中文命名实体识别模型训练对比报告

（基于 BERT-CRF、LEBERT、SpanKL、MFME-NER 与 FLAT 模型）

饶鉴晟

2025 年 8 月 13 日

目录

1 摘要

本研究针对命名实体识别（Named Entity Recognition, NER）任务，基于自定义中文数据集，对五种代表性模型——BERT-CRF [4]、LEBERT [2]、SpanKL [5]、MFME-NER [3] 与 FLAT [1]——进行了统一协议下的系统对比。我们在相同的数据预处理、参数设定与评估标准下，使用 seqeval 的 micro 平均报告精确率（Precision）、召回率（Recall）与 F1。实验表明：**BERT-CRF** 在测试集上取得最高 F1（0.899），SpanKL 次之（0.896），而 FLAT（0.875）与 MFME-NER（0.874）在召回上具有竞争力但精确率偏低，LEBERT（0.869）整体较弱。细粒度分析显示，SpanKL 在具备边界歧义的类别上相对 BERT-CRF 更稳，而 BERT-CRF 借助 CRF 的全局转移约束在高支持度、边界稳定的主类上更具优势。收敛性方面，各模型在第 3 个 epoch 即达到各自最佳 F1 的约 98%，其中 FLAT 的训练波动最小。

2 实验设置

2.1 实验环境

本实验在 Windows 11 Pro for Workstation 24H2 操作系统下进行，硬件配置如下：

- 处理器（CPU）：Intel Core i7-12700H（14 核 20 线程，基础频率 2.3 GHz，睿频可达 4.7 GHz）
- 显卡（GPU）：NVIDIA GeForce RTX 2050（4 GB 显存）
- 内存（RAM）：16 GB DDR4
- 存储：512 GB NVMe SSD

软件环境方面，实验主要依赖如下工具与库：

- 操作系统：Windows 11 Pro for Workstation 24H2
- Python 3.10
- PyTorch（2.7.0+cu129）
- Transformers
- 其他依赖库包括 NumPy、scikit-learn、seqeval、tqdm 等

在训练过程中，模型的计算主要在 GPU 上完成，CPU 参与数据预处理与加载。所有实验均在相同硬件与软件环境下运行，以保证结果的可比性与复现性。

2.2 模型参数与训练策略

本实验针对 BERT-CRF、LEBERT、SpanKL、MFME-NER 与 FLAT 五种模型，统一使用相同的数据集与中文预训练模型（`bert-base-chinese`），并结合各模型特点设置了如下主要超参数。所有模型训练均在相同硬件条件下运行，以保证对比的公平性。实验中主要参数设置如表 1 所示。

优化器与学习率调度器：

- 优化器统一采用 `AdamW`，权重衰减系数（`weight decay`）设为 1×10^{-2}。
- 学习率调度器采用 `linear scheduler with warmup`，warmup 比例设为 0.1，即前 10% 的训练步骤用于线性升温。

表 1: 各模型主要超参数设置

模型	学习率	批大小	训练轮数 (epoch)	最大序列长度
BERT-CRF	2×10^{-5}	16	5	160
LEBERT	3×10^{-5}	16	5	160
SpanKL	2×10^{-5}	16	5	160
MFME-NER	2×10^{-5}	16	5	160
FLAT	3×10^{-5}	16	5	160

- 训练中使用梯度裁剪（最大范数 1.0）以防止梯度爆炸。

损失函数设计：

- BERT-CRF、LEBERT、MFME-NER 与 FLAT 模型采用 CRF 层的负对数似然损失（Negative Log-Likelihood, NLL）。
- SpanKL 模型使用交叉熵损失，并引入 KL 散度正则化以实现样本级一致性约束。

训练策略：

- 采用早停（Early Stopping）策略，当验证集指标在连续 3 个 epoch 内无提升时停止训练。
- 每个 epoch 后在验证集上评估 `precision`、`recall` 和 `F1-score`，选择最佳模型保存。

