arXiv:2305.01526v1 [cs.CL] 2 May 2023

Huatuo-26M，大规模中文医学问答数据集

Jianquan Li[β*], **Xidong Wang**[α,β*], **Xiangbo Wu**[β], **ZhiyiZhang**[β], **Xiaolong Xu**[β], **JieFu**[γ], **XiangWan**[α], **Benyou Wang**[α,β✉] 深圳大数据研究院[α]香港中文大学（深圳）[β] 北京人工智能研究院[γ] wangbenyou@cuhk.edu.cn

摘要

在本文中，我们发布了一个**迄今规模最大的**医学问答数据集，其中包含**2600万**个问答对。我们针对检索和生成，对数据集上的多种现有方法进行了基准评测。实验结果表明，现有模型的表现远低于预期，在预训练语言模型时代，发布的数据集仍然具有挑战性。此外，我们还通过实验展示了该数据集在多个方面的价值：(i) 以零样本方式训练其他问答数据集的模型；(ii) 作为检索增强生成（RAG）的外部知识；以及 (iii) 在持续训练过程中，将问答对用作预训练语料库，从而改进现有预训练语言模型。我们相信，该数据集不仅有助于医学研究，也能为患者和临床医生提供便利。请参阅
https://github.com/FreedomIntelligence/Huatuo-26M。

1 引言

预训练语言模型在自然语言处理（NLP）领域取得了巨大进展，并显著提升了自然语言理解和自然语言生成能力。这促使研究人员将预训练语言模型应用于医学等以往并非自然语言处理核心领域的方向。然而，与人工智能领域的大多数其他突破一样，医学领域应用预训练语言模型面临的首要瓶颈是数据，而突破往往始于数据收集。

为打破这一瓶颈，本研究收集了规模最大的中文医学问答数据集，也可能有助于医学研究。值得注意的是，有 14 亿人以中文为母语；更重要的是，他们所获得的医疗服务（尤其是中国大陆）总体上远低于西方对应水平（例如，英语国家和发达国家）[1]。

数据集 我们从以下多个来源收集了规模最大的医疗问答数据集：(i) 从在线医疗咨询网站收集；(ii) 从医学百科全书中自动提取；(iii) 从医学知识库中自动提取。经过文本清理和数据去重后，我们获得了规模最大的中文医学问答数据集，其中包含 **2600万** 个问答对。我们将该数据集命名为“Huatuo-26M”，以纪念生活在公元200年左右的中国伟大医师华佗。如表1 所示，本研究将现有医疗领域问答数据集的规模扩大了两个数量级以上，甚至超过了大多数通用领域的问答数据集。

基准 基于收集的数据集，我们对检索领域中的经典方法进行了基准测试：对于稀疏检索，我们测试了 BM25 (Robertson 等人，2009年) 和 DeepCT (Dai 和 Callan，2019年) 的性能；对于稠密检索，我们测试了 DPR (Karpukhin 等人，2020年) 的性能。同时，我们还训练了一些自回归语言模型，即 GPT2 (Brown 等人，2020年) 和 T5 (Raffel 等人，2020年)。结果表明，使用预训练语言模型完成该任务仍然具有挑战性，这可能是因为医学领域涉及的专家知识比通用领域更多。

为进一步展示所收集数据集的实用性，我们将其应用于三种场景：(i) 迁移到其他问答数据集；(ii) 作为检索增强生成的外部知识；以及 (iii) 作为预训练语料。

用例 I：迁移到其他问答数据集由于 Huatuo-26M 数据集规模庞大，我们也期望通过该数据集训练的模型能够

*前两位作者对本文贡献相同。

[1]参见 https://en.wikipedia.org/wiki/世界_卫生_组织_对_卫生_系统_的排名_，位于_2000

数据集	Lang	领域来源		#Q
MedHop（Welbl 等，2018年）	英语	医学	MEDLINE	2.5K
BiQA（Lamurias 等，2020年）	英语	医疗在线医学论坛		7.4K
HealthQA（Zhu 等，2019年）	英语	医学	医疗服务网站	7.5K
MASH-QA（Zhu 等，2019年）	英语	医学	医学文章网站	35K
MedQuAD（Ben Abacha和Demner-Fushman，2019年）	英语	医学	美国国立卫生研究院（NIH）	47K
ChiMed（Tian 等人，2019年）	中文	医疗在线医学论坛		47K
MedQA（He 等人，2020年）	英语和中文	医学	医学考试	60K
webMedQA（He 等人，2019年）	中文	医学	医疗咨询网站	63K
CliCR（Šuster 和 Daelemans，2018年）	英语	医学临床病例报告		100K
cMedQA2 (Zhang等人，2018年)	中文	医学医疗在线医学论坛		108K
Huatuo-26M	**中文**	**医疗咨询记录、百科全书、知识库 2600万**		
TriviaQA（Joshi等人，2017年）	英语	通用	冷知识	96K
HotpotQA（Yang等人，2018年）	英语	通用	维基百科	113K
SQuAD（He 等人，2016年）	英语	通用	维基百科	158K
DuReader（He 等人，2017年）	中文	通用	网络搜索	200K
NaturalQuestions(Kwiatkowski 等人，2019年)	英语	通用	维基百科	323K
MS MARCO（Nguyen 等人，2016年）	英语	通用	网络搜索	1.0M
CNN/《每日邮报》（See 等人，2017年）	英语	通用	News	1.3M
PAQ（Lewis等，2021）	英语	通用	维基百科	65M

表1：现有问答数据集。

掌握通用医学知识。因此，我们将训练好的模型应用于两个现有的医学问答数据集，即cMedQA2（Zhang等人，2018年）和webMedQA（He 等人，2019年）。实验结果表明，即使在少样本或零样本的情况下，该模型也能取得有竞争力的表现。

用例II：作为检索增强生成的外部知识大规模医学问答数据集本身明确包含丰富的医学知识，我们将其用作检索增强生成（Lewis等，2020年）中的外部知识。cMedQA2和webMedQA数据集上的实验结果表明，将该数据集用作外部知识库可以显著提升生成文本的质量。

用例 III：作为预训练语料库 考虑到该数据集的规模与通用预训练语言模型的预训练语料库规模相当，我们将 Huatuo-26M 的文本语料库用作预训练语料库，通过预训练向模型注入隐性知识。我们以持续训练的方式，在该数据集上使用问答对作为预训练语料库，对 Bert 和 RoBERTa 进行改进。实验结果表明，将 Huatuo-26M 用作额外的预训练语料库，可以大幅提升预训练模型在生物医学任务上的表现。

本文的贡献如下：**(i)** 我们发布了规模最大的中文医学问答数据集（包含 **26,504,088** 个问答对）；**(ii)** 我们对一些现有模型在所提出方法中的检索和生成效果进行了基准测试；以及 **(iii)** 我们探索了该数据集的一些其他用法，例如迁移至其他问答数据集、将其作为检索增强生成的外部知识进行训练，以及将其作为预训练语料进行训练。

2 数据集创建

我们从多个来源收集了各类医学知识文本，并统一整理为医学问答对的形式。主要来源包括在线医学问答网站、医学百科全书和医学知识库。不同来源的具体示例见附录B 。下文将介绍从上述三个数据来源收集数据的详细情况。

2.1 在线医疗咨询记录

数据来源 我们从一个医疗咨询网站[2]收集了数据，该网站包含许多由医学专家提供的在线咨询记录。每条记录都是一个问答对：患者提出问题，医生回答问题。我们记录了医生的基本信息（包括姓名、医院机构和科室）。

数据处理 我们直接抓取患者的问题和医生的回答作为问答对，共获得 31,677,604 个问答对。随后，我们删除了

[2]Qianwen Health 位于 https://51zyzy.com/

包含特殊字符的问答对，并删除了重复的问答对。最终，我们得到了 25,341,578 个问答对。

2.2 在线医学百科

数据来源 我们从纯文本（例如医学百科全书和医学文章）中提取医疗问答对。我们从中文维基百科收集了 8,699 条疾病百科词条和 2,736 条药物百科词条 [3]。此外，我们从 Qianwen Health 网站抓取了 226,432 篇高质量医学文章[4]。

数据处理 我们首先对文章进行结构化处理。每篇文章都会被划分为标题—段落对。例如，关于药物的文章中，标题可能包括用法、禁忌证和营养；关于疾病的文章中，标题可能包括诊断、临床特征和治疗方法。我们删除出现次数少于五次的段落标题，最终得到733个不重复的标题。基于这些标题，我们人工设计模板，将每个标题转换为相应段落可以回答的问题。请注意，疾病名称或药物名称可以作为模板中的占位符。详情请参见附录 C 。

2.3 在线医学知识库

数据来源 一些现有知识库明确存储了经过良好组织的知识，我们从中提取医疗问答对。我们从以下三个医学知识库收集数据：

CPubMed-KG [5] 是一个面向中国医学文献的知识图谱，基于中华医学会的大规模医学文献数据构建； **39Health-KG**[6] 和**Xywy-KG**[7] 是两个开源知识图谱。基本信息见表2。

