MedAttr: Cost-Effective, Diverse Attribution of LLM Responses in Medical Conversation Systems to Reliable Sources

Abstract

Conversational systems that emulate medical professionals are increasingly used to triage patients and interpret clinical reports, but their safety depends on grounding responses in reliable evidence rather than unchecked model guesses. Existing attribution pipelines often retrieve many near-duplicate passages, wasting a limited passage budget and driving up computational cost without improving trust. In this work, we focus on cost-effective diversity in attribution: given a fixed budget, how can we support a model's response with a set of passages that is both highly relevant and meaningfully diverse? We introduce **MedAttr**, a two-stage attribution system that first uses inexpensive embedding-based retrieval to build a high-recall candidate pool, and then applies targeted submodular optimization to explicitly trade off relevance and diversity when selecting the final supporting passages. This hybrid design reduces redundancy without extra model calls or heavyweight re-ranking, making MedAttr efficient and scalable. In medical dialogue and report-interpretation settings, experiments with AutoAIS and disease label matching show that MedAttr improves attribution quality and label coverage under the same passage budget, while maintaining or reducing computational cost compared to top-k retrieval and re-ranking baselines, demonstrating a practical path toward more trustworthy medical conversational systems.[1]

Keywords: Attribution for medical LLMs, Diverse evidence selection, Cost-efficient retrieval

1. Introduction

Conversational systems that use natural language to communicate (i.e chatbots) are in big demand in many industries [(Breazu and Katsos, 2024; Bartz and Bartz, 2023; Kitamura, 2023)]. These systems can take human queries and generate responses conditioned on domain knowledge and past queries [(Caldarini et al., 2022; Dam et al., 2024)]. With the advent of LLMs, these systems can reason and are easier to implement and deploy.

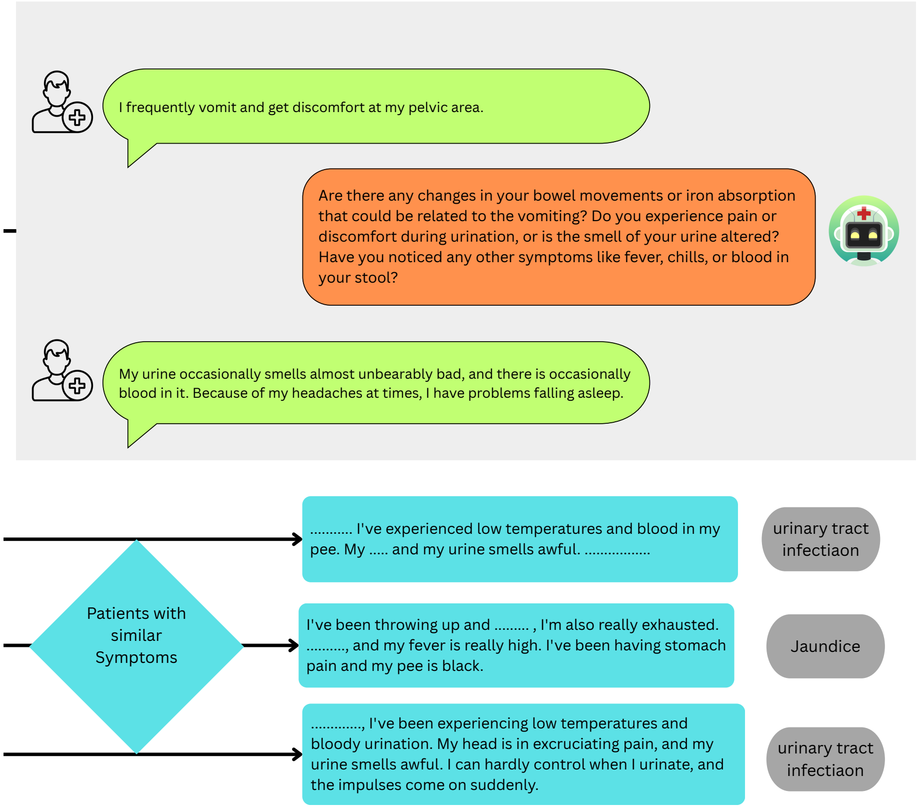
Large language model (LLM)-based chatbots have shown strong performance across a range of medical tasks. Systems like ChatGPT have demonstrated success in medical question answering, particularly when interpreting clinical notes, and have even passed medical licensing examinations – highlighting their capacity for comprehension and reasoning. These generative models for dialogue are also increasingly integrated into patient care, engaging in conversations about symptoms, medication dosages, and safety precautions, thereby marking a significant advancement in digital health technologies [(Nori et al., 2023; Thirunavukarasu et al., 2023; Dam et al., 2024)].

A key challenge with LLM-based chat systems is their tendency to generate factually incorrect or unsupported responses [(Roller et al., 2020; Shuster et al., 2019)]. Since these

[1] We release our code here

responses are very fluent and organic, they might mislead the users. This reduces users' trust and prevents these systems from being deployed in critical domains.

This highlights the importance of attribution – linking generated responses to supporting source documents – for conversational systems in the medical domain. In this work, we explore methods for efficient attribution in responses generated by medical dialogue systems. We also examine how to select multiple, diverse supporting texts within a predefined budget to back the generated response. Our experiments are conducted on both public and private datasets. For evaluation, we use AutoAIS, (Bohnet et al., 2022), on private data and label matching on public data.



I frequently vomit and get discomfort at my pelvic area.

Are there any changes in your bowel movements or iron absorption that could be related to the vomiting? Do you experience pain or discomfort during urination, or is the smell of your urine altered? Have you noticed any other symptoms like fever, chills, or blood in your stool?

My urine occasionally smells almost unbearably bad, and there is occasionally blood in it. Because of my headaches at times, I have problems falling asleep.

.... I've experienced low temperatures and blood in my pee. My ..... and my urine smells awful. ....

I've been throwing up and .... , I'm also really exhausted. ...., and my fever is really high. I've been having stomach pain and my pee is black.

............., I've been experiencing low temperatures and bloody urination. My head is in excruciating pain, and my urine smells awful. I can hardly control when I urinate, and the impulses come on suddenly.

