Evaluating Online Medical Information Quality in Colon Surgery Using Pre-Tuned Large Language Models.

Vincent Ochs, Sébastien Muheim, Florian Ponholzer, Dietmar Öfner, Antonio J. Rodríguez-Sánchez, Mariana Flaifel, Amit Kumar, Julia Wolleb, Florentin Bieder, Stephanie Taha-Mehlitz, Michael Drew Honaker, Robert Rosenberg, Philippe C. Cattin, Anas Taha

Submitted to: Journal of Medical Internet Research
on: August 14, 2025

Table of Contents

Evaluating Online Medical Information Quality in Colon Surgery Using Pre-Tuned Large Language Models.

Vincent Ochs[1*]; Sébastien Muheim[1]; Florian Ponholzer[2]; Dietmar Öfner[2]; Antonio J. Rodríguez-Sánchez[3]; Mariana Flaifel[4]; Amit Kumar[5]; Julia Wolleb[1]; Florentin Bieder[1]; Stephanie Taha-Mehlitz[6]; Michael Drew Honaker[7]; Robert Rosenberg[8]; Philippe C. Cattin[1*]; Anas Taha[1, 2, 8*]

[1]Department of Biomedical Engineering University of Basel Allschwil CH
[2]Department of Visceral, Transplant and Thoracic Surgery Center of Operative Medicine Medical University of Innsbruck Innsbruck AT
[3]Department of Computer Science, Intelligent and Interactive Systems University of Innsbruck Innsbruck AT
[4]Department of Surgery, St. George's University of London London GB
[5] Data Since and Statistic Medoc Basel CH
[6]Department of Visceral Surgery, University Center for Gastrointestinal and Liver Diseases St. Clara Hospital and University Hospital Basel CH
[7]Department of Surgery Brody School of Medicine East Carolina University Greenville US
[8]Department of Visceral Surgery Cantonal Hospital Basel-Land Liestal CH
[*]these authors contributed equally

Corresponding Author:
Anas Taha
Department of Biomedical Engineering
University of Basel
Hegenheimermattweg 167c
Allschwil
CH

Abstract

Background: Online medical information in colon surgery, particularly in robotic-assisted procedures, is frequently inconsistent, outdated, or misleading, which can hinder patient education and informed decision-making. Large Language Models (LLMs) have shown promise in complex text analysis, but their potential for assessing the quality of medical web content has not been systematically explored. The Ensuring Quality Information for Patients (EQIP) 36 criteria offer a standardized framework for evaluating patient-facing materials, but manual assessments are time-consuming and resource-intensive.

Objective: To systematically evaluate the ability of multiple transformer-based LLMs to classify the quality of online information on robotic colon surgery according to the EQIP 36 criteria, and to determine whether embedding-oriented or generative models provide better performance for this task.

Methods: English-language web pages on robotic colon surgery were collected in December 2023 via automated searches on Google, Bing, and Yahoo using predefined keywords. From 450 retrieved URLs, duplicates, irrelevant content, and professional/academic sites were removed, resulting in 47 unique patient-targeted websites with expert EQIP ratings. Due to input length limitations, texts were split into overlapping 512-token chunks (50-token overlap) before embedding extraction. Models evaluated included base BERT, custom pre-trained BERT, BiomedBERT, pre-trained BiomedBERT, SBERT, LLaMA-2 (concatenated and averaged embeddings), and Longformer. Chunk embeddings were concatenated into document-level embeddings and classified via a feedforward multilabel neural network predicting 19 EQIP items. Hyperparameters (batch size, learning rate, dropout) were optimized via grid search. Performance was assessed using stratified 5-fold cross-validation, with accuracy, precision, recall, F1-score, and ROC-AUC as metrics.

Results: SBERT achieved the highest performance (F1-score = 80.16, accuracy = 79.05%, AUC = 0.79), outperforming all other models. Pre-trained BiomedBERT and custom pre-trained BERT performed comparably well (F1 ? 0.79), indicating benefits from domain-specific adaptation. LLaMA-2 with concatenated embeddings (F1 = 0.784) outperformed the averaged embedding version (F1 = 0.765) but remained below SBERT. Longformer, despite handling longer contexts, had the lowest performance (F1 = 70.07). Across models, those optimized for generating semantically meaningful embeddings consistently surpassed generative architectures.

Conclusions: Transformer-based models optimized for embedding generation, particularly SBERT, are more effective than large

generative models in automated EQIP-based quality assessment of online medical content. Domain-specific adaptation improves performance, but embedding quality is a more critical determinant than model size. These findings support the integration of embedding-focused LLMs into scalable, AI-driven tools for medical content evaluation, potentially enhancing the accessibility and reliability of online patient education. Future research should expand datasets, explore ensemble approaches, and address regulatory and ethical considerations for deployment in healthcare contexts.

(JMIR Preprints 14/08/2025:82400)
DOI: https://doi.org/10.2196/preprints.82400

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <a href="https:/

Original Manuscript

Title Page *1*

Evaluating Online Medical Information Quality in Colon Surgery Using Pre-Tuned Large Language Models.

