MED-OMIT: Extrinsically-Focused Evaluation of Omissions in Medical Summarization

Anonymous ACL submission

Abstract

Summarization is designed to condense text by focusing on the most critical information drawn from a source document (Paice, 1990; Kupiec et al., 1995). Generative large language models (LLMs) outperform previous summarization methods, yet traditional metrics struggle to capture resulting performance in more powerful LLMs (Goyal et al., 2022). In safety-critical domains such as medicine, rigorous evaluation is required given the potential for errors. We propose MED-OMIT, a new omission benchmark for medical summarization. MED-OMIT focuses on omissions as these have not been as widely studied as other types of errors. Given a doctor-patient conversation and a generated summary, MED-OMIT categorizes the chat into a set of facts and identifies which are omitted from the summary. We further determine fact importance by simulating the impact of each fact on a downstream clinical task: differential diagnosis (DDx) generation. MED-OMIT leverages LLM prompt-based approaches which categorize the importance of facts and cluster them as supporting or negating evidence to the diagnosis. We evaluate MED-OMIT on a publicly-released dataset of patient-doctor conversations. Based on expert evaluations, we find that MED-OMIT captures omissions and does so in cases where traditional metrics cannot. We highlight which models perform well out-of-the-box on this task, such as gpt-4, and which would likely require additional training.

1 Introduction

Powerful pretrained autoregressive large language models (LLMs) (OpenAI, 2023; Chowdhery et al., 2022; Touvron et al., 2023; Jiang et al., 2023) have led to high-quality summarization within a disparate set of domains such as news text (Liu and Healey, 2023), academic literature (Pu and Demberg, 2023), and food science (Shi et al., 2023). Within medicine previous studies have shown that medical providers face major challenges in maintaining documentation (Payne et al., 2015; Arndt et al., 2017). Therefore, applying these approaches to medicine (Ben Abacha et al., 2023a; Nair et al., 2023b) promises to allow providers to focus on other tasks.

Yet powerful, LLM-powered summarization for medical text is still error-prone (Ben Abacha et al., 2023b). Two critical types of errors are omissions, in which important information is excluded from the summary, and hallucinations, in which information is fabricated and erroneously included. Hallucinations can be detected using comparisons against ground truth from the document or external sources (Min et al., 2023; Umapathi et al., 2023; Vu et al., 2023; Ji et al., 2023b; Cohen et al., 2023; Peng et al., 2023). In contrast, detecting erroneous omissions is especially challenging as some information is always omitted from the summary. Yet *important* omissions can mislead a reader by incorrectly portraying the source document. Therefore, detecting omissions requires detecting what information is omitted *and* how important that information is.

We focus on the challenges of omissions in one common medical summarization task – the *subjective*. Subjective is a summary of everything the patient says or feels that is relevant to their health issue and informs the provider how to assess the patient's condition and design a treatment plan (see Section 2). Working from this common approach, we take inspiration from the common framework of intrinsic vs. extrinsic (or external) evaluation (Walter, 1998). The provider commonly uses the subjective summary to determine a differential diagnosis (DDx), which consists of possible diagnoses and their likelihood. As a result, the subjective must contain information that supports the most likely diagnosis and less likely ones to inform the provider fully.

We use this extrinsic task, the DDx, to focus on

Subjective
Chief Complaint (CC): The patient, Stephanie, a 49-year-old female, has been experiencing increased fatigue and dizziness over the past couple of months. She reports feeling worn out from daily activities that she used to handle without issue.
History of Present Illness (HPI): Stephanie's symptoms have been ongoing for a few months. She has not noticed any blood in her stools, nor have they been dark or tarry. She denies heavy menstrual bleeding, weight loss, loss of appetite, or fainting. She has been feeling dizzy but has not passed out. She has had some nasal congestion due to seasonal allergies.
Past Medical History (PMH): Stephanie has a significant past medical history of congestive heart failure, kidney stones, and a colonoscopy due to blood in her stools three years ago. The colonoscopy revealed mild diverticulosis, but she has had no issues since then. She has been struggling with her salt intake due to her congestive heart failure, and she admits to not doing well recently due to travel and eating fast food. She has noticed some swelling in her legs but nothing extreme. She has not had any recent flare-ups of her kidney stones, back pain, or blood in her urine.

Omitted Facts
Stephanie has slightly reduced heart function
The summary does not mention Stephanie's slightly reduced heart function (Score: 0.5).
Stephanie went to Vermont to explore the mountains
The summary does not mention Stephanie's recent travel to Vermont (Score: 0.1).
Stephanie ate two cheeseburgers at McDonald's during her travel
The summary does not mention Stephanie's specific food intake during her travel (Score: 0.1).
Stephanie has not experienced any shortness of breath or problems lying flat at night
The summary does not mention Stephanie's lack of shortness of breath or problems lying flat at night (Score: 0.1).
Stephanie's hemoglobin is low
The summary does not mention Stephanie's low hemoglobin levels (Score: 1).

Figure 1: Example GPT-4 generated subjective paired with the list of omitted facts and their weight. The facts are generated from the original patient-provider dialogue and their importance is scored using the MED-OMIT pipeline. The full chat is shown in Figure 7. For a longer example including the dialogue, see Appendix Figures 8 and 9.

our intrinsic subjective metric as is done in other machine learning approaches (Kao et al., 2018; Zhu et al., 2023). We propose MED-OMIT, to identify and weigh omitted information. We formulate a multi-prompt pipeline to produce an omission metric. As shown in Figure 2, we generate a summary using common LLM-based approaches from the patient-provider chat. Separately, we generate a list of facts from the conversation, which are designed to be atomic pieces of medical information. Using the list of facts paired with the summary, we can detect which facts are omitted.

Yet this alone does not inform us of which omissions are important. To accomplish this, we propose a *fact importance weight* which quantifies the criticality of each omitted fact, illustrated in Figure 3. We calculate this weight in two ways. First, we do so by categorizing the importance of all facts as a group. Second, we separately cluster facts that support and refute each diagnosis in an LLM-simulated DDx, and further sub-cluster these by their underlying medical function (or *pathophysiological mechanism*). This second approach allows us to highlight facts that uniquely point to any diagnosis – including rare or unlikely ones. While many facts are highly correlated, this seeks to surface non-correlated facts to the provider even if they are judged unimportant overall. Figure 1 shows an example of the resulting list. Using a simple weight scheme, we generate an importance score for each omitted fact and a cumulative score representing all omitted facts in a summary.

We also explore a metric that calculates the change in LLM generation score between the top diagnosis and the next possible diagnosis for both the chat and diagnosis and the summary and diagnosis. This represents a proxy for the information loss between the chat and the summary and directly represents the extrinsic metric. In both cases, our approach does not require a gold standard reference. We compare these metrics against referenced-based automated summarization metrics such as BERTScore (Zhang et al., 2019) and ROUGE (Lin, 2004) that are not designed to capture more nuanced errors.

In a human annotation analysis, we find that MED-OMIT reflects expert opinion on what is and is not an omission and how important each omission is. We find that our reference-free approach reflects the summarization performance of LLMs as they increase in size. We further find that for larger LLMs, previously reported metrics do not correlate well with the number of omissions in the summary. We plan to release the codebase upon paper publication.

