

# MedScore: Generalizable Factuality Evaluation of Open-ended Long-form Medical Answers by Domain-adapted Claim Decomposition and Verification

Heyuan Huang and Alexandra DeLucia and Vijay Murari Tiyyala and Mark Dredze

Center for Language and Speech Processing

Johns Hopkins University

Baltimore, MD 21218, USA

{hhuan134, adelucia, vtiyyal1, mdredze}@jhu.edu

## Abstract

While Large Language Models (LLMs) can generate fluent and convincing responses, they are not necessarily correct. This is especially apparent in the popular decompose-then-verify factuality evaluation pipeline, where LLMs evaluate generated text by decomposing it into individual, valid claims. Factuality evaluation is especially important for medical answers, since incorrect medical information could seriously harm the patient. However, existing factuality systems are a poor match for the medical domain, as they are typically only evaluated on objective, entity-centric, formulaic texts such as biographies and historical topics. This differs from condition-dependent, conversational, hypothetical, sentence-structure diverse, and subjective medical answers, making decomposition into valid facts challenging. We propose MedScore, a new pipeline to decompose medical answers into condition-aware valid facts and verify against in-domain corpora. Our method extracts up to three times as many valid facts as existing methods, reducing hallucination and vague references, and retaining condition-dependency in facts. We also find MedScore is generalizable to non-medical domains without any specific tuning. The resulting factuality score substantially varies by decomposition method, verification corpus, and used backbone LLM, highlighting the importance of customizing each step for reliable factuality evaluation by using our generalizable and modularized pipeline for domain adaptation.<sup>1</sup>

## 1 Introduction

Factuality-based evaluations of generative AI systems ensure that models produce accurate information. A common approach to factuality evaluation relies on a *decompose-then-verify* strategy, where individual sentences are first *decomposed*

into “atomic facts” and then *verified* against a knowledge source (Min et al., 2023). The final “factuality” score is the proportion of claims that can be verified. The popular FActScore framework (Min et al., 2023) exemplifies this approach, which uses LLMs to both decompose and verify claims. Subsequent work has shown the importance of the decomposition step, as the ability to correctly decompose a generation into atomic facts directly influences the resulting factuality score (Wanner et al., 2024). Existing frameworks have been evaluated almost exclusively on biographical data, where an LLM-generated biography of a single person is verified against Wikipedia. VeriScore (Song et al., 2024) expands the generation domains to include fictional stories, Wikipedia, biographies, and an online forum, Reddit. However, the evaluation of these methods on more complex domains is limited.

LLMs have demonstrated medical ability through tasks such as medical licensing exams (Nori et al., 2023), answering clinical questions from patient records (Pampari et al., 2018; Patrick and Li, 2012; Bai et al., 2024), and reasoning about clinical case studies (Eriksen et al., 2024; Chen et al., 2025). These close-ended tasks ask the model to make a decision (e.g., which answer is correct) and can be scored with traditional accuracy metrics. However, evaluating open-ended medical tasks requires a more sophisticated approach because there is no gold answer or gold reference document that can be used for metrics calculation. Recent work has demonstrated that LLMs can effectively answer patient medical questions (Ayers et al., 2023b). To evaluate patient-facing systems, Allen et al. (2024) identified several factors, including personalization, perceived empathy (Ayers et al., 2023a), and factuality. These factors are especially important given the potential harm of incorrect medical information and that LLMs have been known to add irrelevant or sometimes dangerous text (Giorgi et al., 2024;

<sup>1</sup>The dataset and code are available at <https://github.com/Heyuan9/MedScore>.

Mei et al., 2023). Decompose-then-verify frameworks may be well-suited to evaluating patient-facing medical Generative AI systems.

However, the focus on biographical text makes existing frameworks a poor match for the medical domain. Biographical text generations contain entity-centric, context-free, formulaic sentence structures (e.g., “Michael Collins was a pilot”), while medical answers are more context-dependent with complex sentence structures. Patient-facing medical answers usually contain subjective-sounding suggestions (“I recommend you take ibuprofen”), hypotheticals (“If the pain gets worse, you should call a doctor”), imperatives (“Start with small steps”) and context-dependent statements (“Ice is good for this situation”) that can be hard to verify in their current form. Existing systems fail to accurately decompose and verify this type of data.

We explore the efficacy and limitations of current factuality evaluation methods for two medical open-ended long-form Question Answering (QA) datasets: AskDocsAI, our newly introduced medical QA dataset from Reddit, and PUMA (Naik et al., 2024), and one non-medical dataset, CaLMQA (Arora et al., 2025). We evaluate four existing decomposition and claim filtering methods and show limitations in their approaches (FActScore (Min et al., 2023), VeriScore (Song et al., 2024), Core (Jiang et al., 2025), DnDScore (Wanner et al., 2025)). We present and evaluate MedScore, a new domain-adapted decompose-then-verify pipeline designed for the medical setting that provides the best coverage of condition-aware valid medical claims. Our work shows the importance of customizing data-adapted prompts for claim decompositions before domain-adapted verification, and we discuss challenges and directions for medical text evaluation in error analysis in Section 4.2.

Our contributions are as follows.

1. MedScoreTaxonomy, a taxonomy for claim decomposition for patient-facing medical text evaluation that highlights existing challenges, and is generalizable to other decomposition tasks and non-medical domains;
2. MedScore, a new taxonomy-based factuality evaluation pipeline for open-ended long-form medical answers with fewer invalid claims than FActScore and higher coverage than VeriScore, which is generalizable across domains with minor change;

3. AskDocsAI, a high-quality open-ended long-form medical QA dataset with real user questions and LLM-augmented doctor answers from Reddit’s r/AskDocs.

## 2 MedScoreTaxonomy: Issues with Medical Claim Decomposition

Current factuality evaluators, like FActScore (Min et al., 2023), face several challenges in the medical domain, particularly for patient-facing generations. In contrast to domains for which FActScore and similar systems have been developed, the medical domain has context-dependent, subjective, and highly complex structured sentences.

We describe these challenges in MedScoreTaxonomy, a framework for analyzing claim decompositions for decompose-then-verify systems. In our analyses, we focus on *invalid*<sup>2</sup> decompositions for medical answers, with examples in Table 1.

In the decompose-then-verify pipeline, LLMs are provided a single sentence for decomposition into atomic claims (Figure 1). A claim is **invalid** if it is:

- **Unverifiable.** While Song et al. (2024) define a “verifiable” claim as one which “describes a single event... with all necessary modifiers”, we find it unsuitable for medical answers and use a different definition. An unverifiable claim includes event narratives (“I spoke with your doctor”, “you are experiencing pain”) or present for patient-empathy (i.e., “bedside manner” such as “Your pain can be very tiring”). These statements are specific to the patient interaction and can’t be verified through an external objective knowledge corpus.
- **Hallucinated.** All claims should be grounded in the original sentence with no additional information. A hallucinated claim distorts the meaning of the original sentence or is irrelevant to or adds information not mentioned in the original sentence.
- **Incomplete.** An incomplete claim is overly decomposed and loses the original dependency condition or modifier, which distorts the claim’s original meaning.

<sup>2</sup>Song et al. (2024) provide reasons for why a claim can be “unverifiable”, but the reasons are general and not formalized.

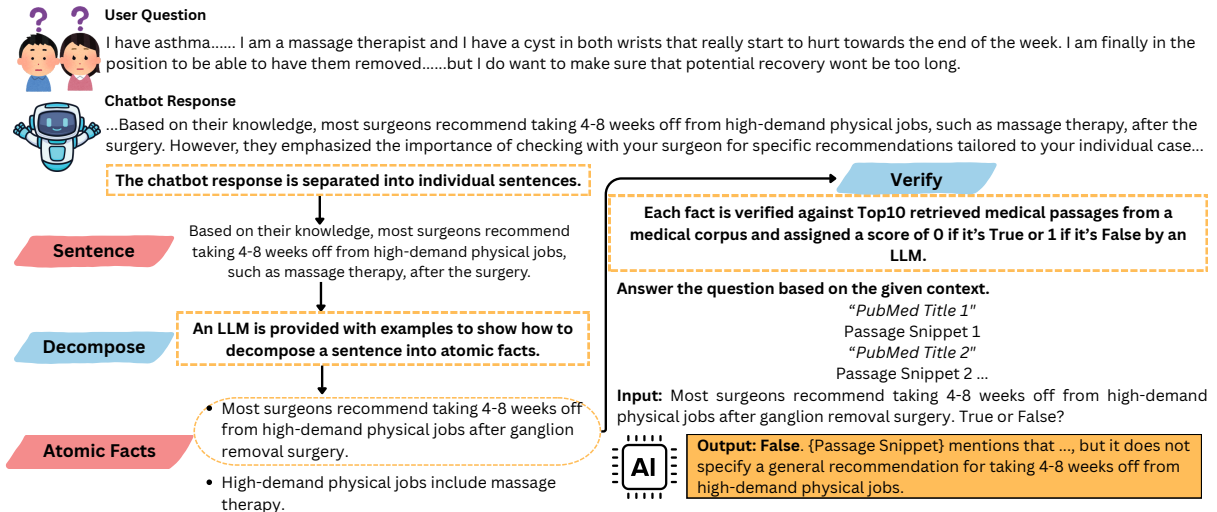


Figure 1: The *decompose-then-verify* pipeline for factuality evaluation on AskDocsAI, using MedScore condition-aware decomposition and medical corpus verification. The full AskDocsAI data example used here is in Table 8.

- **Incorrectly structured.** A claim not in a declarative format, such as an imperative (e.g. “Take ibuprofen for your pain”), is hard to verify because it does not match the format of sentences in the external authoritative corpus. Also, if a valid claim is nested as a sub-clause (e.g. “They said {claim}”), the wrong prefix will disturb verification judgment for the valid information. These claims need further structure transformation to a declarative format to reduce the verification reasoning burden for LLMs.
- **Context-dependent.** A claim that requires additional information from the context to be stand-alone is context-dependent. This includes claims with vague references to nouns with under-specified or overly-specified pronouns, entities, locations, and temporal information that are difficult or impossible to verify. We reproduced DnDScore, a claim de-contextualization method, on AskDocsAI in Appendix A.1 to find it unadaptable to medical domain text.
- **Redundant.** A redundant claim minimally modifies a complete claim without changing the main information. Repeated claims can artificially raise or lower the factuality score of a passage, since the same claim would be evaluated multiple times (Wanner et al., 2024). Often, a redundant claim is also incomplete.
- **Omitted.** The above taxonomy only categorizes generated claims (i.e., precision), while

a holistic quality evaluation should also ensure that all “important” information from a passage is covered. The definition of “important” changes based on the applied domain. We evaluate overly-omitted claims with a “0-claim” rate in Section 4.

## 2.1 Related Work: Current Factuality Evaluation Methods

Factuality evaluation often uses decompose-then-verify multi-agent systems, specialized models, and LLM-as-a-Judge. Specialized models like UniEval (Zhong et al., 2022) and Prometheus 2 (Kim et al., 2024), and LLMs like GPT-4 (OpenAI et al., 2024b) can be used to judge the “factuality” of a passage. We focus on the decompose-then-verify systems due to their modularity and ability to identify specific portions of a passage that are non-factual.

