

“太一”：基于多任务指令微调的中英双语生物医学大模型

罗 凌

信息检索研究室

计算机科学与技术学院

大连理工大学

2024年1月4日



搜人搜物搜信息 重情重义重认知

太一 Taiyi



目录

1

项目简介

2

“太一”大模型

- 数据构建
- 模型训练
- 能力展示与评测

3

总结与展望

第一部分

项目简介

- 健康医疗信息化和智能化

- 生物医学领域特性

- 数据量大、增长快
- 专业术语多、非标准化
- 句子复杂等

- 困难和挑战

- 领域知识受限
- 性能下降
- 可靠性差

MeSH: D000868

Name: Anterior Compartment Syndrome
(胫骨筋膜室综合征)

Synonym:

- Anterior Tibial Syndrome,
- Tibial Anterior Compartment Syndrome
- Anterior-compartment syndrome



What is the difference between Anterior Compartment Syndrome and Anterior Tibial Syndrome?



I'm not aware of a medical condition called "Anterior Tibial Syndrome". However, I can provide information about "Anterior Compartment Syndrome".....



重新生成

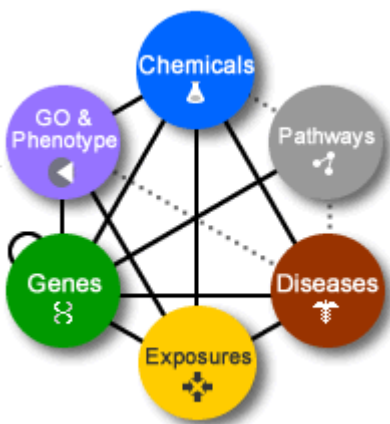
Anterior Compartment Syndrome and Anterior Tibial Syndrome are two distinct medical conditions that affect the lower leg, particularly the anterior (front) compartment. They have different causes, symptoms, and treatments.

● 主要专注于基于**监督微调**的**单语**（中文或英语）**医疗问答对话**

模型	基座	语言	指令数据	训练方式	开发团队	发布时间
GatorTron	Transformer	En	任务数据	PT+SFT	佛罗里达大学等	2022.5
Med-PaLM2	PaLM	En	问答数据	SFT	谷歌团队	2023.5
BenTsao	ChatGLM/LaMA	Zh	问答对话数据	SFT	哈尔滨工业大学	2023.4
DoctorGLM	ChatGLM	Zh	问答对话数据	SFT	上海交通大学	2023.4
HuatuoGPT	BLOOMZ	Zh	问答和对话数据	SFT+RL	香港中文大学	2023.5
ClinicalGPT	BLOOM	Zh	问答和对话数据	SFT+RL	北京邮电大学	2023.6
PULSE	BLOOMZ	Zh	问答、对话和任务数据	SFT	上海人工智能实验室	2023.7
DISC-MedLLM	Baichuan	Zh	问答和对话数据	SFT	复旦DISC	2023.9
ChatDoctor	LLaMA	En	问答和对话数据	SFT	美国德克萨斯大学	2023.3
Medalpaca	LLaMA	En	问答和对话数据	SFT	德国亚琛大学附属医院	2023.4

