Effect of Multilingual and Domain-adapted Continual Pre-training on Few-shot Promptability

Ken Yano[1], **Makoto Miwa**[2,1]
[1]National Institute of Advanced Industrial Science and Technology
[2]Toyota Technological Institute

Abstract

Continual Pre-training (CPT) can help pre-trained large language models (LLMs) effectively adapt to new or under-trained domains or low-resource languages without re-training from scratch. Nevertheless, during CPT, the model's few-shot transfer ability is known to be affected for emergent tasks. We verified this by comparing the performance between the few-shot and fine-tuning settings on the same tasks. We used Llama3-ELAINE-medLLM, which was continually pre-trained on Llama3-8B, targeted for the biomedical domain, and adapted for multilingual languages (English, Japanese, and Chinese). We compared the performance of Llama3-ELAINE-medLLM and Llama3-8B in three emergent tasks: two related domain tasks, entity recognition (NER) and machine translation (MT), and one out-of-domain task, summarization (SUM). Our experimental results show that degradation in few-shot transfer ability does not necessarily indicate the model's underlying potential on the same task after fine-tuning.

1 Introduction

Continual Pre-training (CPT) can help pre-trained large language models (LLMs) effectively adapt to new or under-trained domains or low-resource languages without re-training from scratch.

Because open-source foundation LLMs such as the Llama series (Touvron et al., 2023a,b) are under-trained for the biomedical domain and non-English languages, many studies have been conducted to adapt such base LLMs to the biomedical domain in bilingual and multilingual settings (Singhal et al., 2022; Li et al., 2023; Singhal et al., 2023; Chen et al., 2023). Such LLMs are reported to perform better than the base model on downstream tasks in the target domains and languages.

However, CPT from a base model to endow non-English capability or to adapt to specific domains comes with the issue of degradation of the capabilities of the base model (Scialom et al., 2022; Fujii et al., 2024; Ankit Pal, 2024). Although many previous studies have shown that the incorporation of training datasets that the base model used during CPT significantly mitigates this forgetting (Rolnick et al., 2019; Chen et al., 2023; Lewkowycz et al., 2022; Yano et al., 2025), further analysis is needed to quantify these effects because such training datasets might be inaccessible and private and to determine whether these methods will be valid for a wide range of tasks.

In this work, we conducted experiments on three NLP tasks that were not primarily targeted during CPT. Specifically, we used Llama3-ELAINE-medLLM (Yano et al., 2025), which was adapted from Llama3-8B to the biomedical domain and has trilingual ability, including English (EN), Japanese (JA), and Chinese (ZH). Llama3-ELAINE is a pre-trained model without fine-tuning with instruction datasets.

For the emergent NLP tasks, we selected named entity recognition (NER) and machine translation (MT) tasks in a domain similar to the biomedical domain, and a summarization (SUM) task in the general domain, which were not targeted during CPT. Our experiments found that compared with Llama3-8B, Llama3-ELAINE due to CPT shows some forgetting phenomena that affect the model's promptability even in new tasks in similar domains where the model was trained during CPT.

However, our results also show that after fine-tuning Llama3-ELAINE-medLLM on the same downstream task, the model performs competitively or better than the base model. These results indicate that even though the adapted models' few-shot prompt ability may degrade in an emerging task, even in the relevant domains, the model will perform better after fine-tuning, as it has acquired more in-depth domain knowledge than the base model.

Proceedings of the 24th Workshop on Biomedical Language Processing (BioNLP 2025), pages 18–26

Model	EN	JA	ZH
Llama3-8B	61.68	25.83	45.47
Llama3-ELAINE	59.56 (-2.1)	31.96 (+6.1)	52.25 (+6.8)

Table 1: Comparison of average scores of medical QA benchmarks in English, Japanese, and Chinese between ELAINE-medLLM and the base Llama3-8B

	Training	Validation	Testing	lang
BC5CDR	500	500	500	EN
MedTxt-CR	128	10	10	JA
CMeEE-V2	19,600	400	400	ZH

Table 2: Statistics of NER datasets (# of documents). BC5CDR (en), MedTxt-CR (ja), and CMeEE-V2 (zh)

	Train	Validation	Test
JA-EN	1,000,000	1,790	1,812
JA-ZH	672,315	2,090	2,107

Table 3: Statistics of ASPEC parallel corpora (# of sentence pairs)

