

• Supplementary File •

OpenBA: An Open-Sourced 15B Bilingual Asymmetric Seq2Seq Model Pre-trained from Scratch

Juntao Li^{*}, Zecheng Tang[†], Yuyang Ding[†], Pinzheng Wang[†],
Pei Guo, Wangjie You, Dan Qiao, Chenyu Wang, Wenliang Chen,
Guohong Fu, Qiaoming Zhu, Guodong Zhou[‡], Min Zhang[‡]

Soochow University, Suzhou 215006, China

Appendix A Instruction Template

The task instruction prompts for evaluation are provided here:

Test prompt example for MMLU:

Context:

(Instruction)

Please answer the following multiple choice question:

(Exemplar)

Question: Which of the following occurred first during the separation of the elements of Pangaea through continental drift? Options: A. Gondwana and Laurasia were formed. B. Africa separated from South America. C. India collided with Eurasia to form the Himalayan mountain chain. D. Australia separated from the rest of the continental landmasses. Answer: A

...(Other exemplars, if any)

(Test case)

Question: Experiments on song development in birds have shown that when a young male reared in isolation hears only the song of a different bird species, he will develop an adult song repertoire that lacks certain characteristics typical of his own species. This result shows that the song of his species is most likely Options: A. entirely learned during development B. entirely instinctive C. both instinctive and learned D. dependent upon hormones for proper development Answer:

Response: A

Test prompt example for CMMLU:

Context:

(Instruction)

以下是关于(大学教育学)的单项选择题，请直接给出正确答案的选项。

(Exemplar)

题目：在古代文献记载中，我国西周时期设在王都的小学 and 大学，总称为()

A. 都学 B. 乡学 C. 官学 D. 国学

答案是：D

...(Other exemplars, if any)

(Test case)

以下是关于(大学教育学)的单项选择题，请直接给出正确答案的选项。

题目：教育的本质特征是()

A. 系统性 B. 知识性 C. 科学性 D. 育人性

答案是：

Response: D

^{*}denotes Project Leader, [†] denotes Equal Contribution, [‡] denotes Corresponding Author

Email: {ljt,gdzhou,minzhang}@suda.edu.cn, {zctang,yyding23,pzwang1}@stu.suda.edu.cn

Test prompt example for C-Eval:

Context:

(Instruction)

以下是关于(中国语言文学)的单项选择题，请直接给出正确答案的选项。

(Exemplar)

题目：元朝政府曾经实行残酷的民族政策，把全国人民分为____四个等级。

A. 色目人、蒙古人、汉人、南人 B. 蒙古人、汉人、南人、色目人 C. 蒙古人、南人、色目人、汉人 D. 蒙古人、色目人、汉人、南人
答案是：D

...(Other exemplars, if any)

(Test case)

以下是关于(中国语言文学)的单项选择题，请直接给出正确答案的选项。

题目：《国语》和____，都是国别史。

A. 《左传》 B. 《战国策》 C. 《史记》 D. 《汉书》

答案是：

Response: D

Test prompt example for BBH:

Context:

(Exemplar)

not (True) and (True) is

Answer: False

...(Other exemplars, if any)

False or not not False and True is

Response: True

Test prompt example for $En \Rightarrow Zh$ Machine Translation:

Context:

将以下中文翻译成英文，并输出英文翻译：

Local authorities are warning residents in the vicinity of the plant to stay indoors, turn off air-conditioners and not to drink tap water.

Response:

当地政府警告核电站附近的居民，要待在室内，关掉空调，不要喝自来水。

Test prompt example for $Zh \Rightarrow En$ Machine Translation:

Context:

将以下英文翻译成中文，并输出中文翻译：

当地政府警告核电站附近的居民，要待在室内，关掉空调，不要喝自来水。

Response:

Local government warns residents near nuclear power plant to stay indoors, turn off air conditioning, and do not drink bottled water.