Figure 1: MedAttr grounds conversational responses by first retrieving the top neighbors from the source data, then applying targeted submodular optimization to select a diverse set of highly attributable passages. When an underlying corpus (such as an EHR) is unavailable, attribution is evaluated by matching labels between the classifier-predicted label of the current conversation and the retrieved passages from prior patient conversations.

Existing approaches typically rely on information retrieval techniques [(Weston et al., 2018; Lewis et al., 2020)] to select the top-K relevant neighbors. However, these selected passages are not always fully attributable or diverse. Some methods incorporate Natural Language Inference models to check for entailment [(Huang et al., 2024; Honovich et al., 2022)], but these approaches incur high inference times that can hinder user experience. Furthermore, in scenarios with budget-constraints (limiting the size of text), selecting the top K passages often results in redundant information.

Our method, MedAttr, solves the question *"Can we guarantee that every chatbot response is anchored to evidence a clinician can verify, even when time and memory are scarce?"* using transformer-based vector matching to retrieve top neighbors and applying targeted submodular optimization to select the most relevant and diverse supporting texts.

Our contributions are as follows:

- We propose **MedAttr**, a novel attribution framework for medical conversational systems that combines vector-based retrieval with submodular optimization.
- We design a scalable and efficient selection mechanism that jointly optimizes for relevance and diversity under a fixed passage budget.

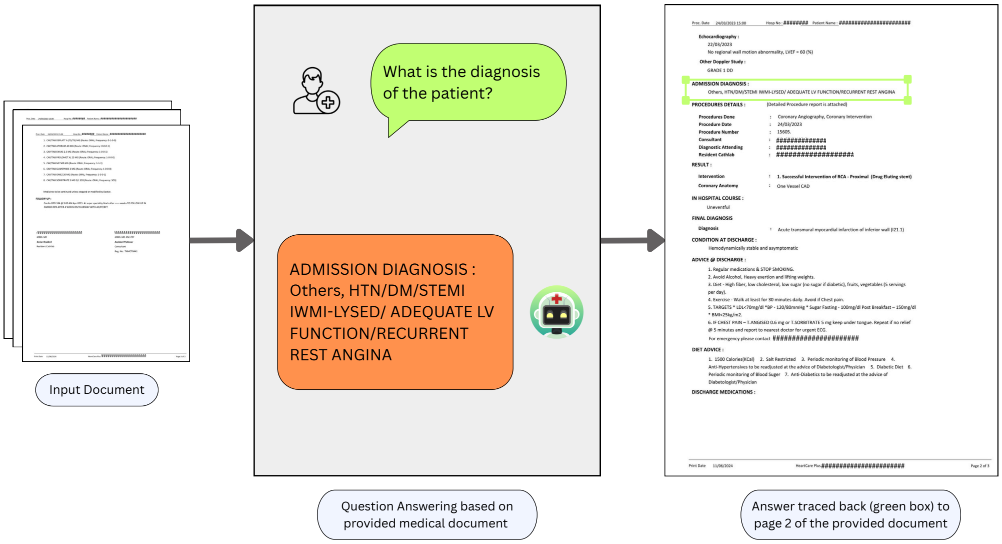
Generalizable Insight

Attribution to source documents for LLM-based chat generated text under a budget can be significantly improved using targeted submodular subset selection. We need relevant as well as diverse text to support these responses.

2. Related Work

Attribution for LLMs Recent years have seen significant research on attribution for large language models (LLMs) [(Rashkin et al., 2023; Bohnet et al., 2022; Yue et al., 2023)]. Most attribution techniques fall into two main categories: Retrieve-Then-Read methods [(Hu et al., 2019; Nishida et al., 2018; Izacard and Grave, 2020)] and posthoc methods [(Huo et al., 2023; Gao et al., 2022)]. Retrieve-Then-Read approaches first extract relevant documents from a corpus, then pass both the retrieved content and the query to the LLM, which generates an answer. In contrast, post-hoc methods let the LLM generate an answer from the query first, and then retrieve supporting passages from the corpus using both the question and the answer—these passages serve as the attribution. A more recent strategy involves fine-tuning the LLM to directly generate attributions (Bohnet et al., 2022), where the model outputs an answer along with a signal that identifies the supporting passages from the source corpus. Our approach, MedAttr, follows the post-hoc framework to effectively link responses to the relevant sources.

Selection of retrieved passages Selection of passages from the retrieved passages are done in a number of ways [(Zhu et al., 2023; Zhuang et al., 2023; Schlatt et al., 2024)]. In most of the cases there will be a groundtruth ordering available, in which case we can train an LLM to be a reranker. In unsupervised scenarios, mostly Top-K retrieved passages are selected. The literature on subset selection has primarily looked at selecting representative


Input Document
Question Answering based on provided medical document
Answer traced back (green box) to page 2 of the provided document

What is the diagnosis of the patient?

ADMISSION DIAGNOSIS : Others, HTN/DM/STEMI IWMI-LYSED/ ADEQUATE LV FUNCTION/RECURRENT REST ANGINA

Proc. Date 24/03/2023 15:00 Hosp No : ######## Patient Name : #######################

Echocardiography :
22/03/2023
No regional wall motion abnormality, LVEF = 60 (%)

Other Doppler Study :
GRADE 1 DD

ADMISSION DIAGNOSIS :
Others, HTN/DM/STEMI IWMI-LYSED/ ADEQUATE LV FUNCTION/RECURRENT REST ANGINA

PROCEDURES DETAILS : (Detailed Procedure report is attached)

Procedures Done	:	Coronary Angiography, Coronary Intervention
Procedure Date	:	24/03/2023
Procedure Number	:	15605.
Consultant	:	##############
Diagnostic Attending	:	##############
Resident Cathlab	:	######################

RESULT :

Intervention	:	**1. Successful Intervention of RCA - Proximal (Drug Eluting stent)**
Coronary Anatomy	:	One Vessel CAD

IN HOSPITAL COURSE :
Uneventful

FINAL DIAGNOSIS

Diagnosis	:	Acute transmural myocardial infarction of inferior wall (I21.1)

CONDITION AT DISCHARGE :
Hemodynamically stable and asymptomatic

ADVICE @ DISCHARGE :

1. Regular medications & STOP SMOKING.
2. Avoid Alcohol, Heavy exertion and lifting weights.
3. Diet - High fiber, low cholesterol, low sugar (no sugar if diabetic), fruits, vegetables (5 servings per day).
4. Exercise - Walk at least for 30 minutes daily. Avoid if Chest pain.
5. TARGETS * LDL<70mg/dl *BP - 120/80mmHg * Sugar Fasting - 100mg/dl Post Breakfast – 150mg/dl * BMI<25kg/m2.
6. IF CHEST PAIN – T.ANGISED 0.6 mg or T.SORBITRATE 5 mg keep under tongue. Repeat if no relief @ 5 minutes and report to nearest doctor for urgent ECG.

