错误驱动增强与跨度一致性学习的《藏药志》命名实体识别

摘要：目的：针对《藏药志》中目标药材完整漏检、与来源名称及相关药名发生语义混淆，以及性味、药效和病症实体边界识别不准的问题，提高命名实体识别的准确性与稳定性。**方法：**以 MacBERT-large 和 GlobalPointer 为基础，提出融合错误驱动增强与跨度一致性学习的方法。通过训练集内五折折外预测定位药材漏检和语义混淆误报，在原文证据约束下利用大语言模型构造困难样本，经程序校验和人工审核后仅监督其中的药材类别；构建师生框架，在真实样本的性味、药效和病症类别中，结合金标准跨度、教师正跨度、边界邻域和高不确定跨度建立候选集合，以教师软目标约束学生预测，并通过指数移动平均更新教师。**结果：**在包含 408 条药材条目、1 505 个文本片段和 5 991 个实体提及的数据集上进行严格实体级评价。完整模型的整体 F_1 值为 90.53%，较基础模型提高 2.36 个百分点；药材 F_1 值为 86.52%，提高 5.64 个百分点。消融实验表明两模块具有互补作用；3 个随机种子下，整体与药材 F_1 均值分别为 $90.60 \pm 0.23\%$ 和 $86.04 \pm 0.42\%$。在《四部医典》外部测试集上，整体微平均 F_1 由 82.29% 提高至 83.29%。**结论：**该方法在保持非药材类别识别能力的同时改善了目标药材漏检与语义混淆，且推断时无需调用大语言模型，可为藏医药文献结构化处理提供支持。

关键词：藏医药文献；命名实体识别；错误驱动增强；跨度一致性学习

Named Entity Recognition in Zangyaozhi Integrating Error-Driven Augmentation and Span Consistency Learning

Abstract: [Objective] Zangyaozhi records medicinal materials, their sources, properties and tastes, therapeutic effects, and indications in natural language. Named entity recognition is challenging because a target medicinal material may recur in a long entry, while names of source plants or animals, related medicinal materials, and comparison objects often have similar forms and contexts. Models may therefore miss a target completely or identify a non-target name as a medicinal material. Properties and tastes, therapeutic effects, and diseases and symptoms also occur in continuous phrases, increasing boundary shifts and category confusion. This study addresses both target-specific errors and difficult non-medicinal-material spans to improve recognition accuracy and stability. **[Methods]** A corpus was built from scanned pages of Zangyaozhi through optical character recognition, manual collation, entry normalization, and entity annotation. It contains 408 medicinal-material entries, 1,505 text segments, and 5,991 mentions in four categories: medicinal materials, properties and tastes, therapeutic effects, and diseases and symptoms. Annotation followed unified guidelines and was cross-checked for entity boundaries and categories. Data were split by entry to prevent segments from the same entry from crossing subsets, and strict entity-level exact matching was used for evaluation. The method adopts MacBERT-large as the encoder and GlobalPointer as the span decoder and contains two complementary modules. The first performs error-driven hard-sample augmentation. Five-fold out-of-fold prediction within the real training set locates complete misses of medicinal materials and false positives caused by semantic confusion. A large language model constructs candidate texts under source-evidence constraints. Candidates are checked for format, target terms, duplication, label boundaries, and prompt residue and are then reviewed manually. All error instances and evidence units originate from the training split, while the test set is excluded from sample construction and parameter selection. Retained samples provide hard-label supervision only for medicinal materials, so unverified annotations of other categories are not treated as negative labels. This design prevents unannotated properties, therapeutic effects, or diseases in generated texts from being optimized as negative examples. The second module performs selective span-consistency learning on real samples. Student and teacher models share the same architecture. For properties and tastes, therapeutic effects, and diseases and symptoms, the candidate set combines

gold spans, teacher-predicted positive spans, one-position neighborhoods of gold boundaries, and additional high-uncertainty spans. Teacher probabilities serve as stop-gradient soft targets. The teacher first follows the student synchronously and is later updated by an exponential moving average. The complete objective combines four-category supervision on real samples, medicinal-material supervision on augmented samples, and consistency loss on selected non-medicinal-material spans. At inference, only the teacher checkpoint selected on the development set is used; neither the large language model nor candidate construction is required. Candidate construction and consistency learning are used only during model training. **[Results]** On the test set, the MacBERT-large–GlobalPointer baseline obtained an overall precision of 84.74%, recall of 91.90%, and F_1 score of 88.17%. The complete method achieved 88.09% precision, 93.10% recall, and a 90.53% F_1 score, improving overall F_1 by 2.36 percentage points. Medicinal-material F_1 increased from 80.88% to 86.52%, a gain of 5.64 points. The F_1 scores for properties and tastes, therapeutic effects, and diseases and symptoms were 93.75%, 93.38%, and 90.29%, respectively. Ablation results showed that error-driven augmentation raised medicinal-material F_1 to 84.91%, whereas span consistency raised overall F_1 to 90.22%; their combination produced the best overall result. Under the same quota, difficult-span selection outperformed random selection. Across three random seeds, the complete model obtained overall and medicinal-material F_1 scores of $90.60 \pm 0.23\%$ and $86.04 \pm 0.42\%$, indicating stable improvements. On a 731-entity cross-document test set from the Four Medical Tantras, overall micro-F_1 increased from 82.29% to 83.29%. **[Conclusion]** Error-driven augmentation and selective span consistency provide complementary supervision: the former targets medicinal-material misses and semantic-confusion errors, whereas the latter constrains informative non-medicinal-material spans. Their combination balances entity categories without requiring a large language model during inference and supports structured information extraction from Tibetan medicine literature when annotated domain data are limited.

