

# Counterfactual Simulatability of LLM Explanations for Generation Tasks

Marvin Limpijankit, Yanda Chen, Melanie Subbiah,  
Nicholas Deas, and Kathleen McKeown

Department of Computer Science, Columbia University  
m14431@columbia.edu

## Abstract

LLMs can be unpredictable, as even slight alterations to the prompt can cause the output to change in unexpected ways. Thus, the ability of models to accurately explain their behavior is critical, especially in high-stakes settings. Counterfactual simulatability measures how well an explanation allows users to infer the model’s output on related counterfactuals and has been previously studied for yes/no question answering. We provide a general framework for extending this method to generation tasks, using news summarization and medical suggestion as example use cases. We find that while LLM explanations do enable users to better predict their outputs on counterfactuals in the summarization setting, there is significant room for improvement for medical suggestion. Furthermore, our results suggest that evaluating counterfactual simulatability may be more appropriate for skill-based tasks as opposed to knowledge-based tasks.

## 1 Introduction

While large language models (LLMs) have proven effective for a diverse range of applications, their outputs still often contain hallucinations of unsupported information (Ji et al., 2023) or biases (Sheng et al., 2021) that hinder their reliability on critical language generation tasks. At the same time, it is important that users themselves can accurately evaluate the model capabilities, knowledge, and reliability (Steyvers et al., 2025). Particularly in high stakes domains, such as medical applications, misunderstandings or the lack of ability to predict model behavior on unseen inputs can pose disastrous risks to users (Michalowski et al., 2024).

To anticipate these risks, recent work has turned attention to evaluating the reliability of LLM explanations (Madsen et al., 2024; Turpin et al., 2023; Kunz and Kuhlmann, 2024). In particular, Chen et al. (2024) evaluates the *counterfactual simulatability*

of LLM explanations for yes/no question answering. Counterfactual simulatability is a measure of how well a model’s explanation allows humans to correctly infer the model’s predictions on simulatable counterfactuals (unseen inputs where the explanation should enable the user to confidently guess the model’s output). Furthermore, according to Chen et al. (2024), the counterfactual simulatability of an explanation can be decomposed into *simulation generality*, a measure of the diversity of simulatable counterfactuals and *simulation precision*, a measure of the proportion of these counterfactuals for which humans correctly infer the model’s output. Ideal model explanations should balance both generality and precision.

Counterfactual simulatability, however, is equally critical to generation tasks, where the larger space of possible outputs makes understanding a model’s decision process more challenging. To fill this gap, we formalize a framework for evaluating counterfactual simulatability in language generation tasks (Figure 1). To evaluate a model’s explanation, we first use a separate LLM to (1) decompose the explanation into atomic units and (2) generate relevant counterfactuals. Then, an annotator (human or the LLM) evaluates the simulatability and precision of each unit of the model’s explanation using the counterfactual and the model’s output on the counterfactual respectively. Finally, generality and precision scores are calculated following §4.4. *Rather than evaluating the factual correctness of explanations, our framework measures the ability of LLMs to accurately describe their behavior in a generative setting. We evaluate whether LLMs’ explanations lead to reliable human mental models that are consistent with their outputs.*

We apply the framework to two tasks: news summarization using CNN/DM (Nallapati et al., 2016) and medical suggestion generation using the Taiwan e-Hospital Dataset (Chen et al., 2022b). These tasks involve different *trade-offs between*

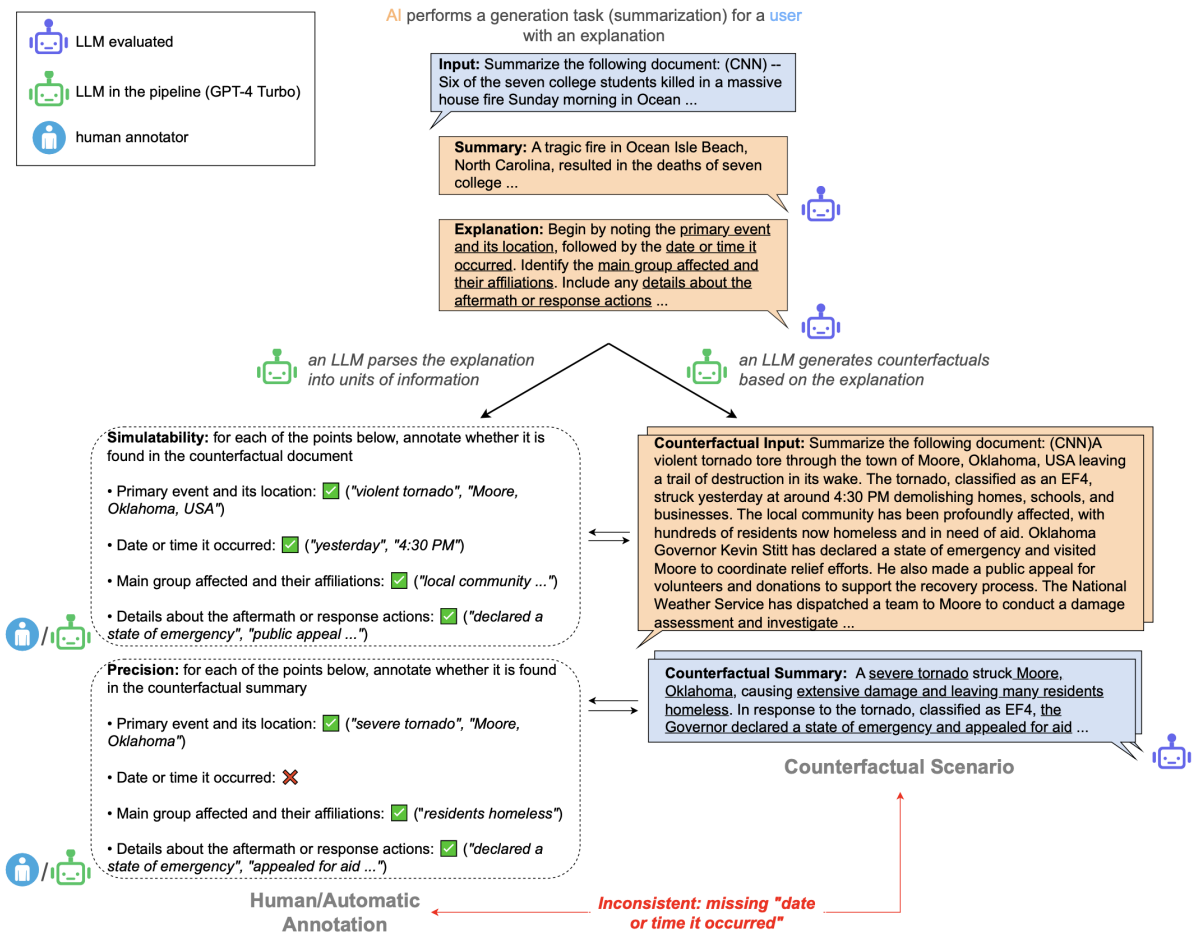


Figure 1: Our evaluation pipeline. Given a model’s explanation, an LLM is prompted to decompose the explanation into atomic units (left) and generate relevant counterfactuals (right). For each unit, an annotator verifies whether the element appears in the counterfactual (simulatability) and the counterfactual output (precision).

*generality and precision.* Explanations for news summarization can be highly general as models may employ similar approaches for documents that contain common elements (e.g., dates, quotes) but differ in content. In contrast, medical suggestion explanations may be less general as the model’s response can be highly dependent on the user input (i.e., minor variances in expressed symptoms can lead to different suggestions). This can lead to a more limited set of simulatable counterfactuals, though potentially being more precise in its outputs. With these generation tasks, we conduct an initial evaluation of LLMs’ explanations and assess where models’ explanations in these generation tasks fail. In summary, our contributions are as follows:

1. We propose a framework to measure the ability of LLMs to accurately describe their behavior in a generative setting by evaluating whether LLMs’ explanations lead to mental models consistent with their outputs (counterfactual simulatability).

2. We assess the feasibility of using an LLM to automate human annotation in our evaluation pipeline and find that doing so achieves agreement with humans comparable to agreement among human annotators.
3. Using our framework, we evaluate Chain-of-thought and Post-hoc explanations for multiple LLMs on two complementary tasks: news summarization and medical suggestion generation. We show that LLM explanations lead to reliable mental models in news summarization, but not medical suggestion.<sup>1</sup>

## 2 Related Work

**Human mental models.** Humans form mental models of the physical world as a whole (Gentner and Stevens, 2014) as well as specific technologies (e.g., Payne, 1991; Du et al., 2018; Lei et al., 2016)

<sup>1</sup>Our code is available at [github.com/mlimpijankit/co-uncertain-simulatability-generation-tasks](https://github.com/mlimpijankit/co-uncertain-simulatability-generation-tasks)

through their past experiences and observations. Specifically with regard to artificial intelligence systems, explanations of model predictions have been considered high quality if they provide users with an accurate and generalizable understanding of the system in the form of mental models (Rutjes et al., 2019; Merry et al., 2021). When such explanations are successful, they can improve users’ ability to effectively use AI models (Vasconcelos et al., 2022; Senoner et al., 2024) as well as to anticipate and correct undesirable model behaviors, such as biases and incorrect predictions (Bansal et al., 2019). We specifically evaluate LLMs’ explanations in generation tasks considering their ability to help form mental models.

**Explanation evaluation.** A variety of different approaches have been used to evaluate the quality and utility of model-generated natural language explanations and rationales. Studies evaluating natural language explanations similar to word attributions (Huang et al., 2023; Madsen et al., 2024) as well as unconstrained explanations (Turpin et al., 2023) have focused on faithfulness measures. In particular, work has proposed metrics for dimensions including comprehensiveness (DeYoung et al., 2020), sufficiency (DeYoung et al., 2020), alignment with human rationales (Fayyaz et al., 2024), plausibility (Wojciechowski et al., 2024), and scrutability (Xu et al., 2023) among others. In contrast to intrinsic measures of model explanations, other work evaluates explanations extrinsically through impact on task performance (e.g., Camburu et al., 2018; Wei et al., 2022; Krishna et al., 2023). Among these topics, relatively few works have investigated natural language explanations in language generation tasks; such studies include dialogue responses (Zhou et al., 2021), dialogue understanding (Gao et al., 2024), and more prominently, open-ended question answering (Ho et al., 2023; Fragkathoulas and Chlapanis, 2024; Lyu et al., 2023). While Chen et al. (2024) introduces *counterfactual simulatability* in a binary classification setting, we propose a novel framework for counterfactual simulatability in language generation settings.

### 3 Counterfactual Simulatability for Generation Tasks

Explanations can be evaluated by considering whether an observer, having seen a model’s explanation for some input, can infer (i.e., simulate) the

model’s output on a related counterfactual input. For instance, if a user asks “*do dolphins swim?*” and a model answers “*yes*” with the explanation “*all aquatic animals swim*”, then the user would infer that when asked the counterfactual “*do starfish swim?*”, the model will similarly answer “*yes*”. If, in reality, the model answers “*no*” to this question, then, as Chen et al. (2024) notes, the explanation is ineffective because it creates a mental model that is inconsistent with the model’s behavior, even though the answer may be factually correct. More specifically, counterfactual simulatability measures how accurate these mental models are on simulatable counterfactuals, unseen inputs where the explanation allows the user to confidently predict the model’s output (Chen et al., 2024).