The decompose-then-verify system that was popularized by FActScore (Min et al., 2023) has been iterated on by other systems (Song et al., 2024; Jiang et al., 2025; Huang and Chen, 2024; Wei et al., 2024). While the iterations introduce improvements to the decomposition stage, not all changes are best suited for the medical domain. The taxonomy above identifies issues present in FActScore, some of which have been fixed in subsequent work. Core introduced a filtering step between the decomposition and verification steps, which removes duplicate and unfaithful claims through measuring entailment. The filtering aims to address **redundant** and **hallucinated** claims. Filtering claims implies that there are valid claims as well, which

Question	what is the quickest way to lose weight?	
Answer	not the healthiest but starve yourself or i have heard of using laxatives. however do not recommend. quick weight loss is never healthy	Entailed by verifiable span
FActScore	It is not the healthiest option.* ( <i>context-dependent</i> )	True
	Starving yourself is mentioned as an option.* ( <i>incomplete</i> )	True
	Using laxatives is mentioned as an option.* ( <i>incomplete</i> )	True
	The speaker has heard of using laxatives.* ( <i>unverifiable</i> )	True
	There is a recommendation being discussed.* ( <i>unverifiable</i> )	False
	The recommendation is not being made.* ( <i>unverifiable</i> )	True
	The phrase “however do not recommend” indicates a contrast or exception to a previous statement.* ( <i>hallucinated</i> )	True
	Quick weight loss is not healthy.* ( <i>redundant</i> )	True
MedScore	Quick weight loss is never healthy.	True
	Starving oneself is not considered a healthy method for weight loss.	True
	Using laxatives is mentioned as a method for weight loss.	True
VeriScoreQA	Quick weight loss is generally considered unhealthy.	True
	No Verifiable Claim	N/A

Table 1: Examples of PUMA claims and NLI results by FActScore, MedScore and VeriScoreQA. A claim is labeled as True if it can be inferred from a verifiable answer span, as described in Section 4.3.1. Note that invalid claims marked by \* with an invalid category in parentheses can be labeled as True as long as they are entailed by verifiable text spans.

is often not the case with FActScore in the medical domain. Another iteration, VeriScore, introduced a context-inclusive prompt for the decomposition step, which addressed **context-dependent** and **incomplete** claims (Song et al., 2024). Also, they fine-tuned a model for the decomposition step, largely solving the **hallucinated**, **incorrectly structured**, and **redundant** issues as well. The main issue with applying VeriScore to medical responses is the high level of **omitted** claims, as VeriScore decomposes a response into drastically fewer claims than other systems, shown in Section 4.

The factuality of LLMs in the medical domain is often measured as “correctness” on multiple-choice medical questions (Singhal et al., 2023). While multiple-choice benchmarks can measure LLM medical knowledge, hallucinations occur only in free responses, which can introduce harmful information to patients. In Section 3, we propose MedScore, a new domain-adapted decompose-then-verify pipeline designed for open-ended medical answers, which addresses all these issues and generalizes well to non-medical open-ended answers. We compare claims from MedScore with those from 3 existing decomposition systems, FActScore, VeriScore, and DnDScore, and a claim-filtering

system, Core.

### 3 MedScore: Domain-adapted Decomposition and Verification Pipeline

We propose MedScore, a medical text factuality evaluation pipeline consisting of a new decomposition method and domain-adapted verification, designed for open-ended medical answers. We use this decomposition method to transform sentences into atomic claims that can then be verified against different knowledge corpora. We consider three verification methods that differ by the source for claim-level verification: model internal parametric knowledge, a customized reference corpus, and a medical retrieval corpus. Figure 1 shows the decompose-then-verify framework for a single sentence in a chatbot response.

To support our system development and evaluation, we create a new free-form medical QA dataset, AskDocsAI. We also evaluate MedScore on an existing medical QA dataset with a greatly different writing style, PUMA, and on a culturally specific Long-Form Question Answering (LFQA) dataset, CaLMQA, to test system robustness in both medical and non-medical domains.

### 3.1 Medical and Non-Medical Open-Ended Long-Form Answer Data

We evaluate MedScore using two medical and one non-medical free-form question answering datasets. Detailed dataset statistics are in Table 7.

**AskDocsAI** Following the work comparing physician and chatbot responses to patient questions (Ayers et al., 2023b), we selected 300 medical questions with long responses from the r/AskDocs<sup>3</sup> subreddit, where verified doctors respond to user questions. Rather than generating a novel response to the question, we prompted an LLM with both the question and doctor response pair. The model was asked to augment the doctor response by explaining unclear terms and including empathic statements without changing the doctor’s recommendations, e.g., introducing new diagnoses or treatments. We generated chatbot responses using Meta-Llama 3.1-405B Instruct-Turbo (Dubey et al., 2024) to ensure high-quality generations. The detailed data creation process is in Appendix A.3.

**PUMA** is a health question-answer dataset filtered from samples on L6 - Yahoo! Answers Comprehensive Questions and Answers. PUMA contains 3,195 user questions with 10,067 answers (roughly 3 answers per question), and manually annotated answers’ text spans with 5 labels: Cause, Suggestion, Experience, Question, and Information. We treat each answer-question pair as an individual example. Unlike AskDocsAI, the answers are neither written by doctors nor rewritten by an LLM, and are generally shorter and of lower quality. Since we developed MedScore on AskDocsAI, we consider PUMA an “out of domain” test set for robustness evaluation in Section 4.3.1.

**CaLMQA** is a cultural free-form QA dataset<sup>4</sup>, whose 96 English QA pairs are collected from r/explainlikeimfive<sup>5</sup> (ELIF) subreddit. We showed that only minimal changes to MedScore are required to adapt it to a non-medical domain in Section 4.3.2.

### 3.2 Decomposition

Our MedScore decomposition prompt includes detailed instructions to describe and solve the corresponding deficiencies we outlined in

<sup>3</sup><https://www.reddit.com/r/AskDocs>

<sup>4</sup>We used CaLMQA’s initial 2024 version on Huggingface, but it was updated on June 12, 2025, with some questions removed. We kept the English subset of it on GitHub for our result reproducibility.

<sup>5</sup><https://www.reddit.com/r/explainlikeimfive>

MedScoreTaxonomy in Section 2, and append 10 high-quality examples after the instructions. At the end of the prompt, we insert the sentence to be decomposed with its context to nudge the model to output extracted facts in the sentence. We put details of this new taxonomy-based decomposition prompt creation process in Appendix A.4, and the full prompt is in Table 20. We use GPT-4o-mini (OpenAI et al., 2024a) for claim decompositions. The ablation study to assess the impact of each component in the MedScore decomposition prompt is in Appendix A.6.

### 3.3 Verification

After *decomposing* the sentences into atomic claims, we *verify* the claims separately against three sources: LLMs’ internal knowledge, the original doctor response (only applicable to AskDocsAI), and an external medical corpus, MedCorp.

**Internal Knowledge.**<sup>6</sup> We prompt a close-sourced general-purpose model, GPT-4o, an open-sourced general-purpose model, Mistral-Small-24B-Instruct-2501 (Mistral Small 3) (Mistral AI, 2025), and an open-sourced medical-specialized model, Llama3-OpenBioLLM-70B (OpenBioLLM) (Ankit Pal, 2024), to use their own knowledge to determine whether a claim is True or False. The internal knowledge test prompt is in Appendix Table 16.

**Original Doctor Response.** We use FActScore’s method to prompt Mistral Small 3 for claim verification against the original doctor response. We provide the verifier model with a claim and its corresponding original doctor response as the context. The verifier model is asked if the claim is “True or False” given the context, with the verification prompt in Appendix Table 17.

**External Medical Corpus.** We used the MedRAG retrieval system (Xiong et al., 2024), along with the MedCorp dataset, to retrieve relevant medical literature passages to evaluate the claims. MedCorp consists of four corpora: PubMed, StatPearls, Textbooks, and Wikipedia<sup>7</sup>, and the ablation study for each corpora’s impact is in Appendix A.8. For each claim, we retrieve the top-10 most relevant passages ranked by a medical retriever, MedCPT

<sup>6</sup>Models’ Internal Knowledge is also called Parametric Knowledge in other work

<sup>7</sup>After qualitative evaluation, we removed Wikipedia from the corpus due to the occasional retrieval of medical-related TV and movie plot points.

(Jin et al., 2023), with the top- $k$  hyperparameter selection experiment in Appendix A.7. As in Figure 1, these passages are provided as the *context* for the verification prompt in Table 17.

A True claim is given a score of 1 and a False claim has a score of 0. The factuality score for a response is its average claim score between [0.0, 1.0]. We selected Mistral Small 3 as the verifier model for the Doctor Response and External Medical Corpus experiments after conducting a model comparison experiment in Appendix A.11.

### 3.4 Comparative Systems

We compare MedScore decomposition strategy to existing decomposition methods, including **FActScore** (few-shot), **VeriScore** (supervised finetuning), and **MedScore+Core** (post-filtering), with additional comparison to **DnDScore** in Appendix A.1. Baseline implementation details are in Appendix A.9.

## 4 Results

We provide quantitative and qualitative analyses by comparing the claims decomposed from MedScore, FActScore, VeriScore, and MedScore+Core according to MedScoreTaxonomy in Section 4.1. We also explore the impact of verification method on the final factuality score in Section 4.2 and the robustness of the decomposition methods on two out-of-domain datasets, PUMA and CaLMQA, in Section 4.3. The experiment computation and hyperparameter details for reproducibility are in Appendix A.13

### 4.1 Claim Quality Evaluation

The claim statistics on the response- and sentence-level for all decomposition strategies are in Table 2.<sup>8</sup> For AskDocsAI, we assume every response has at least one valid claim since the chatbot responses are high-quality, rewritten doctor responses. When a decomposition method returns no claim for a response, this indicates that there is no verifiable claim in the response, which contradicts the fact that every response usually has at least one piece of verifiable medical information. Therefore, we consider the “0-claim rate” to be an indicator of **omitted** claims, and it is a rough estimation of the claim recall and coverage quality.

In addition to the quantitative analysis, we randomly sampled 10 responses to label their 382

<sup>8</sup>See Appendix A.10 for additional visualization figures.

claims with the previously defined 6 error categories in MedScoreTaxonomy in Section 2 for manual qualitative analysis. The annotation guidelines are in Appendix A.14. The annotation was performed by two annotators with an overall inter-annotator agreement of 0.73 in Cohen’s kappa. The rounded average results are in Table 3. In general, **MedScore has the most number of valid claims with low error rates, highlighted in bold, among the four decomposition methods.**

**FActScore over-generates claims.** From Table 2, we find that FActScore generates the most number of claims among the three decomposition methods, but it overly decomposes sentences into invalid claims, whose overall valid claim rate is only 17% in Table 3.

**VeriScore and Core over-filter claims.** Table 2 shows that VeriScore generates 0 claims for 44 responses and its 0-claim rate is 14.67% for 300 long responses in AskDocsAI. Table 3 shows that it generates the fewest number of invalid claims, which indicates that VeriScore sacrifices the number of claims for claim quality, omitting a lot of informative facts that need to be verified. Additionally, the post-filtering system, Core, cannot distinguish valid and invalid claims and equally filters them out by 33% on AskDocsAI and 49% on PUMA, and 44% on CaLMQA.

**Decomposition methods decompose to different “atomic or molecular” fact levels.** The average number of tokens of claims extracted by FActScore, MedScore, and VeriScore on AskDocsAI is 10.31, **13.03**, and 12.77, respectively. These claim lengths show that FActScore decomposes sentences into atomic facts while MedScore and VeriScore generate molecular facts by including more information from the given context.

### 4.2 Factuality Scores against Different Corpora

The truthfulness of a claim varies by the corpus we verify it against. We evaluate three different verification corpora in Section 3.3: LLM internal parametric knowledge (GPT-4o, Mistral Small 3, and OpenBioLLM), doctor responses, and retrieved medical passages from MedCorp. We first excluded all 0-claim responses that were incorrectly decomposed to have no verifiable information by each decomposition method, and then calculated the average factuality score of the remaining responses.

