CCL25-Eval任务8总结报告：中文电子病历ICD诊断编码评测

梁镇鹏[1,2], 李传龙[1,2], 廉颖[3], 陈国强[3], 管红娇[1,2*], 鹿文鹏[1,2]
[1]齐鲁工业大学（山东省科学院），山东省计算中心（国家超级计算济南中心），
算力互联网与信息安全教育部重点实验室/ 济南，中国
[2]山东省算力互联网与服务计算重点实验室，
山东省基础科学研究中心（计算机科学）/ 济南，中国
[3]山东第一医科大学第一附属医院/ 济南，中国
{liangzhenpeng8, lianying2074}@163.com, chuanlongli@foxmail.com
15854178612@126.com, {hongjiao.guan, wenpeng.lu}@qlu.edu.cn

摘要

中文电子病历国际疾病分类（ICD）诊断编码评测依托第二十四届中国计算语言学大会（CCL）举办。该评测聚焦于自然语言处理技术在智能医疗领域的应用，旨在从真实脱敏的电子病历文本中自动分析关键临床表征，实现主诊断及其他诊断ICD编码的精准预测与分配，从而辅助临床医生与专业编码员提升编码工作的准确性和效率。本次评测在阿里云天池平台进行，获得了学术界与工业界的广泛关注和积极参与。数据显示，共有445支队伍报名参赛，其中A榜和B榜分别有85支和36支队伍成功提交了有效结果。最终，8支表现优异的队伍受邀撰写并分享了其技术报告，为推动该领域的技术进步与方法创新贡献了宝贵经验。本次评测的详细信息可参见相关发布页面[1]。

关键词： ICD编码 ； 评测任务 ； 大语言模型

Overview of CCL25-Eval Task 8: Chinese Electronic Medical Record ICD Diagnosis Coding Evaluation

Zhenpeng Liang[1,2], Chuanlong Li[1,2], Ying Lian[3], Guoqiang Chen[3], Hongjiao Guan[1,2*], Wenpeng Lu[1,2]
[1]Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences) Jinan, China
[2]Shandong Provincial Key Laboratory of Computing Power Internet and Service Computing, Shandong Fundamental Research Center for Computer Science Jinan, China
[3]First Affiliated Hospital of Shandong First Medical University Jinan, China
{liangzhenpeng8, lianying2074}@163.com chuanlongli@foxmail.com
15854178612@126.com, {hongjiao.guan, wenpeng.lu}@qlu.edu.cn

Abstract

The Chinese Electronic Medical Record International Classification of Diseases (ICD) Diagnostic Coding Evaluation was organized as part of the 24th China National Conference on Computational Linguistics (CCL). This evaluation focused on the application of Natural Language Processing (NLP) techniques in the field of intelligent healthcare. It aimed to automatically analyze key clinical manifestations from authentic, de-identified EMR texts, to achieve precise prediction and assignment of ICD codes for primary and other diagnoses, thereby assisting clinicians and professional coders in improving the accuracy and efficiency of their coding work. The evaluation was conducted on the Alibaba Cloud Tianchi platform and garnered significant attention and active participation from both academia and industry. Statistics showed that a total

[1]https://github.com/QLU-NLP/icdevaluator

of 445 teams registered for the competition, with 85 and 36 teams successfully submitting valid results on Leaderboard A and Leaderboard B, respectively. Ultimately, eight top-performing teams were invited to write and share their technical reports, contributing valuable insights and experiences towards advancing technological progress and methodological innovation in this field. The detailed information of this evaluation can be found on the related release page[1].

Keywords: ICD Code , Evaluation Task , Large Language Model

1 引言

近年来，随着人口老龄化加剧和健康意识的普遍提升，全球医疗体系正面临日益严峻的服务压力(Teng et al., 2022)。在此背景下，医疗信息化进程的持续推进，特别是电子健康记录（EHR）和医院信息系统（HIS）等数字化平台的广泛应用，使得在患者诊疗过程中积累了海量的医疗数据(Huang et al., 2018)。这些数据蕴藏着巨大的价值，但其多样性和复杂性也对有效管理和利用提出了挑战。

