Clinfo.ai: An Open-Source Retrieval-Augmented Large Language Model System for Answering Medical Questions using Scientific Literature

Alejandro Lozano[1,*,†], Scott L. Fleming[1,*], Chia-Chun Chiang[2,3], Nigam Shah[4,5,6]
[1] *Department of Biomedical Data Science, Stanford University, Stanford, CA, USA*
[2] *Department of Neurology, Mayo Clinic, Rochester, MN*
[3] *Human Centered Artificial-Intelligence Institute, Stanford University, Stanford, CA, USA*
[4] *Department of Medicine, Stanford School of Medicine, Stanford, CA, USA*
[5] *Clinical Excellence Research Center, Stanford School of Medicine, Stanford, CA, USA*
[6] *Technology and Digital Solutions, Stanford Health Care, Palo Alto, California, USA*
[*] *Equal Contribution*
[†] *E-mail: lozanoe@stanford.edu*

The quickly-expanding nature of published medical literature makes it challenging for clinicians and researchers to keep up with and summarize recent, relevant findings in a timely manner. While several closed-source summarization tools based on large language models (LLMs) now exist, rigorous and systematic evaluations of their outputs are lacking. Furthermore, there is a paucity of high-quality datasets and appropriate benchmark tasks with which to evaluate these tools. We address these issues with four contributions: we release Clinfo.ai, an open-source WebApp that answers clinical questions based on dynamically retrieved scientific literature; we specify an information retrieval and abstractive summarization task to evaluate the performance of such retrieval-augmented LLM systems; we release a dataset of 200 questions and corresponding answers derived from published systematic reviews, which we name PubMed Retrieval and Synthesis (PubMedRS-200); and report benchmark results for Clinfo.ai and other publicly available OpenQA systems on PubMedRS-200.

Keywords: Large Language Models, Abstractive Summarization, Artificial Intelligence, Clinical Medicine, Generative AI, Interactive Systems, ChatGPT

1. Introduction

The aggregation and distribution of medical knowledge, facilitated by platforms such as PubMed or Cochrane, enables healthcare professionals and medical researchers to stay abreast of the latest scientific discoveries and make informed decisions based on up-to-date scientific evidence.[1] However, the staggering influx of more than 1 million papers each year into PubMed alone (equivalent to two papers per minute as of 2016)[2] highlights the daunting task of keeping up with scientific findings.[3] This is especially true for practicing clinicians, who face the challenge of keeping track of the most updated research findings in all areas related to their patient care duties.[4]

Existing technologies fail to adequately satisfy the information needs of health care profes-

sionals and researchers. In daily practice, clinicians have on average one care-related question for every other patient seen[5] and they refer to sources like PubMed or UpToDate to obtain summarized information answering these questions.[6] Questions that cannot be answered within 2 to 3 minutes are often abandoned, potentially negatively impacting patient care and outcomes.[5,7] While systematic review (SR) articles can provide quick answers to clinical questions, many questions are not answerable through existing reviews. On the other hand, manually synthesizing findings from multiple primary sources without the help of a published review article can be extraordinarily time consuming. Review articles take on average 67.3 weeks to complete,[8] and those written reviews may not even include the most updated research published in the literature. Question-answering tools that leverage frequently updated external electronic resources would enable researchers and clinicians to obtain up-to-date information in a more efficient way that benefits scientific discovery and quality of patient care.[9–13]

In previous decades, applications that integrated clinical systems with on-line information to answer users' information needs (e.g., "infobuttons")[14] were typically driven by semantic networks. Other works such as CHiQA proposed a combination of knowledge-based, machine learning, and deep learning approaches to develop a question-answering system using patient-oriented resources to answer consumer health questions.[15]

The new capabilities of agents powered by large language models (LLM) has accelerated the development of automated literature summarization tools. Most of these solutions tend to be privately developed, closed-source solutions based on retrieval-augmented[16] (RetA) LLMs[17] (e.g. Scite,[18] Elicit,[19] GlacierMD,[20] Consensus,[21] OpenEvidence,[22] Statpearls semantic search[23]). However, the paucity of publicly available technical reports describing these systems and the lack of appropriate guidelines, regulations, and evaluations to ensure their safe and responsible usage is an urgent concern.[24]

This Natural Language Generation (NLG) problem has been exacerbated by a lack of (1) representative datasets and associated tasks, and (2) automated metrics for evaluating RetA LLMs on said tasks.

Fortunately, developments in the LLM evaluation space have shown that a number of automated metrics correlate moderately with human preference, even in domain-specific scenarios (including medicine).[25–27]

Building on these advancements, we provide four contributions:

(1) Clinfo.ai [a], the first publicly available, open-source, end-to-end retrieval-augmented LLM-based system for querying and synthesizing the clinical literature. The system is hosted as a publicly available WebApp at https://www.clinfo.ai/.
(2) An open information retrieval and abstractive summarization task specification designed to evaluate an algorithm's ability to both retrieve relevant information and adequately synthesize it. In the task setup, both the information retrieval and abstractive summarization sub-tasks are compared to gold standard (human generated but pragmatically retrieved)

[a]https://github.com/som-shahlab/Clinfo.AI

references and answers. Furthermore, our task is defined to truly resemble RetA deployment conditions (enabling the evaluation of already deployed but potentially closed-source systems).

(3) PubMed Retrieval and Synthesis (PubMedRS-200), a publicly available dataset of 200 questions structured in Open QA format, paired with answers derived from systematic reviews and corresponding references.

(4) Benchmark results for Clinfo.ai and other publicly available OpenQA systems on PubMedRS-200).

2. Related Work

LLMs in healthcare The remarkable performance of LLMs in the general domain has brought about a revolution in the field of natural language processing,[28] showcasing exceptional capabilities in tasks like summarization, question-answering, and NLG.[29] Given their wide utility, researchers are now actively exploring applications of LLMs in healthcare.[30–33] Several LLMs have achieved human-level performance on numerous medical professional licensing exams such as the United States Medical Licensing Exam (USMLE).[34] Other works have demonstrated promise in various healthcare-inspired tasks, such as automated clinical note generation and reasoning about public health topics.[30–33] However, NLG tasks and publicly available benchmarks that directly address true medical needs are still underrepresented in the literature. Such tasks and benchmarks are especially important for estimating the capabilities and risks of LLMs in the clinical domain.