Test prompt example for BoolQ:

Context:

Parity (mathematics) – In mathematics, parity is the property of an integer's inclusion in one of two categories: even or odd. An integer is even if it is evenly divisible by two and odd if it is not even. For example, 6 is even because there is no remainder when dividing it by 2. By contrast, 3, 5, 7, 21 leave a remainder of 1 when divided by 2. Examples of even numbers include -4, 0, 82 and 178. In particular, zero is an even number. Some examples of odd numbers are -5, 3, 29, and 73.

question: can an odd number be divided by an even number?

answer:

Response: yes

Test prompt example for RTE:

Context:

Yet, we now are discovering that antibiotics are losing their effectiveness against illness. Disease-causing bacteria are mutating faster than we can come up with new antibiotics to fight the new variations.

Can we say the following?

Bacteria is winning the war against antibiotics.

OPTIONS: - yes - no

Response: yes

Test prompt example for ReCoRD:

Context:

Tracy Morgan hasn't appeared on stage since the devastating New Jersey crash that nearly ended his life last summer, but all that will change this fall when he returns to host Saturday Night Live. NBC announced on Twitter Monday that Morgan, an SNL alum with seven seasons as a cast member under his belt, will headline the third episode of Season 41 airing October 17. For Morgan, 46, it will be a second time hosting the long-running variety show, the first since the June 2014 pileup on the New Jersey Turnpike that killed his friend and mentor James 'Jimmy Mack' McNair.

@highlight

Morgan, 46, will host third episode of season 41 of SNL airing October 17

@highlight

He tweeted to his fans: 'Stoked to be going home...#SNL'

@highlight

For the SNL alum who had spent seven years as cast member, it will be a second time hosting the show

@highlight

Morgan has been sidelined by severe head trauma suffered in deadly June 2014 crash on New Jersey Turnpike that killed his friend

@highlight

First episode of new SNL season will be hosted by Miley Cyrus, followed by Amy Schumer

Response:

On October 10, acclaimed comedian and star of the summer box office hit Trainwreck Amy Schumer will make her SNL debut, followed by Morgan a week later.

Test prompt example for WSC:

Context:

Bernard , who had not told the government official that he was less than 21 when he filed for a homestead claim, did not consider that he had done anything dishonest. Still, anyone who knew that he was 19 years old could take his claim away from him . "him" refer to what?

Response: anyone

Test prompt example for WiC:

Context:

An emerging professional class.

Apologizing for losing your temper, even though you were badly provoked, showed real class. The word 'class' used in the same way in the two sentences above, True or False?

answer:

Response: False

Test prompt example for CoPA:

Context:

The man turned on the faucet, so ...

A. The toilet filled with water. B. Water flowed from the spout.

answer:

Response: B

Test prompt example for MultiRC:

Context:

Please read the following article and judge whether the answer to the question is correct: What causes a change in motion? The application of a force. Any time an object changes motion, a force has been applied. In what ways can this happen? Force can cause an object at rest to start moving. Forces can cause objects to speed up or slow down. Forces can cause a moving object to stop. Forces can also cause a change in direction. In short, forces cause changes in motion. The moving object may change its speed, its direction, or both. We know that changes in motion require a force. We know that the size of the force determines the change in motion. How much an objects motion changes when a force is applied depends on two things. It depends on the strength of the force. It also depends on the objects mass. Think about some simple tasks you may regularly do. You may pick up a baseball. This requires only a very small force.

questions: Would the mass of a baseball affect how much force you have to use to pick it up?

answer: No.

Is this answer True or False?

Response: False

Test prompt example for AX_b :

Context:

Read the sentence below and answer the question: The cat sat on the mat.

Question: The cat did not sit on the mat. True or False?

Answer:

Response: False

Test prompt example for AX_g

Context:

Read the sentence below and answer the question: The taxpayer met with the accountant to get help filing his taxes.

