A Zero-shot and Few-shot Study of Instruction-Finetuned Large Language Models Applied to Clinical and Biomedical Tasks
Yanis Labrak, Mickaël Rouvier, Richard Dufour

To cite this version:
Yanis Labrak, Mickaël Rouvier, Richard Dufour. A Zero-shot and Few-shot Study of Instruction-Finetuned Large Language Models Applied to Clinical and Biomedical Tasks. Fourteenth Language Resources and Evaluation Conference (LREC-COLING 2024), Nicoletta Calzolari; Min-Yen Kan, May 2024, Torino, Italy. hal-04470883

HAL Id: hal-04470883
https://hal.science/hal-04470883
Submitted on 21 Feb 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons CC0 - Public Domain Dedication| 4.0 International License
A Zero-shot and Few-shot Study of Instruction-Finetuned Large Language Models Applied to Clinical and Biomedical Tasks

Yanis Labrak∗1,2, Mickael Rouvier∗1 and Richard Dufour∗1,3

1LIA, Avignon Université 2Zenidoc
3Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France
{first.last}@{univ-avignon.fr, univ-nantes.fr}

Abstract
The recent emergence of Large Language Models (LLMs) has enabled significant advances in the field of Natural Language Processing (NLP). While these new models have demonstrated superior performance on various tasks, their application and potential are still underexplored, both in terms of the diversity of tasks they can handle and their domain of application. In this context, we evaluate four state-of-the-art instruction-tuned LLMs (ChatGPT, Flan-T5 UL2, Tk-Instruct, and Alpaca) on a set of 13 real-world clinical and biomedical NLP tasks in English, including named-entity recognition (NER), question-answering (QA), relation extraction (RE), and more. Our overall results show that these evaluated LLMs approach the performance of state-of-the-art models in zero- and few-shot scenarios for most tasks, particularly excelling in the QA task, even though they have never encountered examples from these tasks before. However, we also observe that the classification and RE tasks fall short of the performance achievable with specifically trained models designed for the medical field, such as PubMedBERT. Finally, we note that no single LLM outperforms all others across all studied tasks, with some models proving more suitable for certain tasks than others.

Keywords: NLP evaluation, Benchmarking, Medical domain, Biomedical, Clinical, Large Language Models, BERT, Transformers

1. Introduction

Medical domain is currently benefiting greatly from significant progress in Natural Language Processing (NLP), thanks to the availability of massive textual databases and the use of deep learning techniques that allow for more efficient exploitation of this data. Traditionally, the approach involved training a generic masked language model (MLM) and then adapting it to a specific domain or task, such as BERT models (Devlin et al., 2019). However, the latest approaches aim to develop Large Language Models (LLMs) that can directly process a wide range of NLP tasks and domains. They can then handle tasks as classification or entity extraction, as well as more complex generative tasks like machine translation or question-answering.

While there is clear enthusiasm for LLMs among both scientists and the general public, the evaluation of these models, also known as foundation models, is still in its infancy. The initial evaluations demonstrate the usefulness of these models in performing various NLP tasks, including classification and generation tasks on general domains (Liu et al., 2023; Bang et al., 2023). However, in the medical field, these models have been evaluated to a lesser extent, often on a limited number of tasks (Rehana et al., 2023; Chen et al., 2023; Lamichhane, 2023; Singhal et al., 2022; Chowdhery et al., 2022). This is mainly due to the scarcity of tasks and data, particularly sensitive data that is difficult to obtain, compared to other fields.

To evaluate how well LLMs encode medical knowledge and to demonstrate their capabilities in specific domains, a wide range of tasks that closely resemble real-world applications and require appropriate medical knowledge and expert reasoning were considered. Unlike other studies (Fries et al., 2022; Jin et al., 2021) that have compared performances of these models using automatic metrics (BLUE (Papineni et al., 2002), ROUGE (Lin, 2004) or BertScore (Zhang et al., 2020)) or only accuracy on a small set of tasks, we decide to showcase their relevance in various evaluation contexts by using more commonly used metrics (Accuracy and F1) which are allowing a fair direct comparison with BERT-based models. In overall, we curate a collection comprising 13 real-world medical tasks, including classification (CLS), question-answering (QA), relation extraction (RE), natural language inference (NLI) and named-entity recognition (NER). The main contributions of the paper are:

- Evaluation of four state-of-the-art instruction-tuned models (ChatGPT, Flan-T5 UL2, Tk-Instruct, and Alpaca) on a broad range of medical tasks in English language beyond those typically addressed by generative models.
- Assessment of the ability of the studied language models to perform zero- and few-shot inference and comparison of their performance on the tasks with that of a fine-tuned
PubMedBERT model.

- Introduction of a novel method called Recursive Chain-of-Thought (RCoT) that enables performing the NER task on all types of LLMs thanks to the use of a prompt sequentially enriched to mimic human reasoning.

