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

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]
[1]LIA, Avignon Université [2]Zenidoc
[3]Nantes 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 such 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.

Task	Dataset	Eval	Metric	Reference
CLS	HoC	Test	F1-measure	Baker et al. (2016)
	LitCovid	Test	F1-measure	Chen et al. (2021)
	PubHealth	Test	Accuracy	Neema and Toni (2020)
	N2C2 2006 Smokers	Test	Accuracy	Uzuner et al. (2008)
QA	BioASQ 7b	Test	Accuracy	Tsatsaronis et al. (2015)
	MedMCQA	Dev	Accuracy	Pal et al. (2022)
	SciQ	Test	Accuracy	Welbl et al. (2017)
	Evidence Inference 2.0	Test	Accuracy	DeYoung et al. (2020)
RE	GAD	Test	Accuracy	Bravo et al. (2015)
NLI	SciTail	Test	Accuracy	Khot et al. (2018)
	MedNLI	Test	Accuracy	Shivade (2017)
NER	BC5CDR	Test	F1-measure	Li et al. (2016)
	NCBI-disease	Test	F1-measure	Dogan et al. (2014)

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 Sentence-Transformers (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

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

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`.

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 $\mathcal{O}^N$ complexity, with N being the number of tokens in the sequence, compared to the method used for ChatGPT, which performs at $\mathcal{O}^1$ 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

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136, virtual+Dublin. Association for Computational Linguistics.

Shan Chen, Yingya Li, Sheng Lu, Hoang Van, Hugo JWL Aerts, Guergana K Savova, and Danielle S Bitterman. 2023. Evaluation of chatgpt family of models for biomedical reasoning and classification. *arXiv preprint arXiv:2304.02496*.

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

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022a. Scaling instruction-finetuned language models.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Jason Alan Fries, Leon Weber, Natasha Seelam, Gabriel Altay, Debajyoti Datta, Samuele Garda, Myungsun Kang, Ruisi Su, Wojciech Kusa, Samuel Cahyawijaya, Fabio Barth, Simon Ott, Matthias Samwald, Stephen Bach, Stella Biderman, Mario Sänger, Bo Wang, Alison Callahan, Daniel León Periñán, Théo Gigant, Patrick Haller, Jenny Chim, Jose David Posada, John Michael Giorgi, Karthik Rangasai Sivaraman, Marc Pàmies, Marianna Nezhurina, Robert Martin, Michael Cullan, Moritz Freidank, Nathan Dahlberg, Shubhanshu Mishra, Shamik Bose, Nicholas Michio Broad, Yanis Labrak, Shlok S Deshmukh, Sid Kiblawi, Ayush Singh, Minh Chien Vu, Trishala Neeraj, Jonas Golde, Albert Villanova del Moral, and Benjamin Beilharz. 2022. Bigbio: A framework for data-centric biomedical natural language processing. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. Maieutic prompting: Logically consistent reasoning with recursive explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1266–1279, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Bishal Lamichhane. 2023. Evaluation of chatgpt for nlp-based mental health applications. *arXiv preprint arXiv:2303.15727*.

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

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. 2022a. Reframing instructional prompts to GPTk's language. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 589–612, Dublin, Ireland. Association for Computational Linguistics.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022b. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.

Pedro Javier Ortiz Suarez, Benoit Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut f"ur Deutsche Sprache.

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

Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Bo Peng, Emmanuele Chersoni, Yu-Yin Hsu, and Chu-Ren Huang. 2021. Is domain adaptation worth your investment? comparing bert and finbert on financial tasks. In *Proceedings of the Third Workshop on Economics and Natural Language Processing*, pages 37–44.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Hasin Rehana, Nur Bengisu Çam, Mert Basmaci, Yongqun He, Arzucan Özgür, and Junguk Hur. 2023. Evaluation of gpt and bert-based models on identifying protein-protein interactions in biomedical text. *arXiv preprint arXiv:2303.17728*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2022. Large language models encode clinical knowledge.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, Elton Zhang, Rewon Child, Reza Yazdani Aminabadi, Julie Bernauer, Xia Song, Mohammad Shoeybi, 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.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Siamak Shakeri, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2023. Ul2: Unifying language learning paradigms.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022a. Self-instruct: Aligning language model with self generated instructions.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi,

Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022b. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

9. Language Resource References

Simon Baker, Ilona Silins, Yufan Guo, Imran Ali, Johan H"ogberg, Ulla Stenius, and Anna Korhonen. 2016. Automatic semantic classification of scientific literature according to the hallmarks of cancer. *Bioinform.*, 32(3):432–440.

Àlex Bravo, Janet Piñero, Núria Queralt-Rosinach, Michael Rautschka, and Laura I Furlong. 2015. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC Bioinformatics*, 16(1).

Qingyu Chen, Alexis Allot, Robert Leaman, Rezarta Islamaj Doğan, and Zhiyong Lu. 2021. Overview of the biocreative vii litcovid track: multi-label topic classification for covid-19 literature annotation. In *Proceedings of the seventh BioCreative challenge evaluation workshop*.

Jay DeYoung, Eric Lehman, Benjamin Nye, Iain Marshall, and Byron C. Wallace. 2020. Evidence inference 2.0: More data, better models. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 123–132, Online. Association for Computational Linguistics.

Rezarta Islamaj Dogan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. In *AAAI*.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. Biocreative V CDR task corpus: a resource for chemical disease relation extraction. *Database J. Biol. Databases Curation*, 2016.

Kotonya Neema and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. *arXiv preprint arXiv:2010.09926*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Chaitanya Shivade. 2017. Mednli — a natural language inference dataset for the clinical domain.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):138.

Ozlem Uzuner, Ira Goldstein, Yuan Luo, and Isaac Kohane. 2008. Identifying patient smoking status from medical discharge records. *Journal of*

the American Medical Informatics Association, 15(1):14–24.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.

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 do 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 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 ».

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) .

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 ŏ3bcM caused a pronounced reduction in cell survival rates five days after treatment , whereas concentrations higher than 25 ŏ3bcM were effective in reducing the survival rates to However , the maximum apoptosis frequency was 20.4% for 25 ŏ3bcM 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 beetween 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.