However, in contrast to the classification example, generation tasks have a much larger output space (e.g., there exists many possible summaries for a news document). This makes simulation extremely challenging as it is impossible to precisely identify a single possible output based on an explanation. For instance in news summarization, if a user is shown the explanation “*the summary should include the key event*” and a counterfactual document on “*the opening ceremony of the 2024 Paris Olympics*”, while the user can logically infer that the summary will include the opening ceremony, it is impossible to predict the exact wording in which it will appear. Explanations typically cannot enable humans to pinpoint a single model output. However, they are still very useful if they help humans narrow down the possible outputs (e.g., refining “all possible summaries” to “summaries that mention the opening ceremony”).

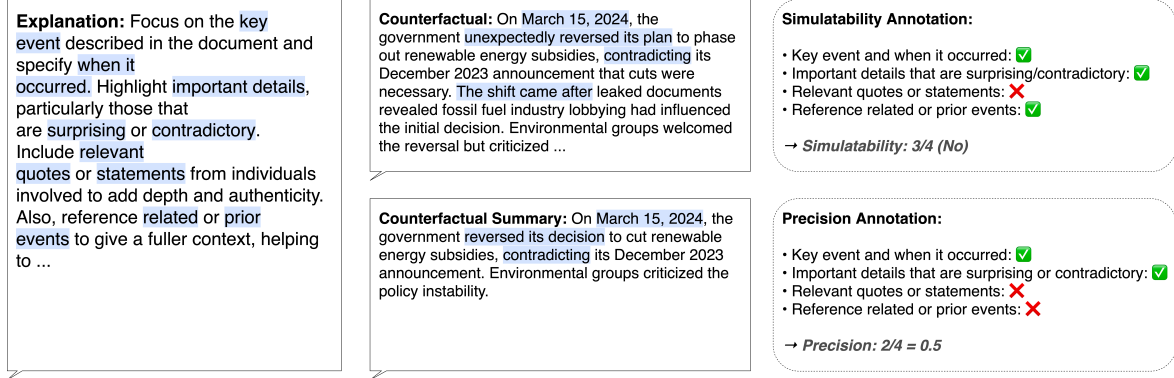
#### 3.1 Notation

For a given generation task, a model  $M$  takes an input  $x \in X$  and produces an output  $o_x \in O$  and a corresponding explanation  $e_x$ . Here, the input, output, and explanation are all natural language and, in the case of generation,  $|O|$  may be arbitrarily large. A human observes  $x, e_x$ , and forms a one-to-many mental model  $h_{x,e_x} : X \rightarrow \mathcal{P}(O)$ , where  $\mathcal{P}(O)$  denotes the power set of  $O$ , and  $h_{x,e_x}(x')$  denotes what the human infers to be  $M$ ’s possible outputs on a counterfactual  $x'$ . For simplicity,  $h_{e_x}(x')$  is used to denote  $h_{x,e_x}(x')$ .

#### 3.2 Simulatability

Simulatability is a metric on counterfactuals that describes whether the model’s output on that coun-

## News Summarization



## Medical Suggestion

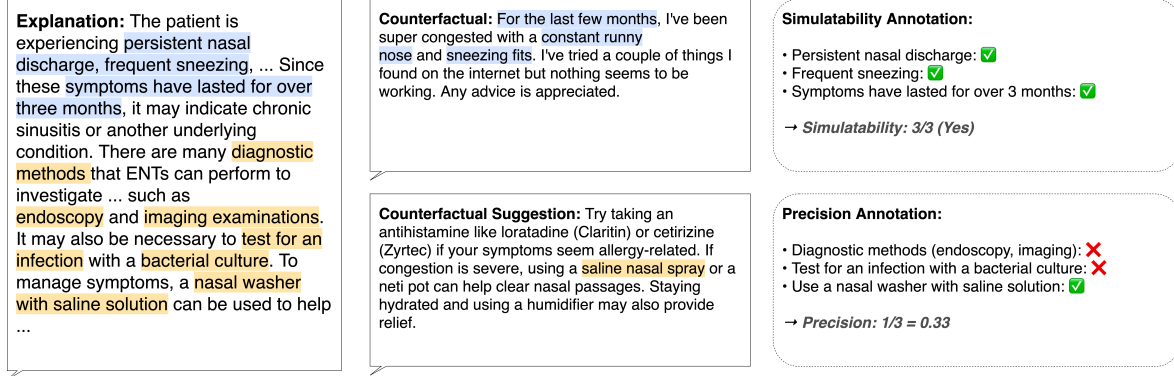


Figure 2: Example explanations, counterfactuals, counterfactual outputs, and annotations for news summarization and medical suggestion. Atomic units of the explanation are highlighted (for medical suggestion, blue: patient information, orange: suggestions).

terfactual can be inferred given an explanation. More specifically, a simulatable counterfactual is one where an observer, using the mental model produced by an explanation, can refine their expectation of the model’s output (i.e.,  $|h_{x,e_x}(x')| \ll O$ ). Non-simulatable counterfactuals are omitted from our evaluation as the explanations are not helpful in these cases.

### 3.3 Simulation generality

Simulation generality is a metric on explanations that measures the diversity of simulatable counterfactuals an explanation leads to. For instance, the explanation “*the summary should include the key event*” is more general compared to “*the summary should mention the opening ceremony of the 2024 Paris Olympics*” since the former can be applied to documents covering a broader range of topics. It is calculated as one minus the average pairwise similarity between simulatable counterfactuals:

$$\text{generality} = 1 - \mathbb{E}_{x',x'' \sim p, x' \neq x''} [\alpha(x', x'')] \quad (1)$$

where  $p$  is the distribution of simulatable counterfactuals and  $\alpha$  is a similarity metric such as cosine similarity. Explanations with higher generality are more useful since they enable users to infer the model’s behavior on a wider range of possible inputs.

### 3.4 Simulation precision

Simulation precision is another metric on explanations and measures whether the model’s actual output  $M(x')$  is within the observer’s inferred output space:

$$\text{precision} = \mathbb{E}_{x' \sim p} [\mathbf{1}[M(x') \in h_{e_x}(x')]] \quad (2)$$

Explanations with higher precision scores indicate that the mental models they produce are more aligned with the model’s actual behavior. In practice, it is unfeasible for humans to enumerate all valid outputs in  $h_{e_x}(x')$  due to the large output space. Thus, in §4.4 we propose a method to estimate mental models by evaluating the atomic units of a model’s explanation.



## 4 Methods

### 4.1 Datasets

We select two generative tasks that represent practical LLM use cases and whose natural language explanations differ in generality and precision. For the first task of **news summarization**, we use the CNN/DM dataset (Nallapati et al., 2016). Summarization is chosen as it is a well-studied application for language generation. Additionally, this task allows us to study high-level, abstract explanations since summarization does not rely on facts learned during the LLM’s pre-training, rather its ability to extract key elements from the input document.

Conversely, we also investigate **medical suggestion**, a domain where explanations are critical. For this task, we use the Taiwan e-Hospital Dataset, a collection of 86,399 Mandarin question answer pairs from the online health website Taiwan e-Hospital (Chen et al., 2022b). The data was translated into English using the Google Translate API. We select this specific dataset because each sample consists of a question, suggestion, explanation triplet, ensuring that the questions are complex enough such that explanations are helpful. Furthermore, the dataset also mirrors the practical setting we aim to study, where everyday users interact with LLMs to seek general medical advice without any specialized knowledge.

The contrast between explanations for these tasks is illustrated in Figure 2. Medical suggestions are more knowledge-based, requiring the LLM to identify key aspects of the query (e.g. expressed symptoms, relevant medical history) and generate suggestions using knowledge encoded in the model. As such, the explanations are highly specific to the input content. In the example above, the summarization explanation applies to counterfactuals containing any “*key event*” or “*quotes or statements*” whereas the medical suggestion explanation can only be used to infer the model’s output when a user expresses “*persistent nasal discharge*”, “*frequent sneezing*”, and “*symptoms [that] have lasted for over three months*”. While medical suggestion explanations likely apply to a more limited set of counterfactuals, they allow the user to infer specific pieces of information rather than high-level elements as is the case for summarization. Therefore, these tasks are complementary for studying counterfactual simulatability as they represent different requirements for generality and precision.

### 4.2 Models

Our proposed framework uses two models: the LLM being evaluated and the LLM used to automate parts of the pipeline (e.g., counterfactual generation, explanation decomposition, annotation). We evaluate Anthropic’s Claude 3.7 Sonnet, OpenAI’s GPT-4 and GPT-4 Turbo, and Meta’s Llama 3.3 70B Instruct. These represent a few popular proprietary and open-source models. To automate steps in the pipeline, we opt for GPT-4 Turbo due to its competitive performance at a low cost.

### 4.3 Explanations

Through prompting (instructions, few-shot examples), we guide the LLMs to generate explanations that emphasize the difference in generality and precision across tasks. For news summarization, to be highly generalizable, we encourage the model to explain its decision process at a high-level, identifying abstract elements while avoiding reference to specific topics. For medical suggestion, we instruct the model to first identify important details in the user’s question, propose possible underlying causes, and suggest recommended actions accordingly. These approaches reflect the skill-based vs. knowledge-based nature of our tasks. Additionally, following Chen et al. (2024), we experiment with both Chain-of-thought and Post-hoc prompting to produce explanations (Camburu et al., 2018). The prompts used are provided in Appendix D.

### 4.4 Estimating mental models

Generation tasks are challenging for counterfactual simulatability because mental models are extremely difficult to estimate given the large space of possible outputs. Motivated by previous work on atomic units (Wright et al., 2022; Chen et al., 2022a; Kamoi et al., 2023), we introduce a method for estimating mental models by decomposing explanations into units of information as a proxy.

Intuitively, explanations  $e_x$  identify key pieces of information that the model considers essential for the output. Thus, an observer may simulate that if the counterfactual input  $x'$  includes a piece of information  $a$  deemed important in  $e_x$ , the model should similarly take into account  $a$  in its output for  $x'$ . Based on this explanation  $e_x$ , an observer’s mental model  $h_{e_x}$  can be formalized as:

$$h_{e_x}(x') = \{o \in O \mid \forall a \in e_x \cap x', a \in o\} \quad (3)$$

where the space of possible outputs on the counterfactual is refined to only outputs that contain the elements  $\{a\}$ . With this, we formulate simulatability, generality, and precision for each explanation, counterfactual pair as follows. First, the simulatability of a counterfactual is determined by verifying if all atomic units of the explanation appear in the counterfactual:

$$\text{simulatability} = \mathbb{1}\{\forall a \in e_x, a \in x'\} \quad (4)$$

Then, all non-simulatable counterfactuals are discarded and the generality of an explanation is calculated as one minus the average pairwise cosine similarity between simulatable counterfactuals (Equation 1). Finally, by identifying whether each unit of the explanation is addressed in the model’s output on the counterfactual, a precision score is calculated as:

$$\text{precision} = \frac{|\{a \in e_x \cap M(x')\}|}{|\{a \in e_x\}|} \quad (5)$$

Note that this differs from the formulation in §3.4. Precision scores are averaged across counterfactuals for a given explanation.