## 2. Related work

We first introduce the concept of Large Language Models (LLMs) and their limitations (Section 2.1). Next, we present the concept of instruction-tuning (Section 2.2). Finally, we describe our few-shot learning strategy with prompts (Section 2.3).

### 2.1. Large Language Models (LLMs)

While classical language models like BERT are efficient across various NLP tasks and trained on substantial amounts of unannotated textual data, they still necessitate a significant quantity of annotated data to excel in specific tasks, such as NER, NLI, and RE. Moreover, these models encounter challenges when attempting to generalize their knowledge to other languages or domains after being adapted to a particular task and context (Peng et al., 2021). Collecting such data for any scenario can be costly, demanding highly skilled annotators and giving rise to privacy concerns.

Recently, LLMs have brought additional performance improvements, especially in generative tasks. These models are composed of billion of parameters and trained on gigantic amounts of data, from various natures, domains and languages (Gao et al., 2020; Raffel et al., 2020; Ortiz Suarez et al., 2019). Previous studies have demonstrated in particular that this gigantic number of parameters associated with this massive data allowed the fine modeling of the language, making it possible to achieve this level of performance (Zhang et al., 2022; Black et al., 2022; Hoffmann et al., 2022; Smith et al., 2022).

New approaches leveraging the generative capabilities of LLMs have aimed to align them with instructions (Ouyang et al., 2022) (see Section 2.2), thereby enhancing their capacity to handle a multitude of NLP tasks in multiple languages using zero-shot or few-shot learning (Bang et al., 2023).

### 2.2. Instruction Tuning

Efrat and Levy (2020) and Mishra et al. (2022b) propose the instruction paradigm, in which models can learn new tasks based on natural language instructions only. These instructions are given as inputs to the models, describing how they should behave, what we expect from them, and on which information they can base their thinking on. Wang et al. (2022b) introduced the first large-scale instruction benchmark called SUPER-NATURALINSTRUCTIONS, by collecting crowdsourced instructions based on an existing set of 1600+ NLP datasets and converting them into a uniform format. Sanh et al. (2022) and Wei et al. (2022a) further extend the adoption of instructions by suggesting instruction tuning, in which a LLM is trained on many natural language instructions with the aspiration that it will generalize to new, unseen instruction tasks. Chung et al. (2022b) advance instruction tuning by scaling the number of tasks, scaling the model size, and introducing the concept of chain-of-thought (Wei et al., 2022b), while Ouyang et al. (2022) propose a reinforcement learning approach for instruction tuning and human feedback.

### 2.3. Few-shot Learning with prompts

During inference, a few examples of the task are given to the model as conditioning, without updating its weights. These examples usually comprise an instruction, context, and desired completion (e.g., a premise, hypothesis, and corresponding label for the NLI task). The few-shot technique involves presenting the model with \( k \) examples of context and completion, followed by a final example of context, for which the model should provide the completion. The value of \( k \) typically ranges from 3 to 100, which depends on the number of examples that can fit within the model’s context window (for instance, Flan-UL2 has a context window of 2,048 tokens).

## 3. Experimental Protocol

In this section, we describe the models utilized and the datasets used to benchmark the various models.

### 3.1. Studied Models

Our evaluation involves four distinct generic LLMs (ChatGPT, Flan-UL2, Tk-Instruct and Alpaca) and a specific biomedical masked language model (PubMedBERT) for comparison purposes.

**Flan-T5 UL2** abbreviated to Flan-UL2, is an encoder-decoder model based on UL2 20B parameters model (Tay et al., 2023) and was fine-tuned using the Flan instruction tuning tasks collection (Chung et al., 2022b).

**Tk-Instruct** is based on the T5 encoder-decoder model (Raffel et al., 2020) and has been fine-tuned on the 1,600+ NLP tasks from the SUPER-NATURALINSTRUCTIONS dataset (Wang et al., 2022b).
In our study, we chose the 3B parameter setting, since our preliminary comparison with Flan-T5-XL (Chung et al., 2022a) using the 3B parameter setting showed that Tk-Instruct performed better on QA tasks, which is considered to be one of the most suited tasks for LLMs.

ChatGPT is built upon GPT-3.5 Turbo, fine-tuned with a set of proprietary instructions, and continuously refined through reinforcement learning from human feedback (RLHF) techniques. Access to its weights is restricted, and the model can only be accessed via a paid API. These restrictions raise privacy concerns regarding its application in medical contexts, and it cannot ensure that the evaluated data has not been previously encountered.

Stanford Alpaca is built upon LLaMA with 7B parameters (Touvron et al., 2023) and utilizes a dataset of 52K instructions, which were automatically generated in the style of self-instruct using OpenAI’s text-davinci-003 model (Wang et al., 2022a). Due to its base model and data sources, it is exclusively intended for academic research purposes and non-commercial use.

PubMedBERT is a biomedical-specific BERT-based model with 110M parameters (Gu et al., 2021). It was trained entirely from scratch on the 3.1 billion words of the PubMed corpus. We chose it as our baseline for comparison with the zero-shot and few-shot performance of generative models.