PT: 预训练; SFT: 监督微调; RL: 强化学习

● 面向生物医学多场景的多语多任务生物医学大模型



面向生物医学研究者

- ✓ 自动信息抽取
- ✓ 知识发现

辅助构建和管理生物医学知识库，提高信息建设和管理的效率。生物医学知识发现，以便验证得到新的科学发现。



面向临床医生

- ✓ 自动电子病历报告生成
- ✓ 疾病预测辅助临床决策

对实现疾病的辅助诊断、预防和治疗等具有重要的作用。



面向患者用户

- ✓ 智能导诊
- ✓ 健康咨询

这有助于患者更好地理解他们的健康状况和引导到适当的医疗部门，并减少不必要的医疗咨询。

第二部分

“太一”大模型

“太一” 整体框架



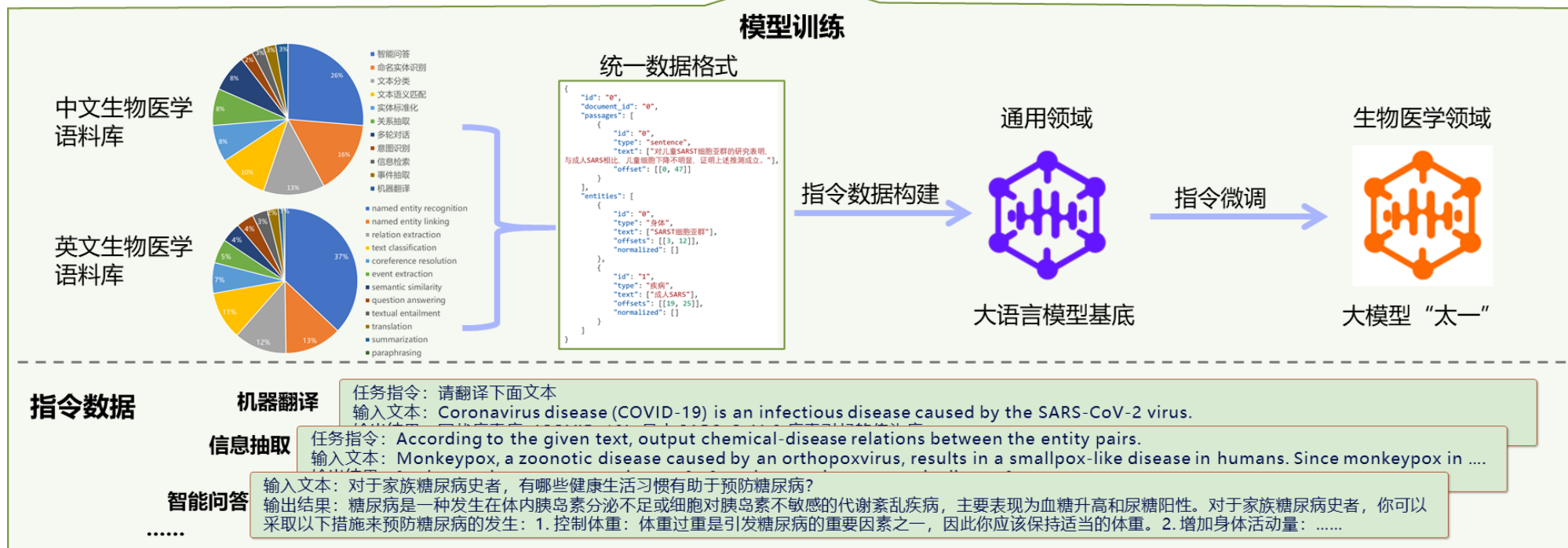
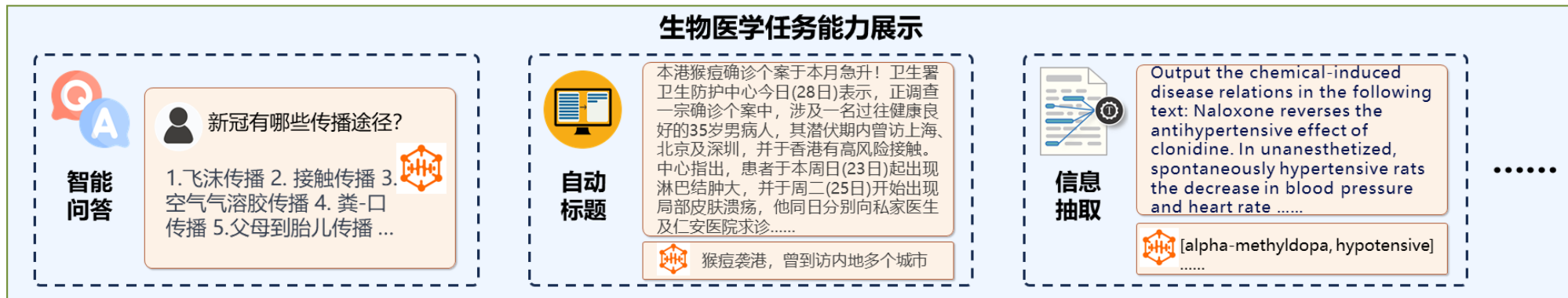
“太一”