为了实现医疗数据的标准化管理、促进信息共享与深度分析，一个统一且国际通用的编码体系至关重要。为此，世界卫生组织（WHO）制定并持续维护着国际疾病分类（International Classification of Diseases, ICD）标准(Shull, 2019)。ICD旨在将全球范围内纷繁复杂的疾病、症状、体征、伤害以及导致这些情况的外部原因等，系统性地转化为一套具有明确层级结构、标准化的字母数字编码(Aarseth et al., 2017)。例如，当患者在入院过程中出现“急性心肌梗死”的诊断时，在ICD-10编码体系中，这一诊断会被编码为“I21.900”，其中字母“I”代表循环系统疾病这一大类，数字“21”代表疾病类别为缺血性心脏病，数字“900”则进一步细化为缺血性心脏病的具体类别。该标准历经了多次修订和发展，从早期的版本演进至目前全球广泛采用的第十版（ICD-10）以及最新的、功能更为强大的第十一版（ICD-11）(周强等, 2021)。许多国家和地区还会基于ICD进行本地化调整以适应其特定的医疗信息需求，例如中国目前主要采用基于ICD-10的《疾病分类与代码国家临床版2.0》进行疾病诊断编码，而在手术操作方面则广泛使用如《手术操作分类代码国家临床版》（通常基于ICD-9-CM进行编制和修订）等专门的编码体系(刘娟等, 2019)。尽管ICD版本和具体应用存在差异，其核心目标始终是为跨地域、跨机构的医疗数据交换、临床研究、流行病学监测以及卫生资源管理奠定基础(Zhang et al., 2022; Avati et al., 2018)。

然而，将海量的、通常以非结构化文本形式存在的临床文档（如电子病历中的主诉、现病史、诊疗计划等）准确高效地转换为ICD编码，传统上完全依赖于经过专业培训的编码员进行人工操作(Vu et al., 2020)。这一过程不仅劳动强度大、耗时费力、导致编码成本居高不下，而且极易受到编码员个人经验、对编码规则理解深度以及临床知识掌握程度等主观因素的影响，从而引入编码的不一致性和潜在错误(Williamson et al., 2024; Cao et al., 2020; Yoo and Kim, 2025)。此外，人工编码的滞后性也常常导致医疗数据无法及时应用于需要快速响应的临床分析与决策支持场景，进而限制了医疗大数据潜能的充分释放和价值挖掘。

随着人工智能（AI）技术的迅猛发展，特别是在自然语言处理（NLP）、机器学习（ML）、深度学习（DL）以及大语言模型（LLM）等领域的显著进步，为ICD自动编码这一难题提供了新的解决方向和实现可能(Vu et al., 2020; Yuan et al., 2022; Boukhers et al., 2024)。利用先进的AI算法，自动编码系统有望显著提升编码的效率、准确性和一致性，并有潜力辅助处理更复杂、更细致的编码需求(Baksi et al., 2025)。这不仅能够极大减轻临床编码人员的繁重工作负担，使其能更专注于处理疑难病例和质量控制，还将为临床决策支持系统、精准医疗、流行病学研究、医疗服务质量评估、公共卫生突发事件监测与响应等诸多关键领域提供更为及时、可靠、高质量的结构化数据支撑，从而有力推动整个智能医疗保健生态系统的创新与发展(Zhu et al., 2025)。

本文主要包含如下内容：第二节主要介绍了任务的定义以及评测使用的指标;第三节介绍本次评测使用的数据集，包括数据的示例、处理流程以及数据的统计信息；第四节介绍本次评测的参赛情况并对参赛队伍使用的方法进行总结；最后，第五节对本次评测进行总结。

第二十四届中国计算语言学大会论文集，第338页-第346页，济南，中国，2025年8月11日至14日。

2 任务描述

本评测的核心任务是开发一个框架，旨在从非结构化的电子病历文本中自动识别并分配ICD编码，以辅助医生或编码员更精准地完成ICD编码过程。具体而言，任务目标在于运用先进的自然语言处理方法，对电子病历中以自然语言形式详细记录的患者临床信息进行深度语义分析与理解。这些临床信息通常较为复杂，涵盖了患者的主诉、现病史（包括症状、体征、病程演变）、既往史、家族史、个人史、体格检查等多个方面。系统需要能够从这些冗长且充满医学术语、缩写和潜在歧义的文本中，精准地提取出与患者病情直接相关的关键症状、体征及其他临床表征信息，并为其分配一组ICD编码。

