（命名）实体识别 NER 已有古文预训练模型测评实验报告

数据集

数据集 A 取自《新五代史》《北史》等古籍，训练集，验证集和测试集分别为 21128，2831，2148，平均句长为 26，采用 BIO，共有 5 种实体类型，[O，B-NOUN_BOOKNAME，I-NOUN_BOOKNAME，B-NOUN-OTHER ，I-NOUN-OTHER]

数据集 B 取自《春秋谷梁传》等古籍，训练集，验证集和测试集分别为 7528，865，877，平均字长为 85，采用 BIO，共有 13 种实体类型，[O ,B-PER ,I-PER, B-ORG ,I-ORG ,B-LOC ,I-LOC, B-JOB ,I-JOB ,B-WAR, I-WAR, B-BOO, I-BOO]

实验数据下载：
https://github.com/jizijing/C-CLUE
https://github.com/Ethan-yt/CCLUE

评价指标

准确率(Accuracy)

Accuracy 是从整体上衡量模型的性能，即模型预测正确的样本占全部样本的比例。它的计算公式如下：

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

其中，TP（True Positive）表示真正例，即模型预测为正例且实际上也为正例的样本数量；FP（False Positive）表示假正例，即模型预测为正例但实际上为负例的样本数量；TN（True Negative）表示真负例，即模型预测为负例且实际上也为负例的样本数量；FN（False Negative）表示假负例，即模型预测为负例但实际上为正例的样本数量。

精确率(Precision)

Precision 关注的是模型预测为正例的样本中有多少是真正的正例。换句话说，它衡量的是模型预测的正例中有多少是正确的，计算公式为：

$$Precision = \frac{TP}{TP + FP}$$

其中，TP（True Positive）表示真正例，即模型预测为正例且实际上也为正例的样本数量；FP（False Positive）表示假正例，即模型预测为正例但实际上为负例的样本数量。

F1 值(F1 Score)

F1 分数是精确率和召回率的调和平均值，用于衡量模型的准确性。它的计算公式如下：

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

其中，精确率（Precision）是指模型预测为正例的样本中真正为正例的比例，召回率（Recall）是指所有真正的正例样本中被模型预测为正例的比例。

召回率(Recall)

召回率是指在所有真正的正例样本中，被模型正确预测为正例的比例。它的计算公式如下：

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

其中，TP（True Positive）表示真正例，即模型预测为正例且实际上也为正例的样本数量；FN（False Negative）表示假负例，即模型预测为负例但实际上为正例的样本数量。

Micro-Average

Micro-Average 把所有类别的结果汇总起来计算平均值。它将所有类别的贡献视为等同，因此对于样本量大的类别，Micro-Average 更加敏感。

$$\text{Micro-Average Precision} = \frac{\sum TP}{\sum TP + \sum FP}$$

$$\text{Micro-Average Recall} = \frac{\sum TP}{\sum TP + \sum FN}$$

$$\text{Micro-Average F1} = 2 \times \frac{(\text{Micro Precision} \times \text{Micro Recall})}{(\text{Micro Precision} + \text{Micro Recall})}$$

Macro-Average

Macro-Average 分别计算每个类别的性能指标，然后计算这些指标的算术平均值。它给予每个类别同等的权重，无论类别的样本量大小，因此对于样本量小的类别更加敏感。

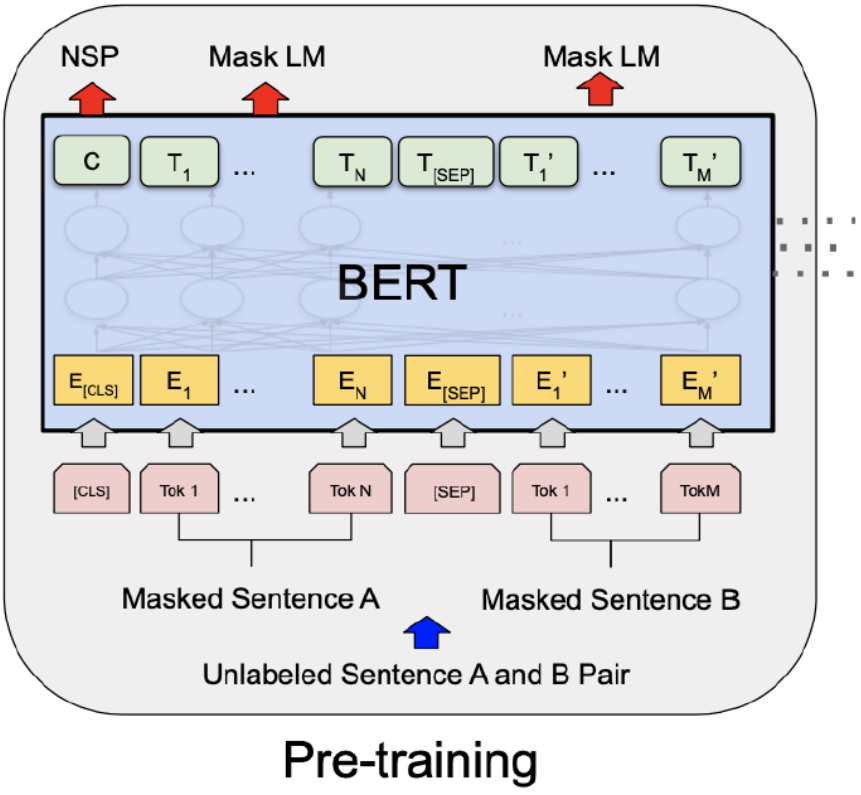
$$\text{Macro-Average Precision} = \frac{1}{N} \sum \frac{TP_i}{TP_i + FP_i}$$