**Explanation Decomposition.** The atomic units of an explanation should highlight aspects that are considered important to the model, and thus, a user would expect in the output accordingly. These effectively describe a user’s mental model and should ideally be separate, non-overlapping points. One key distinction between the tasks is that while each atomic unit of a summarization explanation is linked to an expected item in the input (for simulatability) and its reference in the output (for precision), medical suggestion explanations do not follow this one-to-one mapping of units between input and output. For example in Figure 2, while “*Persistent nasal discharge*” is identified as important, it is unclear which suggested action the model has generated in response to this symptom. To address this, for the medical domain, we further instruct the LLM when decomposing the explanation to classify units into two categories, patient information and suggestions. Atomic units from these categories are then used to evaluate simulatability and precision respectively during annotation (Figure 2).

#### 4.5 Experimental setup

We run two main experiments: (1) a human evaluation where GPT-4 Turbo explanations are eval-

uated using human annotations and (2) a larger-scale, fully automatic evaluation of multiple models (Claude 3.7 Sonnet, GPT-4, Llama 3) using GPT-4 Turbo as the annotator. This allows us to assess the feasibility of using an LLM as an annotator before adopting the automatic approach that can scale to larger experiments.

**Human evaluation.** Using randomly sampled inputs from our dataset, we prompt GPT-4 Turbo to *generate outputs and explanations*. Note that here, GPT-4 Turbo is both the model being evaluated and the model being used in the pipeline. For each explanation, we use GPT-4 Turbo to *generate 3 relevant counterfactuals*, providing the explanation in the context. Then, we use GPT-4 Turbo to *decompose the explanation into atomic units* (i.e.,  $e_x \rightarrow \{a\}$ ) and, in the medical suggestion use case, instruct it to extract units into two groups for patient details or suggested actions. Finally, we *generate counterfactual outputs* using the same prompt as the first step and have humans examine the counterfactual and counterfactual output to *annotate simulatability and precision* for each atomic unit of the explanation.

Additionally, we assess the ability of GPT-4 Turbo to break down explanations by instructing the annotators to indicate whether each unit was extracted correctly and whether details that should have been extracted are missing. We also briefly investigate if increasing the number of counterfactuals generated per explanation affects generality, but find no noticeable differences (Appendix C).

We assigned 3 student annotators per task to evaluate 15 explanations along with 3 generated counterfactuals each (45 explanation, counterfactual pairs total). Each annotator was given an overlapping set of 3 explanations (to measure human-human agreement) plus an additional 4. Since they annotated the atomic units of each explanation, the exact number of annotations varied but approximately resulted in 260 annotations per person (see Table 2). The annotators were not required to have any specialized domain knowledge. One annotator (news summarization annotator #3 in Table 1) did not complete all annotations, leading to a slight discrepancy in annotations across tasks. The prompts and annotations instructions are provided in Appendix D and E respectively.

**Comparison of human-LLM annotations.** We evaluate the feasibility of using an LLM to automate human annotation by prompting GPT-4 Turbo

Annotator	News Summarization					Medical Suggestion				
	1	2	3	GPT-4 Turbo	n	1	2	3	GPT-4 Turbo	n
1	-	0.35	0.71	0.64	263	-	0.76	0.74	0.68	273
2	0.35	-	0.57	0.48	213	0.76	-	0.73	0.54	261
3	0.71	0.57	-	0.71	73	0.74	0.73	-	0.74	258

Table 1: Inter-annotator agreement (Cohen’s Kappa) between human annotators and GPT-4 Turbo for news summarization and medical suggestion.

with similar instructions and measuring the inter-annotator agreements between human-human and human-LLM pairs.

**Automatic evaluation.** Finally, using GPT-4 Turbo as the annotator, we scale up our experiment to evaluate three LLMs using more data and generating five counterfactuals per explanation. For the automatic evaluation, 50 explanations along with 5 generated counterfactuals each (250 explanation, counterfactual pairs total) were evaluated.

## 5 Results

### 5.1 Intermediate results

**GPT-4 Turbo is able to parse explanations well for summarization but not for medical suggestion.** Annotators find that the LLM successfully breaks down explanations with 96% and 57% accuracy for news summarization and medical suggestion respectively. For medical suggestion errors, the LLM either fails to identify a key detail or extracts an atomic unit incorrectly (extracted a unit it should not have or classified it as the incorrect type of information). Further investigation reveals that errors often involve minor details whereas the main points are successfully extracted (see Appendix A).

**GPT-4 Turbo is able to generate simulatable counterfactuals for summarization but not for medical suggestion.** While almost all (74/76) generated counterfactuals are deemed simulatable in the summarization setting, only slightly more than half are for medical suggestion (52/90). Most of these non-simulatable counterfactuals contain 0.4 – 0.8 of the atomic units of the explanation, indicating that counterfactual generation is more challenging in this setting. The distribution of generated counterfactuals, sorted by the proportion of atomic units that appear is shown in Appendix B.

### 5.2 Human evaluation

GPT-4 Turbo explanations lead to generalizable counterfactuals and consistent mental models in the case of summarization, with approximately 0.8 of the inferred information appearing in the counterfactual output. In contrast, explanations in the medical suggestion setting are less generalizable and less precise (approximately 0.5), indicating that LLMs may struggle more to reliably explain their behavior for this task. Chain-of-thought and Post-hoc explanations lead to similar results (see Table 2). For each pair of settings, we fit an independent sample t-test and find that the difference in both metrics when compared across tasks are significant ( $p < 0.05$ ), but not when compared across explanation types within the same task.

Task	Explanation	Generality	Precision
News Summarization	Chain-of-thought	0.52	0.81
	Post-hoc	0.49	0.89
Medical Suggestion	Chain-of-thought	0.20	0.51
	Post-hoc	0.26	0.59

Table 2: Generality and precision scores for GPT-4 Turbo from the human evaluation.

As a sanity check for the pipeline, we generate counterfactual outputs conditioned on the original explanation and verify that these result in a precision score of 1.00. This validates that our framework captures how well models follow their explanations. Therefore, the low precision scores for medical suggestion are due to the model’s behavior not adhering to its explanation on that task rather than the evaluation setup.

Task	Model	Chain-of-thought				Post-hoc			
		# Expl	# Samples	Generality	Precision	# Expl	# Samples	Generality	Precision
News Summarization	Claude 3.7 Sonnet	48	189	0.67	0.93	45	173	0.62	0.84
	GPT-4	46	160	0.59	0.84	48	193	0.65	0.78
	Llama 3	47	172	0.67	0.74	43	153	0.67	0.66
Medical Suggestion	Claude 3.7 Sonnet	24	55	0.21	0.48	24	78	0.20	0.66
	GPT-4	36	103	0.20	0.46	36	122	0.19	0.65
	Llama 3	24	69	0.19	0.56	30	110	0.20	0.66

Table 3: Generality and precision metrics across models, explanation types, and tasks from the automatic evaluation.

### 5.3 Automatic evaluation

**GPT-4 Turbo is able to approximate human annotation for our tasks.** Table 1 reports the pairwise Cohen’s Kappa between each pair of annotators and the LLM. We find that overall, GPT-4 Turbo achieves similar inter-annotator agreement to humans, with an average of 0.61 for human-LLM pairs compared to 0.54 for human-human pairs in news summarization and 0.65 for human-LLM pairs compared to 0.74 for human-human pairs in medical suggestion. We calculate a two-sided p-value for ratings between all annotation pairs and find that the observed agreement is significant.

The results of the automatic evaluation, presented in Table 3, are in line with our findings from the human evaluation. Namely, models are better at accurately explaining their behavior for summarization compared to medical suggestion while also remaining general such that these explanations apply to diverse counterfactuals. Chain-of-thought explanations lead to better mental models for summarization, which may reflect the skill-based nature of this task. On the other hand, Post-hoc explanations lead to more precise explanations for medical suggestion. Models may also differ in their ability to describe their behavior in a generative setting, for instance, Claude 3.7 Sonnet demonstrates noticeably better precision scores for news summarization compared to other models.

## 6 Discussion

**LLM explanations can be helpful for skill-based generation tasks but may struggle for knowledge-based generation tasks.** While model explanations do enable users to infer pieces of information that will appear in counterfactual outputs in both settings, the mental models pro-

duced are more accurate for summarization compared to medical suggestion. Furthermore, summarization explanations, which employ a high-level approach to their explanation, are also more generalizable. This may suggest that LLMs are better able to describe their behavior for skill-based tasks, where users can reliably infer the elements that will appear in the outputs, compared to knowledge-based tasks, where users are less capable of inferring specific points (e.g., suggestions). Additionally, utilizing Chain-of-thought prompting, which aligns with the skill-based nature of summarization, may lead to more precise explanations. It is important to consider that our evaluation only examined news summarization and medical suggestion specifically. As such, the results may vary for other skill-based and knowledge-based tasks.

These findings may reflect the predictability of model behavior on different tasks types. For medical suggestion, minor variances in the question may lead to very different answers. For instance, if a user describes “*I experience chest pain sometimes when I exercise.*” a model might respond “*consider reducing exercise intensity, try warming up thoroughly before exercise ...*”. However, if a user changes their question slightly to “*I experience chest pain sometimes when I exercise. I started taking a new pre-workout supplement with high caffeine recently.*”, although they are expressing the same symptoms, because of the presence of additional information the model might respond “*discontinue the supplement or use smaller dosages, the chest pain is likely related to caffeine-induced heart palpitations ...*”. These differences in the sensitivity of the model’s output relative to the input may lead to the observed differences in precision and generality in our experiments.



## **Our counterfactual simulatability evaluation framework is effective in the summarization setting but less suited for medical suggestion.**

Although LLMs are able to automate the human annotation steps in our evaluation, they demonstrate issues in other aspects for the medical suggestion task. Specifically, our annotators identify many errors in the parsing of explanations into atomic units, where the LLM misses key information or mis-classifies the information. In one instance, “possible heat-induced asthma” is incorrectly extracted as a key symptom when in reality it is a potential cause. Unlike the summarization setting, this added requirement of classifying the type of information introduces more complexity, making the LLM a less effective tool in the evaluation pipeline. Additionally, we found that the LLM is unable to produce as many simulatable counterfactuals in the medical domain compared to summarization. There is significant room for improvement towards adapting counterfactual simulatability for knowledge-based tasks like medical suggestion, where the explanation and decision process for an LLM is less explicit and relies heavily on knowledge encoded in the model.