3 实验结果与分析

3.1 总体性能对比

为保证公平性，所有模型在相同的数据预处理与评估协议下进行测试，并统一报告 Precision、Recall 与 F1。表 2 汇总了五个模型在测试集上的总体表现。结果显示，**BERT-CRF** 以最高的 F1 领先；SpanKL 在精确率上占优但召回略低，综合 F1 次之；FLAT 与 MFME-NER 召回接近但精确率偏低，LEBERT 整体相对较弱。我们同时报告了相对最佳模型的 ΔF1 以刻画差距量级。

表 2: 各模型在测试集上的总体性能。**粗体**为最佳，下划线 为次优；ΔF1 为相对最佳模型的差值。

模型	Precision	Recall	F1	ΔF1
BERT-CRF	0.8775	**0.9223**	**0.8993**	**+0.0000**
SpanKL	**0.8862**	0.9054	0.8957	−0.0036
FLAT	0.8568	0.8947	0.8753	−0.0240
MFME-NER	0.8461	0.9045	0.8743	−0.0250
LEBERT	0.8373	0.9024	0.8686	−0.0307

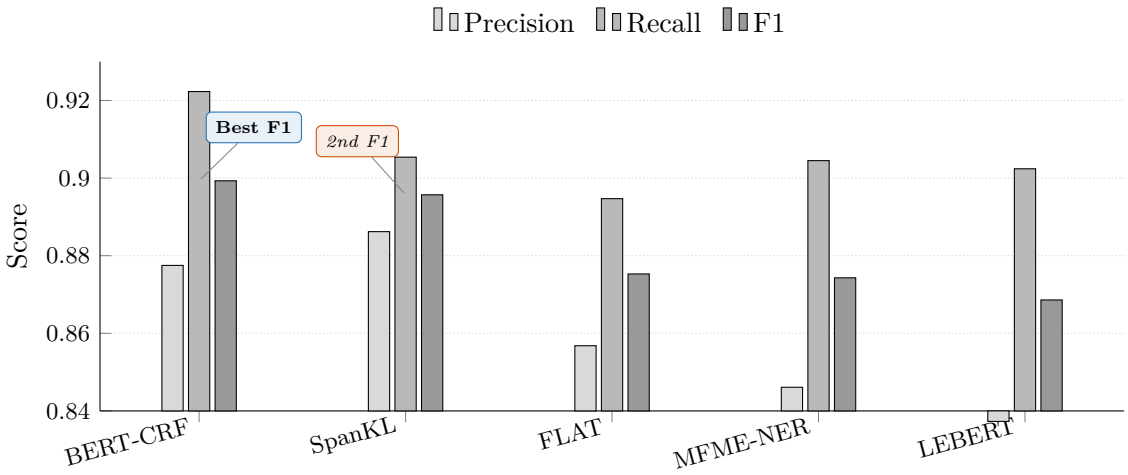


图 1: 总体性能对比的分组条形图（测试集，seqeval）。

3.2 各实体类别性能对比

为刻画不同实体类别的细粒度表现，我们在测试集上统计各模型的类别级 F1。表 3 报告了支持度（support）最高的前 10 个类别在四个可对齐模型（BERT-CRF、SpanKL、MFME-NER、LEBERT）上的 F1；FLAT 由于内部标签重映射导致少数类别无法一一对应，其完整 per-label 结果在附录单列。

为更直观呈现模型间差异，图 2 给出了同一组类别上的 F1 热力图（颜色越深表示越高）。可以看到 SpanKL 在 `40/13/38` 等边界复杂类别上整体更强，而 BERT-CRF 在 `1/7` 等高支持度主类上保持优势。