2 Related work

Numerous medLLMs (Singhal et al., 2022; Li et al., 2023; Singhal et al., 2023; Chen et al., 2023) have been proposed using CPT, adapted from open-source LLMs such as Llama (Touvron et al., 2023a,b). However, CPT can potentially degrade few-shot learning performance, hindering its ability to adapt to new tasks quickly. There have been many studies to prevent this issue, such as replaying pre-trained data and careful selection of the training dataset during CPT (Chen et al., 2023; Lewkowycz et al., 2022; Yano et al., 2025). The negative impact of CPT can be addressed in post-processing, such as task-specific pre-training, which involves further fine-tuning the pre-trained model on a small dataset related to the target few-shot task (Ke et al., 2022). Prompt engineering is another solution, involving the design of prompts during fine-tuning to guide the model toward the desired task with few-shot examples (Radford et al., 2019).

3 Experiments

To evaluate the effect of multilingual and domain-adapted continual pre-training on few-shot prompt-ability for NLP tasks, we used Llama3-ELAINE-medLLM (Yano et al., 2025), which was continually pre-trained without instruction fine-tuning on Llama3-8B, targeted for the biomedical domain, and adapted for multiple languages (English, Japanese, and Chinese). Table 1 shows the average scores on several medical QA benchmarks in English, Japanese, and Chinese. We can see that Llama3-ELAINE-medLLM shows much better in-context learning (ICL) capabilities for medical QA tasks than Llama3-8B for Japanese and Chinese while slightly sacrificing English capability.

In this work, we used named entity recognition (NER) and machine translation (MT) tasks related to the biomedical domain and a summarization task in the general domain as the emergent tasks for our experiments.

3.1 Datasets

3.1.1 NER dataset

We used BC5CDR (Li et al., 2016) for the English NER dataset, which defines "Disease" and "Chemical" entities. For the Japanese NER dataset, we used MedTxt-CR (Yada et al., 2022), which annotates various medical expression entities such as "disease", "anatomical part", etc. This experiment only used the "disease/symptoms" entity labeled as *d* in the corpus. Note that their annotation method does not delineate adjacent entity mentions, such as 呼吸困難、黄疸、下腿浮腫(dyspnea, jaundice, leg edema), which were labeled as one single, continuous entity rather than three independent entities as seen in other corpora.

For the Chinese NER dataset, we used CMeEE-V2 (Du et al., 2024), which annotates nine medical entity types, such as "disease", "clinical manifestations", "drugs", etc. This work used only disease and clinical symptoms labeled as "*dis*" and "*sim*" in the corpus, respectively. Table 2 summarizes the number of samples (documents) for each split of the corpus. Note that we randomly split the training datasets for the Japanese and Chinese datasets.

3.1.2 MT dataset

We used ASPEC (Nakazawa et al., 2016), consisting of two corpora from scientific paper abstracts: Japanese-English and Japanese-Chinese parallel corpora. Table 3 summarizes the number of samples (sentence pairs) for each split of the corpus. We used a four-way language pair for evaluation by reversing the source and target languages.

3.1.3 Summarization dataset

We used XLSum (Hasan et al., 2021), a diverse dataset of professionally annotated news article

	Train	Validation	Test
EN	306,522	11,535	11,535
JA	7,113	889	889
ZH	37,362	4,670	4,670

Table 4: Statistics of XLSum summarization dataset (# of text and summarization pairs) for English, Japanese, and Chinese

summary pairs from BBC that cover 45 languages. We used the English, Japanese, and Chinese splits of the dataset for evaluation. Table 4 summarizes the number of samples (text and summarization pairs) for each language dataset.

3.2 Evaluation

For each task, we evaluate the performance of ELAINE-medLLM and Llama-8B in the zero- or few-shot and fine-tuning settings. A sample of the instruction format for the training dataset for each task is described in Appendix A. The details of the settings are as follows.

3.2.1 Zero or few-shot settings

We used in-context learning (ICL) to evaluate each task's performance in the zero- or few-shot settings. For the few-shot settings, we evaluated one-shot, three-shot, five-shot, and ten-shot scenarios. ICL samples were selected from the training split, with the top N most similar to the input, where N is the number of few-shot samples. We used the text embeddings calculated by SentenceTransformer (Reimers and Gurevych, 2019) to compute similarity.

