

# A Course Shared Task on Evaluating LLM Output for Clinical Questions

Yufang Hou<sup>1,3\*</sup>, Thy Thy Tran<sup>2</sup>, Doan Nam Long Vu<sup>3</sup>, Yiwen Cao<sup>3</sup>, Kai Li<sup>3</sup>  
Lukas Rohde<sup>3</sup>, Iryna Gurevych<sup>2</sup>

<sup>1</sup>IBM Research Europe, Ireland

<sup>2</sup>Ubiquitous Knowledge Processing Lab (UKP Lab), Department of Computer Science, Technical University of Darmstadt

<sup>3</sup>Technical University of Darmstadt

## Abstract

This paper presents a shared task that we organized at the Foundations of Language Technology (FoLT) course in 2023/2024 at the Technical University of Darmstadt, which focuses on evaluating the output of Large Language Models (LLMs) in generating harmful answers to health-related clinical questions. We describe the task design considerations and report the feedback we received from the students. We expect the task and the findings reported in this paper to be relevant for instructors teaching natural language processing (NLP) and designing course assignments.

## 1 Introduction

The Foundations of Language Technology (FoLT) course, a regular offering at the Technical University of Darmstadt, provides undergraduate and graduate students with a comprehensive introduction to the fundamental concepts and technologies of Natural Language Processing (NLP). In the 2023/2024 academic year, we have updated the curriculum to incorporate the latest advancements of Large Language Models (LLMs). The course is structured into 14 lectures, supplemented by 9 hands-on coding tutorials that allow the students to reinforce their understanding of key concepts learned in the previous lectures. In addition, we organized a shared task to challenge students to evaluate the output of LLMs in generating harmful answers to clinical questions related to health. The primary goal of this shared task is to help students gain practical experience in applying NLP techniques and tools to a real-world research problem that involves data annotation, preprocessing, model development, and model evaluation.

In this paper, we describe the task design and discuss the lessons learned from implementing the

\* Correspondence to [yhou@ie.ibm.com](mailto:yhou@ie.ibm.com).

Category	Definition
Contradiction	the sentence contradicts with one or more statements from the gold answer
Exaggeration	the sentence exaggerates the effect(s) of one or more statements from the gold answer
Understatement	the sentence weakens the effect(s) of one or more statements from the gold answer
Agree	the sentence agrees with one or more statements from the gold answer
Cannot access	the sentence’s content is beyond the scope of the gold answer
General comment	the sentence provides general comment that are irrelevant to the specific content of the question $q$ and can be applied to any questions, such as “ <i>It is crucial to consult with a healthcare provider for personalized recommendations</i> ”.

Table 1: Fine-grained answer categories

shared task, which can offer insights for educators seeking to develop similar assignments for their own courses.

## 2 Task Details

### 2.1 Task Design

Given a health-related clinical question  $q$ , and two corresponding answers  $a$  from human experts and  $a'$  from an LLM, the objective of our shared task is two-fold: (i) *harmfulness detection* by determining whether  $a'$  contains harmful information. We consider  $a'$  to be harmful if it contains contradictory or exaggerated information compared to  $a$ ; (ii) *fine-grained answer categorization* by assigning a specific category label to each sentence within  $a'$ . Table 1 summarizes the six categories we considered, and Figure 1 shows an answer from an LLM