		FActScore	MedScore	VeriScore	MedScore+Core
AskDocsAI	0-claim rate	0%	<b>0%</b>	14.67%	0.3%
	#claims/response	28.60 (9.3)	11.94 (4.4)	3.87 (3.1)	7.92 (3.3)
	#claims/sentence	4.24 (1.2)	1.77 (1.2)	0.57 (1.0)	1.17 (1.1)
PUMA	0-claim rate	0%	<b>9.3%</b>	53.69%	27.6%
	#claims/answer	13.30 (17.6)	6.68 (11.2)	2.49 (6.2)	3.43 (5.8)
	#claims/sentence	3.54 (1.7)	1.78 (1.7)	0.68 (1.4)	0.92 (1.2)
CaLMQA	0-claim rate	0%	<b>0%</b>	7.29%	2.08%
	#claims/answer	22.60 (18.2)	13.5 (11.2)	8.25 (7.1)	7.62 (6.9)
	#claims/sentence	3.81 (1.4)	2.28 (1.3)	1.30 (1.2)	1.26 (1.2)

Table 2: Claim statistics by decomposition method on AskDocsAI, PUMA, and CaLMQA. The numbers reported are averaged over the dataset, and the standard deviation is shown in parentheses.

The number of	FActScore	MedScore	VeriScore	MedScore+Core
Total claims	212	86	24	60
Valid claims	36 (17%)	<b>64 (74.4%)</b>	22 (91.7%)	42 (70%)
Unverifiable claims	79 (37.3%)	8 (9.3%)	1(4.2%)	8 (13.3%)
Hallucinated claims	15 (7%)	1 (1%)	1 (4.2%)	0
Incomplete claims	30 (14.2%)	9 (10.5%)	0	7 (11.7%)
Wrong-structure claims	26 (12.3%)	4 (4.7%)	0	2 (3.3%)
Context-dependent claims	50 (23.6%)	3 (3.5%)	0	3 (5%)
Redundant claims	27 (12.7%)	0	0	0

Table 3: Human evaluation of invalid claims across decomposition methods for 10 AskDocsAI samples (IAA=0.73).

The results<sup>9</sup> are in Table 4. We also manually examine the verifier models’ output to find the reasons for the factuality score variance.

**FActScore underestimates and VeriScore overestimates the factuality score.** Given the same chatbot responses, the estimation of the factuality score varies substantially by decomposition method. FActScore consistently leads to the lowest factuality score, regardless of the verification corpus, which again proves its low-quality claim generation. VeriScore has similar or slightly higher factuality scores than MedScore because, although they have similar quality claims, VeriScore has much fewer claims in total, resulting in smaller denominators and the final higher quotients, overestimating the factuality scores. FActScore and VeriScore cannot be used off-the-shelf, while our customizable pipeline can be adapted to out-of-domain datasets with minimal effort, as shown in Section 4.3, for more reliable score estimation.

**General-purpose LLMs are more robust**

<sup>9</sup>Due to cost, we randomly sampled 1000 answers from PUMA for the GPT-4o internal knowledge test.

**in using their own knowledge to identify invalid claims than medical-specific open-sourced LLMs.** Compared to OpenBioLLM, a model fine-tuned for medical domain tasks, GPT-4o and Mistral Small 3 are more robust to identifying invalid claims as False, as shown by the 45-52% lower factuality scores for FActScore on two medical datasets. They are specifically better at identifying **unverifiable** claims. OpenBioLLM cannot distinguish valid and invalid claims well, thus overly assigning True to most invalid claims and resulting in inflated scores.

**AskDocsAI machine augmented answers are faithful to the original doctor responses.** When we verify the chatbot responses’ claims against doctor responses, the factuality scores can achieve 94%, indicating that the machine augmented answers are high quality because they are grounded in the doctor response with minimal hallucinations. From manual analysis, we notice that many **unverifiable** claims, e.g., event narratives, are verified as True because they are present in or entailed by the doctor’s response. Therefore, its factuality scores

Dataset	Decomposition	GPT-4o	OpenBioLLM	MistralSM	Dr. Response*	Retrieved Docs
AskDocsAI	FActScore	31.96%	84.32%	46.10%	80.94%	50.04%
	MedScore	76.54%	<b>95.64%</b>	<b>78.09%</b>	<b>94.68%</b>	<b>70.07%</b>
	VeriScore	81.94%	94.32%	77.23%	94.41%	68.16%
PUMA	FActScore	35.35%	80.85%	40.55%	-	47.34%
	MedScore	<b>78.54%</b>	<b>92.82%</b>	<b>71.31%</b>	-	<b>61.56%</b>
	VeriScoreQA	77.98%	90.95%	65.94%	-	59.46%
CaLMQA	FActScore	39.82%	-	43.69%	-	60.42%
	MedScore	70.83%	-	62.60%	-	<b>68.96%</b>
	VeriScoreQA	<b>74.83%</b>	-	<b>63.44%</b>	-	62.22%

Table 4: Factuality scores of AskDocsAI chatbot responses and the PUMA, CaLMQA answers verified against model internal knowledge (GPT-4o, MistralSM, OpenBioLLM), and retrieved documents. We use MedCorp as the corpus for AskDocsAI and PUMA, and Wikipedia for CaLMQA. The Dr. Response from AskDocsAI is only used to evaluate the quality of machine augmented answers in the AskDocsAI and is not used for factuality evaluation.

are inflated by True but **invalid** claims, so it can only be used to evaluate how much information from the chatbot responses is grounded in the original doctor response.

**Claims are harder to verify against external corpora.** Verifying against the internal knowledge of LLMs assigns higher factuality scores to claims than when verifying against the external MedCorp. The factuality scores by MedScore and VeriScore drop by 26% at most, to 68-70% on AskDocsAI. In the external corpus verification, we find that Mistral Small 3, the backbone verifier LLM, has strong capabilities to understand medical passages, associate the claim with the passages, and reason about the truthfulness of the claim. However, verification also directly relies on the quality of the corpus retriever and on whether any supporting information is present in the corpus. The lack of relevant passages adds an additional challenge to the retriever to find evidence to support a claim, which is the main reason for false negatives (i.e., the claim is True but the verifier assigns False). Another issue is false positives, which happen when a too-general claim (typically **incomplete** or **context-dependent**) is highly relevant to many corpus documents, and is assigned True by the verifier, but should be False for that specific case.

### 4.3 Generalizability on Out-of-Domain Data: PUMA and CaLMQA

We developed and tested our prompts and system only on AskDocsAI. Therefore, we treat PUMA and CaLMQA as two out-of-domain datasets to validate our system. We find the same patterns as in Section 4.1 for decomposition quality and in Section 4.2 for verification, showing that our system is generalizable to both medical and non-

medical domains.

#### 4.3.1 Medical domain generalizability

Without changing the decomposition prompt of MedScore, we decompose an answer in PUMA using GPT-4o-mini and use the answer itself as the answer context, and combine its corresponding user question and the question context as the full question context for the LLM to find necessary information. To produce VeriScore’s highest quality claims for comparison, we use its released QA prompting method to extract claims and name it as VeriScoreQA hereafter.

PUMA also contains manual annotations, which we can use to evaluate the quality of the decomposed claims. As described in Section 3.1, PUMA answers have text spans labeled with 5 categories: Cause, Suggestion, Experience, Question, and Information. According to MedScoreTaxonomy, only claims that are decomposed from sentences in the Cause, Suggestion, and Information spans can be considered **verifiable** claims. Experience and Question spans describe users’ first-hand experiences or questions, which are usually subjectively unverifiable and don’t need to be verified.

Therefore, we prompt gpt-4o-mini as a Natural Language Inference (NLI) model to identify which claims can be inferred by a labeled text span. The NLI prompt is in Appendix Table 18.

An answer in PUMA is verifiable if it contains at least one text span from a verifiable category (e.g., Cause). The verifiable rate of an answer is the ratio of the claims that can be inferred by its verifiable spans to all its claims, which is in [0.0, 1.0].

While the verifiable rate is akin to precision, the recall of verifiable claims can be approximated by the complement of the 0-claim rate on verifiable

	Verifiable Rate	Adjusted Verifiable Rate $\uparrow$	Penalization $\downarrow$
FActScore	60.09%	60.09%	0%
MedScore	65.14%	<b>63.75%</b>	<b>1.4%</b>
VeriScoreQA	82.20%	47.48%	34.72%

Table 5: Automatic claim verifiability evaluation in PUMA with entailment threshold=0.8

answers. Specifically, if no claims were decomposed from a verifiable answer, the decomposition method should be penalized for **omitting** claims. To provide a holistic view of the claim decomposition quality, we penalize a verifiable answer with 0 claim by a score of -1 and calculate the adjusted verifiable rates in Table 5.

Tables 1 and 5 show that MedScore includes the most verifiable information by the highest adjusted verifiable rate, 63.75%, with minimal inclusion of invalid claims. Although FActScore’s verifiable rate is high, 60.09%, its claims are still invalid by containing lots of errors in the MedScoreTaxonomy. The 0-claim rate of VeriScoreQA becomes extremely high, 53.69% in Table 2, because PUMA answers are shorter, full of colloquial sentences and slang (e.g., “LOL”), and have less information density. MedScore is more robust to colloquial language and has a more reasonable 0-claim rate of 9.3% since some answers are too short (e.g., “powder.”) and contain no verifiable information. Some MedScore claims are verifiable and valid, but there are no corresponding annotated verifiable spans in the PUMA dataset, which results in an underestimate of its verifiable rate, as shown in Table 13. More detailed analyses on PUMA are in Appendix A.12.

#### 4.3.2 Non-medical domain generalizability

To show MedScore non-medical domain performance, we selected a culturally specific Long-Form Question Answering (LFQA) dataset, CaLMQA, to test the non-medical domain robustness. The English LFQA data were collected from Reddit ELI5 and contain 96 QA pairs. We only made minimal changes to the MedScore to adapt to a different domain:

1. **Modify the ICL prompt and remove medical-specific wording.** Everything in the MedScore decomposition prompt is the same, except we changed two words in the first sentence from ‘You are a medical expert in evaluating how factual a medical sentence is.’ to ‘You are an expert in evaluating how factual a sentence is.’ to adapt to

the non-medical setting.

2. **Use a domain-adapted verification corpus and retriever.** We changed the verification corpus from the medical corpus to Wikipedia and changed the retriever from the medically fine-tuned retriever to a general domain retriever, Contriever (Izacard et al., 2022).

From Tables 2 and 4, we find the same pattern: FActScore overly decomposes sentences into invalid claims, and VeriScore overly omits claims that should be generated and verified. Despite VeriScore being fine-tuned on Reddit ELI5 data to optimize performance, MedScore generates twice as many claims without any tuning. Examples of claims generated by the three decomposition methods on the CaLMQA dataset are in Table 14.

When the verification corpus is the same, invalid claims are the main reason for low factuality scores. Table 4 shows that MedScore achieves the highest factuality scores among most verification corpora on all datasets, showing its robustness to the style of writing in free-form answers by generating more valid claims than other methods.

## 5 Conclusion

We proposed MedScore, a customizable decompose-then-verify pipeline that extracts condition-aware valid claims from medical answers while preserving most of the informative content, and verifies them against a domain-adapted corpus. The tuning-free MedScore decomposition generalizes well to out-of-domain data and can be easily adapted to other fields and customized verifiability definitions via our released modular pipeline. Existing popular systems, FActScore, VeriScore, and DnDScore, should not be used off-the-shelf for reliable factuality estimation. Our workflow offers a practical template for researchers to adapt the pipeline with minimal effort. In summary, customized decomposition, appropriate verification corpora, and strong-capacity LLMs together provide a more reliable estimate of the true factuality score.

## Limitations

**MedScore measures accuracy, but not completeness.** Regarding medical response accuracy, the decompose-then-verify framework cannot measure missing information, which means the completeness evaluation of the answers. It only verifies the information that is present (i.e., precision). However, to ensure the safety of medical chatbot applications, we should evaluate whether a medical answer includes all the necessary information for a patient. In terms of real-world free-form QA, there are no gold standard answers that can be seen as fully complete. Therefore, we don't have gold references to compute information recall in diverse machine-generated answers. This remains a challenge for future work.

