CBLUE： 一个中文生物医学语言理解评估基准

张宁宇 1*，陈莫沙 2*，毕振 1*，梁小转 1*，李雷 1*，尚新 3 尹康平 2 谭传奇 2 徐建 2 黄飞 2 施罗思 2 倪元 4 谢国通 4、5、6 隋志方 7、13 常宝宝 7、13 宗会 8、14 袁征 9 李林峰 10 严俊 10 赞红英 11、13 张坤利 11、13 唐布周 12、13†陈青才 12、13†
1 阿里巴巴-浙江大学前沿技术联合研究中心，浙江大学 2 阿里巴巴集团 3 浙江大学数学科学学院 4 平安健康科技有限公司 5 平安健康云公司 6 平安国际智慧城市科技股份有限公司 7 教育部计算语言学重点实验室，北京大学 8 同济大学生命科学技术学院 9 清华大学
10 亿度云技术有限公司 11 郑州大学信息工程学院
哈尔滨工业大学（深圳） 13 鹏城实验室，14 飞利浦中国研究部

摘要

随着生物医学语言理解基准的发展，人工智能在医疗领域的应用越来越广泛。然而大多数基准仅限于英语，这使得将许多英语的成功复制到其他语言中变得具有挑战性。为了促进这一方向的研究，我们收集了真实世界的生物医学数据，并提出了第一个中文生物医学语言理解评估（CBLUE）基准：一个包括命名实体识别、信息提取、临床诊断规范化等自然语言理解任务的集合以及与之相关的在线平台用于模型评价、比较和分析。为建立对这些任务的评估，我们在当前的 11 个预训练中文模型上进行了实证结果报告，实验结果显示最先进的神经网络模型的表现远低于人类天花板水平。我们的基准发布在 https://tianchi.aliyun.com/dataset/dataDetail？dataId=95414&lang=en-us。

1.介绍

人工智能正在逐渐改变医疗保健的面貌，生物医学研究 (Yu 等人，2018 年）。随着生物医学数据集的快速进步，生物医学自然语言处理（BioNLP）已促进了广泛的应用。
的应用，如生物医学文本挖掘，利用电子健康记录 (EHR) 中的文本数据。
推动这些改进和模型快速迭代的关键驱动力之一是使用通用评估数据集和基准（Gijsbers 等人，2019 年）。例如，BLURB (Gu 等人，2020)、PubMedQA (Jin 等人，2019) 和其他先驱基准为我们提供了研究生物医学语言理解和开发实际应用的机会。不幸的是，大多数这些基准都是在英语中开发的，这使得相关的机器智能具有盎格鲁-撒克逊中心主义的特点。与此同时，其他语言，如中文，拥有独特的语言特征和类别需要考虑。尽管中国人口占世界总人口的四分之一，但还没有现有的中文生物医学自然语言理解评价基准。
为解决这一问题并促进中文自然语言处理研究，我们首先提出一个全面的中文生物医学语义理解评估（CBLE）基准。该基准包含八项生物医学语义理解任务：命名实体识别、信息抽取、临床诊断规范化、短文本分类、问答（在迁移学习设置下）、意图分类和语义相似度等。我们在 CBLE 上对几个预训练的中文语言模型进行评估，并报告其性能。当前模型的表现远远低于单个人类的标准

，为未来改进留有空间。我们还通过案例研究进行综合分析，以表明中文生物医学语言理解面临的挑战和语义差异。我们的目标是开发一个通用的 GLUE 风格的开放平台来促进中国生物自然语言处理社区的研究，并且这项工作有助于加速该方向上的研究。总体而言，本研究的主要贡献如下：

- 我们提出了第一个中文生物医学语言理解基准，这是一个由社区驱动的开放项目，具有多种任务。所提出的基准为中文 BioNLP 社区提供了一个平台，并鼓励新的数据集贡献。
- 我们对 11 个中国预训练语言模型进行了系统评估，以了解这些任务带来的挑战。我们发布了基准的源代码作为未来研究的工具包。

2 相关工作

过去几年中，已经开发了几个基准来评估一般语言理解。GLUE（Wang 等人，2019b）是第一个正式挑战的框架之一，该框架提供了任务无关转移学习技术之间的直接比较。SuperGLUE (Wang 等人，2019a)，类似于 GLUE ，引入了一组更困难的语言理解数据集的新集合。其他类似的基准包括 DecaNLP（McCann 等人，2018），它将一组目标任务重新表述为通用问题回答格式，并禁止任务特定参数；SentEval (Conneau 和 Kiela，2018)，它评估显式固定大小句子嵌入。非英语基准包括俄罗斯 SuperGLUE (Shavrina 等人，2020)和 CLUE (Xu 等人，2020)，这是一个社区驱动的基准，包含九个中文自然语言理解任务。这些在通用领域的基准为我们研究人员提供了一个北星目标，并且是我们可以自信地说我们在我们的领域取得了巨大进步的部分原因。

对于 BioNLP，已经提出了许多数据集和基准（Wang 等，2020；Li 等，2016；Wu 等，2019），这些都促进了生物医学语言理解的发展 (Beltagy 等人，2019；Lewis 等人，2020；Lee 等人，2020）。Tsatsaronis et al.(2015) 提出生物医学语言理解数据集以及大规模生物医学语义索引和问答竞赛。Jin 等人(2019 年)提出 PubMedQA，这是一个从 PubMed 摘要中收集的新型生物医学问答数据集。Pappas 等人（2018 年）提出了 BioRead，这是一项公开可用的基于掩码的生物医学机器阅读理解 (MRC) 数据集。Gu 等 人（2020 年） 创建了一个包含生物医学语言理解和推理基准测试 BLURB 的排行榜。不同于通用领域的语料库，生物医学语料库需要专家干预，并且劳动密集型且耗时。此外，大多数基准都是基于英语；忽略其他语言意味着可能有价值的信息可能会丢失，这对于泛化是有帮助的。

本研究旨在填补中文生物医学领域自然语言处理的空白，开发首个中文生物医学语义理解基准。需要注意的是，中文生物医学文本在语言学上不同于英文，并且具有其特有的领域特征，因此需要设计一个专门针对中文的 BioNLP 基准。

3 CBLUE 概述

3.1 设计原则

CBLUE 包含 8 个生物医学语言理解任务，表 1 显示了 CBLUE 的任务描述和统计数据。与 CLUE（Xu 等，2020）不同的是，CBLUE 的数据源更加多样化(标注成本高)，任务设置更丰富，因此对 NLP 模型提出了更高的挑战性。我们介绍

CBLUE 的设计原则如下：
1） 多样化任务：CBLUE 包含广泛的任务，包括单词级、序列级和序列对。
2）不同分布的数据种类：CBLUE 从各种来源收集数据，包括临床试验、电子健康记录（EHR）、医学论坛、教科书和搜索引擎日志等，并以真实世界的分布进行收集。
3）长期维护的质量控制：我们请了顶级医院的专家（三甲医院医生）对数据集进行标注，并且仔细审查数据，以确保数据质量。

Dataset	Task	Train	Dev	Test	Metrics
CMeEE	NER	15,000	5,000	3,000	Micro F1
CMeIE	Information Extraction	14,339	3,585	4,482	Micro F1
CHIP-CDN	Diagnosis Normalization	6,000	2,000	10,192	Micro F1
CHIP-STS	Sentence Similarity	16,000	4,000	10,000	Macro F1
CHIP-CTC	Sentence Classification	22,962	7,682	10,000	Macro F1
KUAKE-QIC	Intent Classification	6,931	1,955	1,994	Accuracy
KUAKE-QTR	Query-Document Relevance	24,174	2,913	5,465	Accuracy
KUAKE-QQR	Query-Query Relevance	15,000	1,600	1,596	Accuracy

Table 1: Task descriptions and statistics in CBLUE. CMeEE and CMeIE are sequence labeling tasks. Others are single sentence or sentence pair classification tasks.

Benchmark	Language	Domain	Data Distribution	Label Distribution
CBLUE	Chinese	medical	long-tailed (CMeEE)	non-i.i.d (CHIP-STS)
CLUE	Chinese	general	uniform	i.i.d
BLURB	English	medical	uniform	i.i.d

Table 2: Difference between CBLUE, CLUE and BLURB. There are three major differences: a) CBLUE has a much more diverse task setting with different data sources in the biomedical domain including clinical trials, EHRs medical forum, text books and search engine logs; b) CBLUE has a long-tailed distribution which is challenging c) CBLUE contains a specific transfer learning scenario supported by the CHIP-STS dataset, in which the testing set has a different distribution fromthe training set.

表 2：CBLUE、CLUE 和 BLURB 之间的差异。主要有三个主要的差异：a）CBLUE 在生物医学领域包括临床试验，电子健康记录（EHR），医疗论坛，教科书以及搜索引擎日志等不同数据源中具有更加多样化的任务设置；b）CBLUE 具有长尾分布，这使得它更具挑战性；c）CBLUE 包含一个由 CHIP-STS 数据集支持的具体迁移学习场景，在该场景中测试集与训练集的分布不同。

3.2 任务

CMeEE 对于这项任务，数据集首先在 CHIP 2020 中发布。给定一个预先定义的模式，该任务是识别并从给出的句子中提取实体，并将其分类为九类：疾病、临床表现、药物、医疗设备、医疗程序、身体部位、医学检查、微生物和部门。
CMeIE 对于这项任务，数据集也在 CHIP2020（Guan 等人，2020）中发布。该任务的目标是根据模式约束在句子中识别实体和关系。该数据集中定义了 53 个关系，包括 10 个同义子关系和 43 个其他子关系。
CHIP-CDN 为此任务，数据集是标准化中国电子病历最终诊断中的术语。给定

原始短语，该任务将根据北京临床版 ICD-10 标准将其规范化为标准术语。
CHIP-CTC 对于这项任务，数据集是分类临床试验的资格标准，这些是定义为识别受试者是否符合临床试验的基本指南（Zong 等人，2021）。所有文本数据均来自中国临床试验注册中心 (ChiCTR) 的网站，并且总共定义了 44 类。该任务类似于文本分类；尽管这不是一项新任务，但对中国的临床试验准则的研究和语料库仍然有限，我们希望促进未来研究以造福社会。
CHIP-STS 对于这项任务，数据集是用于非独立同分布（i.i.d）设置下的句子相似性。具体来说，该任务旨在评估中文疾病问题和答案数据中不同疾病的泛化能力。给定与 5 种不同的疾病相关的语句对(训练和测试集中包含的疾病类型不同)，任务需要确定两个句子之间的语义是否相似。
KUAKE-QIC 对于这项任务，数据集是用于意图分类。给定搜索引擎查询任务是将它们中的每一个分类到 KUAKE-QIC 中定义的 11 个医疗意图类别之一。这些包括诊断、病因分析、治疗计划、医学建议和测试结果分析等。
KUAKE-QTR 对于这项任务，数据集用于估计查询文档标题的相关性。给定一个查询（例如，“维生素 B 缺乏的症状”），该任务旨在找到相关的标题（例如，“维生素 B 缺乏的主要表现形式”。
KUAKE-QQR 为此任务，数据集用于评估两个查询中表达的内容的相关性。类似于 KUAKE-QTR，该任务旨在估计查询-查询相关性，在现实世界搜索引擎中这是一个重要的且具有挑战性的任务。