Question: The accountant sought help filing taxes. True or False?

Answer:

Response: False

Test prompt example for BELEBELE:

Context:

Please read the following article and answer the questions:

Make sure your hand is as relaxed as possible while still hitting all the notes correctly - also try not to make much extraneous motion with your fingers. This way, you will tire yourself out as little as possible. Remember there's no need to hit the keys with a lot of force for extra volume like on the piano. On the accordion, to get extra volume, you use the bellows with more pressure or speed.

Please answer the following multiple-choice questions:

According to the passage, what would not be considered an accurate tip for successfully playing the accordion? Options: (A) For additional volume, increase the force with which you hit the keys (B) Keep unnecessary movement to a minimum in order to preserve your stamina (C) Be mindful of hitting the notes while maintaining a relaxed hand (D) Increase the speed with which you operate the bellows to achieve extra volume Answer: (

Response: B

Test prompt example for CLTS:

Context:

简要概括以下文本:

温布尔登网球锦标赛官方微博发布消息称,全英草地网球俱乐部今天非常遗憾地决定,由于冠状病毒疫情持续蔓延产生的公共卫生安全问题,将取消2020年温布尔登网球锦标赛,第134届锦标赛将于2021年6月28日至7月11日举行。(原题为《2020年温网锦标赛取消将于2021年6月至7月举行》)(本文来自澎湃新闻,更多原创资讯请下载“澎湃新闻”APP)

概括为:

Response:

由于冠状病毒疫情持续蔓延产生的公共卫生安全问题,将取消2020年温布尔登网球锦标赛,第134届锦标赛将于2021年6月28日至7月11日举行。

Test prompt example for OpenBA-Chat:

Context:

Human: 你叫什么名字啊?

Assistant: 我叫OpenBA,是一款由苏州大学的自然语言处理实验室开发的大语言模型。

Human: 你是GPT架构的模型吗?

Assistant: 我不基于GPT架构开发,而是基于浅编码器-深解码器的架构开发的模型。

Human: 你会哪些语言?

Assistant:

Response:

在预训练及微调数据中,我学习了中英两种语言。正如我的名字OpenBA,我是一个双语(Bilingual)的模型。

Test prompt example for OpenBA-InstructGen:

Context:

Please generate the instruction according to the context I provide:

Nice to meet you! I am an AI assistant, how can I help you today?

Response:

Hello, nice to meet you!

Test prompt example for OpenBA-Tool:

Context:

Who wrote i got a brand new pair of rollerskates?

Response:

WikiSearch(“Brand New Key”) → People will take it any way they want to take it.” In a 2013 interview with music journalist Ray Shasho, Melanie revealed the true origin of “Brand New Key”: Of course I can see it symbolically with the key, but I just thought of roller skating. I was fasting with a twenty seven day fast on water. I broke the fast and went back to my life living in New Jersey and we were going to a flea market around six in the morning. On the way back ...and I had just broken the fast, from the flea market, we]Melanie

Appendix B Chinese Flan Collection

Example of Unified Json Format:

```
{
  "instruction": "请将输入的英语句子翻译成中文",
  "input": "Yet not nearly enough has been invested in this effort.",
  "output": "但目前这方面的投入还远远不够。",
  "domain": "多领域",
  "task": "机器翻译"
}
```

Figure B1 An example of the unified format of Chinese Flan Dataset.

As shown in Fig. B1, the unified format of each data includes “instruction”, “input”, “output”, “domain” and “task”, where “instruction” denotes the description of the task that provides LLM with a clear purpose. “input” and “output” are the question and answer respectively. “domain” is the topic of the task, such as medicine, news, etc. “task” indicates the type categorized into one of eighteen task types. Table B1 shows all the tasks and the source of the instruction datasets in each task.