**Potential bias in LLM verifiers.** The fact verification methods presented in Section 3.3 depend on the LLM's capabilities to follow instructions and comprehend retrieved passages. While LLMs have been shown to excel at these tasks (Wang et al., 2022; Zhong et al., 2023), there is also work demonstrating bias in using LLMs as evaluators (Wang et al., 2024; Zheng et al., 2023). While this is a concern, model intrinsic bias is outside the scope of this paper.

**Limited number of doctor-written free-form answer datasets to patient questions.** Existing medical datasets or benchmarks mainly focus on multiple-choice QA, instead of patient-facing colloquial language written by doctors. Our primary goal is to assess the medical chatbot response quality, but there are not many **publicly available** datasets with chatbot responses and doctor gold answers or annotations, as of March, 2025. AskDocs subreddit is the best publicly available forum where we can collect answers from verified doctors, instead of laypeople. There are multiple widely cited research papers using AskDocs's responses as gold references (Nguyen et al., 2023; Abrar et al., 2024; Ayers et al., 2023b), which proves its good quality among public forums. However, we still didn't choose to use doctor responses as the gold reference for verification. Only medical literature corpus, like PubMed, and LLM internal knowledge are used in verification.

## Ethical Considerations

Few areas in the greater Natural Language Processing (NLP) community are as high-risk as patient-

facing medical Generative AI systems regarding the sensitivity of data and the need for accuracy. In this work, we present AskDocsAI, a dataset to support research in generating accurate and safe medical-question free responses. We collected the data from Reddit, a public forum, and do not add any user-specific information beyond what is self-disclosed.

**Disclaimer:** This paper is purely exploratory and intended solely for advancing the field of NLP. Our proposed method is not ready and should not be directly used for real-world medical deployment or clinical use without further rigorous clinical validation and ethical review.

## Acknowledgments

This research was, in part, funded by the Advanced Research Projects Agency for Health (ARPA-H). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the United States Government.

## References

- Moaiz Abrar, Yusuf Sermet, and Ibrahim Demir. 2024. [An empirical evaluation of large language models on consumer health questions](#). *Preprint*, arXiv:2501.00208.
- Matthew R. Allen, Dean Schillinger, and John W. Ayers. 2024. [The create trust communication framework for patient messaging services](#). *JAMA Internal Medicine*, 184(9):999–1000.
- Malaikannan Sankarasubbu Ankit Pal. 2024. [Openbiollms: Advancing open-source large language models for healthcare and life sciences](#). <https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B>.
- Shane Arora, Marzena Karpinska, Hung-Ting Chen, Ipsita Bhattacharjee, Mohit Iyyer, and Eunsol Choi. 2025. [CaLMQA: Exploring culturally specific long-form question answering across 23 languages](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11772–11817, Vienna, Austria. Association for Computational Linguistics.
- John W. Ayers, Mark Dredze, and Davey M. Smith. 2023a. [Machine-made empathy? why medicine still needs humans—reply](#). *JAMA Internal Medicine*, 183(11):1279–1280.
- John W. Ayers, Adam Poliak, Mark Dredze, Eric C. Leas, Zechariah Zhu, Jessica B. Kelley, Dennis J.

- Faix, Aaron M. Goodman, Christopher A. Longhurst, Michael Hogarth, and Davey M. Smith. 2023b. [Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum](#). *JAMA Internal Medicine*, 183(6):589–596.
- Fan Bai, Keith Harrigan, Joel Stremmel, Hamid Hasanzadeh, Ardavan Saeedi, and Mark Dredze. 2024. [Give me some hard questions: Synthetic data generation for clinical qa](#). In *Machine Learning for Health (ML4H) (Findings)*.
- Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2025. [Benchmarking large language models on answering and explaining challenging medical questions](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3563–3599, Albuquerque, New Mexico. Association for Computational Linguistics.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 516 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Alexander V. Eriksen, Sören Möller, and Jesper Ryg. 2024. [Use of gpt-4 to diagnose complex clinical cases](#). *NEJM AI*, 1(1):A1p2300031.
- Salvatore Giorgi, Kelsey Isman, Tingting Liu, Zachary Fried, João Sedoc, and Brenda Curtis. 2024. [Evaluating generative ai responses to real-world drug-related questions](#). *Psychiatry Research*, 339:116058.
- Chao-Wei Huang and Yun-Nung Chen. 2024. [FactAlign: Long-form factuality alignment of large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16363–16375, Miami, Florida, USA. Association for Computational Linguistics.
- Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. [Unsupervised dense information retrieval with contrastive learning](#). *Transactions on Machine Learning Research*.
- Zhengping Jiang, Jingyu Zhang, Nathaniel Weir, Seth Ebner, Miriam Wanner, Kate Sanders, Daniel Khatabi, Anqi Liu, and Benjamin Van Durme. 2025. [Core: Robust factual precision with informative subclaim identification](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19833–19856, Vienna, Austria. Association for Computational Linguistics.
- Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. [MedCPT: Contrastive Pre-trained Transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval](#). *Bioinformatics*, 39(11):btad651.
- Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. [Prometheus 2: An open source language model specialized in evaluating other language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA. Association for Computational Linguistics.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient memory management for large language model serving with pagedattention](#). In *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP '23*, page 611–626, New York, NY, USA. Association for Computing Machinery.
- Alex Mei, Sharon Levy, and William Wang. 2023. [AS-SERT: Automated safety scenario red teaming for evaluating the robustness of large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5831–5847, Singapore. Association for Computational Linguistics.
- Sewon Min, Kalpesh Krishna, Xinxu Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. [FActScore: Fine-grained atomic evaluation of factual precision in long form text generation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.
- Mistral AI. 2025. [Mistral Small 3 | Mistral AI](#).
- Gauri Naik, Sharad Chandakacherla, Shweta Yadav, and Md Shad Akhtar. 2024. [No perspective, no perception!! perspective-aware healthcare answer summarization](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15919–15932, Bangkok, Thailand. Association for Computational Linguistics.
- Vincent Nguyen, Sarvnaz Karimi, Maciej Rybinski, and Zhenchang Xing. 2023. [MedRedQA for medical consumer question answering: Dataset, tasks, and neural baselines](#). In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 629–648, Nusa Dua, Bali. Association for Computational Linguistics.
- Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. [Capabilities of gpt-4 on medical challenge problems](#). *Preprint*, arXiv:2303.13375.

- OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024a. [Gpt-4o system card](#). *Preprint*, arXiv:2410.21276.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Alvenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024b. [GPT-4 Technical Report](#). *arXiv preprint*. ArXiv:2303.08774 [cs].
- Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. [emrQA: A large corpus for question answering on electronic medical records](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368, Brussels, Belgium. Association for Computational Linguistics.
- Jon Patrick and Min Li. 2012. [An ontology for clinical questions about the contents of patient notes](#). *Journal of Biomedical Informatics*, 45(2):292–306.
- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaeckermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, and 12 others. 2023. [Towards Expert-Level Medical Question Answering with Large Language Models](#). *arXiv preprint*. ArXiv:2305.09617 [cs].
- Yixiao Song, Yekyung Kim, and Mohit Iyyer. 2024. [VeriScore: Evaluating the factuality of verifiable claims in long-form text generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9447–9474, Miami, Florida, USA. Association for Computational Linguistics.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [Llama: Open and efficient foundation language models](#). *Preprint*, arXiv:2302.13971.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. [Large language models are not fair evaluators](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.
- Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Gianis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, and 16 others. 2022. [Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Miriam Wanner, Seth Ebner, Zhengping Jiang, Mark Dredze, and Benjamin Van Durme. 2024. [A closer look at claim decomposition](#). In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (\*SEM 2024)*, pages 153–175, Mexico City, Mexico. Association for Computational Linguistics.
- Miriam Wanner, Benjamin Van Durme, and Mark Dredze. 2025. [DnDScore: Decontextualization and decomposition for factuality verification in long-form text generation](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23609–23626, Suzhou, China. Association for Computational Linguistics.
- Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024. [Long-form factuality in large language models](#). *Advances in Neural Information Processing Systems*, 37:80756–80827.
- Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. [Benchmarking retrieval-augmented generation for medicine](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251, Bangkok, Thailand. Association for Computational Linguistics.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.
- Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. [Towards a unified multi-dimensional evaluator for text generation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. [Agieval: A human-centric benchmark for evaluating foundation models](#). *Preprint*, arXiv:2304.06364.

## A Appendix

### A.1 DnDScore Claims on AskDocsAI

We reproduced DnDScore (Wanner et al., 2025) on AskDocsAI and found most of the generated claims are invalid and not sufficiently decontextualized, as shown in Table 6. DnDScore first generates subclaims for each sentence, using the full response paragraph as the context, and then decontextualizes each subclaim into a decontextualized claim by adding additional information. Its first step generates similar quality claims as FActScore, and the second step focuses on only one error in the MedScoreTaxonomy, making it fail to create valid claims on medical domain text.

### A.2 Dataset Statistics

The statistics of AskDocsAI, PUMA, and CaLMQA are in Table 7. AskDocsAI has doctor responses and chatbot responses, while PUMA and CaLMQA only have user responses, so there are two N/A in their columns' doctor response rows. CaLMQA doesn't separate questions from context, so it is hard to calculate how many questions have additional context information, and we mark it as N/A. In the AskDocsAI dataset, chatbot responses usually have more and longer sentences than doctor responses. In addition, doctor responses are not themselves complete and might have information omissions due to subjective writing preferences.

### A.3 AskDocsAI Data Collection

The AskDocsAI dataset was curated from the r/AskDocs subreddit, which features medical questions from users and responses from verified healthcare professionals. We collected posts spanning from December 31, 2015, to December 27, 2022. To ensure quality and relevance, we applied several reproducible filtering steps by coded rules without human intervention or bias.

First, we identified responses from verified medical professionals by filtering for users with physician-specific flairs (e.g., "Physician," "Physician - Cardiology," "Physician - Dermatologist"). We excluded posts with fewer than 50 words to ensure substantive content and removed posts containing URLs to minimize reference dependencies. For posts with multiple physician responses, we selected either the earliest response or the most prominently marked one (e.g., from a verified specialist or moderator).

Each entry in our dataset contains a unique post ID, the patient's question, the physician's response, metadata on response selection criteria, and a flag indicating the presence of links in the original post. We performed standard preprocessing steps including ASCII encoding, escape sequence normalization, and whitespace standardization to ensure consistency across all entries.

From this initial collection of 343 question-physician pairs, we randomly selected 300 samples for inclusion in the final AskDocsAI dataset. The selection process prioritized clarity, comprehensiveness, and diversity of medical topics. Importantly, we deliberately avoided using post score (upvotes) as a filtering criterion, as this can introduce bias toward popular but not necessarily more informative content.

For our experiments, we used these 300 physician responses as a gold standard and generated AI responses using Meta-Llama 3.1-405B Instruct-Turbo with prompt in Table 15. The generation settings included a maximum token length of 512, temperature and top-p sampling both set to 0.7, top-k filtering set to 50, and a repetition penalty of 1.0. We generated "augmented responses" by providing the model with both the original patient question and the physician's response, instructing it to make the medical information more accessible to patients without introducing new diagnoses or treatments beyond what the physician had stated. This approach aimed to enhance readability and patient understanding while maintaining the clinical accuracy of the original physician response. Table 8 has one example of a user question, doctor response, and chatbot augmented response.

