Assessing Reliability of Artificial Intelligence Transformation of Online Resources on Gracilis Free Muscle Transfer for Facial Reanimation

Davis S. Chong, BS[a,†], William M. Swift, MD[b], Nina D. Ham[a], BS, Travis T. Tollefson, MD, MPH[b], Alexander P. Marston, MD[b]

[a] University of California Davis School of Medicine
[b] Department of Otolaryngology – Head and Neck Surgery, University of California Davis School of Medicine, Sacramento, California, USA

UCDAVIS HEALTH | SCHOOL OF MEDICINE

UCDAVIS HEALTH | Department of Otolaryngology Head and Neck Surgery

UCDAVIS UNIVERSITY OF CALIFORNIA

ABSTRACT

Introduction
Generative language models (GLMs) such as ChatGPT stand to revolutionize health literacy by improving accessibility to patient educational materials (PEMs). As a proof of concept, gracilis free muscle transfer (GFMT) exists today as a versatile solution for facial reanimation. However, as a complex and niche procedure, online information on GFMT may exceed the 6th grade reading level set by the American Medical Association (AMA) for PEMs. Analyzing the readability of patient-directed websites on GFMT and assessing the readability and integrity of GLM output will help assess the accessibility of education for this highly subspecialized surgery.

Methods
The first nine non-repeated patient-directed websites from a Google Search for "gracilis free tissue transfer" and other related searches were evaluated by five standardized readability formulas: Automated Readability Index (ARI), Coleman Liau Index (CLI), Gunning Fog Score, Simple Measure of Gobbledygook (SMOG), and Flesch Kincaid Grade Level (FKGL).These websites were transformed by GLMs (ChatGPT 4, Claude 3.5, Llama 2) and reanalyzed for readability. Semantic similarity of GLM output to source was assessed by pairwise latent semantic analysis (LSA).

Results
The mean pooled FRE of the 9 websites was 44 (age 18+). The mean grade level by each test used was: FKGL 12.2 (age 17-18), GFS 14.6 (age 18+), SMOG 10.9 (age 14-15), CLI 13.5 (age 18+), and ARI 12.3 (age 17-18). After GLM transformation, the mean pooled FRE increased significantly from 44 to a mean of 64 ($p < 0.05$), indicating improved readability. Average FRE was similar between all three GLMs (67.2, 64.3. 59.7), but semantic similarity decreased with Claude 3.5 (cosine similarity [scale 0 to 1, 1 = identical] =0.37, vs. =0.52 [ChatGPT 4], =0.56 [Llama 2]).

Conclusions
GLMs can improve readability of current online educational resources of GFMT, but while readability improves, content integrity may suffer depending on complexity of original source and GLM utilized.

CONTACT

DAVIS S. CHONG, BS
University of Davis School of Medicine
Email: davchong@health.ucdavis.edu

Introduction

- Generative language models (GLM) are popular tools to make patient educational material (PEM) more accessible[1,2], and could be applied for PEM on gracilis free muscle transfer (GFMT) for facial reanimation[3,4].
- However, GLMs have a concerning tendency to "hallucinate" and provide falsified or altered information as fact[4,5].
- This study aims to evaluate multiple GLMs in how effectively they improve readability of PEMs while assessing content fidelity using Latent Semantic Analysis (LSA).

Objectives

- Evaluate the ability of ChatGPT 4, Claude 3.5, and Llama 2 to improve PEMs according to standardized readability formulas.
- Analyze the semantic similarity of transformed PEM to their original texts to approximate the reliability of transformed output from GLMs
- Perform head-to-head comparison of output from GLM to identify potential ideal tool for patient use.