Vincent Ochs[1], Sébastien Muheim[1], Florian Ponholzer[2], Dietmar Öfner[2], Antonio J. Rodríguez-Sánchez[3], Mariana Flaifel[4], Amit Kumar[5], Julia Wolleb[1], Florentin Bieder[1], Stephanie Taha-Mehlitz[6], Michael Drew Honaker[7], Robert Rosenberg[8], Philippe C. Cattin[1††], Anas Taha[1,2,8††]

1. Department of Biomedical Engineering, Faculty of Medicine, University of Basel, 4123 Allschwil, Switzerland.
2. Department of Visceral, Transplant and Thoracic Surgery, Center of Operative Medicine, Medical University of Innsbruck, 6020 Innsbruck, Austria.
3. Intelligent and Interactive Systems, Department of Computer Science, University of Innsbruck, Innsbruck, Austria.
4. Department of Surgery, St. George's University of London, London, UK.
5. Medoc, Data Since and Statistic, Basel, Switzerland
6. Clarunis, Department of Visceral Surgery, University Center for Gastrointestinal and Liver Diseases, St. Clara Hospital and University Hospital, Basel, Switzerland.
7. Department of Surgery, East Carolina University, Brody School of Medicine, Greenville, NC, USA.
8. Department of Visceral Surgery, Cantonal Hospital Basel-Land, Liestal, Switzerland

[††] Anas Taha and Philippe C. Cattin equally contributed and share the last authorship

Corresponding Author:
Anas Taha,
Department of Biomedical Engineering,
Faculty of Medicine,
University of Basel, Hegenheimermattweg 167C,
4123 Allschwil, Switzerland,
anas.taha@unibas.ch. +41 61 207 5402.

Abstract

Background:
Online medical information in colon surgery, particularly in robotic-assisted procedures, is frequently inconsistent, outdated, or misleading, which can hinder patient education and informed decision-making. Large Language Models (LLMs) have shown promise in complex text analysis, but their potential for assessing the quality of medical web content has not been systematically explored. The Ensuring Quality Information for Patients (EQIP) 36 criteria offer a standardized framework for evaluating patient-facing materials, but manual assessments are time-consuming and resource-intensive.

Objective:
To systematically evaluate the ability of multiple transformer-based LLMs to classify the quality of online information on robotic colon surgery according to the EQIP 36 criteria, and to determine whether embedding-oriented or generative models provide better performance for this task.

Methods:
English-language web pages on robotic colon surgery were collected in December 2023 via automated searches on Google, Bing, and Yahoo using predefined keywords. From 450 retrieved URLs, duplicates, irrelevant content, and professional/academic sites were removed, resulting in 47 unique patient-targeted websites with expert EQIP ratings. Due to input length limitations, texts were split into overlapping 512-token chunks (50-token overlap) before embedding extraction. Models evaluated included base BERT, custom pre-trained BERT, BiomedBERT, pre-trained BiomedBERT, SBERT, LLaMA-2 (concatenated and averaged embeddings), and Longformer. Chunk embeddings were concatenated into document-level embeddings and classified via a feedforward multilabel neural network predicting 19 EQIP items. Hyperparameters (batch size, learning rate, dropout) were optimized via grid search. Performance was assessed using stratified 5-fold cross-validation, with accuracy, precision, recall, F1-score, and ROC-AUC as metrics.

Results:
SBERT achieved the highest performance (F1-score = 80.16, accuracy = 79.05%, AUC = 0.79), outperforming all other models. Pre-trained BiomedBERT and custom pre-trained BERT performed comparably well (F1 ≈ 0.79), indicating benefits from domain-specific adaptation. LLaMA-2 with concatenated embeddings (F1 = 0.784) outperformed the averaged embedding version (F1 = 0.765) but remained below SBERT. Longformer, despite handling longer contexts, had the lowest performance (F1 = 70.07). Across models, those optimized for generating semantically meaningful embeddings consistently surpassed generative architectures.

Conclusions:
Transformer-based models optimized for embedding generation, particularly SBERT, are more effective than large generative models in automated EQIP-based quality assessment of online medical content. Domain-specific adaptation improves performance, but embedding quality is a more critical determinant than model size. These findings support the integration of embedding-focused LLMs into scalable, AI-driven tools for medical content evaluation, potentially enhancing the accessibility and reliability of online patient education. Future research should expand datasets, explore ensemble approaches, and address regulatory and ethical considerations for deployment in healthcare contexts.

Keywords: Large Language Models (LLMs); Medical Text Classification; Online Health Information Quality; Transformer-Based Models; EQIP 36 Criteria

1. Introduction

In the past two decades, the internet has become a pivotal platform for patients seeking medical information. The various health-related data offered by the digital revolution has significantly transformed how patients interact with healthcare services (1, 2). Notably, the field of robotic colon surgery, introduced with the deployment of the da Vinci robot in 2001, represents an area where online information plays a crucial role in patient education (3). Despite the technological advancements and benefits of robotic surgery in enhancing surgical precision and outcomes, disseminating reliable, high-quality patient information online remains a challenge.

The study conducted by Taha et al., 2023, underscores a critical issue: the quality of online information concerning robotic colon surgery is predominantly low, with a substantial portion of available content being inaccurate or misleading (4). This finding shows a pressing need for healthcare providers and medical institutions to actively engage in the creation and curation of content that accurately reflects the current standards and procedures of robotic colon surgery (5, 6). The disparity in the quality of online patient information not only complicates the decision-making process for patients but also poses a risk of misinformation, potentially influencing patients' perceptions and choices regarding their treatment options (4, 7).