3.3 数据收集

由于机器学习模型大多基于数据驱动，因此数据起着至关重要的作用，并且通常以静态数据集的形式存在（Gebru 等，2018）。我们从各种来源收集不同任务的数据，包括临床试验、电子健康记录 (EHR) 、医学书籍以及来自真实世界搜索引擎的搜索日志。生物医学数据可能包含患者姓名、年龄和性别等私人信息，所有收集到的数据集均进行了匿名化处理并由每个数据提供方的 IRB 委员会进行审查，以保护隐私。接下来我们将介绍数据采集细节。

来自临床试验的收集

临床试验资格标准文本是从 ChiCTR 收集的，这是一个非营利组织，为公众研究提供有关临床试验注册的信息。在每个试验登记文件中，资格标准文本以包括标准和排除标准的形式作为段落进行组织。一些无意义的文字被排除在外，并且剩余文字进行了注释生成 CHIP-CTC 数据集。

从电子病历中收集

我们从几家三级甲等医院的病历中获取最终诊断，从中抽取不同科室的几项诊断项目构建 CHIP-CDN 数据集用于研究。这些诊断项目是从不包含在通用医学同义词字典中的项目中随机抽取的。
最终诊断不涉及任何隐私信息。

来自医学论坛和教科书的收集

由于新冠疫情，通过互联网进行在线咨询越来越流行。为了促进数据多样性，我们选择患者在线提问构建 CHIP-STS 语料库。请注意大多数问题是主诉。为确保语料库的权威性和实用性，我们也选择了儿科教材（王等，2018）、临床儿科学（沈和桂，2013）以及临床实践指南第四版作为来源来收集数据以构建 CMeIE 和 CMeEE 语料库。

来自搜索引擎日志的收集

我们还从阿里巴巴 KUAKE 搜索引擎等真实世界的搜索引擎中收集了搜索日志。首先，通过医疗标签过滤原始搜索日志中的查询以获取候选医学文本；然后，为每个查询采样具有非零相关性分数的文档（即确定文档是否与查询相关）。具体来说，我们将所有文档分为高、中和尾部文档，并且均匀地采样数据以保证多样性。我们利用搜索日志的数据构建 KUAKE-QTC,KUAKE-QTR 和 KUAKE-QQR 数据集。

3.4 注释

每个样本由三到五名领域专家标注，投票最多的标注被用来估计人类表现。在标注阶段，我们添加控制问题以防止领域专家的不诚实行为。因此，我们拒绝任何在培训阶段失败并无法通过低性能控制任务的领域专家的标注，并且不采用那些在控制任务中表现不佳的结果。我们在批准和审核时保持严格的标准，至少从每位工作者那里选择 10 个随机样本来决定是否批准或拒绝所有他们的 HITs。我们也使用 Fleiss' Kappa 分数（Fleiss， 1971）计算注释者之间的平均互评一致性，发现六个注释中有五个几乎完美一致 ($\kappa = 0.9$) 。

3.5 特性

保留有用信息的匿名化生物医学数据可能被视为对个人隐私的侵犯，因为它们通常包含敏感信息。因此，在发布基准之前，我们采用（Lee 等，2017）提出的保留有用信息的匿名化方法来匿名化数据。

真实世界分布 为了促进模型的泛化，我们在 CBLUE 基准中使用的所有数据都遵循真实世界的分布而没有进行上采样或下采样。如图 1（a）所示，我们的数据集遵循 Zipf 定律所描述的长尾分布，并且不可避免地是长尾分布。然而，长尾分布对性能几乎没有显著影响。此外，一些数据集，例如 CMedIE，具有粗粒度和细粒度关系标签的层次结构，如图 1（b）所示。

多样化的任务设置我们的 CBLUE 基准包括八个不同的任务，包括命名实体识别、关系抽取和单句/句子对分类。除了独立的和 i.i.d.场景之外，我们的 CBLUE 基准还包含一个由 CHIP-STS 数据集支持的具体迁移学习场景，在该场景中测试集与训练集具有不同的分布。

3.6 领跑榜

我们为用户在 CBLUE 上提交自己的结果提供了一个排行榜。当用户提交预测结果时，评估系统将对每个任务给出最终分数。平台从阿里云 6 免费提供了 60 个 GPU 小时来帮助研究人员开发和训练他们的模型。

3.7 分发和维护

我们的 CBLUE 基准于 2021 年 4 月 1 日在线发布，至今已有超过三百位研究人员使用了数据集，并有八十多个团队向我们平台提交了模型预测结果，包括医疗机构（北京协和医院等）、高校（清华大学、浙江大学等）以及企业（百度、京东等）。我们将持续维护该基准，关注新的需求并添加新任务。

3.8 可重复性

为了使 CBLUE 基准更容易使用，我们还提供了一个在 PyTorch（Paszke 等人，2019 年）中实现的工具包以确保可重复性。我们的工具包支持主流预训练模型和广泛的目标任务。
CMeIE STS
61.7% 70.1%
62.1％ 70.9％
62.4 53.7 69.4 85.5
61.8％，55.9％。
47.6
55.7 85.2 85.3 62.8
50.5 35.9 50.2 75.8
61.8％，37.5％，84.8％。
50.1 57.8 68.6 83.2
53.2 67.7 59.7
62.4 68.6 85.6 82.7
83.8 84.3
66.0 65.0 78.0 93.0 88.0 71.0

4 个实验

基线 我们基于不同的中文预训练语言模型进行实验。我们为每个 CBLUE 任务添加一个额外的输出层（例如，MLP），并微调预训练模型。
模型 我们在以下公共可用的中文预训练模型上评估 CBLUE：
• BERT-base（Devlin 等人，2018 年）。我们使用了具有 12 层、768 个隐藏单元、12 个头和 1.1 亿个参数的基模型。
• BERT-wwm-ext-base（崔等，2019）。一种带有整词掩码的中文预训练 BERT 模型。
• RoBERTa-large（Liu 等，2019）。与 BERT 相比，RoBERTa 删除了下一个句子预测目标，并动态更改应用于训练数据的掩码模式。
• RoBERTa-wwm-ext-base/large。RoBERTa-wwm-ext 是一个高效的预训练模型，它结合了 RoBERTa 和 BERT-wwm 的优点。
• ALBERT-tiny/xxlarge（Lan 等人，2019）。AL-BERT 是一种预训练模型，有两个目标：Masked 语言建模 (MLM) 和句子排序预测 (SOP)。
• ZEN（Diao 等，2019）。一种基于 BERT 的中文文本编码器，通过 n-gram 表示增强，在训练过程中考虑了不同字符的组合。
• Mac-BERT-base/large（Cui 等，2020）。Mac-BERT 是一种改进的 BERT 模

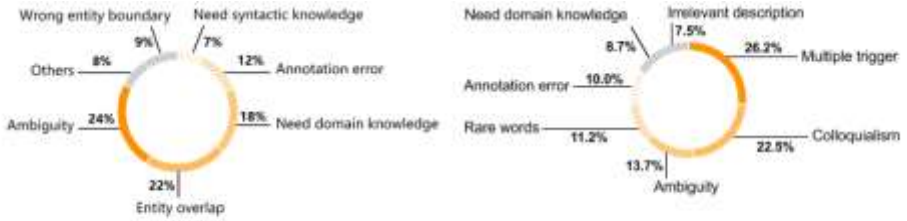
型，在 MLM 任务上进行了预训练。
• PCL-MedBERT7。由鹏城实验室提出的预训练医学语言模型。
我们使用 PyTorch（Paszke 等人，2019 年）实现所有基线。有关训练细节，请参阅附录。

4.1 基准结果

我们在表 3 中报告了我们基准模型在 CBLUE 基准上的结果。我们注意到，预训练的大型模型表现更好。由于中文文本由术语组成，精心设计的掩码策略可能有助于表示学习。然而，我们观察到，在一些任务（如 CTC、QIC、QTR 和 QQR）中，使用整词掩码的模型并不总是比其他模型表现出更好的性能，表明我们的基准中的任务具有挑战性，并且需要开发更高级的技术。此外，我们发现 ALBERT-tiny 在 CDN、STS、QTR 和 QQR 任务上与基线模型的表现相当，这说明较小的模型也可能在特定的任务中表现良好。我们认为这是由于预训练语料库和中国医学文本之间的不同分布造成的；因此，大型 PTLM 可能无法获得令人满意的性能。最后，我们注意到 PCL-MedBERT，它倾向于成为中文生物医学文本处理任务的最佳实践之一，但其表现不如预期的好。这进一步证明了该领域的困难程度。

		CMeEE	CMeIE	CDN	CTC	STS	QIC	QTR	QQR
Trained annotation	annotator 1	69.0	62.0	60.0	73.0	94.0	87.0	75.0	80.0
	annotator 2	62.0	65.0	69.0	75.0	93.0	91.0	62.0	88.0
	annotator 3	69.0	67.0	62.0	80.0	88.0	83.0	71.0	90.0
	avg	66.7	64.7	63.7	76.0	91.7	87.0	69.3	86.0
	majority	**67.0**	**66.0**	**65.0**	**78.0**	**93.0**	**88.0**	**71.0**	**89.0**
	best model	62.4	55.9	59.3	70.9	85.6	85.5	62.9	84.7

Table 4: Human performance of two-stage evaluation scores with the best-performed model. "avg" refers to the mean score from the three annotators. "majority" indicates the performance taken from the majority vote of amateur humans. Bold text denotes the best result among human and model prediction.



(a) Error Analysis in CMeEE　　(b) Error Analysis in KUAKE-QIC

Figure 2: We conduct error analysis on datasets CMeEE and QIC. For CMeEE, we divide error cases into 6 categories, including ambiguity, need domain knowledge, entity overlap, wrong entity boundary, annotation error and others (long sequence, rare words, etc.). For KUAKE-QIC, we divide error cases into 7 categories, including multiple triggers, colloquialism, ambiguity, rare words, annotation error, irrelevant description, and need domain knowledge.