该任务通常被视为一个复杂的多标签文本分类问题(Huang et al., 2022; Duan et al., 2023; Luo et al., 2024)。然而，随着大型语言模型（LLMs）的崛起及其在理解和生成复杂文本方面的卓越能力，越来越多的研究倾向于将自动化ICD编码视为一个序列到序列（Sequence-to-Sequence）的生成问题(Li et al., 2024; Kwan, 2024; You et al., 2025; Baksi et al., 2025)。在这种新范式下，模型不再局限于从一个固定的、预定义的标签集合中进行选择，而是将输入的临床文本作为上下文条件，直接以生成式的方式产生对应的ICD编码序列或编码集合。这种转变不仅为处理具有庞大且动态变化的编码空间（尤其是当考虑到编码的组合与顺序时）提供了更灵活的解决思路，也使得模型能够更好地捕捉编码之间的复杂依赖关系，并有潜力生成在训练数据中未曾显式组合过但临床上合理的编码结果，从而更贴近人类编码专家的复杂推理过程。

2.1 任务定义

在形式化层面，该任务的输入（Input）是将患者在一次完整就诊或住院期间所产生的各类临床叙述性文档（如入院记录、病程记录、手术记录、出院小结等），经过适当预处理后，整合并拼接形成的单一长文本字符串（str类型字段）。其输出（Output）则是针对该输入临床文本，模型预测出的一组或一个有序序列的ICD编码。这组编码需要明确区分出对患者本次医疗事件的资源消耗和诊疗决策起决定性作用的“主诊断编码”（Primary Diagnosis Code），以及反映患者其他并存健康状况的“其他诊断编码”（Other Diagnosis Codes），同时这些预测的编码必须属于特定ICD编码体系（例如ICD-10国家临床修订版）中的有效编码。

2.2 评测指标

中文电子病历ICD诊断编码任务的性能评估采用一个综合准确率指标（Acc, Accuracy Score）来衡量。该指标综合考量了模型对主诊断编码预测的精确性以及对其他诊断编码集合预测的整体质量，其计算公式定义如下：

$$Acc = \frac{1}{N}\sum_{i=1}^{N}\{0.5 \cdot I(y_{\text{main}} = \hat{y}_{\text{main}}) + 0.5 \cdot F1(y_{\text{other}}, \hat{y}_{\text{other}})\}_i$$

公式中各符号含义如下：

- N：表示测试集中样本的总数量。
- $\{\cdot\}_i$：表示对测试集中第i个样本计算得到的单样本综合评估分数。
- $I(\cdot)$：为指示函数，当且仅当括号内的条件成立时，其值为1，否则为0。在此用于判断预测的主诊断编码$\hat{y}_{\text{main}}$是否与真实的主诊断编码y_{main}完全一致。
- y_{main}和 $\hat{y}_{\text{main}}$：分别指人工标注的真实主诊断编码和模型预测的主诊断编码。
- $F1(y_{\text{other}}, \hat{y}_{\text{other}})$：表示针对其他诊断编码预测的F1分数（F1-score）。用于评估真实的其他诊断编码集合y_{other}与模型预测的其他诊断编码集合$\hat{y}_{\text{other}}$之间的一致性。F1分数是精确率（Precision）和召回率（Recall）的调和平均值，其计算公式为：

$$F1(y, \hat{y}) = 2 \cdot \frac{\text{Precision}(y, \hat{y}) \cdot \text{Recall}(y, \hat{y})}{\text{Precision}(y, \hat{y}) + \text{Recall}(y, \hat{y})}$$

第二十四届中国计算语言学大会论文集，第338页-第346页，济南，中国，2025年8月11日至14日。

其中，精确率和召回率分别定义为：

$$\text{Precision}(y, \hat{y}) = \frac{NUM(y \cap \hat{y})}{NUM(\hat{y})}$$

$$\text{Recall}(y, \hat{y}) = \frac{NUM(y \cap \hat{y})}{NUM(y)}$$

- $\hat{y}_{\text{other}}$和y_{other}：分别表示模型预测的其他诊断编码的集合和人工标注的真实其他诊断编码的集合。
- $NUM(S)$：表示计算集合S 中元素数量的函数（即集合的基数）。在Precision和Recall的计算中，$NUM(y \cap \hat{y})$ 表示预测正确且真实存在的其他诊断编码的数量，$NUM(\hat{y})$ 表示模型预测的其他诊断编码总数，而$NUM(y)$ 表示真实的其他诊断编码总数。