	# 实体类型	#关系	#实体	#三元组
CPubMed-KG	-	40	1.7M	4.4M
39Health-KG	7	6	36.8K	210.0K
Xywy-KG	7	10	44.1K	294.1K

表 2：三个知识库的基本统计信息。

[3]zh.wikipedia.org/wiki/ [4]https:/51zyzy.com/ [5] https://cpubmed.openi.org.cn/graph/wiki [6] https://github.com/zhihao-chen/QASystemOnMedicalGraph [7] https://github.com/baiyang2464/chatbot-base-on-Knowledge-Graph

组成	# 对	问题长度	答案长度
Huatuo-26M 训练集	26,239,047	44.6	120.7
Huatuo-26M 测试集	265,041	44.6	120.6
数据来源：			
咨询记录	25,341,578	46.0	117.3
百科全书	364,066	11.5	540.4
知识库	798,444	15.8	35.9
All	26,504,088	44.6	120.7

表3：Huatuo-2600万的基本统计信息。

数据处理 我们通过移除无效字符对这三个知识图谱进行清理，随后合并这三个知识图谱中的实体及实体间的关系，最终得到 43 个类别。每个类别对应实体间的一种关系或实体的一项属性。随后，我们手动设计模-板，将每个类别转换为一个问题。问题 可以是：1）根据主体实体查询客体实体；或 2）查询某个实体的属性。在这两种情况下，客体实体都是该问题对应的答案 。最后，我们利用相应的模板构造问题和答案，共获得 798,444 个问答对。详情请参阅附录 D 。

3 数据统计与分析

Huatuo-2600万的基本统计信息如表3所示，大多数问答对来自在线咨询记录。数据集问题的平均长度为44.6，答案的平均长度为120.7。问题可能较长（例如咨询记录中的问题），也可能较短（例如百科全书和知识库中的问题）。答案既有较长的（例如百科全书中的答案），也有较短的（例如咨询记录和知识库中的答案）。我们随机抽取1%的问答对作为测试集，其余问答对构成训练集。

问题较为口语化，而答案较为专业 Huatuo-2600万由大量口语化的问答对组成，这些问答对更接近线下医学诊断，并包含大量医学知识。如表9所示，在在线医学咨询的示例中，患者的问题包含患者特征和日常症状，并伴随着贴近生活的场景，而医生的回答具有针对性，并且在语义上保持上下文连贯性。

问题种类繁多为了更好地了解数据集的特征，我们对问题进行了启发式分析，从第一个

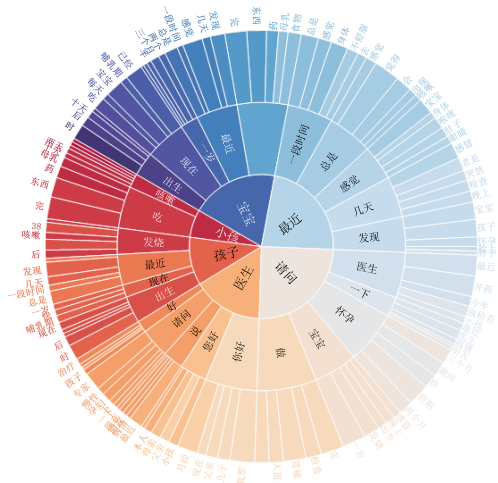


图 1：以患者为中心的问题分布。我们采用启发式方法进行分析，根据问题中第一个有意义的短语进行计数。我们展示了这些反复出现的问题的相对分布及其后续分布。

有意义的短语开始统计问题。如图所示1，我们得到了重复问题的相对分布及其后续词汇分布。患者提出的问题很大一部分与新生儿、孕妇和儿童有关，这也合乎情理，因为父母缺乏育儿经验，而儿童免疫力尚未成熟，容易患病。同时，代年迈的父母进行的问诊也占了相当大的比例，这表明在线问诊在一定程度上帮助解决了中国人口老龄化的问题。

Huatuo-26M 中的重要主题 为了更细致地了解该数据集，我们还在附录中绘制了不同数据来源的词云A。我们发现，对于来源于在线医疗咨询记录的问答对，问题不仅涉及感冒和咳嗽等常见疾病，也涉及高血压和糖尿病等慢性疾病，甚至还有少量囊肿和癫痫等复杂疾病。答案主要集中于开具药物、提供饮食和日常生活方面的注意事项，以及建议在医院接受进一步治疗。对于来自维基百科的问答对，与来自咨询记录的问答对相比，男科学、妇科学和传染病的占比有所上升。对于来自知识库的问答对，关于淋巴瘤等疑难病的问题占比和白血病的问题占比上升，而答案则包含更多专业的诊断和治疗措施，这些措施需要在医院线下进行，例如磁共振成像、计算机断层扫描和穿刺。

4 基准测试

我们从答案检索和生成两个方面，对 Huatuo-26M 数据集上的一些主流方法进行了基准测试。

4.1 基于检索的基准

本节将对 Huatuo-26M 数据集上的主流检索方法进行基准测试，包括稀疏检索和稠密检索方法。

4.1.1 基线与实验设置

对于给定的问题，我们从答案池中对排名靠前的 1000 个相关答案进行排序，答案池由训练集和测试集中的答案组成。对于百科全书和知识库，我们使用 90% 的问题进行训练，其余用于测试。对于咨询记录或所有类别，我们使用 99% 的问题进行训练，其余用于测试，因为使用 1% 的问题进行测试已足够，而且相比使用 10% 的问题进行测试，可以节省更多评估时间。

BM25 BM25 是一种词袋检索函数，它根据每篇文档中出现的查询词对一组文档进行排序。考虑到中文预训练模型都基于单个字符，为便于比较，我们使用单个字符而非词语作为单位来构建索引。我们使用 Lucene 代码库，并将 k1 设为 1.2 、 b 设为 0.9。

DeepCT （Dai和Callan，2020年）使用BERT [8] 来确定上下文感知术语权重。我们使用 Adam 训练模型 3 个训练轮次，学习率为2×10^{-5}。批量大小设为 72，最大序列长度设为 256。

DPR (Karpukhin 等人，2020年）通过简单的双编码器框架学习嵌入。我们实验中使用的 DPR模型采用批内负样本设置进行训练，批量大小为 192，并额外加入 BM25负样本。我们对问题编码器和段落编码器训练了 2 个训练轮次，学习率为 10^{-5} 使用 Adam，采用带预热的线性调度，并将 Dropout率设为 0.1。

[8]我们使用 Google 发布的 BERT检查点，可从 https://huggingface.co/bert-base- chinese 下载

数据来源	模型	召回率 @5	召回率 @20	召回率 @100	召回率 @1000	平均倒数排名@10
医学咨询记录	BM25	4.91	6.99	10.37	17.97	3.82
	DeepCT	**7.60**	10.28	14.28	22.85	**6.06**
	DPR	6.79	**11.91**	**20.96**	**42.32**	4.52
百科全书	BM25	4.58	8.71	17.82	39.91	3.10
	DeepCT	**20.33**	26.92	36.61	53.41	**16.25**
	DPR	16.01	**27.25**	**45.33**	**78.30**	11.20
知识库	BM25	0.52	1.02	1.82	3.51	0.38
	DeepCT	1.05	1.46	2.10	3.29	0.71
	DPR	**2.66**	**5.25**	**11.84**	**33.68**	**1.83**
ALL	BM25	4.77	6.83	10.21	17.84	3.71
	DeepCT	**7.58**	10.24	14.22	22.68	**6.04**
	DPR	6.79	**11.92**	**21.02**	**42.55**	4.53

表 4: 基于检索的 Huatuo-26M 数据集基准。结果按不同数据来源分别列出 .