### 3.2. Downstream evaluation tasks

We conducted an evaluation of the models’ capabilities by encompassing the test set of the 13 diverse tasks listed in Table 1. These tasks were chosen to facilitate a comprehensive assessment spanning both clinical and biomedical domains, including tasks suitable for both generative and classical model evaluations.

<table>
<thead>
<tr>
<th>Task</th>
<th>Dataset</th>
<th>Eval</th>
<th>Metric</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLS</td>
<td>HoC</td>
<td>Test</td>
<td>F1-measure</td>
<td>Baker et al. (2016)</td>
</tr>
<tr>
<td>LiD</td>
<td>Test</td>
<td>F1-measure</td>
<td>Chen et al. (2021)</td>
<td></td>
</tr>
<tr>
<td>Health</td>
<td>Test</td>
<td>Accuracy</td>
<td>Neema and Toni (2020)</td>
<td></td>
</tr>
<tr>
<td>NLI 2008 Smokers</td>
<td>Test</td>
<td>Accuracy</td>
<td>Uzuner et al. (2009)</td>
<td></td>
</tr>
<tr>
<td>QA</td>
<td>BioASQ 7b</td>
<td>Test</td>
<td>Accuracy</td>
<td>Tsatsaronis et al. (2015)</td>
</tr>
<tr>
<td>MedQA</td>
<td>Dev</td>
<td>Accuracy</td>
<td>Pal et al. (2022)</td>
<td></td>
</tr>
<tr>
<td>SciQ</td>
<td>Test</td>
<td>Accuracy</td>
<td>Welbl et al. (2017)</td>
<td></td>
</tr>
<tr>
<td>Evidence Inference 2.0</td>
<td>Test</td>
<td>Accuracy</td>
<td>DeYoung et al. (2020)</td>
<td></td>
</tr>
<tr>
<td>RE</td>
<td>QAID</td>
<td>Test</td>
<td>Accuracy</td>
<td>Bane et al. (2015)</td>
</tr>
<tr>
<td>NLI</td>
<td>SciTe</td>
<td>Test</td>
<td>Accuracy</td>
<td>Khair et al. (2018)</td>
</tr>
<tr>
<td>MedNLI</td>
<td>Test</td>
<td>Accuracy</td>
<td>Shrivastava et al. (2017)</td>
<td></td>
</tr>
<tr>
<td>NER</td>
<td>BioASQ</td>
<td>Test</td>
<td>F1-measure</td>
<td>Li et al. (2016)</td>
</tr>
<tr>
<td>NCBI-disease</td>
<td>Test</td>
<td>F1-measure</td>
<td>Dogan et al. (2014)</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: List of evaluation tasks and their metrics. CLS: Classification, QA: Question Answering, RE: Relation Extraction, NLI: Natural Language Inference, NER, Named-Entity Recognition.

### 3.3. Evaluation of generative outputs

Evaluating the outputs of generative models presents a challenge due to their free-text nature, which may not necessarily conform to a predefined set of classes. Instead, we are confronted with noisy outputs that may contain correct answers. To address this challenge, we manually developed parsing scripts tailored to each task and model, aligning them with their respective output styles. This approach enables us to capture most of the answers and compute metrics that can be compared with our baseline model (PubMedBERT).

### 3.4. Instruction Format

Previous studies (Wei et al., 2022b; Jung et al., 2022; Mishra et al., 2022a) have demonstrated the effectiveness of using task-specific prompts for each model. Consequently, we chose to construct the input instruction prompt by concatenating three elements: (1) an instruction that outlines the task, describes the nature of the data, and specifies our expectations from the model, (2) the input argument, which provides essential information for the task, and (3) the constraints on the output space, which guide the model during output generation. Lastly, the output serves as a reference point during the few-shot strategy evaluation. More information about the instruction formats in Appendix A.

### 3.5. Few-shot Examples using Semantic Retriever

To enhance few-shot performance compared to randomly sampled examples, we introduced an additional retrieval module based on SentenceTransformers (Reimers and Gurevych, 2019). The objective is to identify the $k$ most semantically similar examples from the training set. To accomplish this, we first populate a vector space with sentence representations of each individual instruction prompt from the training set, obtained using a pre-trained and fixed PubMedBERT (Gu et al., 2021) model. Subsequently, we compute the cosine distance between the query of the current test instance and all the elements within the vector space to retrieve the top $k$ closest examples. In our case, we set the value of $k$ to 5.

### 3.6. Recursive Chain-of-Thought

We performed NER using two inference methods. The first one is based on the method introduced by Ye et al. (2023) and can only be applied using ChatGPT. It consists of giving the model a sequence of words separated by double vertical bars for word separation and single vertical bars for the
Table 2: 0- and 5-shot versus finetuning evaluation on clinical and biomedical tasks. Bold values are the highest scores obtained for the task and in underlined the seconds ones. Not allowed experiments are replaced by NaN.