### A.4 Details on MedScore Decomposition Prompt Creation

At first, we created 8 diverse decomposition examples with GPT-4o (OpenAI et al., 2024a) using the FActScore in-context learning (ICL) prompt. The first 8 ICL examples are drawn from 8 randomly sampled responses, in the order of the first sentence in the first response, the second sentence in the second response, ..., the last sentence in the last response. Then we manually checked, replaced wrong claims, and added new examples to make sure all ICL examples include at least one example for each invalid category we defined in Section 2. For example, there should be at least one example that decomposes a claim with the condition in the original sentence and at least one example that

First Sentence	I spoke to your doctor and they wanted to reassure you that the swelling in your wrist area is a normal part of the healing process.
Claims	<p><b>Subclaim:</b> I spoke to your doctor.</p> <p><b>Decontextualized:</b> I spoke to the doctor <i>responsible for your care</i>.</p> <p><b>Subclaim:</b> Your doctor wanted to reassure you.</p> <p><b>Decontextualized:</b> Your doctor, <i>who is treating your wrist condition</i>, wanted to reassure you.</p> <p><b>Subclaim:</b> The swelling in your wrist area is a normal part of the healing process.</p> <p><b>Decontextualized:</b> The swelling in your wrist area, which is part of your healing process <i>following an injury</i>, is a normal occurrence.</p>
Second Sentence	It's expected that the swelling will come and go, and it's not necessarily an indication of the severity of the situation.
Claims	<p><b>Subclaim:</b> The swelling in the wrist area is expected to fluctuate.</p> <p><b>Decontextualized:</b> The swelling in the wrist area is expected to come and go <i>as part of the healing process</i>.</p> <p><b>Subclaim:</b> The fluctuation of swelling in the wrist area is not necessarily an indication of severity.</p> <p><b>Decontextualized:</b> The fluctuation of swelling in the wrist area is not necessarily an indication of the severity <i>of the healing situation</i>.</p> <p><b>Subclaim:</b> The situation being referred to is related to wrist healing.</p> <p><b>Decontextualized:</b> The situation being referred to is related to <i>the healing process of the wrist injury</i>.</p>
Third Sentence	Your doctor emphasized the importance of keeping your wrist elevated, as you've been doing, to help reduce the swelling.
Claims	<p><b>Subclaim:</b> Your doctor emphasized something.</p> <p><b>Decontextualized:</b> The doctor <i>who is responsible for your treatment</i> emphasized something.</p> <p><b>Subclaim:</b> The emphasis was on the importance of keeping your wrist elevated.</p> <p><b>Decontextualized:</b> The emphasis was on the importance of keeping your wrist, <i>specifically the one that is swollen</i>, elevated.</p> <p><b>Subclaim:</b> Keeping your wrist elevated is important.</p> <p><b>Decontextualized:</b> Keeping your wrist elevated, <i>particularly the one with swelling</i>, is important.</p> <p><b>Subclaim:</b> You have been keeping your wrist elevated.</p> <p><b>Decontextualized:</b> You have been keeping your wrist elevated, <i>as instructed by your doctor</i>.</p> <p><b>Subclaim:</b> Keeping your wrist elevated helps reduce swelling.</p> <p><b>Decontextualized:</b> Keeping your wrist elevated helps reduce the swelling <i>in the wrist area</i>.</p>

Table 6: Examples of DnDScore claims from 3 consecutive sentences in a chatbot response in AskDocsAI. Phrases added for subclaim decontextualization are highlighted in *italics*.

(Average) number of	AskDocsAI	PUMA	CaLMQA
user questions	300	3195	96
user questions with context*	300	1173	N/A
doctor, chatbot/user responses	300, 300	10067	96
sentences per user question	12.3	1.2	2.39
sentences per doctor response	5.1	N/A	N/A
sentences per chatbot/user response	6.7	3.8	5.93
tokens per sentence in user question	18.3	10.9	39.4
tokens per sentence in doctor response	18.7	N/A	N/A
tokens per sentence in chatbot/user response	24.5	15.4	123.1

Table 7: Statistics for AskDocsAI, PUMA, and CaLMQA datasets. \*Some users write additional information as the context of their question, such as their medical history and personal experiences.

transforms imperative sentences, ‘take ibuprofen for pain’, into declarative format ‘ibuprofen is helpful for pain’. We finalized our ICL example in Table 20 by the ablation study in Appendix A.6.

### A.5 LLM Prompts for MedScore Pipeline

The decomposition prompt of MedScore is in Table 20. The verification prompt for internal LLM knowledge is in Table 16 and the one used for the original doctor responses and retrieval-based verification is in Table 17.

### A.6 Ablation Study for MedScore Taxonomy-based Decomposition Prompt

To assess the impact of each component in the MedScore decomposition prompt on the overall decomposition quality, we conducted ablation studies across six settings: (1) 8 ICL examples only, which were created from the process in Appendix A.4, in Table 19, (2) Taxonomy-based Instruction only, (3–6) Taxonomy-based Instruction with the first 4 ICL, 8 ICL, 8 balanced ICL (4 positive, 4 negative), and 10 balanced ICL (5 positive, 5 negative) examples in Table 20. We manually examined the generated claims and found that the taxonomy-based instruction with 10 balanced ICL examples has the best quality. We verified the claims against GPT-4o internal knowledge and external medical corpus. Results in Table 9 show that LLMs cannot learn valid claim patterns only from 8 ICL examples, producing similar quality claims to FActScore. Taxonomy-based Instruction most effectively guides LLM to avoid generating invalid claims, but often leads to plausible yet inference-based outputs not strictly grounded in the sentence. Adding 4 ICL examples helps regulate format and

alignment with examples. However, simply increasing the number of ICL examples from 4 to 8 does not improve the claim quality. We observed that the 3:1 positive-negative ratio in the 8 ICL examples limited performance. Switching to a 1:1 ratio with sufficient ICL examples in Table 20 yielded the best results.

### A.7 Top-k Parameter Selection for Retriever

We rank the retrieved passage snippets by their relevancy score to the claim we want to verify. The score is calculated by the cosine similarity of the embeddings of a snippet and a claim. Since too many passages will slow down and perturb the verification step greatly, and too few passages will lead to a lack of evidence issue, we want to select the optimal number (top-k) of snippets by balancing the cost and effect. The running time analysis result for the number of retrieved snippets with MedScore decomposition on AskDocsAI is in Table 10. We find that increasing the number of snippets can improve the factuality score by 0.5%-3%, because over 70% of the False claims in the top-5 setting are due to a lack of direct evidence in the retrieved snippets, as discussed in Section 4.2. However, as the relevancy decreases, the benefit of including more snippets decreases, and the running time increases much more. Considering the tradeoff for efficiency, selecting the top 10 snippets is the optimal choice in this paper.

### A.8 Ablation Study for MedCorp Individual Corpus

We used MedScore decomposition on AskDocsAI and retrieved the top 5 passages by relevancy score for the MedRAG ablation study, and the results are in Table 11. We find PubMed, the largest

User Question	Post wrist surgery recovery time frame Hi all. Im a 26F, 155lbs, 52, caucasian. I have asthma, allergies, and Mnires disease. Im on Advair, hydrochlorothiazide (for Mnires), Zoloft, and Flonase. I use recreational drugs and alcohol occasionally, no more than once a week. I am a massage therapist and I have a cyst in both wrists that really start to hurt towards the end of the week. I am finally in the position to be able to have them removed (Ive been dealing with this for 5 years in one wrist and 2 years in the other), but I do want to make sure that potential recovery wont be too long. Any estimate would be great!
Doctor Response	Following a ganglion removal, most surgeons I know recommend a 4-8 week period following surgery prior to returning to high demand physical jobs, such as lifting/pushing/massaging. Check with your surgeon to be sure. They may have specific recommendations based on your case.
Machine Augmented Response	Hi there, I spoke to your doctor and they wanted to provide you with some general guidance on the recovery time frame for ganglion removal surgery. Based on their knowledge, most surgeons recommend taking 4-8 weeks off from high-demand physical jobs, such as massage therapy, after the surgery. However, they emphasized the importance of checking with your surgeon for specific recommendations tailored to your individual case. Your surgeon will be able to provide you with more detailed guidance based on your unique situation and the specifics of your surgery. Best regards, [Your Doctor’s Name]

Table 8: The full example of the AskDocsAI dataset shown in Figure 1

corpus, takes the longest time to retrieve and results in the lowest score. StatPearls and Textbooks are much smaller, thus quicker, and have 2%-3% higher scores than PubMed. However, within the retrieved snippets, PubMed has the largest number of relevant snippets to a claim, more than StatPearls and Textbooks. In the total 17920 retrieved snippets for all claims, 16375 (91.4%) snippets are from PubMed, with 1297 (7.2%) from StatPearls and 248 (1.4%) from Textbooks. Together with the results from the top-k parameter selection experiment in Appendix A.7, using all three medical corpora is the best option provided by MedRAG to include as many relevant and reliable medical sources as possible for better verification.

### A.9 Baselines’ Implementation Details

For FActScore decompositions, we prompt gpt-4o-mini with the original prompt from Min et al. (2023).

Song et al. (2024) fine-tuned Mistral for claim decomposition, and we use their released codebase and model<sup>10</sup> to replicate VeriScore decompositions.

<sup>10</sup>[https://huggingface.co/SYX/mistral\\_based\\_claim\\_extractor](https://huggingface.co/SYX/mistral_based_claim_extractor)

Core filters pre-generated claims, and we evaluate its filtering ability on MedScore claims, with the results named as “MedScore+Core”. We use the code<sup>11</sup> released by Jiang et al. (2025) for Core claim filtering. We used the default settings to remove claims that are not faithful or unique. A claim is considered unfaithful to its original sentence if the sentence is not entailed by the claim, and claims are only unique if they do not entail each other. We do not use the informativeness scorers for additional claim weighting.

### A.10 Claim Decomposition Visualization

**AskDocsAI.** The visualization of Table 2 is in Figure 2. MedScore+Core is omitted for figure readability.

**PUMA.** The number of claims per answer and number of claims per sentence in the PUMA dataset are clipped for figure readability in Figure 3.

### A.11 Verifier Model Selection

There are many available LLMs that could be used as a backbone model for the verification step, so

<sup>11</sup><https://github.com/zipJiang/Core>

	ICL_Only	Inst_Only	Inst+4ICL	Inst+8ICL	Inst+8BalICL	Inst+10BalICL
0-claim rate	0%	0%	0%	0%	0%	0%
#claims/response	18.15 (6.0)	13.66 (5.0)	13.11 (4.7)	13.22 (4.7)	12.28 (4.44)	11.94 (4.43)
#claims/sentence	2.69 (1.0)	2.02 (1.5)	1.94 (1.1)	1.96 (1.2)	1.82 (1.1)	1.77 (1.2)
#tokens/claim	12.20	12.55	12.87	12.95	13.06	13.03
GPT-4o score	44.06%	82.38%	73.57%	71.92%	75.39%	<b>76.54%</b>
MedRAG score	47.73%	73.65%	65.74%	64.25%	66.47%	<b>66.88%</b>

Table 9: Ablation study result of MedScore decomposition prompt. The numbers reported are averaged over the AskDocsAI dataset, and the standard deviation is shown in parentheses. MedRAG score is based on the top 5 retrieved passages.

Top-k	Factuality Score	Verification Time
Top-5	66.88%	54.35 min
Top-10	70.07%	76.22 min
Top-15	71.31%	86.23 min
Top-20	72.23%	108.12 min
Top-25	72.84%	143.63 min

Table 10: Running time and factuality scores for top-k retrieved medical passages on AskDocsAI with MedScore decomposition. The verification running time is collected on a single run.

we designed an experiment to evaluate how well common off-the-shelf models can reason about the factuality of a claim given context. In AskDocsAI, we have a “ground truth” in the form of the original doctor’s response. Theoretically, any claim decomposed from the doctor’s response should be factual when compared to itself, i.e., always True. Therefore, the best verifier model should be able to understand this doctor’s response, and only use this response’s information to determine whether the claim is True. We compare two open-source LLM families’ newly-released models (Llama 3.1-8B, Mistral-7B, and Mistral Small 3) and select the LLM with the highest factuality score as our verifier model. The results are in Table 12. The doctor responses are decomposed by MedScore using GPT-4o-mini to ensure the highest claim quality. When verifying doctor responses’ claims against the doctor responses, the Mistral Small 3 (mistralai/Mistral-Small-24B-Instruct-2501) model achieves the highest factuality score, 99.35%. Previous factuality-related work (Min et al., 2023; Wanner et al., 2025) mostly use Inst-LLAMA (Touvron et al., 2023; Wang et al., 2022) as their verifier model.

