

A Multistakeholder Assessment of Generative Artificial Intelligence Platforms for Creating Plain Language Summaries

POSTER #14

Amy Shaberman,¹ Andy Shepherd,² Ann Marjolein Hepping,³⁻⁴ Catherine Richards Golini,⁵ Jonathan Patience,⁶ Tom Smith,⁷ Sophie Randall,⁸ Joanne Walker,⁹ Trishna Bharadia¹⁰

¹Boehringer Ingelheim Pharmaceuticals Inc., Ridgefield, CT, USA. ²Envision Pharma Group, Wilmslow, UK. ³Roessingh Centre for Rehabilitation, Enschede, The Netherlands. ⁴Department of Orthopedics, University Medical Center, Groningen, The Netherlands. ⁵Karger Publishers, Basel, Switzerland. ⁶Taylor & Francis Group, London, UK. ⁷Thomas Smith Associates, Newcastle, UK; Patient Author. ⁸Patient Information Forum, London, UK. ⁹Becaris Publishing Ltd, Royston, UK. ¹⁰The Spark Global, Buckinghamshire, UK; Patient Author.



Background

- The last decade has seen multiple initiatives aimed at removing paywalls to make medical research publications free to access, but physical access alone is not enough to make this information accessible
- Technical terms and jargon frequently used in these publications form a language barrier to non-specialists, creating a need for plain language summaries (PLS) of published articles
- Could generative artificial intelligence (AI) help to lower the barriers by assisting the rapid generation of PLS?

Objectives

- This pilot study assessed different generative AI platforms and approaches for the creation of 250-word PLS of medical research publications
- Here we present the results for PLS generated by a single simple prompt

Methods

- PLS were generated from two different manuscripts^{1,2} using 18 generative AI platforms (free and paid), including general, document-specific, and science-specific platforms
- Prompt wording varied according to the platform requirements, but specified creation of a 250-word summary for each of the two scientific articles using language suitable for a 12-year-old reading age
- All PLS underwent initial technical assessment and were scored for content relevance, accuracy, and reading level^{3,4}
- The highest scoring PLS were then assessed by various stakeholders (author or pharma sponsor of the original article, health care professionals, patient advocates, medical journal editors, general public) using a Likert scale to rate each for understandability, ease and appropriate level of reading, content inclusion and usefulness, and ranked by preference

Results

- Of the 18 platforms initially assessed, the following were excluded:
 - 1 general platform (Summarize) was nonfunctional at the time of the assessment
 - 4 science-specific (Stork, Elicit, Scite, Cohere) and 4 general platforms were unable to generate content in an appropriate format (Classgist, Writeful, Editpad, Bard)
 - 1 document-specific tool (HiPDF) returned content irrelevant to the source articles
- The format returned varied greatly between platforms, from bulleted lists and titled sections to block text, with some platforms also including links to Wikipedia pages for key terms
- AI generally appeared to perform well when extracting key themes and producing broad topline summaries, but many failed to extract specific data, hence avoiding inaccuracy (Data errors in table below) through exclusion

Summary of technical assessments for PLS sent for stakeholder review

	Word count	Data errors ^a	Reading age (range) ^b	FDA readability score (/5) ^c	EMA readability score (/5) ^d
Paper A	NA	NA	21–22 year olds (18, 22)	0	NA
Original abstract	354	NA	21–22 year olds (17, 18+)	0	1
ChatPDF	252	No	17–18 year olds (14, 19+)	1	4
Claude	252	No	11–13 year olds (10, 14)	5	5
GPT3-5	276	No	10–11 year olds (10, 13)	5	5
GPT4	225	No	12–14 year olds (10, 18)	5	5
Perplexity	171	Yes	14–19 year olds (13, 18)	3	5
Scholarcy	218	Yes	17–18 year olds (15, 18+)	1	2
SciSummary	139	No	18–19 year olds (15, 22)	0	3
Paper B	NA	NA	14–15 year olds (12, 16)	3	NA
Original abstract	435	NA	12–14 year olds (10, 15)	5	0
ChatPDF	194	Yes	13–15 year olds (13, 17)	4	5
Claude	230	Yes	17–18 year olds (13, 18)	1	5
GPT4	309	No	12–14 year olds (11, 18)	5	5
PDFelement	245	No	14–15 year olds (11, 18)	3	3
Perplexity	134	No	14–15 year olds (14, 17)	3	4
SciSummary	253	Yes	College graduates (18+)	0	2