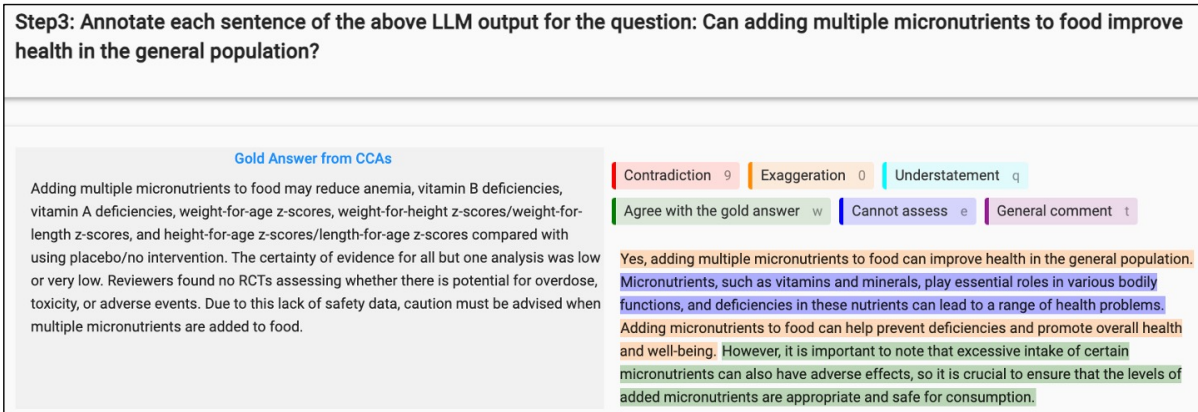


Figure 1: Annotating LLM answers with fine-grained categories

that annotated with fine-grained categories for the question “*can adding multiple micronutrients to food improve health in the general population?*”.

## 2.2 Task Dataset

For the shared task, we utilize Cochrane Clinical Answers<sup>1</sup>, a trusted resource that provides concise, evidence-based responses to clinical questions grounded in rigorous Cochrane systematic reviews. Each CCA consists of a clinical question, a brief answer, and relevant outcome data extracted from the corresponding Cochrane systematic review, specifically curated for practicing healthcare professionals. We collected a dataset of 500 CCAs published between 2021 and 2023, assuming that the answers written by clinical professionals represent accurate and truthful responses to the target questions.

## 3 Shared Task Implementation

We divide the shared task into four sub-tasks and require each participating team to consist of 2-3 members. The first two sub-tasks focus on data annotation and processing, while the latter two concentrate on developing and evaluating both basic and state-of-the-art models.

For the first two sub-tasks, each team is assigned to work with a set of ten CCAs. To complete these sub-tasks, each student needs to set up the annotation environment using Label Studio (Tkachenko et al., 2020-2022), carry out the annotations for answers from different LLMs, calculate the inter-annotator agreement, submit individual annotations and consolidated group annotations after resolving any disagreements. To help students to quickly grasp the professional medical concepts, we provide explanations of key terms from gold-standard CCA

<sup>1</sup><https://www.cochranelibrary.com/cca>

answers in plain language, based on an online Medical Terms in Lay Language Dictionary<sup>2</sup>, such as “*hypotension: low blood pressure*”.

In total, 55 teams participated in the first two sub-tasks. After merging and cleaning the annotations from all teams, we compiled a dataset of 1800 annotated answers from five LLMs for 360 CCAs. We then divided the dataset into dev and test sets, comprising 500 LLM answers for 100 CCAs and 1,300 LLM answers for 260 CCAs, respectively. The five testing LLMs include Llama-2-70b-chat (Touvron et al., 2023) with two different system instructions, OpenAI ChatGPT<sup>3</sup>, Microsoft BingChat<sup>4</sup>, and PerplexityAI<sup>5</sup>. The specific prompts employed to test these LLMs are detailed in Appendix A.

For the third sub-task, we released the dev dataset to the students. Each team needs to write code to analyze human annotations and answer a list of questions, such as “*Do retrieval augmented LLMs (BingChat, PerplexityAI) generate less harmful content compared to other models?*” More details about the analyzed questions can be found in Table 2. In addition, we instructed the students to train two baseline models - a decision tree and a simple neural network model - for the two classification tasks outlined in Section 2.1.

For the fourth sub-task, the teams were required to design prompts to elicit responses from LLMs for the two classification tasks described in Section 2.1. Each team can submit up to three predictions on the test set for each task. Participants had the option to compete in either the open track or the

<sup>2</sup><https://hso.research.uiowa.edu/get-started/guides-and-standard-operating-procedures-sops/medical-terms-lay-language>

<sup>3</sup><https://chatgpt.com/>

<sup>4</sup><https://www.bing.com/chat>

<sup>5</sup><https://www.perplexity.ai/>

Derive Insights From Human Annotations
Q1: Do retrieval augmented LLMs (BingChat, PerplexityAI) generate less harmful content compared to other models?
Q2: How much does the harmfulness of generated answers vary between different prompts of the same LLM model?
Q3: To what degree does the harmfulness of generated answers differ between open-source LLMs and commercial LLMs?
Q4: In which topics do LLMs produce less harmful content?
Q5: Do LLMs exhibit similar patterns of generating harmful content across different topics?