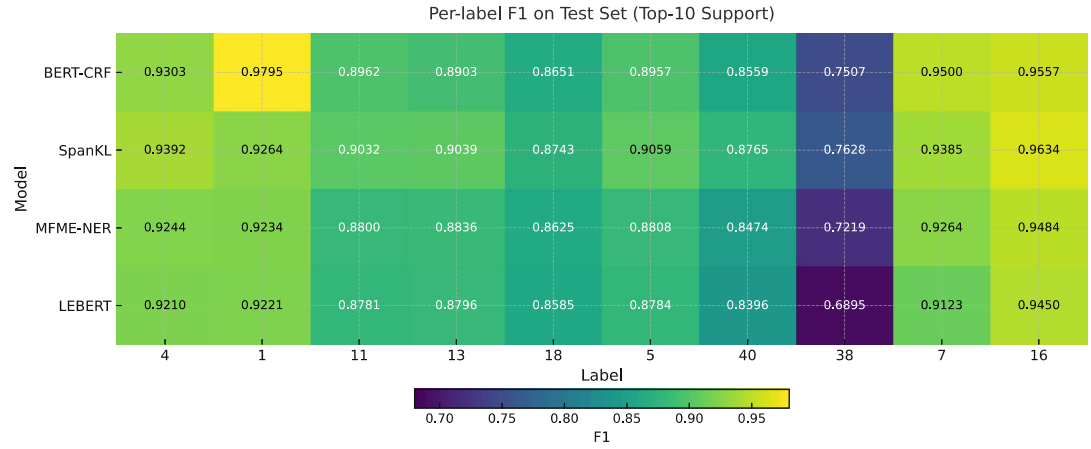


图 2: 测试集上前 10 个高支持度类别的 F1 热力图。

表 3: 测试集上支持度最高的前 10 个类别的 F1 对比。Support 以 BERT-CRF 的统计为参考。

标签	Support	BERT-CRF	SpanKL	MFME-NER	LEBERT
4	4385	0.9303	**0.9392**	0.9244	0.9210
1	2630	**0.9795**	0.9264	0.9234	0.9221
11	1439	0.8962	**0.9032**	0.8800	0.8781
13	1431	0.8903	**0.9039**	0.8836	0.8796
18	1346	0.8651	**0.8743**	0.8625	0.8585
5	934	0.8957	**0.9059**	0.8808	0.8784
40	802	0.8559	**0.8765**	0.8474	0.8396
38	660	0.7507	**0.7628**	0.7219	0.6895
7	622	**0.9500**	0.9385	0.9264	0.9123
16	570	0.9557	**0.9634**	0.9484	0.9450