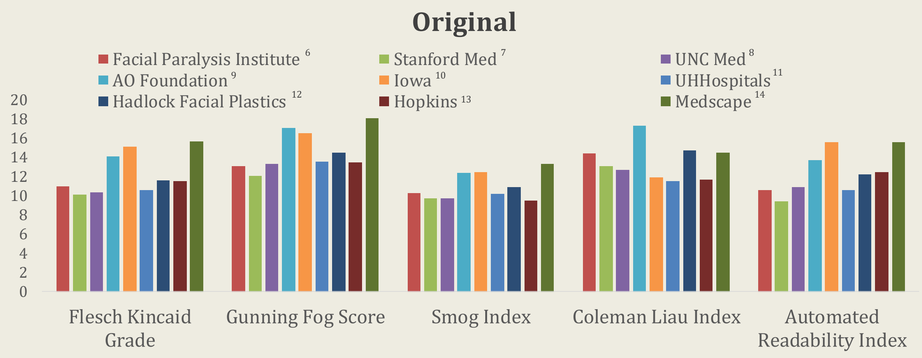
Methods

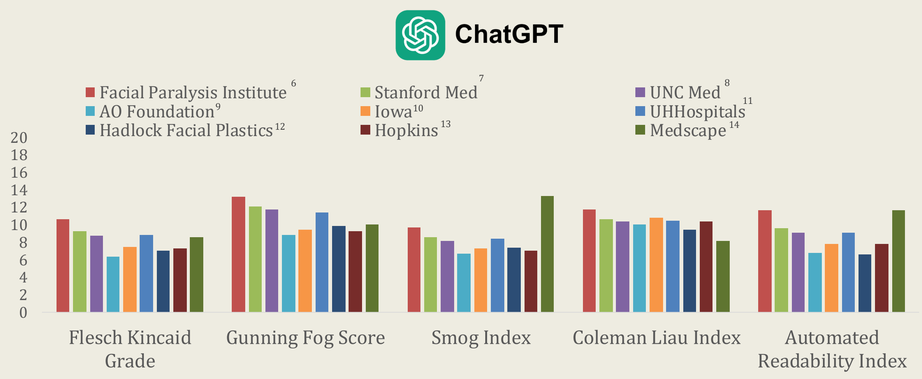
- The first nine non-repeated patient-directed websites from a Google Search for "gracilis free muscle transfer" OR "gracilis free flap" AND "facial reanimation" OR "facial paralysis" were evaluated by five standardized readability formulas as described below. These websites were transformed by GLMs (ChatGPT 4, Claude 3.5, Llama 2) using a standardized prompt and reanalyzed for readability. Semantic similarity of GLM output to source was assessed by pairwise latent semantic analysis (LSA).

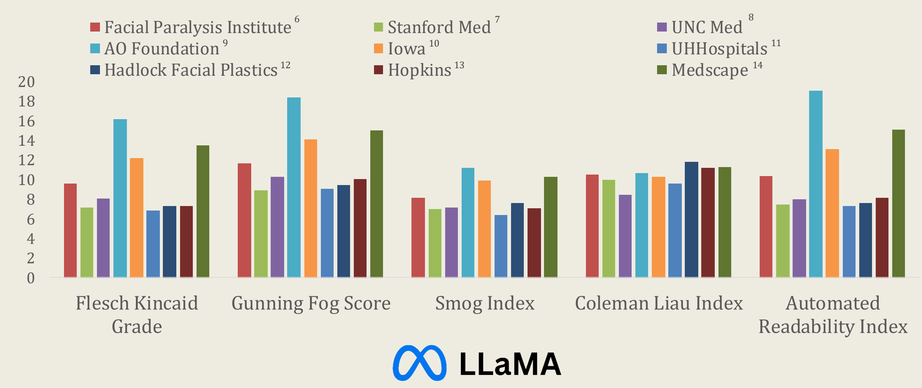
		Formula	Score Type	Goal
Automated Readability Index (ARI)	Approximates the American school grade required to comprehend material	G = (4.71*(C/W)) + 0.5*(W/S)) – 21.43	US School Grade Level	≤6-8
Coleman Liau Index (CLI)	Approximates the American school grade required to comprehend material	G = ((5.88*C)/W) – ((29.5*S)/W) – 15.8	US School Grade Level	≤6-8
Gunning Fog Score (GFS)	Estimates the years of formal education a person needs to understand material (6–17)	G = 0.4*(W/S + ((X/W)*100))	US School Grade Level	≤6-8
Simple Measure of Gobbledygook (SMOG)	Estimates the number of years of education needed to understand 100% of a piece of writing (5–18)	G = 1.0430* √(X + (30/S)) + 3.1291	US School Grade Level	≤6-8
Flesch Kincaid Grade Level (FKGL)	Approximates the American school grade required to comprehend material	G = (11.8*(B/W)) + (0.39*(W/S)) – 15.59	US School Grade Level	≤6-8