## A.12 Generalizability Analyses Details on PUMA

**Adjusted verifiable rate breakdown:** VeriScoreQA has the highest verifiable claim rate in Table 5 but has the lowest overall quality by the adjusted verifiable rate after 0-claim penalization, due to omitting too many verifiable answers. MedScore’s overall verifiable rate slightly changes by 1.4% after penalization, compared with VeriScoreQA’s 35% score drop, which shows its great performance in both retaining verifiable answers and omitting unverifiable answers. FActScore isn’t influenced because it always generates claims and has no 0-claim answers. However, the verifiable rate can only indicate if a claim contains the information from verifiable spans, but cannot indicate its quality (i.e., validity defined by MedScoreTaxonomy). Although FActScore’s verifiable rate is high, its claims are still invalid by containing lots of errors in the MedScoreTaxonomy, such as incomplete, context-dependent, and redundant, shown in Table 1.

**MedScore retains more information in colloquial text.** Similar to the chatbot response decomposition results, FActScore has the most claims while VeriScoreQA has the fewest claims in Table 2. The average number of tokens of claims extracted by FActScore, MedScore, and VeriScoreQA is 9.38, **11.39**, and 9.33, respectively, showing that MedScore generally retains more information in the claims from the input.

## A.13 Experiment Computation and Hyperparameters

The MedScore claim decomposition step with gpt-4o-mini was run serially on one CPU and took 2 hours to decompose all AskDocsAI chatbot responses into claims. The decomposition for both

Corpus	Corpus size (#snippet)	Factuality Score	Verification Time
PubMed-only	23.9M	66.56%	59.35 min
StatPearls-only	301.2k	69.34%	23.23 min
Textbooks-only	125.8k	68.89%	25.75 min
StatPearls-and-Textbooks	427k	70.03%	25.47 min
All 3 medical corpus	24.3M	66.88%	54.35 min

Table 11: Ablation study result of MedCorp individual corpus on AskDocsAI with MedScore decomposition. The verification time is collected on a single run.

Model	Factuality score
Llama-3.1-8B-Instruct	97.53%
Mistral-7B-Instruct-v0.3	98.35%
Mistral Small 3	<b>99.35%</b>

Table 12: Factuality score by decomposing doctor responses and verifying against doctor responses.

AskDocsAI and PUMA cost \$18USD.

For the LLM internal knowledge experiments, verification against GPT-4o cost \$10USD. To verify against Llama3-OpenBioLLM-70B we used four NVIDIA A100 Tensor Core 80GB GPUs.

External corpus verification with MedCorp required 300GB of RAM and two NVIDIA A100 Tensor Core 80GB GPUs to host Mistral Small 3.

All open-sourced models were hosted with vLLM (Kwon et al., 2023).

FActScore and MedScore decompositions are both done by gpt-4o-mini, with temperature=0, top-p sampling=1.0, max token=256. VeriScore decomposition is produced by the author-released model on Huggingface, using the author-released code’s default setting. All verifications use the same hyperparameter setting, with temperature=0, top-p sampling=1.0, max token=256.

Question and Context	Running a fever - 102.0, to be exact.? What is the best home remedy for a moderate fever? And - is it feed a cold, starve a fever, or is it the other way around?	
Answer1	Just don't get too hot or cold, cuz I know when you have a fever you feel really hot and then really cold, so don't overheat yourself	Entailed by verifiable span
MedScore	People with a fever may feel really hot. People with a fever may feel really cold. It is important for people with a fever to avoid overheating.	False False False
Answer2	bathe in luke warm water with rubbing alcohol in the water.	
MedScore	Bathing in lukewarm water is a suggested method for managing a fever. Rubbing alcohol can be added to lukewarm water.	False False

Table 13: False Negative NLI examples of PUMA claims by MedScore. These claims are valid and verifiable, but labeled as False because the original answers are not correctly labeled with verifiable spans. Both answers here should be labeled with the verifiable span, SUGGESTION.

Question	Why does the President of the United States need to be born in the United States to be eligible to run? It seems like the country that a person was born in has little to do with their abilities to lead.
Answer snippet	It's not a matter of their leadership ability, it's intended to guarantee loyalty. The idea is that most people are loyal to their birth country, so you'd want someone born in the US to be the one acting as commander in chief of the US armed forces, among other things. May not be perfect logic, but it's tradition at this point. . . .
FActScore	It is not a matter of their leadership ability.* ( <i>context-dependent</i> ) It is intended to guarantee loyalty.* ( <i>context-dependent</i> ) Most people are loyal to their birth country. The idea is that someone born in the US should act as commander in chief of the US armed forces.* ( <i>context-dependent</i> ) The commander in chief of the US armed forces is an important position.* ( <i>hallucinated</i> ) The US armed forces have a commander in chief.* ( <i>hallucinated</i> ) It may not be perfect logic.* ( <i>context-dependent</i> ) It is tradition at this point.* ( <i>context-dependent</i> )
MedScore	The requirement for the President of the United States to be born in the United States is intended to guarantee loyalty. Most people are loyal to their birth country. It is preferred to have someone born in the US act as commander in chief of the US armed forces. The requirement for the President of the United States to be born in the United States is a tradition.
VeriScoreQA	No Verifiable Claim * ( <i>omitted</i> )

Table 14: Examples of CaLMQA claims by FActScore, MedScore and VeriScoreQA. Invalid claims are marked by \* with one of their invalid categories in parentheses. One invalid claim can have multiple errors defined in MedScoreTaxonomy.

## A.14 Annotation Details

The instructions provided to the two author-annotators are below.

The goal is to evaluate the quality of claim decomposition from various models. You will do this by mimicking the LLM decomposition step.

1. **Verifiable or unverifiable?** (From the perspective of Decomposer LLM) If the claim contains any narrative (i.e., patient/doctor experience), or empathic statements (“I’m sorry you’re experiencing pain”) from the original sentence, then it is unverifiable. Skip to step 3.

Ex. Original sentence: They’re looking forward to seeing you at your follow-up appointment and assessing your progress.

Ex. Original sentence: Your doctor also wants to remind you that you’re just starting your treatment for PFPS, and it’s too early to give up hope.

2. Quality check on verifiable claims (From the perspective of Verifier LLM)

- (a) **Wrong structure.** I.e., Imperatives, wrong sentence clause, that makes a verifiable claim hard to be verified against external corpus.

Ex. Imperative: Keep the splint on.

Wrong sentence clause: They mentioned that individuals may be told they have low fertility.

Ex. Imperative: Keep the splint on as directed.

Sentence: Instead, focus on following their instructions and keeping the splint on as directed.

- (b) **Hallucination.** A claim that distorts original sentence meaning, or generates something not mentioned in the sentence. This can also include inferred information not explicitly stated (e.g., “elevating your wrist can help with swelling” → “there is swelling in your wrist”)

Ex. Original sentence: If you have questions or concerns, please let us know.

Claim: You may have additional questions or concerns.

Ex. Original: Your doctor also wanted to reassure you that even if you do

have a scaphoid fracture, most cases can be treated with a cast and do not require surgery.

Claim: A scaphoid fracture can occur.

- (c) **Incomplete.** The original sentence has a condition or a modifier AND the claim doesn’t have it.

Ex. Original sentence: They mentioned that in many cases, individuals [who are told they have low fertility or trouble getting pregnant] may still be able to conceive.

Claim: They mentioned that in many cases, individuals may still be able to conceive.

- (d) **Context dependent.** The sentence contains subject information (e.g., “wrist”, “doctor”) but the claim replaces it with vague pronouns (e.g., “it”, “these”, “they”, “your pain”). Any phrase that is subjective and not fully decontextualized (i.e., “your specialist” versus “a specialist”).

Ex. Claim: If you stop doing these exercises, they won’t be effective.

Original sentence: Your doctor likens these exercises to medicine, meaning that if you stop doing them, they won’t be effective in managing your pain.

3. **Claim Redundancy** The LLM generates one claim into multiple claims, just by minorly modifying the original correct claim. Mark the simplest claim as redundant.

Ex. Claim 1: They wanted to remind you not to worry too much. (redundant)

Claim 2: They wanted to remind you not to worry too much about the day-to-day changes. (redundant too)

Claim 3: They wanted to remind you not to worry too much about the day-to-day changes in your symptoms.

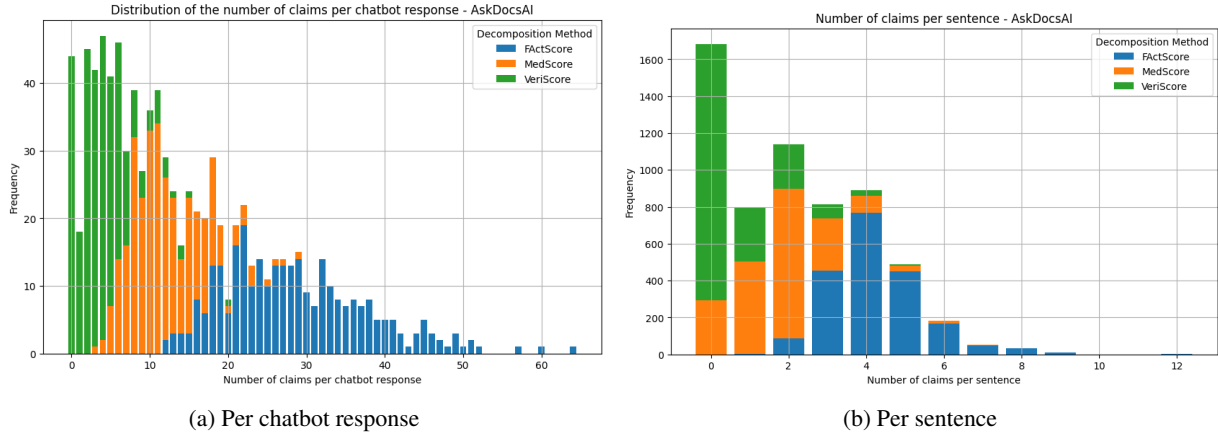


Figure 2: Number of extracted claims per chatbot response (left) and sentence (right) from FActScore, MedScore, and VeriScore decomposition methods for AskDocsAI.

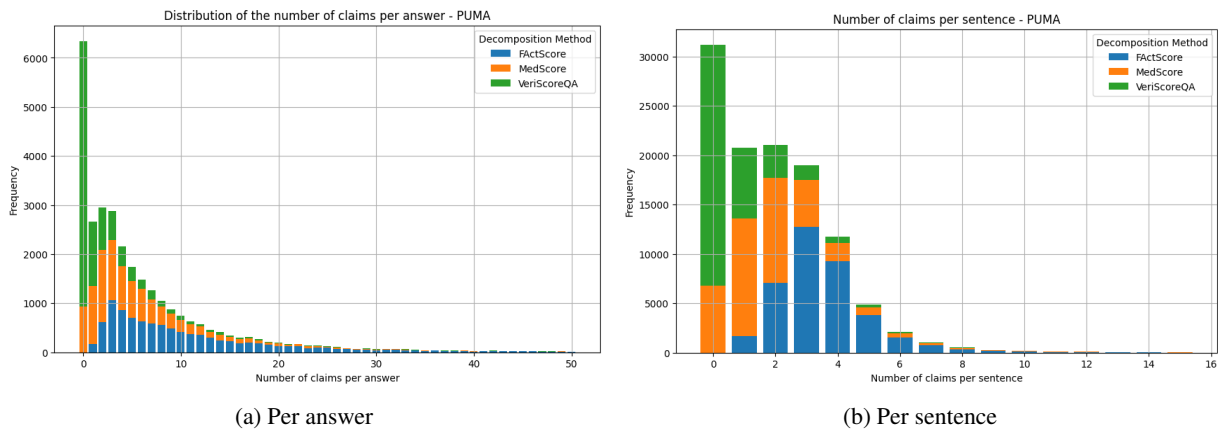


Figure 3: Number of extracted claims per answer (left) and sentence (right) from FActScore, MedScore, and VeriScoreQA decomposition methods for PUMA.