● 太医：
生物学

● 大模型上多
一点：

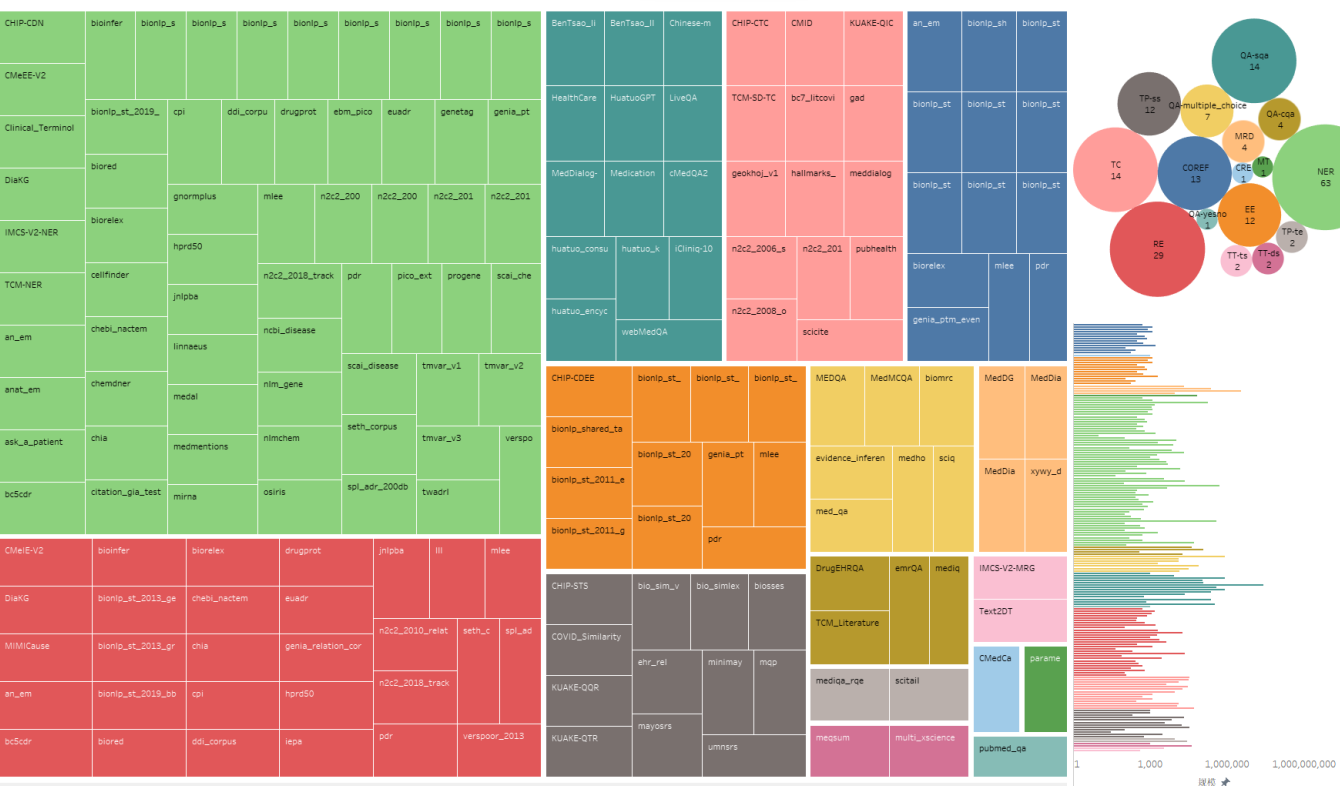
指令微调

● 浑然为一：
双语多任务





- 收集整理了数据集地址、论文地址、任务描述、数据规模等
 - 中文数据集38个、英文数据集102个，覆盖问答、对话、报告生成、信息抽取、机器翻译、标题生成、文本分类等10余种大类任务。



Overview of the Biomedical Datasets

- Overview of the Biomedical Datasets
 - Chinese Datasets
 - English Datasets

Chinese Datasets

Num	Dataset	Task	Size	URL	Task Description
1	CMedCausal	CRE	Train:800; Dev:200	[Dataset] [paper]	本次任务主要目标是从大量的医疗的问题中可以挖掘抽取医疗因果关系构建因果图。具体来说，细分的医学概念片段分为如下3类关系：1、因果关系是指某种原因直接导致某种结果；2、条件关系是指医学概念片段中一些条件对于修饰特定的因果关系，注意：条件并不一定导致结果的发生；3、上下位关系：上下位关系是指医学概念片段中一个概念是另一个概念的上位或下位关系的情况。关系具有方向性。
2	CHIP-CDEE	EE	Train:1,587; Dev:384	[Dataset] [paper]	有的事件属性不是实体。本次任务主要目标是从病历中挖掘出临床发现事件。即给定一段病历报告，要求从中抽取临床发现事件，包括事件主体、客体、描述词，以及发生状态。
3	MedDialog-zh	MRD	Train:2,725,989; Dev:340,748; Test:340,754	[Dataset] [paper]	原始的中文对话来自于haodf.com、英文对话来自于healthcaremagic.com和icliniq.com。
4	xywy_disease	MRD	Train:8,808	[Dataset] [paper]	疾病百科。
					任务具体定义如下：给定医生和患者对话H，其中xk为患者当前的对话语句，并假设H中xk+1包含标注的实体列表E，模型要