图 2：我们对 CMeEE 和 QIC 数据集进行错误分析。对于 CMeEE，我们将错误案例分为 6 类，包括歧义、需要领域知识、实体重叠、错误的实体边界、注释错误和其他（长序列、罕见单词等）。对于 KUAKE-QIC，我们将错误案例分为 7 类，包括多个触发器、口语化、歧义、罕见单词、注释错误、无关描述以及需要领域知识。
我们的基准，当代模型可能很难快速实现卓越的性能。

4.2 人类表现

对于 CBLUE 中的所有任务，我们要求没有医学经验的业余人类标注者从测试集实例中进行标注，并计算其对由专家标注的黄金标签的多数投票。与 SuperGLUE (Wang 等人，2019 年) 类似，我们需要在他们开始处理测试数据之前先训练这些标注者。首先需要让标注者们从开发集中标注一些数据；然后，他们的标注会被验证为黄金标准。标注者需要反复纠正自己的错误标记以便掌握特定的任务。最后，他们会对测试数据中的实例进行标注，而这些标注将用于计算最终的人类分数。结果如表 4 和表 3 的最后一行所示，在所有任务中，人类的表现都更好。

4.3 案例研究

我们选择两个数据集：CMeEE 和 KUAKE-QIC，一个序列标注与分类任务。分别进行案例分析。如图 2 所示，我们报告了各种错误类型占总错误的比例 8；对于 CMeEE，我们注意到实体重叠 9、歧义 10、需要领域知识 11 和标注错误 12 是导致预测失败的主要原因。此外，在 CMeEE 中存在许多实体重叠的实例，这可能会对命名实体识别任务造成混淆。而在 KUAKE-QIC 的分析中，几乎一半的坏例都是由于多个触发词 13 和口语化造成的。口语化在搜索查询中是很自然的现象，这意味着一些中文医学文本中的描述过于简单、口语化或不准确。

We present several examples of CMeEE in Table 5. In the second row, we observe that "rash can be caused by host producing specific antitoxin antibodies" is an example of CMeEE.

8 请参阅附录中的错误定义。

在实例中存在多个重叠实体。

10 该实例具有相似的上下文但不同的含义，这误导了预测。

在实例中存在生物医学术语，需要专业知识才能理解。

注释标签是错误的。

存在多个指示词误导预测。

14 这个例子与书面语言（例如，有许多缩写）大不相同。

包括缺失、转位和倒位...

表 5：CMeEE 的案例研究。我们对 3 个采样句子评估了 roberta-wwm-ext 和 PCL-MedBERT，以及它们的黄金标签和模型预测结果。Ite（医学检查项目）、Pro（医疗程序）、Bod（身体）和 Sym（临床症状）被标记为医学命名词。O 表示模型无法从句子中提取实体。RO = roberta-wwm-ext，MB = PCL-MedBERT。

Query	Model			Gold
	BERT	BERT-ext	MedBERT	
请问淋巴细胞比率偏高、中性细胞比率偏低有事吗? Does it matter if the ratio of lymphocytes is high and the ratio of neutrophils is low?	病情诊断 Diagnosis	病情诊断 Diagnosis	指标解读 Test results analysis	指标解读 Test results analysis
咨询：请问小孩一般什么时候出水痘? Consultation: When do children usually get chickenpox?	其他 Other	其他 Other	其他 Other	疾病表述 Disease description
老人收缩压160，舒张压只有40多，是什么原因？怎么治疗? The systolic blood pressure of the elderly is 160, and the diastolic blood pressure is only more than 40. What is the reason? How to treat?	病情诊断 Diagnosis	病情诊断 Diagnosis	病情诊断 Diagnosis	治疗方案 Treatment

Table 6: Case studies in KUAKE-QIC. We evaluate the performance of baselines with 3 sampled instances. The correlation between Query and Title is divided into 3 levels (0-2), which means '*poorly related or unrelated*', '*related*' and '*strongly related*'. BERT = BERT-base, BERT-ext = BERT-wwm-ext-base, MedBERT = PCL-MedBERT.

"Reduction (Rash can be reduced by the host producing specific anti-toxin antibodies.)", ROBERTA and PCL-MedBERT get different predictions. The reason is that there are medical terms like "antitoxin antibody". ROBERTA cannot recognize these tokens correctly, while PCL-MedBERT trained with a large amount of medical corpora can handle them well. Furthermore, PCL-MedBERT can extract entities like "deletion, translocation, inversion" from long sentences, which is difficult for other models to do so.
We also present several examples from KUAKE-QIC in Table 6. For the first example, we observe that both BERT and BERT-ext failed to get the intent label for the query "Please ask whether there's something wrong with a high lymphocyte ratio and a low neutrophil ratio?," whereas MedBERT could correctly predict this. Because "lymphocyte ratio" and "neutrophil ratio" are medical terms, and the general pre-trained language model needs to leverage domain knowledge to understand these phrases.
如表 5 和表 6 所示，与其它语言相比，在医学文本中汉语非常口语化。此外，汉语多义现象普遍，一个词的含义会随着语调的变化而变化，这通常会导致机器阅读的困惑和困难。综上所述，我们认为 CBLUE 任务并不容易解决，因为汉语具有独特的特性，并且需要开发更强大的模型。

结论

本文提出了一种中文生物医学语言理解评估 (CBLUE) 指标。我们对当前的 11 种语言表示模型在 CBLUE 上进行了评估，并分析了其结果。这些结果表明，最先进的模型处理某些更具挑战性的任务的能力是有限的。与英语基准如 GLUE / SuperGLUE 和 BLURB 相比，它们的模型性能已经达到了人类水平，而我们观察到对于中文生物医学语言理解而言，这还远远不够真实。

致谢

我们对匿名审稿人辛勤工作和善意的评论表示感谢。本研究由国家自然科学基金（U1813215，61876052）资助；广东省自然科学基金项目

（2019A1515011158），深圳市战略性新兴产业发展专项资金（20200821174109001）、工业和信息化部、国家卫生健康委员会"5G+医疗健康应用试点"项目（5G+罗湖医院集团：居民健康管理新模式探索）支持；郑州市重大科技专项（20XTZX11020）；浙江省自然科学基金（LGG22F030011）；宁波市自然科学基金（2021J190）以及永江人才引进计划（2021A-156-G）。

道德考虑

我们从拥有数据的组织那里获得了授权，并签署了协议，然后发布基准 CC BY-NC 4.0 许可证。所有收集的数据集均进行了匿名化处理，并由每个数据提供方的 IRB 委员会进行审查，以保护隐私。由于我们遵循真实世界的分布来收集数据，可能存在无法忽略的流行度偏见。

参考文献

Iz Beltagy，Kyle Lo 和 Arman Cohan。2019 年。Scib-ert：一种用于科学文本的预训练语言模型。在第 2019 届自然语言处理实验方法会议和第九届国际联合自然语言处理会议上发表，EMNLP-IJCNLP 2019，香港，中国，2019 年 11 月 3 日至 7 日，页码为 3613 - 3618 。计算语言学协会。

Alexis Conneau 和 Douwe Kiela。2018 年。Senteval：一种用于通用句子表示的评估工具包。在第十一届国际语言资源与评价会议，LREC 2018，日本宫崎市，2018 年 5 月 7 日至 12 日。欧洲语言资源协会（ELRA）。

崔一鸣，车万祥，刘婷，秦冰，王世金和胡国平。2020 年。重新审视中文自然语言处理的预训练模型。arXiv 预印本：2004.13922。

崔一鸣，车万祥，刘婷，秦冰，杨子清，王世金和胡国平。2019 年。用于中文 BERT 的整词掩码预训练。arXiv 预印本：1906.08101。

Jacob Devlin，Ming-Wei Chang，Kenton Lee 和 Kristina Toutanova。2018 年。Bert：用于语言理解的深度双向变压器预训练。在 NAACL-HLT 上发表。

Diao Shi-Zhe， Bai Jia-Xin， Song Yan， Zhang Tong 和 Wang Yong-Gang。2019 年。 Zen：通过 n-gram 表示增强的中文文本编码器预训练。arXiv 预印本 arXiv：1911.00720。

Joseph L Fleiss。1971。许多评估者之间名义尺度的一致性测量。心理公报，76（5）：378。

Timnit Gebru，Jamie Morgenstern，Briana Vecchione，Jennifer Wortman Vaughan，Hanna M. Wallach、Hal Daumé III 和 Kate Crawford。2018 年。数据集的数据表。CoRR，abs / 1803.09010。

Pieter Gijsbers，Erin LeDell，Janek Thomas，Sébastien Poirier，Bernd Bischl 和 Joaquin Vanschoren。 2019 年。开源自动机器学习基准。arXiv 预印本：1907.00909。

Yu Gu，Robert Tinn，Hao Cheng，Michael Lu-cas，Naoto Usuyama，Xiaodong Liu，Tristan Nau-mann，Jianfeng Gao 和 Hoifung Poon。2020 年。针对特定领域的语言模型预训练
生物医学自然语言处理。CoRR，abs/2007.15779。