In the broader context of colon surgery, patients and healthcare professionals face a challenging information landscape, with the internet serving as a pivotal resource for understanding health options and staying updated on medical advancements (8, 9, 10, 11). However, despite the benefits of widespread access, a significant portion of online medical information remains misleading, outdated, or outright false, complicating informed decision-making and posing risks to patient safety and clinical outcomes (12). This challenge is particularly critical in colon surgeries, where procedural complexity and patient variability demand precise and accurate medical knowledge (13). Patients exposed to flawed information may develop misconceptions about their treatment options, leading to poor choices or unrealistic expectations, while healthcare providers must navigate vast quantities of online content to identify trustworthy and up-to-date resources.

Given these challenges, there is a need for tools capable of evaluating available information with both speed and accuracy. Large Language Models (LLMs), such as those based on the Generative Pre-trained Transformer (GPT) architecture, offer a promising solution (14, 15). These advanced neural networks excel in analyzing and interpreting large sets of unstructured text (16). By leveraging their extensive training on diverse datasets, LLMs can differentiate between reliable and dubious information with a level of precision that traditional methods, hindered by rigid rule-based systems and limited adaptability, cannot match (17).

This research endeavors to harness the potential of LLMs to assess online medical information and its alignment with expert evaluations. The overarching research question probes whether LLMs, with their deep understanding of language and context, can reach an expert level in evaluating the quality of medical texts aimed at patients.
This study compares the classification performance of various transformer-based LLMs in evaluating online medical texts against the Ensuring Quality Information for Patients (EQIP) 36 criteria (18). Specifically, it seeks to determine how well models such as BERT (the base variant as well as a custom-pretrained variant), domain-adapted BiomedBERT, LLama-2, SBERT, and Longformer can classify website content based on the EQIP framework (19, 20, 21, 22, 23).
By leveraging a document chunking approach for models with token limitations and concatenating the chunk embeddings, this investigation will assess the accuracy and F1-score of these models in capturing expert-level discernment in quality assessment. The goal is to develop a robust methodology for automated evaluation of health-related web content, thereby enhancing the reliability and accessibility of high-quality medical information through AI-driven analysis.

2. Methods

This study employed a quantitative research design to assess the quality of online medical information using LLMs. Transformer-based models, including BERT, BiomedBERT, LLama-2, SBERT, and Longformer, were evaluated as multilabel classifiers to determine their effectiveness in assessing medical website content based on the EQIP 36 criteria. Furthermore, the study adhered to Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines (24).

The dataset focused on online information regarding robotic colon surgery. Website content was assessed using a modified Ensuring Quality Information for Patients (EQIP) scale that evaluated both content quality and structural integrity as rated by medical experts.

A web scraping algorithm using Python libraries Beautiful Soup (Version 4.10) and Selenium (Version 4.0) was executed in December 2023 (25, 26). Beautiful Soup facilitated parsing HTML and XML documents, while Selenium enabled automated browsing of interactive web content.
Data were collected across major search engines—Google, Bing, and Yahoo—using specific search phrases such as 'Da Vinci Colon-Rectal Surgery,' 'Colon Robotic Surgery,' and 'Robotic Bowel Surgery.' The algorithm retrieved the first 150 unique URLs per search engine, based on the assumption that users rarely consult results beyond this range. After initial retrieval of 450 English-language web pages, 163 duplicates were identified and removed via a customized Python script. Bilingual experts then excluded 39 irrelevant pages and 41 academic or professional sites, focusing on retaining websites with coherent, patient-relevant content.

Following manual verification by two investigators, the dataset included 207 websites, which were further refined based on structural consistency, yielding 159 suitable text samples. Only records with a single set of EQIP labels were retained to ensure unique labeling per document, resulting in a final dataset comprising 47 unique websites. The final flow of dataset construction is depicted in Figure 1, and the EQIP questionnaire items are detailed in Table 1.

#	EQIP Item	Positive (count)	Positive (percentage)
1	Initial definition of subjects will be covered	47	87.0
3	Description of the medical problem	46	85.2
4	Definition of the purpose of the surgical intervention	42	77.8
5	Description of treatment alternatives	40	74.1
6	Description of the sequence of the surgical procedure	32	59.3
7	Description of the qualitative benefits to the recipient	22	40.7
8	Description of the quantitative benefits to the recipient	6	11.1
9	Description of the qualitative risks and side effects	15	27.8
10	Description of the quantitative risks and side effects	0	0.00
11	Addressing quality-of-life issues	16	29.6
12	Description of how complications are handled	10	18.5
13	Description of the precautions that the patient may take	37	68.5
14	Mention of alert signs that the patient may take	43	79.6
15	Addressing medical intervention costs and insurance issues	11	20.4
17	Specific details of other sources of reliable information	35	64.8
18	Coverage of all relevant issues for the topic	12	22.2
30	Clear information (no ambiguities or contradictions	52	96.3
31	Balanced information on risks and benefits	6	11.1
32	Presentation of information in a logical order	42	77.8

Table 1. EQIP Questionnaire Items

Nineteen quality indicators from the EQIP scale were assessed as binary outcomes (present or absent).

Due to the input length limitations of standard transformer architectures, a document chunking strategy was implemented. Website texts were split into overlapping segments, each with a maximum length of 512 tokens (aligned with typical model constraints). To maintain semantic coherence, a 50-token overlap was applied between consecutive chunks. Tokenization was performed using the default tokenizer of each respective LLM, and chunk boundaries were aligned with sentence endings where possible to minimize context disruption. Each chunk was converted into an embedding via the respective LLM. Subsequently, all chunk embeddings belonging to the same document were concatenated to form a comprehensive document-level embedding representation, enabling effective downstream

multilabel classification. To expand the dataset and evaluate potential performance gains, data augmentation techniques were applied. A paragraph-level dataset was constructed by deconstructing websites into discrete paragraphs and applying transformations before recombining them into coherent documents. The augmentation methods included backtranslation (27), summarization (28), and paraphrasing (29), aiming to preserve content integrity while broadening the training corpus. Through this process, the dataset expanded from 47 to 79 records.