- 借鉴BigBio项目 (Fries et al., 2022)
- 将不同数据集根据任务类型进行格式转换

```
{
  "id": "ABCDEFGH",
  "document_id": "XXXXXX",
  "passages": [...],
  "entities": [...],
  "events": [...],
  "coreferences": [...],
  "relations": [...],
}

{
  "entities": [
    {
      "id": "0",
      "offsets": [[0, 8]],
      "text": ["Naloxone"],
      "type": "Chemical",
      "normalized": [{"db_name": "MESH",
        "db_id": "D009270"}]
    },
    ...
  ],
}
```

- 信息抽取
 - 文本段落
 - 实体
 - 关系
 - 因果关系
 - 事件
 - 指代消解
- 文本分类
- 问答
 - 多项选择
 - 简单问答
 - 基于上下文问答
- 多轮对话
- 机器翻译
- 文本对
 - 文本蕴含
 - 语义相似度计算
- 文本到文本/结构
 - 文本摘要
 - 文本到结构

- 数据集质量筛选和过滤
- 借鉴PromptCBLUE项目(Zhu et al., 2023)
指令模板多样性, 每个任务约20个模板
- 任务拆解, 减低难度, 增加多样性

从下面文本中识别出指定的实体类型:
治疗以选用大环内酯类抗生素, 沙眼衣原体肺炎也可用磺胺二甲基异唑, 年长儿和成人用氟喹诺酮类效果也很好。
实体类型: 疾病, 药物

从下面文本中识别疾病实体类型:

从下面文本中识别药物实体类型:

任务类型	中文数据规模	英文数据规模
命名实体识别	44,667	28,603
关系抽取	26,606	17,279
事件抽取	2,992	2,022
文本分类	37,624	40,339
文本对任务	45,548	11,237
机器翻译	74,113	
单轮问答	129,562	57,962
多轮对话	16,391	10,000
其他额外任务	9,370	
通用思维链数据	50,000	7,473
通用对话数据	390,000	
总计	1,001,788	



- 模型基座

- 通义千问 (Qwen-7B-base, Bai et al., 2023)

- 高效微调

- Qlora指令微调 (Dettmers et al., 2023)
- 8张A40 GPU

- 分阶段训练

- 阶段1：非对话任务数据 (5epoch)
- 阶段2：全部数据 (3epoch)

- 模型推理

- 1张3090/4090 GPU

任务类型	数据规模
信息抽取	122,169
文本分类	77,993
文本对匹配	56,785
机器翻译	74,113
其他任务	9,370

任务类型	数据规模
信息抽取	122,169
文本分类	77,993
文本对匹配	56,785
机器翻译	74,113
其他任务	9,370
问答	187,524
多轮对话	586,391

类型1

类型2



模型主要能力

- 智能问答
- 医学对话
- 医疗报告生成
- 生物医学信息抽取
 - 命名实体识别
 - 关系抽取
 - 事件抽取
- 机器翻译
- 标题生成
- 文本分类
- 文本相关性判断