评估指标 我们使用 Recall@k 和 MRR@10 作为评估指标。Recall@k 衡量前 k 个检索段落中包含答案的比例。MRR@10 计算首个相关文档被检索到时其排名倒数的平均值。

4.1.2 结果

实验结果如表4所示。DeepCT 和 DPR 均优于 BM25，证明了神经信息检索模型的有效性。在大多数情况下，DPR 的表现优于 DeepCT，这可能是因为稠密信息检索模型通常比稀疏神经信息检索模型更强大。值得注意的是，在涉及咨询记录的实验中，召回率相对较低，因为检索候选项池（即 2600万）过大，难以召回所需文档。有趣的是，我们发现即使未能召回所需答案，排名靠前的答案仍然很有参考价值。有关具体样例分析，请参阅附录 E。

这些检索模型通常无法在从知识库中提取的问题问答任务中取得良好的表现。这是因为知识库中的问题通常较为简洁，因此模型需要深入理解知识（例如医学实体及其相互关系）。预训练语言模型中的知识表示（例如在检索场景中）仍然具有挑战性；而在医学领域，这一问题更具挑战性，因为医学领域的知识密集程度更高。

⚠ 值得注意的是，基于检索的医疗问答解决方案假设：1）所有医学问题都应有预先定义的答案；2）对于给定问题，答案应是固定的，且不受患者不同背景的影响。这两个假设有时并不成立。首先，医学领域总会出现一些新的突发情况，例如新冠肺炎，在其刚出现时，人们对它知之甚少。其次，对于给定的医学问题，答案（如建议和治疗方案）取决于个人情况，例如年龄、性别、症状和并发症，以及症状处于早期还是晚期。因此，固定的答案可能不足以满足医疗咨询的需求。

4.2 基于生成的基准

我们使用 Huatuo-26M 的训练集对生成式语言模型（例如 T5 和 GPT2）进行微调，并在测试集上对其进行评估。

4.2.1 基线模型与实验设置 ngs

我们报告了原始 T5 和 GPT2 的结果，以及在 Huatuo-2600万 训练集上进行微调后的结果。

T5 在统一的文本到文本框架中训练多种基于文本的语言任务。我们使用批量大小为 8、学习率为 10^{-4}，采用 Adam 优化器，并使用预热比例为 0.1 的线性调度，在完整训练集上对 T5 持续训练 1 个周期。中文 T5 模型有 12 层 T5 [9]。

GPT2 是一种仅解码器的生成式语言模型。我们使用批量大小为 12、学习率为 10^{-4}、采用 Adam 优化器和预热率为 0.1 的线性调度，在完整训练集上对 GPT2 微调 1 个 epoch。对于 T5 和 GPT2，问题和答案的最大长度分别设为 256 和 512。中文 GPT 是原始的 12 层 GPT2 [10]。

[9] https://huggingface.co/imxly/t5-pegasus [10] 下载自 https://huggingface.co/ uer/GPT2-中文-cluecorpussmall

模型	BLEU-1	BLEU-2	BLEU-3	BLEU-4	GLEU	ROUGE-1	ROUGE-2	ROUGE-L	Distinct-1	Distinct-2
T5	0.33	0.18	0.12	0.07	0.10	0.67	0.19	0.63	0.01	0.02
T5（微调）	**26.63**	**16.74**	**11.77**	**8.46**	**11.38**	**33.21**	**13.26**	**24.85**	**0.51**	**0.68**
GPT2	10.04	4.60	2.67	1.62	3.34	14.26	3.42	12.07	0.17	0.22
GPT2（微调）	23.42	14.00	9.35	6.33	9.47	30.48	11.36	23.15	0.43	0.58

表5：基于生成的 Huatuo-2600万 基准。

评估指标 我们使用双语评估替补、ROUGE、GLEU 和 Distinct 作为评估指标。**双语评估替补** 通过计算生成话语与参考文本之间的 k 元组重叠，评估生成句子与参考句子的相似度。**ROUGE-N** 衡量生成句子与参考文本之间的 N 元组重叠；ROUGE-L 则利用最长公共子序列衡量词语匹配的最长序列。**GLEU**通过考察不同的解析器，自动评估句子层面的流畅度。**Distinct-1/2**是一种辅助指标，通过计算不同 n 元组的数量来评估生成回复的文本多样性。

4.2.2 结果

生成基准测试的结果总结于表5中。显然，与未经微调的原始 T5 和 GPT2 模型相比，微调后的 T5 和 GPT2 模型均有显著提升，尤其是微调后的 T5 在所有评估指标上都取得了最佳结果。需要注意的是，生成方法的性能似乎相对较弱（在这些生成指标上的得分相对较低），这可能是因为预期答案通常较长，生成完全相同的长答案比生成简短答案更困难（例如某些通用问答任务中的实体，如 Natural Questions（Kwiatkowski 等人，2019年））。

⚠ 我们提醒大家，基于生成的医疗问答具有风险。由于难以验证生成内容的正确性，医学领域中的误导性信息可能会引发更严重的伦理问题。我们对这些生成方法进行基准测试，是因为得益于 ChatGPT 的成功，如今问答中的生成方法比检索方法更具发展前景。然而，它们尚未准备好部署到现实世界中。

5 应用

本节将从多个方面展示所提出数据集的实用性：迁移方面的应用。其他问答数据集，作为外部知识，以及作为第5.1节、5.2节和第 5.3节。

5.1 迁移到其他问答数据集

本节将说明 Huatuo-26M 如何有助于现有的问答数据集。

问题设定 本节中，我们直接将基于 Huatuo-26M 数据集预训练的模型应用于其他答案生成数据集，并对其进行评估。T5-CBQA（Roberts 等人，2020年）中也采用了类似的配置。

实验设置 我们选取了两个现有的中文医学问答数据集作为示例，分别是 cMedQA2（Zhang等人，2018年）和 webMedQA（He 等人，2019年）。**cMedQA2** 是一个公开可用的数据集，基于中文医学问题和答案构建，包含 108,000 个问题和 203,569 个答案。**webMedQA**是一个从在线健康咨询网站收集的真实世界中文医学问答数据集，包含 63,284 个问题。我们从这两个数据集中筛选出正确的问答对，用于训练生成模型。T5 和 GPT2 的模型设置参见第 4.2.1节。

结果 如表6所示，在 Huatuo-26M 数据集上预训练的模型性能远高于原始模型。特别是，在 Huatuo-26M 上进行额外训练，使原始 T5 模型在 cMedQA2 上提升了 25.42 个绝对百分点，在 webMedQA 上提升了 22.73 个绝对百分点。此外，在 cMedQA2 数据集上，从未接触过 cMedQA2 训练集或测试集的 Huatuo-26M 训练 T5，在 BLEU-1 方面优于通过 cMedQA2 训练的 T5。这表明 Huatuo-26M 包含广泛的医学知识，有利于下游医学任务。此外，将 Huatuo-26M 作为训练集，在 cMedQA2 上取得了比使用其自身训练集更好的性能，这可能是因为 Huatuo-26M 规模庞大，其中可能包含与 cMedQA2 相关的信息。这显示了 Huatuo-26M 在迁移

数据集	模型	BLEU-1	BLEU-2	BLEU-3	BLEU-4	GLEU	ROUGE-1	ROUGE-2	ROUGE-L	Distinct-1	Distinct-2
cMedQA2	GPT2（原始）	9.96	4.30	2.33	1.33	3.18	13.85	3.07	11.60	0.175	0.218
	T5（原始）	0.23	0.12	0.07	0.04	0.07	0.53	0.13	0.50	0.014	0.015
	GPT2（经 Huatuo-26M 微调）	23.34	13.27	8.49	5.55	8.97	29.10	9.81	21.27	0.462	0.611
	T5（经 Huatuo-26M 微调）	**25.65**	**14.94**	**9.79**	**6.64**	**10.03**	**30.64**	**10.49**	**21.48**	**0.543**	**0.727**
	T5（经 cMedQA2 微调）†	20.88	11.87	7.69	5.09	7.62	27.16	9.30	20.11	0.418	0.526
webMedQA	GPT2（原始）	7.84	3.51	1.99	1.16	2.56	12.00	2.70	10.07	0.120	0.150
	T5（原始）	0.47	0.21	0.13	0.08	0.13	1.04	0.20	0.97	0.009	0.009
	GPT2（经 Huatuo-26M 微调）	19.99	11.54	7.51	4.97	7.80	28.19	9.69	21.30	0.363	0.494
	T5（经 Huatuo-26M 微调）	**23.20**	**13.80**	9.21	6.29	**9.22**	30.68	**10.90**	22.26	**0.462**	**0.633**
	T5（经 webMedQA 微调）†	21.42	13.79	**10.06**	**7.38**	8.94	**31.00**	**13.85**	**25.78**	0.377	0.469

表 6：在 Huatuo-26M 上训练的模型的零样本性能。† 表示微调，其余为零样本。

中医学习。

5.2 Asan 外部知识

问题设定 检索增强生成（Lewis 等人，2020年）通过结合预训练的参数化记忆和非参数化记忆（即外部知识）来进行生成，从而将记忆与泛化解耦。此处，我们将 Huatuo-26M 用作检索增强生成中的外部知识资源。对于给定的问题 q，我们使用训练好的 DPR 作为检索模型，从问答数据集中获取排名最高的问答对 (q增强 a增强)，并将其作为额外输入。

实验设置 考虑到 T5 在零样本场景下的表现优于 GPT2，我们使用 T5 而非 GPT2，根据拼接后的文本生成答案 (q增强 a增强 q)。由于 RAG 模型依赖检索模型，我们首先使用我们的数据集训练一个中文 DPR 模型。然后，我们使用文档编码器为每个文档计算嵌入，并使用 FAISS 构建单个最大内积搜索索引（Johnson 等人，2017年），以实现快速检索。在 RAG 训练中，我们为每个问题检索最接近的问答对，并将其拆分为(q增强 a增强 q) 格式。我们将拼接后的最大文本长度设为 400，训练 10 轮，批量大小为 24，学习率为 3e-05。**T5** 与 **T5（Huatuo-26M）**的区别在于，后者先在 Huatuo-26M 数据集上训练，再在目标数据集（即 cMedQA2 或 webMedQA）上训练。