LLMs have several documented disadvantages and risks. First, updating LLMs with new knowledge and information is challenging and inefficient.[35] Second, the training objective of LLMs to predict the most probable next token can cause these models to generate inaccurate information (hallucination), requiring costly and imperfect post-hoc model adjustments like reinforcement learning with human feedback (RLHF).[36] More importantly, most popular consumer-facing LLMs (e.g., OpenAI's GPT-4,[29] Meta's Llama 2,[37] Anthropic's Claude 2[38]) do not provide references pointing to their source of information, even when the model's output is factual. This can engender distrust with users in many scientific domains, including healthcare. Prior work has proposed ReTA LLMs[16] to solve the information provenance issue and have shown promising results. These ReTA LLMs do not require post-hoc model editing in order to incorporate new knowledge.

Retrieval Augmentation Question Answering LLMs in Medicine Hiesinger et al.[39] introduced Almanac, a novel LLM integrated with a vector database and calculator, designed to answer 130 clinical questions generated by a panel of five board-certified clinicians and resident physicians. The results showed that Almanac surpassed a standard LLM (GPT-4) in factuality, safety, and correctness, indicating that retrieval systems lead to more accurate and reliable responses to clinical inquiries. Soong et al.[40] evaluated GPT-3.5 and GPT-4 models against a custom RetA LLM using a set of 19 questions. The evaluation, based solely on human judgments, revealed that both GPT-3.5 and GPT-4 exhibited more hallucinations in all 19 responses compared to the RetA model. While these works on RetA LLM systems represent significant progress, they suffer from at least two shortcomings: (1) they typically

require human evaluation, making systematic benchmarking of new systems challenging and unscaleable; (2) they often focus solely on evaluating an LLM's output, disregarding the relevance of the information retrieved to generate an answer. Deciding which "relevant" sources should be summarized can be just as challenging as generating the actual summary. Hence there is a need for a benchmark that enables integrated evaluation of both a system's ability to select relevant documents as well as its ability to summarize these documents.

3. Materials and Methods

3.1. *Dataset Generation*

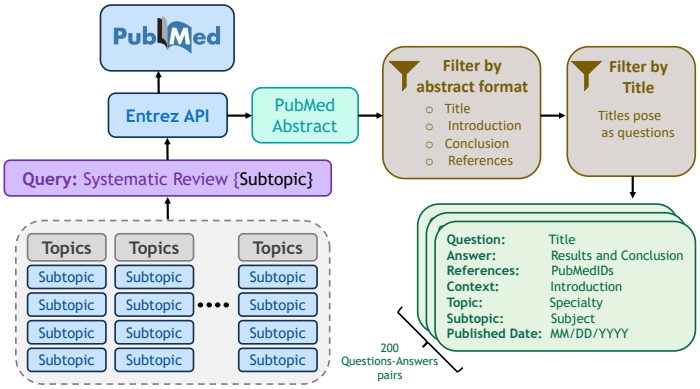

Fig. 1: Schematic Representation of the Protocol for Retrieving Abstracts from PubMed and Generating Title-Based Questions

PubMed is a free resource supporting search and retrieval of biomedical literature. As prior work has demonstrated, a large quantity of research papers available in this index are phrased as questions, and it is possible to structure them in a question-answer format.[41,42] Extending this idea, we created an open information retrieval and abstractive summarization dataset, using SR as a proxy for inquiries of medical interest. The rationale is that SRs are structured reviews written by human experts which summarize the pertinent literature related to a question of interest in an evidence-based manner.[43] In writing a SR, experienced authors (1) screen the published literature in a systematic way and include studies in a standardized manner; (2) critically evaluate methodology and reported outcomes of the included studies; and (3) carefully extract data, summarize original research findings, and in some instances, conduct additional statistical analysis of extracted results from studies including randomized controlled trials, observational cohort studies, case series and other qualitative studies on a specific topic. Furthermore, SRs are extensively used to provide evidence for various purposes, including policy-making, clinical practice guidelines, health technology assessment, and decision making in healthcare.[44] As SRs unify and present a comprehensive overview of a given subject by human experts, we chose to leverage published SRs as gold standards when building our database.

To populate such a dataset, we employed E-utilities, a public API to the NCBI Entrez system[45] , to access PubMed and construct question-answer pairs with their respective references. Figure 1 illustrates our process in detail. First, we established a comprehensive selection of medical specialties and subspecialties. Second, we formulated a query to retrieve Systematic Reviews relevant to each medical specialty/subspecialty. Upon constructing the specialty-specific queries and retrieving associated abstracts, we retrieved all papers structured in a format that can be easily converted to questions-answer pairs (as noted by Jin et al 2019[41]) namely Title, Introduction, Conclusion, and References. Third, we applied another filtering process, narrowing down to solely those publications whose titles included an explicit question (i.e., publications whose titles including question marks). The questions from these titles were extracted.

Finally, two human evaluators (AL and SF) manually reviewed the retrieved questions and extracted an answer to each question using minimally modified text from the results and conclusions section of the corresponding SR abstract. Concretely, in order to generate each answer, the human reviewers removed from the Results and Conclusions section of the abstract any text describing the structure or design of the systematic review (e.g., "We used PubMed to retrieve 100 papers"), leaving only text that directly addressed the question extracted from the SR's title. In the process, abstracts that were lacking substantive results and abstracts that merely described research proposals (e.g. descriptions of future work) were entirely removed.

3.2. *Clinfo.ai: An LLM Chain for Information Retrieval and Synthesis*

Our proposed RetA LLM system, Clinfo.ai, consists of a collection of four LLMs working conjointly (an LLM chain[46]) coupled to a Search Index (either PubMed or Semantic Scholar) as depicted in Figure 2. Previous works have observed that very large language models (e.g., 100B parameters or more) exhibit zero-shot reasoning capabilities, where task-specification prompts can be used to guide the LLM output without further fine-tuning.[47,48] We leverage the zero-shot reasoning capabilities of two LLMs, specifically OpenAI's GPT-3.5 and GPT-4 models, to complete each step in the LLM chain depicted in Figure 2. All prompts used in each step of the chain are available in the supplemental material [b]. We use LangChain's API to send prompts and receive outputs from GPT-3.5 and GPT-4. While different models could technically be used through this entry point, our experiments are limited to OpenAI's GPT-3.5 and GPT-4 models (snapshots gpt-3.5-turbo-0613 , gpt-4-0613 respectively). For both models, we employ a temperature of 0.5 and a max token generator limit of 1024.