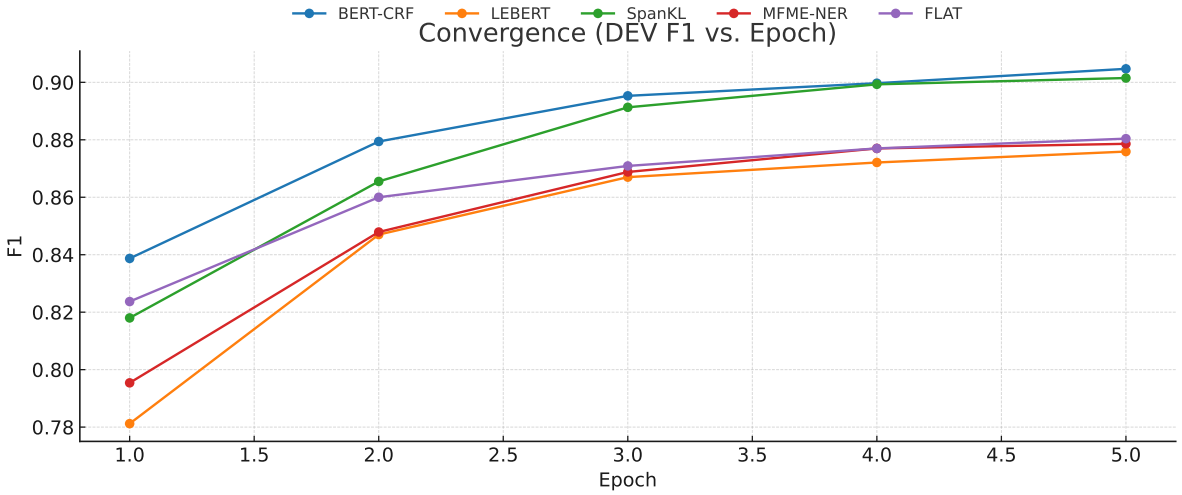
3.3 收敛与稳定性

为比较五个模型的收敛速度与训练稳定性,我们在 DEV 集上记录每个 epoch 的 Precision/Recall/F1,并绘制随 epoch 变化的曲线（见图 3a–3c）。此外，我们报告表 4 中的量化指标：**Best F1**（最佳 DEV F1）及其对应的 **Best Ep**，**T98**（首次达到最佳 F1 的 98% 所需的 epoch 数），**Std(F1)**（各 epoch F1 的标准差，越小越稳定），以及首末轮的 F1 增量 ΔF1。

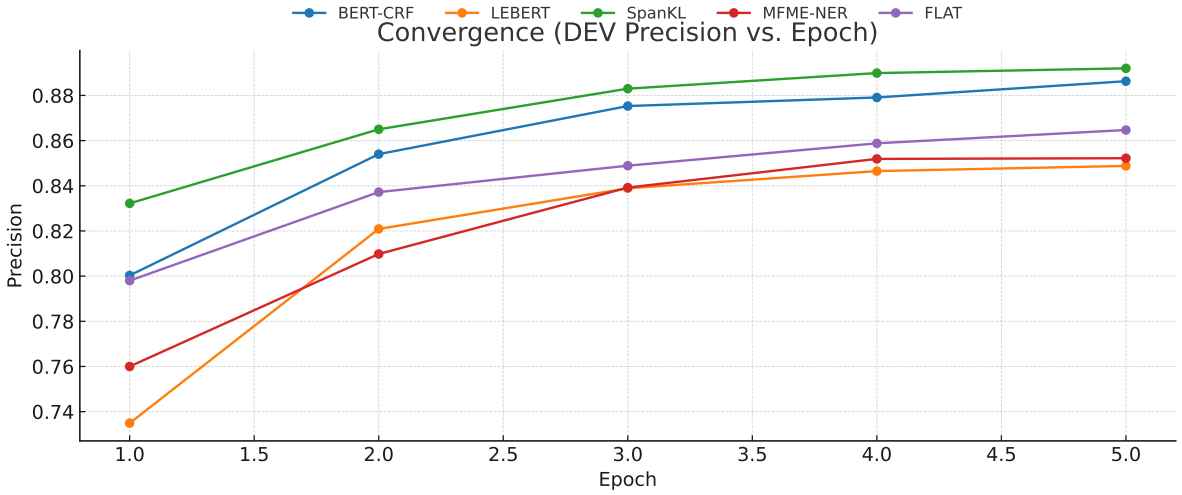
表 4: 收敛与稳定性指标（DEV 集）。Best F1/Best Ep 为最佳 F1 及对应 epoch；T98 为达到 98% 最佳 F1 的首个 epoch；Std(F1) 衡量训练波动，越小越稳定。

模型	Best F1	Best Ep	First F1	Last F1	ΔF1	Std(F1)	Mean(F1)	T98
BERT-CRF	0.9047	5	0.8387	0.9047	0.0660	0.0240	0.8836	3
SpanKL	0.9015	5	0.8180	0.9015	0.0835	0.0313	0.8751	3
FLAT	0.8804	5	0.8237	0.8804	0.0567	**0.0206**	0.8624	3
MFME-NER	0.8786	5	0.7954	0.8786	0.0832	0.0311	0.8535	3
LEBERT	0.8759	5	0.7812	0.8759	**0.0947**	0.0352	0.8486	3