3 数据集介绍

评测数据基于医院脱敏病历构建，共1485条数据。数据分为训练集、验证集和测试集，数据量分别为800条、200条和485条。

3.1 数据示例

本次评测所用数据包含丰富的临床文本信息，数据示例如图1所示，其中主要标注字段及其具体含义如下：

- 病案标识: 患者在医院就诊的唯一病案编号。
- 主诉: 患者在就诊时向医生描述的最主要、最直接的症状，通常用一句简短的话概括，是患者就医的主要原因。
- 现病史: 对患者当前疾病的发生、发展、演变过程的详细描述，内容涵盖起病情况、症状特征、治疗经过及效果等。
- 既往史: 患者以往的健康状况和疾病史，记录了既往的主要疾病、手术史、外伤史、过敏史等信息。
- 个人史: 涉及患者的生活习惯、职业暴露史、疫区接触史等相关信息。
- 婚姻史: 包括患者的婚姻状况、结婚年龄、配偶健康状况、有无子女以及性生活情况等。
- 家族史: 描述患者直系亲属中是否存在遗传性疾病或特定疾病的病史。
- 入院情况: 描述患者入院时的具体症状、体征以及一般身体状况。
- 入院诊断: 患者入院时，医生依据其现病史和初步检查结果得出的初步诊断。
- 诊疗经过: 详细记录了患者在住院期间所接受的各项检查、具体治疗方案以及病情的动态变化过程。
- 出院情况: 简要描述患者出院时的健康状况和疾病恢复情况。
- 出院医嘱: 医生针对患者出院后的日常生活、药物使用、定期复查等方面给出的指导性建议。
- 主诊断编码: 患者本次住院期间，对其健康状况和医疗资源消耗起决定性作用的主要诊断所对应的标准化ICD编码。
- 其他诊断编码: 患者本次住院期间，除主诊断外的其他并存疾病、并发症或重要相关情况所对应的标准化ICD编码集合。

```
{
    "病案标识": "ZY0*********97",
    "主诉": "胸闷、喘7天。",
    "现病史": "患者于7天前无明显诱因出现胸闷、喘，呈阵发性，活动及情绪激动后明显加重，不能从事日常活动，时有心慌、心前区疼痛不适，无发热、头痛、头痛，无左肩及后背部放射痛，无咳嗽、咳痰，无黑朦及意识障碍，病后至“***”住院治疗，完善相关辅助检查：化验检查（不详）",
    "既往史": "有“冠状动脉粥样硬化性心脏病”10余年，20**年于****行“冠状动脉移植术”（具体不详）",
    "个人史": "生长于原籍，否认疫区及地方病流行区长期居住史，生活规律。否认放射性物、工业粉尘、化学性物质接触史，否认冶游史，吸烟史40年，20支/日，已经戒烟1年余，否认饮酒史。",
    "婚姻史": "适龄结婚，育有1子，配偶及孩子身体健康。",
    "家族史": "父母已逝，具体不详。否认家族性遗传病及传染病史。",
    "入院情况": "患者老年男性，76岁，因“胸闷、喘7天”入院。有“冠状动脉粥样硬化性心脏病”10余年，20**年于***行“冠状动脉移植术”（具体不详），术后规律服用“酒石酸美托洛尔、螺内酯、单硝酸异山梨酯、阿司匹林”等药物治疗，平素活动耐量差。“高血压病”10余年，血压最高180/100mmHg，平素口服“沙坦氨氯地平”，血压控制在130-140/80-90mmHg。",
    "入院诊断": "1.急性失代偿性心力衰竭心功能II级（Killip分级）2.肺炎3.急性呼吸衰竭（I型）",
    "诊疗经过": "入院后完善相关辅助检查，凝血常规：凝血酶原时间：13.1秒，D-二聚体：1.74mg/L；血酮体测定：弱阳性；心梗三联：肌红蛋白：143.15ng/mL；NT-proBNP：8199pg/mL；血气分析：PH值：7.47，二氧化碳分压：26mmHg，氧分压：68mmHg，血二氧化碳总量：19.7mmol/L，细胞外液剩余碱：-4.8mmol/L：尿常规检查加沉渣粒细胞：+-，尿蛋白：+，酮体：2+，尿潜血：+2，红细胞：174/ul；血细胞分析：单核细胞计数：0.83×109/L，血红蛋白：109.0g/L：糖化血红蛋白：6.80%：2024-01-01床边胸片DR：双肺炎症、渗出性改变，请结合临床；双侧胸腔积液。心电图：1.快速型心房纤颤2.ST-T改变3.异常Q波。",
    "出院情况": "双侧瞳孔等大等圆，对光反射及调节反射存在，前胸部可见长约15cm纵行手术切口，左侧斜肋部带状疱疹素色沉着，胸廓对称无畸形，双肺呼吸音粗糙，未闻及湿性啰音。心前区无隆起及凹陷，心界无扩大，节律规则，各瓣膜听诊区未闻及病理性杂音。腹部膨隆，腹软，无压痛，无反跳痛。肝、脾肋下未触及，Murphys征阴性，肝、肾区无叩痛，肠鸣音无亢进，移动性浊音阴性。双下肢无水肿。",
    "出院医嘱": "1、低盐低脂饮食，注意休息，避免劳累，按时服药；2、出院后继续药物治疗：r阿托伐他汀钙片（齐鲁），每次量：20mg，睡前，每次一片单硝酸异山梨酯片（鲁南），每次量：20mg，每天一次，每次一片1托拉塞米片，每次量：5mg，每天一次，每次一片螺内酯片（长江），每次量：20mg，每天一次，每次一片1琥珀酸美托洛尔缓释片（合源），每次量：47.5mg，每天一次，每次一片门冬氨酸钾镁片，每次量：1片，每天两次，每次一片|普瑞巴林胶囊（赛维）。",
    "主诊断编码": "J81.x00x002",
    "其他诊断编码": "I50.907; I50.903; I25.103; I20.000; I49.900; I48.x01;E11.900"
}
```