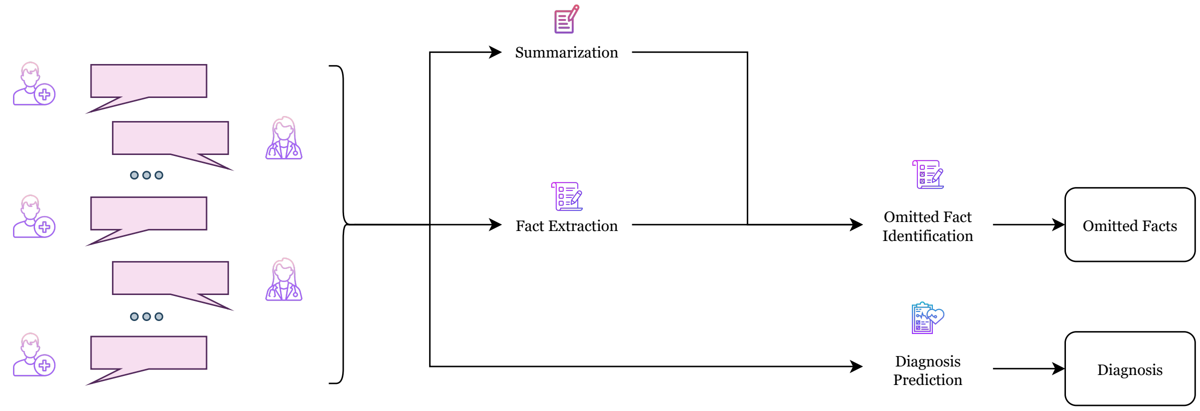


Figure 2: Given a patient-provider dialogue (left), we compute a summary and use a *fact extraction* module to extract facts from the conversation. We use the extracted facts from the conversation to identify if any facts are omitted from the summary. We also compute a differential diagnosis using the conversation data.

2 MED-OMIT

We begin by generating a subjective using the patient-doctor chat. The subjective is one of four standard note types (Podder et al., 2022) that are used in the SOAP framework. The subjective consists of the chief complaint (the most pressing medical issue), history of present illness (details about the chief complaint), medical and social history (details about previous medical issues), and current medications and allergies. We solely focus on the LLM generation of the subjective, given their superior performance on the task. We adopt the summarization prompt included in (Nair et al., 2023a). The original prompt contains section headers corresponding to the presence, absence, or unknown state of medical findings for the current encounter and medical history. We altered the section headers to only include information present in the subjective (see Prompt 1). While research has shown going beyond zero-shot improves performance (Ben Abacha et al., 2023a), we focus on using a zero-shot prompt to highlight the model's inherent summarization ability.

Reflective of how providers use subjectives, we generate a differential diagnosis (DDx) which lists potential medical diagnoses for the specific encounter. We do this from the chat instead of the summary to provide the most information possible. Separately, we generate a list of facts from the chat, similar to that in (Min et al., 2023) but medically focused. This allows us to represent what is present within the encounter discretely. We can then detect which fact(s) are excluded from the summary. We define an omission as a fact that is entirely or partially excluded from the resulting summary. We further assign an importance score to each omitted fact to better represent the effect of that fact being omitted from the summary. This is done first using a straightforward classification task, which assigns importance to each fact given the DDx. In addition, we also cluster the facts by whether they support or refute diagnoses in the DDx to highlight facts that uniquely point to a specific diagnosis. Accumulating the importance scores from all omitted facts produces a document-level metric, serving as a weighted count of all the omissions present. Below, we outline the details of each component in our pipeline. An example of the output of select pipeline components is included in Appendix Figure 8 and 9. We also include selected prompts in the Appendix.

DDx We prompt the LLM to generate a differential diagnosis given the summary. We separately do the same given the chat. This DDx includes at most ten potential medical conditions that might be relevant to the encounter. Each condition is ranked by order of likelihood, assigned a likelihood category (probable, possible, or unlikely), and given a short explanation. Note that a patient may have multiple medical issues in a given encounter, so multiple probable conditions may be true. We use the DDx generated from the chat in all prompts and only use the summary-generated DDx in Section 2.2.

Fact Identification We extract a list of facts from the dialogue using a prompt. This creates a discretized set of facts that is separate from the summary. The prompt is structured to categorize them as medical, related to care access or social determi-

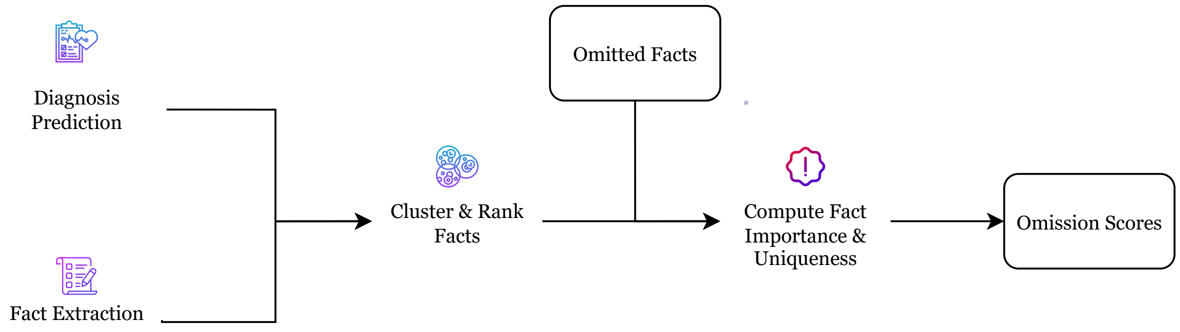


Figure 3: Given the previous outputs of the diagnosis prediction and fact extraction modules, we cluster facts that either support or refute a diagnosis. We also categorize each fact w.r.t. each diagnosis. With the clustered & categorized facts and the previously computed fact omissions, we assign an importance and uniqueness score to each fact.

nants of health, or non-medical. We do not leverage these groups but include them in the prompt to produce high-quality facts.

Fact Omission Detection Given the list of facts and the summary, we can then detect which facts are omitted from the summary. The resulting facts can either be unimportant or very important to clinical decision-making. However, at this stage, we only make the binary decision of present or omitted. We adopt a strict definition of a fact being omitted – if even some portion of the fact (e.g., 'severe' from 'severe pain') is omitted, it is counted as an omission. We hope future work will explore quantifying the degree of omission. We create the omission list by using Prompt 2.

Fact Importance Providing a list of omitted facts is an incomplete picture. The utility of each fact varies significantly – a fact such as *The patient has a fever* is likely much more important than *The patient loves iceberg lettuce*. Yet definite determinations can only be made with reference to the differential diagnosis. In a different scenario, *The patient loves iceberg lettuce* may be a critical fact if the doctor suspects a Listeria infection. Therefore, we prompt our LLM to categorize the importance of the facts with respect to the generated DDx.

We use three categories, including *critical*, *important*, and *other* (Prompt 3). We adopt this categorization for several reasons over other explored approaches. Finer-grained methods we explored, such as ranking or scoring each fact individually, result in pairwise decisions that are challenging to judge. Conversely, binary categorization is insufficiently nuanced.

Fact Uniqueness Categorizing facts only by their surface-level importance obfuscates other aspects of how a fact might be important. Specifically, facts that uniquely support or refute a specific diagnosis are also critical. For example, if the only supporting fact for Listeria is *The patient ate iceberg lettuce*, it is relevant even if the DDx determines that Listeria is unlikely. Ultimately, the provider should be provided with all evidence for any relevant diagnosis and empowered to make the final determination.

Therefore, we cluster each fact as supporting or refuting evidence with respect to each potential diagnosis (e.g., Prompt 4). This enables us to create a supporting and refuting evidence list. For example, *The patient has a fever* would be a supporting fact of a diagnosis of *Influenza*, whereas fever would be inconsistent with *Seasonal Allergies*.

In addition to the first-level clustering approaches, we create sub-clusters for supportive and refuting clusters. For each group of facts that support a single diagnosis, we prompt the model to cluster facts that suggest the same *pathophysiological mechanism*. This is designed to identify facts that are correlated because they are related to the same underlying issue. For example, the facts *Pain at the site of the bursa* and *Swelling at the site of the bursa* both point to potential *Inflammation*. As they are correlated, supporting evidence for inflammation would still be present even if only one fact were included. However, if only one were present, inflammation would be less likely to be considered. This intuition leads us to frame the uniqueness as an inverse frequency. Therefore, a fact's uniqueness would be scored as $\frac{1}{|S|}$, where S is the facts in the subcluster.