3.2.1. *Query Generator*

In our Clinfo.ai system, the input is the question submitted by the user. Once a question is submitted, the primary task of the query generator (labeled "Question2Query" in Figure 2) is to construct a PubMed (or Semantic Scholar) query that efficiently retrieves a substantial number of relevant articles pertaining to the posed question. This is achieved by instructing

[b]https://github.com/som-shahlab/Clinfo.AI/tree/main/SupplementalMaterial

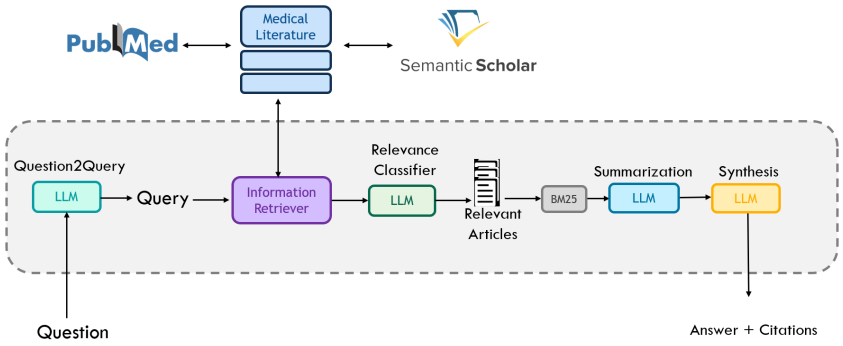

Fig. 2: Clinfo.ai: A RetA LLM system for retrieving and summarizing scientific articles

the model to incorporate the most crucial and relevant keywords that accurately represent the query's context and requirements.

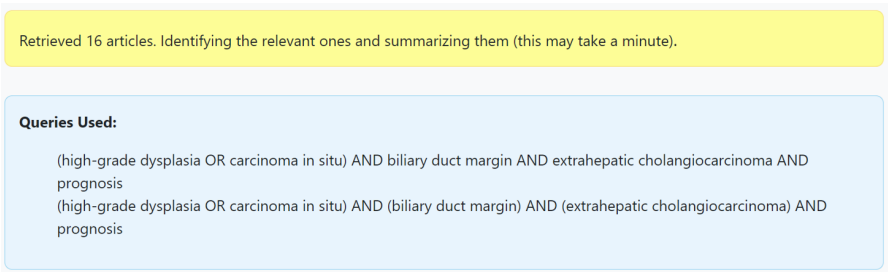

Fig. 3: Query Generated by Clinfo.ai for question: "Does high-grade dysplasia/carcinoma in situ of the biliary duct margin affect the prognosis of extrahepatic cholangiocarcinoma?"

3.2.2. *Information Retriever*

In a similar fashion to the Dataset Generation process, we utilize the Entrez API to fetch abstracts from PubMed using the output generated by the Query Generator. By leveraging the Entrez API, we are able to programmatically access and retrieve the relevant abstracts that match the constructed PubMed queries. Because LLM output is stochastic and different queries may capture different aspects of the literature, we take the union of all papers returned by three LLM-generated queries (each with the same prompt but different seeds).

3.2.3. *Relevance Classifier*

Since the query generator emphasizes recall over precision (i.e., it retrieves as many potentially relevant articles as possible), it is crucial to classify the relevancy of the retrieved articles. To achieve this, we adopt an LLM-enabled binary classification approach, wherein each article is categorized as either relevant or not relevant to the posed question using GPT-3.5. Once the relevant articles are identified, we make use of the full abstract metadata of each article to construct their citations in the IEEE format. If more than 35 relevant articles are deemed relevant, the user can decide to re-rank and filter them using BM25.[49]

3.2.4. *Summarization*

The penultimate step in Clinfo.ai uses an LLM to summarize each relevant abstract within the context of the user-submitted question.

3.2.5. *Synthesis*

In the final step of Clinfo.ai, the relevant article summaries are organized as an ordered list, with each number in the list corresponding to a citation. This structured list of article summaries is then fed to a LLM with the task of constructing a concise and informative summary. The LLM is also instructed to utilize only the provided article summaries and no other additional information, relying on the structured list of citations to reference and accurately attribute each finding.

3.3. *www.clinfo.ai: A Clinfo.ai User Interface via Web Application*

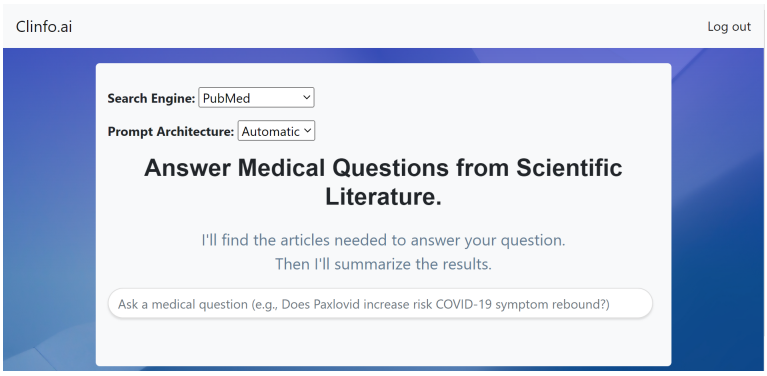

Fig. 4: Clinfo.ai user interface

To facilitate interaction with our system, we developed a web application that allows users to submit their own questions and/or customize the prompts. The latter enables users to tailor the system according to their individual preferences and needs, as illustrated in Figure 4.The entire process provides real-time access, displaying the queries generated during the search (as shown in Figure 3), the number of retrieved articles, a concise summary of each important article, and a final "Literature Summary" (or "Synthesis", to distinguish it from the individual article summaries) accompanied by an abbreviated answer to the question ("TL;DR"). Additionally, the references are presented as hyperlinks, enabling users to verify both the validity of the reference and the information captured from it. It is possible that even after summarizing an article's abstract, Clinfo.ai may not include that article in final Literature Summary or "TL;DR". Nevertheless, we ensure that all relevant articles are presented to the user so that they can access and explore them as needed. An example of a final Literature Review constructed with Clinfo.ai is shown in Figure 5.