结果 如表 7所示，我们发现，RAG 策略在一定程度上提高了文本生成的质量。特别是在 cMedQA2 上，无论是否在 Huatuo-26M 数据集上进行预训练，模型都能持续从 RAG 策略中受益。对于 RAG，我们还可以在微调之前，按照第 5.1节中的介绍，先在 Huatuo-26M 上对骨干模型进行额外训练；这种额外预训练带来的提升可以在 cMedQA2 上观察到（相比纯 RAG 提升了 3 个百分点），但在 webMedQA 上没有观察到（下降了近 6 个百分点）；这可能取决于目标数据集的特征。

5.3 作为预训练语料

问题设定 我们使用 Huatuo-26M 作为预训练语料，继续训练现有的预训练语言模型，例如 BERT 和 RoBERTa。

5.3.1 实验设置

BERT BERT（Devlin 等人，2018年）是一种基于 Transformer的语言表示模型。**BERT-base** 是原始的 12 层 BERT，中文BERT 下载自 https://huggingface.co/bert-base-chinese。**BERT-base（Huatuo-26M）**是以**BERT-base** 初始化，并使用掩码语言模型在 Huatuo-26M 数据集上持续训练得到的模型。我们对模型进行了 10 个训练轮次，学习率为 5^{-5}，批大小为 64。问题和答案拼接在一起，最大长度为 256。

RoBERTa RoBERTa（Liu 等人，2019年）是一种经过更好优化的 BERT 模型。中文 Roberta 下载自https://huggingface.co/hfl/chinese-roberta-wwm-ext。**RoBERTa-base** 有 12 层，而**RoBERTa-large** 有 24 层。 **RoBERTa-base (Huatuo-2600万)** 是由 **RoBERTa-base** 初始化的模型，并使用 Huatuo-26M 数据集通过掩码语言模型进行持续训练。我们以学习率 5^{-5} 和 64 的批大小训练模型，共训练 10 轮。问题和答案拼接在一起，最大长度为 256。

ZEN (Diao 等人，2019年）一种基于 BERT 的中文文本编码器，通过 N 元语法表示进行增强

模型	BLEU-1	BLEU-2	BLEU-3	BLEU-4	GLEU	ROUGE-1	ROUGE-2	ROUGE-L	Distinct-1	Distinct-2
cMedQA2 微调										
T5	20.88	11.87	7.69	5.09	7.62	27.16	9.30	20.11	0.418	0.526
T5-RAG	25.86	18.48	15.26	13.02	14.27	34.24	17.69	27.54	0.395	0.516
T5(Huatuo-26M)	28.76	17.08	11.67	8.41	10.45	29.79	10.23	20.68	**0.647**	**0.831**
T5(Huatuo-26M)-RAG	**31.85**	**22.77**	**18.70**	**15.96**	**17.08**	**37.01**	**19.23**	**28.72**	0.573	0.760
webMedQA 微调										
T5	21.42	13.79	10.06	7.38	8.94	31.00	13.85	25.78	0.377	0.469
T5-RAG	20.30	13.29	9.97	7.61	9.40	32.40	14.88	27.25	0.285	0.377
T5(Huatuo-26M)	**31.47**	**20.74**	**15.35**	**11.60**	**12.96**	34.38	15.18	26.72	**0.651**	**0.832**
T5(Huatuo-26M)-RAG	25.56	16.81	12.54	9.58	11.80	**34.88**	**15.59**	**27.43**	0.447	0.611

表 7：使用或不使用 Huatuo-26M 作为外部检索增强生成语料库的比较。与表 6 的区别在于，这里我们最终在目标数据集上微调这些模型。

模型	CMedEE	CMedIE	CDN	CTC	STS	QIC	QTR	QQR	平均-全部
BERT-base	**62.1**	**54.0**	55.4	69.2	83.0	84.3	60.0	**84.7**	69.1
BERT-base (Huatuo-26M)	61.8	53.7	**56.5**	**69.7**	**84.6**	**86.2**	**62.2**	**84.7**	**69.9**
RoBERTa-base	62.4	53.7	56.4	69.4	83.7	85.5	60.3	82.7	69.3
RoBERTa-large	61.8	**55.9**	55.7	69.0	**85.2**	85.3	**62.8**	84.4	70.0
RoBERTa-base（Huatuo-26M）	**62.8**	53.5	**57.3**	**69.8**	84.9	**86.1**	62.0	**84.7**	**70.1**
ZEN（Diao 等人，2019年）	61.0	50.1	57.8	68.6	83.5	83.2	60.3	83.0	68.4
MacBERT（Cui 等人，2020年）	60.7	53.2	57.7	67.7	84.4	84.9	59.7	84.0	69.0
MC-BERT（Zhang 等人，2020年）	61.9	54.6	57.8	68.4	83.8	85.3	61.8	83.5	69.6

表8：CBLUE 评测测试集上的性能。我们使用 Huatuo-26M 作为预训练语料库。包括 ZEN、MacBERT 和 MC-BERT 在内的结果来自官方网站。

它在训练过程中会考虑不同的字符组合。因此，ZEN 将字符序列以及其中包含的词语或短语的全面信息结合起来。

MacBERT (Cui 等人，2020年) 通过用与其词汇相似的词语进行遮盖，缩小了预训练与微调阶段之间的差距，这对下游任务很有效。它用掩码语言模型纠错（Mac）任务取代了原有的掩码语言模型任务，并缓解了预训练与微调阶段之间的差异。

MC-BERT (Zhang等人，2020年研究预训练语言模型 BERT 如何适应中文生物医学语料库，并提出一种新的概念表示学习方法，通过由粗到细的加密策略将实体和语言领域知识注入表示学习中。

5.3.2 实验数据

我们在 CBLUE（Zhang等人，2022）基准上评估了使用 Huatuo-26M 数据集训练的 BERT 和 RoBERTa。CBLUE 是首个中文医疗语言理解评测基准平台，包含一系列自然语言理解任务，例如命名实体识别、信息抽取，以及单句分类和句对分类。

5.3.3 结果

如表格 8所示，在 Huatuo-26M 数据集上训练的 BERT 和 RoBERTa 显著提升了 CBLUE 的性能。训练后的 12 层 RoBERTa（Huatuo-26M）模型在平均分方面优于 24层 RoBERTa 模型，这表明 Huatuo-26M 数据集包含丰富的医疗信息。RoBERTa-base（Huatuo-26M）模型的平均分比 RoBERTa-base 模型高 0.8 个百分点，比 MC-BERT-base 模型高 0.5 个百分点。

6 结论

本文提出了迄今规模最大的中文医学问答数据集，其中包含**2600万**个医疗问答对，其规模比现有数据集扩大了两个数量级以上。同时，我们基于该数据集对许多现有研究进行了基准测试，发现这些方法在医疗问答场景中仍有很大的改进空间。我们还展示了该数据集在实践中的潜在用途。实验结果表明，该数据集包含丰富的医学知识，能够为现有

数据集和任务提供很大帮助。我们希望Huatuo-26M数据集不仅能够促进医疗问答研究，还能在实际应用中帮助医生和患者。

局限性

⚠ ! 由于该数据集规模庞大，包含2600万问答对，且现阶段几乎不可能由专家进行人工核查，因此其中可能包含一些错误的医学信息。为了更好地维护该数据集，我们计划建立一个在线网站，供临床医生或专家修改这些问答对。我们可以通过招募兼职医生，先对这些数据进行核查，再定期更新。

这个数据集可能会被翻译成其他语言，尤其是资源较少的语言。需要注意的是，翻译可能会引入一些额外的错误。此外，还应注意中医和西医之间的一些基本差异。

对于医疗咨询，治疗方案/建议因人而异。换句话说，这可能高度取决于个人情况，例如年龄和性别、疼痛等主要症状是否伴有其他症状，或者症状处于早期还是晚期。这些信息可能需要通过多轮对话来确认，而非单轮问答。未来，我们将探索用于医疗问答的对话系统。

伦理声明

正如我们在局限性部分所提到的，收集的数据中仍可能包含错误的医疗信息，这主要源于两个方面：1）医生在在线医疗咨询中可能会犯错，尤其是患者可能提供不完整信息；2）问答对的自动提取也可能引入一些不准确的信息。尽管数据规模太大，无法由医学专家逐条人工核查，但我们已采取一些措施来减轻其负面影响。我们在本文的多个部分强调了这些问题，并提醒读者注意。