Rather than fine-tuning a single generative model (e.g., GPT-4), this study focused on evaluating document-level embeddings derived from chunked website texts. These embeddings, obtained from BERT, BiomedBERT, LLama-2, SBERT, and Longformer, were concatenated and fed into a lightweight feedforward neural network for multilabel classification, predicting the 19 EQIP items.

It is important to note that the classification model primarily served as a probe: the core evaluation centered on the quality and consistency of the embeddings generated by each LLM. To ensure fair assessment, classifier hyperparameters—hidden layer size, learning rate, and dropout rate—were optimized through a grid search, with settings detailed in Table 2. Stratified 5-fold cross-validation was employed throughout the training process to maintain robust model evaluation.

Model	Batch Size	Learning Rate	Epochs	Weight Decay	Masked Token Probability
Base BERT	8	1e-5	2000	0.001	NA
PreTrained BERT – Pretraining	32	2e-5	3	0.01	0.15
PreTrained BERT – Classification	8	1e-5	2000	0.001	NA
BioMed BERT	8	1e-5	2000	0.001	NA
PreTrained BioMed BERT – Pretraining	32	2e-5	3	0.01	0.15
PreTrained BioMed BERT – Classification		1e-5	2000	0.001	NA
Llama2 (Concatenated)	8	1e-5	2000	0.001	NA
Llama2 (Averaged)	8	1e-5	2000	0.001	NA
S-BERT	8	1e-5	2000	0.001	NA
Longformer	2	2e-5	50	0.01	NA

Table 2. Hyperparameter Grid Search

Performance metrics included accuracy, precision, recall, and F1-score, with particular emphasis on F1-score to account for potential label imbalance.
An experimental setup was established where each LLM's embeddings were tested via the classifier, comparing performance against expert-rated EQIP labels. In contrast to approaches solely relying on generative inference, this method directly optimized downstream classifiers using pre-trained embeddings for better generalizability (30). A dataset split of 70% training, 10% validation, and 20% test was employed. Hyperparameter tuning was conducted using the validation set, and final evaluation was performed on the unseen test set to ensure unbiased performance reporting. While recent universal embedding models such as nomic-embed-text or OpenAI's text-embedding-ada-002 offer strong general-purpose representations, our comparison focused on benchmarking models with varying domain specialization and architectural types (31,32). Future studies could explore how these universal embeddings perform on EQIP-based medical information evaluation tasks.

3. Results

A comprehensive analysis was conducted to evaluate the ability of multiple transformer-based models to generate meaningful semantic embeddings to assess the quality of online medical information against the EQIP criteria. Embeddings extracted via chunked document processing were used for multilabel classification with a feedforward neural network.

The experimental results revealed substantial differences in model performance. The SBERT model outperformed all

others, achieving the highest F1-score of 80.16 and accuracy of 79.05%, demonstrating its strength in meaningfully converting text into representative embeddings. This aligns with its architecture, which is designed for a training goal of converting text into its representational embeddings while retaining its semantic meaning (similar texts have similar embeddings and vice versa, unlike other LLMs, which are trained using either MaskedLanguageModelling or CausalLanguageModelling approaches. Among the BERT-based models, the Pre-trained BiomedBERT model demonstrated comparative performance to our custom Pre-trained BERT, particularly in capturing domain-specific information. The fine-tuned BiomedBERT model showed an increase in the F1-score by 0.82%, reflecting its ability to better identify quality indicators in medical texts.

For the LLaMA-based approaches, the concatenated embedding method showed a notable performance (of an F1 score of 0.784) improvement over the averaged embedding approach (F1 score of 0.765), suggesting that retaining a richer representation of segmented text provides a more informative document-level embedding. However, both versions underperformed in comparison to SBert, custom-pretrained BERT and BiomedBERT, likely due to the lack of domain-specific pre-training and challenges in creating meaningful and consistent embeddings effectively. The Longformer model, despite its ability to process longer pieces of text, exhibited lower recall and precision, struggling in comparison to the other models. This resulted in an F1-score of 70.07, which was surpassed by models better suited for handling medical texts.

Hyperparameter optimization through grid search focused on encoder learning rates, batch sizes, and dropout rates, aiming to balance model generalization and overfitting prevention. The final configurations optimized for each model resulted in performance improvements, particularly for custom pretrained BERT and BiomedBERT, where increased domain-specific adaptability and hidden layer sizes proved beneficial respectively.

The best-performing models showcased a balance between precision and recall, with the SBERT model achieving the highest average precision across 5 runs of 77.4, signifying its ability to identify true positives while minimizing false positives effectively. Results are shown in Table 3. The corresponding ROC-AUC figure can be seen in Figure 2 (Fig.2).
The classification results of the best performing model are further summarized through normalized confusion matrices in Figure 3, illustrating the extent to which each model captured positive and negative instances across the EQIP items (Fig. 3). These findings indicate that domain adaptation and meaningful embeddings of the texts are critical factors in improving the automated evaluation of medical information quality.