2.1 Document-Level Scores

While the above section results in a list of omitted facts and their importance, it does not provide us with a document-level metric for all omissions in a given document. Therefore, we propose a measure for creating a document-level metric for the omitted facts in the summary. In addition, we explore an alternative metric that seeks to measure the difference in the DDx generated from the chat and the DDx generated from the subjective. While the subjective is generated from the chat, the information loss can result in changes in the DDx. Therefore, we use an LLM to score the difference in DDxes between the chat and summary.

Fact Cumulative Score To achieve a single score representing the number of omissions and their importance, we begin by individually scoring each omitted fact. We first assign a *importance score* i for each omitted fact. If the fact omitted was critical, it receives a penalty of 1, a penalty of 0.5 for important, and a penalty of 0.1 for other. We separately assign a *uniqueness score* u. We assume that facts that uniquely support or contradict a diagnosis are the most important, compared with several facts that point to the same conclusion. Therefore, we use inverted scoring, where the fact is assigned a score of $\frac{1}{|S|}$ for each score it is present in. We take the maximum value of all potential penalties for an overall cluster score. To achieve a fact score for the entire document, we sum all of the individual score of all omitted facts;

$$\sum_{f \in \text{omissions}} max(i_f, u_f^0 ... u_f^k))$$

This represents a weighted count of the number of omissions in the document.

2.2 DDx Completion Score

In addition to our intrinsic metric, we explore an extrinsic diagnosis metric. We use an LLM to separately score the likelihood of the DDx conditioned on the summary and the chat. We then use the change in the likelihood score as a proxy for the effect of the information reduction in the subjective.

For the DDx generated by the chat, we compute the LLM's completion score of the highest ranked diagnosis l_{c0} [1] and the completion score of the highest ranked non-probable diagnosis l_{c1} [2]. In both cases, the LLM is prompted with the entirety of the chat, followed by the phrase *The patient most likely has*, and the diagnosis in question [3].

For the summary, we calculate the LLM completion score of the same highest-ranked diagnosis generated by the chat l_{s0} and the completion score of the highest ranked-non probable diagnosis coming from the summary l_{s1}. This is calculated with the same prompting approach, except that the summary is substituted for the chat. Ideally, the margin of the top diagnosis over the second-ranked diagnosis would be maintained when comparing the summary and the chat. Therefore, our overall score is;

$$\frac{(l_{c0} - l_{c1})}{(l_{s0} - l_{s1})}$$

As a score increases over 1, this would indicate that the chat margin is significantly higher than the summary and information is lost during summarization. This approach would be further strengthened by diagnosis-specific fine-tuning data, which is unavailable at this time. This approach also requires access to completion probabilities, which are unavailable for gpt models. After evaluating the ability of several models on development data, we selected Llama 2 (Touvron et al., 2023).

3 Dataset and Experiments

ACI-Bench We use the ACI-Bench dataset (Yim et al., 2023) to validate our omission metric. This consists of patient-doctor conversation and notes at the encounter level. The dialogues are role-played and not taken from actual patient data (no deidentification is required). The notes were generated using a note-generating system but edited by doctors afterward. We use the training set of 67 chats to calibrate our scoring system and use the three test sets of 118[4] chats to evaluate. We truncate the chats using a gpt-4 prompt to exclude non-subjective information (see Appendix A for details).

3.1 Quantitative Experimental Setup

Our approach enables us to select different LLMs for different components of the pipeline. For the summary prompt (Prompt 1; see the beginning of Section 2), we can select any LLM whose performance we wish to evaluate. Separately, we can select an LLM for MED-OMIT, which powers all

[1] This calculation could also be done with a gold standard diagnosis, but this was unavailable for our dataset.

[2] As many chats have multiple correct diagnoses, we do not use the second-ranked diagnosis.

[3] We tested different prefix phrases and found this worked best on training data.

[4] Two examples from the test dataset were excluded as their truncated chats were too small to generate a robust subjective.

Summary LLM	MED-OMIT Count mean	σ	MED-OMIT Weight mean	σ	LLM Comp. Margin mean	σ
gpt-4	3.72	3.07	2.41	2.31	1.34	4.63
gpt-3.5-turbo	5.61	3.56	3.63	2.57	1.72	12.26
mistral-7b	7.79	4.21	5.03	3.01	—	—
llama-70b	9.82	4.33	6.45	3.19	—	—

Table 1: For each summary LLM, we calculate the mean and standard deviation of both the number of MED-OMIT omissions and the cumulative weight, and the LLM completion margin.

other prompts in Section 2. For *summary* models, we use gpt-4 (OpenAI, 2023) and gpt-3.5-turbo, models shown to perform highly on several medical benchmarks (Ben Abacha et al., 2023a; Nair et al., 2023b).

In addition, we also explore the performance of two non-GPT open-source models – llama-70B (Touvron et al., 2023) and mistral-7b (Jiang et al., 2023). For the MED-OMIT model, we use gpt-4 given its high performance. We separately select the LLM for the DDx Completion Score (see Section 2.2). Finally, we also calculate correlation scores with referenced based metrics BERTScore and ROUGE using the same implementation as used in the dataset paper's code. As these are referenced based, we use the ACI gold-standard summaries .Unlike in our generated subjectives, the gold standard notes had access to the entire chat which discussed the final diagnosis.

3.2 Human Evaluation

Separately, we ask a group of three medical annotators to validate our fact omission detection and fact importance approaches for 20 conversations each (60 total). Given the output of the gpt-4 checking gpt-4 MED-OMIT, we ask them to validate the following information. First, was this fact included in the summary? *("Yes", "Partially", "No")*. Second, how many diagnoses are supported by this fact? Third, how many diagnoses are refuted by this fact? Note that these two questions are simplified forms of the MED-OMIT approach, as we do not ask them to do any clustering. Finally, if this fact were omitted, how much of an effect would it have on the differential diagnosis? *("Critical", "Important", or "Other")* . We randomly selected facts to annotate in each encounter which resulted in 330 fact annotations. We separately calculated inner annotator agreement and found that it was generally high (see Appendix Table 4 for full details). For additional annotation details, see Appendix B.

4 Results and Discussion

LLM performance We report MED-OMIT metrics on several Summary - Metric LLM configurations in Table 1. We separately report the number of omissions (MED-OMIT Count), the summation of the omission weights (MED-OMIT Weight), and the Llama-powered completion margin. For each, we report the mean and standard deviation over the test set. In all metrics, we find that gpt-4 performs best. However, the performance margin between gpt-4 and gpt-3.5-turbo isn't substantial. It is further remarkable that the MED-OMIT count margin of gpt-3.5-turbo over gpt-4 is larger than that for the MED-OMIT weight, suggesting gpt-3.5-turbo isn't omitting information that is more critical than gpt-4 summaries.

The Llama-powered LLM completion margin results show a similar pattern – the loss ratio is smaller for gpt-4 than for gpt-3.5-turbo. Yet in both cases, the standard deviation is large, suggesting that this is a much noisier metric than MED-OMIT. Finally, we also find that the open-source models trail gpt models in performance. However, we found that the mistral-7b model outperforms llama-70b on this task, which is remarkable given the mistral's smaller size. Yet in both cases, additional work in bridging the gap between closed-source and open-source performance is needed.

Expert evaluation of MED-OMIT As outlined in Section 3.2, we also compare the output of MED-OMIT (gpt-4) to that of medical annotators on four questions. As shown in Table 3, we see broad agreement between our medical annotators and MED-OMIT. First, we find that annotators agree 80% of the time with MED-OMIT's determination with whether a fact is omitted or not. Second, we find that the agreement on the fact importance question was even higher at 89.3%. For confusion matrices on these questions, see Appendix Figures 4 and 5. For selected examples, see Appendix Section

B.1. Additionally, we asked annotators to count the number of diagnoses each fact both supports and refutes. The absolute difference between the annotator's count and GPT-4's count was less than one in both cases. Histograms of the full distributions are available in Appendix Figure 6. In summation, these results show that MED-OMIT accurately captures the identifying and quantifying the importance of omissions.