For emergency please contact #######################

DIET ADVICE :
1. 1500 Calories(KCal) 2. Salt Restricted 3. Periodic monitoring of Blood Pressure 4. Anti-Hypertensives to be readjusted at the advice of Diabetologist/Physician 5. Diabetic Diet 6. Periodic monitoring of Blood Suger 7. Anti-Diabetics to be readjusted at the advice of Diabetologist/Physician

DISCHARGE MEDICATIONS :

Print Date 11/06/2024 HeartCare Plus####################### Page 2 of 3

Figure 2: MedAttr grounds conversational responses by retrieving the top neighbors from the underlying source—such as an EHR—and then applying targeted submodular optimization to select a diverse, highly attributable set of supporting passages within the EHR

subsets of data points [(Mirzasoleiman et al., 2020; Killamsetty et al., 2021b,a)]. Recently, several approaches have been developed where subsets of K points are selected based on a specific target query (Kothawade et al., 2021b).

Chatbots in Medicine Chatbots in medical domain is an extensively studied topic. Most of the existing approaches focus on chatbots' accuracy in diagnosing diseases [(Jegadeesan et al., 2023; Tjiptomongsoguno et al., 2020; Laranjo et al., 2018)]. But these methods do not incorporate grounding. Similarly, a lot of work has gone to QA systems without grounding [(Wu et al., 2024; Yang et al., 2022; Han et al., 2023)].

3. Background

3.1. GuideQ

GuideQ, (Mishra et al., 2024), is an LLM based conversational system that refines classification through conversational dialogue. It gradually narrows down the list of candidate labels (e.g., diseases) using an intelligent prompting technique. Specifically, GuideQ accepts a partial input x – which consists of a sequence of queries and responses between the user and the system and employs a classifier to assign one of the n possible labels.

3.2. Attribution

We invoke the definition from (Rashkin et al., 2023) –
'Attribution - A sentence/paraphrase of a sentence that is interpretable in the linguistically

empty context and that preserves the truth-conditional meaning of the sentence in context c is attributable to a set of parts P of some underlying corpus K if and only if: A generic hearer will, with a chosen level of confidence, affirm the following statement: 'According to P, s'. This measure is termed as Attributable to Identifiable sources/ AIS.'

This definition of attribution should not be confused with attributing to a training data point. The provided attribution (support document/passage) may or may not have affected the generation of the answer. (Bohnet et al., 2022) proposed AutoAIS in order to remove the human component from the evaluation. Here the task of AIS is modeled as a Natural Language Inference Task that checks whether the passage entails the answer given the question.

3.3. Submodular Functions

A set function $F : 2^{\mathcal{P}} \rightarrow \mathbb{R}$ is submodular if it satisfies *diminishing returns*:

$$\forall A \subseteq B \subset \mathcal{P}, \forall p \notin B : \Delta(p|A) \geq \Delta(p|B)$$

Where $\Delta(p|S) = F(S \cup \{p\}) - F(S)$ is the marginal gain.

4. Methods

4.1. Background

We define a conversational dialogue as a sequence of queries and responses, $q_1, r_1, q_2, r_2, ...q_l, r_l$. Queries are generated by the user and responses are generated by the system (i.e the LLM). Our task is to find attribution for any response r_c given $q_1, r_1, q_2, r_2, ...q_{c-1}, r_{c-1}, q_c$.

In cases where electronic health records (EHRs) or any other corpus D is provided with a set of documents $d_1, d_2, ...d_n$, let us assume each document d_i is made up of passages (some paragraphs) $p_1, p_2...p_{|d_i|}$. Given a passage budget K, the goal is to select a subset $A \subseteq \{p_1, p_2, \ldots, p_{100}\}$ of size $|A| = K$ such that each passage $p_i \in A$ maximally supports the response r_c given the dialogue history $(q_1, r_1, \ldots, q_c, r_c)$, subject to constraints on semantic relevance and diversity. We attribute the response, r_c to an underlying corpus (like EHR) by retrieving most relevant passages $p_1, p_2, ...p_{100}$. We then use our algorithm MedAttr over $p_1, p_2...p_{100}$ and select diverse passages p_is given a budget K and this forms our attribution.

In case of some public data no corpus like EHR is available, but there is text data about similar symptoms and the corresponding label of other patients. In this case we define attribution as a label matching problem, whether we are able to retrieve the text of a patient with the exact same label as in the ongoing conversation.

4.2. Method

MedAttr consists of 2 parts. First part is the retrieval of relevant passages and the second part is the selection of relevant passages from the retrieved ones.

4.3. Retrieval Of Relevant Passages

This part consists of 2 stages. In the first stage, a neural network is used to create embeddings of the passages. These embeddings are formed in the vector space such that semantically similar passages will be close to each other. The second part consists of a vector index that stores the vectors in a compressed form. When a query vector is given as input, the vector index returns the corpus vector (and its neighbours) that forms the closest match.

4.4. Targeted Subset Selection

Given a set V of passages $p_1, p_2, p_3, ...p_{100}$ (returned by the retriever), we have to select K passages that provide adequate grounding to a response r_c. For this we used a combination of different submodular mutual information functions that ensured query relevance and diversity of the selected passages.

