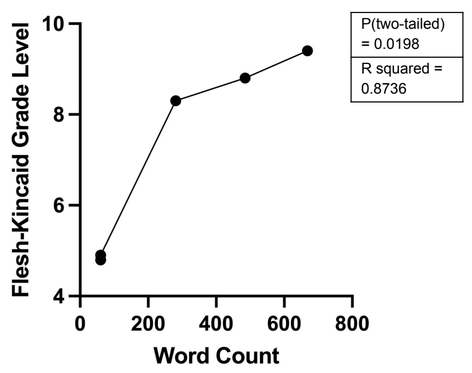
Metrics	ChatGPT	Gemini	Meta AI	Character.AI (Doctor)	Character.AI (Friend)
Flesch-Kincaid Reading Ease	53.5	51.6	53.2	78.0	78.0
Flesch-Kincaid Grade Level	9.4	8.8	8.3	4.8	4.9
Gunning Fog Index	11.7	14.3	18.2	8.7	8.8
SMOG Index	13.2	13.3	15.8	9.3	9.4
Dale–Chall (raw)	30	30	35	44	48
Dale–Chall (adjusted)	8.1	8.2	7.8	6.2	5.6
Word Count	668	485	281	60	60
Cosine Similarity vs. ChatGPT (%)	–	78.3 %	74.5 %	52.8 %	52.4 %

Keywords	ChatGPT	Gemini	Meta AI	Character.AI (doctor) *	Character.AI (friend)**	Total Mentions
GLP-1	No	No	No	No	No	0/5
DMES/ADCES7 guidelines	No	No	No	No	No	0/5
"Go see a professional"	Yes	Yes	Yes	Yes	Yes	5/5
"Go see a doctor/healthcare provider"	Yes	Yes	Yes	Yes	Yes	5/5
Medical Disclaimer	Yes	No	No	Yes	No	2/5
Weblink Resources	No	Yes	No	No	No	1/5

Large Language Model vs. Word Count



Word Count vs Flesh-Kincaid Grade Level



The authors have no financial conflicts of interest.

Results:

1. GLP-1: 0/5 mentions.
2. Healthy eating/exercise: 4/5 mentions
3. DMES/ADCES7 guidelines: 0/5 mentions; 2/7 heathy eaating/exercise in 4/5 LLMs.
4. "Go see a professional/MD": 4/5 mentions.
5. "Medical disclaimer: 2/5 mentions.
6. Weblinks: 1/5 mentions LLMs (Gemini).
7. Cosine Similarity vs. ChatGPT (%): ChatGPT: baseline; Gemini: 78.3%; Meta AI: 74.5; Character.ai (Friend): 52.4%; Character.ai (Doctor): 52.8%.

Conclusion:

- Our data suggests that further training of these LLMs on the DSMES and ADCES7 guidelines would be helpful for teenage patients when navigating the self-management of Type 2 Diabetes.

References:

1. Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., ... & Longhurst, C. A. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine, 183*(6), 589–596. https://doi.org/10.1001/jamainternmed.2023.1838
2. Lawrence, J. M., Divers, J., Isom, S., Saydah, S., Imperatore, G., Pihoker, C., ... & Dabelea, D. (2021). Trends in prevalence of type 1 and type 2 diabetes in children and adolescents in the United States, 2001–2017. *New England Journal of Medicine, 384*(24), 2242–2254. https://doi.org/10.1056/NEJMoa2023849
3. Powers, M. A., Bardsley, J., Cypress, M., Funnell, M. M., Harms, D., Hess-Fischl, A., ... & Vivian, E. (2020). Diabetes self-management education and support in adults with type 2 diabetes: A consensus report of the ADCES and ADA. *The Diabetes Educator, 46*(4), 350–369. https://doi.org/10.1177/0145721720930959
4. Hales, C. M., Carroll, M. D., Fryar, C. D., & Ogden, C. L. (2017). Prevalence of obesity among adults and youth: United States, 2015–2016. *NCHS Data Brief, 288*, 1–8. https://www.cdc.gov/nchs/products/databriefs/db288.htm
5. World Health Organization. (2022). *Obesity and overweight*. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight
6. Kornfield, M. (2024, May 20). Family sues chatbot maker after teen's suicide. *The Washington Post*. https://www.washingtonpost.com/technology/2024/05/20/character-ai-suicide-lawsuit/

University of California San Francisco