3.2.2 Fine-tuning settings

For each task, we fine-tuned the model by using the training split of each dataset. We used full-parameter tuning using DeepSpeed stage-3 and trained the model for 6, 3, and 3 epochs for NER, MT, and Summarization, respectively. We used the following training parameters:

- per_device_batch_size: 6
- gradient_accumulation_steps: 2
- learning_rate: 1e-6
- weight_decay: 0.001
- warmup_ratio: 0.1
- lr_scheduler_type: cosine

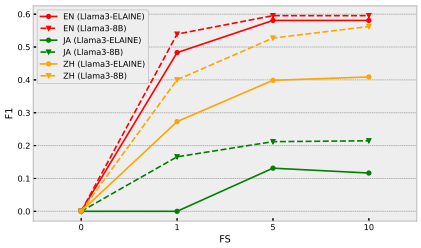

Figure 1: NER: few-shots performance in F1 (EN: BC5CDR, JA: MedTxt-CR, ZH: CMeEE-V2

		Precision	Recall	F1
EN	Llama3-ELAINE	0.825	0.802	0.813
	Llama3-8B	**0.833**	**0.831**	**0.832**
JA	Llama3-ELAINE	0.678	**0.701**	**0.689**
	Llama3-8B	**0.682**	0.667	0.674
ZH	Llama3-ELAINE	**0.766**	**0.792**	**0.779**
	Llama3-8B	0.764	0.789	0.776

Table 5: NER: fine-tuning performance (EN: BC5CDR, JA: MedTxt-CR, ZH: CMeEE-V2

4 Results

4.1 Named entity recognition (NER)

We adopt the TANL format (Paolini et al., 2021) to solve NER by LLM. In this format, the input text is copied to the output by annotating entity names and enclosing them in brackets by suffixing the detected entity type (see Appendix A). Figures 1 show the performance of language-dependent NER tasks in few-shot and Table 5 shows the performance of these NER tasks under the fine-tuning settings. The scores were computed by converting from TANL to IOB format (Ramshaw and Marcus, 1995). During conversion, we regulated the output by removing all parts that did not conform to our defined format, which made the zero-shot scores zero in all cases.

For few-shot settings, Llama performs better than Llama3-ELAINE-medLLM in all cases. This indicates the adverse effects of continual pre-training on the promptability of the base model. However, in fine-tuning settings, Llama3-ELAINE performs competitively with LLama in Japanese and Chinese. This result suggests that the degradation of promptability by CPT may not reveal the model's latent performance when the same task is fine-tuned.

4.2 Machine translation (MT)

Figures 2 and 3 show the few-shot performance of the MT task between Japanese and English and

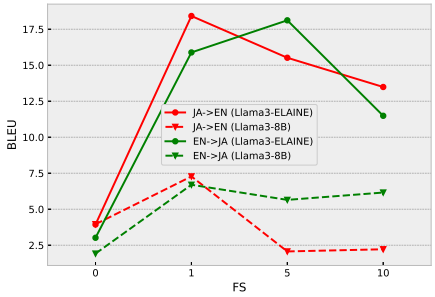

Figure 2: Machine Translation: few-shots (JA→EN, EN→JA) performance in BLEU (ASPEC)

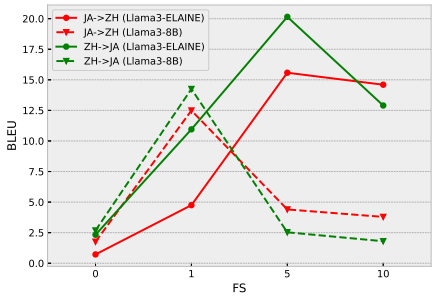

Figure 3: Machine Translation: few-shots (JA→ZH, ZH→JA) performance in BLEU (ASPEC)

	JA→EN	EN→JA
Llama3-ELAINE	**28.10**	**45.20**
Llama3-8B	27.92	44.36
	JA→ZH	ZH→JA
Llama3-ELAINE	34.25	**49.55**
Llama3-8B	**34.28**	48.67

Table 6: Machine Translation: fine-tuning performance in BLEU (ASPEC)

Japanese and Chinese, and Table 6 shows the fine-tuning performance of the same MT task measured in BLEU (Papineni et al., 2002). Unlike the performance of NER tasks, the performance of MT tasks, both in few-shot and fine-tuning, shows that ELAINE-medLLM is similar or superior to Llama3-8B.