3.4. *Task Description and Evaluation*

The task is defined in a three step manner:

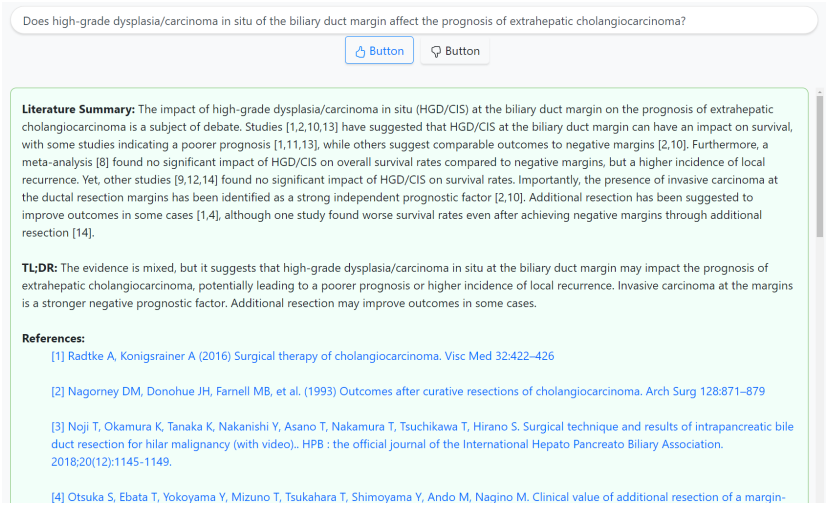

Fig. 5: "Literature Summary" (Synthesis) and "TL;DR" constructed with Clinfo.ai for the question, "Does high-grade dysplasia/carcinoma in situ of the biliary duct margin affect the prognosis of extrahepatic cholangiocarcinoma?" (not all references are included in figure)

(1) Given a question, generate a query to retrieve a set of articles;
(2) Given the provided articles, determine their relevancy to the question;
(3) Given relevant articles, summarize the findings.

Step (2) is evaluated based on precision and recall. Considering the set of all documents D, $RET(D,k)$ denotes the set of k retrieved documents deemed relevant and $REL(D,q)$ the set of all documents referenced by a SR. We define precision and recall in this context as follows:

$$\text{precision} = \frac{|RET(D,k) \cap REL(D,q)|}{|RET(D,k)|} \tag{1}$$

$$\text{recall} = \frac{|RET(D,k) \cap REL(D,q)|}{|REL(D,q)|} \tag{2}$$

Step (3) is conducted using both source-free (SF) and source-augmented (SA) automated metrics. Source-free metrics compare a model's output to a gold standard reference summary, without including any information from the articles used to generated the gold standard summary. For our evaluation purposes, the gold standard is the human-curated answer (derived from conclusions and/or results of each SR). On the other hand, SA metrics additionally consider relevant context to evaluate the quality of model-generated outputs. For our experiments, context is constructed by concatenating a SR's introduction, results, and conclusion sections. The SA metrics we employed (and the LMs they use) include UniEval[26] (T5 -large), COMET (XLM-RoBERTa),[50] and CTC Summary Consistency (BERT).[51]

UniEval is a multi-dimensional evaluator designed for summarization tasks and takes into account four key dimensions (and their corresponding overall average):

- **Coherence:** Assesses whether the summary forms a cohesive and rational body of text;

- **Consistency:** Evaluates the factual alignment between the information presented in the summary and the content of the source document;
- **Fluency:** Assesses the readability and linguistic fluency of a summary;
- **Relevance:** Measures whether the summary contains only the important information from the source document.

COMET is an evaluation metric developed to assess the quality of Machine Translation (MT) systems. Despite being trained on multilingual MT outputs, it performs remarkably well in monolingual settings, when predicting summarization output quality.[52] CTC is an evaluation framework, based on information alignment between input, output, and context, for compression (e.g summary), transduction (e.g translation), and creation (e.g. conversation).

Finally we perform an evaluation using SF metrics, including BERTScore,[53] ROUGE-L,[54] METEOR,[55] chrF[56] , GoogleBLEU, CTC Summary (without providing context) , and CharacTer.[57] The majority of these metrics have shown moderate correlation with human preference and are widely reported in NLG tasks.[25,26]

The multi-dimensional evaluation based on source-augmented metrics makes the assumption that an LLM+RetA model is able to (1) retrieve abstracts of works that were deemed relevant by an author of a SR and (2) synthesize them in a similar fashion. We acknowledge that if this assumption is not met, the evaluation would heavily penalize the output. Conversely, if the system retrieves an article that was not considered by a SR but bears a similar semantic meaning to an article present in the references of a SR, the evaluation would not penalize the generated text. For our proposed method, both behaviors are desired.

4. Baselines and Experiments

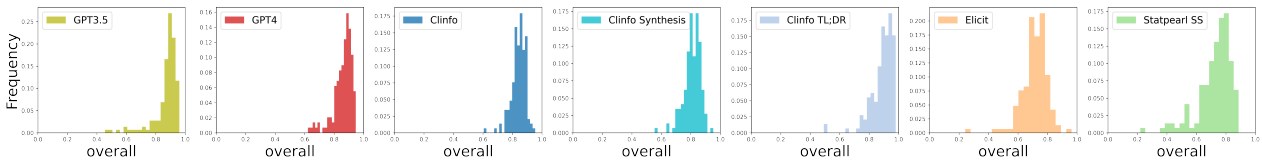

Fig. 6: UniEval Overall Score of 146 questions (unconstrained by published date) from PubMedRS-200 distribution across Unrestricted Search (GPT3.5 and GPT4 zero-shot performance is added)

Using our proposed task, we evaluated the performance of GPT-4 and GPT-3.5 without retrieval augmentation, Clinfo.ai (our GPT-enabled RetA LLM system), and two deployed tools: Elicit (an AI research assistant based on LLMs, designed for facilitating literature review generation, accessed on 07-02-2023), and Statpearls Semantic Search (a free search tool for medical knowledge, accessed on 07-25-2023). While other automated literature summarization systems are available, at the time of this study the vast majority require a subscription to answer multiple questions. Additionally, a subset of these systems refused to provide an answer to a significant number of the PubMedRS-200 questions as posed, making evaluation for these systems fraught and difficult to interpret. We exclude these systems from our analysis.