数据集下载

所有数据均爬取自开源资源。对于从中提取问答对的数据资源，即在线百科全书和知识库，我们直接提供全文问答对。对于我们爬取并整理为问答对的原始数据，例如在线咨询记录，我们提供两个版本：一种是**URL版本**，提供与问答对相关联的网址；另一种是**全文版本**，直接提供问答对的全文。Huatuo-26中提供在线咨询记录URL链接的部分完全开源[11]。而Huatuo-26中提供所有问答对全文的部分，仅向研究机构或大学开放，前提是其同意签署许可协议，承诺仅将其用于研究目的。

参考文献

Asma Ben Abacha 和 Dina Demner-Fushman。2019年。一种用于问答的蕴含方法。*BMC*生物信息学，20(1):1–23。

Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。2020年。语言模型是少样本学习者。神经信息处理系统进展，33:1877–1901。

Yiming Cui、Wanxiang Che、Ting Liu、Bing Qin、Shi- jin Wang 和 Guoping Hu。2020年。重新审视中文自然语言处理的预训练模型。收录于 计算语言学协会论文集：*EMNLP 2020* 研究成果，第657–668 页，线上举行。计算语言学协会。

Zhuyun Dai 和 Jamie Callan。2019年。用于第一阶段检索的上下文感知句子/段落术语重要性估计。*arXiv* 预印本 *arXiv:1910.10687*。

Zhuyun Dai 和 Jamie Callan。2020年。用于第一阶段段落检索的上下文感知的术语加权。收录于第*43*届国际 *ACM SIGIR* 大会信息检索研究与发展论文集，第 1533–1536 页。

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换器预训练。*arXiv* 预印本 *arXiv:1810.04805*。

Shizhe Diao、Jiaxin Bai、Yan Song、Tong Zhang 和 Yonggang Wang。2019年。ZEN：通过 N 元语法表示增强的中文文本编码器预训练。*arXiv*，abs/1911.00720。

[11] 临时下载链接位于 https://drive.google.com/file/d/1SKsU8owLt3IWZPLlnPytpCwm8-EH3iW6/view；来自百科全书和知识库的问答对是完整的全文，但在线医疗咨询记录中的问答对需要自行抓取。这是为了避免一些公司或个人滥用数据。

Junqing He、Mingming Fu 和 Manshu Tu。2019年。将深度匹配网络应用于中文医疗问答：一项研究及一个数据集。*BMC Medical Informatics and Decision Making*，19(2):91–100。

Wei He、Kai Liu、Jing Liu、Yajuan Lyu、Shiqi Zhao、Xinyan Xiao、Yuan Liu、Yizhong Wang、Hua Wu、Qiaoqiao She 等。2017年。Dureader：一个源自真实应用场景的中文机器阅读理解数据集。*arXiv* 预印本 *arXiv:1711.05073*。

Di Jin、EileenPan、NassimOufattole、Wei-Hung Weng、Hanyi Fang 和 Peter Szolovits。2020年。该患者患有什么疾病？来自医学考试的大规模开放领域问答数据集。*arXiv* 预印本 *arXiv:2009.13081*。

Jeff Johnson、Matthijs Douze 和 Hervé Jégou。2017年。使用图形处理器进行十亿规模的相似性搜索。*CoRR*，abs/1702.08734。

Mandar Joshi、Eunsol Choi、Daniel S Weld 和 Luke Zettlemoyer。2017年。Triviaqa：一种用于阅读理解的大规模远程监督挑战数据集。*arXiv* 预印本 *arXiv:1705.03551*。

Vladimir Karpukhin、Barlas O˘guz、Sewon Min、Patrick Lewis、Ledell Wu、Sergey Edunov、Danqi Chen 和 Wen-tau Yih。2020年。《用于开放域问答的稠密段落检索》。*arXiv* 预印本 *arXiv:2004.04906*。Tom Kwiatkowski、J ennimaria Palomaki、Olivia Red- field、Michael Collins、Ankur Parikh、Chris Alberti、Danielle Epstein、Illia Polosukhin、Jacob Devlin、Kenton Lee 等人。2019年。《Natural questions：问答研究的基准》。计算语言学协会会刊，7：453–466。Andre Lamurias、Diana Sousa 和 Francisco M Couto。2020年。《从问答论坛生成生物医学问答语料库》。*IEEE Access*，8：161042–161051。Patrick Lewis、Ethan Perez、Aleksandra Piktus、Fabio Petroni、Vladimir Karpukhin、Naman Goyal、Heinrich Küttler、Mike Lewis、Wen-tau Yih、Tim Rock-täschel 等人。2020年。《面向知识密集型自然语言处理任务的检索增强生成》。神经信息处理系统进展，33：9459–9474。Patrick Lewis、Yuxiang Wu、Linqing Liu、Pasquale Minervini、Heinrich Küttler、Aleksandra Piktus、Pontus Stenetorp 和 Sebastian Riedel。2021。《PAQ：6500万个可能被问到的问题，以及你可以用它们做什么》。计算语言学协会会刊，9：1098–1115。Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Mandar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettlemoyer 和 Veselin Stoyanov。2019年。《RoBERTa：一种稳健优化的 BERT 预训练方法》。*arXiv* 预印本 *arXiv:1907.11692*。

Tri Nguyen、Mir Rosenberg、Xia Song、Jianfeng Gao、Saurabh Tiwary、Rangan Majumder 和 Li Deng。2016年。《MS MARCO：一个由人工生成的机器阅读理解数据集》。发表于*CoCo@ NIPs*。

Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li、Peter J Liu 等。2020年。通过统一的文本到文本转换器探索迁移学习的极限。机器学习研究期刊，21(140):1–67。

Pranav Rajpurkar、Jian Zhang、KonstantinLopyrev 和 Percy Liang。2016年。SQuAD：用于文本机器理解的 100,000+ 个问题。*arXiv* 预印本 *arXiv:1606.05250*。

Adam Roberts、Colin Raffel 和 Noam Shazeer。2020年。语言模型的参数中能容纳多少知识？载于自然语言处理实证方法会议（*EMNLP*）。

Stephen Robertson、Hugo Zaragoza 等。2009年。概率相关性框架：BM25 及其拓展。信息检索基础与趋势®，3(4):333–389。

Abigail See、Peter J Liu 和 Christopher D Manning。2017年。直击要点：使用指针生成网络进行摘要生成。*arXiv* 预印本 *arXiv:1704.04368*。Simon Šuster 和 Walter Daelemans。2018年。CliCR：用于机器阅读理解的临床病例报告数据集。*arXiv* 预印本 *arXiv:1803.09720*。Yuanhe Tian、Weicheng Ma、Fei Xia 和 Yan Song。2019年。ChiMed：用于问答的中文医学语料库。载于生物医学自然语言处理研讨会及共享任务第 *18* 届论文集，第 250–260 页。Johannes Welbl、Pontus Stenetorp 和 Sebastian Riedel。2018年。构建用于跨文档多跳阅读理解的数据集。计算语言学协会汇刊，6：287–302。Zhilin Yang、Peng Qi、Saizheng Zhang、Yoshua Ben- gio、William W Cohen、Ruslan Salakhutdinov 和 Christopher D Manning。2018年。HotpotQA：用于多样化、可解释多跳问答的数据集。*arXiv* 预印本 *arXiv:1809.09600*。Ningyu Zhang、Mosha Chen、Zhen Bi、Xiaozhuan Liang、Lei Li、Xin Shang、Kangping Yin、Chuanqi Tan、Jian Xu、Fei Huang、Luo Si、Yuan Ni、Guo- tong Xie、Zhifang Sui、Baobao Chang、Hui Zong、Zheng Yuan、Linfeng Li、Jun Yan、Hongying Zan、Kunli Zhang、Buzhou Tang 和 Qingcai Chen。2022年。CBLUE：中文生物医学语言理解评测基准。载于计算语言学协会第 *60* 届年会论文集（第 *1* 卷：长篇论文），第 7888–7915 页，爱尔兰都柏林。计算语言学协会。

Ningyu Zhang、Qianghuai Jia、Kangping Yin、Liang Dong、Feng Gao 和 Nengwei Hua。2020年。面向中文生物医学文本挖掘的概念化表示学习。*arXiv* 预印本 *arXiv:2008.10813*。

S. Zhang、X. Zhang、H. Wang、L. Guo 和 S. Liu。2018年。用于中文医疗问答选择的多尺度注意力交互网络。*IEEEAccess*，6：74061–74071。

Ming Zhu、AmanAhuja、Da-Cheng Juan、Wei Wei 和 Chandan K Reddy。2020年。长多跨度答案的问答。载于 计算语言学协会年会论文集：*EMNLP 2020*，第 3840–3849 页。

Ming Zhu, Aman Ahuja, Wei Wei, and Chandan K Reddy. 2019年. 一种用于医疗健康问答的层次注意力检索模型. In 万维网会议, 第2472–2482页.