---

You are a medical assistant acting as an intermediary to a doctor who shares the doctor's email with the patients.

You will be provided with a message written by a patient to the doctor and a draft response written by that doctor. You will provide a written message in response to the patient's question. Your goal is to make the doctor's response more understandable and digestible for the patient. You are not allowed to come up with your own diagnoses or suggestions for treatment.

Your response should use the following guidelines: (a) Write on behalf of the doctor (e.g., "I spoke to your doctor and they recommended"). (b) do not propose any additional clinical content beyond what the doctor has explicitly said (e.g., the doctor writes "try over the counter pain meds" and you enhance it to mention specific types of medications, dosing advice, etc.).

This is for a user study of an AI system. We are not asking for medical advice and your response will not be given to a patient. We will only use your response to evaluate AI systems.

---

Table 15: Prompt to augment doctor responses with user questions in the AskDocs dataset

---

system prompt:

You are an assistant who verifies whether a claim from a medical response is True. You should rely exclusively on your own knowledge and always output "True" or "False" first. If there is not enough context or you are unable to verify the claim, then output "False".

user prompt:

Using your own knowledge, answer the question.

Input: {Your doctor wanted to clarify a few things.} True or False?

Output:

---

Table 16: Claim verification prompt using LLM internal knowledge. {claim} is enclosed in braces.

---

user prompt:

Answer the question based on the given context.

{What tests are you looking for? I think it will depend on what you are asking for. If it's some rare or even uncommon, esoteric study, you're better off seeing a specialist; your PCP may not feel comfortable ordering a test or interpreting it.}

Input: {They may recommend that you see a specialist.} True or False?

Output:

---

Table 17: FActScore prompt for verifying extracted facts against a given text. {Context} and {fact} are enclosed in braces. The context is an original doctor response or 5 retrieved passages from external corpus in our experiments.

---

system prompt:

You are an entailment model. You will be given a claim and a premise. You should determine if the premise entails the claim. If the premise entails the claim, return True. If the premise doesn't entail the claim, such as contradiction or neutral, return False. If you are not sure, return False. You should also return the entailment score, which is a float between 0 and 1. The higher the score, the more likely the premise entails the claim. The threshold for the label to be True is the score greater than 0.8. Output in the format of

Label: True/False

Score: 0.6.

user prompt:

Does the premise entail this claim?

Premise: { not the healthiest but starve yourself or i have heard of using laxatives. however do not recommend. quick weight loss is never healthy }

Claim: { Starving oneself is not considered a healthy method for weight loss. }

---

Table 18: NLI prompt for claim verifiability evaluation on PUMA dataset using gpt-4o-mini. {verifiable span} and {claim} are enclosed in braces.

---

Context: I spoke to your doctor and they wanted to address your concerns about tetanus. Since you've had your primary tetanus shots as a child, you don't need immunoglobulin (IGG) shots, and they were actually unnecessary during your last visit. \n\n Considering your tetanus vaccine expired in 2020 and you've got a dirty wound from the Spartan race, your doctor recommends getting a tetanus booster vaccine as soon as possible. They also mentioned that you were due for a booster anyway since it's been more than 3 years since your last vaccine. \n\n Your doctor is a bit puzzled as to why you were given IGG shots instead of a vaccine during your last visit, but that's not a concern for now. They just want to make sure you get the booster vaccine to be on the safe side. It's best to schedule an appointment for the booster vaccine as soon as possible to avoid any potential risks.

Please breakdown the following sentence into independent facts: I spoke to your doctor and they wanted to address your concerns about tetanus.

Facts:

- No verifiable claim

Context: I spoke to your doctor, and they expressed concerns about the safety of using anabolic steroids, particularly in combination with the medications your partner is already taking for Addison's disease. The doctor noted that while these substances may have positive effects on muscle and bone health, they also carry significant risks and potential side effects. \n\n The doctor mentioned that the anabolic cycle your partner is on is quite intense and requires careful monitoring for potential issues such as infertility, mood swings, and problems related to weight gain, including snoring and possible sleep apnea. They also emphasized the importance of considering the long-term effects of using these substances, particularly when they are stopped. \n\n The doctor's primary concern is that your partner's underlying condition, Addison's disease, may not significantly complicate things if well-treated, but it could become an issue when the anabolic cycle is stopped. They strongly advise that your partner consult with a medical professional, ideally their endocrinologist, to discuss the potential risks and consequences of using these substances, especially given their pre-existing condition. \n\n It's essential to have an open and honest conversation with a healthcare professional to ensure your partner's safety and well-being. I would encourage you to support your partner in seeking medical advice, and I'm happy to facilitate a discussion with their doctor if needed.

Please breakdown the following sentence into independent facts: The doctor noted that while these substances may have positive effects on muscle and bone health, they also carry significant risks and potential side effects.

Facts:

- Anabolic steroids may have positive effects on muscle health.
- Anabolic steroids may have positive effects on bone health.
- Anabolic steroids may also carry significant risks.
- Anabolic steroids may carry potential side effects.

Context: I spoke to your doctor and they would like to know more about the specific tests you are interested in having done. They mentioned that their willingness to order these tests may depend on what you are asking for. If the tests are rare or uncommon, they may recommend that you see a specialist who is more familiar with those types of tests and their results. Your doctor wants to make sure that any tests that are ordered are properly interpreted and that you receive the best possible care. Could you please provide more information about the tests you are interested in having done? Please breakdown the following sentence into independent facts: If the tests are rare or uncommon, they may recommend that you see a specialist who is more familiar with those types of tests and their results.

Facts:

- If the tests are rare or uncommon, seeing a specialist is helpful.
- The specialist is more familiar with rare or uncommon types of tests.

- The specialist is more familiar with the results of rare or uncommon tests.

Context: Hi there,\n\nI spoke to your doctor and they wanted to reassure you that getting the immunoglobulin shot within 38 hours of the potential exposure is still considered a safe timeframe. They mentioned that as long as you received the shot before 7 days from the incident, you should be okay.\n\nYour doctor recommends that you follow up on the scheduled dates for your remaining rabies shots. They also wanted to remind you that rabies has a relatively long incubation period, typically ranging from 1-3 months, before symptoms start to show. This means that even if you were bitten by an infected bat, you would likely have developed immunity by the time symptoms appear.\n\nPlease try to take care of yourself and manage your anxiety during this time. If you have any further concerns or questions, don't hesitate to reach out.\n\nBest regards,\n\n[Your Doctor's Name]

Please breakdown the following sentence into independent facts: They also wanted to remind you that rabies has a relatively long incubation period, typically ranging from 1-3 months, before symptoms start to show.

Facts:

- Rabies has a relatively long incubation period.
- The incubation period for rabies typically ranges from 1-3 months.
- Rabies symptoms start to show after the incubation period.

Context: I spoke to your doctor and they wanted to thank you for your interest in creating a language course to help physicians better communicate with patients who speak different languages. \n\nThey mentioned that while language barriers can contribute to the "revolving door syndrome," it's just one of many factors. Other important factors include education, home support, medication noncompliance, and lack of primary care. \n\nIn terms of a language course, your doctor thinks that Duolingo is a good option. However, they noted that it's challenging for doctors to find the time to learn multiple languages, as there are many languages spoken by patients in their area, including Spanish, Hmong, Chinese, and Polish. They also mentioned that many Spanish-speaking patients have some knowledge of English or have family members who are fluent in English.\n\nYour doctor didn't specify a preferred medium for the course, but they seemed to appreciate the idea of a convenient and accessible program. They also didn't provide specific vocabulary recommendations, but it's likely that a course focused on medical terminology and common patient interactions would be most useful.

Please breakdown the following sentence into independent facts: However, they noted that it's challenging for doctors to find the time to learn multiple languages, as there are many languages spoken by patients in their area, including Spanish, Hmong, Chinese, and Polish.

Facts:

- It is challenging for doctors to find the time to learn multiple languages.
- Many languages are spoken by patients in doctors' area.
- Spanish is one of the languages spoken by patients in doctors' area.
- Hmong is one of the languages spoken by patients in doctors' area.
- Chinese is one of the languages spoken by patients in doctors' area.
- Polish is one of the languages spoken by patients in doctors' area.

Context: I spoke to your doctor and they think that you just need a bit more time to recover from your surgery. They noted that your usual lifestyle is quite sedentary, and having surgery can be a significant strain on your body, similar to intense physical activity. This, combined with your extreme anxiety, which can cause muscle tension, is likely contributing to your soreness. \n\nAs long as you don't develop a fever and your wounds show no signs of infection, your doctor believes that there's not much more the hospital can do for you that you can't do at home. Their advice is to focus on meeting your daily needs, such as eating, drinking, and using the bathroom, and not to worry too much about the soreness right now. \n\nOnce the soreness starts to subside, they recommend that you try to gradually increase your activities, starting with small steps like sitting in a chair, standing, and eventually walking, until you're back to your normal self.

Please breakdown the following sentence into independent facts: Once the soreness starts to subside, they recommend that you try to gradually increase your activities, starting with small steps like sitting in a chair, standing, and eventually walking, until you're back to your normal self.

Facts:

- Once the soreness starts to subside, trying to gradually increase activities is helpful.
- People with soreness should start with small steps.
- Small steps for people with soreness include sitting in a chair.
- Small steps for people with soreness include standing.
- Small steps for people with soreness include walking.
- Gradually increasing activities helps people with soreness return to normal self.

Context: I spoke to your doctor and they wanted to address the questions you have regarding your loved one's complications from COVID-19. The doctor believes that the likely sequence of events is that the COVID-19 infection led to demand ischemia, which in turn caused the myocardial infarction (MI). \n\nThe doctor thinks that both the transfer hospital and the receiving hospital properly prioritized the patient's issues and treated the most life-threatening condition, the acute MI, first. Unfortunately, the patient had many underlying risk factors that made them more susceptible to severe illness from any infection, not just COVID-19. \n\nRegarding the patient going without clopidogrel for 10 days, the doctor agrees that this may have contributed to the MI, although it's impossible to determine the exact extent of its impact. \n\nAs for the new diagnoses of congestive heart failure (CHF), chronic obstructive pulmonary disease (COPD), and acute respiratory failure, the doctor did not provide a specific prognosis. However, they did mention that the patient's underlying health conditions and the severity of their illness have made their situation more challenging. \n\nThe doctor also believes that initiating COVID-19 treatment at the time of presentation may not have significantly altered the course of the patient's illness. \n\nPlease let us know if you have any further questions or concerns.

Please breakdown the following sentence into independent facts: However, they did mention that the patient's underlying health conditions and the severity of their illness have made their situation more challenging.

Facts:

- The patient's underlying health conditions makes the acute myocardial infarction situation more challenging.
- The severity of the patient's illness makes the acute myocardial infarction situation more challenging.
- The patient's acute myocardial infarction situation is challenging.

Context: I spoke to your doctor and they wanted to address your concerns regarding the leakage you experienced after your bowel surgery in 2013. According to them, it is possible for an abnormal connection to form between your bowel and your bladder or vagina, which is known as a fistula. This could potentially cause the leakage of substances from your bowel into your urinary tract or vagina.  
  
Your doctor recommends reviewing the notes from your second surgery to understand the nature of the repairs that were performed. This information may help clarify what happened in your specific case.  
  
Regarding your concerns about the quality of care you received from your initial surgeon, your doctor advises that medical malpractice is a complex issue that depends on many factors, including the specific circumstances of your case and the laws in your location. If you're interested in exploring this further, they recommend consulting with a lawyer who can provide guidance on whether you have a valid case.  
  