Table 1: Performance on 146 questions from PubMedRS-200 using source-augmented (SA) metrics: UniEval (T5-large), COMET (XLM-RoBERTa), CTC summary (BERT)

Model	Unified Multi-Dimensional Evaluator (UniEval) Coherence ↑	Consistency ↑	Fluency ↑	Relevance ↑	Overall ↑	COMET ↑	CTC (SA) Consistency ↑	Avg. Length
LLM								
GPT-3.5	0.908 (0.149)	0.694 (0.144)	0.947 (0.059)	0.939 (0.101)	0.872 (0.082)	0.676 (0.075)	0.865 (0.017)	104.834 (47.778)
GPT-4	0.915 (0.099)	0.655 (0.145)	0.942 (0.051)	0.929 (0.078)	0.86 (0.062)	0.677 (0.075)	0.866 (0.017)	84.214 (39.772)
LLM + RetA								
Restricted Search								
Synthesis & TL;DR	**0.949 (0.065)**	0.466 (0.105)	0.903 (0.104)	**0.964 (0.053)**	0.82 (0.055)	0.704 (0.055)	0.84 (0.014)	205.579(46.181)
Synthesis	0.925 (0.066)	0.394 (0.11)	0.893 (0.119)	0.939 (0.101)	0.788 (0.059)	0.693 (0.057)	0.842 (0.015)	165.814 (40.749)
TL;DR	0.866 (0.143)	0.787 (0.161)	0.954 (0.018)	0.826 (0.159)	0.858 (0.098)	0.665 (0.078)	0.874 (0.018)	38.766 (11.682)
Source Dropped								
Synthesis & TL;DR	0.942 (0.092)	0.465 (0.104)	0.918 (0.085)	0.962 (0.059)	0.822 (0.055)	0.706 (0.056)	0.843 (0.014)	204.248 (38.394)
Synthesis	0.925 (0.066)	0.398 (0.112)	0.912 (0.096)	0.943 (0.055)	0.795 (0.055)	0.695 (0.059)	0.845 (0.016)	164.938 (33.221)
TL;DR	0.829 (0.202)	0.763 (0.197)	0.953 (0.029)	0.796 (0.194)	0.835(0.13)	0.672 (0.078)	0.876 (0.017)	38.31 (10.726)
Unrestricted Search								
Our Models								
Synthesis & TL;DR	0.945 (0.064)	0.539 (0.127)	0.912 (0.096)	0.962 (0.059)	0.84 (0.052)	**0.721 (0.055)**	0.852 (0.017)	214.338 (44.173)
Synthesis	0.916 (0.092)	0.48 (0.142)	0.904 (0.098)	0.935 (0.069)	0.809 (0.06)	0.712 (0.057)	0.855 (0.019)	173.379 (38.492)
TL;DR	0.896 (0.123)	**0.81 (0.159)**	**0.955 (0.012)**	0.857 (0.135)	**0.88 (0.081)**	0.681 (0.072)	**0.88 (0.016)**	39.959 (11.754)
Deployed Models								
Elicit[19]	0.854 (0.136)	0.352 (0.147)	0.743 (0.151)	0.902 (0.117)	0.713 (0.085)	0.7 (0.066)	0.866 (0.017)	130.566 (22.946)
Statpearls SS[23]	0.753 (0.225)	0.383 (0.129)	0.93 (0.053)	0.845 (0.159)	0.728 (0.112)	0.633 (0.075)	0.841 (0.016)	118.172 (26.603)

Lastly, since our framework generates two outputs — "TL;DR" and "Literature Summary" (also referred to as "Synthesis") — we conducted evaluations of three forms of Clinfo.ai's output: (1) the synthesis of the articles retrieved and deemed relevant ("Synthesis"); (2) the abbreviated summary distilling the proposed "Synthesis" into one or two sentences ("TL;DR"); (3) the combined "Synthesis" and "TL;DR".

We recognize that the usage of scientific literature to extract question-answer pairs comes with the possibility that an answer deemed correct at the time of acquisition may be incorrect as new discoveries are published. To ensure that a system is not rewarded for simply copy-pasting the text of a retrieved source SR nor penalized when new relevant articles are published, we consider three evaluation regimes:

(1) **Restricted Search (RS)**: The retrieval process is constrained to include publications up to one day before the publication date. While this approach may not guarantee the retrieval of all publications considered important by the authors of each source systematic review, it effectively narrows down the search space to the subset of publications that could have been retrieved and deemed relevant during the review's preparation.
(2) **Source Dropped (SD)**: The retrieval process can retrieve articles published both before and after the source systematic review. However, if the source SR is retrieved, it is removed from the set of relevant articles and not used in the subsequent steps of the summarization process.
(3) **Unrestricted Search (US)** No restriction is applied; the source SR may (but need not)

Table 2: Performance on 146 questions from PubMedRS-200 using source-free (SF) metrics