关天，占海，周晓，徐浩，张凯。2020 年。CMeIE：中文医学信息抽取数据集的构建与评估。自然语言处理和中国计算第九届中国计算机学会国际会议，NLPCC 2020，郑州，中国，2020 年 10 月 14 日至 18 日，论文集，第一部分。
乔金，Bhuwan Dhingra，郑平刘，威廉 W。科恩和兴华卢。2019 年 PubMedQA：生物医学研究问答的数据集。在第 2019 届自然语言处理的实验方法会议和第九届国际联合会议上的程序文本处理，EMNLP-IJCNLP 2019，香港，中国，2019 年 11 月 3 日至 7 日，页码为 2567 - 2577。计算语言学协会。
Zhenzhong Lan，Mingda Chen，Sebastian Goodman，Kevin Gimpel，Piyush Sharma 和 Radu Soricut。2019 年。Albert：一种用于语言表示自监督学习的轻量级 BERT。arXiv 预印本 arXiv：1909.11942。
李赫基，金秀永，金炯伍和中庸忠。2017 年。用于健康数据发布的保留有用性的匿名化方法。BMC 医学信息学与决策支持，第 17 卷（1）：104：1-104：12。
金赫利，尹勇俊，金松东，金东铉，金顺圭，苏灿浩和康在宇。2020 年。Biobert：一种用于生物医学文本挖掘的预训练生物医学语言表示模型。Bioinformatics，36（4）：1234-1240。
Patrick Lewis，Myle Ott，Jingfei Du 和 Veselin Stoyanov。2020 年。用于生物医学和临床任务的预训练语言模型：理解并扩展最先进的技术。在第 3 届临床自然语言处理研讨会论文集上，页码为 146-157，在线发布。计算语言学协会。
李杰，孙悦平，约翰·罗宾·约翰逊，丹妮拉·斯卡伊基，魏志轩，莱曼，A.P.戴维斯，C.马廷利，托马斯·威格尔斯和朱志强。 2016 年。生物创造 v cdr 任务语料库：化学疾病关系提取资源。数据库：生物学数据库与收藏杂志，2016 年。
林帅，周潘，梁小丹，唐建恒，赵瑞慧，陈子良和林亮。 2020 年。低资源医疗对话生成的图进化元学习。CoRR，abs / 2012.11988。
刘文革，唐建恒，秦静慧，徐林，李振，梁晓丹。2020 年。MedDG：用于构建医疗对话系统的大型医疗咨询数据集。CoRR，abs / 2010.07497。
Yinhan Liu，Myle Ott，Naman Goyal，Jingfei Du，Man-dar Joshi，Danqi Chen，Omer Levy，Mike Lewis，Luke Zettlemoyer 和 Veselin Stoyanov。2019 年。Roberta：一种经过优化的 BERT 预训练方法。预印本 arXiv：1907.11692。
Bryan McCann，Nitish Shirish Keskar，Caiming Xiong 和 Richard Socher。2018 年。自然语言脱十项全能：多任务学习作为问题回答。CoRR，abs / 1806.08730。
Dimitris Pappas，Ion Androutsopoulos 和 Haris Pagelogeorgiou。2018 年。Bioread：一种新的生物医学阅读理解数据集。在第十一届国际语言资源和评估会议论文集中，LREC 2018，日本长崎市，2018 年 5 月 7 日至 12 日。欧洲语言资源协会（ELRA）。
Adam Paszke，Sam Gross，Francisco Massa，Adam Lerer，James Bradbury，Gregory Chanan，Trevor Killeen，Zeming Lin，Natalia Gimelshein，Luca Antiga 等人。2019 年。Pytorch：一种命令式风格、高性能

的深度学习库。在神经信息处理系统进展中，第 8024-8035 页。
塔蒂亚娜·沙维里纳，阿列娜·费诺格诺娃，安东·埃梅利扬诺夫，德尼斯·谢韦列夫，叶卡捷琳娜·阿尔莫托瓦，瓦连京·马利赫，弗拉基米尔·米哈伊洛夫，玛丽娅·季霍诺娃，安德烈·切尔托克和安德烈·耶夫兰皮耶夫。2020 年。俄罗斯超级胶水：一种俄语理解评估基准。在第 20 届自然语言处理实证方法会议上的论文集（EMNLP 2020），在线，2020 年 11 月 16 日至 20 日，页码为 4717-4726。计算语言学协会。
沈夏明，桂永浩。2013 年。临床儿科学第 2 版。人民卫生出版社。
乔治·塔萨拉尼斯，乔治斯·巴利卡斯，普罗多莫斯·马拉凯西奥斯，伊奥尼奥斯·帕特拉斯，马蒂亚斯·茨申克，迈克尔·阿尔弗尔斯，迪克·韦森博恩，安娜斯塔西亚·克拉里哈，塞吉奥斯·佩特里斯，季米特里斯·波利克龙普洛斯，雅尼斯·阿米尔安提斯，约翰·帕夫洛普洛斯，尼古拉斯·巴斯基奥蒂斯，帕特里克·加林纳里，蒂埃里·阿蒂埃尔，阿克塞尔-西尔维勒·恩戈诺·恩戈莫，诺曼·海诺，埃里克·高赛耶，莉莉安娜·巴拉约-阿尔弗尔斯，迈克尔·施罗德，伊昂·安德鲁托普洛斯和乔治斯·帕里乌拉斯。2015 年。大型生物医学语义索引与问答竞赛的概述。BMC 生物信息学，第 16 卷：138：1 - 138：28。
Alex Wang，Yada Pruksachatkun，Nikita Nangia，Amanpreet Singh，Julian Michael，Felix Hill，Omer Levy 和 Samuel R. Bowman。2019 年 a。Superglue：通用语言理解系统的一个更粘的基准。在第 32 届神经信息处理系统年度会议（Advances in Neural Information Processing Systems 32）上发表：2019 年神经信息处理系统大会(NeurIPS 2019)，2019 年 12 月 8 日至 14 日，加拿大不列颠哥伦比亚省温哥华市，页码为 3261 - 3275。
Alex Wang，Amanpreet Singh，Julian Michael，Felix Hill，Omer Levy 和 Samuel R. Bowman。2019 年 b。GLUE：自然语言理解的多任务基准和分析平台。在第 7 届国际学习表示会议，ICLR 2019，新奥尔良，美国路易斯安那州，2019 年 5 月 6 日至 9 日。OpenReview.net。
Lucy Lu Wang，Kyle Lo，Yoganand Chandrasekhar，Russell Reas，Jiangjiang Yang，Darrin Eide，Kathryn Funk，Rodney Kinney，Ziyang Liu，William Merrill，Paul Mooney，Dewey A. Murdick，Devvret Rishi，Jerry Sheehan，Zhihong Shen，Brandon Stillson，Alex D. Wade，Kuansan Wang，Chris Wilhelm，Boya Xie，Douglas Raymond，Daniel S. Weld，Oren Etzioni 和 Sebastian Kohlmeier。2020 年 CORD-19：COVID-19 开放研究数据集。CoRR，abs/2004.10706。
王玮平，宋坤，常丽文。2018 年版《儿科学》人民卫生出版社。
魏中宇，刘千龙，彭宝林，陶华晓，陈婷，黄璇晶，翁家培和戴向英。2018 年。面向任务的对话系统用于自动诊断。在第 56 届计算语言学协会年度会议论文集上发表，ACL 2018，墨尔本，澳大利亚，2018 年 7 月 15 日至 20 日，卷 2：短篇论文，页码为 201 - 207。计算语言学会。
吴 Y，罗瑞邦，梁 H，丁 H 和 LamT。 2019 年。 Renet：从文献中提取基因-疾病关联的深度学习方法。 在 RE-COMB 上发表。
梁旭，胡海，张轩威，李璐，曹晨杰，李玉东，徐业琛，孙凯，于典，余聪，田印天，董千茜，刘伟棠，石博，崔一鸣，李俊毅，曾军，王荣钊，谢卫建，李艳婷，李怡娜·帕特森，田左宇，张艺文，周和，刘少华，赵哲，赵启鹏，岳春月，张新瑞，杨正亮，理查德·基尔，兰振中。2020 年。CLUE：一个中文语

言理解评估基准。在第 28 届国际计算语言学会议（COLING 2020）上发表的论文，巴塞罗那，西班牙（在线），2020 年 12 月 8 日至 13 日，页码为 4762 - 4772。国际计算语言学委员会。
Yu K-H，Beam A L，Kohane I S。2018 年。医疗保健中的人工智能。自然生物医学工程，2（10）：719-731。
曾广涛，杨文绵，居泽乾，杨悦，王世成，张瑞思，周梦，曾嘉琪，董向宇，张若雨，方红超，朱鹏辉，陈书，谢鹏韬。2020 年。MedDialog：大规模医疗对话数据集。在第 20 届会议论文集中。
在自然语言处理中的经验方法，EMNLP 2020，在线，2020 年 11 月 16 日至 20 日，第 9241 至 9250 页。计算语言学协会。
胡宗，杨金轩，张泽宇，李佐峰，张晓燕。2021 年。使用机器学习方法对临床试验中中国入选标准进行语义分类。BMC 医学信息学决策支持，21（1）：128。

更广泛的影响

COVID-19（新型冠状病毒病）大流行对社会产生了重大影响，这不仅是因为 COVID-19 的严重健康后果以及为减缓其传播而实施的公共卫生措施。缺乏信息根本上导致了在爆发期间所经历的各种困难；试图解决这些需求造成了研究人员和公众的信息过载。生物医学自然语言处理——人工智能的一个分支，可以解释人类的语言 — 可以应用于解决许多由 COVID-19 大流行引起的紧迫信息需求。不幸的是，大多数语言基准都是英文的，并且目前还没有中文的生物医学基准。我们的基准 CBLUE 作为第一个中文生物医学语义理解基准，可作为模型评估的开放测试平台，促进该技术的进步。

B 负面影响

尽管我们要求领域专家和医生对所有语料进行标注，但仍存在一些实例的标注标签有误。如果模型是基于基准上的数字选择出来的，则可能会造成实际危害。此外，我们的基准降低了与生物医学数据合作的门槛。虽然这通常是一件好事，但它可能比已经存在的更稀释了生物医学领域的数据驱动工作池，并使专家难以发现相关的工作。

C 限制条件

尽管我们的 CBLUE 提供了多种设置，但仍有一些任务未被基准覆盖，例如医疗对话生成（Liu 等，2020；Lin 等，2020；Zeng 等，2020）或 医疗诊断 (Wei 等，2018）。我们鼓励学术界和工业界的研究人员贡献新的数据集。此外，我们的基准是
静态；因此，模型可能在任务上表现突出但失败于简单的挑战示例，并且在现实世界场景中表现不佳。我们将其作为未来的工作来构建一个平台，包括数据集创建、模型开发和评估，从而产生更稳健和信息丰富的基准。

D CBLUE 背景

标准数据集和共享任务在促进人工智能技术的发展中发挥了重要作用，以中国生物信息学社区为例，CHIP（中国健康信息技术）会议每年都会发布与医学相关的共享任务，这极大地促进了中文医学自然语言处理技术的进步。然而，在共享任务结束后，一些数据集不再可用，这引发了对数据获取以及未来研究的担忧。
近年来，借助预训练语言模型的帮助，我们可以在许多下游任务中获得最先进的性能。一个显著的趋势是出现了多任务排行榜，例如 GLUE（通用语言理解评估）和 CLUE（中文语言理解评估）。这些排行榜为公平的基准提供了吸引众多研究人员的关注，并进一步促进了语言模型技术的发展。例如，微软在 2020 年底发布了医疗领域的 BLURB （生物医学语言理解和推理评估），最近天池平台在 CHIP 学会的指导下推出了 CBLUE (中国生物医学语言理解评估) 公共基准。我们认为 CBLUE 的发布将进一步吸引研究者对医疗 AI 领域关注并促进社区发展。
CBLUE 1.015 包括 CHIP 会议的共享任务和阿里巴巴 QUAKE 搜索引擎的数据集，包括命名实体识别、信息提取、临床诊断规范化以及单句/句子对分类。