## **7 Limitations**

First, the generality metric relies on our LLM (GPT-4 Turbo) being able to generate diverse, yet simulatable counterfactuals. Although we briefly experimented with counterfactual generation, more work can be done to assess whether other models or prompting strategies may lead to higher quality counterfactuals. Additionally, cosine similarity might not be the optimal way to judge diversity across different tasks as it is heavily content-based. For example, while many counterfactuals for summarization followed the same structure (e.g., a headline followed by when/where it occurred followed by a quote ...), they were still rewarded with high generality because the specific topics and words differed.

Second, in a few cases there was ambiguity when determining whether atomic units were ‘present’ in the counterfactual/the counterfactual output. Since these units of information may appear with different wording or even be implied, we encouraged annotators to use their best judgement and leave notes. We leave investigation into ways of improving this matching to future work.

Finally, we focused on demonstrating our frame-

work using explanations generated from a few popular LLMs. Testing newer types of models (e.g., reasoning models) and different task setups (e.g., multiple interactions) are a promising direction. Furthermore, expanding the set of metrics to consider aspects such as explanation comprehensiveness, faithfulness, and plausibility, and assessing how they vary with counterfactual simulatability, may also be worthwhile.

**Supplementary Materials Availability Statement:** Our source code as well as an anonymized version of the data used in the human evaluation is made available at [github.com/mlimpijankit/counterfactual-simulatability-generation-tasks](https://github.com/mlimpijankit/counterfactual-simulatability-generation-tasks).

## **8 Acknowledgments**

This research is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract #2022-22072200005 and the National Science Foundation and by DoD OUSD (R&E) under Cooperative Agreement PHY-2229929 (The NSF AI Institute for Artificial and Natural Intelligence). The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense, the National Science Foundation or the U.S. Government. One of the authors holds an equity interest in OpenAI.

## **References**

- Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. 2019. *Beyond accuracy: The role of mental models in human-ai team performance*. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7:2–11.
- Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. *e-snli: Natural language inference with natural language explanations*. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.
- Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. 2022a. Generating literal and implied subquestions to fact-check complex claims. *arXiv preprint arXiv:2205.06938*.
- Wei-Lin Chen, An-Zi Yen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2022b. *Learning to generate explanation from e-hospital services for medical suggestion*. In

- Proceedings of the 29th International Conference on Computational Linguistics*, pages 2946–2951, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Yanda Chen, Ruiqi Zhong, Narutatsu Ri, Chen Zhao, He He, Jacob Steinhardt, Zhou Yu, and Kathleen McKeown. 2024. Do models explain themselves? counterfactual simulatability of natural language explanations. In *Proceedings of the 41st International Conference on Machine Learning, ICML'24*. JMLR.org.
- Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. [ERASER: A benchmark to evaluate rationalized NLP models](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.
- Yuemeng Du, Jingyan Qin, Shujing Zhang, Sha Cao, and Jinhua Dou. 2018. [Voice User Interface Interaction Design Research Based on User Mental Model in Autonomous Vehicle](#), page 117–132. Springer International Publishing.
- Mohsen Fayyaz, Fan Yin, Jiao Sun, and Nanyun Peng. 2024. [Evaluating human alignment and model faithfulness of llm rationale](#). *Preprint*, arXiv:2407.00219.
- Christos Fragkathoulas and Odysseas Spyridon Chlapanis. 2024. [Local explanations and self-explanations for assessing faithfulness in black-box llms](#). In *Proceedings of the 13th Hellenic Conference on Artificial Intelligence, SETN '24*, New York, NY, USA. Association for Computing Machinery.
- Haoyu Gao, Ting-En Lin, Hangyu Li, Min Yang, Yuchuan Wu, Wentao Ma, Fei Huang, and Yongbin Li. 2024. [Self-explanation prompting improves dialogue understanding in large language models](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14567–14578, Torino, Italia. ELRA and ICCL.
- Dedre Gentner and Albert L. Stevens. 2014. *Mental Models*. Psychology Press.
- Matthew Ho, Aditya Sharma, Justin Chang, Michael Saxon, Sharon Levy, Yujie Lu, and William Yang Wang. 2023. [Wikiwhy: Answering and explaining cause-and-effect questions](#). In *The Eleventh International Conference on Learning Representations*.
- Shiyuan Huang, Siddarth Mamidanna, Shreedhar Jangam, Yilun Zhou, and Leilani H. Gilpin. 2023. [Can large language models explain themselves? a study of llm-generated self-explanations](#). *Preprint*, arXiv:2310.11207.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.
- Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. [Wice: Real-world entailment for claims in wikipedia](#). *arXiv preprint arXiv:2303.01432*.
- Satyapriya Krishna, Jiaqi Ma, Dylan Slack, Asma Ghandeharioun, Sameer Singh, and Himabindu Lakkaraju. 2023. [Post hoc explanations of language models can improve language models](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 65468–65483. Curran Associates, Inc.
- Jenny Kunz and Marco Kuhlmann. 2024. [Properties and challenges of LLM-generated explanations](#). In *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 13–27, Mexico City, Mexico. Association for Computational Linguistics.
- Tian Lei, Xu Liu, Lei Wu, Ziliang Jin, Yuhui Wang, and Shuaili Wei. 2016. [The Influence of Matching Degree of the User’s Inherent Mental Model and the Product’s Embedded Mental Model on the Mobile User Experience](#), page 320–329. Springer International Publishing.
- Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. [Faithful chain-of-thought reasoning](#). 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 305–329, Nusa Dua, Bali. Association for Computational Linguistics.
- Andreas Madsen, Sarath Chandar, and Siva Reddy. 2024. [Are self-explanations from large language models faithful?](#) In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 295–337, Bangkok, Thailand. Association for Computational Linguistics.
- Michael Merry, Pat Riddle, and Jim Warren. 2021. [A mental models approach for defining explainable artificial intelligence](#). *BMC Medical Informatics and Decision Making*, 21(1).
- Martin Michalowski, Szymon Wilk, Jenny M. Bauer, Marc Carrier, Aurelien Delluc, Grégoire Le Gal, Tzu-Fei Wang, Deborah Siegal, and Wojtek Michalowski. 2024. Manually-curated versus llm-generated explanations for complex patient cases: An exploratory study with physicians. In *Artificial Intelligence in Medicine*, pages 313–323, Cham. Springer Nature Switzerland.
- Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. [Abstractive text summarization using sequence-to-sequence RNNs and beyond](#). In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

- Stephen J. Payne. 1991. [A descriptive study of mental models†](#). *Behaviour & Information Technology*, 10(1):3–21.
- Heleen Rutjes, Martijn Willemsen, and Wijnand IJsselstein. 2019. Considerations on explainable ai and users’ mental models. In *Where is the Human? Bridging the Gap Between AI and HCI*, United States. Association for Computing Machinery, Inc. CHI 2019 Workshop : Where is the Human? Bridging the Gap Between AI and HCI ; Conference date: 04-05-2019 Through 04-05-2019.
- Julian Senoner, Simon Schallmoser, Bernhard Kratzwald, Stefan Feuerriegel, and Torbjørn Netland. 2024. [Explainable ai improves task performance in human–ai collaboration](#). *Scientific Reports*, 14(1).
- Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. [Societal biases in language generation: Progress and challenges](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4275–4293, Online. Association for Computational Linguistics.
- Mark Steyvers, Heliodoro Tejeda, Aakriti Kumar, Catarina Belem, Sheer Karny, Xinyue Hu, Lukas W. Mayer, and Padhraic Smyth. 2025. [What large language models know and what people think they know](#). *Nature Machine Intelligence*.
- Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don’t always say what they think: unfaithful explanations in chain-of-thought prompting. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS ’23*, Red Hook, NY, USA. Curran Associates Inc.
- Helena Vasconcelos, Matthew Jörke, Madeleine Grundle-McLaughlin, Ranjay Krishna, Tobias Gerstenberg, and Michael S Bernstein. 2022. When do xai methods work? a cost-benefit approach to human-ai collaboration. In *CHI Workshop on Trust and Reliance in AI-Human Teams*, pages 1–15.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS ’22*, Red Hook, NY, USA. Curran Associates Inc.
- Adam Wojciechowski, Mateusz Lango, and Ondrej Dusek. 2024. Faithful and plausible natural language explanations for image classification: A pipeline approach. *arXiv preprint arXiv:2407.20899*.
- Dustin Wright, David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Isabelle Augenstein, and Lucy Lu Wang. 2022. Generating scientific claims for zero-shot scientific fact checking. *arXiv preprint arXiv:2203.12990*.
- Zhichao Xu, Hansi Zeng, Juntao Tan, Zuohui Fu, Yongfeng Zhang, and Qingyao Ai. 2023. [A reusable model-agnostic framework for faithfully explainable recommendation and system scrutability](#). *ACM Trans. Inf. Syst.*, 42(1).
- Pei Zhou, Pegah Jandaghi, Hyundong Cho, Bill Yuchen Lin, Jay Pujara, and Xiang Ren. 2021. [Probing commonsense explanation in dialogue response generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4132–4146, Punta Cana, Dominican Republic. Association for Computational Linguistics.

## A Explanation Decomposition

We categorize the errors reported by our annotators for explanation decomposition using GPT-4 Turbo during the human evaluation. Examples of each type of error are also provided below. Note that errors are subjective and often involve small details.

Parsed Explanations	News Summarization (n = 26)	Medical Suggestion (n = 30)
Accuracy	0.96	0.57
<i>Breakdown of Incorrect Examples:</i>		
Missing Extraction	1	8
Incorrect Extraction	0	4
Missing and Incorrect Extraction	0	1

Table 4: Explanation parsing errors across tasks.

### Example of a Missing Extraction

*In the example below, “anxiety and stress” are mentioned in the explanation but not extracted.*

**Explanation.** The symptoms you describe, including chest tightness, difficulty breathing, and the presence of white phlegm, along with discomfort in the stomach and worsening symptoms throughout the day, suggest a complex interaction between gastrointestinal and possibly psychosomatic factors. Superficial gastritis diagnosed previously might not fully explain the severity and range of your symptoms, especially if the treatment provided little relief. The persistence and nature of your symptoms indicate the need for a more comprehensive gastrointestinal evaluation to rule out other conditions such as gastroesophageal reflux disease (GERD), which can cause acid reflux, chest discomfort, and respiratory symptoms. Additionally, the significant impact of anxiety and stress on your physical health suggests that addressing these psychological factors is crucial. Cognitive behavioral therapy (CBT) is effective in managing anxiety and may reduce the intensity of the symptoms by helping you develop better coping mechanisms.