Correlation with other Metrics In Table 2, we report the Spearman and Pearson correlations between commonly reported summarization metrics ROUGE and BERTScore with our Omission Weight and Counts. Additional metrics are included in Appendix Table 5. We do not find any significant correlation between the LLM Completion metric and ROUGE or BertScore. We find that for the less powerful LLM, traditional summarization metrics do correlate with our omission metrics. Unsurprisingly, higher omission weight and count scores inversely correlate with higher BertScore and ROUGE metrics. However, for summaries generated by more powerful LLMs, there is no statistically significant correlation. This finding is supported by previous research (Goyal et al., 2022) that showed smaller incremental changes in summarization quality are challenging to detect with traditional metrics and highlights the importance of targeted metrics.

Areas of potential improvement We performed a qualitative analysis by randomly sampling ten training examples from the ACI dataset. While we found MED-OMIT was broadly accurate, there are areas of improvement. First, we found that while the LLM was able to consistently detect which facts were omitted from the summary, it did so in a strict manner. Consider the example in Figure 8 and 9 – a fact (F8) was correctly identified as excluded. However, the summary only omitted the specific foods the patient was excluding from their diet but did include the overall point that she was trying to apply a low-sodium diet. Capturing the amount of a fact that was excluded remains an open question. As shown in the expert annotation results, the task is challenging for gpt-4 to perform given the complexity of what constitutes an omission.

Perhaps the most challenging task is generating the clusters and sub-clusters of supporting and refuting evidence. Specifically within the framework of the sub-clustering, accurately clustering the facts around symptoms, tests, treatments, and social determinants of health was a challenging prompt to engineer. While we find that it does well at selecting the correct category and the correct pathophysiological mechanism for the common categories, it can make mistakes. For example, in Figure 9, there is a "NONE" category for symptoms within *Well-managed Congestive Heart Failure*, which is not an actual pathophysiological mechanism.

Additionally, the refuting sub-clustering step occasionally makes broad inferences given the full set of facts. For example, one refuting sub-cluster noted that *[NAME] has chronic back pain that bothers her when she sits for long periods of time at her desk at work* is a refuting fact for Fibromyalgia because *Fibromyalgia typically presents with widespread pain* even though this is not explicitly stated. Both LLMs and medical providers make inferences based on what is absent from a medical case, but the degree of alignment is unclear.

Finally, we find that the weighting system does sort summaries pairwise in a sensible manner. Consider the example in Figure 9 and another case with only a single omission. In the single omission case, the fact *Edward experiences swelling in his ankles, mainly near the end of the day* was omitted from a subjective. This was categorized as critical as it speaks to potential fluid retention which potentially supports several diagnoses. By contrast, the example in Figure 9 has five omissions. Yet they are all judged to be less important, and none receive a max score. This illustrates the importance of going beyond binary judgments on omitted facts.

5 Related Work

Work in large language models, such as gpt-4 (OpenAI, 2023), PaLM (Chowdhery et al., 2022), Llama (Touvron et al., 2023), and Mistral (Jiang et al., 2023), have enabled advances in text generation performance. Compared to earlier LLMs such as BERT (Devlin et al., 2019), these models are autoregressive and can condition generation on a set of input instructions (Reynolds and McDonell, 2021; Brown et al., 2020).

Summarization tools built on LLMs have shown increasing performance that is equivalent to human-written summaries (Zhang et al., 2023). Yet quantifying the performance of such approaches has also increased in difficulty (Goyal et al., 2022) as common summarization metrics such as BLEU (Papineni et al., 2002), ROUGE (Zhang et al., 2019), METEOR (Banerjee and Lavie, 2005), and

Summary LLM	Comparison Metric Name	Value	MED-OMIT weight spear.	pear.	MED-OMIT count spear.	pear.
gpt-4	Rouge LSum	0.363	0.003	-0.041	-0.046	-0.057
gpt-4	BertScore F1	0.651	-0.044	-0.130	-0.130	-0.043
gpt-3.5-turbo	Rouge LSum	0.333	**-0.242**	**-0.220**	**-0.244**	**-0.200**
gpt-3.5-turbo	BertScore F1	0.627	**-0.338**	**-0.299**	**-0.338**	**-0.281**

Table 2: For the two best models, we compare MED-OMIT mean count and weight to reference-based metrics BERTScore and Rouge. We report the Spearman and Pearson correlation between each reference-based and MED-OMIT metric. Bolded values are significant with a two-sided test $p < 0.05$. For additional metrics, see Appendix Table 5.

Fact Missing Agreement	Fact Importance Agreement	# Diagnoses Supp. Mean Abs. Diff.	σ	# Diagnoses Ref. Mean Abs. Diff.	σ
80.0%	89.3%	0.439	1.184	0.447	1.224

Table 3: Agreement statistics for comparing MED-OMIT (gpt-4 checking gpt-4) with expert annotators decisions on four questions. For confusion matrices and distribution plots, see Appendix Figures 4, 5, and 6. For inner-annotator agreement, see Appendix Table 4.

BertScore (Zhang et al., 2019) don't align with human judgments. Further studies of LLM summarization have also highlighted issues with hallucinations (Ji et al., 2023a).

Therefore, there has been a major focus on developing ways to identify and remediate hallucinations in LLM generations (Vu et al., 2023; Ji et al., 2023b; Cohen et al., 2023; Peng et al., 2023; Shuster et al., 2021; Liu et al., 2022). For example, Min et al. (2023) proposes to automatically extract atomic facts from the generated text and verify them against an external knowledge source. In contrast to our work, they weigh each hallucination equally and do not discuss omissions. In addition, there have been domain-focused hallucination studies in safety-critical domains such as medicine (Umapathi et al., 2023). Other work has looked at evaluating medical texts using different extrinsic metrics (Moramarco et al., 2021). Relatedly, there is also a line of work that seeks to reduce the risk of harmful LLM output (Glaese et al., 2022; Ouyang et al., 2022; Scheurer et al., 2022; Bai et al., 2022) which is especially important in safety-critical domains such as medicine.

6 Conclusion

We propose MED-OMIT, which identifies omitted facts within a summary and quantifies their importance. We first extract facts from the source chat and thus can detect which are omitted from the summary. However, we also go beyond simple detection by assigning an importance weight to each fact to reflect the potential impact of that fact on downstream decision-making. Therefore, we prompt an LLM to categorize the list of facts by their importance concerning a simulated differential diagnosis. To complement this, we also separately cluster all facts by whether they support or refute each diagnosis, and further cluster by pathophysiological mechanism. This highlights facts uniquely pinpointing any diagnosis, no matter how unlikely, to inform patients and providers best.

Using these individual fact weights, we then can provide a MED-OMIT weight for each document in addition to the omission count. We find that the MED-OMIT metrics conform to the expectation that larger LLMs outperform smaller ones. In addition, while we find a small but statistically significant correlation with Rouge and BertScore on smaller LLMs such as gpt-3.5-turbo, we do not with larger LLMs. This suggests that focused metrics are essential in better capturing performance as LLMs increase in size.

7 Limitations

We focus solely on omissions in this work but hope that future work can extend it to other potential summarization errors. While using an extrinsic metric is less critical in detecting hallucinations given the presence of ground truth, summaries adding unneeded information could be detected using a similar approach. Additionally, we detect an omission on a binary scale - either some part of the fact is omitted or none of it is. An extension could include attempting to capture what percentage of the fact is omitted and the importance of the omitted aspect.

Our metric pipeline performs best with a large LLM, such as gpt-4. This adds increased cost when evaluating summaries. While we feel this is warranted in safety-critical domains such as medicine, we also look towards exploring fine-tuned models to accomplish the same goals with reduced cost. This is further an important goal as gpt-4 is a closed LLM, and having a high-performance open-source alternative is useful.

In addition, as a community, we need to invest in building more publicly available datasets. We were able to do this study only because of access to the openly available medical dialogue dataset.

