· 数据智能与知识服务 ·

中医古文相似度计算研究：一种以生成式 AI 融合领域知识的 SimCSE 方法

张君冬[1] 刘江峰[1,2] 邓景鹏[3] 刘艳华[4] 黄 奇[1*]
（1. 南京大学信息管理学院，江苏 南京 210023；
2. 南京大学数据智能与交叉创新实验室，江苏 南京 210023；
3. 武汉大学信息管理学院，湖北 武汉 430064；
4. 南京中医药大学卫生经济管理学院，江苏 南京 210023）

摘 要：［目的/意义］为构建专门适用于中医古籍文本的相似度计算模型，解决 BERT 在中医古籍文本上语义表征困难和数据标注成本高昂的问题。［方法/过程］本文在多个模型增量预训练的基础上，利用生成式 AI 生成全部任务数据，结合 SimCSE 方法，对比不同训练方式、预训练模型、正负样本构造方法、正样本混合策略的作用。［结果/结论］研究结果显示，无监督学习模型性能普遍偏低，引入 AI 生成的正负样本对后性能明显提升。其中，使用 AI 构建的语义不同的、相似性较低的负样本，并与采用 AI 辅助的同义词替换方法构建的正样本混合而成的训练集上，TCM-Gujiroberta 模型性能最佳，达到 90.9%；此外，选择相似性较低的负样本并随机混合不同类型正样本的数据集可进一步提升模型性能。本研究在零样本情境下，设计出一种融合中医古籍知识的 SimCSE 相似度计算模型，可为古籍研究和应用提供支持，未来考虑在数据集构建策略方面进一步优化。

关键词：中医古籍；相似度计算；预训练语言模型；SimCSE；AIGC
DOI：10.3969/j.issn.1008-0821.2025.04.005
〔中图分类号〕TP18；TP391.1 〔文献标识码〕A 〔文章编号〕1008-0821（2025）04-0049-11

Research on Similarity Calculation of Traditional Chinese Medicine Classical Prose：A SimCSE Approach to Fusing Domain Knowledge With Generative AI

Zhang Jundong[1] Liu Jiangfeng[1,2] Deng Jingpeng[3] Liu Yanhua[4] Huang Qi[1*]
（1. School of Information Management，Nanjing University，Nanjing 210023，China；
2. Data Intelligence and Cross-Innovation Laboratory，Nanjing University，Nanjing 210023，China；
3. School of Information Management，Wuhan University，Wuhan 430064，China；
4. School of Health Economics and Management，Nanjing University of Chinese Medicine，Nanjing 210023，China）

Abstract：［Purpose/Significance］The study aims to construct a similarity computation model specifically applicable to TCM ancient texts，solve the problems of difficulty in semantic characterization and the high cost of data annotation of BERT on TCM ancient texts.［Method/Process］On the basis of incremental pre-training of multiple models，this study

收稿日期：2024-07-15
基金项目：江苏省研究生科研与实践创新计划项目“图模驱动的在线医疗健康智慧问答服务研究”（项目编号：KYCX24_0107）；江苏高校哲学社会科学研究重大项目“中医古籍文献预训练模型构建及其应用研究”（项目编号：2023SJZD084）。
作者简介：张君冬（1996-），男，博士研究生，研究方向：医疗数据再组织。刘江峰（1998-），男，博士研究生，研究方向：科学计量，科技文本挖掘，计算人文。邓景鹏（1996-），男，博士研究生，研究方向：古籍整理与数字人文。刘艳华（1985-），女，副教授，博士，研究方向：社会科学研究评价与数字人文研究。
通信作者：黄奇（1961-），男，教授，博士，博士生导师，研究方向：网络信息资源的语义化。

used generative AI to generate all task data，and combined SimCSE method to compare the effects of different training methods，pre-training models，positive and negative sample construction methods，and positive sample mixing strategies.［Result/Conclusion］The research results show that the performance of unsupervised learning models is generally low，and the introduction of positive and negative samples generated by AI significantly improves the performance. On the training set composed of semantically different and low similarity negative samples constructed using AI，and mixed with positive samples constructed using AI assisted synonym replacement method，the TCM Gujiroberta model showed the best performance，reaching 90.9%. In addition，selecting negative samples with low similarity and randomly mixing datasets with different types of positive samples can further improve model performance. This study designs a SimCSE similarity calculation model that integrates knowledge of traditional Chinese medicine ancient books in a zero sample scenario，which can provide support for the research and application of ancient books. In the future，further optimization would be considered in terms of dataset construction strategies.

Key words：traditional Chinese medicine ancient books；similarity calculation；pre-trained language model；SimCSE；AIGC

中医古籍，中华文化之瑰宝，载千年医道之精粹，古籍所书，不独医药方剂，更含养生之术，辩证之法，治病之经，皆以不朽之经典，传承至今。文本相似度技术对古籍整理、文献溯源、文献查找等方面具有重要意义：①集注集释整理，可精准比对不同古籍中的相似文段，极大提升整理古籍时的效率和准确性，从而为研究者呈现更为清晰、完整的中医知识体系；②文本生成溯源方面，则助力追踪和分析特定医学理论或治疗方法的发展历程，揭示中医学术思想的演变和流变；③对于重出文献的追寻和查找，文本相似度计算能有效识别并对比古籍中的相似或重复内容，便利版本比较与校勘工作。然而，中医古籍的文本内容涵盖多个世纪的医学知识和实践经验，包含了大量特殊术语和古代汉字，这也使得传统的自然语言处理(NLP)方法不能胜任，因此，如何构建适用于中医古籍领域的相似度计算模型已成为一个重要的研究问题。