华佗-26M 数据集的词云

如图 2、 3和 4所示，我们基于词频-逆文档频率提取了排名前1000的关键词，并为Huatuo-26M的不同来源绘制了词云。在线咨询记录中的问答对使用了更多“婴儿”（即“对婴儿可爱的昵称”）等日常词语，因此显得更加非正式；而在其他资源中，它们使用了更多专业医学词汇，因此更加正式。正式与非正式问题的结合使该数据集更加多样化。

B Huatuo-26M 数据集示例

表9 展示了数据集不同来源的示例，通过这些示例可以大致了解各数据来源的特点。对于源自在线医疗咨询记录的问答对，问题更口语化，答案也更有针对性。对于源自在线医学维基和专家文章的问答对，问题更简洁，很少涉及具体的患者信息，而答案则更详尽、专业。对于来自在线医学知识库的问答对，问题简洁，答案准确，且答案与问题之间完全相同的文本较少。

C 从百科页面中提取问答对

如图所示 5 ，对于给定的维基百科页面，我们使用 HTML 解析工具根据页面内容提取其结构化信息。因此，我们根据内容获取一个标题，该标题与一个或多个段落相关联。接下来，我们按照类似表 10 中的人工设计模板，将每个标题转换为可由其关联段落回答的问题。

D 知识库问题模板

表 10展示了为所有知识图谱问题生成的模板。每个问题模板都对应一个实体间关系或实体属性。每个问题模板均以主体实体为条件，例如表 10中的 疾病和 药物等实体占位符。请注意，问题的答案应为客体实体或主体实体的属性。共有43种问题类型。我们人工检查了500个答案可能与问题相匹配的随机示例；结果显示，几乎所有问答对都是正确的。

E 基于检索的基准示例

我们选择 DPR 进行案例研究，因为它作为检索模型的总体表现最佳。表11、12、13 展示了使用 DPR 得到的检索结果。有趣的是，排名靠前的答案都相关且通常有效，尤其是表 13 中在线咨询记录的第一个案例，因为问答数量庞大，其中许多可能重复，从而导致假阴性。因此，尽管检索指标（例如召回率 5）相对较低，但其检索质量仍令人比较满意。





图 2：根据在线咨询记录中的问答对绘制的词云。





图 3：根据百科全书中的问答对绘制的词云。





图 4：根据知识库中的问答对绘制的词云。

	来自 Online Medical Consultant 咨询记录：
问题：我可能得了戊型肝炎，因为我的饮食越来越差，而且吃饭时总是想吐。想问一下，戊型肝炎早期有哪些症状？	我可能得了戊型肝炎，因为我的饮食越来越差，而且吃饭时总觉得恶心。戊型肝炎的早期症状有哪些？
答案：肝脏早期的症状包括脂肪无法消化，导致大量脂肪堆积在体内；同时胃功能也会紊乱，从而出现厌食，看到油腻食物就感到恶心、皮肤色素沉着、指甲颜色改变、毛发改变等症状。	肝脏出现早期症状时，脂肪可能无法消化，导致大量脂肪堆积在体内。同时，胃部功能也会紊乱，出现食欲不振以及看到油腻食物就恶心的症状。还可能出现色素沉着和指甲颜色变化等症状。
问题：3岁婴儿把整颗水果糖咽了，怎么才能知道是咽下去了呢？	3岁的婴儿把整颗水果糖吞了下去，怎么才能知道婴儿已经咽下去，而不是卡在喉咙里？
答案 只要婴儿吞咽后没有屏气的现象，那就是咽下去了。	只要婴儿吞咽后没有屏气，就说明婴儿已经把东西咽下去了。
	摘自在线医学百科：
问题 前列腺钙化灶该如何治疗？	前列腺钙化该如何治疗？
答案 钙化灶是通过X射线检测到的前列腺内的钙沉积。前列腺内的钙化灶有大小之分：粗大的钙化灶通常为前列腺内的良性病变，如前列腺内动脉的老化、陈旧性损伤以及炎症等，不需要进一步活检。细小的钙化灶通常位于细胞生长与分裂较快的部分。出现前列腺钙化或结石必须治疗，钙化会发展成结石，引发各种症状，有的症状长期无法消除，需要做全面检查，看看是否有结石钙化。不治疗结石钙化，前列腺疾病难以彻底治愈。保养治疗需注意劳逸结合，避免过度疲劳；适当进行体育运动，尤其要加强盆腔肌肉锻炼。避免久坐和长时间骑车。，避免久蹲，排便时间控制在3到5分钟，避免坐在潮湿的地方。便后清洁肛门。注意饮食，多喝水，忌酒及辛辣食物。多吃蔬菜、水果和坚果类食物。坚果类食物富含铜和锌，对前列腺有益。	钙化是前列腺内的钙沉积，可通过X射线检测到。前列腺内的钙化可分为不同大小：粗大钙化通常是前列腺内的良性病变，例如前列腺内动脉老化、陈旧性损伤和炎症，通常无需进一步活检。细小钙化通常位于细胞生长和分裂更快的部位。前列腺钙化或结石必须治疗。钙化会发展成结石并引起各种症状。有些症状长期无法钙化会发展成结石，并引发各种症状。有些症状长期无法消除。应进行全面检查，以确定是否存在结石钙化。前列腺疾病无法彻底治愈，钙化的治疗必不可少。维持治疗需要劳逸结合，避免过度疲劳，并进行适当的体育锻炼，尤其是加强盆底肌肉的锻炼。避免久坐、长时间骑自行车和长时间蹲坐。排便时间应控制在3 至 5 分钟内。避免坐在潮湿的地方。排便后清洁肛门。注意饮食，多喝水，避免饮酒和食用辛辣食物。多吃蔬菜、水果和坚果。坚果富含铜和锌，对前列腺有好处。
问题 什么是生物药剂学？	3岁宝宝把整颗水果糖吞下去了，怎么才能知道宝宝已经吞下去了，而且没有卡在喉咙里？
答案 生物药剂学是研究给药后药物吸收的整个体内过程，包含各种制剂因素和生物因素对这一过程与药效的影响响。此外，生物药剂学通过研究药物对生物细胞产生反应的过程，来实现施药者预期的目的。20世纪50年代初，人们普遍认为“化学结构决定药效”，药剂学只是为了改善外观、掩盖不良气味，以便于服用。随着大量临床实践的证明，人们逐渐认识到剂型和生物因素对药效的影响。因此，研究药物在代谢过程中的各种机制和理论，以及各种剂型和生物因素对药效的影响，对于控制药物制剂的内在品质、确保最终药品安全有效，以及为新药开发和用药提供严格评价，都具有重要意义。	生物药剂学研究给药后药物吸收的全过程，包括各种因素的影响——多种制剂因素和生物因素对这一过程及药物疗效的影响。生物药剂学利用药物对生物细胞的响应来达到预期目的。20世纪50年代初，人们普遍认为“化学结构决定药物疗效”，而药学的作用仅仅是改善药物的外观，并掩盖难闻的气味，使其更容易服用。随着大量临床实践的开展，人们逐渐开始重新认识到剂型和生物因素对药物疗效的影响。研究药物在代谢过程中的各种机制和理论，以及各种剂型和生物因素对药物的影响，非常重要疗效，控制药物的内在质量，确保最终药物的安全性和有效性，并提供严格的评估-用于新药研发。
	来自在线医学知识库：
问题 脓腔穿刺的辅助治疗有哪些？	脓腔穿刺有哪些辅助治疗？
答案 消毒隔离；皮肤护理；营养支持	消毒隔离；皮肤护理；营养支持
问题 气道吸痰的辅助治疗有些什么？	气道吸痰的辅助治疗有哪些？
答案	足量补液 充分补液

表9：该数据集的各类来源示例



发热 [编辑]

维基百科，自由的百科全书

"发烧"重定向至此。关于张惠妹的专辑，请见"发烧 (张惠妹专辑)"。

关于与"发热"标题相近或相同的条目，请见"发热 (消歧义)"。

此条目需要补充更多可靠的医学来源，或过于依赖第一手来源。

请协助补充可靠来源以改善这篇条目。无法查证的内容可能会因为异议提出而移除。

维基百科中的医学内容仅供参考，并不能视作专业意见。如需获取医疗帮助或意见，请咨询专业人士。详见医学声明。

发烧（英语：fever），又称作为**发热**（英语：pyrexia）或**发热反应**（英语：febrile response）[6]，其定义为：体温在调节时超过了平常体温。[4][11][5]现在医界并没有一致认可的正常体温上限，文献从37.3到38.3℃都有。[6][7]下丘脑的体温调节中心将原本的体温设定（set point）调高，即为发烧，并让人感到寒冷。[1]这使得身体为了生产更多热而出现肌肉收缩，且开始试图保存热量。[2]当体温调定点回到正常值时（即为退烧时），患者就会开始感到燥热，出现脸红，也可能开始流汗。[2]发热导致的热性痉挛（较强烈的肌肉收缩）较为少见，然而这在年轻孩童患者之中较为常见。[3]发热通常不会高到41至42 °C（105.8至107.6 °F）[5][1]。