Task	Source	Dataset
Question Answering	[10]	XQA
	[5]	ChineseDBQA
	[1]	Xquad
	https://www.luge.ai/#/luge/dataDetail?id=40	ChineseBiomedicalQA
Text Classification	[3]	XNLI
	http://tcci.ccf.org.cn/conference/2014/dldoc/evtestdata6.zip	Chinese News Categorization
	https://tianchi.aliyun.com/dataset/133838?spm=a2c22.28136470.0.0.6e5a6a23SPZMrX&from=search-list	TNEWS
	https://huggingface.co/datasets/dirtycomputer/ChnSentiCorp.html https://storage.googleapis.com/cluebenchmark/tasks/iflytek_public.zip	ChnSentiCorp.html iflytek
Sentiment Classification	[22]	FewCLUE EPRSTMT
	https://www.luge.ai/#/luge/dataDetail?id=25	ChnSentiCorp
	https://www.heywhale.com/mw/dataset/5e09a9eb2823a10036b126c0/file https://www.luge.ai/#/luge/dataDetail?id=20	BDCI 2019 NLPCC14-SC
Named Entity Recognition	https://huggingface.co/datasets/msra_ner	MSRA_NER
	https://storage.googleapis.com/cluebenchmark/tasks/cluener_public.zip	CLUE Fine-Grain NER
Text Matching	[21]	CLUE WSC 2020
	[21]	CMNLI
	[8]	OCNLI
	https://www.luge.ai/#/luge/dataDetail?id=39 https://tianchi.aliyun.com/dataset/106411	CINLID AFQMC
Text Summarization	[12]	CSL
	[7] http://tcci.ccf.org.cn/conferen	WikiLingua Chinese Weibo Oriented Chinese News Summarization NLPCC2015
Reading Comprehension	[21]	C3
	[4]	CMRC2018
	[14]	DRCD
Question Generation	[17]	DuReader.QG
	https://tianchi.aliyun.com/dataset/dataDetail?dataId=86895	TCM Literature Question Generation
Dialogue	[19]	douban
	[24]	kdconv
	https://www.luge.ai/#/luge/dataDetail?id=38	Chinese Persona Chat
Machine Translation	[6]	mttt
	[20]	translation 2019 zh
	[2] https://www.kaggle.com/datasets/garyongguanjie/wikititles-zhen	WMT19 en-zh wikititles_en-zh
Cloze Test	[23]	ChiD
Text Generation	[15]	AdvertiseGen
Semantic Analysis	[18]	Math23K
Relation Extraction	[11]	DuIE2.0
Grammatical Error Correction	[13]	MD-SCS
Fact-checking	[9]	CHEF
Interpretable Evaluation	[16]	DuExplain
Event Extraction	https://tianchi.aliyun.com/dataset/dataDetail?dataId=110904	tianchi_event_doclevel_attr

Table B1 All types of tasks and the source of the instruction datasets in each task.

