A Zero-shot and Few-shot Study of Instruction-Finetuned Large Language Models Applied to Clinical and Biomedical Tasks

Yanis Labrak, Mickael Rouvier and Richard Dufour

COLING 2024 Short Paper

Abstract

- LLM의 출현으로 medical domain 데이터의 효율적인 처리가 가능해지고, 다양한 작업에서 우수한 성능을 보임

- 13개의 실제 임상/생물의학 NLP task 세트에 대해 Instruction-Finetuned 된 LLM의 성능을 평가

- 결과적으로는 LLM이 대부분의 task에서 의료 전용 모델과 비슷한 성능을 보였고, 특히 QA Task에서 두드러진 성과를 보임

- 하지만 분류/RE(Related Extraction) task에서는 의료 분야 맞춤형 모델인 PubMedBERT 의 성능에는 미치지 못함

Dataset

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.

Used Model

- TK-Instruct – T5 encoder-decoder based model, Flan-T5-XL 모델과 비교하였을 때 QA 작업에서 더 좋은 성능을 보였기 때문에 사용
- ChatGPT – GPT 3.5 Turbo based model
- Standard Alpaca – LLaMA with 7B parameter based model
- PubMedBERT – 생물의학 전용 BERT based model. PubMed 말뭉치의 31억 단어를 기반으로 훈련됨

Result

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`.

느낀 점

- 간단한 논문이지만 특정 도메인 / task 에서는 LLM과 LLM이 아닌, 모델을 비교하면서 논문을 적어나가는 방법을 배움

- 현재 연구 중인 부분에서 데이터셋을 평가하는 부분이 빈약했는데, 이를 평가하기 위한 여러가지 방법들을 알 수 있었음

Thank you