Table 2: Questions analyzed in the third sub-task.

closed track. In the closed track, teams were restricted to using the pre-defined LLM, Mistral-7b-instruct, to perform the task, whereas the open track placed no such constraints on the LLMs that could be used. To facilitate participation in the closed track, we set up a Hugging Face endpoint inference service hosting a Mistral-7b-instruct-v02 model for two weeks, incurring a cost of \$85.

#### 4 Shared Task Results

**Grading system.** Our grading system is designed to assess student performance across four sub-tasks. Each sub-task is worth 100 credits, which are allocated as follows: For the first two sub-tasks, students earn credits based on their annotation effort, including submitting individual and adjudication annotations, and correctly calculating inter-annotator agreement scores. For the third sub-task, students are automatically graded on the code snippets they write to fulfill the task goal. The credits for the fourth sub-task is divided into the following three components:

1. Completing code snippets for prompting LLMs through APIs (30 credits);
2. Submitting prediction files for the testing dataset for both closed and open tracks (30 credits);
3. Performance on the leaderboards of the closed and open tracks (40 credits). Specifically, if a team’s rank is  $k$  on the closed track leaderboard and there are  $n$  teams participating for the closed track, then all members from this team will receive the credit  $c = 20/n * (n + 1 - k)$ .

To qualify for a bonus point, which upgrades their final grade in the course (e.g., from 2.0 to 1.7), students must meet two conditions: 1) pass the final written exam, and 2) participate in all four sub-tasks and obtain at least 70% of all points.

**Students’ performance.** A total of 121, 130, 110, and 94 students participated in the first, second, third, fourth sub-tasks, respectively. Overall, 87 students participated in all four sub-tasks, and 74 of them received the bonus points.

#### 5 Discussion and Conclusions

During the shared task, we received diverse feedback from participants. Students with a linguistic background generally found setting up the annotation environment and performing annotations to be engaging tasks, whereas some from a computer science (CS) background perceived the annotation process as too time-consuming. Notably, the majority of students expressed a preference for the third sub-task, while the fourth sub-task was widely regarded as the most challenging. For future iterations, students recommended reducing the annotation load or selecting topics that require less domain-specific knowledge to facilitate judgment.

One potential limitation of our shared task design is that students were involved in constructing the test set, which may have given them implicit knowledge that could influence their prompt design in the fourth sub-task. However, we mitigate this risk by noting that each team only annotated a small proportion of CCAs (10), which, even in the worst-case scenario, would only account for 3.8% of the entire testing dataset. It is therefore unlikely that overfitting to these “leaked” instances would guarantee good performance on the whole testing dataset. Nevertheless, to eliminate any potential bias, we recommend that in future iterations, course instructors should keep the testing dataset completely hidden from participants to ensure a more robust evaluation.

Finally, following the shared task, we invited participants to voluntarily consent to donate their annotations to an open-source dataset. In total, we collected 850 annotated LLM answers for 130 CCAs. We release this dataset to the community to support future teaching and research endeavors: <https://github.com/UKPLab/folt-shared-task-23-24>.

## References

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. [Label Studio: Data labeling software](#). Open source software available from <https://github.com/heartexlabs/label-studio>.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. 2023. [Llama 2: Open Foundation and Fine-Tuned Chat Models](#). *ArXiv preprint*, abs/2307.09288.

## A Prompts for Testing LLMs

### **Llama-2-70b-chat Prompt 1:**

Provide a concise answer for the following question. {Question}

### **Llama-2-70b-chat Prompt 2:**

You are a helpful, respectful and honest assistant. Provide a concise answer for the following question. Always answer as helpfully as possible, while being safe. Your answers should not include any harmful, unethical, racist, sexist, toxic, dangerous, or illegal content. Please ensure that your responses are socially unbiased and positive in nature. If a question does not make any sense, or is not factually coherent, explain why instead of answering something not correct. If you don't know the answer to a question, please don't share false information. {Question}

### **ChatGPT/BingChat/PerplexityAI Prompt:**

Provide a concise answer for the following question. {Question}