8 Ethical Considerations

We use a publicly de-identified dataset for evaluating our approach. However, given that this method can be applied to publicly available data, we caution that care should be taken when using non-hosted LLMs with sensitive data. Additionally, while our approach is designed to aid a medical practitioner in making medical diagnoses, this is not designed to replace a medical practitioner overall.

References

Brian G Arndt, John W Beasley, Michelle D Watkinson, Jonathan L Temte, Wen-Jan Tuan, Christine A Sinsky, and Valerie J Gilchrist. 2017. Tethered to the ehr: primary care physician workload assessment using ehr event log data and time-motion observations. *The Annals of Family Medicine*, 15(5):419–426.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Asma Ben Abacha, Wen-wai Yim, Griffin Adams, Neal Snider, and Meliha Yetisgen. 2023a. Overview of the MEDIQA-chat 2023 shared tasks on the summarization & generation of doctor-patient conversations. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 503–513, Toronto, Canada. Association for Computational Linguistics.

Asma Ben Abacha, Wen-wai Yim, George Michalopoulos, and Thomas Lin. 2023b. An investigation of evaluation methods in automatic medical note generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2575–2588, Toronto, Canada. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. 2023. Lm vs lm: Detecting factual errors via cross examination.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. 2022. Improving alignment of dialogue agents via targeted human judgements.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023a. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Ziwei Ji, Zihan Liu, Nayeon Lee, Tiezheng Yu, Bryan Wilie, Min Zeng, and Pascale Fung. 2023b. Rho (ρ): Reducing hallucination in open-domain dialogues with knowledge grounding.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Hao-Cheng Kao, Kai-Fu Tang, and Edward Chang. 2018. Context-aware symptom checking for disease diagnosis using hierarchical reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Julian Kupiec, Jan Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '95, page 68–73, New York, NY, USA. Association for Computing Machinery.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Sengjie Liu and Christopher G Healey. 2023. Abstractive summarization of large document collections using gpt. *arXiv preprint arXiv:2310.05690*.

Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2022. A token-level reference-free hallucination detection benchmark for free-form text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6723–6737, Dublin, Ireland. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation.

Francesco Moramarco, Alex Papadopoulos Korfiatis, Aleksandar Savkov, and Ehud Reiter. 2021. A preliminary study on evaluating consultation notes with post-editing. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 62–68, Online. Association for Computational Linguistics.

Varun Nair, Elliot Schumacher, and Anitha Kannan. 2023a. Generating medically-accurate summaries of patient-provider dialogue: A multi-stage approach using large language models. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 200–217, Toronto, Canada. Association for Computational Linguistics.

Varun Nair, Elliot Schumacher, Geoffrey Tso, and Anitha Kannan. 2023b. Dera: Enhancing large language model completions with dialog-enabled resolving agents.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Chris D Paice. 1990. Constructing literature abstracts by computer: techniques and prospects. *Information Processing & Management*, 26(1):171–186.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Thomas H Payne, Sarah Corley, Theresa A Cullen, Tejal K Gandhi, Linda Harrington, Gilad J Kuperman, John E Mattison, David P McCallie, Clement J McDonald, Paul C Tang, et al. 2015. Report of the amia ehr-2020 task force on the status and future direction of ehrs. *Journal of the American Medical Informatics Association*, 22(5):1102–1110.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback.

V Podder, V Lew, and S Ghassemzadeh. 2022. Soap notes.[updated 2021 sep 2]. *StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing*.

Dongqi Pu and Vera Demberg. 2023. ChatGPT vs human-authored text: Insights into controllable text summarization and sentence style transfer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 1–18, Toronto, Canada. Association for Computational Linguistics.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Jérémy Scheurer, Jon Ander Campos, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. 2022. Training language models with language feedback.

Yiwen Shi, Ping Ren, Jing Wang, Biao Han, Taha ValizadehAslani, Felix Agbavor, Yi Zhang, Meng Hu, Liang Zhao, and Hualou Liang. 2023. Leveraging gpt-4 for food effect summarization to enhance product-specific guidance development via iterative prompting. *Journal of Biomedical Informatics*, 148:104533.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Logesh Kumar Umapathi, Ankit Pal, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2023. Freshllms: Refreshing large language models with search engine augmentation.

Sharon M. Walter. 1998. Book reviews: Evaluating natural language processing systems: An analysis and review. *Computational Linguistics*, 24(2).

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Nature Scientific Data*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2023. Benchmarking large language models for news summarization.

Jian Zhu, Ilya Valmianski, and Anitha Kannan. 2023. Dialogue-contextualized re-ranking for medical history-taking.

A Dataset details

Our approach is targeted to subjective note, which encapsulates the early part of the encounter where the diagnosis is not necessarily known. However, the ACI chats discuss the full patient encounter, and include physician-determined diagnoses, outcomes of physical examinations, and test results. Therefore, we truncate the chats to exclude any information that would point to a diagnosis to better simulate when a subjective would be generated. We find the last relevant line in the chat that discusses any subjective-related information and truncate the chat to the next line using Prompt 5. We will release the truncation indices with our codebase.

B Annotation Details

The selected facts consist of all omitted facts in the summary, plus a randomly selected set of facts that were not omitted. We select all omitted facts and add n more non-omitted facts to annotate at most 5 per encounter. All values except for the first question were precomputed and presented to the annotator for validation. The annotators were instructed to change any precomputed value if they believed it appropriate.

The instructions given to the annotators were as follow; The following sheets contain encounter information from an external dataset. Each encounter consists of

- A generated subjective.
- A generated differential diagnosis
- A list of all facts extracted from the encounter

Before answering any questions, please read the above information.

A specific fact from the list is included for consideration. With respect to this fact, we'd like you to validate the following questions. The values in the first three are pre-computed. However, you are free to change them if you think appropriate.

- Is this fact included in the summary? Rate as No (it is completely excluded), Partially (some element, even a non-medically important one, is excluded), or Yes (it is included)
- If this fact is a positive finding, how many diagnosis does it support? This should be a value between 0 and the total number of diagnoses.
- If this fact is a negative finding, how many diagnoses does it refute? This should be a valute between 0 and the total number of diagnoses.
- If this fact were ommitted from the list of facts, what would the impact be on the differential diagnosis? Please rate as Critical (highest), Important (moderate), Other (lowest).
  - The impact" of the diagnosis can include a variety of factors. These include but are not limited to adding a new diagnosis to the list or removing an diagnosis currently present in the list. Alternatively, would a diagnosis be more or less likely?

B.1 Differences between annotators and gpt-4

While there is generally agreement between gpt-4 and annotators, there are several instances where they disagree. The following are several examples taken from the development data. We report the fact, the relevant sentence(s) from the summary, and the judgements.

Fact: Vincent experienced dizziness and lightheadedness.

Relevant Summary: He reported experiencing lightheadedness but denied any noticeable bleeding.

Is Included?: No (gpt-4, 2 annotators); Partially (1 annotator)

The above example shows the challenge of detecting whether a fact is omitted from the summary. The summary includes most of the important text, but does exclude *dizziness*. While related to *lightheadedness*, it is not the same thing. Since gpt-4 is only allowed to make binary judgements, it says its not included. Our annotators have the option to select 'Partially'; one decides to do so while the others agree fully with gpt-4.

Fact: Rachel's depression has moments of highs and lows

Relevant Summary: Her depression is managed with Effexor, but she still experiences periods of low mood.

Is Included?: Yes (1 annotator); Partially (2 annotators); No (gpt-4)

This example further illustrates the challenge in determining whether a fact was included. The majority of the fact is included in the summary. However, the "highs" work is excluded, which may be informative for the patient's condition. Since gpt-4 only has a binary choice, it selects No, while the annotators alternatively select Yes or Partially.

Relevant Fact: Mrs. Peterson would avoid going upstairs or downstairs.