$$\text{Macro-Average Recall} = \frac{1}{N} \sum \frac{TP_i}{TP_i + FN_i}$$

$$\text{Macro-Average F1} = \frac{1}{N} \sum 2 \times \frac{(Precision_i \times Recall_i)}{(Precision_i + Recall_i)}$$

模型

BERT



Pre-training

BertModel 部分包含了 BERT 的所有主要组件，包括词嵌入层、位置嵌入层、令牌类型嵌入层、层归一化层和 Dropout 层。
BertEncoder 部分负责处理输入序列，通过多个 BERT 层来进行特征提取。每个 BERT 层都包含注意力机制和前馈神经网络。
BertForTokenClassification 模型的最后一部分是一个线性分类器，它将 BERT 编码器的输出映射到目标类别的概率分布上。

1.Transformer 架构： BERT 建立在 Transformer 模型的基础上，这是一种使用自注意力机制（Self-Attention Mechanism）的深度神经网络。Transformer 允许模型在处理序列数据时同时关注序列中的所有位置，而不是像传统的循环神经网络（RNN）或卷积神经网络（CNN）那样逐步处理。
2.预训练策略： BERT 采用了无监督的预训练策略，通过大规模的语言模型预训练来学习丰富的语义表示。该模型通过对大量文本数据进行“遮蔽语言模型”（Masked Language Model，MLM）任务的预训练，使得模型能够理解词汇和语法结构，并捕捉单词之间的关系。
3.双向性： BERT 在预训练时考虑了双向信息，即使用上下文信息来理解每个词的语义。这种双向性有助于模型更好地理解文本中的语境和关联，提高了对上下文相关性的捕捉。
4.Fine-tuning： 预训练后，BERT 模型可以通过微调（fine-tuning）来适应特定的下游任务，如命名实体识别、情感分析等。这种能力使得 BERT 在各种 NLP 任务中都表现出色，无需从零开始训练新的模型。
5.Contextual Embeddings： BERT 生成的词向量是上下文相关的，每个词的表示取决于整个输入句子的上下文，而不是简单地从固定的嵌入中获取。这种上下文敏感的嵌入有助于更准确地捕捉语义信息。

实验结果

Bert-ancient-Chinese

训练最终结果

数据集	Eval Precision	Eval Recall	Eval F1	Eval Accuracy	Eval Loss	Eval Runtime
数据集 A	0.971361393018 5667	0.9717125 382262997	0.9715153 155843004	0.9717125382262997	0.1591310352087021	4.3364
数据集 B	0.960773262497 1154	0.9653663 548752834	0.9627669 359301028	0.9653663548752834	0.1790240257978439 3	8.1976

数据集	Micro-Average Precision	Micro-Average Recall	Micro-Average F1	Macro-Average Precision	Macro-Average Recall	Macro-Average F1
数据集 A	0.9717125 382262997	0.9717125 382262997	0.9717125 382262997	0.8673289276579453	0.8478130526942544	0.8573962308701384

数据集B	0.9911334325396826	0.9911334325396826	0.9911334325396826	0.24364633330184382	0.21699121039347588	0.21620752290443754

模型介绍和分析

这是一个基于 BERT 的模型，专门训练用于古汉语处理。它适用于古文的语义理解和文本生成，旨在提高对古汉语的处理能力。

训练过程分析

数据集 A

Epoch	Eval Precision	Eval Recall	Eval F1	Eval Accuracy	Eval Loss	Eval Runtime
2	0.965859360425647	0.9673042704626335	0.9661755533216788	0.9673042704626335	0.10091111063957214	3.5747
4	0.9695756273928118	0.9709185943060499	0.9698405783513404	0.9709185943060499	0.09781475365161896	3.5827
6	0.9748397118458895	0.9753113879003559	0.9749971259961178	0.9753113879003559	0.10068126767873764	3.5825
8	0.9729364759002894	0.9735042259786477	0.9731642014794695	0.9735042259786477	0.11649686098098755	3.5887
10	0.9743676182461154	0.9747553380782918	0.974540298619832	0.9747553380782918	0.13155880570411682	3.5785