E 详细任务介绍

E.1 中国医学命名实体识别数据集（CMeEE）

任务背景：作为信息抽取的重要子任务，实体识别近年来取得了显著的成果。医学文本如教科书、百科全书、临床试验、医疗文献、电子健康记录和体检报告等包含丰富的医学知识。命名实体识别是将疾病症状等术语从上述无结构或半结构化文本中提取出来的过程，并且可以显著提高科学研究效率。CMeEE 数据集正是为了这个目的而提出的，原始数据集在 CHIP 2020 会议上发布。
任务描述：本任务定义为给定预定义的语义槽和输入句子，识别医疗实体并将其分类到 9 个类别中，包括疾病（dis）、临床症状(sym)、药物 (dru)、医疗器械(equ)、医疗程序(pro)、身体(bod)、医学检查项目(ite)、微生物(mic) 和部门(dep)。关于详细注释说明，请参阅 CBLUE 官方网站，并在表 7 中显示示例。
标注过程：由两家三甲医院的两位医学专家进行指导，优化了在试标注过程中。共参与标注人员 32 人，包括两位指南制定者也是医学专家、四位生物医学信息学领域专家、六位临床医生和二十二名计算机科学专业硕士研究生。历时三个月（从 2018 年 10 月到 2018 年 12 月），另外一个月时间用于校对。总费用约五万元人民币。
注释过程分为两个阶段。
• 第一阶段：这一阶段被称为试注释阶段。医疗专家对注释者进行了培训，以确保他们全面了解任务。注释者进行了两轮试注释，目的是熟悉注释任务以及

Entity type	Entity subtype	Label	Example
疾病 disease	疾病或综合症 disease or syndrome 中毒或受伤 poisoned or injured 器官或细胞受损 damage to organs or cells	dis	尿潴留者易继发泌尿系感染 Patients with urinary retention are prone to secondary infections of the urinary system.
临床表现 clinical manifestations	症状 symptom 体征 physical sign	sym	逐渐出现呼吸困难、阵发性喘憋，发作时呼吸快而浅，并伴有呼气性喘鸣，明显鼻扇及三凹征 Then dyspnea and paroxysmal asthma may occur, along with shortness of breath, expiratory stridor, obvious flaring nares, and three-concave sign.
医疗程序 medical procedure	检查程序 check procedure 治疗 treatment 或预防程序 or preventive procedure	pro	用免疫学方法检测黑种病原体的特异抗原很有诊断价值，因其简单快速，常常用于早期诊断，诊断意义常较抗体检测更为可靠 It is of great diagnostic value to detect the specific antigen of a certain pathogen with immunoassay a simple and quick assay that is intended for early diagnosis and proves more reliable than the antibody assay.

Table 7: Examples in CMeEE

表 7：CMeEE 示例

发现指南中不明确的地方，以及标注问题，并根据反馈对标注指南进行迭代改进。

• 第二阶段：对于第一阶段，每条记录由两名标注者独立进行标记，并且医学专家和生物医学信息学专家会及时提供帮助。通过开发的用于 CMeEE 和 CMeIE 任务的注释工具对注释结果自动比较，任何不一致都会被记录并移交到下一阶段。在第二阶段，医学专家与标注者就分歧记录以及其他注释问题进行了讨论，并由标注者做出修正。经过两个阶段后，IAA 分数（Kappa 分数）为 0.8537，满足要求。

研究目标。

PII 和 IRB，语料库来自授权的医学教科书或临床实践，并且文本中不包含任何个人可识别信息或冒犯性内容。

上述资源中不包含任何 PII。该数据集未涉及伦理问题，已由提供者的 IRB 委员会进行检查。

原始数据集格式为自定义的纯文本格式，为了简化数据预处理步骤，CBLUE 团队在获得数据提供方授权的情况下将数据格式转换成了统一的 JSON 格式。

评估指标 该任务使用严格的微 F1 度量。

数据集统计 这个任务有 15，000 个训练集数据、5，000 个验证集数据和 3，000 个测试集。
数据集包含 938 个文件和 47,194 句。每个文件平均含有 2,355 个单词，该数据集包含 504 种常见儿科疾病、7,085 个体部位、12,907 种临床症状以及 4,354 种医疗程序。
数据集提供者 数据集由以下人员提供：
- 教育部计算语言学重点实验室，北京大学
- 自然语言处理实验室，郑州大学，中国
- 中国鹏城实验室人工智能研究中心
- 哈尔滨工业大学，中国深圳

E.2 中国医学信息抽取数据集（CMeIE）
任务背景 实体和关系抽取是自然语言处理（NLP）与知识图谱 (KG) 中的一项重要信息提取任务，用于从非结构化文本中检测实体对及其关系。该技术可应用于医疗领域，例如，通过实体和关系抽取，可以将非结构化或半结构化的医学文本构建为医学知识图谱，并服务于众多下游任务。
Task Description Given a schema and sentence, where it defines the relationship (Predicate) and its associated subject and object, such as ("subject_type": "disease", "predicate": "drug treatment" "object_type": "drug"). The task requires the model to automatically analyze the sentence and then extract all triples = [(S1, P1, O1), (S2, P2, O2)...] from the sentence. Table 8 shows some examples in the dataset, and there are 53 schemas including 10 types of genus relations and 43 other sub-relations. Details can be found in the 53_schema.json file. For more detailed annotation instructions, please refer to the CBLUE official website, and examples are shown in Table 8.
注释过程 注释指南由两家三级甲等医院的两位医学专家编写，并在试注过程中进行了优化。共有 20 名注释员参与了注释过程，包括两名
参与标注的专家均为指南编写组成员，其中生物医学信息学领域 2 人、临床医生 6 名（含主任医师 3 名）、计算机专业硕士研究生 14 名。标注工作持续约四个月(从 2018 年 10 月到 2018 年 12 月)，包含标注时间及审核时间。总费用约为人民币 4 万元。
与 CMeEE 数据集类似，CMeIE 的数据标注过程也包含轨迹标注阶段和正式标注阶段，并且遵循相同的过程。此外，为这个数据集增加了一个额外的步骤——中文分词验证步骤。数据提供者开发了一种用于医学文本的分词工具，该工具可以生成段落以及 POS 标记，并且一些特定的 POS 类型（如"疾病"、"药物"）可以帮助自动验证是否存在潜在的缺失命名实体，这有助于协助标注人员检查遗漏标签。最终的 IAA 值为 0.83，能够满足研究目的。
PII 和 IRB，语料库来自授权的医学教科书或临床实践，并且文本中不包含任何个人可识别信息或冒犯性内容。
上述资源中不包含任何 PII。该数据集未涉及伦理问题，已由提供者的 IRB 委员会进行检查。
评估指标：参与者提供的 SPO 结果需要准确匹配。使用严格的 Micro-F1 进行评估。
数据集统计：该任务有 14，339 个训练集数据、3，585 个验证集数据和 4，482 个测试集数据。 数据集来自儿童语料库和常见疾病语料库。 儿童语料库来源于 518 种儿科疾病，而常见疾病语料库则来源于 109 种常见疾病。 数据集包

含近 75，000 条三元组、28，000 条疾病句子以及 53 个模式。
数据集提供者 数据集由以下人员提供：
- 教育部计算语言学重点实验室，北京大学
- 自然语言处理实验室，郑州大学，中国

Relation type	Relation subtype	Example
疾病_其他 disease_other	预防 prophylaxis	{'predicate': '预防-prevention', 'subject': '麻风病-Leprosy', 'subject_type': '疾病-disease', 'object': '利福-rifampicin', 'object_type': '其他-others'}
	阶段 phase	{'predicate': '阶段-phase', 'subject': '肿瘤-tumor', 'subject_type': '疾病-disease', 'object': 'I期-phase_', 'object_type': '其他-others'}
	就诊科室 treatment department	{'predicate': '就诊科室-treatment_department', 'subject': '腹主动脉瘤-abdominal_aortic_aneurysm', 'subject_type': '疾病-disease', 'object': '初级医疗保健医处-primary_medical_care_clinic', 'object_type': '其他-others'}
疾病_其他治疗 disease_other treatment	辅助治疗 adjuvant therapy	{'predicate': '辅助治疗-adjuvant_therapy', 'subject': '皮肤鳞状细胞癌-utaneous_squamous_cell_carcinoma', 'subject_type': '疾病-disease', 'object': '非手术破坏-non_surgical_destructio', 'object_type': '其他治疗-other_treatment'}
	化疗 chemotherapy	{'predicate': '化疗-chemotherapy', 'subject': '肿瘤-tumour', 'subject_type': '皮肤鳞状细胞癌-cutaneous_squamous_cell_carcinoma', 'object': '局部化疗-local_chemotherapy', 'object_type': '其他治疗-other_treatment'}
	放射治疗 radiotherapy	{'predicate': '放射治疗-radiation_therapy', 'subject': '非肿瘤性疼痛-non_cancer_pain', 'subject_type': '疾病-disease', 'object': '外照射-external_irradiation', 'object_type': '其他治疗-other_treatment'}
疾病_手术治疗 disease_surgical treatment	手术治疗 surgical treatment	{'predicate': '手术治疗-surgical_treatment', 'subject': '皮肤鳞状细胞癌-cutaneous_squamous_cell_carcinoma', 'subject_type': '疾病-disease', 'object': '传统手术切除-surgical_resection(traditional_therapy)', 'object_type': '手术治疗-surgical_treatment'}