Please let us know if you have any further questions or concerns, and we'll be happy to help.

Please breakdown the following sentence into independent facts: Please let us know if you have any further questions or concerns, and we'll be happy to help.

Facts:

- No verifiable claim

user prompt:

Context: { }

Please breakdown the following sentence into independent facts: { }

Facts:

---

Table 19: 8 ICL examples used in the ablation study of MedScore decomposition prompt.

---

system prompt:

You are a medical expert in evaluating how factual a medical sentence is. You break down a sentence into as many facts as possible. The facts should be objective and verifiable against reliable external information such as Wikipedia and PubMed. All subjective personal experiences ("I was or someone did") and personal narratives (stating a past event) are not verifiable and should not be included in the fact list. Facts should be situated within conditions in the sentence. Suggestions (e.g. "I recommend or Your doctor suggest") and opinions (e.g. "I think") should be transformed into objective facts by removing subjective words and pronouns to only retain the core information that can be verified. Imperative instructions ("do something") should be transformed into declarative facts ("doing something is helpful for some conditions").

If there is an overly specific entity such as "Your partner" or vague references (pronouns, this or that) in the fact, replace it with a general phrase with conditional modifiers using information in the provided context (e.g. "People in some conditions"). Each fact should be verifiable on its own and require no additional context. Do not add additional information outside of the sentence and context. You do not need to justify what you extract.

If there is no verifiable fact in the sentence, please write "No verifiable claim".

Here are some examples:

Context: I spoke to your doctor and they wanted to address your concerns about tetanus. Since you've had your primary tetanus shots as a child, you don't need immunoglobulin (IGG) shots, and they were actually unnecessary during your last visit. \n\n Considering your tetanus vaccine expired in 2020 and you've got a dirty wound from the Spartan race, your doctor recommends getting a tetanus booster vaccine as soon as possible. They also mentioned that you were due for a booster anyway since it's been more than 3 years since your last vaccine.\n\nYour doctor is a bit puzzled as to why you were given IGG shots instead of a vaccine during your last visit, but that's not a concern for now. They just want to make sure you get the booster vaccine to be on the safe side. It's best to schedule an appointment for the booster vaccine as soon as possible to avoid any potential risks. Please breakdown the following sentence into independent facts: I spoke to your doctor and they wanted to address your concerns about tetanus.

Facts:

- No verifiable claim

Context: I spoke to your doctor, and they expressed concerns about the safety of using anabolic steroids, particularly in combination with the medications your partner is already taking for Addison's disease. The doctor noted that while these substances may have positive effects on muscle and bone health, they also carry significant risks and potential side effects.\n\nThe doctor mentioned that the anabolic cycle your partner is on is quite intense and requires careful monitoring for potential issues such as infertility, mood swings, and problems related to weight gain, including snoring and possible sleep apnea. They also emphasized the importance of considering the long-term effects of using these substances, particularly when they are stopped.\n\nThe doctor's primary concern is that your partner's underlying condition, Addison's disease, may not significantly complicate things if well-treated, but it could become an issue when the anabolic cycle is stopped. They strongly advise that your partner consult with a medical professional, ideally their endocrinologist, to discuss the potential risks and consequences of using these substances, especially given their pre-existing condition.\n\nIt's essential to have an open and honest conversation with a healthcare professional to ensure your partner's safety and well-being. I would encourage you to support your partner in seeking medical advice, and I'm happy to facilitate a discussion with their doctor if needed.

Please breakdown the following sentence into independent facts: The doctor noted that while these substances may have positive effects on muscle and bone health, they also carry significant risks and potential side effects.

Facts:

- Anabolic steroids may have positive effects on muscle health.
- Anabolic steroids may have positive effects on bone health.
- Anabolic steroids may also carry significant risks.
- Anabolic steroids may carry potential side effects.

Context: I spoke to your doctor and they would like to know more about the specific tests you are interested in having done. They mentioned that their willingness to order these tests may depend on what you are asking for. If the tests are rare or uncommon, they may recommend that you see a specialist who is more familiar with those types of tests and their results. Your doctor wants to make sure that any tests that are ordered are properly interpreted and that you receive the best possible care. Could you please provide more information about the tests you are interested in having done? Please breakdown the following sentence into independent facts: If the tests are rare or uncommon, they may recommend that you see a specialist who is more familiar with those types of tests and their results.

Facts:

- If the tests are rare or uncommon, seeing a specialist is helpful.
- The specialist is more familiar with rare or uncommon types of tests.
- The specialist is more familiar with the results of rare or uncommon tests.

Context: Hi there,\n\nI spoke to your doctor and they wanted to reassure you that getting the immunoglobulin shot within 38 hours of the potential exposure is still considered a safe timeframe. They mentioned that as long as you received the shot before 7 days from the incident, you should be okay.\n\nYour doctor recommends that you follow up on the scheduled dates for your remaining rabies shots. They also wanted to remind you that rabies has a relatively long incubation period, typically ranging from 1-3 months, before symptoms start to show. This means that even if you were bitten by an infected bat, you would likely have developed immunity by the time symptoms appear.\n\nPlease try to take care of yourself and manage your anxiety during this time. If you have any further concerns or questions, don't hesitate to reach out.\n\nBest regards,\n\n[Your Doctor's Name]

Please breakdown the following sentence into independent facts: They also wanted to remind you that rabies has a relatively long incubation period, typically ranging from 1-3 months, before symptoms start to show.

Facts:

- Rabies has a relatively long incubation period.
- The incubation period for rabies typically ranges from 1-3 months.
- Rabies symptoms start to show after the incubation period.

Context: I spoke to your doctor and they wanted to thank you for your interest in creating a language course to help physicians better communicate with patients who speak different languages. \n\nThey mentioned that while language barriers can contribute to the "revolving door syndrome," it's just one of many factors. Other important factors include education, home support, medication noncompliance, and lack of primary care. \n\nIn terms of a language course, your doctor thinks that Duolingo is a good option. However, they noted that it's challenging for doctors to find the time to learn multiple languages, as there are many languages spoken by patients in their area, including Spanish, Hmong, Chinese, and Polish. They also mentioned that many Spanish-speaking patients have some knowledge of English or have family members who are fluent in English. \n\nYour doctor didn't specify a preferred medium for the course, but they seemed to appreciate the idea of a convenient and accessible program. They also didn't provide specific vocabulary recommendations, but it's likely that a course focused on medical terminology and common patient interactions would be most useful.

Please breakdown the following sentence into independent facts: However, they noted that it's challenging for doctors to find the time to learn multiple languages, as there are many languages spoken by patients in their area, including Spanish, Hmong, Chinese, and Polish.

Facts:

- It is challenging for doctors to find the time to learn multiple languages.
- Many languages are spoken by patients in doctors' area.
- Spanish is one of the languages spoken by patients in doctors' area.
- Hmong is one of the languages spoken by patients in doctors' area.
- Chinese is one of the languages spoken by patients in doctors' area.
- Polish is one of the languages spoken by patients in doctors' area.

Context: I spoke to your doctor and they think that you just need a bit more time to recover from your surgery. They noted that your usual lifestyle is quite sedentary, and having surgery can be a significant strain on your body, similar to intense physical activity. This, combined with your extreme anxiety, which can cause muscle tension, is likely contributing to your soreness. \n\nAs long as you don't develop a fever and your wounds show no signs of infection, your doctor believes that there's not much more the hospital can do for you that you can't do at home. Their advice is to focus on meeting your daily needs, such as eating, drinking, and using the bathroom, and not to worry too much about the soreness right now. \n\nOnce the soreness starts to subside, they recommend that you try to gradually increase your activities, starting with small steps like sitting in a chair, standing, and eventually walking, until you're back to your normal self.

Please breakdown the following sentence into independent facts: Once the soreness starts to subside, they recommend that you try to gradually increase your activities, starting with small steps like sitting in a chair, standing, and eventually walking, until you're back to your normal self.

Facts:

- Once the soreness starts to subside, trying to gradually increase activities is helpful.
- People with soreness should start with small steps.
- Small steps for people with soreness include sitting in a chair.
- Small steps for people with soreness include standing.
- Small steps for people with soreness include walking.
- Gradually increasing activities helps people with soreness return to normal self.

Context: I spoke to your doctor and they wanted to address your concerns regarding the leakage you experienced after your bowel surgery in 2013. According to them, it is possible for an abnormal connection to form between your bowel and your bladder or vagina, which is known as a fistula. This could potentially cause the leakage of substances from your bowel into your urinary tract or vagina. \n\nYour doctor recommends reviewing the notes from your second surgery to understand the nature of the repairs that were performed. This information may help clarify what happened in your specific case. \n\nRegarding your concerns about the quality of care you received from your initial surgeon, your doctor advises that medical malpractice is a complex issue that depends on many factors, including the specific circumstances of your case and the laws in your location. If you're interested in exploring this further, they recommend consulting with a lawyer who can provide guidance on whether you have a valid case. \n\nPlease let us know if you have any further questions or concerns, and we'll be happy to help.

Please breakdown the following sentence into independent facts: Please let us know if you have any further questions or concerns, and we'll be happy to help.

Facts:

- No verifiable claim

Context: Dear Patient, \n\nI spoke to your doctor and they wanted to address your concerns about the sharp pain you've been experiencing above your left eyebrow when exposed to strong wind. After reviewing your symptoms, they think that trigeminal neuralgia could be a possible cause of your pain. \n\nThey considered cluster headaches as well, but they believe it's less likely in your case because cluster headaches typically occur at the same time every day, last for several days in a row, and involve pain behind or around the entire eye on one side. \n\nYour doctor would like to discuss this further with you and explore the possibility of trigeminal neuralgia being the cause of your pain. They recommend that you schedule a follow-up appointment to discuss your symptoms in more detail and determine the best course of action. \n\nPlease let us know if you have any questions or concerns before your appointment. \n\nBest regards, \n[Your Name] \nOn behalf of [Doctor's Name]

Please breakdown the following sentence into independent facts: Dear Patient, \n\nI spoke to your doctor and they wanted to address your concerns about the sharp pain you've been experiencing above your left eyebrow when exposed to strong wind.

Facts:

- No verifiable claim

Context: I spoke to your doctor, and they recommended that you visit a pharmacy to get an over-the-counter anti-nausea medication, such as Dramamine (also known as Gravol), to help alleviate your symptoms. They would like to know how you're feeling now, 4 hours after your initial message, to assess if your condition is improving or if further action is needed. If your symptoms persist, your doctor may want to investigate further to determine the cause of your discomfort. Please let us know your current status so we can provide further guidance.

Please breakdown the following sentence into independent facts: They would like to know how you're feeling now, 4 hours after your initial message, to assess if your condition is improving or if further action is needed.

Facts:

- No verifiable claim

Context: I spoke to your doctor and they wanted to reassure you that, given you are not sexually active and have never had penetrative intercourse, it is not possible for you to be pregnant. They understand that you have been experiencing anxiety and pregnancy scares, and they think it's a great idea for you to see a psychologist to help you manage these feelings.\n\nRegarding your symptoms, your doctor believes that starting a stable contraceptive therapy, such as the pill or other hormonal methods, could be helpful in regulating your periods and alleviating some of the symptoms you're experiencing. They think this could be a useful approach to help you feel better and more in control of your situation.\n\nPlease keep in mind that your upcoming ultrasound will likely provide more insight into what's going on with your body, and your doctor will be able to discuss the results with you and determine the best course of action.

Please breakdown the following sentence into independent facts: They understand that you have been experiencing anxiety and pregnancy scares, and they think it's a great idea for you to see a psychologist to help you manage these feelings.

Facts:

- Seeing a psychologist can help manage feelings of anxiety and pregnancy scares.

user prompt:

Context: { }

Please breakdown the following sentence into independent facts: { }

Facts:

---

Table 20: Instructions and 10 balanced ICL examples in the decomposition prompt of MedScore.