Model	Accuracy	F1 Score	AUC-ROC
Base BERT	0.77	0.78	0.77
PreTrained BERT – Pretraining	0.78	0.79	0.78
PreTrained BERT – Classification	0.78	0.79	0.78
BioMed BERT	0.78	0.79	0.78
PreTrained BioMed BERT – Pretraining	0.79	0.8	0.79
PreTrained BioMed BERT – Classification	0.78	0.79	0.78
Llama2 (Concatenated)	0.77	0.78	0.77
Llama2 (Averaged)	0.76	0.76	0.77
S-BERT	0.79	0.80	0.79
Longformer	0.68	0.70	0.68

Table 3. Metrics for the multilabel classifiers. The accuracy, f1-score and the AUC-ROF are shown on the test set

4. Discussion

The evaluation of various transformer-based models for assessing the quality of online medical information against the EQIP criteria has demonstrated notable differences in performance. Unlike previous studies that relied on a single model like GPT-4, this study systematically compared multiple models, including BERT, Pre-trained BERT, BiomedBERT, Pre-trained BiomedBERT, LLaMA concatenated, LLaMA averaged, SBERT, and Longformer, to identify the most effective

approach for this classification task.

Among the tested models, the SBERT exhibited the highest accuracy (79.05%) and F1-score (80.16), significantly outperforming the others in creating meaningful embeddings from medical texts. As shown in our results, models optimized for generating coherent and semantically meaningful embeddings, such as SBERT, consistently outperformed larger generative models (33). Additionally, Pre-trained BERT outperformed standard BERT-based model and matched the performance of BioMedBERT, reinforcing the importance of domain-specific adaptation in medical text analysis (34). The LLaMA-based approaches showed mixed results, with the concatenated embedding method achieving higher classification performance compared to the averaged approach. However, both struggled in capturing semantic representational embeddings as effectively as SBERT or BiomedBERT. The Longformer model, despite its longer context window, demonstrated lower recall, indicating challenges in maintaining creating meaningful embeddings (35).
Overall, LLMs like SBERT, which were explicitly trained to generate coherent and logical semantic embeddings, outperformed larger models like Llama2 (which is trained with the objective of text generation) since the latter fails at consistently converting text into meaningful embeddings. This highlights how the training objective plays a large role in deciding an LLM's performance in a downstream task as opposed to its massive size and/or complexity (36).

Hyperparameter tuning, including grid search for batch sizes, dropout rates, and learning rates, played a crucial role in optimizing classifier performance. While models trained with pre-trained domain-specific encoders exhibited improved F1 scores, class imbalance remained challenging for EQIP items with sparse positive instances.

Given the evolving nature of LLM-based medical text assessment, these findings must be interpreted with caution. The inherent token limitations of LLMs necessitate document chunking, as described previously, which may introduce slight noise but is currently unavoidable. The inconsistency in grading partially present items indicates that transformers require more sophisticated reasoning capabilities to accurately assess nuanced medical quality indicators (37).

Several limitations were observed, including the challenge of class imbalance for some labels, the complexity of multi-label classification, and computational constraints associated with processing long texts across multiple models. Although Llama2 and BiomedBERT demonstrated strong classification capabilities, these models' high computational cost and memory requirements highlight the need for more efficient architectures tailored for medical NLP tasks (38). Additionally, newer domain-specific models such as MedLLaMA3 could not be included due to limited availability of embedding APIs or fine-tuning interfaces at the time of analysis. Future research should evaluate the performance of such specialized models once they become more accessible for downstream applications

Despite these challenges, this study lays the foundation for future research in leveraging transformer-based models for medical information assessment.
The results emphasize the necessity for open-source, fine-tuned models that align with regulatory requirements and ethical considerations in healthcare AI. As LLMs continue to evolve, it is imperative to integrate them into medical quality assessment frameworks while ensuring their alignment with human expert evaluations. Future work should explore larger, more balanced datasets, improved domain-specific fine-tuning strategies, and ensemble approaches to further enhance the reliability of AI-driven medical information evaluation.

Declarations

Ethics approval and consent to participate: Not Applicable.

Consent for publication: All authors have read and agreed to the published version of the manuscript.

Data Availability: The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest: Non declared.

Funding Statement: This research received no external funding.

Authors Contributions: Conceptualization A. Taha, V. Ochs; data collection M.D. Honaker; analysis S. Muheim, V. Ochs; writing—original draft preparation V. Ochs; writing—review and editing J. Wolleb, F. Bieder, A. Taha, M.D. Honaker, P.C. Cattin, S. Taha-Mehlitz, M.Flaifel, A.J. Rodríguez-Sánchez, D. Öfner, F. Ponholzer, A. Kumar; R. Rosenberg, Supervisor: A. Taha, P.C. Cattin

Acknowledgements: Not applicable.

Clinical trial number: Not applicable.

Abbreviations:
Large Language Models - (LLMs)
The Ensuring Quality Information for Patients - (EQIP)
Strengthening the Reporting of Observational Studies in Epidemiology - (STROBE)

References

[1] Fox, S. and Duggan, M. (2013) Health Online 2013. Pew Internet & American Life Project. https://www.pewinternet.org/2013/01/15/health-online-2013

[2] Eysenbach G. What is e-health? J Med Internet Res. 2001 Apr-Jun;3(2):E20. doi: 10.2196/jmir.3.2.e20. PMID: 11720962; PMCID: PMC1761894.

[3] Lanfranco AR, Castellanos AE, Desai JP, Meyers WC. Robotic surgery: a current perspective. Ann Surg. 2004 Jan;239(1):14-21. doi: 10.1097/01.sla.0000103020.19595.7d. PMID: 14685095; PMCID: PMC1356187.