SimCSE(Supervised and Unsupervised Improved Contrastive Sentence Embedding)作为主流的相似度计算方法，已在多个领域文本相似度计算任务中取得显著效果，主要分为有监督和无监督两种。有监督的 SimCSE 相较于无监督能够更准确地捕捉语义信息，效果也更为可靠，但需标注一定规模的高质量数据作为训练集，而对于本文的中医古文相似度任务，其痛点在于，一方面，市面上并无开源的中医古文相似数据集，若采取人工标注则需标注者在具备古文理解力的同时具备强大的中医知识基础，速度慢，产能低；另一方面，SimCSE 方法基于预训练语言模型，而现有的语言模型多以通用古籍类为主，针对中医古籍这一细分领域，尚未有相关的模型。

随着以 ChatGPT、ChatGLM 为代表的生成式大语言模型取得飞速突破，自然语言处理也迎来新的研究范式和多样化选择。大语言模型能够根据用户输入的 Prompt 提示词，利用自身强大的语言理解和生成能力给出流畅通顺的回答[1]。在此情境下，采用 AI 生成的自监督标注(Automated Supervision by AI)方法来取代传统有监督人工标注下游任务训练集成为一大可能。

结合上述情况，本文在多个通用古籍模型增量预训练的基础上，利用 AIGC 技术生成全部下游任务数据，在此基础上结合 SimCSE 对比学习方法，设计出一种针对中医古籍领域的古文相似度计算模型。本文主要贡献在于：①对现有多个通用古籍 BERT 模型进行增量预训练，获得适用于中医古籍领域的 BERT 模型，以更好地表示中医古籍语义文本特征。②针对中医古籍领域暂无公开数据集且标注成本高的情况，利用生成式 AI 技术，构建适用于中医古籍领域对比学习的正负样本训练集，极大地减轻了人工标注工作量。③首次提出针对中医古籍领域的古文相似度计算模型，实验比较了不同训练方式、不同预训练语言模型、不同种类 Prompt 提示词构建的正负样本进行对比学习的效果，探讨了不同正样本混合方式对模型性能的提升策略，证明了在零样本训练集条件下，基于 AIGC 的样本训练数据构造方法具备一定的可行性，效果显著优于传统的无监督对比学习。

1 相关研究

1.1 从浅层距离到深度语义探索：文本相似度研究历程

文本相似度是一种用于确定两个或多个文本之间语义或结构相似性的任务。早期的文本相似度方法大多是通过度量文本间的距离进行计算，如 SimHash[2]、BM25[3]等。随着特征工程的兴起，文本相似度领域开始通过构建合适的特征来将文本表示为词向量或句向量，并使用向量之间的距离或相似性度量来衡量文本的相似程度，如词袋模型、TF-IDF[4]、N-gram[5]等。这类方法在一定程度上提高了文本相似度的效果，但难以表示文本中的全部语义信息，因此实际效果并不显著。再后来，利用诸如 Word2vec[6]、GloVe[7]等词向量模型进行文本表示更具便捷性，可以自动学习语义特征表示，逐渐取代了相对繁琐的特征工程方法。近年来，随着预训练语言模型技术的迅猛发展，研究者们开始利用 BERT 模型[8]提取文本的上下文语义信息，并使用其生成的语义向量进行相似度量，这也使得文本相似计算效果得到进一步提高。已有实验证实[9]，相较于利用 BERT 直接获取语义向量表示的方法，SimCSE 方法可有效解决向量表达存在各向异性以及向量分布不均匀的情况，能更好地学习到句向量表征，可进一步优化 BERT 模型在文本相似度计算中的应用效果，提高模型的鲁棒性和泛化能力，这也为文本相似度计算提供了更为强大和可靠的工具。

1.2 零样本资源下的智慧启迪：AIGC 赋能 NLP 经典任务

自 ChatGPT 问世以来，凭借其深厚的语义理解和智能推理能力，实现了对复杂语境的准确把握与敏锐回应，从而在对话生成和智能问答中显露出独特优势，同时也为 NLP 经典任务注入了新的活力与可能。当前，就生成式 AI 技术能否直接应用于 NLP 经典任务，相关学者以此为契机进行了探索。如，张华平等[10]在零样本资源情况下使用 9 个数据集评估 ChatGPT 的中文表现性能，发现在 NLP 经典理解式任务上表现较好，在情感分析上具有 85%以上的准确率，在闭卷问答上出现事实性错误的概率较高。鲍彤等[11]评估 ChatGPT 在典型中文信息抽取任务中的性能，发现 ChatGPT 在事件抽取中具有较好的表现，在命名实体识别、关系抽取中的效果与中文预训练模型存在较大差距。

上述研究表明，生成式 AI 技术在传统理解式任务上表现出优异性能，但对于复杂场景下特定领域的判别式、抽取式任务上，效果并不领先，此后部分学者开始尝试利用 AIGC 技术简化传统 NLP 方法处理流程，如，张恒等[12]针对研究流程段落识别任务，在 SciBERT 模型的基础之上，利用 ChatGPT 通过数据增强，显著提高了分类的准确率和 F1 值。因而，本研究认为，在大语言模型无法很好地胜任且传统 NLP 处理方法又缺少标注数据的情境下，利用生成式 AI 技术构建样本训练集具备一定的可行性。

2 研究介绍

2.1 研究框架

本文的研究框架，如图 1 所示，主要分为语料收集及预处理、融合领域知识的继续预训练、基于对比学习的 SimCSE 相似度计算 3 个部分；①语料收集及预处理，通过模拟鼠标键盘点击的方式爬取《中华医典》数据库中的所有数据，之后通过进行数据清洗以形成继续预训练所需的中医古籍纯文本语料；②融合领域知识的继续预训练，选择多个通用古籍语言模型进行继续预训练，采用 10%的中医古籍语料作为验证集，使用困惑度指标(PPL，Perplexity)初步评估模型的性能；③基于对比学习的 SimCSE 相似度计算，通过设计不同的 Prompt 模板，采用 AI 技术构建不同种类的正负样本对作为训练集，在多个模型继续预训练的基础上结合 SimCSE 方法进行多次实验对比，同时探讨不同正样本混合策略对模型性能的提升。

