Clinical leaflets vs AI: Benchmarking eye health advice from generative artificial intelligence in terms of factual accuracy, safety, comprehensiveness and readability

Aleksander Stupnicki[1], Bernardo Souza Mendes[1], Maxwell J.B. Reinstein[1], Ariel Yuhan Ong[2,3], Andrew Malem[4,5], Pearse A. Keane[2,3], Arun J. Thirunavukarasu[6]

1. University College London Medical School, University College London, London, UK
2. Institute of Ophthalmology, University College London, London, UK
3. Moorfields Eye Hospital NHS Foundation Trust, London, UK
4. Ophthalmology, Integrated Surgical Institute, Cleveland Clinic Abu Dhabi, Abu Dhabi Emirate, UAE
5. Department of Ophthalmology, Cleveland Clinic Lerner College of Medicine, Case Western Reserve University, Cleveland, Ohio, US
6. International Centre for Eye Health, London School of Hygiene and Tropical Medicine, London, UK

INTRODUCTION

BACKGROUND

- **32% of adults use generative AI (genAI) for healthcare advice**[1], and 25% of ChatGPT users ask health questions weekly[2]
- On-demand, tailored responses from genAI chatbots can **improve patients' health literacy and engagement**[3,4]
- However, **general-purpose genAI chatbots lack regulatory approval**[5] for medical use, and **significant risks** have been raised in the literature [6]
- Prior evaluations in ophthalmology focus on exam performance, evidence on patient-facing advice quality is limited[7]

AIM

- **Can leading general-purpose LLMs produce eye health advice of comparable 'quality' to clinical patient information leaflets?**
- **Aim**: Benchmark GPT-5 and Gemini 3 against NHS ophthalmology patient information leaflets across four domains: **accuracy, comprehensiveness, safety, and readability**

METHODOLOGY

DATASET

- **Control arm:** Patient information leaflets (NHS Manchester Royal Eye Hospital)
- **Eye diseases evaluated (n = 9):** age-related macular degeneration (AMD), cataract, Charles Bonnet syndrome (CBS), diabetic retinopathy (DR), dry eye disease, glaucoma, nystagmus, posterior vitreous detachment (PVD), and retinal detachment (RD)

GENERATIVE AI ADVICE GENERATION

- **Models:** GPT-5 (OpenAI, USA) and Gemini 3.0 (Google DeepMind, UK); latest, most popular models at the time of study; default settings
- **Prompts:** Outputs were generated using NHS leaflet subheadings as verbatim prompts with word-count constraints to match control text length
  - Example: *"What is age-related macular degeneration (AMD)? Restrict your answer to 154 words"*

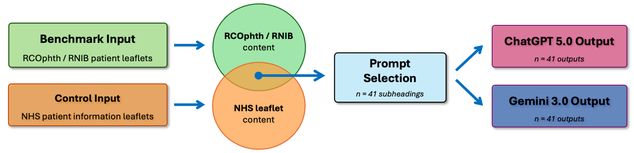
EVALUATION

- **Scoring framework:** Comprehensiveness, accuracy, safety evaluation framework (CASEF); -2 to +2 bidirectional scale to quantify concordance between benchmark and genAI / NHS leaflet text
- **Factual accuracy:** Normalised total CASEF scores (-100% to +100%)
- **Comprehensiveness:** Number of omissions (CASEF score = 0)
- **Safety:** Number of checklist items with negative CASEF scores (-1, -2) + qualitative evaluation by a senior ophthalmology resident (AYO) and a consultant ophthalmologist (AM)
- **Readability:** Simple Measure of Gobbledygook (SMOG) Index, Flesch-Kincaid Grade Level (FKGL), and Automated Readability Index (ARI)
- **Scoring panel:** Hybrid, semi-automated panel of 5 LLMs and 1 human evaluator (LLM-as-a-judge methodology)
  - **Inter-rater reliability:** Intraclass correlation coefficient (ICC(2,1)) and preferential scoring were evaluated → ICC=0.843; no preferential scoring

BENCHMARK

- **Source material:** Patient information leaflets (by the Royal College of Ophthalmologists (RCOphth) and the Royal National Institute of Blind People (RNIB))[7]
- **Benchmark development:** The benchmark was manually derived from RCOphth/RNIB leaflets by translating each statement into a checklist point

#1 CONTROL TEXT SELECTION AND GENAI OUTPUT GENERATION



#2 BENCHMARK CREATION

#3 BENCHMARKED EVALUATION (GENAI vs CLINICAL LEAFLETS)

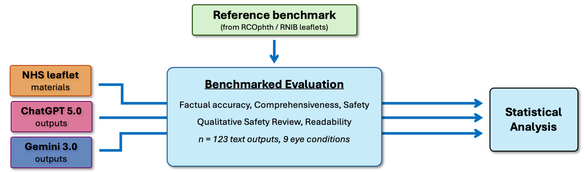


Figure 1: Study workflow overview showing three-phase design.