Keywords: Tibetan medicine literature; named entity recognition; error-driven augmentation; span consistency learning

藏医药文献以自然语言记录药材来源、性味、药效及主治病症等知识。命名实体识别（named entity recognition，NER）用于确定实体边界和类别，是关系抽取、知识组织与检索等任务的基础[1]，这对藏医药文献的结构化处理具有重要意义。

《藏药志》以药材条目组织内容，同一条目可能同时出现目标药材、原植物或原动物名称及相关药名。它们词形和语境相近但语义角色不同，易造成目标药材漏检与非目标名称误报；性味、药效和病症又常连续出现，易发生边界偏移或类别混淆。因此，模型既要区分目标药材与干扰项，也要准确判断其他类别的实体跨度。

现有序列标注和跨度方法仍受领域样本覆盖限制。大语言模型虽可离线构造文本，但无依据生成可能引入事实与标签错误；师生一致性能够补充监督，却易被跨度矩阵中的大量简单负例稀释。

为此，本文以 MacBERT-large 和 GlobalPointer 为基础，利用训练集内五折折外预测定位药材漏检与语义混淆误报，在原文证据约束下构造仅监督药材（Drug）类别的困难样本；同时筛选真实样本中的非药材（non-Drug）候选跨度，以教师软目标进行约束。推断阶段仅使用教师模型。

本文的主要贡献如下：

（1）提出错误驱动的药材困难样本增强方法。利用训练集内折外预测确定完整漏检和语义混淆误报，在原文证据约束下构造困难增强样本，并结合程序校验、人工审核和药材选择性监督，增强模型对目标药材及其干扰项的判别能力。

（2）提出面向非药材类别的选择性跨度一致性学习方法。在真实样本中，结合金标准跨度、教师预测正跨度、金标准边界邻域及额外高不确定跨度构建候选集合，利用师生软目标约束辅助跨度判别，并与药材定向增强共同构成联合训练框架。

在自建《藏药志》数据集上开展对比实验、消融实验及 3 个随机种子重复实验，并补充典型错误案例分析与《四部医典》跨文献评价。严格实体级匹配结果表明，完整模型的整体 F_1 值达到 90.53%，较基础模型提升 2.36 个百分点；药材 F_1 值达到 86.52%，提升 5.64 个百分点。

1 相关研究

1.1 命名实体识别研究

命名实体识别主要采用序列标注和跨度建模。BiLSTM-CRF 结合双向长短时记忆网络与条件随机场建模上下文及标签转移[2]，BERT[3]和 MacBERT[4]等预训练模型进一步增强了语境表示。针对中文缺少显式词边界的问题，FLAT[5]将字符与词汇组织为扁平格结构，LEBERT[6]通过适配机制融入词典信息。这些方法能够利用词汇线索，但仍受词典覆盖和领域适配影响，对低频且语义相近的药材名称难以充分消歧。

跨度方法直接预测边界与类别。机器阅读理解（machine reading comprehension，MRC）方法将 NER 转化为片段抽取[7]，GlobalPointer[8]构建类别跨度矩阵，W2NER[9]利用词对关系表示实体。本文采用 MacBERT-large 与 GlobalPointer，并从训练数据和跨度监督两方面改进。

1.2 大语言模型辅助的数据增强研究

GPT-NER[10]将 NER 转化为带标记的文本生成任务，UniversalNER[11]以大模型生成数据蒸馏开放域识别模型；范西朋等[12]通过边界预测与动态裁剪缓解样本不平衡。然而，大模型生成的语言流畅性不能保证专业事实及标签正确，若直接加入训练可能放大噪声。本文不进行无目标扩增，而是依据折外错误确定增强对象，以原文证据约束内容，经程序检查与人工审核后仅监督药材类别。

1.3 一致性学习与实体跨度约束研究

一致性学习要求模型在扰动或不同参数状态下给出相近预测。Mean Teacher[13]以学生参数的指数移动平均更新教师，可在不改变推断结构的情况下提供稳定软目标。吕书宁等[14]利用分词、词性与实体标签间的一致性，冯嘉琦等[15]以教师筛选未标注样本并生成伪标签。边界平滑[16]与 DiFiNet[17]分别从边界目标和跨度语义改善定位，说明边界邻域及高不确定跨度具有额外判别价值。

鉴于硬伪标签可能放大错误且跨度矩阵含大量简单负例，本文仅在真实样本的性味、药效和病症类别上，从金标准跨度、教师正跨度、边界邻域和高不确定跨度构建候选集合，并以教师软目标进行一致性约束。

2 方法

本文方法以 MacBERT-large 和 GlobalPointer 为基础，整体框架见图1。M1 根据训练集折外预测发现药材完整漏检与语义混淆误报，在原文证据约束下构造困难正、负样本，并实施 Drug 选择性监督，以增强目标药材和干扰名称之间的判别。M2 在真实样本的非 Drug 类别中综合金标准、教师预测、边界邻域与不确定性筛选候选跨度，以教师软目标约束学生。训练时联合计算真实样本四类监督、增强样本 Drug 监督和跨度一致性损失，教师由同步更新逐步转为指数移动平均；推断时仅使用开发集选出的教师模型，不调用大语言模型。

2.1 任务定义与总体框架

本文识别药材、性味、药效和病症四类实体，类别集合定义为

$$\mathcal{Y} = \{\text{Drug}, \text{Taste}, \text{Effect}, \text{Disease}\}.$$

其中，Drug、Taste、Effect 和 Disease 分别表示药材、性味、药效和病症。Drug 包括标题中的目标药材及其正文重复指称，不包括来源名称、相关药名和比较对象。本文采用平面实体识别；实体三元组 (s, e, c) 的起点、终点和类别均与金标准一致时才判定正确。

以“嘎贝折吾”条目为例，目标药材“嘎贝折吾”在标题和正文中的 3 处出现均标注为 Drug，来源植物“葫芦”和“木桔”不标注。前者可能被完整漏检，后者可能被误报为 Drug，对应 M1 的两类目标错误。短语“能愈疮，治肺病，止热痢，尤止大、小肠热泻”中，“愈疮”为 Effect，“肺病”“热痢”和“大、小肠热泻”为 Disease；排除“治肺病”等扩展边界属于 M2 关注的候选跨度判别。

给定第 b 个《藏药志》文本序列

$$X_b = \{x_{b,1}, x_{b,2}, \ldots, x_{b,n_b}\},$$

其中，$x_{b,i}$ 为第 i 个字符，n_b 为片段长度。模型不额外拼接标题；仅含标题的短片段与正文合并后，标题作为序列的一部分参与编码和评价。记字符位置集合为 $\mathcal{I}_b$，合法跨度集合为

$$\Omega_b = \{(i, j) \mid i, j \in \mathcal{I}_b,\ i \le j\}.$$

特殊标记和填充位置不参与预测，候选实体记为 (i, j, c)。

训练集由真实样本 $\mathcal{D}_r$ 和增强样本 $\mathcal{D}_a$ 组成，分别提供四类完整监督和仅 Drug 监督。M2 以同构师生模型对真实样本的候选非 Drug 跨度计算一致性损失。