Model	BERTScore ↑	ROUGE-L ↑	METEOR ↑	chrF ↑	GoogleBLEU ↑	CTC (SF) ↑	CharacTer↓	Avg. Length
LLM								
GPT-3.5	0.781 (0.037)	0.165 (0.053)	0.181 (0.073)	30.2 (10.5)	0.077 (0.036)	0.575 (0.065)	0.912 (0.102)	104.834 (47.778)
GPT-4	0.78 (0.037)	0.157 (0.049)	0.192 (0.07)	31.6 (9.06)	0.074 (0.031)	0.571 (0.064)	0.89 (0.099)	84.214 (39.772)
LLM + RetA								
Restricted Search								
Synthesis & TL;DR	0.77 (0.028)	0.135 (0.043)	0.121 (0.055)	21.5 (9.98)	0.058 (0.03)	0.527 (0.059)	0.993 (0.029)	205.579(46.181)
Synthesis	0.773 (0.028)	0.141 (0.044)	0.133 (0.059)	24.3 (10.4)	0.063 (0.032)	0.533 (0.06)	0.976 (0.056)	165.814 (40.749)
TL;DR	0.784 (0.041)	0.145 (0.068)	0.221 (0.089)	32.7 (7.67)	0.061 (0.043)	0.594 (0.068)	0.833 (0.086)	38.766 (11.682)
Source Dropped								
Synthesis & TL;DR	0.773 (0.028)	0.136 (0.037)	0.119 (0.054)	21.4 (9.69)	0.057 (0.028)	0.53 (0.06)	0.989 (0.036)	204.248 (38.394)
Synthesis	0.775 (0.026)	0.143 (0.038)	0.132 (0.057)	24.1 (9.91)	0.061 (0.043)	0.536 (0.06)	0.976 (0.056)	164.938 (33.221)
TL;DR	0.787 (0.041)	0.148 (0.064)	0.218 (0.078)	33 (6.98)	0.06 (0.039)	0.6 (0.066)	0.83 (0.092)	38.31 (10.726)
Unrestricted Search								
Our Models								
Synthesis & TL;DR	0.786 (0.029)	0.167 (0.06)	0.145 (0.073)	23.5 (11.2)	0.079 (0.046)	0.546 (0.067)	0.989 (0.036)	214.338 (44.173)
Synthesis	0.789 (0.03)	0.178 (0.067)	0.164 (0.084)	26.7 (12)	0.088 (0.051)	0.555 (0.07)	0.975 (0.065)	173.379 (38.492)
TL;DR	0.793 (0.038)	0.169 (0.076)	**0.252 (0.092)**	**35.5 (7.95)**	0.076 (0.049)	**0.61 (0.067)**	**0.825 (0.094)**	39.959 (11.754)
Deployed Models								
Elicit[19]	**0.807 (0.04)**	**0.218 (0.095)**	0.206 (0.093)	31.6 (12.5)	**0.127 (0.085)**	0.596 (0.07)	0.938 (0.096)	130.566 (22.946)
Statpearls SS[23]	0.77 (0.028)	0.136 (0.037)	0.149 (0.057)	26.5 (9.8)	0.062 (0.026)	0.536 (0.06)	0.939 (0.09)	118.172 (26.603)

Table 3: Clinfo.ai Precision and Recall on PubMedRS-200

Evaluation Regime	Precision ↑	Recall ↑	Source Included
Restricted Search	0.224 (0.239)	0.057 (0.061)	0.0 (0.0)
Source Dropped	0.186 (0.22)	0.064 (0.064)	0.0 (0.0)
Unrestricted Search	0.162 (0.175)	0.052 (0.064)	0.965 (0.185)

be included in the set of relevant articles retrieved by the system. Because we could not control the set of articles retrieved and summarized by closed-source tools like Elicit and Statpearls SS, they effectively fall within this evaluation regime.

Finally, to ensure that conformity with the SD regime would not prevent direct comparison with the other evaluation regimes, we removed questions from all other training regimes for which Clinfo.ai could only retrieve the source article (resulting in zero articles remaining after exclusion under the SD regime). This yielded 145 SRs (80 after October 2021 and 65 before).

5. Experimental Results and Analysis

Is RetA associated with significant improvements in automated metric evaluation?

As reported in previous studies,[34,39,58] both GPT-3.5 and GPT-4 without RetA demonstrated strong zero-shot performance using both source-augmented (Table 1) and source-free (Table 2) metrics. Notably, there was no substantial performance drop observed when these

models were presented with questions based on source SRs published after September 2021 (Comparing Table 1 and Table S1 in the Supplement). While more studies are necessary, we postulate that this can be attributed to the models' exposure to prior published works during training. Since SRs are built upon existing literature ranging across multiple years, it is plausible that the models have been trained on relevant information that aids them in providing accurate responses to questions based on newer research. However, comparing all LLM against LLM + RetA models, the inclusion of RetA leads to a slight improvement in the overall performance of the models when evaluated with SF and SA automated metrics, irrespective of the publication date of the source SR. Previous works based on human evaluation have observed a similar trend, corroborating our automated evaluation framework.

How does Clinfo.ai perform compared to other systems?

As depicted in Table 1, Clinfo.ai exhibited better performance in overall UniEval compared to other RetA systems, irrespective of the chosen output strategy (Synthesis, TL;DR, or a concatenation of the two). This improvement in performance remained consistent regardless of the average length of the output, with Clinfo.ai achieving better results for both approximately 3x shorter (TL;DR) and around 2x longer outputs (Synthesis). Furthermore, this performance persisted across all different evaluation regimes, even when the source SR was dropped. This improvement amounted to at least 6.2% and at most 14.9% in UniEval Overall performance. These results suggest two significant points: (1) Our system is not merely copying and pasting information from an SR review. Instead, it demonstrates a genuine ability to process and present the information effectively, resulting in enhanced performance compared to other available tools; and (2) even in the absence of a source SR, Clinfo.ai can still provide conclusions that are better aligned with a source SR's conclusion (compared to tools that might include the source SR).

TL;DR or Synthesis?

Clinfo.ai TL;DR demonstrates significantly better performance compared to Synthesis and Synthesis & TL;DR, even though they all utilize the same relevant retrieved articles. It is worth noting that while Synthesis provides evidence to answer the question based on the retrieved articles, this evidence may not align with the original evidence reported by a Systematic Review (SR). However, the increased performance of TL;DR could be attributed to the LLM's capability to correctly identify the most salient points of the relevant articles and effectively summarize them. On the other hand, using only source-free (SF) metrics (Table 2), Elicit performs better under BERTScore, ROUGE-L and GoogleBLEU, while Clinfo.ai TL;DR performs better under METEOR, chrF, CTC (SF), and CharacTer.

These results highlight a potential limitation of automated evaluation . For instance, SF metrics tend to reward short responses, which may not necessarily be accurate or comprehensive. On the other hand, several SA metrics can assign the best score to considerably larger generations (UniEval's Coherence and Relevance, and Comet), acknowledging their quality and relevance. This discrepancy in evaluation metrics raises concerns about the fair assessment of model performance and emphasizes the need for a comprehensive evaluation approach.

Comparing different evaluation regimes, the best performance was observed under the Unrestricted Search evaluation regime, possibly due to the fact that the source SR was retrieved

on 96.5% of the questions. As expected given the restricted set of retrievable documents, Clinfo.ai's precision was highest under the Restricted Search regime (Table 3).