2.2 融合领域知识的继续预训练

BERT 模型作为一种自监督学习的语言表示模型，已在许多自然语言处理任务中取得显著的成就。然而，当被应用到具有特定领域知识的任务时，其通用的预训练框架可能不足以捕获领域特有的语义细节，领域知识融合则是将任务相关的数据或特定领域的知识引入预训练模型，使其能够更好地理解语境和上下文，从而提升其特定任务的性能表现。如，赵一鸣等[13]将医学信息查询相关的语料对 BERT

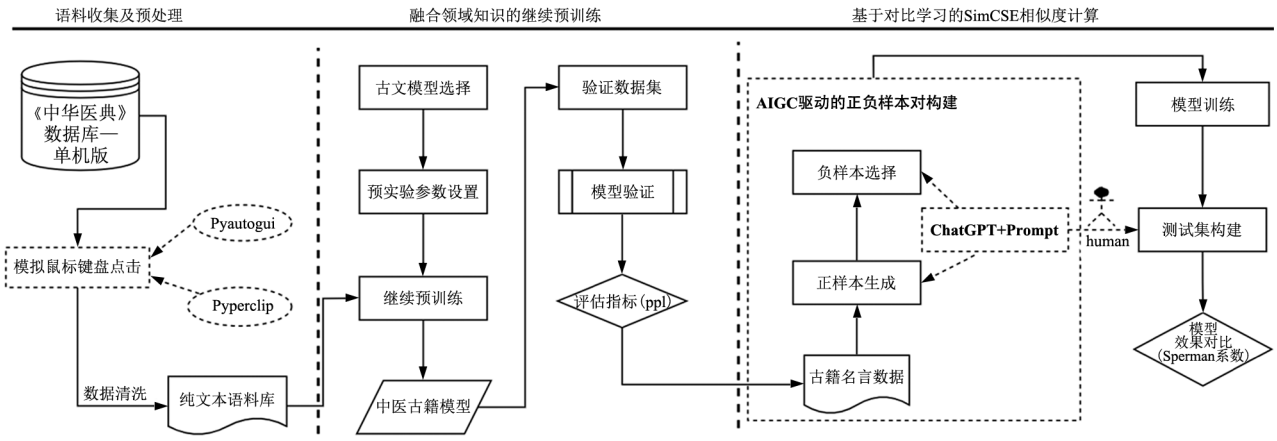

图 1 研究框架图

Fig. 1 Research Framework Diagram

模型进行继续预训练，在较低的资源和时间成本下获得预训练模型 MQ-BERT，使其更好地表征医学信息查询式的词向量，以适应意图强度识别任务。

2.3 基于对比学习的 SimCSE 相似度计算

SimCSE 是一种在预训练语言模型的基础上，通过对比学习来提高相似度计算效果的方法，其训练主要分为无监督和有监督两种方式。无监督的 SimCSE 采用 Dropout 作为简单的数据增强技术，通过对同一个输入句子进行两次前向传播以产生两个略有差异的正样本，同时使用与输入句子长度不同的负样本来进行训练，其弊端在于训练完成的模型倾向于认为长度相近的句子在语义上也更为相似。有监督的 SimCSE 需要一定规模的精加工标签数据集，使用预先定义的正样本对和负样本对来训练。由于直接从标记数据中学习，有监督的 SimCSE 能够更准确地捕捉语义信息，与无监督相比，效果更为可靠。

3 融合中医古籍知识的继续预训练实验

3.1 实验语料收集

本文所进行的继续预训练实验语料来源为“九五”国家重点电子出版规划项目的重要成果《中华医典》数据库。该数据库按图书馆分类法将历代中医古籍分为医经、诊法、本草等 12 个大类，条理清晰、泾渭分明，涵盖到民国为止的中国传统医学文化建设的主要成就，卷帙上万，是目前市面上规模最为宏大的中医古籍类电子丛书[14]。

表 1 列出《中华医典》各类目具体数量及字数。从字数统计结果来看，共67 346 246个汉字，单本古籍字数最少的为临证各科类目，最多的为方书；不重复汉字共8 628个，各个类目不重复汉字数均占 50% 以上，反映出中医古籍用词凝练度高，专业术语集中性强。

表 1 《中华医典》书籍类目及字数概况

Tab. 1 Overview of Book Categories and Word Count in the Chinese Medical Code

古籍类目	数量	汉字数：67 346 246			不重复汉字数：8 628		
		汉字数	最小值	最大值	汉字数	最小值	最大值
1. 医 经	52	3 486 166	4 863	604 003	4 675	482	3 096
2. 诊 法	50	1 365 141	2 133	170 794	6 061	514	3 231
3. 本 草	76	8 286 760	8 115	1 339 977	6 368	862	4 836
4. 方 书	145	802 282	2 329	1 611 436	4 708	565	3 606
5. 针灸推拿灸	53	1 378 101	1 824	274 529	4 639	387	3 348
6. 伤寒金匮	78	5 985 719	3 162	355 688	6 138	517	3 441
7. 温 病	49	10 202 425	3 494	115 383	6 023	645	2 687

表 1（续）

古籍类目	数量	汉字数：67 346 246			不重复汉字数：8 628		
		汉字数	最小值	最大值	汉字数	最小值	最大值
8. 综合医书	107	2 054 613	8 118	1 259 730	4 822	1 119	4 495
9. 临证各科	300	3 229 364	1 120	861 878	4 761	377	3 904
10. 养生食疗外治	45	1 330 476	1 249	173 437	4 643	357	4 629
11. 医论医案	194	14 337 320	1 661	756 521	6 220	707	4 248
12. 其　他	7	14 887 879	23 691	542 201	6 287	1 555	4 392