All facts:
F0: Mrs. Peterson is a 43-year-old patient.
F1: Mrs. Peterson is experiencing right leg pain.
F2: Mrs. Peterson injured her right leg while bowling.
F3: Mrs. Peterson's bowling ball hit her right leg.
F4: Mrs. Peterson's right leg has a little bit of bruising on the back end.
F5: Mrs. Peterson is able to walk on her right leg, but very carefully.
F6: Walking on her right leg is very sore for Mrs. Peterson.
F7: Mrs. Peterson would avoid going upstairs or downstairs.
F8: Mrs. Peterson has a history of atopic eczema.
F9: Mrs. Peterson uses fluocinonide for her eczema when it gets really itchy.
F10: Mrs. Peterson has a previous surgical history of a colectomy.
F11: Mrs. Peterson had diverticulosis which turned into diverticulitis, leading to the removal of a part of her colon.
F12: Mrs. Peterson was bowling when she injured her leg.

DDx: Contusion (Bruise) : Probable
Muscle Strain : Probable
Fracture : Possible
Soft Tissue Injury : Possible
Hematoma : Possible
Bursitis : Unlikely
Tendon Rupture : Unlikely
Nerve Damage : Unlikely
Deep Vein Thrombosis (DVT) : Unlikely
Compartment Syndrome : Unlikely
Is Important?: Critical (2 annotator and gpt-4); Important (1 annotator)

While there is less disagreement for fact importance, there are still some tricky cases. Consider the above case; the fact that the patient is unable to walk up and down stairs should be of obvious concerns to the provider given the hindrance to mobility. While 2 annotators and gpt-4 decide that it's a critical fact, one annotates it as important. This is potentially because there are other facts that encapsulate that the patient has trouble walking, and it isn't of strict criticality that she has trouble walking on the stairs.

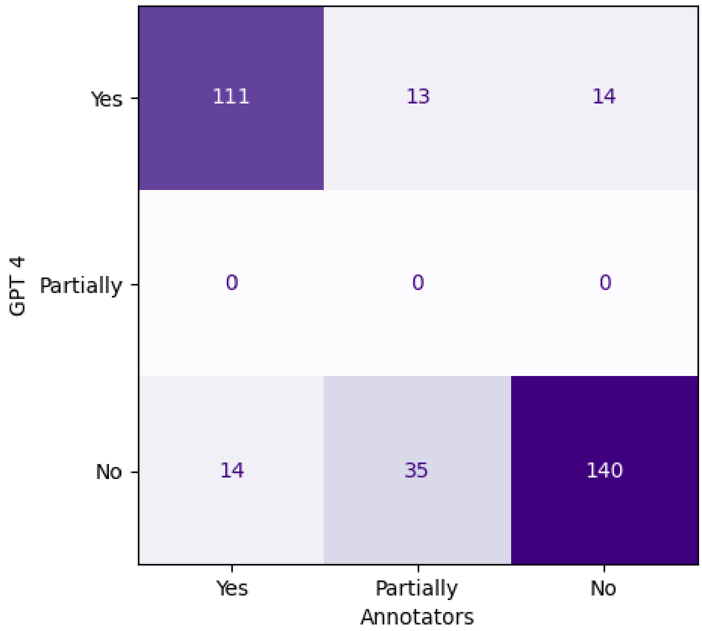


Figure 4: Confusion Matrix for annotator agreement with GPT-4 for the Fact Omission task. The overall agreement was 80%. Note that while we give annotators three labels to choose from, MED-OMIT only uses a binary judgement (and excludes the "Partially" option). Therefore, we count annotators selecting "Partially" as correct if MED-OMIT selects "Yes"). We believe work capturing the degree of omission would provide further insight.

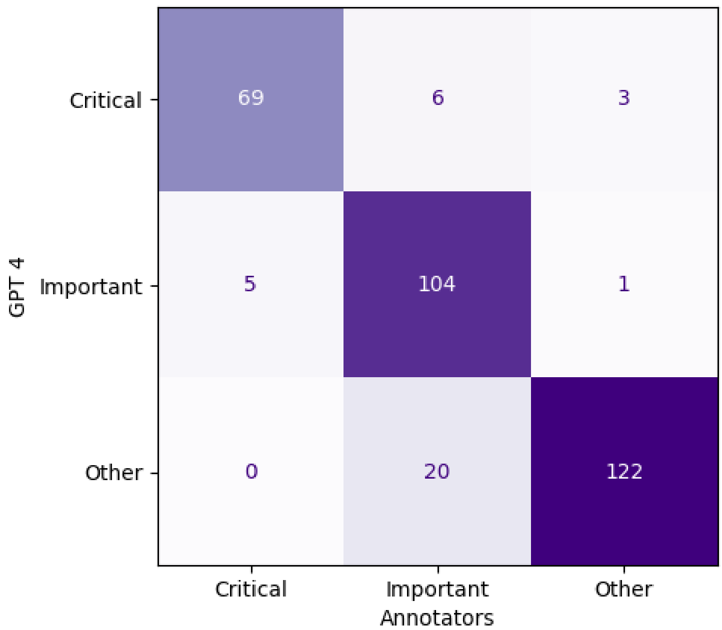
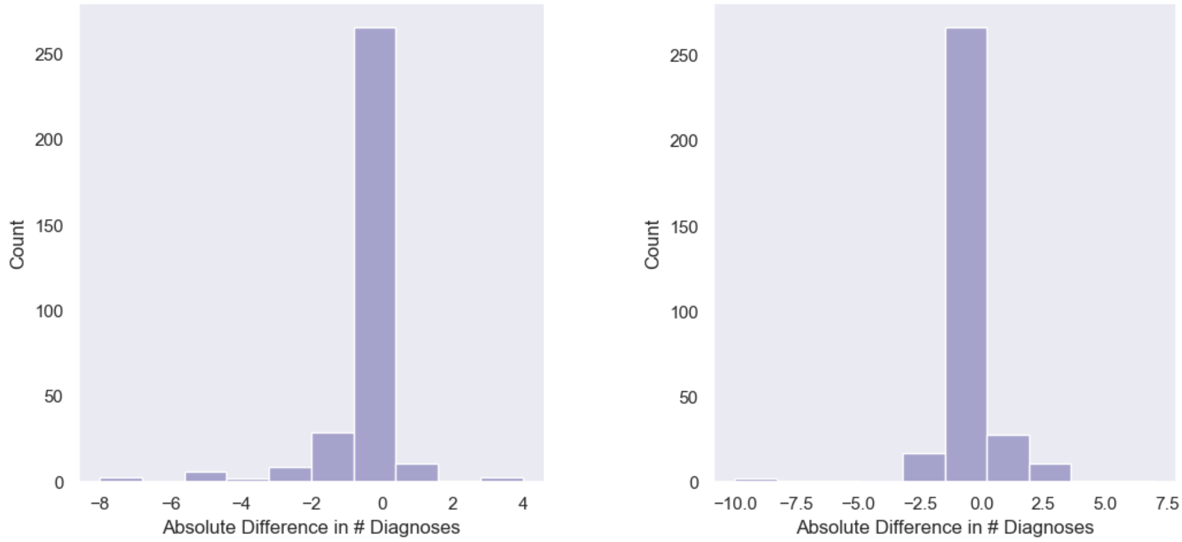


Figure 5: Confusion Matrix for annotator agreement with GPT-4 for the Fact Importance categorization task. The strict agreement is 89.4%.



(a) Supporting

(b) Refuting

Figure 6: Distribution of absolute differences between number of diagnoses supported (a) and refuted (b) by each fact as determined by MED-OMIT and expert annotators.

Question	Exact Match	Cohen's Kappa	Cohen's Kappa (Linear)
Is Included?	82.35%	0.70,0.70,0.74	0.76,0.80,0.80
Is important?	78.43%	0.68,0.65,0.62	0.74,0.71,0.71
Question	Exact Match	Max Diff Mean	Max Diff σ
Supports how many diagnosis?	64.71%	0.78	1.29
Refutes how many diagnoses?	58.82%	1.84	2.80

Table 4: Inner-annotator agreement statistics for a separate dataset of 51 facts that were annotated by all three annotators. We find that for both the ommission and importance questions, the exact match rate (defined as all three annotators agreeing with each other) was high. Separately, we calculated Cohen's kappa for each annotator pair, using both an unweighted and linearly weighted approach. For the supporting and refuting fact questions, we found that the exact match rate between the three annotators is unsurprisingly lower. We additionally calculated the margin between the highest number of diagnoses and the lowest, representing the maximum disagreement between annotators. We found that the mean difference was than 1 for supporting diagnoses. For the refuting question, it was less than 2, which while larger than supporting, is within a reasonable range given the complexity of the question.