<table>
<thead>
<tr>
<th>Task</th>
<th>Dataset</th>
<th>ChatGPT zero-shot</th>
<th>ChatGPT 5-shot</th>
<th>Flan-UL2 zero-shot</th>
<th>Flan-UL2 5-shot</th>
<th>Tk-Instruct zero-shot</th>
<th>Tk-Instruct 5-shot</th>
<th>Alpaca zero-shot</th>
<th>Alpaca 5-shot</th>
<th>PubMedBERT</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLS</td>
<td>HoC</td>
<td>62.24</td>
<td>38.34</td>
<td>56.36</td>
<td>54.86</td>
<td>50.77</td>
<td>25.48</td>
<td>1.21</td>
<td>38.78</td>
<td>82.75</td>
</tr>
<tr>
<td></td>
<td>LitCovid</td>
<td>67.20</td>
<td>72.77</td>
<td>51.48</td>
<td>46.95</td>
<td>36.42</td>
<td>57.49</td>
<td>1.58</td>
<td>64.09</td>
<td>90.60</td>
</tr>
<tr>
<td></td>
<td>PubMedHealth</td>
<td>63.20</td>
<td>66.29</td>
<td>72.46</td>
<td>50.53</td>
<td>53.70</td>
<td>66.04</td>
<td>52.80</td>
<td>55.64</td>
<td>75.39</td>
</tr>
<tr>
<td></td>
<td>N2C2 2006 Smokers</td>
<td>NaN</td>
<td>NaN</td>
<td>22.12</td>
<td>42.31</td>
<td>16.35</td>
<td>37.50</td>
<td>10.57</td>
<td>31.73</td>
<td>60.58</td>
</tr>
<tr>
<td>QA</td>
<td>BioASQ 7b</td>
<td>89.24</td>
<td>92.03</td>
<td>90.97</td>
<td>91.64</td>
<td>88.09</td>
<td>86.36</td>
<td>79.05</td>
<td>79.82</td>
<td>73.39</td>
</tr>
<tr>
<td></td>
<td>MedMCGA</td>
<td>48.91</td>
<td>56.37</td>
<td>41.05</td>
<td>43.34</td>
<td>33.85</td>
<td>33.18</td>
<td>24.91</td>
<td>29.50</td>
<td>38.15</td>
</tr>
<tr>
<td></td>
<td>SciQ</td>
<td>90.10</td>
<td>93.50</td>
<td>87.00</td>
<td>88.40</td>
<td>55.30</td>
<td>47.00</td>
<td>24.90</td>
<td>36.80</td>
<td>74.20</td>
</tr>
<tr>
<td></td>
<td>Evidence Inference 2.0</td>
<td>59.98</td>
<td>63.83</td>
<td>66.45</td>
<td>65.06</td>
<td>41.33</td>
<td>38.79</td>
<td>32.49</td>
<td>94.18</td>
<td>65.47</td>
</tr>
<tr>
<td>RE</td>
<td>GAD</td>
<td>47.75</td>
<td>52.25</td>
<td>49.81</td>
<td>53.37</td>
<td>48.88</td>
<td>57.87</td>
<td>51.12</td>
<td>57.68</td>
<td>79.78</td>
</tr>
<tr>
<td>NLI</td>
<td>SciTail</td>
<td>73.57</td>
<td>65.62</td>
<td>93.51</td>
<td>92.66</td>
<td>57.53</td>
<td>71.31</td>
<td>39.60</td>
<td>40.26</td>
<td>93.51</td>
</tr>
<tr>
<td></td>
<td>MedNLi</td>
<td>NaN</td>
<td>NaN</td>
<td>77.00</td>
<td>79.18</td>
<td>33.19</td>
<td>34.81</td>
<td>33.47</td>
<td>34.45</td>
<td>83.76</td>
</tr>
<tr>
<td>NER</td>
<td>BC5CDR</td>
<td>92.12</td>
<td>93.12</td>
<td>68.26</td>
<td>83.32</td>
<td>84.54</td>
<td>83.23</td>
<td>82.11</td>
<td>84.07</td>
<td>97.65</td>
</tr>
<tr>
<td></td>
<td>NCBI-disease</td>
<td>90.97</td>
<td>92.27</td>
<td>90.75</td>
<td>87.65</td>
<td>87.91</td>
<td>87.50</td>
<td>11.58</td>
<td>92.27</td>
<td>98.72</td>
</tr>
</tbody>
</table>

separation between words and labels. For the second method, we introduce a method called Recursive Chain-of-Thought (RCoT). It is very close to human reasoning and works for all the generative models we have tried. It is derived from the Chain-of-Thought (CoT) concept (Wei et al., 2022b) and the work of Wang et al. (2022b). It involves iterating over the sequence of tokens and giving the current state of the prediction as input to the model, asking for the generation of the label of the $N^{th}$ token. This method guarantees an entity for each token of the sequence and prevents forgotten tokens during generation. However, the only drawback we have been able to identify with this method is its very high computation cost due to its $O^N$ complexity, with $N$ being the number of tokens in the sequence, compared to the method used for ChatGPT, which performs at $O^1$ complexity.