2.2 基础实体识别模型

本文采用 MacBERT-large 获取输入序列的上下文表示：

$$H_b = \text{MacBERT}(X_b) = \{h_{b,1}, h_{b,2}, \ldots, h_{b,n_b}\},$$

其中，n_b 为输入序列长度，$h_{b,i}$ 为位置 i 的上下文向量。

GlobalPointer 根据上下文表示，为每个实体类别构建起止位置之间的跨度得分矩阵。其打分过程记为

$$z_{b,c,i,j} = \text{GP}_c(h_{b,i}, h_{b,j}; \text{RoPE}), \tag{1}$$

其中，RoPE 表示旋转位置编码[18]，用于在起止位置交互中引入相对位置信息；$z_{b,c,i,j}$ 为跨度 (i, j) 属于类别 c 的未归一化得分。模型仅在 Ω_b 内计算有效跨度预测。

记类别 c 的金标准跨度集合为 $\mathcal{G}_{b,c}$，基础模型的跨度监督损失为

$$\ell_{\text{GP}}(b, c) = \ell_{\text{GP}}\big(z_{b,c}, \mathcal{G}_{b,c}; \Omega_b\big). \tag{2}$$

该损失依据金标准标签区分合法区域内的实体跨度与非实体跨度。真实样本中的四个类别均参与监督；增强样本的类别监督范围由 M1 中的选择性监督策略确定。

2.3 错误驱动的困难样本增强

2.3.1 基于折外预测的错误发现

本文按条目编号对真实训练集进行五折折外预测：每次以四折训练基础模型并预测剩余一折，同一条目的全部片段始终属于同一折。由此，每个样本的错误均来自未使用该条目训练的模型，较训练集内直接预测更能反映泛化难点。各折使用正式开发集早停；错误统计、原文证据和增强构造对象仅来自训练集，测试集不参与错误归纳、样本生成或参数选择。

依据折外预测与人工标注之间的差异，本文选取两类错误作为增强来源。

（1）完整漏检：金标准 Drug 未被预测实体覆盖，作为保留目标药材及相关上下文的困难正例来源。

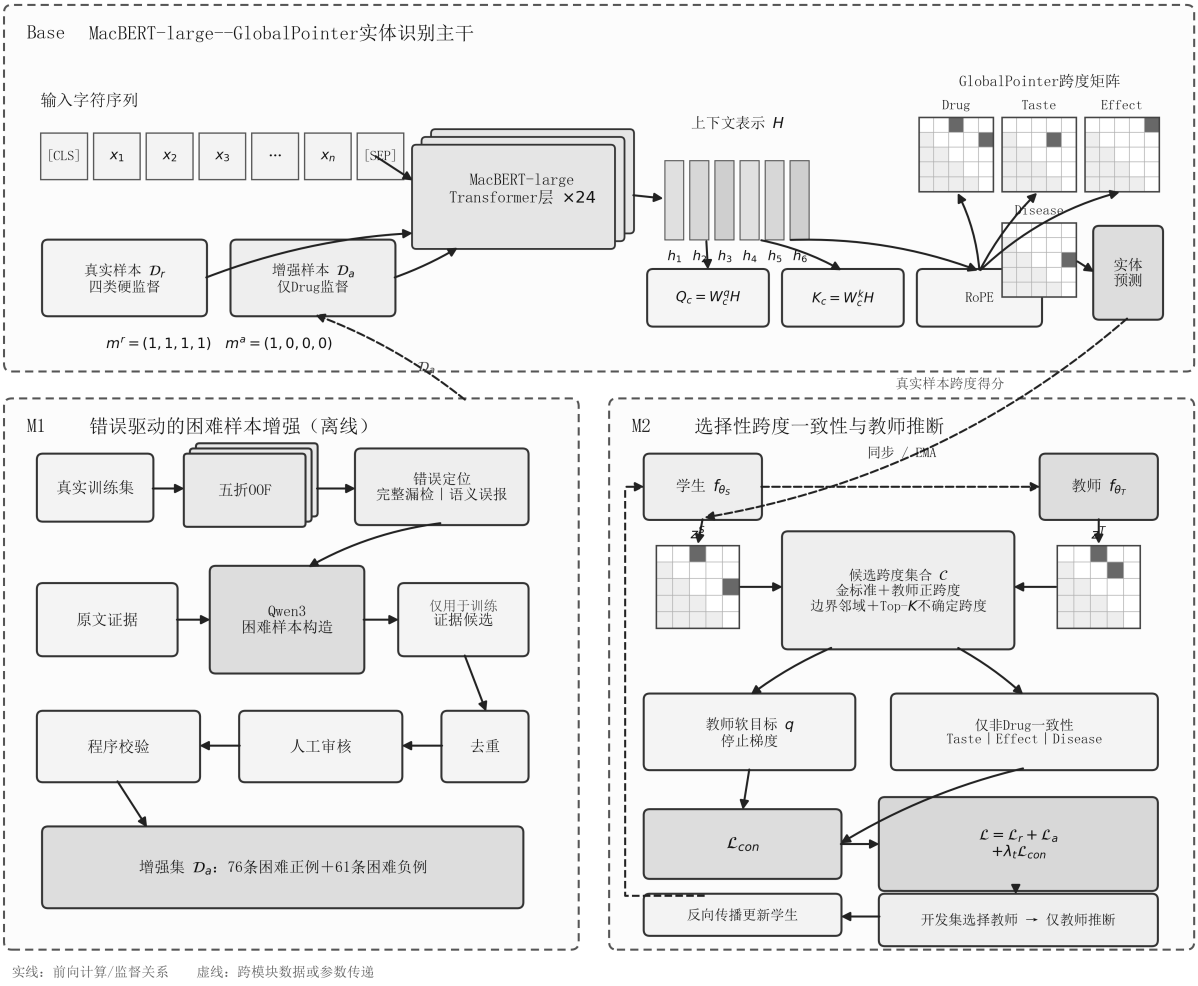


图 1 融合错误驱动增强与跨度一致性学习的命名实体识别框架

Fig. 1 Framework of named entity recognition integrating error-driven augmentation and span consistency learning

注：OOF 表示折外预测（out-of-fold prediction），EMA 表示指数移动平均（exponential moving average），RoPE 表示旋转位置编码（rotary position embedding）；Drug、Taste、Effect 和 Disease 分别表示药材、性味、药效和病症。