Doctor-Patient Chat

...

Doctor: hi , stephanie . how are you ?

Patient: i'm doing okay . how are you ?

Doctor: i'm doing okay . um , so i know the nurse talked to you about dax . i'd like to tell dax a little bit about you , okay ?

Patient: okay .

Doctor: so , stephanie is a 49-year-old female with a past medical history significant for congestive heart failure , kidney stones and prior colonoscopy who presents today for an abnormal lab finding . so , stephanie , i called you in today because your hemoglobin is low . um , how have you been feeling ?

Patient: over the past couple of months , i've been really tired and dizzy . lately , i've been really just worn out , even just , you know , walking a mile or going to work , doing things that i've done in the past every day that have been relatively okay , and i have n't gotten tired . and now , i've been getting tired .

Doctor: okay , yeah . i , you know , the nurse told me that you had called with these complaints . and i know that we have ordered some labs on you before the visit . and it did , it c- you know , your , your , your hemoglobin is your red blood cell count . and now , and that came back as a little low on the results , okay ? so , have you noticed any blood in your stools ?

Patient: uh , no , i have n't . i did about three years ago , um , and i did a colonoscopy for that , but nothing since then .

Doctor: okay , yeah . i remember that , okay . and how about , you know , do your stools look dark or tarry or black or anything like that ?

Patient: no , nothing like that .

Doctor: okay . and have you been , um , having any heavy menstrual bleeding or anything like that ?

Patient: no , not that i've noticed .

Doctor: okay , all right . and any , have you passed out at all , or anything like that ? any weight loss ?

Patient: no , no weight loss or passing out . i have felt a bit dizzy , but it has n't l- led to me passing out at all .

Doctor: okay . so , you endorse some dizziness . you endorse some fatigue . have you , but you have n't had any weight loss , loss of appetite , anything like that ?

Patient: no , nothing like that .

Doctor: okay , all right . so , you know , let's talk a little bit about that colonoscopy . i know you had a colonoscopy about three years ago and that showed that you had some mild diverticuli- diverticulosis . um , no issues since then ?

Patient: nope , no issues since then .

Doctor: okay , all right . and then i know that , uh , you know , you have this slightly reduced heart function , you know , your congestive heart failure . how have you been doing watching your salt intake ? i know that that's kind of been a struggle for you .

Patient: um , it's been more of a struggle recently . i've been traveling a lot . i went up to vermont , um , to go , um , explore the mountains . and along the way i stopped at , you know , mcdonald's and got two cheeseburgers . and so , i , i could be doing better . i've noticed some swelling in my , my legs . um , but nothing too extreme that where i thought i should call .

Doctor: okay , all right . and any shortness of breath or problems lying flat at night , anything like that ?

Patient: no , nothing like that .

Doctor: okay , all right . and then in terms of the kidney stones , i know that you had those a couple years ago , as well . any recent flare ups ? have you had any , any back pain , flank pain , anything like that ?

Patient: no , nothing like that .

Doctor: okay . any blood in your urine that you've seen ?

Patient: no .

Doctor: okay , all right . um , okay . well , i know that the nurse did a review of system sheet when you came in . and we've just talked a lot about your , your s- your symptoms , you know , your dizziness , your fatigue and that type of thing . anything else that i might have missed , fever chills , any nasal congestion , sore throat , cough ?

Patient: uh , i've had a little bit of nasal congestion just because with the seasons changing , i , i get seasonal allergies . but everything else has been okay ." ...

Figure 7: Full chat for Figure 1

Summary LLM	Metric LLM	Comp. Metric	Comp. Mean	Omm. Weight		Omm. Count	
				Spear.	Pear.	Spear.	Pear.
gpt4	gpt4	rouge1	0.4495	0.002	0.001	-0.042	-0.016
gpt4	gpt4	rouge2	0.1929	0.016	-0.007	-0.046	-0.036
gpt4	gpt4	rougeL	0.2652	-0.067	-0.082	-0.148	-0.129
gpt4	gpt4	rougeLsum	0.3630	0.003	-0.041	-0.046	-0.057
gpt4	gpt4	bertscore-P	0.6684	-0.105	-0.038	-0.130	-0.043
gpt4	gpt4	bertscore-R	0.6368	0.055	0.027	0.061	0.058
gpt4	gpt4	bertscore-f1	0.6507	-0.044	-0.009	-0.061	0.000
gpt-3.5-turbo	gpt4	rouge1	0.3892	**-0.273**	**-0.239**	**-0.279**	**-0.212**
gpt-3.5-turbo	gpt4	rouge2	0.1566	**-0.253**	**-0.270**	**-0.255**	**-0.252**
gpt-3.5-turbo	gpt4	rougeL	0.2404	**-0.301**	**-0.291**	**-0.314**	**-0.289**
gpt-3.5-turbo	gpt4	rougeLsum	0.3329	**-0.242**	**-0.220**	**-0.244**	**-0.200**
gpt-3.5-turbo	gpt4	bertscore-P	0.6450	**-0.333**	**-0.294**	**-0.338**	**-0.260**
gpt-3.5-turbo	gpt4	bertscore-R	0.6127	**-0.305**	**-0.246**	**-0.299**	**-0.259**
gpt-3.5-turbo	gpt4	bertscore-f1	0.6270	**-0.338**	**-0.299**	**-0.338**	**-0.281**

Table 5: Full correlation results between the omission weight and count, and all Rouge and BertScore components. The values in bold are found to be significant with a two-sided test $p < 0.05$.

```
Below is a medical encounter between a patient and a doctor done over chat .
----
Medical Encounter
----
{{dialogue}}
----
Summary Instructions
----
Provide a summary of the medical encounter between the doctor and the patient.

Separate the note into separate sections, with divisions were inspired by the
    SOAP standard.
-The "Subjective" includes items taken during verbal exam and typically written
     in the form of chief complaint (CC), history of present illness (HPI), and
     past social history
DO NOT INCLUDE THE FOLLOWING SECTIONS;
-You should not include any "Objective Exam" includes content from the physical
     examination on the day of the visit
-You should not include any "Objective Results", which includes diagnostics
    taken prior to the visit, including laboratory or imaging results
-You should not include any "Assessment and Plan", which includes the doctor's
    diagnosis and planned tests and treatments

If there is no information for a section, please omit it.

Summary of Medical Encounter:
```

Prompt 1: Prompt for generating summary

Doctor-Patient Chat

...

Doctor: martha is a 50-year-old female with a past medical history significant for congestive heart failure , depression and hypertension who presents for her annual exam . so , martha , it's been a year since i've seen you . how are you doing ?

Patient: i'm doing well . i've been traveling a lot recently since things have , have gotten a bit lighter . and i got my , my vaccine , so i feel safer about traveling . i've been doing a lot of hiking . uh , went to washington last weekend to hike in northern cascades, like around the mount baker area .

Doctor: nice . that's great . i'm glad to hear that you're staying active , you know . i , i just love this weather . i'm so happy the summer is over . i'm definitely more of a fall person .

Patient: yes , fall foliage is the best .

Doctor: yeah . um , so tell me , how are you doing with the congestive heart failure ? how are you doing watching your diet ? i know we've talked about watching a low sodium diet . are you doing okay with that ?

Patient: i've been doing well with that . i resisted , as much , as i could , from the tater tots , you know , the soft pretzels , the salty foods that i , i love to eat . and i've been doing a really good job .

Doctor: okay , all right . well , i'm glad to hear that . and you're taking your medication ?

Patient: yes .