Table 8: Examples in CMeIE

表 8：CMeIE 中的示例
• 中国鹏城实验室人工智能研究中心
• 哈尔滨工业大学，中国深圳

E.3 CHIP- 临床诊断正常化数据集（CHIP-CDN）

任务背景 临床术语规范化是研究和工业应用中一个至关重要的任务。在临床方面，可能有数百个不同的同义词来描述相同的诊断、症状等。
或流程；例如，"心肌梗死"和"MI"都代表标准术语"急性心肌梗死"。本任务的目标是找到给定临床术语的标准短语（即 ICD 代码）。借助标准代码，可以帮助研究人员减轻统计分析临床试验的负担；同时也可以帮助保险公司进行 DRG 或 DIP 相关应用。该任务是为了这个目的而提出的，并且原始共享任务在 CHIP 2020 会议上发布。
任务描述：本任务旨在标准化中国电子病历最终诊断中的术语。在最终诊断中不涉及隐私信息，给定原始术语，需要预测其对应的 ICD-10 北京临床版 v601 标准词汇表的标准短语。详细注释说明请参考 CBLUE 官方网站。示例见表 9。
标注过程：医渡云医疗团队对 CHI-P-CDN 进行了标注，所有人员均具有医学背景和临床医师资格证书。此项工作历时约两个月，由于是内部员工完成的，预计成本在十万元左右。
中国诊断标准化数据集（CHIP-CDN）由一轮标注、一轮全审核和一轮随机质量抽检完成，标注及审核工作由具有临床资质的普通标注人员完成，随机质量抽检由高级术语专家完成。
PII 和 IRB，该语料库来自 EMR（电子病历），仅选择最终诊断部分进行研究。数据集不涉及伦理问题。
如示例表所示，最终诊断不包含任何 PII。
原始数据集格式为自定义的 xlsx 格式，为了统一数据预处理步骤，CBLUE 团队在获得数据提供方授权的情况下将数据格式转换成了 JSON 格式。
评估指标 F1 分数是计算的（原始诊断术语，标准短语）
对，如果测试集有 m 个金标准对，预测结果有 n 个对，其中 k 个对被正确预测，则：
数据集统计，提供有 8,000 个训练实例和 10,000 个测试实例。我们把原始的训练集分成 6,000 个用于训练集、2,000 个用于验证集。
数据集提供者：该数据集由易度云技术有限公司提供。

E.4 临床试验标准数据集（CHIP-CTC）

任务背景：临床试验是指由人类志愿者进行的科学研究，以确定药物或治疗方法的有效性、安全性和副作用。它在促进医学发展和改善人类健康方面发挥着关键作用。根据实验目的，受试者可能是患者或健康的志愿者。本任务的目标是预测一个受试者是否符合一项临床试验的要求。招募临床试验的受试者通常通过手动比较医疗记录和临床试验筛选标准来进行，这需要花费大量的时间和精力，并且效率低下。近年来，在许多生物医学应用中，基于自然语言处理的方法取得了成功。因此，我们提出了一种自动分类中文临床试验资格标准的任务，并于 CHIP2019 会议上发布了原始数据集。所有数据均来自中国临床试验

注册中心（ChiCTR）网站上收集的真实临床试验，这是一个非营利组织，为公共研究提供登记服务。每个
任务描述：为本任务定义了总共 44 个预定义的语义类别，目标是预测给定文本到正确的类别。有关详细注释说明，请参阅 CBLUE 官方网站。标记数据示例如表 10 所示。
注释过程 CHIP-CTC 语料库由三位标注者进行标注。第一位标注者是李佐峰，他是飞利浦中国研究的首席科学家，在研究方面有十多年的经验
原始条款

Original terms	Normalization terms
右肺结节转移可能大 Possible nodule metastasis in the right lung	肺占位性病变## Space-occupying Lesion of the Lung 肺继发恶性肿瘤## Secondary Malignant Neoplasm of the Lung 转移性肿瘤 Metastatic Tumor
右肺结节住院 Hospitalization after detection of nodules in the right lung	肺占位性病变 Space-occupying Lesion of the Lung
左上肺胸膜下结节待查 Subpleural nodule in the left upper lung to be examined	胸膜占位 Space-occupying Lesion within the Pleural Space

Table 9: Examples in CHIP-CDN

ID	Clinical trial sentence	Category
S1	年龄>80岁 Age: > 80	Age
S2	近期颅内或椎管内手术史 Recent intracranial/intraspinal surgery	Therapy or Surgery
S3	血糖<2.7mmol/L Blood glucose < 2.7 mmol/L	Laboratory Examinations

Table 10: Examples in CHIP-CTC

在生物医学领域的工作经验。其他标注者是来自同济大学的生物医学信息学领域的博士生张泽宇和杨金轩。该标注工作于 2019 年 7 月开始，大约持续了 1 个月。此外，语料库被用于 CHIP2019 共享任务中。该标注与标注者的科研项目相关，并不需要支付报酬。
一位经验丰富的生物医学研究人员（ZL）和两位生物医学领域的人工智能博士生（ZZ、JY）对 CHIP-CTC 语料库中的 44 个类别进行了标注。首先，他们研究

了这些类别的定义，并调查了大量的标准句表达模式，然后选择每个类别的标准例句；其次，两名评估者独立地为同一 1000 条句子进行标注，之后检查并讨论与 ZL 的矛盾之处直到达成共识。这个步骤重复了 20 次迭代，共标注了 20000 条标准句，随后用于构建模型。
计算了标注者间的一致性得分（Cohen's kappa 为 0.9920）。最后，剩余的 18341 个句子被分配给两个标注者进行标注。
PII 和 IRB 该语料库来自中国临床试验注册中心（ChiCTR）网站，这是一个非营利组织为公众研究提供登记的机构。对于本网站上每个已注册的临床试验案例，它已经由组织伦理委员会批准。此外，注释和语料库还经过飞利浦内部生物医学实验委员会 (ICBE) 的审查和批准。鼓励使用此语料库进行学术研究。对于每个注册的临床试验报告，不包括 PII。
原始数据集格式为自定义的 csv 格式，为了统一数据预处理步骤，CBLUE 团队在获得数据授权的情况下将数据格式转换成了 JSON 格式
提供者。
评估指标 本任务的评估使用 Macro-F1。假设我们有 n 个类别，C1,…,Ci,…,Cn。准确率 Pi 是记录正确预测到类 Ci 的数量/预测为类 Ci 的记录数量。召回率 Ri = 记录中正确预测为类 Ci 的数量 / 实际 Ci 类别的记录数。
数据集统计：该任务有 22，962 个训练样本、7，682 个验证样本和 10，000 个测试样本。
数据集提供者：该数据集由同济大学生命科学与技术学院和飞利浦中国研究院提供。

E.5 语义文本相似性数据集（CHIP-STS）

任务背景 CHIP-STS 任务旨在基于中文在线医疗问题学习不同疾病类型之间的相似知识。具体来说，给定来自 5 种不同疾病的句子对，需要确定两个句子的语义是否相似或不相似。原共享任务在 CHIP2019 会议上发布。
任务描述：类别代表疾病类型的名称，包括糖尿病、高血压、肝炎、艾滋病和乳腺癌。标签表示问题的语义是否相同。如果相同，则标记为 1，否则标记为 0。示例标注如表 11 所示。
标注过程：CHIP-STS 语料库由五名医学本科学生在一名外科医生和一名内科医生的指导下进行标注，由于是二分类任务，所以标注工作相对简单；标注过程以及验证时间持续两周。共标注了 3 万句对子，花费标注费用 2.5 万元人民币。
有五种疾病，所以每个注释者被分配了两种疾病的标签来保证每一种疾病都有两个评估员进行标注。在试用期间的注释
过程，每个标注者被分配了 100 条记录进行标记，旨在测试他们是否能够彻底理解任务。随后，标注者开始对过程进行标记，并且医学专家会提供必要的帮助，例如解释疾病机制以协助评估人员。最后，每一条记录由两名不同的标注者进行标记，对于存在分歧的对子则选择讨论和案例研究；根据专家反馈，标注者重新检查之前标记的结果。一致性系数为 0.93。
PII 和 IRB，该语料库来自医疗论坛的在线问题，并且不涉及伦理学方面的问题，这些已经由提供者的 IRB 委员会进行了检查。
在注释步骤中，标注人员手动丢弃包含 PHI 信息的句子。CBLUE 团队还逐条验

证了数据集以确保没有 PII 包括其中。
原始数据集格式为自定义的 csv 格式，为了统一数据预处理步骤，CBLUE 团队在获得数据提供方授权的情况下将数据格式转换成了 JSON 格式。
评估指标：该任务的评估为 Macro-F1。
数据集统计，该任务有 16,000 个训练集、4,000 个验证集和 10,000 个测试集的数据。
数据集提供者：该数据集由平安科技提供。

E.6 KUAKE-查询意图分类数据集（KUAKE-QIC）

任务背景 在医疗搜索场景中，理解查询意图可以显著提高搜索结果的相关性。特别是医学知识高度专业，分类查询意图也可以帮助整合医学知识以提升搜索结果的表现。本任务为此目的而提出。
任务描述 医疗意图标签有 11 个类别，包括诊断、病因分析、治疗计划、医疗建议、测试结果分析、疾病描述、后果预测、注意事项、预期效果、治疗费用和其他。对于详细的注释说明...
请参见 CBLUE 官方网站。示例如表 12 所示。
标注过程：KUAKE-QIC 语料库由六位来自医学院的毕业生进行标注，他们被阿里巴巴聘为全职员工。在开始标注之前，他们必须通过指定任务测试。这个任务大约需要两周时间，标注费用是 6,600 元人民币，有 22,000 个标记记录，也就是说每条记录 0.3 元人民币。
注释过程分为三步：
第一步是轨迹标注步骤；本阶段选取了 2000 条记录。标注人员分为两组，每组三人。数据提供方对质量控制有严格的标准，比如同一组的三个人之间的 IAA 必须超过 0.9。
第二阶段是正式注释阶段，在此阶段，6 名标注员被分为三组，每组两人。总共对 20,000 条记录进行了注释；该步骤的 IAA 为 0.9230。
最后一步是质量控制步骤，采用抽样策略，抽取了 300 条记录进行验证；医疗专家提出了部分常见注释问题，并在批量模式下对数据进行了修正。此外，一些有争议的案例由医疗专家最终决定。
PII 和 IRB，语料库来自 KUAKE 搜索引擎用户查询，并且不涉及伦理问题，已经由提供商的 IRB 委员会进行了检查。
在注释步骤中，带有 PHI 信息或冒犯性信息（例如性问题）的句子将由标注者手动丢弃。
该数据集也通过了阿里巴巴的数据披露流程。
CBLUE 团队还逐条验证了数据集，以确保其中不包含任何个人身份信息。
评估指标准确度用于此任务的评估。
数据集统计：该任务有 6，931 个训练集数据、1，955 个验证集数据和 1，994 个测试集数据。
数据集提供者：该数据集由阿里巴巴地震搜索引擎提供。

E.7 KUAKE- 查询标题相关性数据集（KUAKE-QTR）

任务背景 KUAKE 查询标题相关性是一个用于查询文档（标题）相关性的数据集。例如，给定查询"维生素 B 缺乏的症状"，相关的标题应该是"维生素 B 缺乏的主要表现"。
任务描述：查询与标题的相关性分为四个等级（0-3），其中 0 表示最差，3 表示最佳匹配。详细注释说明请参考 CBLUE 官方网站。示例如表 13 所示。
标注过程：KUAKE-QTR 语料库由九名标注员进行标注，其中七人来自第三方众包医学本科学生，两人来自阿里全职员工。众包标注员需要经过培训并通过标注测试后才能执行任务。标注持续了两周，共花费 28,000 元人民币。

Intent	Sentences
病情诊断 disease diagnosis	最近早上起来浑身无力是怎么回事? Why do I always feel weak after I get up in the morning? 我家宝宝快五个月了，为什么偶尔会吐清水带? Why does my 5-month-old baby occasionally vomit clear liquid?
注意事项 precautions	哮喘应该注意些什么 What should patients with asthma pay attention to? 孕妇能不能吃榴莲 Can a pregnant woman eat durians? 柿子不能和什么一起吃 Which food cannot be eaten together with persimmons? 糖尿病人饮食注意什么啊? What should patients with diabetes pay attention to about their diet?
就医建议 medical advice	糖尿病该做什么检查? What examination should patients with diabetes receive? 肚子疼去什么科室? Which department should patients with stomachache visit?