（2）语义混淆误报：预测 Drug 与任何金标准实体均不重叠，且经审核不是当前目标药材，作为涉及来源名称、相关药名或比较对象的困难负例来源。

Drug 边界错误和实体类型混淆不作为 M1 的增强目标，避免将不同性质的识别问题混入同一增强过程。

2.3.2 原文证据约束的样本构造

本文采用 Qwen3-30B-A3B-Instruct-2507-AWQ-4bit 构造困难样本。该混合专家（mixture of experts，MoE）模型总参数约 30B、每个 Token 激活约 3B[19]；使用 Transformers 4.57.1 和 vLLM 0.10.2 加载四位量化检查点，线性层采用组大小 32 的对称 4 位整数量化（INT4），激活与计算保持脑浮点数格式（BF16）。

样本构造包括自由生成与原文证据约束。首次生成的温度、top-p、top-k、重复惩罚和最大长度分别为 0.72、0.88、40、1.06 和 320；失败重试时分别为 0.55、0.85、30、1.04 和 480。

自由生成包含 55 个困难正例目标和 49 个困难负例目标，每个目标生成 5 条候选，共 520 条。正例提示要求目标药材自然出现并正确标注为 Drug，同时保留与历史漏检相关的上下文；负例提示要求目标药材与指定干扰词面共现，仅真实 Drug 被标注，使模型学习区分来源名称、相关药名和比较对象。两类文本均须符合《藏药志》的表述方式，不得引入原文之外的来源、性味、药效、病症或因果关系。

证据约束阶段先从真实训练原文选择相关证据单元，再对候选窗口排序。两步的温度和 top-p 分别为 0.25/0.85 与 0.10/0.75，最大长度均为 120；每个可用目标最多构造 4 个证据组合，总上限 392，事实审核采用温度 0 的确定性解码。该阶段的作用是把

生成内容限制在可核验的原文范围内，而非要求模型补充外部医学知识。

候选先经程序检查 JSON 格式、文本长度、指定词项及其次数、重复内容、标签跨度和提示残留，未通过者直接删除。其余样本由 1 名研究人员依据统一标注规范审核，重点核对原文依据、目标药材与干扰词面的语义角色以及新增医学事实；为控制重复，每个困难负例词面最多保留两条。

各阶段候选相互衔接，故不累加阶段数量或计算统一拒绝率。最终保留 137 条增强文本：76 条困难正例覆盖 18 种目标药材、含 206 个 Drug 标注；61 条困难负例覆盖 33 个干扰词面、含 96 个真实 Drug 标注，共增加 302 个 Drug 标注。

2.3.3 药材选择性监督

M1 围绕 Drug 构造和审核，故增强文本仅参与 Drug 监督，不将其他未标注类别视为负例。按 Drug、Taste、Effect 和 Disease 顺序，监督掩码为

$$\begin{aligned} m^r &= (1,1,1,1), \\ m^a &= (1,0,0,0). \end{aligned} \tag{3}$$

记真实样本的四类实体监督损失为 $\mathcal{L}_r$，增强样本的 Drug 监督损失为 $\mathcal{L}_a$，则 M1 的训练目标为

$$\mathcal{L}_{\text{M1}} = \mathcal{L}_r + \mathcal{L}_a. \tag{4}$$

其中，$\mathcal{L}_r$ 为真实样本四类损失，$\mathcal{L}_a$ 为增强样本 Drug 损失。增强样本仍更新共享编码器，但不产生其他类别的监督损失。

2.4 选择性跨度一致性学习

2.4.1 师生预测与困难度度量

M2 在真实样本的非 Drug 类别上引入一致性约束，作用类别集合为

$$\mathcal{Y}_n = \{\text{Taste}, \text{Effect}, \text{Disease}\}.$$

学生与教师结构相同，参数为 θ_S 和 θ_T。学生开启 Dropout，教师关闭 Dropout 且不参与梯度更新；二者在跨度 (i,j) 上的得分记为 $z^S_{b,c,i,j}$ 和 $z^T_{b,c,i,j}$。

教师预测概率为

$$p^T_{b,c,i,j} = \sigma(z^T_{b,c,i,j}),$$

其中，σ 为 Sigmoid 函数。跨度困难度定义为

$$d_{b,c,i,j} = 1 - 2\left|p^T_{b,c,i,j} - 0.5\right|. \tag{5}$$

概率越接近 0.5，困难度越高；实现时按 $|z^T_{b,c,i,j}|$ 升序选择。

2.4.2 候选跨度集合构建

仅选择高不确定跨度可能遗漏教师已确信的真实实体和误报实体。为兼顾这些情形，本文对每个真实样本和非 Drug 类别构建

$$\begin{aligned} \mathcal{C}_{b,c} = \mathcal{G}_{b,c} &\cup \mathcal{P}^T_{b,c} \\ &\cup \mathcal{N}(\mathcal{G}_{b,c}) \cup \mathcal{U}^K_{b,c}, \end{aligned} \tag{6}$$

其中，四部分定义如下。

（1）金标准跨度 $\mathcal{G}_{b,c}$：类别 c 的人工标注跨度。

（2）教师正跨度 $\mathcal{P}^T_{b,c}$：合法区域内得分大于 0 的跨度，

$$\mathcal{P}^T_{b,c} = \{(i,j) \in \Omega_b \mid z^T_{b,c,i,j} > 0\}.$$

（3）边界邻域 $\mathcal{N}(\mathcal{G}_{b,c})$：将金标准跨度 (i,j) 的单侧边界移动一位，

$$\begin{aligned} \mathcal{N}(i,j) = \{&(i-1,j),(i+1,j), \\ &(i,j-1),(i,j+1)\} \cap \Omega_b. \end{aligned}$$

对同类跨度取邻域并集，并排除越界或起点大于终点的跨度。

（4）高不确定跨度 $\mathcal{U}^K_{b,c}$：按教师得分绝对值升序选取最多 K 个合法跨度。开发集确定 $K = 64$，并对每个样本、每个非 Drug 类别分别计算。

四类集合取并集并去重，因此 K 仅限制额外高不确定跨度，最终候选数可以超过 64。边界邻域等候选不会被统一视为负例，而是保留各自的教师概率作为软目标，从而降低硬伪标签错误被放大的风险。