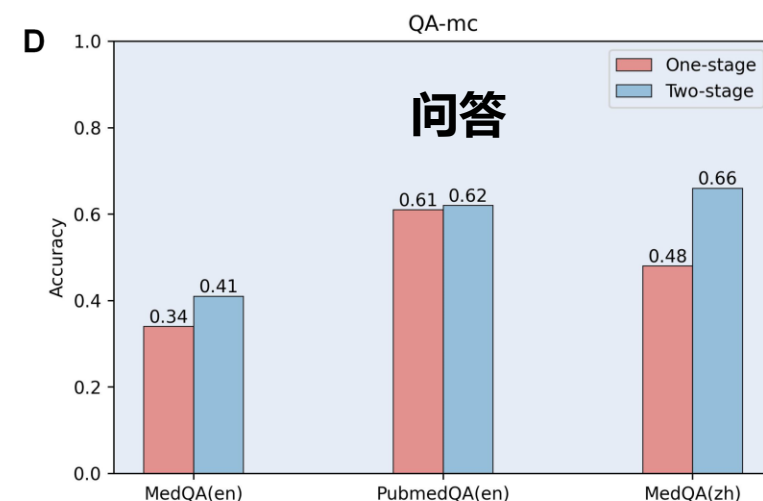
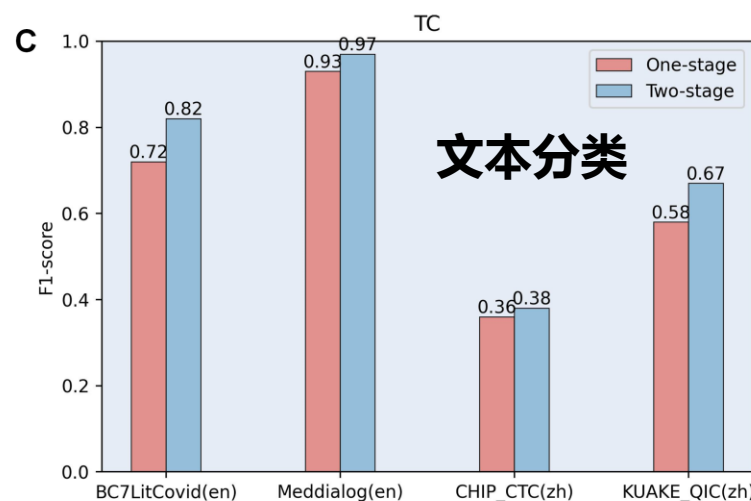
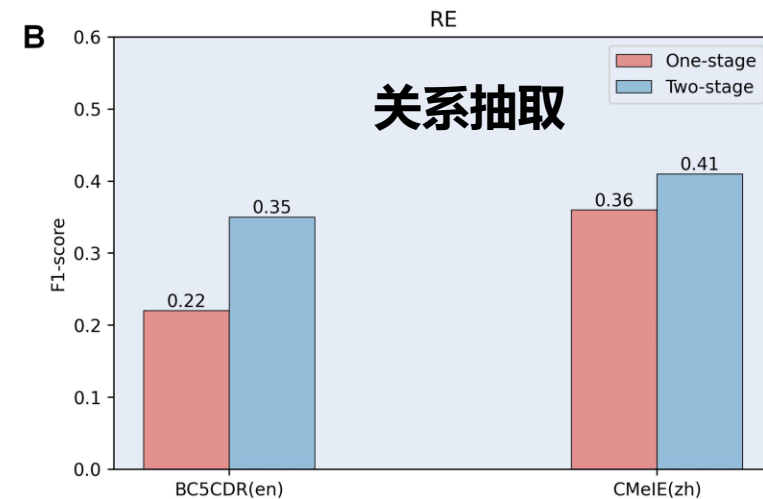
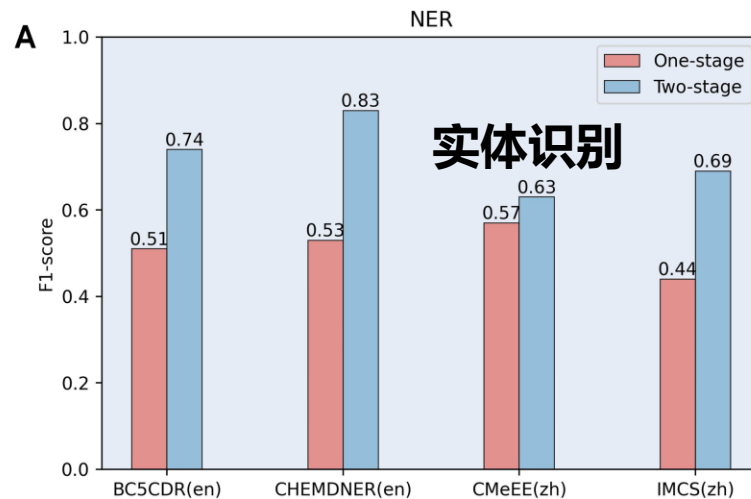
模型性能评测

- 选择题问答
- 命名实体识别
- 端到端三元组关系抽取
- 文本分类

实验分析-单阶段训练 vs. 两阶段训练



- 单阶段 (One-stage)
- 两阶段 (Two-stage)
- 测试集：随机选取200个样本
- 单阶段训练，两种任务类型收敛不一致
- 分阶段后，所有任务平均提升约10%以上