4. Results and Discussions

Table 2 reports performance obtained on each task by the studied LLMs in zero- and few-shot scenario, as well as PubMedBERT fine-tuned. Results are reported by taking the best run out of four.

Zero-shot scenario Compared to PubMedBERT, the zero-shot scenario results show a clear deficit for the generative models on all the tasks except for QA, in which LLMs obtain better performance. ChatGPT and Flan-T5 UL2 are particularly perform better than Tk-Instruct and Alpaca on average, except for GAD dataset (RE task) for which Alpaca reaches the best performance. We can also observe extremely poor performance from Alpaca in zero-shot scenario on the two CLS tasks (HoC and LitCovid). These low scores are attributed to the model generating hallucinated responses, including the label evading growth suppressors across the entire test set of HoC. However, this behavior does not appear to occur in the few-shot scenario, where the model appears to comprehend our expectations.

Few-shot capabilities Unlike the zero-shot scenario, the few-shot inference (5-shots in our experiments) shows impressive behavior. The biggest absolute gains are obtained using Alpaca, which seems to perform much better in few-shot scenarios on all tasks. We suspect this behavior to be correlated with Alpaca’s training data, which does not contain many similar instructions for the tasks we are trying to tackle, allowing it to better understand what we are asking when confronted with dissimilar examples. ChatGPT also benefits from the additional knowledge to further improve the already good results, especially on QA tasks. Flan-T5 UL2 appears to be less affected by the additional context overall, except for the BC5CDR and N2C2 2006 Smokers tasks.

5. Conclusion

In this study, we have demonstrated that generic LLMs are capable of capturing medical knowledge and performing exceptionally well in zero- and few-shot scenarios, despite having no prior exposure to the tasks. Although open-source models such as Flan-T5 UL2 are gradually approaching their closed-source counterparts, such as ChatGPT, their performance still lags behind. We suggest that developing domain-specific models, fine-tuned on a diverse set of tasks and specialized instruction prompts, could help bridge the gap with more robust and performant proprietary models. We also note that domain-specific BERT models remain a viable option, but require a significant amount of data for fine-tuning on targeted languages and tasks. However, BERT-based models offer much lower computational costs compared.
to LLMs, which could be a significant obstacle to developing models in the healthcare domain.

6. Acknowledgements

This work was performed using HPC resources from GENCI-IDRIS (Grant 2022-AD011013061R1). This work was financially supported by ANR MALADES (ANR-23-IAS1-0005) and Zenidoc.

7. Limitations

Through all the experiments we conducted, we have observed that Large Language Models (LLMs) trained based on instructions often exhibit sensitivity to the specific wording used as input, which can influence their ability to generate correct outputs. This finding may not come as a surprise, as LLMs are well-documented to be highly responsive to the prompts they receive, whether in zero-shot or few-shot settings [cite relevant sources]. However, it frequently necessitates tailoring the prompts to suit the models and tasks, or even mapping the classes to more suitable ones. This sensitivity may stem from the limited diversity in the collections of instructions used for their training. One of the primary limitations is our inability to guarantee that the ChatGPT model has not encountered the evaluation data during its training, potentially introducing bias into the results. Similarly, Flan-T5 UL2 and Tk-Instruct have been trained on a broad spectrum of tasks, which could result in the model being exposed to similar or identical data if overlaps are not identified. As a result, we cannot ensure that the training data for certain tasks has never been seen before.

8. Bibliographical References


Jason Wei. 2022b. Scaling instruction-finetuned language models.


Avia Efrat and Omer Levy. 2020. The turking test: Can language models understand instructions?


Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin,


Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhu Lin, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, Elton Zhang, Rewon Child, Reza Yazdani Aminabadi, Julie Bernauer, Xia Song, Mohammad Shoyebi, Yuxiong He, Michael Houston, Saurabh Tiwary, and Bryan Catanzaro. 2022. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model.


Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. Fine-tuned language models are zero-shot learners.


9. Language Resource References
Appendices

A. Instructions Formats

The following sections are giving example of prompts used for training and inference for organized by tasks.