3.2 实验评测指标

困惑度（PPL，Perplexity）作为一种衡量语言模型预测样本概率的指标，被广泛应用于各类预训练任务的评测中。理论上来讲，困惑度越低，模型的性能越好，对数据的不确定性越小，如式（1）所示：

$$PPL(W)=P(w_1,w_2,\cdots,w_N)^{-\frac{1}{N}} \tag{1}$$

其中，W 表示整个词序列，$w_1,w_2,\cdots,w_N$ 是序列中的词，N 是序列中词的总数，$P(w_1,w_2,\cdots,w_N)$ 是模型赋予这个词序列的概率。

3.3 基线模型介绍

尽管现有的古籍语言模型在古籍领域表现出一定的普适性，但应用于更加专业和细分的自然语言处理任务时，其性能往往受到限制。因此，面对中医专业知识密集的中医古籍领域，有必要在通用古籍模型的基础上进行继续预训练。

基线模型选择方面，笔者综合考察了现有古籍方面的 NLP 任务所用模型，发现 guwenbert-base、SikuBERT、SikuRoBERTa 这三类模型所用居多，如刘江峰等[15]对典籍文本进行命名实体识别，张逸勤等[16]针对跨语言典籍进行跨语言风格计算，均采用了上述 3 种模型进行对比。与前人已有研究略有区别的是，本文在选择前面三类模型的基础上新增 Gujibert、Gujiroberta 两种模型进行对比，其主要原因在于这两种模型在继续预训练过程中语料类型较为特殊，为简繁混合型，而本研究的中医古籍语料分布年代各异，简繁体众多，若采用现有软件全部统一为简体或繁体，难免出现遗漏。考虑到上述情况，本研究最终选择以下 5 种基线模型，如表 2 所示，在此基础上进行继续预训练，旨在开发出更加适应中医古籍的预训练模型，以在下游文本相似度计算场景中取得更好的性能表现。

表 2 古籍模型介绍

Tab. 2 Introduction to Ancient Book Models

古籍模型	语料来源	语料类型	语料大小	基线模型
guwenbert-base	殆知阁古文献数据	简　体	约 17 亿	RoBERTa
SikuBERT	《四库全书》	繁　体	约 5.36 亿	BERT
SikuRoBERTa	《四库全书》	繁　体	约 5.36 亿	RoBERTa
Gujibert	殆知阁古文献数据	简繁混合	约 17 亿	BERT
Gujiroberta	殆知阁古文献数据	简繁混合	约 17 亿	RoBERTa

GuwenBERT-base 基于 RoBERTa 模型，由北京理工大学阎覃[17]开发构建古汉语预训练语言模型。该模型使用的训练数据为殆知阁古代文献数据集，包含15 694本古典中文书籍，涵盖佛教、儒家、历史等多个领域，总共有大约 17 亿个字符，同时在继续预训练过程中，所有传统字符都经过简体转换处理。SikuBERT 和 SikuRoBERTa 是由南京农业大学信息管理学院开发的针对古文文本自然语言处理的预训练语言模型[18]，采用校验后的高质量《四库全书》总共约 5.36 亿字繁体语料作为训练集，其中，SikuBERT 基于 BERT 中文模型框架预训练，SikuRoBERTa 则在 RoBERTa 模型的基础上继续预

训练。Gujibert 和 Gujiroberta 两类模型与 SikuBERT 和 SikuRoBERTa 训练过程基本相似[19]，但与 Siku 系列模型相比，两者的训练来源有所不同，与 guwenbert-base 模型相比，其不同点则是在于训练语料类型为简繁混合型。

3.4 实验参数设置

研究设置了一系列超参数，如表 3 所示，其中，学习率(Learning Rate)决定了模型权重更新的速度，将其设置为 5e-05，有助于模型在学习过程中稳定地调整和优化；训练轮数(num_train_epochs)设定为 3，确保模型有足够的时间学习古籍文本的细微特征，防止因过多训练轮次而引起过拟合；设置梯度累积策略(gradient_accumulation_steps)为 4，可有效批量训练，从而优化内存使用并提升模型性能。

表 3 实验超参数设置

Tab. 3 Experimental Hyperparameter Settings

参 数	值
learning_rate	5e-05
num_train_epochs	3.0
per_device_train_batch_size	16
per_device_eval_batch_size	16
gradient_accumulation_steps	4
logging_steps	1 000
save_steps	20 000

3.5 继续预训练实验结果

针对不同模型训练所需的语料类型，通过 OpenCC 包对训练语料进行简繁互换后，在 5 种通用古籍 BERT 模型上进行继续预训练，并在相应的验证数据集上进行了性能评估。继续预训练所获得的模型分别命名为 TCM-guwenbert-base、TCM-SikuBERT、TCM-SikuRoBERTa、TCM-Gujibert、TCM-Gujiroberta。从表 4 实验结果看，各个预训练模型都取得了相对不错的效果，TCM-SikuBERT 最好，为 5.928，TCM-guwenbert-base 最差，为 6.495。

表 4 继续预训练实验结果

Tab. 4 Continuing Pre-training Experiment Results

模 型	验证数据集	困惑度
TCM-guwenbert-base	10%中医古籍语料	6.495
TCM-SikuBERT	10%中医古籍语料	5.928
TCM-SikuRoBERTa	10%中医古籍语料	6.200
TCM-Gujibert	10%中医古籍语料	6.330
TCM-Gujiroberta	10%中医古籍语料	6.041

4 中医古文相似度计算实验

4.1 实验样本来源

古籍中的普通语句较为通俗，缺乏深层次的哲理内涵，这可能不利于模型捕捉语言的深层含义；而古籍名言因其深刻的意义和精辟的表达，往往被后世频繁引用，具有很高的辨识度和丰富的文化背景。因此，本文选择《中华医典》数据库<辞典>类目下的所有古籍名言作为实验样本，使模型更加集中于理解中医古文的语义特征和文化内涵上，而非仅仅是语言形式上的相似性，以增强模型在实际应用中的识别力。