与其他现存模型指标对比



- 4大类任务，13个测试集**全集**上评测
- 平均指标**相比通用领域ChatGPT提升约15%**
- 平均指标**相比先进微调预训练模型差距约8%**

en: 英文; zh: 中文

任务类型	数据集	太一	ChatGP T3.5	SOTA
实体识别 (Micro-F1)	BC5CDR-Chem (en)	80.2	60.3 [11]	93.3 (PubMedBERT [11])
	BC5CDR-Dise (en)	69.1	51.8 [11]	85.6 (PubMedBERT [11])
	CHEMDNER (en)	79.9	36.5 [12]	92.4 (BioBERT [13])
	NCBIdisease (en)	73.1	50.5 [11]	87.8 (PubMedBERT [11])
	CMeEE-dev (zh)	65.7	47.0 [14]	74.0 (CBLUE [15])
关系抽取 (Micro-F1)	BC5CDR (en)	37.5	14.2	45.0 (BioGPT [20])
	CMeIE-dev (zh)	43.2	30.6 [14]	54.9 (CBLUE [15])
文本分类 (Micro-F1)	BC7LitCoivd (en)	84.0	63.9 [16]	91.8 (Bioformer [17])
	HOC (en)	80.0	51.2 [11]	82.3 (PubMedBERT [11])
	KUAKE_QIC-dev (zh)	77.4	48.5 [14]	85.9 (CBLUE [15])
填空问答 (Accuracy)	PubMedQA (en)	54.4	76.5 [11]	55.8 (PubMedBERT [11])
	MedQA-USMLE (en)	37.1	51.3 [18]	36.7 (BioBERT-large [19])
	MedQA-MCMLE (zh)	64.8	58.2 [18]	70.1 (RoBERTA-large [19])
全部	均值AVE	65.1	49.3	73.5

实验分析-泛化和零样本能力



- BioRED数据集 (Luo et al. 2022)
- 同时识别6类实体
- 任务不在训练集中，其中 Species和CellLine实体未在训练集中出现过
- ChatGPT: 39.1% (Zhou et al. 2023)

实体类型	P	R	F1
Chemical	71.7	56.6	63.3
Disease	82.9	53.3	64.9
Gene	93.1	49.0	64.2
Variant	63.3	58.5	60.8
Species	64.0	28.6	39.5
Celline	21.3	56.5	31.0
All	76.0	50.7	60.8

与Qwen7B-Chat对比-问答对话



太一	Qwen-7b-Chat
什么是鼠疫？	
鼠疫是一种由鼠疫杆菌引起的传染病，主要通过跳蚤等媒介传播。鼠疫杆菌主要存在于啮齿类动物体内，当啮齿类动物死亡后，跳蚤等昆虫会吸食其体液，从而将鼠疫杆菌带入人体，引起鼠疫感染。鼠疫的症状包括高热、寒战、淋巴结肿大、咳嗽、呼吸急促等，严重时可导致死亡。鼠疫是一种古老的传染病，历史上曾多次引起大规模的流行病，如14世纪的黑死病，对欧洲造成了巨大的人员和经济损失。	鼠疫是一种由鼠疫杆菌引起的急性传染病，它可以在动物之间传播，并且可以通过接触染病动物的排泄物或携带病毒的动物及其分泌物传播给人类。该病的症状包括发热、寒战、淋巴结肿大、呼吸困难和出血等。
有什么预防方法吗？	
是的，预防鼠疫的方法包括以下几点： 1. 避免接触感染源：尽量避免接触啮齿类动物，如老鼠、土拨鼠等，不要在它们的巢穴附近逗留。 2. 注意个人卫生：保持手部卫生，勤洗手，避免接触有感染风险的物品，如被感染的衣物、被褥等。 3. 食用安全：避免食用被感染的动物或其制品，如鼠肉、鼠肉制品等。 4. 家居环境：保持家居环境清洁卫生，定期消毒，特别是鼠洞、鼠粪等易滋生细菌的地方。 5. 预防接种：对于有鼠疫高发区的人群，可以接种鼠疫疫苗，提高免疫力。总之，预防鼠疫需要从多个方面入手，保持个人卫生、注意饮食安全、保持家居环境清洁卫生等都是预防鼠疫的有效方法。	预防鼠疫的最有效方法是避免接触携带鼠疫杆菌的动物和它们的排泄物，以及不触摸被污染的物品和表面。另外，应保持个人卫生，经常洗手，避免进食未煮熟的食物，以及及时接种疫苗来预防鼠疫的发生。
回答更详细也更准确	