^aFactual errors present in the summary (Y/N). ^bBased on a consensus from the Automated Readability Index, Coleman-Liau Index, Flesch-Kincaid Grade Level, Flesch Reading Ease score, Gunning Fog, Linsear Write Formula, and SMOG Index. ^cBased on a maximum score of 5 minus the number of grades above recommended reading level. ^dBased on assessment of sentence length and syntax, use of unexplained technical terms, use of abbreviations and acronyms, use of the passive voice, and UK government reading age guidance (1 point each).

Conclusions

- AI-generated PLS were generally good at producing coherent and readable text but the content generated was not necessarily complete or appropriate in terms of tone or the language used
- Paid platforms were no guarantee of quality and performed no better than free platforms in this assessment
- If useful, scientifically accurate and appropriate PLS are to be practically produced with limited involvement of subject matter experts:
 - More detailed prompts, sequential, or iterative prompts could address several of the issues identified and have been investigated at subsequent stages of this pilot
 - AI platforms would benefit from further refinements to improve data extraction and accuracy while maintaining an appropriate tone and using language suitable for a lay audience

Stakeholder assessments and rankings of generative AI platforms

		Understanding (higher is better)	Readability (higher is better)	Appropriate level (higher is better)	Presence of desired information (higher is better)	Usefulness (higher is better)	Preference ranking (lower is better)
ChatPDF	Author/Sponsor n=2	5	5	2.5	2.5	2.5	4
	General n=19	5	4.64	2	3.1	3.1	3.6
	HCP n=6	4.75	4.75	3	3	3	3.75
	Journal editor/Publisher n=6	4.88	4.38	2.75	2.75	2.88	4
	Patient advocate n=4	5	5	3.5	4.25	4.25	1.75
	Pop science editor n=2	5	5	2	2	4	5
Claude	Author/Sponsor n=2	5	5	3	4.5	5	2
	General n=19	5	4.82	2.73	4.14	4.14	1.3
	HCP n=5	4.75	3.75	2.5	3.5	3.5	2
	Journal editor/Publisher n=6	4.88	4	2.75	3.88	4.25	1
	Patient advocate n=4	5	4.67	4	4	4.33	3
	Pop science editor n=2	5	5	4	4	4	1
GPT3-5	Author/Sponsor n=2	4	2	1	1	1	5
	General n=10	4.83	4.5	1.5	2.83	3	3.83
	HCP n=2	4.5	3.5	3	2.5	2.5	4
	Journal editor/Publisher n=3	4.5	4	2	2.5	3	3.5
	Patient advocate n=2	4.5	4.5	3	2	3	4.5
	Pop science editor n=1	5	5	3	1	1	6
GPT4	Author/Sponsor n=2	5	5	2	2	5.5	2
	General n=19	5	4.73	1.77	2.36	2.36	5.1
	HCP n=5	4.8	3.4	2	1.6	1.6	5.4
	Journal editor/Publisher n=6	3.13	3.13	1.5	1.63	1.63	5.63
	Patient advocate n=4	4.75	4.75	2	2.75	2.75	5.25
	Pop science editor n=2	5	5	4	4	4	3
PDFelement	Author/Sponsor n=2	5	5	2	2	3	5
	General n=9	5	4.6	2.7	3.4	3.5	3.8
	HCP n=3	4.67	4	3	3	3.67	2.67
	Journal editor/Publisher n=3	4.75	4.5	2.75	2.75	3.25	3.25
	Patient advocate n=2	5	4.5	3	2	4	3
	Pop science editor	–	–	–	–	–	–
Perplexity	Author/Sponsor n=2	5	5	2.5	1.5	1	6
	General n=19	5	4.36	1.82	2.27	2.27	5.4
	HCP n=5	4.75	4.75	3	2	2	4.5
	Journal editor/Publisher n=6	4.5	4	2.25	2.25	2.25	4.25
	Patient advocate n=4	5	5	3	3.5	3.5	3.25
	Pop science editor n=2	5	5	3	4	4	4
Scholarcy	Author/Sponsor n=2	1	1	4	1	1	7
	General n=10	4.29	2.86	3.71	3	2.71	4.67
	HCP n=3	4	1	5	3	1	7
	Journal editor/Publisher n=3	1.5	1	5	2	1	7
	Patient advocate n=2	3.5	2	4.5	3.5	2.5	5.5
	Pop science editor n=1	3	1	4	4	1	7
SciSummary	Author/Sponsor n=2	5	5	2.5	3.5	4	2
	General n=19	5	4.5	3.13	4	3.25	3.33
	HCP n=5	4.75	3.75	3	3.5	3.5	3.25
	Journal editor/Publisher n=6	4.75	3.88	3.38	4	4	2.63
	Patient advocate n=4	5	4	3.33	3.33	3	5.33
	Pop science editor n=2	5	5	4	4	4	2