6. Conclusion

The rapidly expanding medical literature and the capabilities of LLMs to process and summarize vast amounts of information have led to the development of several tools that utilize LLMs to generate on-demand summaries of published scientific literature. However, the lack of high-quality datasets and appropriate benchmarking tasks has hindered rigorous evaluations of these tools. To address this gap, we have introduced Clinfo.ai, an open-source end-to-end LLM-chain workflow designed to query, evaluate, and synthesize medical literature into concise summaries for answering questions on demand. Additionally, we introduce a unique dataset, PubMedRS-200, which consists of questions and answers extracted from systematic reviews, enabling automatic evaluation of LLM performance in Retrieval Augmentation Question Answering. Our tools and benchmarking dataset are publicly available to ensure reproducibility and to facilitate further research in harnessing LLMs for Retrieval Augmentation Question Answering tasks.

7. Limitations

In this study, we employed automated metrics that have demonstrated moderate-to-high correlation with human preferences, but we did not explicitly solicit human preferences to evaluate the RetA LLM systems considered. Future work should consider including human evaluation to ensure alignment of automated metrics and human preferences. Lastly, it is worth noting that prior studies have reported that LLMs demonstrate the ability to generate accurate Boolean operators and syntax, effectively adhering to PubMed query formats. However, our observations revealed that these models also generated hallucinated MeSH terms, which could potentially lead to the exclusion of relevant studies. To overcome this limitation, future research efforts should prioritize improving the query generation process, ensuring that generated MeSH terms are reliable and relevant for better precision and recall in medical literature search tasks.

8. Acknowledgments

AL is funded by Arc Institute. SF is supported by a Stanford Graduate Fellowship. This effort was supported in part by the Mark and Debra Leslie endowment for AI in Healthcare. We thank Will Haberkorn for his aid with Figure S1.

References

1. K. I. Bougioukas, E. C. Bouras, K. I. Avgerinos, T. Dardavessis and A.-B. Haidich, How to keep up to date with medical information using web-based resources: A systematised review and narrative synthesis, *Health Information & Libraries Journal* **37**, 254 (2020).
2. E. Landhuis, Scientific literature: Information overload, *Nature* **535**, 457 (2016).
3. R. Van De Schoot, J. De Bruin, R. Schram, P. Zahedi, J. De Boer, F. Weijdema, B. Kramer, M. Huijts, M. Hoogerwerf, G. Ferdinands *et al.*, An open source machine learning framework for efficient and transparent systematic reviews, *Nature machine intelligence* **3**, 125 (2021).

4. J. E. Andrews, K. A. Pearce, C. Ireson and M. M. Love, Information-seeking behaviors of practitioners in a primary care practice-based research network (pbrn), *Journal of the Medical Library Association* **93**, p. 206 (2005).
5. G. Del Fiol, T. E. Workman and P. N. Gorman, Clinical questions raised by clinicians at the point of care: a systematic review, *JAMA internal medicine* **174**, 710 (2014).
6. A. Daei, M. R. Soleymani, H. Ashrafi-Rizi, A. Zargham-Boroujeni and R. Kelishadi, Clinical information seeking behavior of physicians: A systematic review, *International journal of medical informatics* **139**, p. 104144 (2020).
7. J. W. Ely, J. A. Osheroff, M. L. Chambliss, M. H. Ebell and M. E. Rosenbaum, Answering physicians' clinical questions: obstacles and potential solutions, *Journal of the American Medical Informatics Association* **12**, 217 (2005).
8. R. Borah, A. W. Brown, P. L. Capers and K. A. Kaiser, Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the prospero registry, *BMJ open* **7**, p. e012545 (2017).
9. D. A. Cook, M. T. Teixeira, B. S. Heale, J. J. Cimino and G. Del Fiol, Context-sensitive decision support (infobuttons) in electronic health records: a systematic review, *Journal of the American Medical Informatics Association* **24**, 460 (2017).
10. D. Lobach, G. D. Sanders, T. J. Bright, A. Wong, R. Dhurjati, E. Bristow, L. Bastian, R. Coeytaux, G. Samsa, V. Hasselblad *et al.*, Enabling health care decisionmaking through clinical decision support and knowledge management., *Evidence report/technology assessment* , 1 (2012).
11. P. A. Bonis, G. T. Pickens, D. M. Rind and D. A. Foster, Association of a clinical knowledge support system with improved patient safety, reduced complications and shorter length of stay among medicare beneficiaries in acute care hospitals in the united states, *International journal of medical informatics* **77**, 745 (2008).
12. T. Isaac, J. Zheng and A. Jha, Use of uptodate and outcomes in us hospitals, *Journal of hospital medicine* **7**, 85 (2012).
13. D. A. Reed, C. P. West, E. S. Holmboe, A. J. Halvorsen, R. S. Lipner, C. Jacobs and F. S. McDonald, Relationship of electronic medical knowledge resource use and practice characteristics with internal medicine maintenance of certification examination scores, *Journal of general internal medicine* **27**, 917 (2012).
14. J. J. Cimino, G. Elhanan and Q. Zeng, Supporting infobuttons with terminological knowledge., in *Proceedings of the AMIA annual fall symposium*, 1997.
15. D. Demner-Fushman, Y. Mrabet and A. Ben Abacha, Consumer health information and question answering: helping consumers find answers to their health-related information needs, *Journal of the American Medical Informatics Association* **27**, 194 (2020).
16. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, Retrieval-augmented generation for knowledge-intensive nlp tasks, *Advances in Neural Information Processing Systems* **33**, 9459 (2020).
17. Q. Jin, R. Leaman and Z. Lu, Pubmed and beyond: Recent advances and best practices in biomedical literature search, *arXiv preprint arXiv:2307.09683* (2023).
18. J. M. Nicholson, M. Mordaunt, P. Lopez, A. Uppala, D. Rosati, N. P. Rodrigues, P. Grabitz and S. C. Rife, Scite: A smart citation index that displays the context of citations and classifies their intent using deep learning, *Quantitative Science Studies* **2**, 882 (2021).
19. Ought, Elicit: The ai research assistant (2023).
20. GlacierMD, Glaciermd - a modern physician reference (2023).
21. Consensus, Consensus (2023).
22. OpenEvidence, Openevidence: Making medical knowledge more useful, open, accessible, and understandable (2023).
23. Hippocratic AI, statpearls semantic search (2023).