### GPT-4 Turbo Parsed Explanation (Symptoms).

- chest tightness and difficulty breathing
- presence of white phlegm
- discomfort in the stomach
- symptoms worsen throughout the day
- previously diagnosed with superficial gastritis
- treatment for gastritis provided little relief
- Missing Extraction: anxiety and stress

### GPT-4 Turbo Parsed Explanation (Suggestions).

- conduct a comprehensive gastrointestinal evaluation to rule out conditions like GERD
- address psychological factors such as anxiety and stress
- consider Cognitive Behavioral Therapy (CBT) to manage anxiety and develop coping mechanisms

### Example of an Incorrect Extraction

*In the example below, “possible heat-induced asthma or heat intolerance” is incorrectly extracted in the symptoms category. The annotator noted that this was more of a potential cause rather than an expressed symptom.*

**Explanation.** The patient experiences difficulty breathing, particularly during the summer or in hot temperatures, which improves significantly in cooler environments such as air-conditioned rooms or when using a fan. This pattern suggests that the symptoms may be related to heat-induced asthma or a condition known as heat intolerance. Heat can lead to increased airway resistance and trigger respiratory symptoms in sensitive individuals. It is also possible that environmental allergens, which are often more prevalent in warm weather, could be exacerbating the condition. To manage these symptoms, staying cool and avoiding heat exposure is crucial. However, a thorough evaluation by a healthcare professional is necessary to rule out other respiratory conditions and to confirm the diagnosis.

### GPT-4 Turbo Parsed Explanation (Symptoms).

- difficulty breathing, especially during summer or in hot temperatures
- improvement in symptoms in cooler environments
- possible heat-induced asthma or heat intolerance (Incorrect Extraction)

### GPT-4 Turbo Parsed Explanation (Suggestions).

- stay cool and avoid heat exposure
- undergo a thorough evaluation by a healthcare professional to rule out other respiratory conditions and confirm the diagnosis



## B Counterfactual Generation Simulatability

For each counterfactual in the human evaluation, we measure the extent to which they are simulatable by the proportion of atomic explanation units they contain. This distribution is displayed in Table 5.

Proportion of atomic units present in counterfactual (human annotated)	News Summarization	Medical Suggestion
1.00*	74	52
0.80–0.99	2	0
0.60–0.79	2	20
0.40–0.59	0	17
0.20–0.39	0	1
0.00–0.19	0	0
<b>Total</b>	<b>76</b>	<b>90</b>

Table 5: Distribution of the proportion of atomic units present in GPT4-Turbo generated counterfactuals across tasks. \*Indicates the set of simulatable counterfactuals.

## C Counterfactual Generation Generality

We assess how generality and simulatability changes as the number of counterfactuals generated per explanation is increased. Since news summarizations are expensive to generate due to the prompt length, we only use 10 explanations in this experiment compared to 30 for medical suggestion. Note that in the summarization and 10 counterfactuals setting, not all 100 counterfactuals were generated due to errors parsing the LLMs output.

Metric	Counterfactuals generated per explanation		
	3	5	10
<i>News Summarization</i>			
Generality	0.497	0.512	0.595
Simulatable counterfactuals	18	30	38
Total generated counterfactuals	30	50	80
<i>Medical Suggestion</i>			
Generality score	0.187	0.218	0.227
Simulatable counterfactuals	41	66	95
Total generated counterfactuals	90	150	300

Table 6: Generality and simulatability metrics as the number of generated counterfactuals per explanation increases.

## D LLM Prompts

We provide the prompts used in our evaluation pipeline.

## News Summarization: Chain-of-thought explanations

Your task is to summarize the document provided. First, describe on a high level the essential elements that the summary should include. Crucially, your explanation should only mention high-level, abstract, and generalized themes, and **MUST NOT** leak the topic, actions, subjects, or any details of the document. You will be **HEAVILY** penalized if the explanation is too specific. See below examples for explanations. Following this, compose the summary. Your response should be in the format "Explanation:" followed by "Summary:". Closely follow the format in the examples below.

Example:

Document: (CNN)Share, and your gift will be multiplied. That may sound like an esoteric adage, but when Zully Broussard selflessly decided to give one of her kidneys to a stranger, her generosity paired up with big data. It resulted in six patients receiving transplants. That ...

Explanation: Start by highlighting the subject’s key decision or action. Next, identify the factors that contributed to the event’s success to understand the underlying reasons. Then, outline the major outcomes and include aggregated statistics to provide a data-supported overview of the impact.

Summary: Zully Broussard decided to give a kidney to a stranger. A new computer program helped her donation spur transplants for six kidney patients.

Example:

Document: (CNN)On the 6th of April 1996, San Jose Clash and DC United strode out in front of 31,683 expectant fans at the Spartan Stadium in San Jose, California. The historic occasion was the first ever Major League Soccer match – a brave new dawn for the ...

Explanation: Start by identifying the milestone event and its specific date, as this is the focal point of the docu-

ment. Next, trace the developments and changes that led up to this milestone, noting any significant dates. Additionally, include any ongoing debates or challenges related to the topic, which will help contextualize the current situation and offer insights into possible future developments.

Summary: The 20th MLS season begins this weekend. League has changed dramatically since its inception in 1996. Some question whether rules regarding salary caps and transfers need to change.

Example: ...

Example: ...

Your turn:

Document: [DOCUMENT]

### **News Summarization: Post-hoc explanations**

Your task is to summarize the document provided. First, compose the summary. Following this, describe on a high level the essential elements that the summary should include. Crucially, your explanation should only mention high-level, abstract, and generalized themes, and **MUST NOT** leak the topic, actions, subjects, or any details of the document. You will be **HEAVILY** penalized if the explanation is too specific. See below examples for explanations. Following this, compose the summary. Your response should be in the format "Summary:" followed by "Explanation:". Closely follow the format in the examples below.

Example:

Document: (CNN)Share, and your gift will be multiplied. That may sound like an esoteric adage, but when Zully Broussard selflessly decided to give one of her kidneys to a stranger, her generosity paired up with big data. It resulted in six patients receiving transplants. That surprised and wowed her. "I . . .

Summary: Zully Broussard decided to give a kidney to a stranger. A new computer program helped her donation spur transplants for six kidney patients.

Explanation: Start by highlighting the subject's key decision or action. Next, identify the factors that contributed to the event's success to understand the underlying reasons. Then, outline the major outcomes and include aggregated statistics to provide a data-supported overview of the impact.

Example:

Document: (CNN)French striker Bafetimbi Gomis, who has a history of fainting, said he is now "feeling well" after collapsing during Swansea's 3-2 loss at Tottenham in the Premier League on Wednesday. The worrying incident occurred in the first half at White Hart Lane – after Tottenham scored in the . . .

Summary: Bafetimbi Gomis collapses within 10 minutes of kickoff at Tottenham. But he reportedly left the pitch conscious and wearing an oxygen mask. Gomis later said that he was "feeling well" The incident came three years after Fabrice Muamba collapsed at White Hart Lane.

Explanation: Focus on the key event described in the document and specify when it occurred. Highlight important details, particularly those that are surprising or contradictory. Include relevant quotes or statements from individuals involved to add depth and authenticity. Also, reference related or prior events to give a fuller context, helping to situate the incident within a broader historical framework.

Example: ...

Your turn:

Document: [DOCUMENT]

### **News Summarization: Counterfactual generation**

You will be asked to first read an AI's Decision Process for summarization. Then, you will be asked to craft 5 CNN-style news articles that you can confidently guess the AI's summary to be based on its provided decision process. The content in the crafted news articles should be diverse. Do not use new lines in the crafted

article. Start your crafted medical question with "Crafted Article 1:", "Crafted Article 2:" and so on. Closely follow the format in the examples below.

Example:

AI's Decision Process: Start by highlighting the subject's key decision or action. Next, identify the factors that contributed to the event's success to understand the underlying reasons. Then, outline the major outcomes and include aggregated statistics to provide a data-supported overview of the impact.

Crafted Article 1: (CNN)The world's battle against polio took a historic turn this week as the World Health Organization (WHO) announced the successful eradication of Type 3 poliovirus. The achievement marks a critical milestone in global health, leaving only one strain of wild polio still in circulation . . .

Crafted Article 2: (CNN)A revolutionary breakthrough in clean energy is taking shape in the deserts of Nevada, where scientists have achieved sustained nuclear fusion for the first time. In an announcement that could reshape the future of energy, researchers at the National Ignition Facility (NIF) confirmed . . .

...

Crafted Article 5: (CNN)A decades-old mystery in deep space has finally been unraveled, thanks to NASA's James Webb Space Telescope. Astronomers announced that they have identified the origins of fast radio bursts (FRBs)—intense, millisecond-long pulses of radio waves that have baffled scientists . . .

Example:

AI's Decision Process: Start by identifying the milestone event and its specific date, as this is the focal point of the document. Next, trace the developments and changes that led up to this milestone, noting any significant dates. Additionally, include any ongoing debates or challenges related to the topic, which will help contextualize the current situation

and offer insights into possible future developments.

Crafted Article 1: (CNN)On March 2, 2020, SpaceX launched the Crew Dragon capsule aboard a Falcon 9 rocket from NASA's Kennedy Space Center in Florida, marking a historic moment in the resurgence of American spaceflight. The mission, named Demo-2, carried NASA astronauts Douglas Hurley . . .

Crafted Article 2: (CNN)On June 29, 2007, Apple revolutionized the technology landscape with the release of the first iPhone, a sleek, touchscreen device that combined a phone, music player, and web browser in one. The launch, spearheaded by then-CEO Steve Jobs, marked the beginning of a seismic . . .

Crafted Article 5: (CNN)On December 10, 1948, the United Nations General Assembly adopted the Universal Declaration of Human Rights (UDHR), an unprecedented document that set forth fundamental freedoms and protections for all people, regardless of nationality, race, or gender. Drafted in . . .

Your turn:

AI's Decision Process: [EXPLANATION]

### News Summarization: Explanation decomposition

You will be asked to read an AI's Decision Process for summarization along with a document. Your task is as follows: From the AI's Decision Process, extract information about the elements the AI is using to summarize the document. Closely follow the format in the below examples.

Example: AI's decision process: Start by highlighting the subject's key decision or action. Next, identify the factors that contributed to the event's success to understand the underlying reasons. Then, outline the major outcomes and include aggregated statistics to provide a data-supported overview of the impact.

Answer: \* subject's key decision or action \* factors contributing to the event's

success \* major outcomes of the event \* aggregated statistics

Example: AI's decision process: Focus on the key event described in the document and specify when it occurred. Highlight important details, particularly those that are surprising or contradictory. Include relevant quotes or statements from individuals involved to add depth and authenticity. Also, reference related or prior events to give a fuller context, helping to situate the incident within a broader historical framework.

Answer: \* key event and its timing \* surprising or contradictory details \* relevant quotes or statements from individuals involved \* related or prior events

Example: ...

Example: ...

Your turn:

AI's decision process: [EXPLANATION]

Answer:

### **News Summarization: Simulatability annotation**

You will be given "Document", a document from an online news website. Then, you will be given "AI Summarization Key Details", a bulleted list containing high-level details an AI would deem important in the summarization process. Your task is as follows: For each of the points present in "AI Summarization Key Details", determine whether the point is found in the "Document" or not. Closely follow the format in the below examples.

Example: Document: (CNN)Share, and your gift will be multiplied. That may sound like an esoteric adage, but when Zully Broussard selflessly decided to give one of her kidneys to a stranger, her generosity paired up with big data. It resulted in six patients receiving transplants. That ...