为保证匹配的粒度相对统一，减少句长差异带来的干扰，本研究将名言长度大致限定在 8~30 字的范围内。如果长度超过 30，那么按照句子中间较大语义停顿的标点符号(如句号、感叹号、问号等)进行分句，最终获得3 036条中医古籍名言。表 5 列出部分中医古籍名言示例及出处。

表 5 部分中医古籍名言示例及出处

Tab. 5 Examples and Sources of Famous Sayings in Some Ancient Chinese Medicine Books

名 言	出 处
一阴一阳谓之道，偏阴偏阳谓之疾	宋·成无己《注解伤寒论·辨脉法》
一身之气，皆随四时五运六气兴衰而无相反也	金·刘完素《素问玄机原病式·热类》
二便之开闭，皆肾脏之所主	明·张介宾《景岳全书·泄泻》
十剂以准规矩，七方以明绳墨	清·李用粹《同德堂医案·序》
八脉隶于肝肾	清·叶桂《临证指南医案·调经》

4.2 基于 AIGC 的正负样本对生成

4.2.1 正负样本对构造方式

正样本对：在不改变语义的情况下，基于 AIGC 的方式通过同义词替换(Chat_SR)、随机插入(Chat_RI)、随机交换(Chat_RS)、随机删除(Chat_RD)、混合改写(Chat_RW)5 种方式生成相似样本。

负样本对：当前主流的负样本构造方式往往采用随机选取样本中句子作为负例，其问题在于随机选的负样本太容易区分，无法最大程度提升模型性能，因此本研究采用这一方法的同时新增一种方式进行对比，即通过 AIGC 生成一个句式相同但语义不同的低相似样本。

4.2.2 Prompt 提示词构建步骤

尽管以 ChatGPT 为代表的 AIGC 技术功能强大，但其效能的发挥仍然依赖于精心设计的 Prompt 提示词。Prompt 提示词可以被视为一种机器人响应的指令或问题，引导 AI 沿着用户的意图进行思考，以生成用户期望的回答。一个优质 Prompt 提示词可以减少歧义，提升答复的相关性与准确性，使 AI 机器能够精准把握用户意图，生成包含洞察力的回答。

在 Prompt 提示词工程中，需要考虑问题的背景、语境，以及问题的明确性、信息的完整性、关键词的使用、逻辑的清晰性、期望的回答类型等诸多方面。在多数情况下，Prompt 的性能上限与对“好结果”的理解程度成正比，只有充分理解所谓的“好结果”具体好在哪些“点”，才能将这些“点”形式化为 Prompt，从而把用户的意图更准确地传达给模型。基于以上要求，本文基于种子样本，在不同的任务需求下，根据 ChatGPT-4 构建用于古文正负样本对生成的 Prompt，主要分为以下五步工作，图 2 列出 AI 机器基于“<同义词替换>”这一构建方式生成相似古文样本。

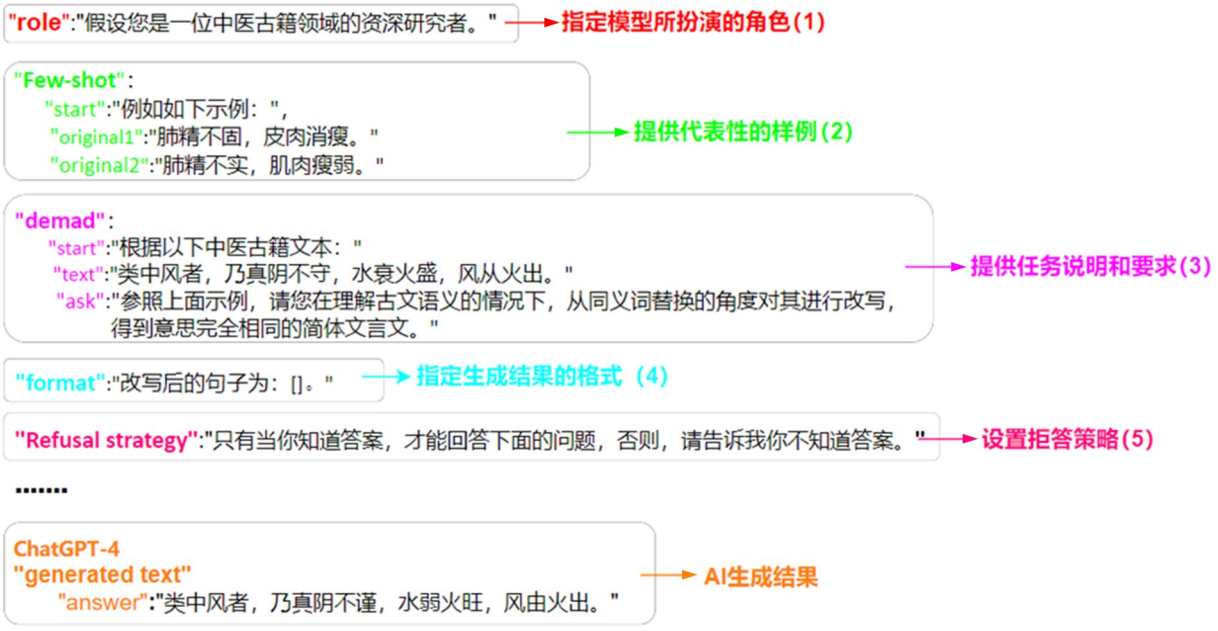

图 2 机器基于“<同义词替换>”这一构建方式生成相似古文样本

Fig. 2 Machine Generated Similar Ancient Text Samples Based on the Construction Method of “<Synonym Replacement>”

