Supplement material

Supplement Table1: Brain aneurysm patients imaging reports utilized in this study

Supplement Table 2 List of equations for quantitative evaluation to assess summary text

Supplement Table 3: Methods for evaluating model-generated summaries

Qualitative evaluation of summaries was performed in four categories: readability, accuracy, comprehensiveness, and redundancy. A score of 1 to 5 was assigned to each category based on the descriptions/number of relevant factors present for each category score.

Supplement Table 4 Metric and expert evaluation of summarization of clinical longitudinal imaging reports

(I-III) Quantitative and qualitative evaluation of clinical imaging reports summary: Performance by model is in bold and underlined, second best is in bold only, and third best is underlined only. (IV) Comparing length reduction by models. Lower percentages indicate shorter summaries.

Supplement Table 5: Summarization of 100 case reports compared vs. two reference standards and through expert evaluation.

Quantitative evaluation with ROUGE-1, ROUGE-2, ROUGEE-L, and BERTscores was performed to compare summaries created by a human and from case report figure captions. Quality evaluation was scored by two experts. Models listed based on overall performance (best at the top).

Supplement Table 6: List of NLP summarization models used in this study

Supplement Table 7:: Example of NLP summary generated by different models

Supplement Table 1: Brain aneurysm patients imaging reports utilized in this study.

	Total
Patient Information	
Gender	
Female	44
Male	8
Race	
American Indian/Alaskan Native	0
Asian	4
Native Hawaiian or Other Pacific Islander	0
Black or African American	5
White	30
More than One Race	1
Unknown or Not Reported	12
Ethnicity	
Not Hispanic or Latino	42
Hispanic or Latino	8
Unknown or Not Reported	2
Smoker	
Yes	20
No	32
Aneurysm Multiplicity	
Yes	15
No	37
Ruptured Aneurysm	
Yes	4
No	48
Personal SAH History	
Yes	4
No	48
Family SAH History	
Yes	1
No	51
Family Aneurysm History	
Yes	5
No	47
Imaging Follow-up Information	
Aneurysm follow-up time	
Months between first and second visit (Mean±SD)	15.35±23.74
Months between second and third visit (Mean±SD)	12.36±10.20
Imaging Modality	
Magnetic resonance angiography	98
Computed tomography angiography	19
Digital subtraction angiography	20

Supplement Table 2 List of equations for quantitative evaluation to assess summary text

ROUGE_1
$Precision = \frac{number\ of\ unigrams\ in\ both\ reference\ and\ model\ summary}{number\ of\ unigrams\ in\ the\ model\ summary}$ $Recall = \frac{number\ of\ unigrams\ in\ both\ reference\ and\ model\ summary}{number\ of\ unigrams\ in\ the\ reference\ summary}$ (Unigram refers to a single word) $ROUGE\_1 = F1 = 2 \times \left(\frac{Precision\ \times\ Recall}{Precision + Recall}\right)$
ROUGE_2
$Precision = \frac{number\ of\ bigrams\ in\ both\ reference\ and\ model\ summary}{number\ of\ bigrams\ in\ the\ model\ summary}$ $Recall = \frac{number\ of\ bigrams\ in\ both\ reference\ and\ model\ summary}{number\ of\ bigrams\ in\ the\ reference\ summary}$ (Bigrams refers to a sequence of text that is two words long.) $ROUGE\_2 = F1 = 2 \times \left(\frac{Precision\ \times\ Recall}{Precision + Recall}\right)$
ROUGE_L
$Precision = \frac{number\ of\ unigrams\ in\ LCS}{number\ of\ unigrams\ in\ the\ the\ model\ summary}$ $Recall = \frac{number\ of\ unigram\ in\ LCS}{number\ of\ unigrams\ in\ the\ referencel\ summary}$ **(** LCS refers to the longest common sequence of text present in both the model and reference summary) $ROUGE\_L = F1 = 2 \times \left(\frac{Precision\ \times\ Recall}{Precision + Recall}\right)$
BERTscore
$Precision = \frac{1}{\lvert\hat{x}\rvert}\sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} x_i^T \hat{x}_j$ $Recall = \frac{1}{\lvert x\rvert}\sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} x_i^T \hat{x}_j$ Reference summary x, and Model summary,$\hat{x}$, are tokenized. Tokens are compared through computing similarity. Highest similarity scores are used to calculate Precision and Recall

$BERTscore = F1 = 2 \times \left(\frac{Precision \times Recall}{Precision + Recall}\right)$
Text Reduction
$Reduction = \frac{Number\ of\ words\ in\ summary}{Number\ of\ words\ in\ original\ text}$