To enforce reliability and control diversity we perform the subset selection with a modified utility function. We select passages based on the greedy algorithm that maximizes $U(S)$

$$U(S) = \alpha \sum_{i \in S} R(S) \quad + (1 - \alpha) Sub(S) \tag{1}$$

- S is the selected subset of passages.
- R_i is the retriever score for item i; the total retrieval relevance is given by $R(S) = \sum_{i \in S} R_i$.
- Sub(S) is the submodular utility function applied to S, promoting diversity and broader coverage.
- $\alpha \in [0, 1]$ is a weighting parameter that controls the trade-off:
  - $\alpha = 1$: selection is based entirely on retriever scores (pure retrieval).
  - $\alpha = 0$: selection is driven solely by the submodular function (pure diversity-based selection).
  - $0 < \alpha < 1$: balances retrieval relevance and diversity.

4.5. Algorithm

The selection of the optimal set S is performed via greedy submodular maximization. The greedy algorithm iteratively selects the context that maximizes the marginal gain in utility $U(S)$, defined as:

$$S^* = \arg \max_{S \subseteq C, |S| = k} U(S)$$

Formally, the greedy algorithm initializes $S = \emptyset$ and, at each step i, selects:

$$c^* = \arg \max_{c \in C \setminus S} [U(S \cup \{c\}) - U(S)]$$

This process is repeated until $|S| = K$, where $U(S)$ is the submodular utility function.

5. Experiments

5.1. Datasets

We experimented on four different datasets from the public and private domains to assess the attribution capabilities and robustness of our suggested approach. Each dataset poses unique challenges and contributes to the comprehensive assessment of our methods in real-world medical and healthcare scenarios. The Datasets are Symptom2Disease (S2D), Human Stress Prediction (Stress), MIMIC-QA, and HOSPITAL A.
The HOSPITAL A dataset is a proprietary collection of electronic health records (EHRs) selected from a top medical institution, whereas S2D, Stress, and MIMIC-QA are openly accessible.

Detailed information about datasets is provided here B

5.2. OCR Pipeline for Medical Documents

Our OCR pipeline is designed to process medical documents from the private dataset by generating HOCR files for each page image, ensuring accurate structure, content, and spatial information extraction. Medical documents often contain critical information such as patient records, diagnostic reports, and lab results, which must be extracted with high precision for effective decision-making. The pipeline begins by splitting the input document into pages. It then identifies words, lines, and blocks (which are not detected as tables) in each page with their positions marked using bounding boxes. For text recognition and HOCR generation, we use Tesseract OCR (Smith, 2007), which converts detected text regions into machine-readable text sequences while preserving their spatial layout.

Given the importance of tables in medical documents, where they provide a concise representation of medical parameters and test results, our pipeline ensures their accurate detection and reconstruction. First, our custom-trained YOLOv8 model (Varghese and M., 2024) identifies the regions of the tables. Once tables are detected, TATR (Smock and Pesala, 2021) determines the positions of the rows and columns, while SPRINT (Kudale et al., 2024) decodes the cellular layout, ensuring that all cells are properly segmented. After cell demarcation, we perform OCR again using Tesseract to extract text cell-wise. Once text from all cells is extracted, the table is fully reconstructed, preserving its original format and structure. To maintain the relative positions of all the elements of the page, our pipeline embeds this reconstructed table sequence in the HOCR file using HTML tags at the appropriate position.

5.3. Models Compared

- Baseline 1 - BM25 - We employ the BM25 (Trotman et al., 2014) to retrieve the top K relevant passages as attributions for a given response r.

- Baseline 2 - gtr-t5-base - Following the approach described in Section 5.4, we first retrieve the top 100 nearest neighbors (i.e., the passages corresponding to the top 100 vectors) and then select the top K passages from this set as attributions.

- Our approach - MedAttr - We use the same initial retrieval method as in Baseline 2 (Section 5.4) but further refine the selection by applying targeted submodular optimization (Kothawade et al., 2021a) to choose K passages as attributions.

5.4. Experiment – Symptom2Disease and Humanstress dataset

The experimental procedure is outlined as follows. First, passage embeddings are generated using a sentence transformer model (gtr-t5-base (Ni et al., 2021)), as described in Section 4.3. Next, vector indices are constructed using FAISS (Douze et al., 2024). Query and response embeddings are similarly obtained with the same GTR model. Furthermore, conversations are formed from the datasets using GuideQ, as detailed in Section 3.1.

To ground these conversations, we apply our MedAttr and other baselinbes 5.3, which extracts relevant passages from the original dataset. This grounding process enriches the interactions by providing both the user and the specialist with contextual information from other patients exhibiting similar symptoms.

5.5. Experiment – HOSPITAL A and MIMIC

HOSPITAL A: Medical-QA datapoints generated using Electronic Health Records provided in PDF format (private data).
Medical-QA datapoints generated using MIMIC: Patient Discharge Diagnoses (PDD) provided in JSON format.

Attribution: We perform task of attribution where we match the generated QA pairs to the most relevant sections from the original EHR/PDD data, ensuring that each QA pair is linked to its corresponding context in the source documents. This structured approach allowed us to generate and validate QA pairs anchored in authentic clinical data.

5.6. Metrics

We use three key metrics to evaluate the quality of attribution across various datasets and tasks.

Exact Match Accuracy

This metric measures whether the retrieved passage exactly matches the ground-truth supporting passage. It is a strict evaluation criterion and is particularly relevant in datasets where gold attribution is explicitly available as text.

Label Match Accuracy

This metric evaluates whether the label inferred from the retrieved passage matches the ground-truth label assigned to the example. It is especially useful in scenarios where the attribution is expected to support a classification decision, such as disease diagnosis or stress level.

AutoAIS

For medical QA settings, we use the **AutoAIS** score, which automatically assesses attribution quality using a Natural Language Inference (NLI) model. Specifically, it tests the degree to which the retrieved context entails the correctness of a generated answer. The model evaluates the following entailment query:

```
Premise: {context}
Hypothesis: The answer to the question '{question}' is '{answer}'.
```

AutoAIS provides a scalable and model-based alternative to manual evaluation, especially in settings where ground-truth answer spans are not available.