图 1: 数据示例

3.2 数据处理

本次评测所使用的数据集在构建过程中主要经历了以下几个处理阶段:

数据采集 原始数据来源于山东省第一医科大学第一附属医院的真实临床病案。每条病案记录均包含了丰富的文本信息，并且其对应的ICD诊断编码均由具有丰富经验的专业编码员人工标注完成，确保了标签的初始质量。

数据筛选与范围界定 为使评测任务既能反映真实医疗场景的复杂性，又能保证任务的可行性和聚焦性，我们对采集到的数据进行了细致的筛选。具体而言，我们首先将数据范围限定在心血管内科（简称心内科）的病案。在此基础上，进一步选取了该科室临床实践中最为常见的五种主诊断编码所对应的病例作为本次评测的基础数据集。关于最终数据集内主诊断编码的详细分布情况，请参见第3.3节的统计分析。

诊断标签处理与规范化 考虑到“其他诊断编码”的多样性和稀疏性问题，我们对其进行了必要的处理。在初步筛选得到的1485条数据样本中，共涉及超过400种不同的“其他诊断编码”。统计发现，其中约70%的编码在整个数据集中仅出现一次，这种高度的稀疏性给模型学习带来了巨大挑战。因此，我们对“其他诊断编码”进行了过滤，保留了出现频率相对较高的常见编码作为模型预测的目标。尽管如此，为了确保评测的全面性和后续研究的参考价值，我们仍然提供了数据集中涉及的所有“其他诊断编码”的完整集合清单。

数据脱敏与隐私保护 为严格遵守医疗数据隐私保护规定，防止患者个人信息泄露与不当传播，我们对筛选后的数据集实施了双重脱敏处理流程。首先，我们收集整理了山东省内所有医

院的官方名称，并辅以人工补充常见的医院简称，构建了一个医院名实体库。基于此库，对病历文本中的相关医院信息进行了初步的自动化脱敏替换。随后，我们组织专业人员对脱敏结果进行了逐条人工核查与补充修正，以确保所有可能涉及患者身份、就诊地点等敏感信息的字段均得到妥善处理，最终形成可供评测使用的脱敏数据集。

3.3 数据统计

为了全面了解本评测所用数据集的特性与规模，本节将从数据集划分、文本长度以及诊断编码的分布等多个维度进行详细的统计与分析。具体的统计数据如表??所示。

统计指标	训练集	验证集	测试集
样本数量(条)	800	200	485
样本占比(%)	53.87	13.47	32.66
平均文本长度(字符)	2427	2434	2426
涉及的诊断编码数	53	51	52
平均其他诊断编码数	4.54	4.62	4.50

表 1: 数据集关键统计指标一览

除了上述量化指标，为了更直观地展示数据集中五种主要诊断编码的分布情况及其在整个数据集（共1485条记录）中所占的比例，我们绘制了如图2所示的扇形图。从图中可以看出，这五种主诊断编码的样本数量分布相对均衡，均占总样本量的20%左右。