A.1. Named-Entities Recognition

A.1.1. Method 1

Prompts

**Instruction:** Do named-entity recognition task for the given text using the categories in candidate list, output using the format as “Word1|Category||Word2|Category||Word3|Category”

**Candidate list:** O, B-Disease or I-Disease

**Text:** Identification|Category || of|Category || APC2|Category || ,|Category || a|Category || homologue|Category || of|Category || the|Category || adenomatous|Category || polyposis|Category || coli|Category || tumour|Category || suppressor|Category || .|Category

**Output:**

**Instruction:** You are a healthcare named-entity recognition expert system and we are giving you a sequence of words that you have to labelized using the following output format ‘Word1|Label||Word2|Label||Word3|Label’

**Labels:** O, B-Disease or I-Disease

**Unfilled sequence:** Identification|Label||of|Label||APC2|Label||,|Label||a|Label||homologue|Label||of|Label ||the|Label ||adenomatous|Label||polyposis|Label||coli|Label||tumour|Label||suppressor|Label||.|Label

**Constraints:** The answer must be one and only one of the given labels.

**Output:**

**Instruction:** As a healthcare named-entity recognition expert, your job is to label a sequence of words provided to you using the following format: ‘Word1|Label||Word2|Label||Word3|Label’. Your goal is to identify all the named entities in the given text. The available labels for this task are: O, B-Disease or I-Disease

**Input:** Identification|Label||of|Label||APC2|Label||,|Label||a|Label||homologue|Label||of|Label ||the|Label ||adenomatous|Label||polyposis|Label||coli|Label||tumour|Label||suppressor|Label||.|Label

**Output:**

Table 3: Sample of three instructions used for the named-entities recognition task with ChatGPT.
A.1.2. Method 2 - Recursive Chain-Of-Thought (RCoT)

Prompt - Recursive Chain-Of-Thought (RCoT)

**Instruction:** You are a highly intelligent and accurate healthcare domain Named-entity recognition (NER) system. You are tasked to do Named-entity recognition (NER) for ‘disease’ and ‘none’ only, please generate the appropriate label.

**Constraints:** You can choose only one label from: *none* or *disease*.

**Examples:**

**Example 1:** Mutations|none|| at|none|| the|none|| ataxia|disease|| -|disease|| telangiectasia|disease|| locus|none|| and|none|| clinical|none|| phenotypes|none|| of|none|| A|disease|| -|disease|| T|disease|| patients|none|| .|none

**Example 2:** Splicing|none|| defects|none|| in|none|| the|none|| ataxia|disease|| -|disease|| telangiectasia|disease|| gene|none|| ,|none|| ATM|none|| ;|none|| underlying|none|| mutations|none|| and|none|| consequences|none|| .|none

**Example 3:** Somatic|none|| mutations|none|| in|none|| the|none|| BRCA1|none|| gene|none|| in|none|| sporadic|disease|| ovarian|disease|| tumours|disease|| .|none

**Example 4:** Malignant|disease|| neoplasms|disease|| in|none|| the|none|| families|none|| of|none|| patients|none|| with|none|| ataxia|disease|| -|disease|| telangiectasia|disease|| .|none

**Example 5:** Founder|none|| mutations|none|| in|none|| the|none|| BRCA1|none|| gene|none|| in|none|| Polish|none|| families|none|| with|none|| breast|disease|| -|disease|| ovarian|disease|| cancer|disease|| .|none

**Considering the sentence:** Clustering of missense mutations in the ataxia - telangiectasia gene in a sporadic T-cell leukaemia.

And considering your precedents predictions: Clustering|none|| of|none|| missense|none|| mutations|none|| in|none|| the|none|| ataxia|disease|| -disease|| telangiectasia|disease|| gene|none|| in|none|| a|none|| sporadic|disease|| T|disease|| -disease|| cell|disease|| leukaemia|Label

**Input:** The label of « leukaemia » at the position 17 of the sentence is ?

**Output:**

Table 4: Example of a 5-shot Recursive Chain-Of-Thought (RCoT) instruction used for the named-entities recognition task of NCBI Disease dataset.

B. Multiple-choice question answering

B.1. Method 1 - One-shot

Prompt

**Instruction:** You are given a science question (easy level) and four answer options (associated with "A", "B", "C", "D"). Your task is to find the correct answer based on scientific facts, knowledge and reasoning. Don’t generate anything other than one of the following characters: 'A B C D'.

**Input:** Heavy forces on periodontal ligament causes: (A) Hyalinization (B) Osteoclastic activity around tooth (C) Osteoblastic activity around tooth (D) Crest bone resorption

**Constraints:** The answer must be one or more of the following letters: 'A','B','C','D'. You must generate one and only one letter for each question. All questions have an answer. No justification is required.

**Output:**

Table 5: Example of a 0-shot instruction used for the Multiple-Choice Question Answering (MCQA) task of MedMCQA dataset.
B.2. Method 2 - Few-shot

In some cases, we mapped the original classes to more effective one's for each of the tasks, based on
tries and errors (e.g: "entailment" has been map to "entails" for ChatGPT and Flan-T5 UL2 based on
noticeable performances gains).

Prompt

**Instruction:** You are given a science question (easy level) and four answer options (associated with "A", "B", "C",
"D"). Your task is to find the correct answer based on scientific facts, knowledge and reasoning. Don’t generate
anything other than one of the following characters: 'A B C D'.

**Constraints:** The answer must be one or more of the following letters: 'A'-'B'-'C'-'D'. You must generate one and only
one letter for each question. All questions have an answer. No justification is required.