6. Results

6.1. Results on conversation attribution

Dataset	Setting	Retrieval	Top-1	Top-3	Top-5	Top-10
S2D	Exact Match	bm25	0.000	0.017	0.017	0.050
		gtr	0.067	0.133	0.233	**0.283**
		MedAttr	0.050	0.117	0.217	0.250
	Label Match	bm25	0.267	0.417	0.483	0.583
		gtr	0.450	0.550	0.583	0.667
		MedAttr	0.450	0.533	0.583	0.**683**
Human Stress	Exact Match	bm25	0.056	0.063	0.070	0.099
		gtr	0.120	0.148	0.155	**0.211**
		MedAttr	0.092	0.113	0.155	**0.211**
	Label Match	bm25	0.296	0.535	0.627	0.831
		gtr	0.472	0.641	0.718	0.845
		MedAttr	0.324	0.704	0.768	**0.859**

Table 1: Top-k accuracy of retrieval methods on S2D and Human Stress datasets with values rounded to three decimal places.

We evaluate attribution across three strategies – BM25, GTR, and MedAttr for two tasks conversation attribution and medical attributed - QA. We show that our method MedAttr performed better than other baselines, especially on higher budget settings.

We evaluate two attribution criteria in Table 1:

- **Exact Match:** Whether the retrieved passage exactly matches the ground-truth supporting text.
- **Label Match:** Whether the retrieved passage aligns with the correct disease label (S2D, Human Stress datasets).

Table 1 presents attribution evaluations for 2 datasets S2D and Human Stress. Across both datasets MedAttr performs competitively at lower-k and consistently outperforms other methods at higher-k, highlighting the strength of its targeted, diverse subset selection.

Label Match: On the Human Stress dataset, MedAttr achieves the highest Top-5 & Top-10 accuracy of **76.8% 85.9%**, outperforming both GTR and BM25. On S2D, MedAttr slightly surpasses GTR at Top-10 (**68.3%** vs. 66.7%), despite similar performance at lower values of k.

Exact Match: GTR achieves marginally better results in S2D at lower values of k, but MedAttr catches up by Top-10. In Human Stress, both GTR and MedAttr achieve the same Top-10 accuracy of 21.1%. BM25 consistently underperforms across all metrics, reflecting its lack of semantic attribution capabilities.

6.2. Results on Medical Attributed QA

Dataset	**Strategy**	**Top-1**	**Top-2**	**Top-3**	**Top-4**	**Top-5**
HOSPITAL A - EHR	bm25	0.5188	0.6015	0.6692	0.6842	0.6992
	gtr-t5-base	0.4812	0.5564	0.6316	0.6617	0.6767
	MedAttr	0.4962	0.5714	0.6241	0.6617	0.6767
MIMIC	bm25	0.3553	0.3832	0.4025	0.4125	0.4165
	gtr-t5-base	0.3786	0.4019	0.4145	0.4218	0.4251
	MedAttr	0.3500	0.3999	0.4152	0.4212	0.4265

Table 2: AutoAIS scores for bm25, gtr and MedAttr on HOSPITAL A dataset and *MIMIC-QA* dataset

In Table 2, we present AutoAIS scores on the MIMIC and HOSPITAL A datasets.

On the HOSPITAL A dataset, **MedAttr** achieves comparable Top-5 attribution accuracy (67.7%) to both GTR and BM25, while showing marginal improvements in more complex QA tasks (*e.g.*, MIMIC), validating its robustness across structured and unstructured clinical data.

BM25 has improved performance on this dataset, which is because of the presence of tables, where keywords and entities matter more and not contextual semantic information. Hence, sparse retrievers perform better in such settings.

On MIMIC, MedAttr slightly outperforms both baselines at Top-5 (42.65%), reflecting its consistent advantage in deeper retrieval settings.

These results show that MedAttr remains effective across both structured and unstructured clinical data, even in high-stakes environments like EHRs.

The results showing diversity results are shown in Table 5 and ablation results are shown in section B.2.

7. Limitations

This approach is more relevant in scenarios where the queries are ambiguous, multiple answers are present or to avoid redundancy. For single hop queries this might add retrieval overhead, hence increasing the time taken for attribution.

References

Diane Bartz and D Bartz. As chatgpt's popularity explodes, us lawmakers take an interest. reuters. *Reuters. Available online: https://www. reuters. com/technology/chatgpts-popularity-explodes-us-lawmakers-take-an-interest-2023-02-13/(accessed on 13 February 2023)*, 2023.

Bernd Bohnet, Vinh Q Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, et al. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*, 2022.

Petre Breazu and Napoleon Katsos. Chatgpt-4 as a journalist: Whose perspectives is it reproducing? *Discourse & Society*, 35(6):687–707, 2024.

Guendalina Caldarini, Sardar Jaf, and Kenneth McGarry. A literature survey of recent advances in chatbots. *Information*, 13(1):41, 2022.

Sumit Kumar Dam, Choong Seon Hong, Yu Qiao, and Chaoning Zhang. A complete survey on llm-based ai chatbots. *arXiv preprint arXiv:2406.16937*, 2024.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. *arXiv preprint arXiv:2401.08281*, 2024.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, et al. Rarr: Researching and revising what language models say, using language models. *arXiv preprint arXiv:2210.08726*, 2022.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*, 2023.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. True: Re-evaluating factual consistency evaluation. *arXiv preprint arXiv:2204.04991*, 2022.

Minghao Hu, Yuxing Peng, Zhen Huang, and Dongsheng Li. Retrieve, read, rerank: Towards end-to-end multi-document reading comprehension. *arXiv preprint arXiv:1906.04618*, 2019.

Lei Huang, Xiaocheng Feng, Weitao Ma, Liang Zhao, Yuchun Fan, Weihong Zhong, Dongliang Xu, Qing Yang, Hongtao Liu, and Bing Qin. Advancing large language model attribution through self-improving. *arXiv preprint arXiv:2410.13298*, 2024.