Heatmap shows unweighted average scores from red to green (low to high).

- With the exception of Scholarcy, all platforms were able to generate coherent text summaries that were rated as readable (average 4.5/5 vs 2.0/5 Scholarcy) and understandable (average 4.8/5 vs 3.3/5 Scholarcy)
 - Scholarcy** was the lowest-rated platform, producing summaries that were well structured but included complex technical terms and abbreviations without explanations, and was generally seen as far too complex for a lay audience
- There was significant variation in the level of details included by the different platforms, as well as in the tone and complexity of language used
- The highest-rated summaries overall were Claude and SciSummary
 - Claude** scored well for the utility of the summary (Av 4.2/5) with clear explanations and a good overall tone
 - The use of some informal language, particularly the term “kids” in place of “children”, was seen as inappropriate, negatively affecting how the summary was rated and perceived
- SciSummary** produced well-structured and accurate summaries but did not provide definitions for some key terms
- Despite scoring very poorly in technical readability assessments, SciSummary was rated well for readability across all stakeholder groups and was the highest-rated platform for pitching summaries at an appropriate level
- Links to Wikipedia pages for key terms were seen as useful but links for some simple terms such as “child” were also included and were seen as patronizing

Study limitations

- This pilot study only included two manuscripts with a limited number of stakeholder assessors, and was not powered for a statistical comparison
- While patient advocates were represented in the stakeholder assessments, no patients with the conditions described in the manuscripts were included in this pilot study

References

1. Hepping AM, et al. Recovery of strength after reduced pediatric fractures of the forearm, wrist or hand: A prospective study. *PLoS One*. 2020 Apr 1;15(4):e0230862;

2. Menter A, et al. Switching between adalimumab reference product and BI 695501 in patients with chronic plaque psoriasis (VOLTAIRE-X): a randomized controlled trial. *Am J Clin Dermatol*. 2022 Sep;23(5):719–728;

3. <https://readabilityformulas.com/>;

4. <https://hemingwayapp.com/>.

Disclosures

In addition to the use of generative AI for the creation of the PLS assessed in this study, AI was also used to assist the collation of responses and for the initial data analysis. Author disclosures are available upon request.

Acknowledgments

Editorial and creative support were provided by Jim Kerr and Patrick Spiller of Envision Pharma Group



Please scan this QR code to view a digital version of the poster