This result indicates that continual pre-training does not always hurt the promptability of the base model for NLP tasks. We hypothesize that the degree of the effect depends on the novelty of the new task and its affinity to the training datasets used during CPT. Since ELAINE-medLLM is trained to harness multilingual ability, it works effectively in MT tasks for the same languages. On the other hand, although the domains of the previous NER tasks are highly aligned to those of the training

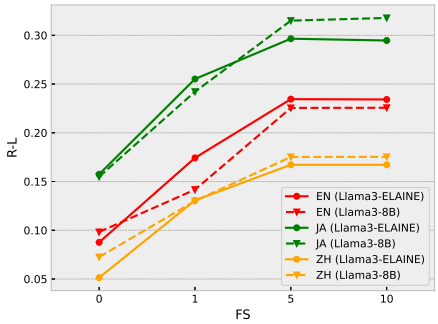

Figure 4: Summarization: few-shot performance in Rouge-L (RL) (XLSum)

		R-1	R-2	R-L
EN	Llama3-ELAINE	0.418	0.192	0.349
	Llama3-8B	**0.421**	**0.194**	**0.352**
JA	Llama3-ELAINE	**0.570**	**0.286**	**0.454**
	Llama3-8B	0.564	0.282	0.450
ZH	Llama3-ELAINE	0.368	0.171	0.319
	Llama3-8B	**0.371**	**0.173**	**0.322**

Table 7: Summarization: fine-tuning performance in Rouge-1 (R-1), Rouge-2 (R-2), Rouge-L (R-L) (XL-Sum)

dataset for ELAINE-medLLM, we assume that the novelty of the TANL output format affects its performance in the few-shot setting.

4.3 Summarization (SUM)

Figure 4 shows the results of the summarization task in few-shot settings measured in ROUGE-L (Lin, 2004). Table 7 shows the performance of the same summarization task in the fine-tuning setting in ROUGE-1, ROUGE-2, and ROUGE-L. Unlike previous NER and MT tasks, the SUM task is in the general domain for each of the three languages.

Unlike the previous two tasks (NER, MT), which can be considered related to the biomedical field, we could not observe noticeable performance differences in the fine-tuning setting. This is probably because the summarization task is in the general domain. We assume that CPT targeted for the biomedical domain does not affect fine-tuning performance in the general domain, though it shows a slight advantage for Llama3-8B for the few-shot setting.

	Precision	Recall	F1
Meditron-7B	**0.824**	0.744	0.783
Llama2-7B	0.808	**0.774**	**0.791**

Table 8: NER: fine-tuning performance (BC5CDR)

	R-1	R-2	R-L
Meditron-7B	**0.402**	**0.182**	**0.334**
Llama2-7B	0.397	0.172	0.330

Table 9: Summarization: fine-tuning performance in Rouge-1 (R-1), Rouge-2 (R-2), Rouge-L (R-L) (XL-Sum)

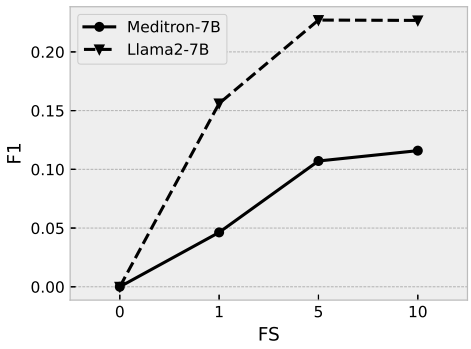

Figure 5: NER: few-shots performance in F1 (BC5CDR)

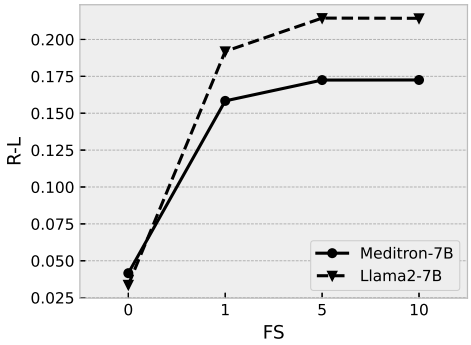

Figure 6: Summarization: few-shots performance in Rouge-1 (R-1) (XLSum)

5 Analysis

This section analyzes whether the phenomenon we found in the previous experiments can be observed in a different experimental setting.