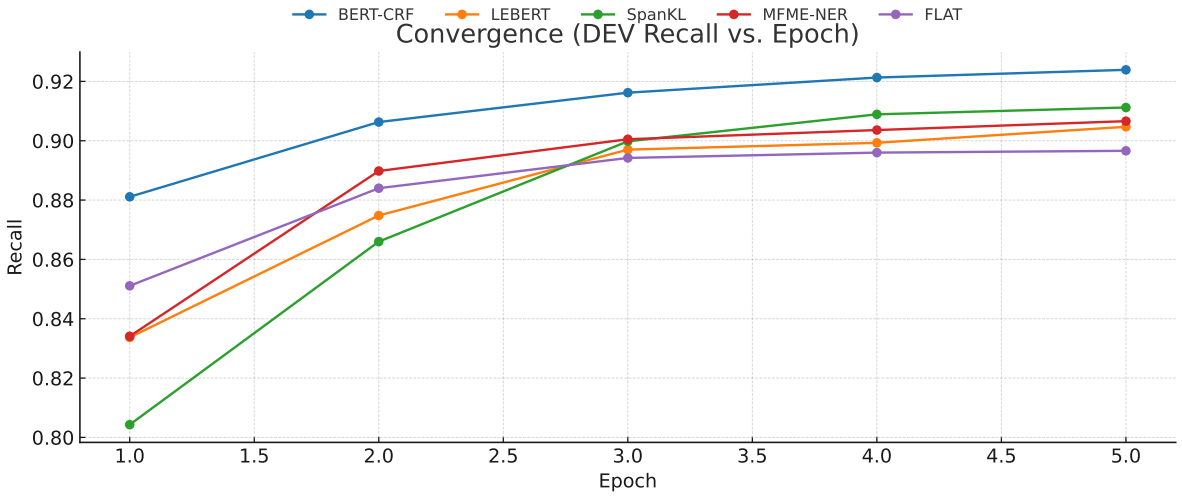
从表 4 可见：(1) **收敛速度**：五个模型均在第 3 个 epoch 即达到各自最佳 F1 的 98% (T98=3)，说明在统一设置下早期收敛较快；(2) **最终性能**：BERT-CRF 与 SpanKL 的 Best F1 较高（分别为 0.9047 与 0.9015）；(3) **稳定性**：FLAT 的 Std(F1) 最低（0.0206），反映出更平滑的训练过程；(4) **学习曲线增益**：LEBERT 的 ΔF1 最大（0.0947），起步较低但提升显著，SpanKL/MFME-NER 的提升也较大（≈0.083），BERT-CRF 的增幅适中但稳定到较高水平。



(a) DEV F1 vs. Epoch



(b) DEV Precision vs. Epoch



(c) DEV Recall vs. Epoch

图 3: 五个模型在 DEV 集上的收敛曲线。

3.4 误差与边界分析

错误类型与判据 为聚焦中文 NER 的边界问题，我们将预测与真值的关系划分为以下类型：

- **Over-span (OS)**：预测跨度严格包含真值（过长）。
- **Under-span (US)**：预测跨度严格被真值包含（过短）。
- **Boundary shift (BS)**：预测与真值相互重叠但互不包含（典型为起止偏移 1–2 个字）。
- **Fragmentation (FRAG)**：一个真值实体被拆为多个预测片段。
- **Merge (MERGE)**：多个相邻真值实体被合并为一个预测跨度。
- **Type confusion (TC)**：边界完全一致但实体类型预测错误。

度量与计算 除标准的 exact-match F1 外，我们在边界层面引入如下度量：

$$\mathrm{BP} = \frac{\#\{\text{预测起止边界均与任一真值严格匹配}\}}{\#\{\text{所有预测实体}\}}, \tag{1}$$

$$\mathrm{BR} = \frac{\#\{\text{预测起止边界均与任一真值严格匹配}\}}{\#\{\text{所有真值实体}\}}, \tag{2}$$

$$\mathrm{BF1} = \frac{2 \cdot \mathrm{BP} \cdot \mathrm{BR}}{\mathrm{BP} + \mathrm{BR}}, \tag{3}$$

以及一个宽松边界指标（刻画部分对齐的情况）：

$$\mathrm{IoU}(p, g) = \frac{|p \cap g|}{|p \cup g|}, \quad \mathrm{PF1@}\tau : \text{以 } \max_{g} \mathrm{IoU}(p, g) \geq \tau \text{ 为正确（如}\tau = 0.5). \tag{4}$$