Siqing Huo, Negar Arabzadeh, and Charles Clarke. Retrieving supporting evidence for generative question answering. In *Proceedings of the Annual International ACM SIGIR*

Conference on Research and Development in Information Retrieval in the Asia Pacific Region, pages 11–20, 2023.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*, 2020.

R Jegadeesan, Dava Srinivas, N Umapathi, G Karthick, and N Venkateswaran. Personal healthcare chatbot for medical suggestions using artificial intelligence and machine learning. *European Chemical Bulletin*, 12(3):6004–6012, 2023.

Krishnateja Killamsetty, Sivasubramanian Durga, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. Grad-match: Gradient matching based data subset selection for efficient deep model training. In *International Conference on Machine Learning*, pages 5464–5474. PMLR, 2021a.

Krishnateja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, and Rishabh Iyer. Glister: Generalization based data subset selection for efficient and robust learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8110–8118, 2021b.

Felipe C Kitamura. Chatgpt is shaping the future of medical writing but still requires human judgment, 2023.

Suraj Kothawade, Vishal Kaushal, Ganesh Ramakrishnan, Jeff Bilmes, and Rishabh Iyer. Submodular mutual information for targeted data subset selection, 2021a. URL https://arxiv.org/abs/2105.00043.

Suraj Kothawade, Anmol Mekala, Mayank Kothyari, Rishabh Iyer, Ganesh Ramakrishnan, Preethi Jyothi, et al. Ditto: Data-efficient and fair targeted subset selection for asr accent adaptation. *arXiv preprint arXiv:2110.04908*, 2021b.

Dhruv Kudale, Badri Vishal Kasuba, Venkatapathy Subramanian, Parag Chaudhuri, and Ganesh Ramakrishnan. *SPRINT: Script-agnostic Structure Recognition in Tables*, page 350–367. Springer Nature Switzerland, 2024. ISBN 9783031705496. doi: 10.1007/978-3-031-70549-6_21. URL http://dx.doi.org/10.1007/978-3-031-70549-6_21.

Liliana Laranjo, Adam G Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie YS Lau, et al. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, 25(9):1248–1258, 2018.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, pages 6950–6960. PMLR, 2020.

Priya Mishra, Suraj Racha, Kaustubh Ponkshe, Adit Akarsh, and Ganesh Ramakrishnan. Guideq: Framework for guided questioning for progressive informational collection and classification. *arXiv preprint arXiv:2411.05991*, 2024.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y Zhao, Yi Luan, Keith B Hall, Ming-Wei Chang, et al. Large dual encoders are generalizable retrievers. *arXiv preprint arXiv:2112.07899*, 2021.

Kyosuke Nishida, Itsumi Saito, Atsushi Otsuka, Hisako Asano, and Junji Tomita. Retrieve-and-read: Multi-task learning of information retrieval and reading comprehension. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 647–656, 2018.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840, 2023.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*, 2020.

Ferdinand Schlatt, Maik Fröbe, Harrisen Scells, Shengyao Zhuang, Bevan Koopman, Guido Zuccon, Benno Stein, Martin Potthast, and Matthias Hagen. Set-encoder: Permutation-invariant inter-passage attention for listwise passage re-ranking with cross-encoders. *arXiv preprint arXiv:2404.06912*, 2024.

Kurt Shuster, Da Ju, Stephen Roller, Emily Dinan, Y-Lan Boureau, and Jason Weston. The dialogue dodecathlon: Open-domain knowledge and image grounded conversational agents. *arXiv preprint arXiv:1911.03768*, 2019.

R. Smith. An overview of the tesseract ocr engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633, 2007. doi: 10.1109/ICDAR.2007.4376991.

Brandon Smock and Rohith Pesala. Table Transformer, 06 2021. URL https://github.com/microsoft/table-transformer.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023.

Andrew Reyner Wibowo Tjiptomongsoguno, Audrey Chen, Hubert Michael Sanyoto, Edy Irwansyah, and Bayu Kanigoro. Medical chatbot techniques: a review. *Software Engineering Perspectives in Intelligent Systems: Proceedings of 4th Computational Methods in Systems and Software 2020, Vol. 1 4*, pages 346–356, 2020.

Andrew Trotman, Antti Puurula, and Blake Burgess. Improvements to bm25 and language models examined. In *Proceedings of the 19th Australasian Document Computing Symposium*, pages 58–65, 2014.

Rejin Varghese and Sambath M. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, pages 1–6, 2024. doi: 10.1109/ADICS58448.2024.10533619.

Jason Weston, Emily Dinan, and Alexander H Miller. Retrieve and refine: Improved sequence generation models for dialogue. *arXiv preprint arXiv:1808.04776*, 2018.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, 31(9):1833–1843, 2024.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona G Flores, Ying Zhang, et al. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records. *arXiv preprint arXiv:2203.03540*, 2022.

Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. Automatic evaluation of attribution by large language models. *arXiv preprint arXiv:2305.06311*, 2023.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*, 2023.

Shengyao Zhuang, Bing Liu, Bevan Koopman, and Guido Zuccon. Open-source large language models are strong zero-shot query likelihood models for document ranking. *arXiv preprint arXiv:2310.13243*, 2023.

Appendix A.