1）指定模型所扮演的角色(Role)：指定模型扮演的角色/身份以帮助模型更好地定位答复的内容与风格。

2）提供代表性的样例(One-shot Prompt)：为 ChatGPT 提供一个答案的参考，使其充分理解要求，提升模型性能表现。

3）提供任务说明和要求(Demand)：提供给 ChatGPT 语句流畅、意图清晰、表达精简的任务描述。

4）指定生成结果的格式(Format)：通过显示规定模型返回结果的格式，以便于后续统计分析。

5）设置拒答策略(Refusal Strategy)：虽然 ChatGPT 设置了诸如“我的知识截至 2021 年 9 月…”“作为一个人工智能模型…”这样的拒答策略，但仍旧无法完全避免大模型胡说八道。本文尝试手动设置拒答策略，即让模型在没有把握的时候拒绝回

答问题，提高生成数据的质量。

4.2.3 基于 AIGC 的正负样本对生成结果

在正负样本对生成结束后，通过 AI 的方式对所有句子对进行排序打分，赋值范围为 0~5，其中 0 代表完全不相似，5 代表完全相似。研究对每对句子进行 5 次评估打分并取平均值，以此抵消单次评估中的随机波动或偏好倾向，在此基础上结合人工 2 次评估调整得到最终评分。表 6、表 7 列出了不同构建方式下正负样本的生成示例，表 8 为训练数据集 AI 打分结果示例。

表 6 不同构建方式的古文正样本生成示例

Tab. 6 Examples of Generating Authentic Ancient Texts Using Different Construction Methods

构建方式	原始样本	生成正样本
同义词替换	七年之病，求三年之艾	七载之疾，求三岁之艾
随机插入	人之一身，脾胃为主	人之于一身，脾胃为主也
随机交换	二便之开闭，皆肾脏之所主	开闭之二便，皆肾脏之所主
随机删除	七情之伤，虽分五脏，而必归本于心	七情伤，虽分五脏，必归本心
混合改写	肺精不固，皮肉消瘦	肺精不交，肌肉消瘦也

表 7 不同构建方式的古文负样本生成示例

Tab. 7 Example of Generating Negative Samples of Ancient Texts Using Different Construction Methods

构建方式	原始样本	生成负样本
随机挑选	一阴一阳之谓道	求穴在乎按经
	七年之病，求三年之艾	嗜欲喜怒之情贤愚皆同
	有胃气则生，无胃气则死	阴不可无阳，阳不可无阴
AIGC 驱动的语义不同的低相似样本	一阴一阳之谓道	一朝一夕之谓彼方
	七年之病，求三年之艾	七度之音，望三方之和
	有胃气则生，无胃气则死	有胃气则死，无胃气则生

表 8 训练数据集 AI 打分结果示例

Tab. 8 Examples of AI Scoring Results for Training Dataset

句子 1	句子 2	标签
一阴一阳之谓道	求穴在乎按经	0
肝肾在下，藏精血也	肝藏肾下，血精也	1.8
一阴一阳之为道	一阴一阳结，谓之喉痹	2.5
呼出之气，由心达肺	呼出为阳，心肺主之	3.7
凡此十二官者，不得相失也	十二官相失，则精神日消	4.3
人之一身，脾胃为主	人之于一身，脾胃为主也	5

4.3 文本相似度计算评估指标

斯皮尔曼相关系数（Sperman Correlation）被用来衡量模型产生的排序结果与数据集中标注的参考排序之间的相关程度[20]。取值范围在-1~1 之间，1 表示完全正相关，-1 表示完全负相关，0 表示没有相关性。其主要优势在于直接从排序角度评价模型的性能，不依赖于具体的阈值设置，避免阈值选择的主观性和不确定性。具体来讲，通过将模型输出的相似度分数转换为一个等级序列，而将数据集中预先标注的“正确”排序作为另一个等级序列，然后通过计算这两个序列之间的相关性，以评估模型排序结果的准确性，如式（2）所示。

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad (2)$$

其中，r_s 是斯皮尔曼相关系数，d_i 是两个等级之间的差值，n 是数据点的数量。

4.4 实验参数设置

文本相似度实验参数设置，如表 9 所示，除学习率(learning_rate)、批量大小(batch_size)等常见指标外，设置 Dropout 比率为 0.1，以减少过拟合的风险，增强模型对未见数据的泛化能力；设置最大长度(max_len)为 100，保证模型可以处理不同长度的文本，同时优化内存利用率和计算效率；设置随机种子(seed)为 42，确保模型训练过程的可重复性，该指标影响着数据集的分割、权重初始化以及模型训练过程中的任何随机性决策，通过固定这个值，确保每次实验在相同的初始条件下进行，从而使不同实验间的比较成为可能。

表 9 实验超参数设置

Tab. 9 Experimental Hyperparameter Settings

参 数	值	参 数	值
learning_rate	3e-5	batch_size_eval	256
dropout	0.1	eval_step	100
epochs	3.0	max_len	100
batch_size_train	64	seed	42

4.5 文本相似度结果与分析

4.5.1 不同训练方式、不同预训练语言模型、不同正负样本构造方法对比学习实验结果

鉴于各个模型的困惑度指标值差距不大，因此，将上述 5 个继续预训练后的模型全部纳入，以更好地比对在下游古文相似度任务中的效果。本研究采用精确的人工标注方法来创建一个测试集，包含了 500 对真实中医古籍中的相似文本及相似得分(0 分~5 分)。表 10 列出了不同训练方式、不同预训练语言模型、不同正负样本构造方法对比学习结果。

表 10 不同训练方式、不同预训练语言模型、不同正负样本构造方法对比学习结果

Tab. 10 Comparison of Learning Results With Different Training Methods，Different Pre-trained Language Models，and Different Positive and Negative Sample Construction Methods