发热是指身体制造过多热能或身体的体温调节失调，导致身体的温度高过温度设定值或温度设定值本身过高。[6]发热可能是由许多疾病造成，从小病到重症都有可能，这包含了病毒、细菌、寄生、普通感冒、流行感冒、泌尿道感染、脑膜炎、疟疾、阑尾炎等。 非感染性的发热成因包含了血管炎、深静脉血栓、药物的副作用、癌症等。[12]另外，发热不等同高热这个类似疾病，不同之处在于，高热（中暑为高热的一种[13][14]）起因于身体累积的热能过多或是身体的散热功能不足，导致体温超越正常体温设定点。

治疗发热本身，一般来说是非必要的。[1][8]然而治疗衍生的疼痛与发炎，有利于患者于生病期间的休养，因为患者会觉得舒服些。[8]布洛芬或对乙酰氨基酚这类药物可能可帮助上述治疗，也可以同时降低体温。[15][8]三岁以下的幼童或是患有像是免疫缺陷这类严重疾病的患者或出现多重并发症状的人必须立刻送医。[16]高热无论如何也必须立刻送医。

发热是常见的医学征象之一。体温上升可强化免疫细胞，增加杀死细菌和病毒的能力，发热也可抑制细菌和病毒在体内繁殖。发热约占看病孩童中的求诊原因的30%[1]；生重病的成人会有75%的几率出现发热这个医学征兆。[10]尽管发热是身体本身的一个强而有力的防卫机制，但是治疗发热本身并不会让身体的抵抗力降低便得潜在病因的病情恶化。[17][18]有些家长与医疗从业人士，常会把发热本身看得太严重，这样的现象被称为**发热恐慌症**（英语：fever phobia）。[1]

目录 [隐藏]

1 测量

2 机理

3 分类 （以口测法为准）

结构化分析


```
{
    "desease": "发热",
    "title": {
        "desc": [
            "-{zh-hant:发烧;zh-tw:发烧;zh-hk:发烧;zh-hans:发烧;zh-cn:发烧}- （），又称作为发热
            "发烧是指身体制造过多热能或身体的体温调节失调，导致身体的温度高过温度设置值或温度设置值本身过
            "治疗发烧本身，一般来说是非必要的。然而治疗衍生的疼痛与发炎，有利于患者于生病期间的休养，因为
            "发烧是常见的医学征象之一。体温上升可强化免疫细胞，增加杀死细菌和病毒的能力，发烧也可抑制细菌
        ],
        "测量": [
            "体温一般用体温计测量。高于下列温度之一可认为是发热：腋下温度等于或高于37.2℃（99℉）；口腔内
            "近年来也有通过红外线传感方式测量耳鼓膜温度的耳温枪测温。因测量快速，于2003年SARS流行期间，
        ],
        "机理": [
            "导致发热的物质称为“热原”（）。主要分两类："
        ],
        "分类 （以口测法为准）": [
```


应用手动模板


```
[
  {
    "desease": "发热",
    "title": "desc",
    "questions": "发热是什么?",
    "answers": "-{zh-hant:发烧;zh-tw:发烧;zh-hk:发烧;zh-hans:发烧;zh-cn:发烧}-较为少见，然而这在
  },
  {
    "desease": "发热",
    "title": "测量",
    "questions": "发热该如何测量",
    "answers": "体温一般用体温计测量。高于下列温度之一可认为是发热：腋下温度等于或高于37.2。\n近年来也有
  },
  {
    "desease": "发热",
    "title": "治疗",
    "questions": "发热的治疗方式为?",
    "answers": "不严重的暂时发热不一定需要治疗。原因如下：\n发热会增加心跳和新陈代谢，所以心脏病患者和年老
  }
]
```


图 5：从维基页面提取问答对的工作流程。

疾病(disease)	症状 (symptom)	[疾病]的症状是什么?	(什么是 [疾病]的症状?)
疾病(disease)	并发症 (complication)	[疾病]的并发症是什么?	([疾病]的并发症是什么?)
疾病(disease)	简介 (简介)	[疾病]的简介是?	([疾病]的简介是什么?)
疾病(disease)	预防 (prevention)	[疾病]的预防措施有哪些?	(什么是 [disease]的预防措施?)
疾病(disease)	病因 (病因)	[疾病]的发病原因?	(疾病 [的发病原因是什么?)
疾病(疾病)	发病率 (发病率)	[疾病]的患病比例是多少?	(疾病 [的患病率是多少?)
疾病(disease)	就诊科室（医学科室）	[疾病]的就诊科室是什么?	([疾病]的就诊科室是什么?)
疾病(disease)	治疗方式（治疗）	[疾病]的治疗方式是什么?	(治疗 [disease]需要什么方式?)
疾病(disease)	治疗周期 (治疗周期)	[疾病]的治疗周期多长?	治疗 [disease]的周期有多长?
疾病（疾病）	治愈率（治愈率）	[疾病]的治愈率是多少?	([疾病]的治愈率是多少?)
疾病（疾病）	检查（一次检查）	[疾病]的检查有哪些?	([疾病]有哪些检查?)
疾病（疾病）	多发群体（多发群体）	[疾病]的多发群体是?	(哪些人群更容易患上 [疾病]?)
疾病(disease)	药物治疗 (药物治疗)	[疾病]的推荐药物有哪些?	(针对 [disease]有哪些推荐药物?)
疾病(disease)	忌食 (不要吃)	[疾病]忌食什么?	([疾病]患者不宜吃什么?)
疾病（疾病）	宜食（适宜食用）	[疾病]宜食什么?	([疾病]患者适宜吃什么?)
疾病(disease)	死亡率（死亡率）	[disease]的死亡率是多少?	([disease]的死亡率是多少?)
疾病(disease)	辅助检查（辅助检查）	[disease]的辅助检查有哪些?	([disease]的辅助检查有哪些?)
疾病（disease）	高发季节（高发季节）	[disease]的高发季节是什么时候?	(人们最可能在哪个季节患上 [disease]?)
疾病(disease)	相关（症状）（相关（症状））	[疾病]的相关症状有哪些?	([疾病的伴随症状有哪些?)
疾病(disease)	发病机制 (pathogenesis)	[疾病]的发病机制是什么?	(什么是 [疾病]的发病机制)
疾病(疾病)	手术治疗 (手术治疗)	[疾病]的手术治疗有哪些?	([疾病]的手术治疗是什么?)
疾病(疾病)	转移部位 (转移部位)	[疾病]的转移部位是什么?	(对于 [疾病]，迁移部位是什么?)
疾病(疾病)	风险评估因素 (风险评估因素)	[disease]的风险评估因素有些什么	([disease]的风险评估因素有哪些）?
疾病(disease)	筛查（筛查）	[disease]的筛查有些什么?	([disease]的筛查有哪些?)
疾病(disease)	传播途径（传播方式）	[disease]的传播途径有哪些?	([disease]的传播途径有哪些?)
疾病（疾病）	发病部位（患病部位）	[disease]的发病部位是什么?	([疾病]的部位是什么?)
疾病（疾病）	高危因素（高危因素）	[疾病]的高危因素有哪些?	([疾病]的高危因素有哪些?)
疾病（疾病）	发病年龄（Age of onset）	[疾病]的发病年龄是多少?	([疾病]的发病年龄是多少?)
疾病(disease)	预后生存率（prognostic survival rate）	[疾病]的预后生存率是多少?	([疾病]的生存预后如何?)
疾病(疾病)	组织学检查（组织学检查）	[疾病]的组织学检查有哪些?	([疾病]的组织学检查有哪些?)
疾病(疾病)	辅助治疗（辅助治疗）	[疾病]的辅助治疗有哪些?	([疾病]有哪些辅助治疗?)
疾病（疾病）	多发地区（高发地区）	[疾病]的高发地区在哪里?	([疾病]的常见发生区域在哪里?)
疾病(disease)	遗传因素 (genetic factors)	[疾病]的遗传因素是什么?	([疾病]的遗传因素是什么?)
疾病(disease)	发病性别倾向（发病性别倾向）	[疾病]的发病性别倾向是啥?	([疾病]的发病是否存在性别倾向?)
疾病（disease）	放射治疗	[疾病]的放射治疗有哪些?	([疾病]的放射治疗是什么?)
疾病(disease)	化疗 (chemotherapy)	[疾病]的化疗有哪些?	([疾病]的化疗是什么?)
疾病(disease)	临床表现（临床表现）	[disease]的临床表现有哪些?	([disease]有哪些临床表现?)
疾病(disease)	内窥镜检查（内窥镜检查）	[disease]的内窥镜检查有哪些?	([疾病]的内窥镜检查是什么?)
疾病(disease)	影像学检查 (影像学检查)	[疾病]的影像学检查有哪些?	([疾病有哪些影像学检查?)
疾病(disease)	相关（导致）（相关（导致））	[疾病]会导致什么样的结果?	([疾病] 会导致什么结果?)
疾病（disease）	治疗后症状（治疗后症状）	[疾病]的治疗后症状是什么?	(治疗后 [疾病]有哪些症状?)
疾病(疾病)	相关(转化) (相关（转化）)	[疾病]会转化成什么?	(治疗后 [疾病]会转化成什么?)
药品(药物)	推荐药物（该药可治愈的疾病）	[药物]能治疗什么疾病?	([药物]能治疗哪些疾病?)