[4] Taha, A. et al. Robotic colorectal surgery: quality assessment of patient information available on the internet using webscraping. Comput. Assist. Surg. 28, 2187275 (2023).

[5] Ben-Or S, Nifong LW, Chitwood WR Jr. Robotic surgical training. Cancer J. 2013 Mar-Apr;19(2):120-3. doi: 10.1097/PPO.0b013e3182894887. PMID: 23528718.

[6] Fröjd C, Swenne CL, Rubertsson C, Gunningberg L, Wadensten B. Patient information and participation still in need of improvement: evaluation of patients' perceptions of quality of care. J Nurs Manag. 2011 Mar;19(2):226-36. doi: 10.1111/j.1365-2834.2010.01197.x. Epub 2011 Feb 1. PMID: 21375626.

[7] Nayyar, G. (2023). Dead Wrong: Diagnosing and Treating Healthcare's Misinformation Illness. John Wiley & Sons.

[8] Beauharnais CC, Aulet T, Rodriguez-Silva J, Konen J, Sturrock PR, Alavi K, Maykel JA, Davids JS. Assessing the Quality of Online Health Information and Trend Data for Colorectal Malignancies. J Surg Res. 2023 Mar;283:923-928. doi: 10.1016/j.jss.2022.10.055. Epub 2022 Dec 8. PMID: 36915020.

[9] Hertling S, Matziolis G, Graul I. Die Rolle des Internets als medizinische Informationsquelle für orthopädische Patienten [The role of the Internet as a source of medical information for orthopedic patients]. Orthopadie (Heidelb). 2022 Jul;51(7):521-530. German. doi: 10.1007/s00132-022-04238-5. Epub 2022 Mar 29. PMID: 35352139; PMCID: PMC8963403.

[10] Ihler F, Canis M. The Role of the Internet for Healthcare Information in Otorhinolaryngology. Laryngorhinootologie. 2019 Mar;98(S 01):S290-S333. English, German. doi: 10.1055/a-0801-2585. Epub 2019 Apr 3. PMID: 31096302.

[11] Bujnowska-Fedak MM, Waligóra J, Mastalerz-Migas A. The Internet as a Source of Health Information and Services. Adv Exp Med Biol. 2019;1211:1-16. doi: 10.1007/5584_2019_396. PMID: 31273574.

[12] Suarez-Lledo V, Alvarez-Galvez J. Prevalence of Health Misinformation on Social Media: Systematic Review. J Med Internet Res. 2021 Jan 20;23(1):e17187. doi: 10.2196/17187. PMID: 33470931; PMCID: PMC7857950.

[13] Loftus TJ, Tighe PJ, Filiberto AC, Efron PA, Brakenridge SC, Mohr AM, Rashidi P, Upchurch GR Jr, Bihorac A. Artificial Intelligence and Surgical Decision-making. JAMA Surg. 2020 Feb 1;155(2):148-158. doi: 10.1001/jamasurg.2019.4917. PMID: 31825465; PMCID: PMC7286802.

[14] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[15] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. Adv Neural Inf Process Syst. 2022; 35: 27730-27744

[16] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

[17] McInerney C, Benn J, Dowding D, Habli I, Jenkins DA, McCrorie C, Peek N, Randell R, Williams R, Johnson OA. Patient Safety Informatics: Meeting the Challenges of Emerging Digital Health. Stud Health Technol Inform. 2022 Jun 6;290:364-368. doi: 10.3233/SHTI220097. PMID: 35673036.

[18] Charvet-Berard AI, Chopard P, Perneger TV. Measuring quality of patient information documents with an expanded EQIP scale. Patient Educ Couns. 2008;70(3): 407-411.

[19] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

[20] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained

biomedical language representation model for biomedical text mining. *Bioinformatics*, *36(4)*, *1234–1240*

[21] Touvron, H., Martin, J., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). LLaMA 2: Open Foundation and Fine-Tuned Chat Models. *Meta AI Research*.

[22] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[23] Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The Long-Document Transformer. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*

[24] von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet. 2007;370(9596):1453-1457. doi: 10.7554/elife.08500.009.

[25] Richardson L. Beautiful soup documentation. April. 2007;.

[26] Salunke SS.. Selenium webdriver in Python: learn with examples. (1st. ed.). North Charleston, SC, USA: CreateSpace Independent Publishing Platform;2014.

[27] Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

[28] Sam Witteveen and Martin Andrews. 2019. Paraphrasing with Large Language Models. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 215–220, Hong Kong. Association for Computational Linguistics.

[29] Jason Wei and Kai Zou. 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

[27] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). *On the opportunities and risks of foundation models*. arXiv preprint arXiv:2108.07258

[28] Gao, T., Yao, X., & Chen, D. (2021). *SimCSE: Simple Contrastive Learning of Sentence Embeddings*. arXiv preprint arXiv:2104.08821.

[29] Peng, Y., Yan, S., & Lu, Z. (2019). *Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets*. arXiv preprint arXiv:1906.05474.

[30] Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontañón, S., Pham, P., Ravula, A., Wang, Q., Yang, L., & Ahmed, A. (2020). *Big Bird: Transformers for longer sequences*. arXiv preprint arXiv:2007.14062

[31] Nussbaum, Z., Morris, J.X., Duderstadt, B., & Mulyar, A. (2024). Nomic Embed: Training a Reproducible Long Context Embedder. ArXiv abs/2402.01613.