Table 12: Examples in KUAKE-QIC

Query	Title	Level
缺维生素b的症状 Symptoms of Vitamin B deficiency	维生素b缺乏症的主要表现 What are the major symptoms of Vitamin B deficiency?	3
大腿软组织损伤怎么办 How can I treat a soft tissue injury in the thigh?	腿部软组织损伤怎么办 What's the treatment for a soft tissue injury in the leg?	2
小腿抽筋是什么原因引起的 What causes lower leg cramps?	小腿抽筋后一直疼怎么办 How can I treat pains caused by lower leg cramps?	1
挑食是什么原因造成的 What is the cause of picky eating?	挑食是什么原因造成的 What is the cause of picky eating?	0

Table 13: Examples in KUAKE-QTR

与 KUAKE-QIC 任务类似，KUAKE-QTR 注释过程分为三个步骤，略有变化：
培训和考试阶段：七位标注员由两位全职专家进行任务讲解，然后每人分配了 200 条数据进行标注，这些数据都经过了全职专家的真值标注。标注精度必须达到 85%以上才能通过测试。
第二步是正式标注步骤，每个标注者被分配了 3000 条记录进行标记，在其中的 100 条中带有金标准。
标注工具会自动比较标注者与金标准之间的标签，如果需要帮助，则会提供帮助。只有精度超过阈值 0.85 的才会进入下一轮。
最后一步是质量控制步骤，采用抽样策略，由 FTE 医学专家对 100 条记录进行验证；有问题的案例会返回给众包标注者进行修正。
PII 和 IRB，语料库来自 KUAKE 搜索引擎用户查询，并且不涉及伦理问题，已经由提供商的 IRB 委员会进行了检查。
在标注步骤中，带有 PHI 信息或冒犯性信息（如性问题）的句子由注释者手动丢弃。该数据集还通过了阿里巴巴的数据披露过程。
CBLUE 团队还逐个验证了数据集，以确保没有包含任何个人身份信息。经提供方同意后，已删除了一条带有空标签的记录。
评估指标与 KUAKE-QIC 任务相同，使用准确率对本任务进行评价。
数据集统计：该任务有 24，174 个训练集数据、2，913 个验证集数据和 54，65 个测试集数据。
数据集提供者：该数据集由阿里巴巴地震搜索引擎提供。

E.8 KUAKE- 查询查询相关性数据集（KUAKE-QQR）

任务背景：KUAKE Query-Query 相关性是一个数据集，用于评估两个给定查询之间的相关性以解决搜索引擎的长尾挑战。与 KUAKE-QTR 类似，查询-查询相关性是现实世界搜索引擎中一个重要的且具有挑战性的任务。
任务描述：查询与标题的相关性分为三个等级（0-2），其中 0 表示最差，2 表示最佳相关。关于详细注释说明，请参阅 CBLUE 官方网站。示例见表 14。
注释过程与 KUAKE-QTR 相同，但费用总计为 2.2 万元。

PII 和 IRB 与 KUAKE-QTR 相同。
评估指标与 KUAKE-QIC 和 KUAKE-QTR 任务相同，使用准确率作为评估指标。
数据集统计：该任务有 15,000 个训练集数据，1,600 个验证集数据和 1,596 个测试集数据。
数据集提供者：该数据集由阿里巴巴地震搜索引擎提供。

F 实验细节

本节详细介绍了每个数据集的训练过程和超参数。我们使用 Pytorch 进行实验，所有运行超参数如表所示。CMeIE 有两个阶段：实体识别（CMeEE-ER）和关系分类（CMeEE-RE）。因此，我们详细说明了 CMeEE-ER 和 CMeEE-RE 中的超参数要求。

- python3
- PyTorch 1.7
- 变形金刚 4.5.1
- jieba
- gensim

特定任务的超参数如表 15-26 所示。

其他任务的 G 错误分析

我们介绍错误定义如下，并在表 27 至 32 中说明其他任务的一些错误情况。
歧义表示实例具有相似的上下文但不同的含义，这会误导预测。
需要领域知识表明实例中存在需要领域知识才能理解的生物医学术语。
需要句法知识表明实例中存在复杂的句法结构，模型无法理解正确的含义实体重叠表示存在多个
实例中的重叠实体。
长序列表示输入实例非常长。
注释错误表示标注的标签是错的。
错误的实体边界表示实例具有错误的实体边界。
稀有词表示实例中存在低频词。
多个触发器表示存在多个误导预测的指示词。
俚语（在搜索查询中非常常见）表明实例相当不同

Query	Query	Level
小孩子打呼噜是什么原因引起的 What causes children's snoring	小孩子打呼噜什么原因 What makes children snore?	2
双眼皮遗传规律 Heredity laws of double-fold eyelids	内双眼皮遗传 Heredity of hidden double-fold eyelids	1
白血病血常规有啥异常 What index of the CBC test will be abnormal for patients with leukemia?	白血病血检有哪些异常 What index of the blood test will be abnormal for patients with leukemia?	0

Table 14: Examples in KUAKE-QQR

Method	Value
warmup_proportion	0.1
weight_decay	0.01
adam_epsilon	1e-8
max_grad_norm	1.0

Table 15: Common hyper-parameters for all CBLUE tasks

从书面语言（例如，有许多缩写）中得出的结论，因此挑战了预测模型。
无关描述表明实例包含大量与预测无关的信息，从而误导了预测。

贡献

浙江大学、知识引擎联合实验室、杭州创新中心的张宁宇，毕振，梁小转和李磊共同撰写了该论文。
阿里巴巴集团陈莫沙、谭传奇，黄飞和罗思以及清华大学统计科学中心郑元共同贡献了 CBLUE 基准排行榜，并将自定义数据格式转换为统一的 JSON 格式。
来自郑州大学信息工程学院、鹏城实验室的张坤利和北京大学计算语言学教育部重点实验室、鹏城实验室的常宝宝贡献了 CMeEE 数据集。
来自郑州大学信息工程学院、鹏城实验室的张红英和北京大学计算语言学教育部重点实验室、鹏城实验室的苏志芳贡献了 CMeIE 数据集。
中国北京易度云技术有限公司的李林峰、严军贡献了 CHIP-CDN 数据集。
同济大学生命科学与技术学院胡宗和飞利浦研究中国贡献了 CHIP-CTC 数据集。

Model	epoch	batch_size	max_length	learning_rate
bert-base	5	32	128	4e-5
bert-wwm-ext	5	32	128	4e-5
roberta-wwm-ext	5	32	128	4e-5
roberta-wwm-ext-large	5	12	65	2e-5
roberta-large	5	12	65	2e-5
albert-tiny	10	32	128	5e-5
albert-xxlarge	5	12	65	1e-5
zen	5	20	128	4e-5
macbert-base	5	32	128	4e-5
macbert-large	5	12	80	2e-5
PCL-MedBERT	5	32	128	4e-5

Table 16: Hyper-parameters for the training of pre-trained models with a token classification head on top for named entity recognition of the CMeEE task.

表 16：在 CMeEE 任务中，对命名实体识别进行预训练模型的 token 分类头上的训练时使用的超参数。

Model	epoch	batch_size	max_length	learning_rate
bert-base	7	32	128	5e-5
bert-wwm-ext	7	32	128	5e-5
roberta-wwm-ext	7	32	128	4e-5
roberta-wwm-ext-large	7	16	80	4e-5
roberta-large	7	16	80	2e-5
albert-tiny	10	32	128	4e-5
albert-xxlarge	7	16	80	1e-5
zen	7	20	128	4e-5
macbert-base	7	32	128	4e-5
macbert-large	7	20	80	2e-5
PCL-MedBERT	7	32	128	4e-5

Table 17: Hyper-parameters for the training of pre-trained models with a token-level classifier for subject and object recognition of the CMeIE task.

Model	epoch	batch_size	max_length	learning_rate
bert-base	8	32	128	5e-5
bert-wwm-ext	8	32	128	5e-5
roberta-wwm-ext	8	32	128	4e-5
roberta-wwm-ext-large	8	16	80	4e-5
roberta-large	8	16	80	2e-5
albert-tiny	10	32	128	4e-5
albert-xxlarge	8	16	80	1e-5
zen	8	20	128	4e-5
macbert-base	8	32	128	4e-5
macbert-large	8	20	80	2e-5
PCL-MedBERT	8	32	128	4e-5

Table 18: Hyper-parameters for the training of pre-trained models with a classifier for the entity pairs relation prediction of the CMeIE task.

表 18：用于训练具有实体对关系预测分类器的预训练模型的超参数。

Model	epoch	batch_size	max_length	learning_rate
bert-base	5	32	128	5e-5
bert-wwm-ext	5	32	128	5e-5
roberta-wwm-ext	5	32	128	4e-5
roberta-wwm-ext-large	5	20	50	3e-5
roberta-large	5	20	50	4e-5
albert-tiny	10	32	128	4e-5
albert-xxlarge	5	20	50	1e-5
zen	5	20	128	4e-5
macbert-base	5	32	128	4e-5
macbert-large	5	20	50	2e-5
PCL-MedBERT	5	32	128	4e-5

Table 19: Hyper-parameters for the training of pre-trained models with a sequence classification head on top fo screening criteria classification of the CHIP-CTC task.

Param	Value
recall_k	200
num_negative_sample	10

Table 20: Hyper-parameters for the CHIP-CDN task. We model the CHIP-CDN task with two stages: recall stage and ranking stage. *num_negative_sample* sets the number of negative samples sampled for the training ranking model during the ranking stage. *recall_k* sets the number of candidates recalled in the recall stage.