表 10：将知识库中的关系转换为问题的模板。

问题	前列腺钙化灶是怎么治呢?
答案	钙化灶是X射线检测到的前列腺内的钙沉积。前列腺内的钙化灶有大小之分：粗大的钙化灶常常是前列腺内的良性病变，如前列腺内动脉的老化、陈旧性损伤以及炎症等，不需要进一步活检。细小的钙化灶通常位于细胞生长与分裂较快的部分。出现前列腺钙化或结石必须治疗，钙化会发展成结石，引发各种症状，有的症状长期无法消除，需要做全面检查，查看是否有结石钙化；不治疗结石钙化，前列腺疾病难以彻底治愈。保养治疗需要劳逸结合，避免过度疲劳，并进行适度的体育运动，尤其要加强盆腔肌肉锻炼；避免久坐和长时间骑车。 ，避免久蹲，排便时间控制在3到5分钟，避免坐在潮湿的地方。便后清洁肛门。注意饮食，多喝水，忌酒及辛辣食物。多吃蔬菜、水果及坚果类食物。坚果类食物富含铜和锌，对前列腺有益。
排名第一：	如果只是单纯的钙化灶，那么不必担心，多数可能是以前患过炎症，治愈后留下的痕迹。可以先做一次前列腺液常规检查。 原问题：再次检查发现前列腺实质内有钙化灶，这是什么问题？需要注意什么？想治好。
排名前二	你好，根据你的情况来看，单纯的钙化灶一般不会有什么症状，也无需治疗。如果合并有前列腺炎，需要针对前列腺炎进行治疗。治疗可使用某些抗生素、中药等，并进行温水坐浴。避免过度劳累、熬夜、饮酒及食用辛辣刺激性食物等。同时也要注意饮食和休息。如果情况严重，请及时就医。祝你早日康复。 原问题：最近到医院检查身体，发现自己有前列腺增生伴钙化，小时候曾经患过前列腺炎，不过后来发现的。虽然已经治愈了，但现在看来是不是以前留下来的后遗症？请问我这种前列腺增生伴钙化怎么治疗？可以治愈吗?
前3名：	前列腺钙化灶这种情况是之前炎症导致的瘢痕，目前要注意再检查一下是否还有尿路感染等情况。的情况，可以通过药物进行治疗。要注意多喝水，不要憋尿，还要避免辛辣刺激性食物，平时的话，注意增强个人体质来改善。 原问题：18岁前列腺钙化灶怎么办?
问题	什么是生物药剂学?
答案	生物药剂学是研究给药后药物吸收的整个体内过程，包含各种制剂因素和生物因素对这一过程与药效的影响。此外，生物药剂学通过研究药物对生物细胞产生的反应过程，达到施药者的预期目的。20世纪50年代初，人们普遍认为“化学结构决定药效”，药剂学只是为了改善外观、掩盖不良气味，便于服用。随着大量临床实践的证明，人们逐渐认识到剂型和生物因素对药效的影响。因此，研究药物在代谢过程中的各种机制和理论，以及各种剂型和生物因素对药效的影响，对于控制药品的内在品质、确保最终药品安全有效，以及为新药开发和用药提供严格的评价，都具有重要意义。
Top 1	生物药理学是生物制药和医药生物技术交叉领域中的一门药理学与生物技术学科，被认为是一门新兴科学。包括在生物反应器中获取具有生物起源的药物。采用这条路线而非化学合成的一个主要优势是可以避免消旋产物，从而获得大量易于纯化的同类产品，提高性能并降低成本。另一个优势是能够获得化合物，几乎无法通过其他方式获得如此大量的重组蛋白。这有助于科学研究以及新疗法的设计与开发。 原问题：什么是生物药理学?
Top 2	生物制药，又译基因制药或药耕，是遗传工程学的一种利用基因改造的动植物生产药物的方法。以此方式生产的通常是重组蛋白质或者其代谢产物。重组蛋白质通常在生物反应器皿中通过细菌和酵母生产，但通过生物制药无需昂贵的基础设备，产能可以以更低的成本按需调整。 原问题：什么是生物制药?
前 3 名	原答案

表 11：DPR模型在从《百科全书》生成的问题上的检索结果示例 a

问题	脓腔穿刺的辅助治疗有哪些?
答案	消毒隔离；皮肤护理；营养支持
排名第1	不留死腔；引流通畅；支管开窗引流；了解脓腔范围 原问题：黏膜下脓肿的辅助治疗有哪些?
排名第2	保持引流通畅；护理干预；严格拔管；严格无菌操作；保持引流瓶高度适宜 原问题：双侧脑室外引流的辅助治疗有哪些?
前3名	破膜;程控降温法;换液;选择合适的供者和采集方法;康复训练;洗血;巴氏消毒;贴壁法;55°C水浴加热 原问题：脐血的辅助治疗有哪些?
问题	气道吸痰的辅助治疗有些什么?
答案	足量补液
排名第1	加温湿化器联合密闭式吸痰法 原问题：神经外科重症监护病房人工气道的辅助治疗有哪些?
前 2 名	平卧位;床头抬高30°到45°体位;按需吸痰原则 原问题：胃肠反流的辅助治疗有哪些?
前 3 名	常规雾化排痰法;气管按压法;吸痰管刺激法 原问题：诱导性主动咳嗽的辅助治疗有哪些?

表 12：DPR模型在基于知识库生成的问题上的检索结果示例

问题：我可能得了戊型肝炎，因为我饮食越来越不好，而且吃饭的时候总是想吐。请问戊型肝炎早期的症状是什么呢? 答案：肝病早期可能会出现脂肪无法消化，导致大量脂肪堆积在体内，同时胃功能也会紊乱，从而出现厌食 ， 看到油腻的食物就感到恶心，还可能出现皮肤色素沉着、指甲颜色改变、毛发改变等症状。 第1条：戊型肝炎通常起病较急，发病初期可能会出现恶心、呕吐以及轻微黄疸等症状。这种疾病主要 通过粪口传播途径传播，并且老年人、孕妇或有乙型肝炎基础的患者发病率通常较高。这种疾病通常早 期应严格卧床休息，直到症状消失后，才可以逐渐恢复正常活动。 原问题：我最近听说我朋友得了戊型肝炎，我不太了解这种疾病，它是不是病毒性肝炎? Top 2 戊型肝炎主要经粪—口途径传播。大多数报道的暴发性流行都与饮用受粪便污染的水有关，大规模暴发常常是 在暴雨与洪水发生后，水源受到污染时出现，多见于秋冬季。在环境与水源卫生状况较差的地区，也可散发 ， 全年都有散发病例。此外，还可通过日常生活接触和输入性传播。症状可能会出现发热、头痛、咽痛、鼻塞、 呕吐、上腹部不适、肝区疼痛、腹胀、腹泻等。每个人的体质和病情不同，症状也会有所不同。 原问题：我最近听说很多人得了戊型肝炎，我也想预防一下，想了解一下戊型肝炎的症状和病因。 排名前三的戊型肝炎其实是由肝炎病毒所致的全身性传染病，主要累及肝脏。其临床表现为食欲减退、恶心、乏力、上 腹部饱胀不适、肝区疼痛、肝肿大、压痛及肝功能损害等，部分病例出现黄疸。 原问题：我体检时检查出戊型肝炎，但是我平时生活挺规律的，想要知道戊型肝炎出现的原因有哪些呢?
问题 3岁婴儿把整个水果糖吞了，怎么才能知道是吞下去了呢? 答案：只要婴儿吞咽后没有屏气的现象，那就是吞下去了。 排名第一 就目前的情况，首先要确认是否已经吞下。一般情况下，婴儿都会有所感觉，比如出现腹痛、呕吐等。 原问题：13个月大的婴儿，昨天发现窗帘上的小挂钩少了一个，怀疑婴儿误吞了，需要去医院做什么检查吗? Top 2 既然能够咽下去，应该就没事。你可以注意观察孩子的呼吸状况和面色状况。如有异常，立即就诊。 原问题：一岁八个月大的婴儿吃果冻时噎住，后来又咽下去了，刚才又喝了点水，有事吗? Top 3 要看果核的大小，一般都可以排出来。可以密切观察孩子的进食情况，只要吃得好、不呕吐，就没问题。如果进 如果食欲不振或呕吐，就要去医院检查了。 原问题：十个月大的婴儿吞下荔枝果核有没有事? 急求答案

表 13：DPR模型对咨询记录中的问题进行检索的结果示例