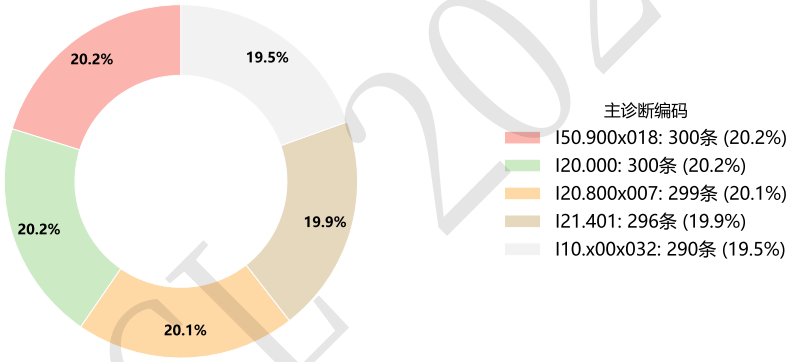

图 2: 各主诊断编码占比情况

4 评测情况

在评测期间，据阿里云天池平台数据显示，共有445支队伍报名参赛，其中A榜和B榜分别有85支和36支队伍成功提交了结果。最终，8支表现优异的队伍受邀撰写并分享了其技术报告。

4.1 评测结果

表??展示了本次评测的成绩排名，排名主要依据各队伍的“Acc得分”高低给出。从表中可以看出，来自北京理工大学的队伍以0.8672分的优异成绩名列榜首，紧随其后的是来自大连理工大学的队伍和来自中国电信重庆分公司的队伍，分别取得了0.8342分和0.8233分的突出成绩。此外，榜单还吸引了包括清华大学、中国石油大学（华东）、澳门大学等知名高校的队伍以及个人参赛者的积极参与，整体展现出了广泛的参与度，各队伍之间的得分也反映出了一定的竞争。

4.2 方法概述

本次评测有8支队伍受邀撰写并分享了其技术报告，其中有5份来自排名前7参赛队伍。本节内容对参赛队伍所采用的技术方案作出概述。

参赛队伍	组织	Acc得分
医路智嗨队	北京理工大学	0.8672
DUTIR-Bio	大连理工大学	0.8342
星辰之力	中国电信重庆分公司	0.8233
天然柠檬酸	个人	0.8159
看不懂题目	清华大学	0.8130
油碟小菜	中国石油大学	0.8089
一二三四1234	澳门大学	0.8082

表 2: 成绩排名

排名第一的来自北京理工大学的队伍针对ICD编码任务提出了基于疾病分组的分层处理框架。其核心技术思路是通过疾病临床相关性分组来重构标签空间，将复杂的ICD编码空间重新组织，根据疾病的临床表现和病理机制将相似疾病归类到同一组别中，确保组内疾病互斥而跨组疾病可共存。在数据处理层面，采用输入压缩策略从完整医疗记录中提取关键诊断信息，过滤冗余内容以提升学习效率。

针对主诊断和次诊断的不同特点，该方案实施了差异化微调策略。主诊断任务采用生成式微调方法，构建基于指令的训练框架增强模型对疾病类别的理解能力；其他诊断任务则选择判别式微调策略，通过专门分类器处理多疾病识别问题。为解决临床数据稀缺性，该队伍开发了知识驱动的数据增强方法，利用大语言模型结合医学知识库生成高质量病历记录扩充训练集。整个框架通过任务分解实现主次诊断的协同学习，在保持子任务独立性的同时促进特征共享，有效降低了学习复杂度并提升了模型专业化程度。

排名第二的来自大连理工大学的队伍也将编码任务进行分解，对主要诊断编码和其他诊断编码设计差异化的处理策略。在主要诊断编码方面，引入了信心引导的语义检索方法，通过构建信心评分机制来评估模型预测的可靠性，并结合语义检索技术来增强主要诊断的准确性。对于其他诊断编码任务，采用了基于命名实体识别的增强策略。通过NER技术从电子病历中提取与疾病相关的关键实体，并将这些实体信息与大型语言模型相结合，形成更为精准的诊断理解能力。在模型集成层面，该队伍实施了多模型投票策略，通过集成学习的方式结合多个模型的预测结果，进一步提升了整体系统的鲁棒性和准确率。

排名第三的来自中国电信重庆分公司的队伍首先系统比较了生成式与判别式微调在中文电子病历ICD诊断编码任务中的性能边界,提出的K-Fold投票集成机制显著提升了诊断多标签预测的稳定性和完整性;同时设计医疗特异性的强化学习奖励函数体系，通过动态难度校准、Token级梯度优化和超长奖励塑造（Overlong Reward Shaping）策略，提升群体相对策略优化算法的训练效率和性能。此外还构建了一个端到端的临床决策优化框架，为基于结果奖励模型（OutcomeReward Model，ORM）的微调提供更好的优化路径，提升了诊断编码推理阶段的稳定性和可靠性，设计了一种温度调节集成共识预测方法(Temperature-Calibrated Ensembleconsensus prediction,TCECP)，最终使LLM在中文电子病历ICD诊断编码任务的性能上限突破了SFT微调方法。