Dataset	Submodlib	Alpha	Label Match Accuracy				Text Match Accuracy			
			Top-1	Top-3	Top-5	Top-10	Top-1	Top-3	Top-5	Top-10
S2D	GC	0	0.167	0.350	0.417	0.533	0.000	0.033	0.033	0.033
		0.25	0.233	0.383	0.433	0.567	0.017	0.033	0.033	0.033
		0.65	0.367	0.433	0.450	0.600	0.067	0.100	0.117	0.183
		0.8	0.417	0.467	0.550	0.617	0.067	0.100	0.150	0.200
		0.95	0.450	0.533	0.583	0.683	0.050	0.117	0.217	0.250
	LogD	0	0.200	0.567	0.633	0.700	0.017	0.017	0.017	0.050
		0.25	0.367	0.637	0.667	0.717	0.067	0.083	0.083	0.250
		0.65	0.450	0.550	0.583	0.667	0.050	0.133	0.200	0.283
		0.8	0.450	0.550	0.583	0.667	0.050	0.133	0.217	0.283
		0.95	0.450	0.550	0.583	0.667	0.067	0.133	0.233	0.283
	FL	0	0.450	0.550	0.583	0.667	0.067	0.133	0.233	0.283
		0.25	0.450	0.550	0.583	0.667	0.067	0.133	0.233	0.283
		0.65	0.450	0.550	0.583	0.667	0.067	0.133	0.233	0.283
		0.8	0.450	0.550	0.583	0.667	0.067	0.133	0.233	0.283
		0.95	0.450	0.550	0.583	0.667	0.067	0.133	0.233	0.283
Human Stress	GC	0	0.338	0.613	0.725	0.817	0.014	0.021	0.028	0.056
		0.25	0.338	0.648	0.732	0.824	0.014	0.021	0.028	0.070
		0.65	0.331	0.655	0.768	0.838	0.042	0.070	0.085	0.106
		0.8	0.437	0.634	0.782	0.845	0.092	0.113	0.155	0.176
		0.95	0.458	0.606	0.704	0.838	0.106	0.148	0.155	0.211
	LogD	0	0.331	0.627	0.718	0.831	0.007	0.021	0.035	0.099
		0.25	0.324	0.704	0.768	0.859	0.028	0.113	0.155	0.211
		0.65	0.458	0.606	0.739	0.845	0.106	0.148	0.162	0.218
		0.8	0.465	0.627	0.718	0.852	0.106	0.148	0.162	0.218
		0.95	0.479	0.641	0.718	0.845	0.120	0.148	0.155	0.211
	FL	0	0.430	0.648	0.725	0.845	0.085	0.148	0.155	0.211
		0.25	0.430	0.648	0.725	0.845	0.085	0.148	0.155	0.211
		0.65	0.472	0.641	0.725	0.845	0.099	0.148	0.155	0.211
		0.8	0.472	0.641	0.718	0.845	0.099	0.148	0.155	0.211
		0.95	0.486	0.641	0.718	0.845	0.113	0.155	0.155	0.211

Table 3: Comparison of Exact label Match and Exact Text Match Accuracy for S2D and Human Stress Datasets. Submodular functions = {graph cut(GC), facility location(FL), log determinant (LogD)}. We show how controlling the relevance parameter α can achieve better results.

Quantitative & Clinical Performance Analysis on the S2D Dataset

The performance of three attribution strategies—MedAttr, GTR, and BM25—on the Symptom2Disease (S2D) dataset is examined in this section. Each sample consists of a clinical text detailing a wealth of symptoms accompanied by its ground truth label (i.e., a disease diagnosis). For each text, each technique produces three ascribed labels, which are then assessed for clinical validity and alignment with the actual diagnosis.

Methodology

From the S2D dataset, we selected five representative clinical samples that span a range of illnesses, including infectious, neurological, dermatological, and respiratory disorders. The attribution outputs from the three approaches (MedAttr, GTR, and BM25) were evaluated based on:

1. Exact label match with the ground truth.
2. Clinical plausibility of alternate attributions.
3. Relevance and consistency within the context of the presented symptoms.

Comparative Attribution Performance: MedAttr vs GTR vs BM25

Table 4: Comparative Attribution Performance on the S2D Dataset

Text ID	Ground Truth	MedAttr Match	GTR Match	BM25 Match	Clinical Comments
17	Chickenpox	YYY	YNN	NNN	MedAttr maintains strong dermatological focus; BM25 retrieves gastric/infectious terms unrelated to rash.
24	Cervical spondylosis	YYY	YYY	NNN	MedAttr & GTR correctly capture the neurological context. BM25 misattributes to systemic infections.
29	Bronchial Asthma	YYY	YYY	NNN	MedAttr/GTR capture the respiratory nature; BM25 errors suggest systemic metabolic conditions.
37	Pneumonia	YNY	YYN	NNN	MedAttr/GTR both reasonable; a mix of asthma and pneumonia is plausible. BM25 misfires toward GI issues.
54	Cervical spondylosis	YYY	YYY	NNN	MedAttr and GTR again consistent with neuromuscular symptoms; BM25 remains clinically irrelevant.

Clinical Interpretation and Insights

1. MedAttr Accuracy and Diversity

Compared to GTR and BM25, MedAttr demonstrates better clinical alignment and yields a wider range of context-aware attributions.

- **Text 17:** Relates chickenpox to skin rash.
- **Texts 24 and 54:** Back pain, phlegm, and limb weakness are correctly linked to cervical spondylosis.
- By recognizing symptom clusters (e.g., *cough + sputum* versus *chest discomfort + dyspnoea*), MedAttr distinguishes between bronchial asthma and pneumonia.

2. GTR as a Compromise

GTR yields reasonably accurate predictions but exhibits occasional semantic drift:

- **Text 17:** Incorrectly assigns a case of rash as impetigo, a disorder lacking systemic symptoms.
- In certain cases, GTR confuses asthma with bronchial asthma, which is clinically plausible.

3. Clinical Irrelevance of BM25

BM25 often returns irrelevant or off-topic attributions:

- **Text 17:** Erroneously assigns a case of rash to Peptic Ulcer Disease.
- **Text 29:** Recommends diabetes or psoriasis for respiratory distress.
- **Text 54:** Inappropriately classifies cases of cervical spondylosis as UTIs, despite the absence of urinary symptoms.

Final Clinical Perspective

Label correctness is only one aspect of attribution quality from a healthcare perspective. It also involves assessing the model's ability to understand complex clinical symptoms and avoid making harmful or misleading recommendations. This is particularly vital in high-stakes scenarios where mislabeling can cause detrimental care delays (e.g., pneumonia or cervical disorders). While GTR performs respectably, it lacks subtle differentiation. BM25, despite its strong lexical matching, fails to capture clinical context and may even be dangerous if used to assist in medical decision-making.