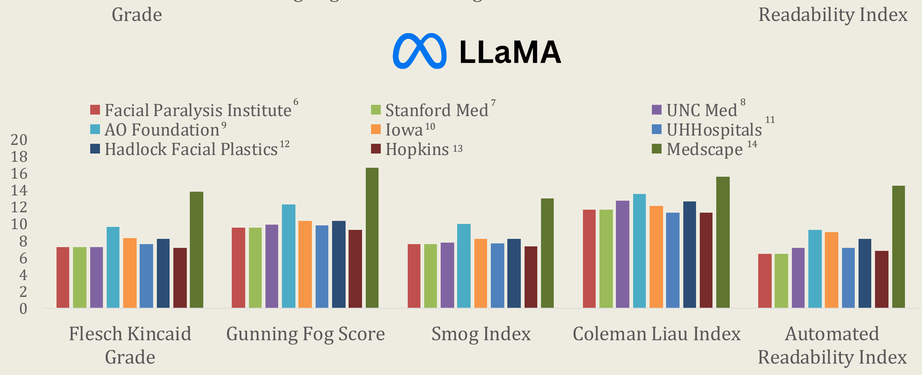
G = Grade; C = characters; W = words; S = sentences; X = complex words; B = syllables; I = Index score

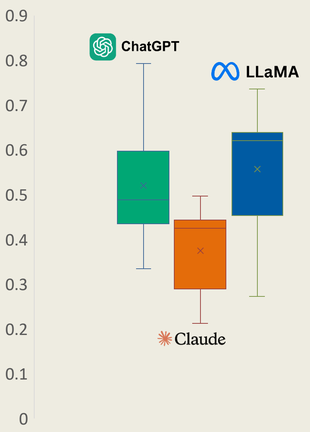
Results

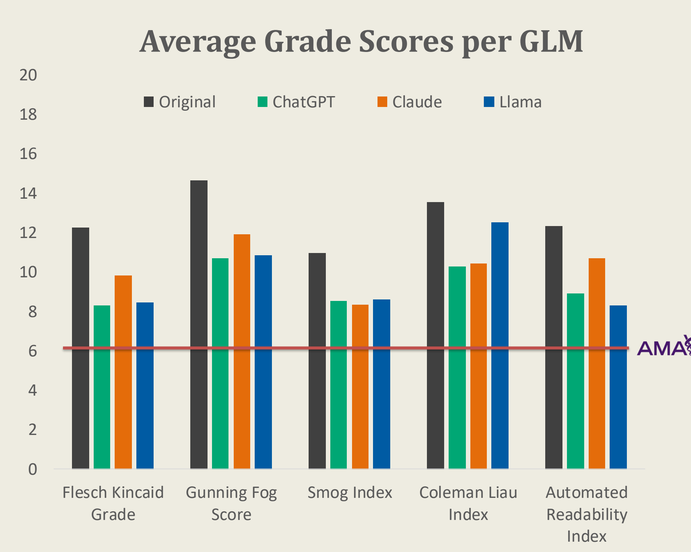












Conclusions

- All three GLMs were able to achieve notable improvement in readability scores but ultimately could not achieve exact thresholds despite specific request.
- Poor readability of original PEM appears to result in difficulty improving readability by GLM.
- ChatGPT and LLaMA were able to achieve comparable levels of semantic similarity to source material, retaining approximately 55% of the same topics in average.
  - Claude appears to suffer significant loss in semantic content while maintaining comparable performance in readability to other GLMs.

References