有部分参赛队伍充分利用提示词工程来完成评测任务。排名第四的参赛选手首先将ICD编码转换为对应的标准名称,并通过实验验证了个人史、婚姻史、家族史的有关内容对提升编码诊断结果的有效性;同时在Qwen3-4b模型测试了多种微调方式在该任务上的效果,并最终选择了全参微调。排名第七的来自澳门大学的参赛队伍同样将主要诊断编码跟其他诊断编码两个任务拆分，并在Qwen2.5-VL-7B-Instruct模型上选用了lora微调的方式。排名第十二的来自云南大学的队伍将复杂且不平衡的ICD编码候选集拆分成若干个更小、更易于管理的子集，然后针对每个子集分别对大型语言模型进行微调，同时结合结构化提示工程和参数高效微调技术。排名第十七的参赛选手在数据处理环节实用术语标准化、同义词替换和病程伸缩等方法，利用Qwen2.5-7B模型对数据进行数据增强；采用基于Qwen2-7B微调的Sunsimiao:7B医疗大模型作为基础，应用RS-LoRA变体解决医学术语长尾分布问题。

同样来自北京理工大学的队伍尝试了多种方法，包括CNN、BERT微调、CNN+Transformer、微调大模型、Prompt工程，最终选择CNN+Transformer的方

案。在数据处理阶段，为了对数据进行有效结构化处理，该队伍抽取高频关键词构建词典，设计词袋模型提取文本特征，同时将不同类型的诊断映射为数字或张量；在模型的构建部分，采用差异化设计主诊断任务使用CNN架构，通过多层卷积和池化捕捉局部文本特征；其他诊断任务则创新性地结合CNN与Transformer，前者提取局部特征，后者通过自注意力机制捕捉全局语义关系，两种特征经融合后输出多标签预测结果。

5 总结

本次中文电子病历ICD诊断编码评测任务由齐鲁工业大学（山东省科学院）和山东第一医科大学第一附属医院联合组织，并作为第24届中国计算语言学大会（CCL2025）的评测任务之一。该评测旨在利用自然语言处理技术提升临床ICD编码的自动化水平和准确性。本次评测在阿里云天池平台进行，吸引了445支队伍的积极参与。

各参赛队在任务中展现出了明显的技术路线分化，从整体技术选择来看，绝大部分队伍采用了大语言模型微调路线，主要基于Qwen系列等预训练模型，通过LoRA微调、全参数微调等策略来适应具体的诊断编码任务。在具体的技术实现上，多数队伍都重视提示工程的设计，针对中文医疗文本的特点开发了专门的提示模板，并且在任务建模上采用了从分类到生成的多样化方法。此外，部分队伍在传统微调基础上引入了更复杂的技术组合，有队伍采用了基于规则奖励的强化学习方法来进一步优化模型性能，还有队伍通过置信度引导的语义检索来增强模型的推理能力。在数据处理和特征工程方面，各队伍都意识到医学术语标准化的重要性，但在具体实现上有不同的侧重点。有的队伍专注于候选集合的分割策略，通过将复杂的编码预测任务分解为更小的子问题来提升准确性；有的队伍则重点关注结构化多类分类方法，将疾病按照临床意义进行分组处理；还有队伍采用了集成学习的策略，通过多模型融合和温度校准来优化最终的预测结果。

期望本次评测的成果能持续推动中文医疗信息处理技术进步，并为智能医疗系统的构建提供支持。后续工作将考虑扩展数据集覆盖面及编码粒度，以应对更复杂的临床编码需求。

致谢

本工作受国家自然科学基金资助项目（No.62376130）、济南市“新高校20条”资助项目（No.202333008）、齐鲁工业大学（山东省科学院）科教产融合试点工程重大创新类项目（No.2024ZDZX08）、山东省科技型中小企业创新能力提升工程项目（No. 2024TSGC0094）资助。

参考文献

Mengxing Huang, Huirui Han, Hao Wang, Lefei Li, Yu Zhang, and Uzair Aslam Bhatti. 2018. A clinical decision support framework for heterogeneous data sources. *IEEE Journal of Biomedical and Health Informatics*, 22(6):1824–1833.

Fei Teng, Yiming Liu, Tianrui Li, Yi Zhang, Shuangqing Li, and Yue Zhao. 2022. A review on deep neural networks for ICD coding. *IEEE Transactions on Knowledge and Data Engineering*, 35(5):4357–4375.

Jessica Germaine Shull. 2019. Digital health and the state of interoperable electronic health records. *JMIR Medical Informatics*, 7(4):e12712.