2.4.3 软目标一致性损失

教师为入选跨度提供停止梯度传播的软目标：

$$q_{b,c,i,j} = \text{stopgrad}\left(\sigma(z^T_{b,c,i,j})\right).$$

学生预测概率为

$$p^S_{b,c,i,j} = \sigma(z^S_{b,c,i,j}).$$

记候选 $s = (b,c,i,j)$，其一致性损失为

$$\begin{aligned} \ell_{\text{con}}(s) = &-q_s \log p^S_s \\ &-(1-q_s)\log(1-p^S_s). \end{aligned} \tag{7}$$

实际采用带 Logits 的二元交叉熵。

设当前批次真实样本的候选集合为

$$\mathcal{C} = \{(b, c, i, j) \mid b \in \mathcal{B}_r,\ c \in \mathcal{Y}_n,\ (i, j) \in \mathcal{C}_{b,c}\},$$

其中，$\mathcal{B}_r$ 为当前批次真实样本。一致性损失为

$$\mathcal{L}_{\text{con}} = \frac{1}{|\mathcal{C}|} \sum_{s \in \mathcal{C}} \ell_{\text{con}}(s). \tag{8}$$

入选跨度等权计算，困难度仅用于选择而不作为损失权重。教师概率经停止梯度后保留其置信程度，相比把候选统一转为正、负硬标签，可减少教师误判对学生的直接放大。该损失只作用于真实样本的 Taste、Effect 和 Disease；增强样本不参与一致性学习，Drug 仍接受真实与增强样本的硬标签监督。

2.5 联合训练与推断

完整模型将真实样本监督、增强样本 Drug 监督与非 Drug 跨度一致性学习相结合，其目标为

$$\mathcal{L} = \mathcal{L}_r + \mathcal{L}_a + \lambda_t \mathcal{L}_{\text{con}}, \tag{9}$$

其中，t 为训练轮次，一致性权重在前三轮逐步增加：

$$\lambda_t = 0.05 \min\left(1, \frac{t}{3}\right). \tag{10}$$

前两轮权重约为 0.0167 和 0.0333，此后固定为 0.05。

训练开始时，教师参数由学生参数初始化。每次学生完成梯度更新后，教师按照下式更新：

$$\theta_T \leftarrow \alpha_t \theta_T + (1 - \alpha_t)\theta_S, \tag{11}$$

其中，

$$\alpha_t = \begin{cases} 0, & 1 \le t \le 5, \\ 0.95, & t \ge 6. \end{cases} \tag{12}$$

前五轮教师在每次学生更新后同步参数，但学生开启 Dropout、教师关闭 Dropout，因此仍能约束扰动预测与确定性预测的一致性。第六轮起教师采用指数移动平均，以整合学生在不同训练时刻的参数。教师更新覆盖共享编码器及全部类别参数；Drug 仅不接受一致性软目标，并未从教师更新中排除。

根据开发集整体微平均 F_1 选择教师检查点，预测实体集合为

$$\widehat{\mathcal{E}}_b = \{(i, j, c) \mid (i, j) \in \Omega_b,\ c \in \mathcal{Y},\ z^{T}_{b,c,i,j} > 0\}.$$

解码阈值固定为 0；推断不调用大语言模型或训练阶段候选集合。

3 实验与分析

3.1 数据集与实验设置

3.1.1 数据来源与标注

《藏药志》扫描文档经光学字符识别（optical character recognition，OCR）、人工整理和条目规范化后得到 408 条有效药材条目，并用 Label Studio 标注药材、性味、药效和病症四类平面实体，保存实体文本、起止位置和类别。药材类别包含标题目标药材及正文重复指称，不包含来源名称、相关药名和比较对象；性味对应性质与味道，药效对应功能与治疗作用，病症对应疾病、症状及病理状态。

两名标注者在制定规范和少量试标后完成标注并交叉复核，对分歧边界和类别经讨论确定，最后以程序检查标签集合、边界和异常重叠。在双标样本上基于字符级 BIO 标签计算两名标注者的 Cohen's Kappa 系数，结果约为 0.81。考虑到条目包含考证、来源、性味和功用等栏目，本文按栏目结构和中文标点切分文本；仅含标题的过短片段与后续正文合并，标题实体按同一规范标注。数据按条目编号分组划分，保证同一条目的片段不跨训练、开发和测试集，避免同源信息泄漏。

数据集统计如表1所示，共包含 1 505 个文本片段和 5 991 个实体。训练集包含 326 条药材条目、1 073 个文本片段和 4 934 个实体，开发集与测试集分别包含 40 条和 42 条药材条目。

表 1 《藏药志》命名实体识别数据集统计

Tab.1 Statistics of the Zangyaozhi named entity recognition dataset

统计项	训练集	开发集	测试集	合计
原始药材条目数	326	40	42	408
文本片段数	1 073	184	248	1 505
药材实体提及数	1 138	125	150	1 413
性味实体提及数	752	89	80	921
药效实体提及数	1 154	99	145	1 398
病症实体提及数	1 890	164	205	2 259
实体提及总数	4 934	477	580	5 991

注：原始药材条目数表示《藏药志》中的独立条目数量；实体提及数按标注出现次数统计，同一实体在不同位置重复出现时分别计数。

表1仅统计真实标注数据。M1 构造的 137 条增强文本仅加入训练集，提供 302 个额外 Drug 标注，

不改变开发集和测试集。加入增强文本后，训练文本总数为 1 210，Drug 标注总数为 1 440；其他类别不增加增强样本监督。

3.1.2 评价指标

本文采用严格实体级完全匹配，预测的起止位置和类别均正确时才计为命中。指标 P、R 和 F_1 为

$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}},$$
$$R=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}},$$
$$F_1=\frac{2PR}{P+R}.$$

其中，TP、FP 和 FN 分别为正确、误报和漏报实体数，边界不完全一致不计为正确。整体指标先汇总四类实体的计数再进行微平均，同时报告各类 F_1 以分析类别差异。所有模型使用相同文本片段、数据划分和评价标准，不额外拼接条目标题字段；开发集仅用于早停、候选配额及检查点选择，测试集仅用于最终评价，不参与困难样本构造或超参数选择。

3.1.3 训练设置

编码器为 MacBERT-large，解码器为带 RoPE 的 GlobalPointer，任务头维度 64；最大长度 192，训练与评估批大小分别为 128 和 192。AdamW[20]的编码器和任务头学习率分别为 2×10^{-5} 和 1×10^{-4}，权重衰减 0.01；前 6% 步数线性预热后余弦衰减。训练采用 BF16、梯度裁剪 1.0，最多 100 轮，开发集 F_1 连续 12 轮不提升时早停。