AI Summarization Key Details: \* subject's key decision or action \* factors contributing to the event's success \* major outcomes of the event \* aggregated statistics

Answer: \* subject's key decision or action: found (Zully Broussard's decision to donate one of her kidneys) \* factors contributing to the event's success: found (e.g., the use of big data) \* major outcomes of the event: found (six patients received transplants) \* aggregated statistics: found (e.g., the ages of donors and recipients ranged from 26 to 70)

Example: Document: (CNN)Urban green spaces serve a multifaceted role in fostering environmental sustainability and enhancing the quality of life for city dwellers. Parks, nature reserves, and community gardens are integral parts of urban landscapes, providing a refuge from the city ...

AI Summarization Key Details: \* key event and its timing \* surprising or contradictory details \* relevant quotes or statements from individuals involved \* related or prior events

Answer: \* key event and its timing: not found \* surprising or contradictory details: not found \* relevant quotes or statements from individuals involved: not found \* related or prior events: not found

Example: ...

Example: ...

Your turn:

Document: [DOCUMENT]

AI Summarization Key Details: [EXTRACTED\_POINTS]

Answer:

### **News Summarization: Precision annotation**

You will be given "Document Summary", a summary of a document from an online news website. Then, you will be given "AI Summarization Key Details", a bulleted list containing high-level details an AI would deem important in the summarization process. Your task is as follows: For each of the points present in "AI Summarization Key Details", determine whether the point is found in the "Document Summary" or not. Closely follow the format in the below examples.



Example: Document Summary: Zully Broussard decided to give a kidney to a stranger. A new computer program helped her donation spur transplants for six kidney patients.

AI Summarization Key Details: \* subject's key decision or action \* factors contributing to the event's success \* major outcomes of the event \* aggregated statistics

Answer: \* subject's key decision or action: found in summary (Zully Broussard decided to give a kidney to a stranger) \* factors contributing to the event's success: found in summary (new computer program) \* major outcomes of the event: found in summary (transplants for six kidney patients) \* aggregated statistics: found in summary (six)

Example: Document Summary: The 20th MLS season begins. League has changed dramatically since its inception in 1996.

AI Summarization Key Details: \* milestone event and its date \* developments and changes leading up to the milestone \* ongoing debates or challenges

Answer: \* milestone event and its date: not found in summary (date not mentioned) \* developments and changes leading up to the milestone: found in summary (League has changed dramatically since its inception in 1996) \* ongoing debates or challenges: not found in summary (ignore)

Your turn:

Document Summary: [SUMMARY]

AI Summarization Key Details: [EXTRACTED\_POINTS]

Answer:

### **Medical Suggestion: Chain-of-thought explanations**

Your task is to give a medical suggestion based on a patient's query. First, provide an overall explanation of the patient's medical situation. This explanation should include multiple actionable suggested next steps. Following this,

return a final medical suggestion. The medical suggestion should be concise. If the medical suggestion involves meeting with a healthcare professional, it **MUST** include a specific description of what should be accomplished with the healthcare professional. Your response should be in the format "Explanation:" followed by "Medical Suggestion:". Closely follow the format in the examples below.

Example:

Patient Query: Ever since I caught a cold, I have been suffering from reflux of nasal discharge. Every day, mucus flows into my mouth or throat. I spit out the mucus many times a day. The symptoms are most obvious when I get up in the morning. The color of the spit is transparent or green, brown or a little bloody. The symptoms have lasted for at least three months. Recently, the number of sneezes has become much more frequent than before. Sometimes the nose will start to run when walking, but it is always a little bit transparent. But I don't have allergies. I would like to know what the problem may be and how to improve it.

Explanation: The patient is experiencing persistent nasal discharge, frequent sneezing, and mucus that varies in color (including transparent, green, brown, or slightly bloody). Since these symptoms have lasted for over three months, it may indicate chronic sinusitis or another underlying condition. There are many diagnostic methods that ENTs (ear, nose, and throat specialists) can perform to help investigate the root cause such as endoscopy and imaging examinations. It may also be necessary to test for the possibility of an infection with a bacterial culture. To manage these symptoms, a nasal washer with saline solution can be used to help clear out mucus and reduce inflammation.

Medical Suggestion: It is necessary to visit an ENT to rule out chronic sinusitis. Diagnostic methods include endoscopy, imaging examinations, etc. It is recommended that you go to the otolaryngol-

ogy department for a bacterial culture. It is recommended that you purchase a "nasal washer" (containing an isotonic saline solution).

Example:

Patient Query: Hello... I have some questions to ask you... For more than a year, I have often felt nausea in my throat. I recently had a gastroscopy and the doctor said there was inflammation of the esophagus, acid reflux, and small ulcers. It was normal after taking the prescribed medicine for 2 months, but then it relapsed. Recently, I have felt nausea in my throat and bloated stomach. If pressed with my fingers, the nausea will become more obvious. When it relapses, my hands and feet will feel cold, my stomach will feel uncomfortable for a while, and the symptoms will increase. Another point is that I feel a little nauseous when I am full, and I can also feel nauseous when I am hungry. The most important thing is that I feel more nauseated, and my stomach feels uncomfortable (a bit like having diarrhea). Please give me some advice. Thank you!

Explanation: The patient is experiencing nausea in the throat, a bloated stomach, and discomfort in the stomach that has lasted more than a year. The patient has previously done a gastroscopy, but after taking the prescribed medicine for 2 months, the symptoms relapsed. The chronic nature of these symptoms are indicative of gastroesophageal reflux disease (GERD) and potentially other gastrointestinal issues. A gastroenterologist can perform tests to rule out potential gastrointestinal diseases and provide a specific treatment plan accordingly. Eating small, frequent meals can help reduce pressure on the abdomen and minimize acid reflux, which often leads to the burning sensation, chest pain, and sore throat that the patient describes. Avoiding greasy food and sweets is also crucial, as these can exacerbate acid reflux symptoms. Additionally, increasing fiber intake and staying hydrated can help improve overall digestive health.

Medical Suggestion: It is recommended that you eat small meals more often, less greasy food, sweets, and chocolate to reduce abdominal pressure and acid reflux that causes burning sensation, chest pain, and sore throat. Also, try eating more fiber and drinking more water. It is best to go to the gastroenterology clinic for further treatment of gastrointestinal problems, they can administer tests to rule out other gastrointestinal diseases and provide treatment accordingly.

Example: ...

Your Turn:

Patient Query: [QUESTION]

### **Medical Suggestion: Post-hoc explanations**

Your task is to give a medical suggestion based on a patient's query. First, return a medical suggestion. The medical suggestion should be concise and contain multiple actionable suggested next steps. If the medical suggestion involves meeting with a healthcare professional, it **MUST** include a specific description of what should be accomplished with the healthcare professional. Following this, provide an overall explanation of the suggestion, explaining the patient's medical situation and how it relates to the suggested steps. Your response should be in the format "Medical Suggestion:" followed by "Explanation:". Closely follow the format in the examples below.

Example:

Patient Query: Ever since I caught a cold, I have been suffering from reflux of nasal discharge. Every day, mucus flows into my mouth or throat. I spit out the mucus many times a day. The symptoms are most obvious when I get up in the morning. The color of the spit is transparent or green, brown or a little bloody. The symptoms have lasted for at least three months. Recently, the number of sneezes has become much more frequent than before. Sometimes the nose will start to run when walking, but it is always a little bit transparent. But I don't have allergies. I would like to know what the problem may be and how to improve it.

**Medical Suggestion:** It is necessary to visit an ENT to rule out chronic sinusitis. Diagnostic methods include endoscopy, imaging examinations, etc. It is recommended that you go to the otolaryngology department for a bacterial culture. It is recommended that you purchase a "nasal washer" (containing an isotonic saline solution).

**Explanation:** The patient is experiencing persistent nasal discharge, frequent sneezing, and mucus that varies in color (including transparent, green, brown, or slightly bloody). Since these symptoms have lasted for over three months, it may indicate chronic sinusitis or another underlying condition. There are many diagnostic methods that ENTs (ear, nose, and throat specialists) can perform to help investigate the root cause such as endoscopy and imaging examinations. It may also be necessary to test for the possibility of an infection with a bacterial culture. To manage these symptoms, a nasal washer with saline solution can be used to help clear out mucus and reduce inflammation.

**Example:**

**Patient Query:** Hello... I have some questions to ask you... For more than a year, I have often felt nausea in my throat. I recently had a gastroscopy and the doctor said there was inflammation of the esophagus, acid reflux, and small ulcers. It was normal after taking the prescribed medicine for 2 months, but then it relapsed. Recently, I have felt nausea in my throat and bloated stomach. If pressed with my fingers, the nausea will become more obvious. When it relapses, my hands and feet will feel cold, my stomach will feel uncomfortable for a while, and the symptoms will increase. Another point is that I feel a little nauseous when I am full, and I can also feel nauseous when I am hungry. The most important thing is that I feel more nauseated, and my stomach feels uncomfortable (a bit like having diarrhea). Please give me some advice. Thank you!

**Medical Suggestion:** It is recommended that you eat small meals more often, less greasy food, sweets, and chocolate to reduce abdominal pressure and acid reflux that causes burning sensation, chest pain, and sore throat. Also, try eating more fiber and drinking more water. It is best to go to the gastroenterology clinic for further treatment of gastrointestinal problems, they can administer tests to rule out other gastrointestinal diseases and provide treatment accordingly.

**Explanation:** The patient is experiencing nausea in the throat, a bloated stomach, and discomfort in the stomach that has lasted more than a year. The patient has previously done a gastroscopy, but after taking the prescribed medicine for 2 months, the symptoms relapsed. The chronic nature of these symptoms are indicative of gastroesophageal reflux disease (GERD) and potentially other gastrointestinal issues. A gastroenterologist can perform tests to rule out potential gastrointestinal diseases and provide a specific treatment plan accordingly. Eating small, frequent meals can help reduce pressure on the abdomen and minimize acid reflux, which often leads to the burning sensation, chest pain, and sore throat that the patient describes. Avoiding greasy food and sweets is also crucial, as these can exacerbate acid reflux symptoms. Additionally, increasing fiber intake and staying hydrated can help improve overall digestive health.

**Example:** ...

**Your Turn:**

**Patient Query:** [QUESTION]

### **Medical Suggestion: Counterfactual generation**

You will be asked to first read an AI's Explanation for giving a particular answer in response to a medical question. This explanation reflects the knowledge that the AI is using to answer the question. Then, you will be asked to craft 5 informal, online forum style medical questions that contain the symptoms reflected in the AI's Explanation. Do not include any of the suggested solutions in

the AI's explanation in the medical question. Start your crafted medical question with "Crafted Medical Question 1:", "Crafted Medical Question 2:" and so on.

Example: AI's Explanation: The patient is experiencing persistent nasal discharge, frequent sneezing, and mucus that varies in color (including transparent, green, brown, or slightly bloody). Since these symptoms have lasted for over three months, it may indicate chronic sinusitis or another underlying condition. There are many diagnostic methods that ENTs (ear, nose, and throat specialists) can perform to help investigate the root cause such as endoscopy and imaging examinations. It may also be necessary to test for the possibility of an infection with a bacterial culture. To manage these symptoms, a nasal washer with saline solution can be used to help clear out mucus and reduce inflammation.