训练方式	负样本	正样本	TCM-guwen-bert-base	TCM-Siku-BERT	TCM-Siku-RoBERTa	TCM-Gujibert	TCM-Gujiroberta
无监督	-	-	0.712	0.734	0.736	0.745	0.748
AI 自监督	随 机	同义词替换	0.805	0.889	0.887	0.888	0.886
		随机插入	0.767	0.882	0.883	0.885	0.883
		随机交换	0.802	0.886	0.887	0.886	0.888
		随机删除	0.787	0.888	0.888	0.889	0.889
		混合改写	0.795	0.887	0.887	0.888	0.887
	AI 构建的语义不同的低相似样本	同义词替换	0.819	0.893	0.896	0.897	0.909
		随机插入	0.775	0.892	0.890	0.896	0.892
		随机交换	0.811	0.884	0.890	0.892	0.894
		随机删除	0.789	0.899	0.895	0.895	0.897
		混合改写	0.795	0.896	0.896	0.898	0.899

1）不同训练方式对比：从实验结果来看，无监督学习中，模型性能普遍最低，表明模型在缺乏明确的正负样本指导时难以捕捉到古文的深层语义信息。相对而言，当引入 AI 自监督学习，特别是结合随机负样本时，性能得到明显的提升。此外，基于 AIGC 技术生成构建语义不同的低相似度样本时，模型的性能得到最大程度的提升。这表明，通过 AIGC 构建的高质量负样本可显著提高模型的区分能力，在提升模型性能方面起到决定性的作用。

2）不同预训练语言模型对比：不同的模型展现出了性能差异，揭示了它们在处理古文语义上的不同能力。与其他模型相比，TCM-Gujiroberta 模型性能往往更好，这可能是由于 RoBERTa 架构在面对 AIGC 生成的高质量负样本时优化了对内部语义

关系的捕捉，从而对古文有更深刻的理解。TCM-SikuBERT 和 TCM-SikuRoBERTa 模型虽然也显示出良好的性能，但相比 Guji 系列模型略显不足，TCM-guwenbert-base 性能最低，这可能意味着该模型的结构、预训练数据或训练策略相对较简单，不足以充分捕捉中医古文的语义复杂性。

3）不同正负样本对构造方式对比：基于 AIGC 的正样本构建方式同样显著影响模型性能，同义词替换和混合改写，在所有模型上都表现出较高的性能，其主要原因在于这两种方法能够在保持原文语义的同时引入适当的变化，可有效帮助模型学习理解不同表达形式下的相同意义。随机插入策略的性能较低，可能因为它在古文中引入额外的噪声，从而降低模型的理解能力。相对而言，随机删除虽然也引入了一定的随机性，但由于它在减少原文内容的时候主要以无实际意义的虚词为主，对模型性能的影响较小。随机交换其性能则介于随机插入和随机删除之间，但这种策略有时可能扰乱文本原有的语义结构。

4.5.2 低相似负样本情况下不同正样本混合策略对各类模型效果的提升

从表 10 实验结果来看，高质量的负样本可显著提升模型的性能，而单一的正样本类型显然不能最大程度提升模型的性能，因此在低相似负样本情况下，选择上述结果较优的同义词替换（Chat_SR）、随机删除（Chat_RD）、混合改写（Chat_RW）3 种正样本构建方式，按照 20%、30%、50%的比例随机抽取混合，形成新的样本训练集进行实验，以更好地探讨不同正样本混合方式对模型性能的提升，具体结果如表 11 所示。

表 11 不同正样本混合策略对模型效果的影响

Tab. 11 The Impact of Different Positive Sample Mixing Strategies on Model Performance

不同正样本混合策略	TCM-guwen-bert-base	TCM-Siku-BERT	TCM-Siku-RoBERTa	TCM-Gujib-ert	TCM-Guji-roberta
20%（Chat_SR）+30%（Chat_RD）+50%（Chat_RW）	0.878	0.903	0.909	0.924	0.928
20%（Chat_SR）+30%（Chat_RW）+50%（Chat_RD）	0.841	0.899	0.901	0.903	0.913
20%（Chat_RD）+30%（Chat_SR）+50%（Chat_RW）	0.864	0.905	0.912	0.932	0.923
20%（Chat_RD）+30%（Chat_RW）+50%（Chat_SR）	0.893	0.912	0.907	0.917	0.921
20%（Chat_RW）+30%（Chat_SR）+50%（Chat_RD）	0.832	0.878	0.886	0.899	0.909
20%（Chat_RW）+30%（Chat_RD）+50%（Chat_SR）	0.891	0.906	0.901	0.917	0.914

实验结果表明，选择低相似负样本，并随机混合不同正样本后，各个模型的性能得到了进一步提升。其中，继续预训练后的 TCM-Gujibert 模型在 20%（随机删除）+30%（同义词替换）+50%（混合改写）的样本组合下效果最好，达到 0.932。此外，相同的混合策略下，各个模型性能表现差异显著，如混合改写（Chat_RW）占据主导地位（50%比例）时，TCM-Gujibert 模型在这种组合下分数最高，分别为 0.924 和 0.932，而 TCM-guwenbert-base 仅为 0.878 和 0.864。

5 总结与展望

文本相似度计算为古籍研究之要点。以此技艺，辨识古文之同异，穷尽文献之深意，如行云流水，得以串联历代典籍之相互关联，揭示古代学术之绵延不绝。相似度之运用，宛如慧眼，洞察文辞之微妙变化，观历史文化之深远脉络。由此览古今之变迁，探思想文化之演进，昭示人文社科研究之新径，开拓中医学术研究之新天地。