Doctor: okay , good . and any symptoms like chest pains , shortness of breath , any swelling in your legs ?

Patient: no , not that i've noticed .

Doctor: okay , all right . and then in terms of your depression , i know that we tried to stay off of medication in the past because you're on medications for your other problems . how are you doing ? and i know that you enrolled into therapy . is that helping ? or-

Patient: yeah , it's been helping a lot . i've been going every week , um , for the past year since my last annual exam . and that's been really helpful for me. ...

Subjective (Prompt 1)

The patient, Martha, is a 50-year-old female with a history of congestive heart failure, depression, and hypertension. She presented for her annual exam and reported that she has been doing well overall. She has been traveling and hiking a lot recently, which indicates an active lifestyle. Regarding her congestive heart failure, she has been adhering to a low sodium diet and taking her medication regularly. She reported no symptoms such as chest pains, shortness of breath, or swelling in her legs. As for her depression, she has been attending therapy sessions weekly for the past year and reported that it has been helpful. She denied any feelings of wanting to harm herself or others. However, she admitted to forgetting to take her blood pressure medication occasionally, especially during stressful periods at work. She has been working long hours and late nights, which has been causing her stress and increasing her blood pressure.

Fact Extraction and Categorization (Prompt 3)

Critical

F1: Martha has a past medical history significant for congestive heart failure, depression and hypertension

F14: Martha is forgetting to take her blood pressure medication

F15: Martha's blood pressure goes up when work gets more stressful

Important

F12: Martha has been going to therapy every week for the past year

F16: Martha has been working long hours, late nights

F17: Martha has a lot of fiscal year end data that she's been having to pull

...

Other

F0: Martha is a 50-year-old female

F2: Martha presents for her annual exam

F3: Martha has been traveling a lot recently

...

Figure 8: Example Provider-Patient chat from ACI training set. We include the generated Subjective and the list and categorization of facts. Note that the chat and facts were truncated for length. We include the unique fact identifiers (F + NUMBER) for reference. For additional output, see Figure 9. All output was generated with GPT-4.

Supportive Sub-Clustering (Prompt 4)
Uncontrolled Hypertension
SOCIAL DETERMINATE OF HEALTH:
  Work-related stress:
    F15: Martha's blood pressure goes up when work gets more stressful
SYMPTOMS:
  Stress-induced hypertension:
    F15: Martha's blood pressure goes up when work gets more stressful
TESTS: NONE
TREATMENTS:
  Medication non-adherence:
    F14: Martha is forgetting to take her blood pressure medication
Well-managed Congestive Heart Failure
OTHER:
  Past medical history:
    F1: Martha has a past medical history significant for congestive heart failure
SOCIAL DETERMINATE OF HEALTH: NONE
SYMPTOMS:
  NONE:
    F10: Martha has no symptoms like chest pains, shortness of breath, any swelling in her legs
TESTS: NONE
TREATMENTS:
  Medication adherence:
    F9: Martha is taking her medication
...

Refuting Sub-Clustering
Uncontrolled Hypertension
SOCIAL DETERMINATE OF HEALTH: NONE
SYMPTOMS: NONE
TESTS: NONE
TREATMENTS:
  Antihypertensive Medication Adherence:
    F14: Martha is forgetting to take her blood pressure medication
    F9: Martha is taking her medication
Well-managed Congestive Heart Failure
SYMPTOMS:
  Cardiac Insufficiency:
    F10: Martha has no symptoms like chest pains, shortness of breath, any swelling in her legs
Social Determinate of Health: NONE
TESTS: NONE
TREATMENTS: NONE
...

Omitted Facts, Generated Explanation, and Omission Weight (Prompt 2 for extraction and explanation)
F17: Martha has a lot of fiscal year end data that she's been having to pull
  The summary does not mention the specific work-related stressors Martha is dealing with. (Score : 0.5)
F18: Martha has a lot of responsibility at work
  The summary does not mention the level of responsibility Martha has at work. (Score : 0.5)
F4: Martha got her vaccine
  The summary does not mention that Martha received her vaccine. (Score: 0.1)
F6: Martha went to Washington last weekend to hike in northern cascades, like around the mount baker area (Score : 0.1)
  The summary does not specify the location of Martha's recent hiking trips. (Score : 0.1)
F8: Martha has been resisting from the tater tots, the soft pretzels, the salty foods (Score : 0.5)
  The summary does not detail the specific foods Martha has been avoiding in her low sodium diet.

Figure 9: Following from Figure 8, the Supportive and Refuting Sub-clustering, and the list of missing facts. Note that there are seemingly conflicting facts in the Refuting sub-clustering example. However, this represents exactly what was discussed in the chat. Initially, the patient says they are taking their medication, and later says they are forgetting their blood pressure medication specifically.

```
Instructions
-The following is a medical summary of a single medical encounter.  In addition, there is a list of facts from that same encounter.
-Acting as a medical expert who is testing medical students on their thoroughness, which facts were omitted from the summary?
-For a fact to be an omission, relevant information from the fact must be omitted.  The fact does not have to be written verbatim.
-Output the list of facts that were omitted, report the fact id, fact, and a short explanation.

--Begin Summary--
{{subjective}}
--End Summary--
--Begin Facts--
{{fact_list}}
--End Facts--

Are there any facts missing from the summary? Report the fact number, the fact, and an explanation for each.

The output should be in a json dictionary, with the following format;
{
"FACT_NUM" : ["FACT", "EXPLANATION"]
...
}
If there are no missing facts, return an empty json dictionary.

Missing facts:
```

Prompt 2: Prompt for detecting fact omissions from summary

```
You are an expert medical data labeler. You will be provided with a differential diagnosis (DDx) for a patient case and a set of medical facts describing the patient. Your task is to group these facts into 3 groups: "critical", "important", and "other". "Critical" facts are absolutely critical in order to arrive at the DDx. If this fact is not present, the DDx would be greatly altered. "Important" facts are helpful in determining the DDX, and may or may not greatly affect the DDx. "Other" facts are facts that are neither "critical" nor "important".

---Differential diagnosis (start)---
{{ddx}}
---Differential diagnosis (end)---

---Medical facts (start)---
{{facts}}
---Medical facts (end)---

Given this information, produce a numbered, ranked list of unique grouped facts.
For each category, output the category name ("Category|[CATEGORY]\n") followed by the list of facts for that category each on its own line ("[Fact_Rank]|[Fact Num]|[Fact]").

Output:
```

Prompt 3: Prompt for assigning categories to each prompt

```
The following is a list of facts extracted from a medical encounter.


Your role is to select which  positive fact(s) support each diagnosis.
Therefore, only report pertinent positives which support each diagnosis.  Do
    not report supportive results that negate the diagnosis, or any other type
    of fact.


A fact can occur in multiple diagnoses.

The classifications should be in reference to this differential diagnosis;
{{ddx}}

Facts:
{{facts}}


Output the results in a json dictionary, such as;
{
"DIAGNOSIS 1" : {"FACT_NUM" : "EXPLANATIION" ...}
...
}
If a diagnosis has no facts, output an empty array.

Clusters:
```

Prompt 4: Prompt for clustering supportive facts by diagnosis

```
The following is a patient-doctor dialogue.


{{dialogue}}

Consider the conversation in the frame of a SOAP medical note framework.
We want to include all dialogue lines that contain information that might be
    relevant to the subjective.
This includes;
-Chief Complaint
-History of Present Illness
--This includes questions about the patient's current health status.
-Past medical history
--The includes any discussion of previously diagnosed medical issues.
This does not include;
-Physical exam
-Laboratory Results
-New diagnoses made by the provider in this conversation
-Assessment or care plan
Return the last line of the conversation that collects this information.

The conversation begins with line number 0.
Output the entire relevant line in a valid json dictionary formatted as follows
    ;
{
[LINE_NUM] : [MSG]
}
Where [LINE_NUM] is a valid integer, and [MSG] is the relevant message.


Output:
```

Prompt 5: Prompt for truncating dialogue