References

- 1 Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. *CoRR*, abs/1910.11856, 2019.
- 2 Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation*, volume 2, pages 1–61. Association for Computational Linguistics, 2019.
- 3 Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018.
- 4 Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. A span-extraction dataset for Chinese machine reading comprehension. 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 5886–5891. Association for Computational Linguistics, 2019.
- 5 Nan Duan. Overview of the nlpcc-iccpol 2016 shared task: Open domain chinese question answering. In *NLPCC/ICCPOL*, 2016.
- 6 Kevin Duh. The multitarget ted talks task. <http://www.cs.jhu.edu/~kevinduh/a/multitarget-tedtalks/>, 2018.
- 7 Claire Cardie Faisal Ladhak, Esin Durmus and Kathleen McKeown. Wikilingua: A new benchmark dataset for multilingual abstractive summarization. In *Findings of EMNLP, 2020*, 2020.
- 8 Hai Hu, Kyle Richardson, Liang Xu, Lu Li, Sandra Kübler, and Lawrence Moss. Ocnli: Original chinese natural language inference. In *Findings of the Association for Computational Linguistics*, page 3512–3526. Association for Computational Linguistics, 2020.
- 9 Xuming Hu, Zhijiang Guo, GuanYu Wu, Aiwei Liu, Lijie Wen, and Philip Yu. Chef: A pilot chinese dataset for evidence-based fact-checking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, page 3362–3376. Association for Computational Linguistics, 2022.
- 10 Zhiyuan Liu Jiahua Liu, Yankai Lin and Maosong Sun. Xqa: A cross-lingual open-domain question answering dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, page 2358–2368, 2019.
- 11 Shuangjie Li, Wei He, Yabing Shi, Wenbin Jiang, Haijin Liang, Ye jiang, Yang Zhang, Yajuan Lyu, and Yong Zhu. Duie: A large-scale chinese dataset for information extraction. In *Tang, J., Kan, MY., Zhao, D., Li, S., Zan, H. (eds) Natural Language Processing and Chinese Computing*, volume 11839. Springer, Cham, 2019.
- 12 Yudong Li, Yuqing Zhang, Zhe Zhao, Linlin Shen, Weijie Liu, Weiquan Mao, and Hui Zhang. Csl: A large-scale chinese scientific literature dataset. In *Proceedings of the 29th International Conference on Computational Linguistics*, page 3917–3923. International Committee on Computational Linguistics, 2022.
- 13 Qi Lv, Ziqiang Cao, Lei Geng, Chunhui Ai, Xu Yan, and Guohong Fu. General and domain adaptive chinese spelling check with error consistent pretraining. In *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* Association for Computing Machinery, 2022.
- 14 Chih Chieh Shao, Trois Liu, Yuting Lai, Yiying Tseng, and Sam Tsai. Dred: a chinese machine reading comprehension dataset. *arXiv preprint arXiv:1806.00920*, 2018.
- 15 Zhihong Shao, Minlie Huang, Jiangtao Wen, Wenfei Xu, and Xiaoyan Zhu. Long and diverse text generation with planning-based hierarchical variational model. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, page 3257–3268. Association for Computational Linguistics, 2019.

- 16 Yaozong Shen, Lijie Wang, Ying Chen, Xinyan Xiao, Jing Liu, and Hua Wu. An interpretability evaluation benchmark for pre-trained language model. *arXiv preprint arXiv:2207.13948*, 2022.
- 17 Hongxuan Tang, Hongyu Li, Jing Liu, Yu Hong, Hua Wu, and Haifeng Wang. Dureader_robust: A chinese dataset towards evaluating robustness and generalization of machine reading comprehension in real-world applications. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, volume 2, page 955–963, 2021.
- 18 Yan Wang, Xiaojiang Liu, and Shuming Shi. Deep neural solver for math word problems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, page 845–854. Association for Computational Linguistics, 2017.
- 19 Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 1, page 496–505. Association for Computational Linguistics, 2017.
- 20 Bright Xu. Nlp chinese corpus: Large scale chinese corpus for nlp, September 2019. URL <https://doi.org/10.5281/zenodo.3402023>.
- 21 Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaowei Hua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. Clue: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772. International Committee on Computational Linguistics, 2020.
- 22 Liang Xu, Xiaojing Lu, Chenyang Yuan, Xuanwei Zhang, Huilin Xu, Hu Yuan, Guoao Wei, Xiang Pan, Xin Tian, Libo Qin, and Hu Hai. Fewclue: A chinese few-shot learning evaluation benchmark. *arXiv preprint arXiv:2107.07498*, 2021.
- 23 Chujie Zheng, Minlie Huang, and Aixin Sun. ChID: A large-scale Chinese IDiom dataset for cloze test. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, page 778–787. Association for Computational Linguistics, 2019.
- 24 Hao Zhou, Chujie Zheng, Kaili Huang, Minlie Huang, and Xiaoyan Zhu. Kdconv: A chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 7098–7108. Association for Computational Linguistics, 2020.