Method	s2d (top-10)	stress (top-3)
BM25	1.837	1.239
GTR	0.961	0.847
TSS	0.965	0.872

Table 5: Label entropy (in bits) of the top-k attributions for the s2d (top-10) and stress (top-3) datasets. Higher values indicate a more even spread of labels in the selected set. The TSS (Targeted Subset Selection) method optimizes a submodular objective that combines a relevance term with a diversity term under diminishing returns. This encourages it to retain highly relevant items (similar to GTR) while penalizing redundant selections, leading to slightly higher label entropy and more balanced coverage of labels compared to pure relevance-based selection.

Appendix B. Datasets

Symptom2Disease (S2D)

A comprehensive collection of symptom-disease mappings. The dataset includes symptom descriptions in plain language that are mapped to the appropriate disease classifications. Each item includes a ground-truth diagnosis along with textual information based on symptoms. This dataset was especially helpful for proving the quality of attribution in categorical label-based classification. Contains over 1200 symptom-to-disease mappings in textual format. We split it into splits of 969 entries for Train, 171 entries for Val, 60 entries for test to generate conversations and guiding information (Mishra et al., 2024). We explicitly assigned text segments to disease labels in order to mimic a conversational system configuration. In this instance, attribution helped ground model outputs in understandable reasoning by highlighting the most instructive symptom expressions associated with each anticipated condition.

Human Stress Prediction (Stress)

The Human Stress Prediction dataset includes data posted on subreddits related to mental health of individual's emotional and psychological states, with corresponding binary labels and subreddit community labels. Annotated with stress levels, this dataset includes narrative accounts of individual encounters. We used the GuideQ technique to divide long-form inputs into two equal halves in order to convert this dataset into a conversational style. The program then identified which textual segments had the most influence on the projected stress level by attributing classification judgements (such as stress category). This framework made it possible to explain emotional states in a way that was easier to understand and more conversationally based. The narrative form of the entries also made this dataset particularly suitable for evaluating attribution in emotionally sensitive domains. The dataset includes 2838 entries of personal diary-style notes, with annotations for different stress levels, where we split into 2,291 entries are for train, 405 entries are for val, and 142 entries are for test, and we have used the test data for attribution evaluations.

MIMIC-QA

The MIMIC-III clinical database, a popular benchmark in healthcare NLP, is the source of the MIMIC-QA dataset. To mimic patient questions and doctor answers, we generated QA pairs using clinical notes and discharge summaries from MIMIC. We have extracted 1503 QA pairs from 511 de-identified clinical notes.

Dataset Generation: Using an semi-automated pipeline, discharge summaries were parsed, and question-answer pairs were formulated. We ensured that the answers could be directly linked to the original note for high-fidelity attribution.

HOSPITAL A Medical Dataset (Private)

The HOSPITAL A dataset consists of medical document images, including prescriptions, diagnostic summaries, and discharge notes, sourced from a tertiary care academic hospital, considering that this data is image-based and unstructured. First, a robust OCR pipeline was used to preprocess scanned images and turn them into text. Concise, document-grounded QA pairs were produced by running the converted inputs through the model. Crucially, every response was guaranteed to be an exact copy of the original paper, which made traceability easier and decreased hallucinations. The QA pairs were refined for linguistic and factual clarity using a final correction procedure. Our attribution system was evaluated within clinical decision support scenarios using this dataset, offering a realistic, complex, and high-stakes environment for rigorous testing. 17 medical documents composed of 50 page images encompassing prescriptions, discharge summaries having 133 extractive question-answer pairs.

Dataset Generation: A semi-automated pipeline with human evaluation was designed to convert medical document images into extractive QA pairs using the Meta Llama-3.2-11B-Vision-Instruct model. This ensured answers were direct excerpts from the documents, supporting high-accuracy attribution.

B.1. Targeted Submodular Subset Selection

B.1.1. Graph Cut Mutual information

Given a set V, a subset A and a query set Q, graph cut mutual information is defined as follows:

$$I_f(A;Q) = 2\lambda \sum_{i\in A}\sum_{j\in Q} s_{ij}$$

where s_{ij} denotes the similarity between passage i and query j, and λ is a scaling factor.

Overall, this function focuses on summing the similarities between the selected subset A and the query set Q , serving as a measure of how relevant (or “mutually informative”) A is with respect to Q

B.1.2. Facility Location Mutual Information

Given a set V, a subset A and a query set Q, Facility Location mutual information is defined as follows:

$$I_f(A;Q) = \sum_{i \in V} \min\big(\max_{j \in A} s_{ij},\ \eta \max_{j \in Q} s_{ij}\big)$$

I_f can be viewed as a submodular function that captures the joint coverage of the ground set V and the query set Q.

B.1.3. Log Determinant Mutual Information

Given a ground set V, a selected subset $A \subseteq V$, and a query set Q, the LogD mutual information is computed over a similarity kernel matrix K constructed using pairwise similarities between items.
Formally, the function is defined as:

$$I_f(A;Q) = \log\det(I + K_{A \cup Q})$$

where $K_{A \cup Q}$ is the submatrix of the similarity matrix K restricted to the union of sets A and Q, and I is the identity matrix of appropriate dimensions.

B.2. Ablation for MedAttr

Table 3 shows the results of an ablation study on the S2D and Human Stress datasets, analyzing the impact of different submodular mutual information functions define in 3.3 and the relevance weight parameter α.
Effect of α: Increasing α starting from $\alpha = 0$, improves both the Label and Text Match metrics. For example, using LogD on S2D, Label Match increases from **63.3% ($\alpha = 0$)** to **68.3% ($\alpha = 0.95$)** at Top-10. Most optimal results are acheived for α lying midway $(0,1)$. As $\alpha \to 1$, the model behaves similarly to GTR. Facility location mutual information's performance seems almost invariant to alpha.

Submodular Function Comparison: Facility location mutual information exhibits stable and consistent performance across α values. In Human Stress, MedAttr using LogD with $\alpha = 0.25$ achieves the best Top-10 Label Match accuracy of **85.9%**.

These findings confirm that MedAttr performs better under larger budgets & at least as good as a pure retrieval-based method and benefits from relevance-focused selection. This makes MedAttr both effective and adaptable for different medical dialogue and QA contexts. Further qualitative analysis has been performed and is present A.