Crafted Medical Question 1: After I got the flu a while back, I've been dealing with a crazy amount of mucus build up in my throat. It's gotten really bad to the point where I have to pack tissues in my bag before going anywhere because I can't even walk properly without my nose constantly dripping. I'm not sure what's causing it. What should I do?

Crafted Medical Question 2: For the past few months, I've been sneezing all the time, and my nose is constantly stuffed or dripping. The mucus changes color a lot—sometimes it's clear, other times it's green or even has a little blood in it. Could this be something serious, or is it just allergies acting up?

...

Crafted Medical Question 5: I've had this annoying runny nose for what feels like forever, and it's not just clear mucus—sometimes it's green or brownish. Sneezing fits are also nonstop, and it's starting to mess with my sleep. Please help.

Example: AI's Explanation: The patient is experiencing nausea in the throat,

a bloated stomach, and discomfort in the stomach that has lasted more than a year. The patient has previously done a gastroscopy, but after taking the prescribed medicine for 2 months, the symptoms relapsed. The chronic nature of these symptoms are indicative of gastroesophageal reflux disease (GERD) and potentially other gastrointestinal issues. A gastroenterologist can perform tests to rule out potential gastrointestinal diseases and provide a specific treatment plan accordingly. Eating small, frequent meals can help reduce pressure on the abdomen and minimize acid reflux, which often leads to the burning sensation, chest pain, and sore throat that the patient describes. Avoiding greasy food and sweets is also crucial, as these can exacerbate acid reflux symptoms. Additionally, increasing fiber intake and staying hydrated can help improve overall digestive health.

Crafted Medical Question 1: hey everyone, I've been having this burning sensation in my chest and throat, especially after eating, and sometimes it even wakes me up at night. My stomach has also been generally uncomfortable lately and I bloat quite a bit. I've had this issue for the past year or so but only recently has it gotten pretty bad. It's starting to negatively affect my lifestyle and my work. Does anyone else have this? what can I do to fix it?

Crafted Medical Question 2: Hi all, I've been feeling really bloated and uncomfortable in my stomach for over a year now. Lately, I've also noticed a weird nausea-like feeling in my throat after meals, and sometimes I get this burning sensation that feels like it's in my chest. I've tried a couple of things, but so far no luck. Has anyone experienced something similar? Any advice on what this could be or how to manage it?

...

Crafted Medical Question 5: Hey, I'm wondering if anyone here has dealt with long-term stomach issues like bloating



and discomfort. For me, it's been going on for about a year. Recently, it's been accompanied by a sore throat and this gross burning feeling, especially after meals. I'm not sure if it's something I'm eating or something else entirely. Should I see a specialist? What's worked for you?

Example: ...

Your Turn:

AI's Explanation: [EXPLANATION]

### **Medical Suggestion: Explanation extraction**

You will be asked to read an AI's Explanation on a medical topic. Then, you will be given a Follow-up Medical Question. Your task is as follows: First, from the AI's Explanation, extract key information about the patient involving symptoms, demographic information, or their relevant medical history. Then, extract medical suggestion points from the AI's Explanation, focusing on the suggested treatments. You will be HEAVILY penalized if multiple medical suggestion points cover the same information. Closely follow the format in the below examples.

Example: AI's Explanation: The patient is experiencing persistent nasal discharge and frequent sneezing. Since these symptoms have lasted for months it may indicate chronic sinusitis or another underlying condition. The patient reported that they recently caught the flu and have no allergies. There are many diagnostic methods that ENTs (ear, nose, and throat specialists) can perform to help investigate the root cause such as endoscopy and imaging examinations. It may also be necessary to test for the possibility of an infection via. a bacterial culture. To manage these symptoms, a nasal washer with saline solution can be used to help clear out mucus and reduce inflammation. Answer: The key information described about the patient in the AI's Explanation include, \* persistent nasal discharge and frequent sneezing \* symptoms have lasted for months \* recently caught the flu \* no allergies The medical suggestions from the AI's Explanation include,

\* diagnostic methods such as endoscopy and imaging examinations \* test for a possible infection via. a bacterial culture \* use a nasal washer with saline solution to clear mucus and reduce inflammation

Example: AI's Explanation: The patient is experiencing a delayed menstrual period which is 9 days late. She mentions having had sexual intercourse during what she considered a "safe period." Additionally, she has been experiencing increased breast tenderness, a symptom she usually associates with the onset of her period. The patient also notes significant changes in her sleep patterns, going to bed very late. Changes in sleep, stress, and daily routine can affect menstrual cycles, it may be helpful to monitor these factors moving forward. It is also advised to conduct a pregnancy test in order to rule out pregnancy, especially given the recent sexual activity, even if it was during the perceived safe period. Answer: The key information described about the patient in the AI's Explanation include, \* delayed menstrual period \* previously had sexual intercourse \* increased breast tenderness \* changes in sleep patterns The medical suggestions from the AI's Explanation include, \* monitor changes in sleep, stress, and daily routine \* conduct a pregnancy test

Example: ...

Your turn:

AI's Explanation: [EXPLANATION]

Answer:

### **Medical Suggestion: Simulatability annotation**

You will be given "Medical Question", a question submitted by a user to an online medical forum. Then, you will be given "AI Extracted Details", a bulleted list containing details an AI extracted from the question. Your task is as follows: For each of the points present in "AI Extracted Details", determine whether the point is found in the user's "Medical Question" or not. Closely follow the format in the below examples.

Example: Medical Question: Ever since I got the flu a couple of months back, I've been dealing with a crazy amount of mucus build up in my throat. It's gotten really bad to the point where I have to pack tissues in my bag before going anywhere because I can't even walk properly without my nose constantly dripping. I'm not sure what's causing it. What should I do?  
AI Extracted Details: \* persistent nasal discharge and frequent sneezing \* symptoms have lasted for months \* recently caught the flu \* no allergies  
Answer: \* persistent nasal discharge and frequent sneezing: found in Medical Question ("nose constantly dripping") \* symptoms have lasted for months: found in Medical Question ("Ever since I got the flu a couple of months back,") \* recently caught the flu: found in Medical Question ("Ever since I got the flu a couple of months back,") \* no allergies: not found in Medical Question

Example: Medical Question: hello doctor, I'm worried because my period hasn't arrived yet even though it should have about a week ago. My boyfriend and I had sex a few weeks ago but I haven't seen him since. I'm also having difficulty falling asleep and I wake up earlier than usual. Could these symptoms be related? Should I take a pregnancy test or is there another possible explanation? Any advice would be appreciated!  
AI Extracted Details: \* delayed menstrual period \* previously had sexual intercourse \* increased breast tenderness \* changes in sleep patterns  
Answer: \* delayed menstrual period: found in Medical Question ("my period hasn't arrived yet even though it should have about a week ago") \* previously had sexual intercourse: found in Medical Question ("My boyfriend and I had sex a few weeks ago") \* increased breast tenderness: not found in Medical Question \* changes in sleep patterns: found in Medical Question ("I'm also having difficulty falling asleep and I wake up earlier than usual")

Example: ...

Your turn:

Medical Question: [QUESTION]

AI Extracted Details:  
[EXTRACTED\_POINTS]

Answer:

### Medical Suggestion: Precision annotation

You will be given "Medical Suggestion", an answer to a user question on an online medical forum. Then, you will be given "AI Extracted Details", a bulleted list containing details an AI extracted from the suggestion. Your task is as follows: For each of the points present in "AI Extracted Details", determine whether the point is found in "Medical Suggestion" or not. Closely follow the format in the below examples.

Example: Medical Suggestion: It is necessary to visit an ENT to rule out chronic sinusitis. Diagnostic methods include endoscopy, imaging examinations, etc. It is recommended that you purchase a "nasal washer" (containing an isotonic saline solution). If further examination is required, visit an otolaryngology department for evaluation.  
AI Extracted Details: \* diagnostic methods (endoscopy, imaging examinations) \* get a bacterial culture \* seek a medical evaluation \* purchase a nasal washer with isotonic saline solution  
Answer: \* diagnostic methods (endoscopy, imaging examinations): found in Medical Suggestion ("Diagnostic methods include endoscopy, imaging examinations, etc.") \* get a bacterial culture: not found in Medical Suggestion \* seek a medical evaluation: found in Medical Suggestion ("visit an otolaryngology department for evaluation") \* purchase a nasal washer with isotonic saline solution: found in Medical Suggestion ("purchase a "nasal washer" (containing an isotonic saline solution)")

Example: Medical Suggestion: It is recommended that you eat small meals more often, less greasy food, sweets, and chocolate to reduce abdominal pressure and acid reflux that causes burning sensation, chest pain, and sore throat. Eat

more fiber and drink more water. It is best to go to the gastroenterology clinic for further treatment of gastrointestinal problems, they can administer tests to rule out other gastrointestinal diseases. AI Extracted Details: \* minimize the consumption of carbohydrates \* eat small, frequent meals \* avoid greasy food and sweets \* increase fiber intake and stay hydrated Answer: \* minimize the consumption of carbohydrates: not found in Medical Suggestion \* eat small, frequent meals: found in Medical Suggestion ("eat small meals more often") \* avoid greasy food and sweets: found in Medical Suggestion ("less greasy food, sweets, and chocolate") \* increase fiber intake and stay hydrated: found in Medical Suggestion ("Eat more fiber and drink more water.")

Example: ...

Your turn:

Medical Suggestion: [SUGGESTION]

AI            Extracted            Details:  
[EXTRACTED\_POINTS]

Answer:

## E Human Evaluation

Details about the human evaluation are provided followed by screenshots of the annotation sheets.

### News Summarization Instructions

- Parsed Explanation Annotations: Check if the explanation (Column B) is parsed correctly in Column C. If yes, put "ok". If not, write which information in the explanation is missing / which info in column C are hallucinating.
- Document Annotations: Check whether each item in the parsed explanation appears in the document or not. Put either "no" or "yes".
- Summary Annotations: Check whether each item in the parsed explanation appears in the summary or not. Put either "no" or "yes"

### Medical Suggestions Instructions

Tasks 1, 3:

- for each of the "Extracted Points", label "Y" or "N" if the point was parsed correctly from the AI explanation

- if there are missing points or other comments, feel free to note them in the EXTRA QUESTION annotation cell
- task 1 focuses on extracting patient details (e.g., symptoms, medical history)
- task 3 focuses on extracting suggested next steps (e.g., recommended treatments)
- since breakdowns are the same across rows, feel free to complete just one per example ID

Tasks 2, 4:

- for each of the "Extracted Points", label "Y" or "N" if the point is reflected in the counterfactual (task 2) or AI suggestion (task 4)
- the matching is flexible, i.e., point does not necessarily need to appear in exactly the same wording to be marked "Y"
- for example, in example 3 the point mentions ""hemoglobin level of 10.5, slightly below normal"" these specific scores do not need to be mentioned, rather ""below normal levels"" is sufficient.