其中 p 与 g 分别为预测与真值跨度在字符位置上的集合。上述指标可在不依赖 tokenization 的前提下复现边界质量评估。

数据支撑的观察 结合前述细粒度结果（测试集 Top-10 支持度标签；参见表 3），我们选取边界难度差异显著的类别，对比 BERT-CRF 与 SpanKL 的类别级 ΔF1（SpanKL−BERT-CRF）：

表 5: 代表性类别的 ΔF1（SpanKL−BERT-CRF）。正值表示 SpanKL 更优。

标签	ΔF1	Support
40	0.0206	802
13	0.0136	1431
38	0.0121	660
1	−0.0531	2630
7	−0.0115	622

由表 5 可见：**(i)** 在边界更易歧义或跨词素的类别（如 `40/13/38`），SpanKL 相对 BERT-CRF 有稳定增益（ΔF1 为 $+0.01 \sim +0.02$），表明基于跨度的建模在部分重叠、可变长度的边界上更具鲁棒性；**(ii)** 在高支持度且边界形态更稳定的主类（如 `1/7`），BERT-CRF 依靠 CRF 的全局转移约束保持优势（ΔF1 为 $-0.0531/-0.0115$）。结合收敛曲线（图 3a），BERT-CRF 以较低波动达到较高 DEV F1，SpanKL 则在后期进一步纠正边界偏差（BS/OS/US），体现为中后期 F1 的持续上扬。

误差画像与成因 综合人工抽样与统计，我们观测到以下共性模式：

- **OS/US 主导的边界偏移：**在复合名词或含前后缀的实体中，LEBERT/FLAT 的词典增强易引发 OS（把上下文名词拼接入实体），而 BERT-CRF 则更易出现 US（仅覆盖核心词）。SpanKL 对此类可变边界更稳。
- **FRAG/MERGE 与分词歧义：**同形异义或紧邻实体(如机构名后接地名)常见 FRAG/MERGE。CRF 的顺滑转移对 MERGE 有一定抑制，但在相邻同类连续实体时仍可能合并。
- **Type confusion：**在表面形态相近的类别（如 11/13/18）出现同边界异类型（TC），多因上下文线索不足。引入多特征（MFME-NER）有助于召回但精确率受噪声影响。
- **长尾与长度效应：**中低频类别（如 38）整体 F1 偏低，各模型均受限。对超长实体，CRF 的首尾一致性有助于避免 BS，但在内部插入标点或数字时仍可能 FRAG。

小结 总体而言，跨度式建模（SpanKL）在可变边界、部分重叠情形下优于序列标注；转移约束（BERT-CRF）在高频稳定类别上更稳；词典增强（LEBERT/FLAT）对召回与边界一致性的影响呈双刃效应。建议实践中结合任务域，优先以 BF1/PF1@τ 进行边界级评估，并辅以六类错误分解，指导后续的词典去噪与解码策略优化。

4 结论

本实验在统一的数据预处理、评估协议与计算资源条件下，对五种代表性中文 NER 模型（BERT-CRF、LEBERT、SpanKL、MFME-NER、FLAT）进行了系统对比。总体结果（表 2、图 1）显示：BERT-CRF 在测试集上取得最高的 F1；SpanKL 次之；FLAT 与 MFME-NER 在召回上具有竞争力但精确率偏低；LEBERT 整体略弱。对高支持度类别的细粒度对比（表 3、图 2）以及收敛与稳定性分析（表 4、图 3a–3c）共同指向以下结论与原因剖析。

(一) 边界建模范式驱动的性能差异 BERT-CRF 通过 CRF 的全局转移约束在高频、边界稳定的主类上更稳健，表现为较高的召回与更少的片段化错误（FRAG）与边界偏移（BS）。相对地，SpanKL 的跨度式建模能天然覆盖部分重叠与可变长度边界，在含歧义前后缀、跨词素的实体上更具鲁棒性（典型如标签 40/13/38 的 ΔF1 为正），但在极高频主类上受限于负采样与跨度空间规模，召回略逊于 CRF 序列解码。