与Qwen7B-Chat对比-实体识别



太一	Qwen-7b-Chat
从以下文本中识别出指定的实体类型： 两肺听诊可闻及干湿性啰音，胸部X线片两肺有斑片状模糊阴影，尤以双下肺明显，严重病例可合并胸腔积液或脓胸。 实体类型：疾病，医疗程序	
疾病：胸腔积液; 脓胸 医疗程序：两肺听诊; 胸部X线片	疾病：肺炎 医疗程序：听诊、胸部X线检查
Identify Chemical, Disease entities in the following text: Phenobarbital-induced dyskinesia in a neurologically-impaired child. A 2-year-old child with known neurologic impairment developed a dyskinesia soon after starting phenobarbital therapy for seizures. Known causes of movement disorders were eliminated after evaluation. On repeat challenge with phenobarbital, the dyskinesia recurred. Phenobarbital should be added to the list of anticonvulsant drugs that can cause movement disorders.	
Chemical: Phenobarbital; phenobarbital Disease: dyskinesia; neurologically-impaired; neurologic impairment; seizures; movement disorders	Chemical: Phenobarbital, Anticonvulsant drugs Disease entities: Neurological impairment, Seizures, Movement disorders

实体识别结果更准确



太一	Qwen-7b-Chat
出下面文本中具有药物治疗关系的实体对： 疟疾感染应当使用氯喹（首选）或羟氯喹治疗感染，共 3 次给药。患者应遵循医生的建议和药物使用说明，以确保药物的最大效力。	
具有药物治疗关系的实体对如下：[疟疾感染, 氯喹]; [疟疾感染, 羟氯喹]	该句包含以下药物实体对： 氯喹：药物名称 羟氯喹：药物名称
Extract the Protein-Protein Interaction from the following text: Such motifs are known to bind to profilins (PFNs), small proteins engaged in the control of actin dynamics.	
[actin, PFNs]; [actin, profilins]	The Protein-Protein Interaction mentioned in the given text is between "motifs" and "profilins" (PFNs).

复杂任务经过指令微调，任务指令理解更准确

● 输入输出指令多样性

从下面文本中识别出指定的实体类型：

治疗以选用大环内酯类抗生素，沙眼衣原体肺炎也可用磺胺二甲基异唑，年长儿和成人用氟喹诺酮类效果也很好。

实体类型：疾病，药物



疾病：沙眼衣原体肺炎

药物：大环内酯类抗生素；磺胺二甲基异唑；氟喹诺酮类



从下面文本中识别疾病，药物实体，并以 “[实体，实体类型]” 格式返回：

治疗以选用大环内酯类抗生素，沙眼衣原体肺炎也可用磺胺二甲基异唑，年长儿和成人用氟喹诺酮类效果也很好。



[沙眼衣原体肺炎，疾病]，[大环内酯类抗生素，药物]，[氟喹诺酮类，药物]，[磺胺二甲基异唑，药物]



● 任务不在相应语言的训练集中

变异 实体识别

从下面文本中抽取变异实体：

在所有脑卒中类型中CVT占0.5% ~ 1%。其形成的风险因素通常与血液淤滞、血管壁变化和血液成分变化的Virchow三联体有关。据报道，中国人群CVT的遗传倾向包括V因子Leiden突变、凝血酶G20210A突变、高同型半胱氨酸血症、蛋白C、蛋白S或抗凝血酶III缺陷等。



上述文本中包含的变异实体为：V因子Leiden突变; 凝血酶G20210A突变

医疗程序 实体识别

Find the medical procedure entity in the following text:

immunomodulator gamma globulin 400 mg/(kg·d), intravenous infusion for 3 to 5 days.



Here are the medical procedure entities mentioned in the text: intravenous infusion

第三部分

总结与展望

- **丰富的生物医学训练资源**

- 140个中英数据集
- 十种以上生物医学任务大类

- **出色的双语多任务能力**

- 百万指令数据，分阶段指令微调，显著提升通用基座在领域任务上的理解能力和性能
- 双语互补，问答对话和多种BioNLP任务能力

- **较好的泛化能力**

- 同一大类任务，不同场景
- 指令输出输入具备多样性

项目地址:

<https://github.com/DUTIR-BioNLP/Taiyi-LLM>



相关资源以及报告PPT已经放在项目上。

arXiv:

<https://arxiv.org/abs/2311.11608>

- 目前“太一”大模型仍存在大模型的一些常见缺点，例如误解、幻觉、信息有限性、偏见、多轮长对话能力较不稳定、话题转换能力弱等。
- 领域知识受限
 - 增量预训练
- 可靠性和安全性弱
 - 检索、知识库增强
 - 提供证据链
- 非生成任务性能上升空间大
 - 领域指令精细化
 - 多维度反馈的强化学习

请介绍一下抗生素中的八环素类。



八环素类是一种广谱抗生素，属于四环素类抗生素的一种。它能够抑制细菌的生长和繁殖，从而达到治疗感染的目的。八环素类抗生素可以用于治疗多种细菌感染，如呼吸道感染、皮肤感染、泌尿道感染等。但是，八环素类抗生素也有一些副作用，如恶心、呕吐、腹泻等，因此在使用时需要谨慎。同时，八环素类抗生素也不应该滥用，因为滥用会导致细菌耐药性的产生，从而影响治疗效果。