Shurui Zhang, Bozheng Zhang, Fuxin Zhang, Bo Sang, and Wanchun Yang. 2022. Automatic ICD coding exploiting discourse structure and reconciled code embeddings. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2883–2891.

Espen Aarseth, Anthony M. Bean, Huub Boonen, Michelle Colder Carras, Mark Coulson, Dimitri Das, Jory Deleuze, Elza Dunkels, Johan Edman, Christopher J. Ferguson, et al. 2017. Scholars' open debate paper on the World Health Organization ICD-11 Gaming Disorder proposal. *Journal of Behavioral Addictions*, 6(3):267–270.

Anand Avati, Kenneth Jung, Stephanie Harman, Lance Downing, Andrew Ng, and Nigam H. Shah. 2018. Improving palliative care with deep learning. *BMC Medical Informatics and Decision Making*, 18:55–64.

Ashton Williamson, David de Hilster, Amnon Meyers, Nina Hubig, and Amy Apon. 2024. Low Resource ICD Coding of Hospital Discharge Summaries. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 548–558.

Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, and Weifeng Chong. 2020. HyperCore: Hyperbolic and co-graph representation for automatic ICD coding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3105–3114.

Youngju Yoo and Sewon Kim. 2025. How to leverage large language models for automatic ICD coding. *Computers in Biology and Medicine*, 189:109971.

Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. 2020. A Label Attention Model for ICD Coding from Clinical Text. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3335–3341.

Zheng Yuan, Chuanqi Tan, and Songfang Huang. 2022. Code Synonyms Do Matter: Multiple Synonyms Matching Network for Automatic ICD Coding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 808–814.

Zeyd Boukhers, AmeerAli Khan, Qusai Ramadan, and Cong Yang. 2024. Large Language Model in Medical Informatics: Direct Classification and Enhanced Text Representations for Automatic ICD Coding. In *Proceedings of the 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3066–3069. IEEE.

Krishanu Das Baksi, Elijah Soba, John J. Higgins, Ravi Saini, Jaden Wood, Jane Cook, Jack I. Scott, Nirmala Pudota, Tim Weninger, Edward Bowen, and Sanmitra Bhattacharya. 2025. MedCodER: A Generative AI Assistant for Medical Coding. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 449–459, Albuquerque, New Mexico. Association for Computational Linguistics.

Yakun Zhu, Shaohang Wei, Xu Wang, Kui Xue, Shaoting Zhang, and Xiaofan Zhang. 2025. MeNTi: Bridging Medical Calculator and LLM Agent with Nested Tool Calling. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5097–5116, Albuquerque, New Mexico. Association for Computational Linguistics.

ChaoWei Huang, ShangChi Tsai, and YunNung Chen. 2022. PLM-ICD: Automatic ICD Coding with Pretrained Language Models. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 10–20.

Junwen Duan, Han Jiang, and Ying Yu. 2023. MHLAT: Multi-Hop Label-Wise Attention Model for Automatic ICD Coding. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023)*, pages 1–5.

Junyu Luo, Xiaochen Wang, Jiaqi Wang, Aofei Chang, Yaqing Wang, and Fenglong Ma. 2024. CoRelation: Boosting Automatic ICD Coding through Contextualized Code Relation Learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3997–4007.

Rumeng Li, Xun Wang, and Hong Yu. 2024. Exploring LLM Multi-Agents for ICD Coding. *arXiv preprint arXiv:2406.15363*.

Keith Kwan. 2024. Large language models are good medical coders, if provided with tools. *arXiv preprint arXiv:2407.12849*.

Xinxin You, Xien Liu, Xue Yang, Ziyi Wang, and Ji Wu. 2025. MKE-Coder: Multi-Axial Knowledge with Evidence Verification in ICD Coding for Chinese EMRs. *arXiv preprint arXiv:2502.14916*.

刘娟, 胡牧, 简伟研, 卢铭. 2019. 医保疾病诊断、手术操作分类与代码标准的研究与制定. 中国医疗保险, (10):39–41. DOI: 10.19546/j.issn.1674-3830.2019.10.008.

周强, 李明, 董全伟, 程磐基, 包来发, 娄月丽, 蒋小贝, 鲍颖洁, 杨丽娜, 朱邦贤, 严世芸. 2021. 《国际疾病分类第十一次修订本(ICD-11)》传统医学章节与新版中医国家标准的比较研究. 上海中医药杂志, 55(05):1–6+23. DOI: 10.16305/j.1007-1334.2021.2101030.