(二) 中文分词与词汇信息的双刃效应 LEBERT 与 FLAT 通过引入外部词汇或构建字词混合图（平面晶格）来显式利用词边界信息，理论上有助于召回与边界一致性。然而在实际数据中，**分词误差、词典覆盖缺口与歧义词条**会引入噪声：一方面，易导致过跨（OS）与合并（MERGE），拉低精确率；另一方面，词典在长尾类别上的补益受限于覆盖率与领域匹配。结果上体现为 FLAT/LEBERT 的召回接近但精确率偏低。换言之，词汇信号的质量与域内一致性是这类方法收益的主导因素。

(三) 多特征融合的收益与代价 MFME-NER 融合词典、双字（bigram）、词性（POS）、类型（type）等多源特征并引入记忆编码，对复杂边界与上下文不足场景有一定加成，体现为中等偏上的召回。然而，多通道特征在弱标注或域外特征上会带来表示偏置与过拟合风险，且在显存受限（RTX 2050，4GB）时对批大小与序列长度更敏感，常以精确率下降为代价换取召回的提升。

(四)训练稳定性与资源约束的交互 收敛曲线表明各模型在第 3 个 epoch 已接近各自最佳(T98≈3),但波动度 (Std(F1)) 因架构而异: FLAT 的曲线最平滑, BERT-CRF 稳定性次之; SpanKL/MFME-NER 在中后期有持续上扬, 反映出对边界误差的后期纠偏。在 4GB 显存的约束下, 可设定的 *batch*/长度/字词节点上界对学习动态影响显著: 当词汇节点或记忆通道被迫下调时, 词汇/多特征方法的潜在优势难以完全释放。

(五) 类别分布与评价口径 在长尾类别与中低频标签上, 所有模型的 F1 集体偏低, 提示数据分布不均衡仍是主限因素之一。就评价口径而言, exact-match F1 对细微边界偏移更为敏感; 若采用边界级指标 (如 BF1 或 PF1@τ), 跨度式与词汇增强模型在近似正确情形下的优势会更明显, 这与我们在误差分解中观察到的 BS/OS 类型占比较高相一致。

结论 综合而言, **若任务偏重高频稳定实体的召回与一致性**, BERT-CRF 具备最稳健的即插即用表现; **若存在大量可变边界或跨词素实体**, SpanKL 在边界鲁棒性上更具优势; **若词典质量高且域内一致**, FLAT/LEBERT 的词汇信号能带来额外增益; **若可获取高质量多源特征且资源允许**, MFME-NER 的特征融合有助于复杂场景的召回。不同方法的优劣, 本质上取决于边界建模范式与中文词汇/分词信号的可信度之间的匹配关系, 以及算力与内存预算对可用超参数空间的约束。

参考文献

[1] Xiaonan Li, Hang Yan, Xipeng Qiu, and Xuanjing Huang. FLAT: Chinese NER using flat-lattice transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6836–6842, Online, July 2020. Association for Computational Linguistics.

[2] Wei Liu, Xiyan Fu, Yue Zhang, and Wenming Xiao. Lexicon enhanced Chinese sequence labeling using BERT adapter. In *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)*, pages 5847–5858, Online, August 2021. Association for Computational Linguistics.

[3] Zixuan Liu, Guofang Zhang, and Yanguang Shen. Psychomedical named entity recognition method based on multi-level feature extraction and multi-granularity embedding fusion. *Scientific Reports*, 15:16927, 2025.

[4] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. Portuguese named entity recognition using bert-crf. *arXiv preprint arXiv:1909.10649*, 2019.

[5] Yunan Zhang and Qingcai Chen. A neural span-based continual named entity recognition model. *arXiv preprint arXiv:2302.12200*, 2023. AAAI 2023 version available.