A	B	C	D	E	F	G	H
Example ID	Explanation	Parsed Explanation	Parsed Explanation Annotation	Document	Document Annotations (yes/no)	Summary	Summary Annotations (yes/no)
0	The summary should begin by noting the key event of accession and the entity involved. It should then mention the implications of this event, particularly in terms of jurisdiction over specific territories. The summary should also include the reactions of significant stakeholders and the broader international community, highlighting both support and opposition. Finally, it should note the ceremonial aspects of the event and the statements made by relevant officials to provide context and indicate the significance of the event in international relations and law.	Key event of accession and entity involved	ok	(CNN)In a historic move on September 1st, the Pacific Island nation of Kiribati officially gained membership in the United Nations, marking a significant milestone in its diplomatic history. After years of striving for international recognition and support, Kiribati's accession was celebrated in a grand ceremony attended by various global leaders and diplomats in New York. Kiribati, which has long been vocal about its struggles with rising sea levels and climate change, now holds the privilege to vote on UN resolutions and partake in international debates fully. "The international community's reception of Kiribati into the United Nations marks a pivotal step in our journey towards global recognition and support," said President Taneti Maamau. "This membership is not just a ceremonial title but a gateway to advocating more intensely for climate action and sustainability on a global stage." Kiribati's primary concern is its territorial integrity, as rising sea levels threaten its very existence. The reception of Kiribati into the UN has garnered a mixed reaction. Many small island nations and climate advocates have shown strong support, viewing Kiribati's membership as a platform to amplify the urgent needs related to climate impacts. However, there have been reservations from some countries, primarily concerning increased voting blocs within the UN that could shift	yes	Kiribati officially joined the United Nations on September 1st, marking a significant step in its diplomatic history. The accession ceremony in New York was attended by global leaders and highlighted Kiribati's ongoing struggles with climate change and rising sea levels. With its new UN membership, Kiribati aims to advocate more effectively for global climate action and sustainability.	yes
2		Implications of the event			yes		yes
3		Reactions of significant stakeholders and the broader international community			yes		no
4		Ceremonial aspects of the event			yes		no
5		Statements by relevant officials			yes		no
6							
0	The summary should begin by noting the key event of accession and the entity involved. It should then mention the implications of this event, particularly in terms of jurisdiction over specific territories. The summary should also include the reactions of significant stakeholders and the broader international community, highlighting both support and opposition. Finally, it should note the ceremonial aspects of the event and the statements made by relevant officials to provide context and indicate the significance of the event in international relations and law.	Key event of accession and the entity involved	ok	(CNN)In a historic move, Catalonia declared independence from Spain on October 1st, marking a significant shift in the landscape of European politics. The Catalan parliament passed a unilateral declaration of independence following a contentious and disputed referendum. The Spanish government, headquartered in Madrid, immediately declared the referendum and subsequent declaration of independence illegal, citing violations of the Spanish Constitution. As a response, the Spanish Prime Minister vehemently opposed the secession and introduced direct rule over the region. International reactions were mixed. The European Union expressed concern over the unilateral move and emphasized the importance of dialogue within the framework of the Spanish Constitution. Meanwhile, thousands of Catalans celebrated in the streets of Barcelona, displaying flags and singing the Catalan anthem as they heralded what they saw as the birth of a new nation. Prominent Catalan leaders delivered impassioned speeches about a long journey towards self-determination and the fulfillment of the (CNN) - In a significant geopolitical development, Scotland formally acceded to full independence from the United Kingdom on January 1, 2025, after a prolonged series of negotiations following the referendum in 2023. This accession marks the first alteration in UK territorial boundaries in over a century and positions Scotland as Europe's newest sovereign nation. The formal ceremony, held at Edinburgh Castle, was attended by leaders from across the globe, symbolizing international acknowledgment of Scotland's new status. United Nations	yes	Catalonia declared independence from Spain on October 1st, following a controversial referendum. The Spanish government declared the move illegal and imposed direct rule over the region. Reactions were mixed globally and domestically, with significant opposition within Catalonia and Spain. The declaration raised questions about the practical implications for Catalonia, particularly regarding EU membership and international recognition.	yes
7		Implications of the event, particularly in terms of jurisdiction over specific territories			yes		yes
8		Reactions of significant stakeholders and the broader international community			yes		yes
9		Ceremonial aspects of the event and statements by relevant officials			yes		no
10							
0	The summary should begin by noting the key event of accession and the entity involved. It should then mention the implications of this event, particularly in terms of jurisdiction over specific territories. The summary should also include the reactions of significant stakeholders and the broader international community, highlighting both support and opposition. Finally, it should note the	Key event of accession and the entity involved	ok	(CNN) - In a significant geopolitical development, Scotland formally acceded to full independence from the United Kingdom on January 1, 2025, after a prolonged series of negotiations following the referendum in 2023. This accession marks the first alteration in UK territorial boundaries in over a century and positions Scotland as Europe's newest sovereign nation. The formal ceremony, held at Edinburgh Castle, was attended by leaders from across the globe, symbolizing international acknowledgment of Scotland's new status. United Nations	yes	Scotland officially became independent from the United Kingdom on January 1, 2025, following a 2023 referendum. The historic event, marked by a ceremony at Edinburgh Castle, was recognized globally with international leaders present. The United Nations Secretary-General congratulated Scotland, highlighting the principle of self-determination. Despite some disputes over North Sea oil reserves, the	yes
11		Implications of this event in terms of jurisdiction over specific territories			yes		yes
12							

Figure 3: Screenshot of the news summarization annotation interface.

	A	B	C	D	E	F	G	H	I	J
1	Example ID		Task 1: Given an AI Explanation extract key points about the patient. This includes any information that would be relevant to providing a medical suggestion. Examples: symptoms, patient's history, and demographic information (e.g. age).			Task 2: Given a set of key points and a counterfactual medical query, determine whether the points are found in the counterfactual or not.				
2		AI Explanation	Extracted Points	Annotation (Y/N)	Extracted Points	Counterfactual Medical Query	Annotation (Y/N)			
3	0	The patient is concerned about a blood test result showing a white blood cell (WBC) count that is slightly elevated, being 1 unit above the normal reference range of 4-10 (presumably $\times 10^9/L$ ). A mild increase in WBC count can be due to a variety of reasons including stress, infection, inflammation, or even normal daily fluctuations. It is important to consider any symptoms the patient might be experiencing, such as fever, pain, fatigue, or signs of infection, which could provide more context to the elevated WBC count. If the patient is asymptomatic, it might not indicate a serious condition. However, it's crucial to monitor the	* concerned about a blood test result	Y	* concerned about a blood test result	Hi everyone, I just found out that my WBC count is a bit elevated, just 1 unit above the normal range. I don't have any major symptoms, maybe just a little fatigue and occasional headaches. Should I be concerned about this, or could it just be due to stress or something minor?	Y			
4	* white blood cell count slightly elevated, 1 unit above normal		Y	* white blood cell count slightly elevated, 1 unit above normal	Y					
5	EXTRA QUESTION: Is anything missing from the above?		N							
6										
7	0	The patient is concerned about a blood test result showing a white blood cell (WBC) count that is slightly elevated, being 1 unit above the normal reference range of 4-10 (presumably $\times 10^9/L$ ). A mild increase in WBC count can be due to a variety of reasons including stress, infection, inflammation, or even normal daily fluctuations. It is important to consider any symptoms the patient might be experiencing, such as fever, pain, fatigue, or signs of infection, which could provide more context to the elevated WBC count. If the patient is asymptomatic, it might not indicate a serious condition. However, it's crucial to monitor the	* concerned about a blood test result		* concerned about a blood test result	Hello all, I need some advice. My recent blood test showed that my white blood cells are slightly elevated. I'm not experiencing any serious symptoms, just some mild pain and fatigue here and there. Is this something that happens normally, or should I get it checked out again?	Y			
8	* white blood cell count slightly elevated, 1 unit above normal			* white blood cell count slightly elevated, 1 unit above normal	Y					
9	EXTRA QUESTION: Is anything missing from the above?				Y					
10										
11										
	A	J	K	L	M	N	O	P	Q	R
1	Example ID		Task 3: Given an AI Explanation extract key points about the medical suggestion. This refers to recommendations for treatment i.e. actionable, suggested next steps.			Task 4: Given a set of key points and a AI generated medical suggestion, determine whether the points are found in the AI's suggestion or not.				
2		AI Explanation	Extracted Points	Annotation (Y/N)	Extracted Points	AI Suggestion	Annotation (Y/N)			
3	0	The patient is concerned about a blood test result showing a white blood cell (WBC) count that is slightly elevated, being 1 unit above the normal reference range of 4-10 (presumably $\times 10^9/L$ ). A mild increase in WBC count can be due to a variety of reasons including stress, infection, inflammation, or even normal daily fluctuations. It is important to consider any symptoms the patient might be experiencing, such as fever, pain, fatigue, or signs of infection, which could provide more context to the elevated WBC count. If the patient is asymptomatic, it might not indicate a serious condition. However, it's crucial to monitor the	* monitor the WBC count	Y	* monitor the WBC count	At this stage, it is advisable to monitor your WBC count and symptoms. Ensure you are maintaining a healthy lifestyle with proper hydration, nutrition, and rest. If you notice any new symptoms or if your current symptoms worsen, or if your WBC count increases further, it would be prudent to consult with your primary care physician. During this consultation, discuss your recent lab results and symptoms to determine if further testing or a referral to a specialist is necessary.	Y			
4	* possibly retest to see if the count changes over time		Y	* possibly retest to see if the count changes over time	Y					
5	EXTRA QUESTION: Is anything missing from the above?		consider related symptoms such as fever pain							
6										
7	0	The patient is concerned about a blood test result showing a white blood cell (WBC) count that is slightly elevated, being 1 unit above the normal reference range of 4-10 (presumably $\times 10^9/L$ ). A mild increase in WBC count can be due to a variety of reasons including stress, infection, inflammation, or even normal daily fluctuations. It is important to consider any symptoms the patient might be experiencing, such as fever, pain, fatigue, or signs of infection, which could provide more context to the elevated WBC count. If the patient is asymptomatic, it might not indicate a serious condition. However, it's crucial to monitor the	* monitor the WBC count		* monitor the WBC count	It is advisable to monitor your symptoms and consider a follow-up blood test to check if the white blood cell count returns to normal. Additionally, evaluate and possibly adjust your diet, stress management, and sleep habits to improve overall health. If the white blood cell count remains elevated or if symptoms worsen, it is recommended to consult with a healthcare professional to rule out any underlying conditions. During this consultation, discuss your initial test results and symptoms in detail to help determine the need for further diagnostic testing or a referral to a specialist if necessary.	Y			
8	* possibly retest to see if the count changes over time			* possibly retest to see if the count changes over time	Y					
9	EXTRA QUESTION: Is anything missing from the above?									
10										
11										

Figure 4: Screenshots of the medical suggestion annotation interface.