本研究设计出一种针对中医古籍领域的古文相似度计算模型，同时解决了通用 BERT 模型在中医古籍领域语义表征困难和下游数据标注成本高昂的问题。研究在现有五类通用古籍模型增量预训练的基础上，结合 SimCSE 方法，对不同训练方式、不同预训练语言模型、不同正负样本构造方法进行对比实验，并探讨低相似负样本情况下不同正样本混合策略对模型性能的提升。实验结果表明，无监督学习中，模型性能普遍偏低，当引入 AI 自监督生成的正负样本对后，模型性能得到明显的提升。其

中，AIGC 驱动的语义不同的低相似负样本结合同义词替换的正样本构成训练集后，TCM-Gujiroberta 模型表现最佳，为 0.909。此外，选择低相似负样本，并随机混合不同正样本，可进一步提升模型效果，如 TCM-Gujibert 模型在 20%(随机删除)+30%(同义词替换)+50%(混合改写)的样本组合下效果最好，达到 0.932。

本文设计了一种巧妙的数据标注方法，其优点在于无需人工标注任何训练数据，并通过大量对比实验验证了该方法的有效性。当然，由于实验和篇幅的限制，本研究仍然存在一些不足，后续将继续开展以下研究以补充和完善本文的工作。

1）在继续预训练语料的选取上，设计自动化的算法，如实施动态选择机制，根据模型在训练过程中的表现反馈调整语料选择，有效识别和选择那些对模型性能提升最有帮助的语料，同时减少对无效或低效语料的依赖，以提高预训练的效果。

2）数据构建策略方面，目前实验仅选择了 5 种 AI 生成的正样本构造方式，虽然这些构造方式有效，但可能无法覆盖中医古文的所有语义复杂性。后续将探索更多样化的 AI 数据构建技术，如针对不同朝代特定语言风格构建等，以更全面地覆盖中医古文的语义特点。

3）目前的分析主要集中在模型的表现层面，对于模型为何在特定数据组合策略下表现更佳的内部机制探索仍不够深入。未来将通过模型可视化和解释性分析，如注意力机制可视化等方法，观察模型在不同数据组合策略下的关键依赖点。

参 考 文 献

[1] 陆伟，刘家伟，马永强，等. ChatGPT 为代表的大模型对信息资源管理的影响 [J]. 图书情报知识，2023，40 (2)：6-9，70.

[2] Charikar M S. Similarity Estimation Techniques from Rounding Algorithms [C] //Proceedings of the Thirty-fourth Annual ACM Symposium on Theory of Computing，2002：380-388.

[3] Robertson S，Zaragoza H. The Probabilistic Relevance Framework：BM25 and Beyond [J]. Foundations and Trends in Information Retrieval，2009，3 (4)：333-389.

[4] Spärck J K. A Statistical Interpretation of Term Specificity and Its Application in Retrieval [J]. Journal of Documentation，2004，60 (5)：493-502.

[5] Shannon C E. A Mathematical Theory of Communication [J]. The Bell System Technical Journal，1948，27 (3)：379-423，623-656.

[6] Mikolov T，Chen K，Corrado G，et al. Efficient Estimation of Word Representations in Vector Space [EB/OL]. https://arxiv.org/abs/1301.3781，2025-02-21.

[7] Pennington J，Socher R，Manning C D. Glove：Global Vectors for Word Representation [C] //Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP)，2014：1532-1543.

[8] Devlin J，Chang M W，Lee K，et al. BERT：Pre-training of Deep Bidirectional Transformers for Language Understanding [EB/OL]. https://arxiv.org/abs/1810.04805，2025-02-21.

[9] Gao T Y，Yao X C，Chen D Q. SimCSE：Simple Contrastive Learning of Sentence Embeddings [EB/OL]. https://arxiv.org/abs/2104.08821，2025-02-21.

[10] 张华平，李林翰，李春锦. ChatGPT 中文性能测评与风险应对 [J]. 数据分析与知识发现，2023，7 (3)：16-25.

[11] 鲍彤，章成志. ChatGPT 中文信息抽取能力测评——以三种典型的抽取任务为例 [J]. 数据分析与知识发现，2023，7 (9)：1-11.

[12] 张恒，赵毅，章成志. 基于 SciBERT 与 ChatGPT 数据增强的研究流程段落识别 [J]. 情报理论与实践，2024，47 (1)：164-172，153.

[13] 赵一鸣，潘沛，毛进. 基于任务知识融合与文本数据增强的医学信息查询意图强度识别研究 [J]. 数据分析与知识发现，2023，7 (2)：38-47.

[14] 裘沛然. 中华医典 [M]. 长沙：湖南电子音像出版社，2014.

[15] 刘江峰，冯钰童，王东波，等. 数字人文视域下 SikuBERT 增强的史籍实体识别研究 [J]. 图书馆论坛，2022，42 (10)：61-72.

[16] 张逸勤，邓三鸿，胡昊天，等. 预训练模型视角下的跨语言典籍风格计算研究 [J]. 数据分析与知识发现，2023，7 (10)：50-62.

[17] 阎覃. GuwenBERT：古文预训练语言模型(古文 BERT) [EB/OL]. https://zhuanlan.zhihu.com/p/275970135，2024-08-19.

[18] 王东波，刘畅，朱子赫，等. SikuBERT 与 SikuRoBERTa：面向数字人文的《四库全书》预训练模型构建及应用研究 [J]. 图书馆论坛，2022，42 (6)：31-43.

[19] Wang D B，Liu C，Zhao Z X，et al. Gujibert and Gujigpt：Construction of Intelligent Information Processing Foundation Language Models for Ancient Texts [EB/OL]. https://arxiv.org/abs/2307.05354，2025-02-21.

[20] Reimers N，Beyer P，Gurevych I. Task-Oriented Intrinsic Evaluation of Semantic Textual Similarity [C] //Proceedings of COLING 2016，the 26th International Conference on Computational Linguistics：Technical Papers，2016：87-96.

(责任编辑：杨丰侨)