RESULTS

SAFETY

Blinded clinician (AYO, AM) review:

- **Missing safety-netting:** critical red flag symptoms were omitted or under-emphasised, including warning signs for endophthalmitis and retinal re-detachment
- **Regional misalignment:** both models provided US-centric advice, including drug recommendations unavailable in the UK and referral pathways inconsistent with the UK system
- **No hallucinations:** no overt hallucinations were found, but sporadic phrasing inaccuracies were identified

Quantitative safety review:

- **Insufficient complication reporting:** NHS leaflets provided more comprehensive information treatment complications (p<0.01 both)

ACCURACY & COMPREHENSIVENESS

- **Accuracy:** GPT-5 > NHS leaflets; Gemini 3 > NHS leaflets (both $p < 0.001$)
- **Comprehensiveness**: GPT-5 > NHS leaflets; Gemini 3 > NHS leaflets (both $p < 0.001$)

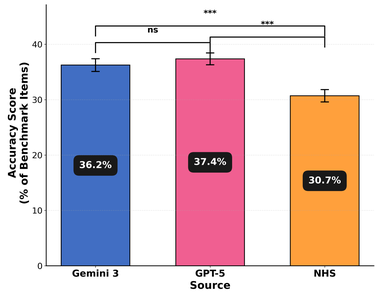


Figure 2: Mean accuracy scores (% maximum CASEF score across 547 benchmark items) for Gemini 3 (36.2%), GPT-5 (37.4%), and NHS leaflets (30.7%). Both genAI models significantly outperformed NHS leaflets (*** $p < 0.001$); no significant difference (ns) was observed between AI models.

READABILITY

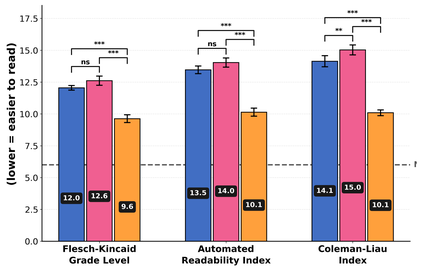


Figure 3: Readability scores across all three sources (blue = Gemini 3; pink = GPT-5; orange = NHS). Lower readability grade = more readable. The dashed line represents NIH 6th grade readability recommendation for all patient-facing materials

- **GenAI requires three additional years of education to comprehend** (fully understood by 14% of UK adults)
- **Reading grade (FKGL)**: GPT-5: 12.6 (A-level), Gemini 3: 12.1 (A-level), NHS: 9.6 (GCSE), all p<0.001 vs. NHS
- **No source met NIH-recommended reading level** for patient materials (6th-8th reading grade)

CONCLUSIONS

KEY FINDINGS

- General-purpose **AI chatbots can match or exceed clinical patient education materials** in **accuracy** and **comprehensiveness**.
- **Significant safety concerns remain** in readability, underreporting of treatment complications, regional guideline alignment and safety netting
- Therefore, general-purpose **genAI models cannot be endorsed as a safe, standalone patient information source**, despite their advantages for patient education

CLINICAL & SCIENTIFIC IMPACT

- Clinicians should **educate patients to approach eye health materials produced by genAI with caution**
- Further **development and regulation** should ensure that genAI health advice meets established healthcare standards
- As genAI accuracy approaches parity with clinical materials and safety risks shift from hallucinations to subtle errors, future genAI research should **implement clinician-in-the-loop evaluation frameworks**

GLOSSARY

- **GenAI:** Software that can generate text or other media forms (e.g. ChatGPT)
- **LLMs:** Large language models; technology that powers genAI chatbots
- **Benchmarking:** Comparing against a trusted standard (here RCOphth leaflets)
- **CASEF score:** Scoring framework used to assess accuracy, safety and comprehensiveness by measuring text alignment with the benchmark
- **Hallucinations:** GenAI confidently stating something entirely made up as true
- **FKGL (reading grade):** Validated metric that estimates years of schooling needed to understand text. Lower score = easier to read.

Contact: aleksander.stupnicki.22@ucl.ac.uk

References are available here: https://www.mybib.com/b/AM0DA2