**Examples:**

**Example 1:** Hyalinisation of the periodontal Ligament, due to excessive orthodontic forces results in (A) Frontal
resorption (B) Undermining resorption (C) Cementum remaining intact (D) Dentine remaining intact
  **Output:** B

**Example 2:** The earliest response of pulpitis is: (A) Cyst formation (B) Calcification (C) Hyalinization (D) Formation
of dental granuloma
  **Output:** C

**Example 3:** Among the secondary changes in tooth the most useful one for age determination is: (A) Attrition (B)
Secondary dentine deposition (C) Root resorption (D) Root transparency
  **Output:** D

**Example 4:** Feature of aging periodontium is (A) Lacunae in bone and cementum (B) Increased cell size (C)
Increased cell number (D) Scalloping of cementum & alveolar bone surface
  **Output:** D

**Example 5:** Bacteria found in gingivitis are localized in (A) Connective tissue fibres (B) Gingival sulcus (C) Alveolar
bone (D) Periodontal ligament
  **Output:** B

**Input:** Heavy forces on periodontal ligament causes: (A) Hyalinization (B) Osteoclastic activity around tooth (C)
Osteoblastic activity around tooth (D) Crest bone resorption
  **Output:**

Table 6: Example of a 5-shot instruction used for the Multiple-Choice Question Answering (MCQA) task
of MedMCQA dataset.
C. Relation Extraction

C.1. Method 1 - One-shot

Prompt

**Instruction:** Your goal is to do relation extraction and identifying if a gene-disease relation exist (positive) or not (negative).

**Input:** These results suggest that the C1772T polymorphism in @GENE$ is not involved in progression or metastasis of @DISEASE$.

**Constraints:** You have to output one label among « negative » or « positive ». Justification and explanations are prohibited.

**Output:**

Table 7: Example of a 0-shot instruction used for the Relation Extraction (RE) task of GAD dataset.

C.2. Method 2 - Few-shot

Prompt

**Instruction:** Your goal is to do relation extraction and identifying if a gene-disease relation exist (positive) or not (negative).

**Constraints:** You have to output one label among « negative » or « positive ». Justification and explanations are prohibited.

**Examples:**

**Example 1:** These findings suggest that the Gly460Trp polymorphism of @GENE$ is not associated with @DISEASE$.

**Output:** Positive

**Example 2:** Our results suggest that deletion polymorphism of the @GENE$ gene is not associated with the pathogenesis of @DISEASE$ in Taiwanese.

**Output:** Positive

**Example 3:** The results suggest that the 5A/6A polymorphism of @GENE$ gene may not be linked with appearance and/or progression of @DISEASE$.

**Output:** Positive

**Example 4:** Our study implies that the G/C polymorphism of the @GENE$ gene may not be directly involved in the development and/or progression of @DISEASE$.

**Output:** Positive

**Example 5:** Our study implies that the G/C polymorphism of the @GENE$ gene may not be directly involved in the development and/or @DISEASE$ of breast cancer.

**Output:** Negative

**Input:** These results suggest that the C1772T polymorphism in @GENE$ is not involved in progression or metastasis of @DISEASE$.

**Output:**

Table 8: Example of a 5-shot instruction used for the Relation Extraction (RE) task of GAD dataset.
D. Natural Language Inference

D.1. Method 1 - One-shot

Prompt

**Instruction:** Your goal is to solve a natural language inference task by identifying if the hypothesis is either « entails » or « neutral » to the premise.

**Input premise:** The liver is divided into the right lobe and left lobes.

**Input hypothesis:** The gallbladder is near the right lobe of the liver.

**Constraints:** You have to output one label among « entails » or « neutral ». Justification and explanations are prohibited.

**Output:**

Table 9: Example of a 0-shot instruction used for the Natural Language Inference (NLI) task of SciTail dataset.
D.2. Method 2 - Few-shot

---

**Prompt**

**Instruction:** Your goal is to do solve a natural language inference task by identifying if the hypothesis is either « entails » or « neutral » to the premise.

**Constraints:** You have to output one label among « entails » or « neutral ». Justification and explanations are prohibited.

**Examples:**

Example 1:
**Premise:** Located primarily on the right side of the abdominal cavity, just above the duodenum, the liver aids in the digestion of fats by secreting bile into the duodenum.
**Hypothesis:** Most digestion is completed in the duodenum.
**Output:** neutral

Example 2:
**Premise:** The brain is divided into the right and left hemisphere and each hemisphere is divided into 4 lobes called the frontal, temporal, occipital and parietal lobes.
**Hypothesis:** Each hemisphere of the cerebrum divided into 4 lobes.
**Output:** entails

Example 3:
**Premise:** The small intestine, where most digestion takes place, is a convoluted tube in the abdomen that begins at the pylorus of the stomach and ends at the opening to the large intestine.
**Hypothesis:** Most of the digestion reactions occur in the small intestine.
**Output:** entails