5.1 Experimental Setting

We use Meditron-7B (Chen et al., 2023), an English medical LLM adapted from Llama2-7B (Touvron et al., 2023c), as the baseline. We selected the monolingual model because we aim to remove the effects of multilingualism on the results. For this experiment, we evaluate performance in few-shot and fine-tuning settings using the same NER task using BC5CDR and SUM task using English XL-Sum as in the previous experiments.

5.2 Results

Fig. 5 and Table 8 show the few-shot and fine-tuning NER results using BC5CDR. These results indicate that domain-adapted training does not benefit the performance of few-shot and fine-tuning results. Especially, Meditron-7B lags far behind Llama2-7B in a few-shot setting. Fig. 6 and Table 9 show the few-shot and fine-tuning Summarization results for English XLSum. In the few-shot setting, Meditron-7B lags far behind Llama2-7B, as in the NER task. However, the model shows competence against the baseline model in the fine-tuning setting.

These results confirm that the performance of the few-shot setting does not always show the model's potential in the fine-tuning setting of the same task. Nonetheless, we do not observe a similar trend in the comparative results between the domain-adapted and the base models. To summarize, domain adaptation works negatively for some tasks that do not depend clearly on acquired domain knowledge in few-shot settings, such as NER, and out-of-domain tasks, such as summarization. However, this degradation does not necessarily indicate the model's potential in fine-tuning settings.

6 Conclusion

CPT can help pre-trained large language models (LLMs) effectively adapt to new, under-trained domains or low-resourced languages without requiring retraining from scratch. Nevertheless, during CPT, the model's few-shot transfer ability is affected for emergent tasks. This also applies to new tasks, even in the relevant domains targeted during CPT. However, our experimental results show that degradation in few-shot transfer ability does not necessarily indicate the model's underlying potential in the same downstream task after fine-tuning. In our experiments, we observe that ELAINE-medLLM, which is adapted to the biomedical domain and endowed with trilingual ability (English, Japanese, and Chinese) by CPT from Llama3-8B, performs competitively with or better than the base model in all emergent tasks after fine-tuning, even though it shows some degradation in some few-shot settings.

Limitations

The prompt inputs used for few-shot evaluations were not optimized, suggesting that an optimal prompt might produce better results, such as prompt tuning or adopting a chain of thoughts. In this work, we only conducted performance analysis of Llama3-ELAINE and Meditron, adapted to the biomedical domain, against their base LLMs on three NLP tasks (NER, machine translation, and summarization). Hence, further experiments will be desired to evaluate the results we found in this study.

Acknowledgement

This paper is based on the results of project JPNP20006, which was commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

References

Malaikannan Sankarasubbu Ankit Pal. 2024. Openbiollms: Advancing open-source large language models for healthcare and life sciences.

Zeming Chen, Alejandro Hernández-Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. Meditron-70b: Scaling medical pretraining for large language models. *Preprint*, arXiv:2311.16079.

Xiaojing Du, Hanjie Zhao, Danyan Xing, Yuxiang Jia, and Hongying Zan. 2024. Mrc-based nested medical ner with co-prediction and adaptive pre-training. *Preprint*, arXiv:2403.15800.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. *Preprint*, arXiv:2404.17790.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Zixuan Ke, Haowei Lin, Yijia Shao, Hu Xu, Lei Shu, and Bing Liu. 2022. Continual training of language models for few-shot learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10205–10216, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Venkatesh Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. Solving quantitative reasoning problems with language models. *ArXiv*, abs/2206.14858.

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*, 2016:baw068.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Preprint*, arXiv:2303.14070.

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.

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. Aspec: Asian scientific paper excerpt corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2204–2208, Portorož, Slovenia. European Language Resources Association (ELRA).

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. *Preprint*, arXiv:2101.05779.

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

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy P. Lillicrap, and Greg Wayne. 2019. Experience replay for continual learning. *Preprint*, arXiv:1811.11682.

Thomas Scialom, Tuhin Chakrabarty, and Smaranda Muresan. 2022. Fine-tuned language models are continual learners. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6107–6122, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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. *Preprint*, arXiv:2212.13138.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards expert-level medical question answering with large language models. *Preprint*, arXiv:2305.09617.