完整模型联合真实样本四类监督、增强样本 Drug 监督和真实样本非 Drug 一致性损失。候选额外配额为 $K=64$，一致性权重在前三轮增至 0.05；教师前五轮同步学生，此后以 0.95 进行 EMA 更新。GlobalPointer 解码阈值固定为 0，不通过开发集或测试集搜索；最终测试使用开发集选出的教师检查点，推断阶段不调用大语言模型。

3.2 主实验结果与分析

本文比较 BiLSTM-CRF、BERT-base-CRF、BERT-base-GlobalPointer、W2NER、MacBERT-large-CRF 及基础模型 MacBERT-large-GlobalPointer，并补充 Qwen 直接提示、少样本和 LoRA 结果，见表2。

表 2 不同模型在《藏药志》数据集上的实验结果（%）

Tab.2 Experimental results of different models on the Zangyaozhi dataset (%)

模型	整体 P	整体 R	整体 F_1	药材 F_1	性味 F_1	药效 F_1	病症 F_1
BiLSTM-CRF	84.88	85.17	85.03	75.32	91.25	87.86	87.92
BERT-base-CRF	81.72	90.17	85.74	77.42	**95.71**	90.07	85.51
BERT-base-GP	83.26	91.72	87.28	77.55	93.59	**93.92**	88.21
W2NER	87.63	87.93	87.78	78.81	94.41	91.22	89.33
MacBERT-large-CRF	87.89	87.59	87.74	79.32	94.34	91.95	88.18
MacBERT-large-GP（Base）	84.74	91.90	88.17	80.88	92.22	92.93	88.73
Qwen 直接提示	41.24	59.66	48.77	19.26	33.52	60.11	81.82
Qwen 少样本	47.16	78.79	59.01	34.31	69.14	68.80	83.50
Qwen-LoRA	80.51	86.90	83.58	74.03	88.89	88.16	85.93
本文方法	**88.09**	**93.10**	**90.53**	**86.52**	93.75	93.38	**90.29**

注：GP 表示 GlobalPointer，粗体表示各列最优结果。

BiLSTM-CRF 和 BERT-base-CRF 的整体 F_1 分别为 85.03% 和 85.74%，但药材 F_1 仅为 75.32% 和 77.42%，且其他模型也呈现药材指标低于多数非药材类别的趋势，说明药材是主要难点。GlobalPointer 相较同编码器 CRF 分别提高 1.54 和 0.43 个百分点；MacBERT-large-GP 又较 BERT-base-GP 提高整体 0.89 个百分点、药材 3.33 个百分点，因此将其作为 Base。Qwen 直接提示、少样本和 LoRA 的整体

F_1 分别为 48.77%、59.01% 和 83.58%，表明少量示例和任务微调虽能改善大模型识别，但本文方法仍较 Qwen-LoRA 提高 6.95 个百分点。

完整模型的精确率、召回率和整体 F_1 分别为 88.09%、93.10% 和 90.53%，较 Base 提高 3.35、1.20 和 2.36 个百分点，说明收益同时来自误报减少与漏检改善，并非只依靠增加预测数量。药材 F_1 由 80.88% 升至 86.52%，性味、药效和病症分别提高 1.53、0.45 和 1.56 个百分点。虽然 BERT-base-CRF 的性味和 BERT-base-GP 的药效单项更高，本文方法仍取得最优整体、药材和病症结果，体现了跨类别的综合平衡。下面通过消融实验分析模块贡献。

3.3 消融实验

为分析 M1 和 M2 的作用，本文比较 Base、Base+M1、Base+M2 和 Full 四种设置，结果如表3所示。其中，Base+M2 不使用 M1 增强数据。

表 3 不同模块组合的消融实验结果（%）

Tab.3 Ablation results of different module combinations (%)

模型	整体 F_1	药材 F_1	性味 F_1	药效 F_1	病症 F_1
Base	88.17	80.88	92.22	92.93	88.73
Base+M1	88.85	84.91	90.91	93.07	88.02
Base+M2	90.22	84.05	**94.41**	93.33	**91.17**
Full	**90.53**	**86.52**	93.75	**93.38**	90.29

M1 使药材 F_1 由 80.88% 升至 84.91%，M2 使整体 F_1 由 88.17% 升至 90.22%；联合后两项达到 90.53% 和 86.52%，表明二者互补。Full 的性味和病症低于 Base+M2，其优势主要在整体效果和药材识别。

3.4 错误驱动增强分析

折外预测得到 98 次 Drug 漏检和 104 次误报，其中 83 次完整漏检、71 次语义混淆误报用于构造困难样本。在不启用 M2 时，比较普通 Qwen 增强、随机过采样、仅困难正例、仅困难负例和 M1。普通增强与 M1 均含 76 条正例和 61 条负例，采用相同审核与 Drug 监督，区别仅在是否依据折外错误构造场景，结果见表4。

表 4 不同样本增强策略的实验结果（%）

Tab.4 Experimental results of different sample augmentation strategies (%)

方法	增强条数	整体 F_1	药材 P	药材 R	药材 F_1
Base	0	88.17	76.33	86.00	80.88
普通 Qwen 增强	137	87.62	79.76	89.33	84.28
随机过采样	137	**88.85**	76.24	**92.00**	83.38
仅困难正例	76	88.27	80.24	89.33	84.54
仅困难负例	61	88.68	73.26	91.33	81.31
正负例联合（M1）	137	**88.85**	**80.36**	90.00	**84.91**

M1 的药材 F_1 为 84.91%，较 Base、普通 Qwen 增强和随机过采样提高 4.03、0.63 和 1.53 个百分点；仅正例和仅负例分别为 84.54% 和 81.31%。与样本数及比例相同的普通增强相比，M1 的药材精确率和召回率提高 0.60 和 0.67 个百分点，说明错误驱动构造取得更好平衡。

为进一步分析错误类型，统计严格实体级 Drug 预测计数。Base 的 TP、FP 和 FN 分别为 129、40 和 21，Base+M1 分别为 135、33 和 15。即正确实体增加 6 个，误报减少 7 个，未正确匹配的金标准实体减少 6 个；其中 FP 和 FN 包含全部边界及类别不匹配情形，不能直接等同于某一种语义混淆错误的修复数量。具体典型纠错实例见表5。