[32] OpenAI. (2022, December 15). *New and improved embedding model*. OpenAI. Retrieved August 7, 2025, from https://openai.com/index/new-and-improved-embedding-model

[33] Bommasani, R., Hudson, D., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., & others. (2021). *On the Opportunities and Risks of Foundation Models*. arXiv preprint arXiv:2108.07258.

[34] Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023). *Capabilities of GPT-4 on Medical Challenge Problems*. arXiv preprint arXiv:2303.13375.

[35] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., ... & Naumann, T. (2021). *Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing*. ACM Transactions on Computing for Healthcare, 3(1), 1-23.

[36] Turc, I., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *Well-read students learn better: On the importance of pre-training compact models*. arXiv. https://arxiv.org/abs/1908.08962

[37] Roberts, M., Driggs, D., Thorpe, M., et. al, (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence*, *3*(3), 199–217. https://doi.org/10.1038/s42256-021-00307-0

[38] Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. N Engl J Med. 2019 Apr 4;380(14):1347-1358. doi: 10.1056/NEJMra1814259. PMID: 30943338.

Figure Titles and Legends

Supplementary Files

Figures

Flow chart of sample size.

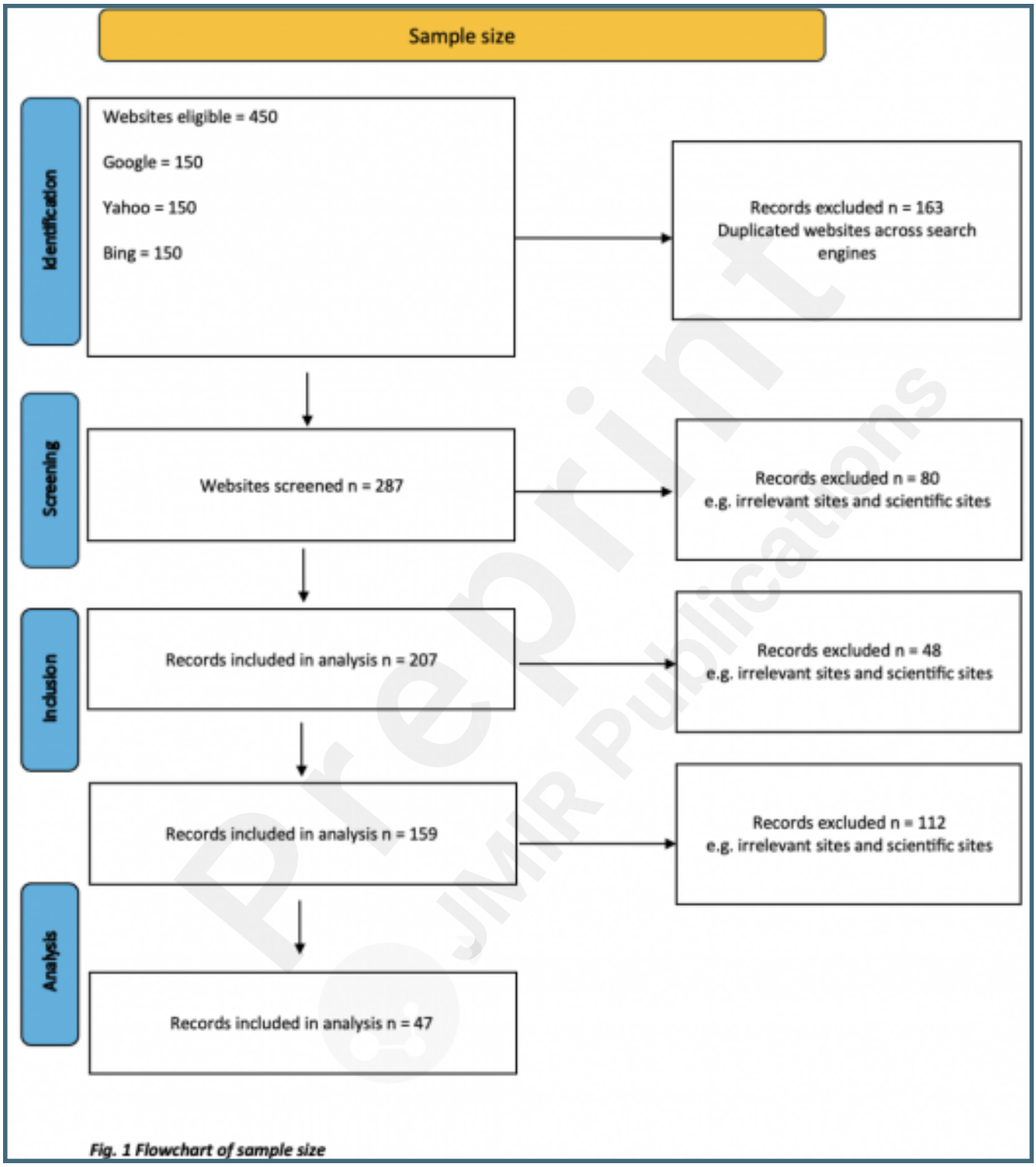

Fig. 1 Flowchart of sample size

ROC-AUC plot for the multilabel classifier.