Supplement Table 3: Methods for evaluating model-generated summaries

Quality Evaluation Score					
Category	1	2	3	4	5
Readability	Lack of sentence structure, and includes many grammatical and spelling errors	Poor sentence structure, with many grammatical and spelling errors	Passable sentence structure with 3-4 errors, several grammatical or spelling errors	Decent sentence structure with 1-2 errors, with minimal grammatical and spelling errors	Good sentence structure, closely mimics human writing style with minimal grammatical and spelling errors
Accuracy	Information provided is completely contradictory or incorrect to the original text	5-6 errors regarding information accuracy	3-4 errors regarding information accuracy	1-2 errors regarding information accuracy	No inaccurate information
Comprehensiveness	Summary does not include any relevant information such as patient info, aneurysms, dates, or imaging modalities. Reader truly cannot comprehend the summary	Many missing key elements, reader cannot comprehend the summary's content without referring to the original report	Some (3-4) key elements are missing, reader is able to somewhat comprehend the content	Few (1-2) key elements missing, but the reader is still able to comprehend the majority content	All relevant information is included within the summary
Redundancy	Summary repeats many words or sentence structure elements. Reading the original report is preferable	Summary is not significantly condensed, with much redundancy in information and/or wording	Summary is wordy, with some redundancy in information and wording	Summary is somewhat concise, with little redundancy in information and wording	Summary is as concise as possible with no unnecessarily repeated information or redundant use of words

Expert qualitative evaluation rubric: Qualitative evaluation of model summaries were evaluated in four categories: readability, accuracy, and comprehensiveness. A score of 1 to 5 is assigned to each category based on the descriptions/number of relevant factors present for each category score.

Supplement Table 6. List of NLP summarization models used in this study

Model	Developer	Version	Base Model	Web Link
GPT3davinci	OpenAI	Nov.2022	(GPT) OpenAI, CA	https://platform.openai.com/playground
BARTcnn	Facebook	Jan 2023	Facebook, Meta AI, CA	https://huggingface.co/facebook/bart-large-cnn
LEDBooksum	P. Szemraj	Dec 2022	Allen Institute for AI, WA	https://huggingface.co/pszemraj/led-large-book-summary
LongT5booksum	P. Szemraj	Feb 2023	Google, CA	https://huggingface.co/pszemraj/long-t5-tglobal-base-16384-book-summary
LEDlegal	C. Ram	Nov 2022	Allen Institute for AI, WA	https://huggingface.co/0-hero/led-large-legal-summary
LEDclinical	G. Adams	Feb 2022	Allen Institute for AI, WA	https://huggingface.co/griffin/clinical-led-summarizer

Supplement Table 7. Example of NLP summary generated by different models

	Human Gold Standard	GPT3Davinci	BARTcnn	LONGT5booksum	LEDBooksum	LEDlegal	LEDclinical
Summary	60-year-old male with worsening headaches, fatigue, memory loss, and behavioral abnormalities for 2-3 weeks	A 60-year-old male with a benign past medical history presented with headaches, fatigue, memory loss, and behavioral abnormalities.	A 60-year-old male with a benign past medical history presented with headaches, fatigue, memory loss, and behavioral abnormalities for 2–3 weeks before admission.	This paper describes a patient with a past medical history who presents with headaches and fatigue for two to three weeks before his admission.	In this brief report, the author describes a 60-year-old man who has been suffering from headaches and fatigue for 2-3 weeks and whose travel history is unremarkable.	the patient is a 60-year-old healthy male with a benign past medical history who has been suffering from headaches for the past few weeks.	Brief Hospital Course:
Comments:	Human-created summary	Understandable and comprehensive, with most information present.	Direct text extraction from source report. High accuracy and comprehensive.	Provides basic information on patient, but frames visit as a paper.	General information is present and accurate. Obtained information regarding travel history from next sentence.	Accurate information, though sentence is incorrectly capitalized.	Lacks any information usable as a summary
ROUGE-1	N/A	0.625	0.667	0.222	0.244	0.286	0.000
ROUGE-2	N/A	0.516	0.514	0.000	0.100	0.056	0.000
ROUGE-L	N/A	0.625	0.667	0.222	0.244	0.286	0.000
BERTScore	N/A	0.9257	0.9416	0.8955	0.902	0.8778	0.8184
Readability	5	5	5	5	5	4	1
Accuracy	5	5	5	5	5	4	1
Comprehensiveness	5	4	5	4	4	4	1
Redundancy	5	5	5	4	4	5	1