表 20：CHIP-CDN 任务的超参数。我们使用两个阶段来建模 CHIP-CDN 任务，即召回阶段和排序阶段。num_negative_sample 设置在排序阶段训练排名模型时使用的负样本数量。recall_k 设置召回阶段召回候选者的数量。

Model	epoch	batch_size	max_length	learning_rate
bert-base	3	32	128	4e-5
bert-wwm-ext	3	32	128	5e-5
roberta-wwm-ext	3	32	128	4e-5
roberta-wwm-ext-large	3	32	40	4e-5
roberta-large	3	32	40	4e-5
albert-tiny	3	32	128	4e-5
albert-xxlarge	3	32	40	1e-5
zen	3	20	128	4e-5
macbert-base	3	32	128	4e-5
macbert-large	3	32	40	2e-5
PCL-MedBERT	3	32	128	4e-5

Table 21: Hyper-parameters for the training of pre-trained models with a sequence classifier for the ranking model of the CHIP-CDN task. We encode the pairs of the original term and standard phrase from candidates recalled during the recall stage and then pass the pooled output to the classifier, which predicts the relevance between the original term and standard phrase.

表 21：CHIP-CDN 任务中排序模型的预训练模型进行训练时序列分类器超参数。我们编码候选人在召回阶段检索到的原始术语和标准短语对，然后将池化输出传递给分类器，该分类器预测原始术语与标准短语之间的相关性。

Model	epoch	batch_size	max_length	learning_rate
bert-base	20	32	128	4e-5
bert-wwm-ext	20	32	128	5e-5
roberta-wwm-ext	20	32	128	4e-5
roberta-wwm-ext-large	20	12	40	4e-5
roberta-large	20	12	40	4e-5
albert-tiny	20	32	128	4e-5
albert-xxlarge	20	12	40	1e-5
zen	20	20	128	4e-5
macbert-base	20	32	128	4e-5
macbert-large	20	12	40	2e-5
PCL-MedBERT	20	32	128	4e-5

Table 22: Hyper-parameters for the training of pre-trained models with a sequence classifier for the prediction of the number of standard phrases corresponding to the original term in the CHIP-CDN task.

Model	epoch	batch_size	max_length	learning_rate
bert-base	3	16	40	3e-5
bert-wwm-ext	3	16	40	3e-5
roberta-wwm-ext	3	16	40	4e-5
roberta-wwm-ext-large	3	16	40	4e-5
roberta-large	3	16	40	2e-5
albert-tiny	3	16	40	5e-5
albert-xxlarge	3	16	40	1e-5
zen	3	16	40	2e-5
macbert-base	3	16	40	3e-5
macbert-large	3	16	40	3e-5
PCL-MedBERT	3	16	40	2e-5

Table 23: Hyper-parameters for the training of pre-trained models with a sequence classifier for sentence similarity predication of the CHIP-STS task.

Model	epoch	batch_size	max_length	learning_rate
bert-base	3	16	50	2e-5
bert-wwm-ext	3	16	50	2e-5
roberta-wwm-ext	3	16	50	2e-5
roberta-wwm-ext-large	3	16	50	2e-5
roberta-large	3	16	50	3e-5
albert-tiny	3	16	50	5e-5
albert-xxlarge	3	16	50	1e-5
zen	3	16	50	2e-5
macbert-base	3	16	50	3e-5
macbert-large	3	16	50	2e-5
PCL-MedBERT	3	16	50	2e-5

Table 24: Hyper-parameters for the training of pre-trained models with a sequence classifier for query intention prediction of the KUAKE-QIC task.

表 24：用于 KUAKE-QIC 任务的查询意图预测序列分类器预训练模型的超参数。

Model	epoch	batch_size	max_length	learning_rate
bert-base	3	16	40	4e-5
bert-wwm-ext	3	16	40	2e-5
roberta-wwm-ext	3	16	40	3e-5
roberta-wwm-ext-large	3	16	40	2e-5
roberta-large	3	16	40	2e-5
albert-tiny	3	16	40	5e-5
albert-xxlarge	3	16	40	1e-5
zen	3	16	40	3e-5
macbert-base	3	16	40	2e-5
macbert-large	3	16	40	2e-5
PCL-MedBERT	3	16	40	3e-5

Table 25: Hyper-parameters of training the sequence classifier for the KUAKE-QTR task.

表 25：训练 KUAKE-QTR 任务的序列分类器时使用的超参数。

Model	epoch	batch_size	max_length	learning_rate
bert-base	3	16	30	3e-5
bert-wwm-ext	3	16	30	3e-5
roberta-wwm-ext	3	16	30	3e-5
roberta-wwm-ext-large	3	16	30	3e-5
roberta-large	3	16	30	2e-5
albert-tiny	3	16	30	5e-5
albert-xxlarge	3	16	30	3e-5
zen	3	16	30	2e-5
macbert-base	3	16	30	2e-5
macbert-large	3	16	30	2e-5
PCL-MedBERT	3	16	30	2e-5

Table 26: Hyper-parameters of training the sequence classifier for the KUAKE-QQR task.

Sentence	Golden	RO	ME
另一项研究显示，减荷鞋对内侧膝骨关节炎也没有效。 Another study showed that load-reducing shoes were not effective for medial knee osteoarthritis.	内侧膝骨关节炎|辅助治疗|减荷鞋 medial knee osteoarthritis, adjuvant therapy, load-reducing shoes	膝骨关节炎|辅助治疗|减荷鞋 medial knee osteoarthritis, adjuvant therapy, load-reducing shoes	膝骨关节炎|辅助治疗|减荷鞋 medial knee osteoarthritis, adjuvant therapy, load-reducing shoes
精神疾病：焦虑和抑郁与失眠症高度相关。 Mental illness: anxiety and depression are related to insomnia.	焦虑|相关（导致）|失眠症 anxiety, related cause, insomnia	无|无|无 None|None|None	焦虑|相关（导致）|失眠症 anxiety, related cause, insomnia
在狂犬病感染晚期，患者常出现昏迷。 In the late stage of rabies infection, patients often appear comatose.	狂犬病|相关（转化）|昏迷 rabies, transform, comatose	无|无|无 None|None|None	无|无|无 None|None|None

Table 27: Error cases in CMeIE. We evaluate roberta-wwm-ext and PCL-MedBERT on 3 sampled sentences, with their gold labels and model predictions. Each label consists of subject | predicate | Object. None means that the model fails to predict. RO = roberta-wwm-ext, MB = PCL-MedBERT.

表 27：CMeIE 中的错误案例。我们对 3 个采样句子评估 roberta-wwm-ext 和 PCL-MedBERT，以及它们的黄金标签和模型预测结果。每个标签由主体|谓词|对象组成。None 表示模型无法预测。RO = roberta-wwm-ext，MB = PCL-MedBERT。

Sentence	Label	RO	MB
右第一足趾创伤性足趾切断 Right first toe traumatic toe cutting	单趾切断 Single toe cut	足趾损伤 Toe injury	单趾切断 Single toe cut
C3-4脊髓损伤 C3-4 spinal cord injury	颈部脊髓损伤 Neck spinal cord injury	脊髓损伤 Spinal cord injury	脊髓损伤 Spinal cord injury
肿瘤骨转移胃炎 Tumor bone metastatic gastritis	骨继发恶性肿瘤##转移性肿瘤##胃炎 Junior malignant tumor##Metastatic tumor##Gastritis	反流性胃炎##转移性肿瘤##胃炎 Reflux gastritis##Metastatic tumor##Gastritis	骨盆部肿瘤##转移性肿瘤##胃炎 Pelvic tumor##Metastatic tumor##Gastritis

Table 28: Error cases in CHIP-CDN. We evaluate roberta-wwm-ext and PCL-MedBERT on 3 sampled sentences with their gold labels and model predictions. There may be multiple predicted values, separated by a "##". RO = roberta-wwm-ext, MB = PCL-MedBERT.

Sentence	Label	RO	MB
既往多次行剖腹手术或腹腔广泛粘连者 Previous multi-time crashed surgery or abdominal adhesive	含有多类别的语句 Multiple	治疗或手术 Therapy or Surgery	治疗或手术 Therapy or Surgery
术前认知发育筛查（DST）发现发育迟缓 Preoperative cognitive development screening test(DST) finds development slow	诊断 Diagnostic	疾病 Disease	诊断 Diagnostic
已知发生中枢神经系统转移的患者 Patients who have been transferred in central nervous system	肿瘤进展 Neoplasm Status	疾病 Disease	疾病 Disease

Table 29: Error cases in CHIP-CTC. We evaluate roberta-wwm-ext and PCL-MedBERT on 3 sampled sentences with their gold labels and model predictions. RO = roberta-wwm-ext, MB = PCL-MedBERT.

Query-A	Query-B	Model			Gold
		BE	BE+	MB	
汗液能传播乙肝病毒吗? Can sweat spread the hepatitis B virus?	乙肝的传播途径? How is hepatitis B transmitted?	0	0	0	1
哪种类型糖尿病? What type of diabetes?	我是什么类型的糖尿病? What type of diabetes am I?	1	1	1	0
如何防治艾滋病? How to prevent AIDS?	艾滋病防治条例。 AIDS Prevention and Control Regulations.	1	0	0	1

Table 30: Error cases in CHIP-STS. We evaluate performance of baselines with 3 sampled instances. The similarity between queries is divided into 2 levels (0-1), which means '*unrelated*' and '*related*'. BE = BERT-base, BE+ = BERT-wwm-ext-base, MB = PCL-MedBERT.

表 31：KUAKE-QTR 中的错误案例。我们使用三个采样实例评估基线性能。查询和标题之间的相关性分为四个等级（0-3），表示"无关"、"较差的相关性"、"相关"和"强相关"。BE = BERT-base，BE+ = BERT-wwm-ext-base，MB = PCL-MedBERT。

表 32：KUAKE-QQR 中的错误案例。我们使用三个采样实例评估基线性能。查询和标题之间的相关性分为三级（0-2），这意味着"不相关或无关"，"相关"和"强烈相关"。BE = BERT-base，ZEN= ZEN，MB=PCL-MedBERT。

中国平安健康科技（上海）有限公司的袁妮和中国平安健康云公司的郭通，以

及中国平安国际智慧城市科技股份有限公司的郭通共同贡献了 CHIP-STS 数据集。
阿里巴巴集团的金旭和浙江大学数学科学学院的王新，贡献了 KUAKE-QIC、KUAKE-QTR 和 KUAKE-QQBR 数据集。周涛和陈青才来自哈尔滨工业大学。
深圳大学、鹏城实验室等单位参与了项目建议，提出了任务需求，并牵头开展了研究。