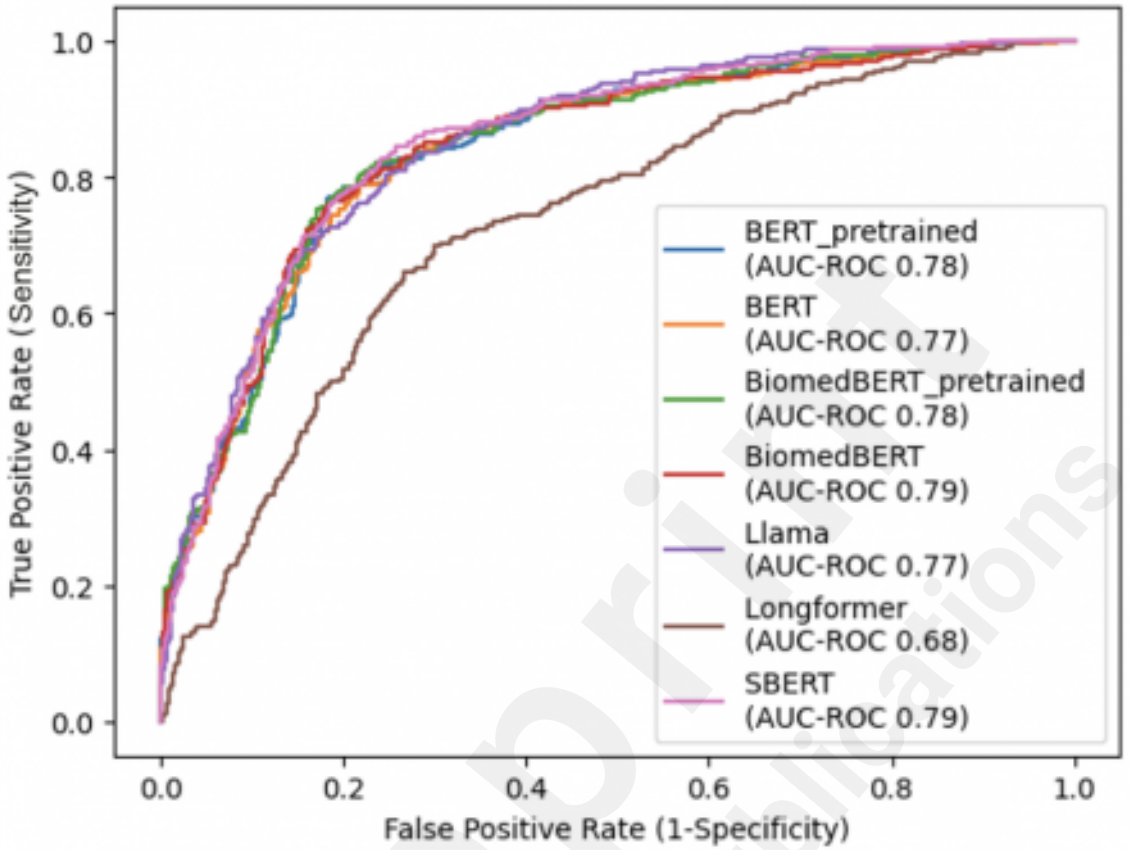

Fig. 2 ROC-AUC plot

Normalized sum of the confusion matrices for the multilabel classifier.

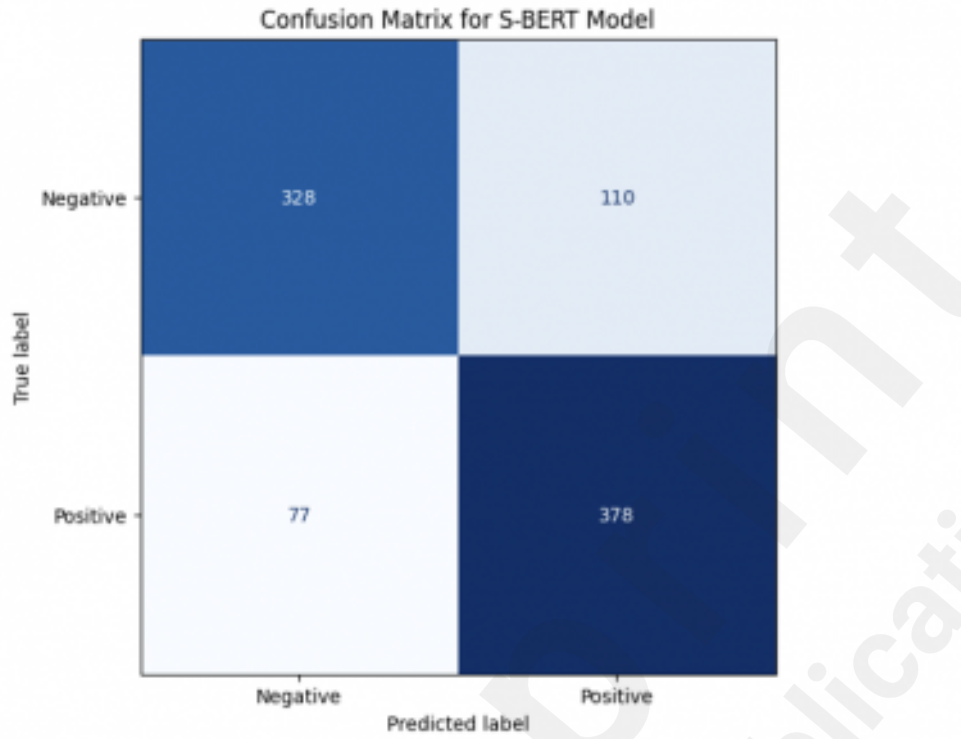

Fig. 3 Normalized sum of the confusion matrices for the multilabel classifier of SBert