并不存在八环素类！

- **大连理工大学信息检索研究室**

- 指导教师：罗凌、杨志豪、王健、孙媛媛、林鸿飞
- 学生成员：宁金忠、赵颖闻、丁泽源、陈鹏、付伟茹、韩钦宇、徐广涛、邱云志、潘丁豪、李记如、汪志军、李浩、凤文铎、涂森博、刘宇奇

- **本项目的工作受到以下开源项目与技术的启发和帮助，在此对相关项目和研究开发人员表示感谢，包括但不限于：**

- Qwen: <https://github.com/QwenLM/Qwen>
- Firefly: <https://github.com/yangjianxin1/Firefly>
- BigBIO: <https://github.com/bigscience-workshop/biomedical>
- PromptCBLUE: <https://github.com/michael-wzhu/PromptCBLUE>
- Moss: <https://github.com/OpenLMLab/MOSS>
- 太一logo 由文心一言AI合成

1. Yang X, Chen A, PourNejatian N, et al. A large language model for electronic health records. NPJ Digital Medicine. 2022 Dec 26;5(1):194.
2. Singhal K, Tu T, Gottweis J, et al. Towards expert-level medical question answering with large language models. arXiv preprint arXiv:2305.09617. 2023.
3. Wang H, Liu C, Xi N, et al. Huatuo: Tuning llama model with chinese medical knowledge. arXiv preprint arXiv:2304.06975. 2023.
4. Xiong H, Wang S, Zhu Y, et al. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. arXiv preprint arXiv:2304.01097. 2023.
5. Zhang H, Chen J, Jiang F, et al. HuatuoGPT, towards Taming Language Model to Be a Doctor. arXiv preprint arXiv:2305.15075. 2023.
6. Wang G, Yang G, Du Z, et al. ClinicalGPT: Large Language Models Finetuned with Diverse Medical Data and Comprehensive Evaluation. arXiv preprint arXiv:2306.09968. 2023 Jun 16.
7. <https://github.com/openmedlab/PULSE>
8. Bao Z, Chen W, Xiao S, et al. Disc-medllm: Bridging general large language models and real-world medical consultation. arXiv preprint arXiv:2308.14346. 2023.
9. Yunxiang L, Zihan L, Kai Z, et al. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. arXiv preprint arXiv:2303.14070. 2023.
10. Han T, Adams LC, Papaioannou JM, et al. MedAlpaca--An Open-Source Collection of Medical Conversational AI Models and Training Data. arXiv preprint arXiv:2304.08247. 2023.

11. Chen Q, Sun H, Liu H, et al. An extensive benchmark study on biomedical text generation and mining with ChatGPT. *Bioinformatics*. 2023 Sep 1;39(9):btad557.
12. Zhou W, Zhang S, Gu Y, et al. UniversalNER: Targeted Distillation from Large Language Models for Open Named Entity Recognition. *arXiv preprint arXiv:2308.03279*. 2023.
13. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020 Feb 15;36(4):1234-40.
14. Zhu W, Wang X, Zheng H, et al. PromptCBLUE: A Chinese Prompt Tuning Benchmark for the Medical Domain. 2023, *arXiv:2310.14151*.
15. <https://tianchi.aliyun.com/dataset/95414>
16. Chen Q, Du J, Hu Y, et al. Large language models in biomedical natural language processing: benchmarks, baselines, and recommendations. *arXiv preprint arXiv:2305.16326*. 2023.
17. Chen Q, Allot A, Leaman R, et al. 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 2021 (Vol. 10)*.
18. Wang Y, Ma X, Chen W. Augmenting Black-box LLMs with Medical Textbooks for Clinical Question Answering. *arXiv preprint arXiv:2309.02233*. 2023 Sep 5.
19. Jin D, Pan E, Oufattole N, et al. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*. 2021 Jul 12;11(14):6421.
20. Luo R, Sun L, Xia Y, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*. 2022 Nov;23(6):bbac409.



太医仁心昭古道
一剂良方复新康

太一 Taiyi

—— 大连理工大学信息检索研究室 ——



大连理工大学

信息检索研究室

Information Retrieval laboratory of DUT

搜人搜物搜信息 重情重义重认知

谢谢!

<https://ir.dlut.edu.cn/>
lingluo@dlut.edu.cn