24. M. Sallam, Chatgpt utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns, *Healthcare* **11** (2023).
25. I. Ni'mah, M. Fang, V. Menkovski and M. Pechenizkiy, Nlg evaluation metrics beyond correlation analysis: An empirical metric preference checklist, *arXiv preprint arXiv:2305.08566* (2023).
26. M. Zhong, Y. Liu, D. Yin, Y. Mao, Y. Jiao, P. Liu, C. Zhu, H. Ji and J. Han, Towards a unified multi-dimensional evaluator for text generation, *arXiv preprint arXiv:2210.07197* (2022).
27. S. L. Fleming, A. Lozano, W. J. Haberkorn, J. A. Jindal, E. P. Reis, R. Thapa, L. Blankemeier, J. Z. Genkins, E. Steinberg, A. Nayak *et al.*, Medalign: A clinician-generated dataset for instruction following with electronic medical records, *arXiv preprint arXiv:2308.14089* (2023).
28. M. Bakker, M. Chadwick, H. Sheahan, M. Tessler, L. Campbell-Gillingham, J. Balaguer, N. McAleese, A. Glaese, J. Aslanides, M. Botvinick *et al.*, Fine-tuning language models to find agreement among humans with diverse preferences, *Advances in Neural Information Processing Systems* **35**, 38176 (2022).
29. S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg *et al.*, Sparks of artificial general intelligence: Early experiments with gpt-4, *arXiv preprint arXiv:2303.12712* (2023).
30. M. Sallam, Chatgpt utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns, in *Healthcare*, (6)2023.
31. G. Eysenbach *et al.*, The role of chatgpt, generative language models, and artificial intelligence in medical education: a conversation with chatgpt and a call for papers, *JMIR Medical Education* **9**, p. e46885 (2023).
32. M. Cascella, J. Montomoli, V. Bellini and E. Bignami, Evaluating the feasibility of chatgpt in healthcare: an analysis of multiple clinical and research scenarios, *Journal of Medical Systems* **47**, p. 33 (2023).
33. S. L. Fleming, K. Morse, A. M. Kumar, C.-C. Chiang, B. Patel, E. P. Brunskill and N. Shah, Assessing the potential of usmle-like exam questions generated by gpt-4, *medRxiv* , 2023 (2023).
34. T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo *et al.*, Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models, *PLoS digital health* **2**, p. e0000198 (2023).
35. E. Mitchell, C. Lin, A. Bosselut, C. D. Manning and C. Finn, Memory-based model editing at scale, in *International Conference on Machine Learning*, 2022.
36. W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong *et al.*, A survey of large language models, *arXiv preprint arXiv:2303.18223* (2023).
37. H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, Llama: Open and efficient foundation language models, *arXiv preprint arXiv:2302.13971* (2023).
38. Anthropic, Claude 2 (2023).
39. W. Hiesinger, C. Zakka, A. Chaurasia, R. Shad, A. Dalal, J. Kim, M. Moor, K. Alexander, E. Ashley, J. Boyd *et al.*, Almanac: Retrieval-augmented language models for clinical medicine (2023).
40. D. Soong, S. Sridhar, H. Si, J.-S. Wagner, A. C. C. Sá, C. Y. Yu, K. Karagoz, M. Guan, H. Hamadeh and B. W. Higgs, Improving accuracy of gpt-3/4 results on biomedical data using a retrieval-augmented language model, *arXiv preprint arXiv:2305.17116* (2023).
41. Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen and X. Lu, Pubmedqa: A dataset for biomedical research question answering, *arXiv preprint arXiv:1909.06146* (2019).
42. H. Scells and G. Zuccon, Generating better queries for systematic reviews, in *The 41st international ACM SIGIR conference on research & development in information retrieval*, 2018.
43. M. Pourreza and F. Ensan, Towards semantic-driven boolean query formalization for biomedical

systematic literature reviews, *International Journal of Medical Informatics* , p. 104928 (2022).
44. K. Khan, R. Kunz, J. Kleijnen and G. Antes, *Systematic reviews to support evidence-based medicine* (Crc press, 2011).
45. E. W. Sayers, E. E. Bolton, J. R. Brister, K. Canese, J. Chan, D. C. Comeau, C. M. Farrell, M. Feldgarden, A. M. Fine, K. Funk *et al.*, Database resources of the national center for biotechnology information in 2023, *Nucleic acids research* **51**, D29 (2023).
46. T. Wu, M. Terry and C. J. Cai, Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts, in *Proceedings of the 2022 CHI conference on human factors in computing systems*, 2022.
47. T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever and D. Amodei, Language models are few-shot learners, in *Advances in Neural Information Processing Systems*, eds. H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan and H. Lin (Curran Associates, Inc., 2020).
48. T. Kojima, S. S. Gu, M. Reid, Y. Matsuo and Y. Iwasawa, Large language models are zero-shot reasoners, in *Advances in Neural Information Processing Systems*, eds. S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho and A. Oh (Curran Associates, Inc., 2022).
49. A. Trotman, A. Puurula and B. Burgess, Improvements to bm25 and language models examined, in *Proceedings of the 19th Australasian Document Computing Symposium*, 2014.
50. R. Rei, C. Stewart, A. C. Farinha and A. Lavie, Comet: A neural framework for mt evaluation, *arXiv preprint arXiv:2009.09025* (2020).
51. M. Deng, B. Tan, Z. Liu, E. P. Xing and Z. Hu, Compression, transduction, and creation: A unified framework for evaluating natural language generation, *arXiv preprint arXiv:2109.06379* (2021).
52. K. Mateusz and P. Pecina, From comet to comes–can summary evaluation benefit from translation evaluation?, in *Proceedings of the 3rd Workshop on Evaluation and Comparison of NLP Systems*, 2022.
53. T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger and Y. Artzi, Bertscore: Evaluating text generation with bert, *arXiv preprint arXiv:1904.09675* (2019).
54. C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in *Text summarization branches out*, 2004.
55. S. Banerjee and A. Lavie, Meteor: An automatic metric for mt evaluation with improved correlation with human judgments, in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005.
56. M. Popović, chrf: character n-gram f-score for automatic mt evaluation, in *Proceedings of the tenth workshop on statistical machine translation*, 2015.
57. W. Wang, J.-T. Peter, H. Rosendahl and H. Ney, CharacTer: Translation edit rate on character level, in *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, (Association for Computational Linguistics, Berlin, Germany, August 2016).
58. H. Nori, N. King, S. M. McKinney, D. Carignan and E. Horvitz, Capabilities of gpt-4 on medical challenge problems, *arXiv preprint arXiv:2303.13375* (2023).