表 5 M1 增强前后药材实体的典型纠错案例

Tab.5 Representative drug-entity corrections before and after M1

类型	典型文本与预测差异
漏检	“《晶珠本草》记载：**夹**的角治妇女经闭。” Base 漏检正文“夹”，M1 正确识别该 Drug 实体。
误报	“其基本特征与**茶多**完全相同，也为绢云母……” “茶多”属于比较对象，Base 误报为 Drug，M1 予以排除。
误报	“形状似**塔尔鲁** (一种大戟), 仅植物大小有区别。” “塔尔鲁”属于比较对象，Base 误报为 Drug，M1 予以排除。

注：仅展示与判断相关的局部文本和所示位置的 Drug 预测。

上述真实预测案例表明，M1 能够纠正部分目标药材的正文漏检以及比较对象引发的错误识别；案例仅提供定性解释，并不能据此断言误报下降全部来自某一类干扰名称。

3.5 选择性跨度一致性学习分析

3.5.1 模块组成分析

在固定 M1 的条件下，本文比较 EMA、跨度选择方式和一致性作用类别，结果如表6所示。全跨度一致性覆盖合法预测区域，困难跨度一致性采用第 2 节定义的候选集合；非 Drug 表示仅约束性味、药效和病症类别。

表 6 固定 M1 条件下的 M2 组件消融结果（%）
Tab.6 Ablation results of M2 components with M1 fixed (%)

设置	整体 F_1	药材 F_1	性味 F_1	药效 F_1	病症 F_1
M1	88.85	84.91	90.91	93.07	88.02
M1+EMA	89.46	84.31	90.32	92.76	90.51
EMA+ 全跨度 + 全类	90.02	85.06	**93.75**	92.67	90.34
EMA+ 全跨度 + 非 Drug	90.13	85.71	**93.75**	92.88	90.05
EMA+ 困难 + 全类	90.36	84.18	**93.75**	**94.31**	**90.91**
Full（困难 + 非 Drug）	**90.53**	**86.52**	**93.75**	93.38	90.29

注：除 M1 外，设置均包含 EMA；“困难”表示困难跨度一致性。

仅加入 EMA 时整体 F_1 为 89.46%，全跨度非 Drug 一致性提高至 90.13%，困难跨度进一步达到 90.53%，故作为最终设置。

各目标共享编码器，性能增益并非线性叠加。相较困难跨度全类别一致性，最终模型仅约束非 Drug 类别，使药材和整体 F_1 提高 2.34 和 0.17 个百分点，性味不变，药效和病症下降 0.93 和 0.62 个百分点；相较 M1，后三类仍提高 2.84、0.31 和 2.27 个百分点。因此，其优势是药材与非 Drug 识别间更好的整体平衡，而非所有单项均最高。

3.5.2 跨度选择策略与候选配额

在相同候选配额下，Random-64 与 Difficult-64 的整体 F_1 值分别为 89.80% 和 90.53%，药材 F_1 值分别为 85.16% 和 86.52%，说明困难跨度选择优于随机选择。开发集上 $K = 32$、64、128 时的整体 F_1 值分别为 90.68%、90.87% 和 90.46%，故最终取 $K = 64$。

3.5.3 学生模型与教师模型对比

学生模型的整体精确率、召回率和 F_1 值分别为 86.61%、90.34% 和 88.44%，EMA 教师模型分别为 88.09%、93.10% 和 90.53%。教师模型的整体 F_1 值提高 2.09 个百分点，且四类实体指标均有改善，因此推断阶段采用教师模型。

3.6 多随机种子稳定性分析

为考察结果稳定性，本文对 Base、Base+M1、Base+M2 和 Full 分别进行 3 次重复实验，结果如表7所示。

表 7 3 个随机种子实验结果（均值 ± 标准差，%）
Tab.7 Results over three random seeds (mean ± standard deviation, %)

指标	Base	Base+M1	Base+M2	Full
整体 F_1	87.98 ± 0.19	88.96 ± 0.29	89.94 ± 0.24	$\mathbf{90.60 \pm 0.23}$
药材 F_1	80.49 ± 0.35	85.02 ± 0.14	84.15 ± 0.45	$\mathbf{86.04 \pm 0.42}$
性味 F_1	93.10 ± 1.15	91.48 ± 0.50	$\mathbf{94.42 \pm 1.18}$	93.52 ± 1.02
药效 F_1	91.73 ± 1.26	92.51 ± 0.68	93.14 ± 0.37	$\mathbf{93.81 \pm 0.50}$
病症 F_1	89.33 ± 0.53	88.38 ± 0.58	90.38 ± 0.71	$\mathbf{90.64 \pm 0.31}$

注：粗体表示各行均值最高的结果，不表示标准差最小。

完整模型的整体和药材 F_1 分别为 $90.60 \pm 0.23\%$ 和 $86.04 \pm 0.42\%$，较 Base 提高 2.62 和 5.55 个百分点，并在整体、药材、药效和病症四项取得最高均值。结果与主实验的变化方向一致，说明改进并非由单次随机初始化造成。以相同随机种子下 Base 与 Full 的整体 F_1 为 3 组成对观测，双侧配对 t 检验得到 $p = 0.0043$；由于样本量有限，该结果仅作为辅助统计证据。

3.7 跨文献泛化实验

为考察方法在不同藏医药文献上的迁移效果，构建《四部医典》外部评价集，包含 28 个文本片段、731 个实体提及，其中药材、性味、药效和病症分别为 264、63、170 和 234 个。采用四类实体标签及严格实体级完全匹配标准，对 Base 与 Full 进行对比评价；这些外部数据未并入原有《藏药志》训练、开发和测试集。需要注意，《四部医典》与按目标药材条目组织的《藏药志》在药材提及的语义角色上可能存在标注范围差异，因此相关评价结果反映初步跨文献迁移表现。

由表8可见，Full 整体微平均 F_1 由 82.292% 升至 83.285%，增加 0.993 个百分点；药材、药效和病症分别提高 1.399、0.765 和 1.196 个百分点，而性味下降 1.172 个百分点。整体精确率由 88.889% 降至 87.976%，召回率由 76.607% 升至 79.070%，说明整体改善主要来自召回率提升。两模型在外部语料上的整体 F_1 均低于《藏药志》测试集表现，外部样本仅来自一部文献且四类实体数量不均衡，不能据此

表 8 Base 与 Full 在《四部医典》外部评价集上的结果（%）

Tab.8 Cross-document results of Base and Full on the Four Medical Tantras (%)