Example 4:
**Premise:** The small intestine is the long, thin segment of bowel that begins at the stomach and ends at the large intestine or colon.
**Hypothesis:** The small intestine begins in the stomach.
**Output:** entails

Example 5:
**Premise:** The small intestine begins at the stomach and ends at the colon (large intestine).
**Hypothesis:** The small intestine begins in the stomach.
**Output:** entails

**Premise:** The liver is divided into the right lobe and left lobes.
**Hypothesis:** The gallbladder is near the right lobe of the liver.
**Output:**

Table 10: Example of a 5-shot instruction used for the Natural Language Inference (NLI) task of SciTail dataset.
E. Classification

E.1. Method 1 - One-shot

**Prompt**

**Instruction:** Your goal is to solve a classification task by identifying if one or more of the following hallmarks of cancer are present in the document: « evading growth suppressors », « tumor promoting inflammation », « enabling replicative immortality », « cellular energetics », « resisting cell death », « activating invasion and metastasis », « genomic instability and mutation », « none », « inducing angiogenesis », « sustaining proliferative signaling » or « avoiding immune destruction ».

**Input:** Cytotoxicity was shown in manganese-treated groups (100, 200, 400, and 800 microM of MnCl(2)) and cell viability was decreased to 58.8% of the control group at 2 days after treatment with 800 microM of MnCl(2).

**Constraints:** You have to output one or more label(s) among « evading growth suppressors », « tumor promoting inflammation », « enabling replicative immortality », « cellular energetics », « resisting cell death », « activating invasion and metastasis », « genomic instability and mutation », « none », « inducing angiogenesis », « sustaining proliferative signaling » or « avoiding immune destruction ». Justification and explanations are prohibited.

**Output:**

Table 11: Example of a 0-shot instruction used for the classification (CLS) task of HoC dataset.
E.2. Method 2 - Few-shot

Prompt

Instruction: Your goal is to do solve a classification task by identifying if one or more of the following hallmarks of cancer are present in the document: « evading growth suppressors », « tumor promoting inflammation », « enabling replicative immortality », « cellular energetics », « resisting cell death », « activating invasion and metastasis », « genomic instability and mutation », « none », « inducing angiogenesis », « sustaining proliferative signaling » or « avoiding immune destruction ».

Constraints: You have to output one or more label(s) among « evading growth suppressors », « tumor promoting inflammation », « enabling replicative immortality », « cellular energetics », « resisting cell death », « activating invasion and metastasis », « genomic instability and mutation », « none », « inducing angiogenesis », « sustaining proliferative signaling » or « avoiding immune destruction ». Justification and explanations are prohibited.

Examples:

Example 1: However, significant cytotoxicity was only observed in PCB 52 concentrations larger than 0.1 microg ml(-1), while there was no significant inhibition in PCB 77-treated cells at concentrations selected.
Output: none

Example 2: In MeT-5A cells, both CNTs caused a dose-dependent induction of DNA damage ( % DNA in comet tail ) in the 48-h treatment and SWCNTs additionally in the 24-h treatment, with a statistically significant increase at 40 u03bcg/cm(2) of SWCNTs and ( after 48 h ) 80 u03bcg/cm(2) of both CNTs.
Output: none

Example 3: Copper-induced DNA strand breakage was first observed after 24 h of exposure, and was recorded again at 96 h, at a copper concentration of 20 microg l(-1).
Output: genomic instability and mutation

Example 4: Drug concentrations of 12.5 to 300 ¹03bcM caused a pronounced reduction in cell survival rates five days after treatment, whereas concentrations higher than 25 ¹03bcM were effective in reducing the survival rates to However, the maximum apoptosis frequency was 20.4% for 25 ¹03bcM cisplatin in cells analyzed at 72 h, indicating that apoptosis is not the only kind of cell death induced by cisplatin.
Output: none

Example 5: In contrast, in MCF 7 cells, molecular iodine ( 100 microM ) inhibited growth from 100% to 83% but delta-iodolactone ( 1, 5 and 10 microM ) dose-dependently decreased growth rate from 100% to 82% and 62%, respectively.
Output: none

Input: Cytotoxicity was shown in manganese-treated groups ( 100, 200, 400, and 800microM of MnCl(2) ) , and cell viability was decreased to 58.8% of the control group at 2days after treatment with 800microM of MnCl(2).
Output:

Table 12: Example of a few-shot instruction used for the classification (CLS) task of HoC dataset.
F. Semantic Textual Similarity

F.1. Method 1 - One-shot

---

**Prompt**

**Instruction:** Give me a similarity score between 0 et 5 and only the similarity score.

**Input:** The original sentence is : "Eviter le contact de l'embout avec l'œil ou les paupières." can you tell me if the sentence is similar to : "Evitez le contact de l'embout du flacon avec l'œil ou les paupières.".

**Output:**

Table 13: Example of a 0-shot instruction used for the Semantic Textual Similarity (STS) task of DEFT-2020 task 1 dataset.