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. 2023a. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023c. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Shuntaro Yada, Yuta Nakamura, Shoko Wakamiya, and Eiji Aramaki. 2022. Real-mednlp: Overview of real document-based medical natural language processing task. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-16)*, pages 285–296.

Ken Yano, Zheheng Luo, Jimin Huang, Qianqian Xie, Masaki Asada, Chenhan Yuan, Kailai Yang, Makoto Miwa, Sophia Ananiadou, and Jun'ichi Tsujii. 2025. ELAINE-medLLM: Lightweight English Japanese Chinese trilingual large language model for biomedical domain. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4670–4688, Abu Dhabi, UAE. Association for Computational Linguistics.

A Instruction format for tasks

A.1 Name entity recognition (NER)

We adopt the TANL format (Paolini et al., 2021) to solve NER by LLM. In this format, the LLM is expected to copy the source context to the target by annotating detected entity mentions, enclosing them in brackets, and suffixing them with the detected entity type. The descriptions of valid entity types and their labels are specified after the task description. The following shows a sample instruction format for the NER task in the case of BC5CDR (Li et al., 2016).

Instruction for NER task

##Task##
You are an AI language model trained to extract entities from sentences. See all possible labels and their descriptions below.
###
Description: Disease name
Label: Disease
###
Description: Chemical name
Label: Chemical
###
<|eot_id|>##Text##
Famotidine-associated delirium. A series of six cases.
Famotidine is a histamine H2-receptor antagonist used in inpatient settings for prevention of stress ulcers and is showing increasing popularity because of its low cost. Although all of the currently available H2-receptor antagonists have shown the propensity to cause delirium, only two previously reported c ases have been associated with famotidine. The authors report on six cases of famotidine-associated delirium in hospitalized pa tients who cleared completely upon removal of famotidine. The pharmacokinetics of famotidine are reviewed, with no change in its metabolism in the elderly population seen. The implications of using famotidine in elderly persons are discussed.
<|eot_id|>##Result##
[Famotidine|Chemical]-associated [delirium|Disease]. A series of six cases.
[Famotidine|Chemical] is a histami ne H2-receptor antagonist used in inpatient settings for prevention of stress [ulcers|Disease] and is showing increasing popula rity because of its low cost. Although all of the currently available H2-receptor antagonists have shown the propensity to caus e [delirium|Disease], only two previously reported cases have been associated with [famotidine|Chemical]. The authors report on six cases of [famotidine|Chemical]-associated [delirium|Disease] in hospitalized patients who cleared completely upon removal of [famotidine|Chemical]. The pharmacokinetics of [famotidine|Chemical] are reviewed, with no change in its metabolism in the e lderly population seen. The implications of using [famotidine|Chemical] in elderly persons are discussed. <|eot_id|>

A.2 Machine translation (MT)

The following shows a sample instruction format for the MT task in the case of ASPEC (Nakazawa et al., 2016).

Instruction for MT task

##Task##
Translate the following sentences from English to Japanese.
<|eot_id|>##Source##
Lattice images were not obtained by high resolution power image observation.
<|eot_id|>##Target##
高分解能像観察での格子像は得られなかった。<|eot_id|>

A.3 Summarization (SUM)

The following shows a sample instruction format for the SUM task in the case of XLSum (Hasan et al., 2021).

Instruction for SUM task

##Task##
Summarize the following paragraph.
<|eot_id|>##Source##
Lewis Williams, 20, died on 11 January from a shotgun wound suffered in Wath Road, Mexborough. South Yorkshire Police said two men aged 20 and 49 were arrested on Friday in connection with his death, bringing the total number of arrests to eight. Two boys, aged 16 and 17, have been charged with Mr Williams' murder and are next due to appear in court on 1 February. Police said one of the men arrested on Friday, a 20-year-old from Barnsley, was arrested on suspicion of murder, while a 49-year-old man from Doncaster was arrested on suspicion of assisting an offender and possession of ammunition. Both are being held in police custody. Four other men, aged between 20 and 32, who have been arrested in connection with Mr Williams'death have been released on bail. Follow BBC Yorkshire on Facebook, Twitter and Instagram. Send your story ideas to yorkslincs.news@bbc.co.uk or send video here.
<|eot_id|>##Target##
Two more people have been arrested in connection with a fatal shooting. <|eot_id|>