类别	Base P	Base R	Base F_1	Full P	Full R	Full F_1	ΔF_1
药材	94.737	68.182	79.295	94.416	70.455	80.694	+1.399
性味	86.957	63.492	73.394	86.667	61.905	72.222	-1.172
药效	92.025	88.235	90.090	91.124	90.588	90.855	+0.765
病症	82.251	81.197	81.720	80.894	85.043	82.917	+1.196
整体（微平均）	88.889	76.607	82.292	87.976	79.070	83.285	+0.993

注：ΔF_1 为 Full 相对 Base 的百分点差值。

断言方法对不同文献类型或 OCR 噪声具有普遍鲁棒性。

4 结论

本文提出融合错误驱动增强与跨度一致性学习的《藏药志》实体识别方法。该方法依据折外预测定位药材漏检和语义混淆误报，在原文约束下构造 Drug 选择性监督样本；同时筛选真实样本的非 Drug 跨度，以教师软目标进行一致性约束。推断仅使用教师模型。

完整模型的整体和药材 F_1 分别为 90.53% 和 86.52%，较 Base 提高 2.36 和 5.64 个百分点。消融实验表明两模块能够互补；典型案例展示了 M1 对部分目标药材漏检和比较名称误报的修正。3 个随机种子下完整模型的整体和药材 F_1 分别为 $90.60\pm0.23\%$ 和 $86.04\pm0.42\%$，整体 F_1 配对 t 检验结果为 $p = 0.0043$。在《四部医典》外部测试集上，完整模型整体微平均 F_1 由 82.29% 提高至 83.29%，提示方法具有一定的跨文献泛化能力，但部分类别仍存在性能波动。

本研究仍存在局限：折外模型和最终模型均使用规模有限的正式开发集选择检查点，可能产生选择偏差；跨文献评价目前仅涉及另一部文献的 28 个片段，且不同文献的药材标注语义可能存在差异；增强样本仅由 1 名研究者审核，尚缺少增强审核者间一致性检验；多随机种子检验的样本量也较小。后续将采用训练集内部验证，扩展外部文献来源与评测规模，完善增强样本多人审核，并探索更灵活的类别约束与候选跨度选择。

参考文献

[1] LI J, SUN A X, HAN J L, et al. A survey on deep learning for named entity recognition[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(1): 50-70.

[2] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition[C]// Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: human language technologies. San Diego, California: Association for Computational Linguistics, 2016: 260-270.

[3] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: human language technologies, volume 1 (Long and short papers). Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 4171-4186.

[4] CUI Y M, CHE W X, LIU T, et al. Revisiting pre-trained models for chinese natural language processing[C]// Findings of the association for computational linguistics: eMNLP 2020. Online: Association for Computational Linguistics, 2020: 657-668.

[5] LI X N, YAN H, QIU X P, et al. FLAT: chinese NER using flat-lattice transformer[C]//Proceedings of the 58th annual meeting of the association for computational linguistics. Online: Association for Computational Linguistics, 2020: 6836-6842.

[6] LIU W, FU X Y, ZHANG Y, et al. Lexicon enhanced chinese sequence labeling using BERT adapter[C]//Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (Volume 1: long papers). Online: Association for Computational Linguistics, 2021: 5847-5858.

[7] LI X Y, FENG J R, MENG Y X, et al. A unified MRC framework for named entity recognition[C]//Proceedings of the 58th annual meeting of the association for computational linguistics. Online: Association for Computational Linguistics, 2020: 5849-5859.

[8] SU J L, MURTADHA A, PAN S F, et al. Global pointer: novel efficient span-based approach for named entity recognition[EB/OL]. (2022-08-05)[2026-09-05]. https://arxiv.

org/abs/2208.03054.

[9] LI J Y, FEI H, LIU J, et al. Unified named entity recognition as word-word relation classification[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(10): 10965-10973.

[10] WANG S H, SUN X F, LI X Y, et al. GPT-NER: named entity recognition via large language models[C]//Findings of the association for computational linguistics: nAACL 2025. Albuquerque, New Mexico: Association for Computational Linguistics, 2025: 4257-4275.

[11] ZHOU W X, ZHANG S, GU Y, et al. UniversalNER: targeted distillation from large language models for open named entity recognition[C/OL]//The twelfth international conference on learning representations. Vienna, Austria: OpenReview.net, 2024[2026-09-05]. https://openreview.net/forum?id=r65xfUb76p.

[12] 范西朋, 刘云飞, 李盛阳, 等. 基于 MRC 动态数据生成的命名实体识别方法[J]. 中文信息学报, 2023, 37(6): 104-114.

[13] TARVAINEN A, VALPOLA H. Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results[C]//Advances in neural information processing systems: v.30. Red Hook, NY: Curran Associates, Inc., 2017: 1195-1204.

[14] 吕书宁, 刘健, 徐金安, 等. 基于多任务标签一致性机制的中文命名实体识别[J]. 中文信息学报, 2023, 37(12): 87-97.

[15] 冯嘉琦, 高见, 王明程. 融合双仿射残差卷积和知识扩展的中文网络威胁情报命名实体识别[J]. 中文信息学报, 2025, 39(9): 100-115,125.

[16] ZHU E W, LI J P. Boundary smoothing for named entity recognition[C]//Proceedings of the 60th annual meeting of the association for computational linguistics (Volume 1: long papers). Dublin, Ireland: Association for Computational Linguistics, 2022: 7096-7108.

[17] CAI Y X, LIU Q, GAN Y L, et al. DiFiNet: boundary-aware semantic differentiation and filtration network for nested named entity recognition[C]//Proceedings of the 62nd annual meeting of the association for computational linguistics (Volume 1: long papers). Bangkok, Thailand: Association for Computational Linguistics, 2024: 6455-6471.

[18] SU J L, AHMED M H M, LU Y, et al. RoFormer: enhanced transformer with rotary position embedding[J]. Neurocomputing, 2024, 568: 127063.

[19] YANG A, LI A F, YANG B S, et al. Qwen3 technical report [EB/OL]. (2025-05-14)[2026-09-05]. https://arxiv.org/abs/2505.09388.

[20] LOSHCHILOV I, HUTTER F. Decoupled weight decay regularization[C/OL]//The seventh international conference on learning representations. New Orleans, Louisiana: OpenReview.net, 2019[2026-09-05]. https://openreview.net/forum?id=Bkg6RiCqY7.