数据集 B

Epoch	Eval Precision	Eval Recall	Eval F1	Eval Accuracy	Eval Loss	Eval Runtime
2	0.9634874417593906	0.9647728883219955	0.9636102287355702	0.9647728883219955	0.08292320370674133	7.6235
4	0.9611707205126476	0.9681919642857143	0.9634537970119926	0.9681919642857143	0.08083999902009964	7.7177
6	0.9619109614420617	0.9659775368480725	0.9631737884030217	0.9659775368480725	0.10151440650224686	7.5857
8	0.9607420133460248	0.9666772959183674	0.9630476118906922	0.9666772959183674	0.13081666827201843	7.5832
10	0.9608151673880866	0.9650031887755102	0.9626049702565307	0.9650031887755102	0.15603597462177277	7.5866

SikuRoberta

训练最终结果

数据集	Eval	Eval Recall	Eval F1	Eval	Eval Loss	Eval

	Precision			Accuracy		Runtime
数据集 A	0.968112015577 6729	0.9683438 455657493	0.9682108 01617711 9	0.9683438455657493	0.15556995 570659637	4.3383
数据集 B	0.961705493661 1624	0.9659598 214285714	0.9635556 888246802	0.9659598214285714	0.15483438 968658447	8.1275

数据集	Micro-Average Precision	Micro-Average Recall	Micro-Average F1	Macro-Average Precision	Macro-Average Recall	Macro-Average F1
数据集 A	0.968343 8455657493	0.968343 8455657493	0.968343 8455657493	0.836226 9475264886	0.835406 7662968905	0.835655 0520943984
数据集 B	0.991106 8594104309	0.991106 8594104309	0.991106 8594104309	0.231851 1566076 3794	0.230711 2719035 1204	0.221999 4128760 5457

模型介绍和分析

这个模型是 Siku Quanshu（四库全书）的基础上训练的 RoBERTa 变体。它对古典文献中的文本理解有很强的能力，适合处理古籍和文献分析任务。

训练过程分析

数据集 A

Epoch	Eval Precision	Eval Recall	Eval F1	Eval Accuracy	Eval Loss	Eval Runtime
2	0.9637076986681025	0.9639401690391459	0.9637398833761893	0.9639401690391459	0.1127683147880478	3.5975
4	0.9678154871484845	0.9688056049822064	0.9674825752554453	0.9688056049822064	0.0996222943067507	3.5792
6	0.9705103710023891	0.9706961743772242	0.9705351202725301	0.9706961743772242	0.1125340014696121 2	3.5679
8	0.9727463190910541	0.9733374110320284	0.9729854307448372	0.9733374110320284	0.1204564943096451	3.5799
10	0.9729097687048657	0.9735042259786477	0.9731459818475224	0.9735042259786477	0.1394817233085632 3	3.5832

数据集 B

Epoch	Eval Precision	Eval Recall	Eval F1	Eval Accuracy	Eval Loss	Eval Runtime
2	0.9612256192641184	0.9679350907029478	0.9630154282453001	0.9679350907029478	0.0801319628953933 7	7.4949
4	0.9612151459441082	0.9696534863945578	0.9622368460374825	0.9696534863945578	0.0794070512056350 7	7.7477

6	0.9617408635243314	0.9676959325396826	0.9637286930779716	0.9676959325396826	0.09067484736442566	7.796
8	0.9619194289796517	0.9665089994331065	0.9638604528771143	0.9665089994331065	0.11791561543941498	7.7083
10	0.9618047987592484	0.9661369756235828	0.9635584547151035	0.9661369756235828	0.13009293377399445	7.9148

AnchiBERT

训练最终结果

数据集	Eval Precision	Eval Recall	Eval F1	Eval Accuracy	Eval Loss	Eval Runtime
数据集 A	0.963940242927 4961	0.9646167 813455657	0.9642436 22737465	0.9646167813455657	0.1848205178976059	3.9459
数据集 B	0.961406071894 0829	0.9663495 606575964	0.9634749 39687441	0.9663495606575964	0.16236752 271652222	7.9751

数据集	Micro-Average Precision	Micro-Average Recall	Micro-Average F1	Macro-Average Precision	Macro-Average Recall	Macro-Average F1
数据集A	0.9646167 813455657	0.9646167 813455657	0.9646167 813455657	0.8273717 706255621	0.8038283103729272	0.8153108407512697
数据集B	0.9911068 594104309	0.9911068 594104309	0.9911068 594104309	0.2512589 632357055	0.24764968 296968584	0.23697334 205426301

模型介绍和分析

AnchiBERT 是针对古代汉语的 BERT 变体，提供了对古汉语语料的更深入的理解。它优化了 BERT 架构以适应古汉语的特殊需求，适合古文翻译和解析。

训练过程分析

数据集 A

Epoch	Eval Precision	Eval Recall	Eval F1	Eval Accuracy	Eval Loss	Eval Runtime
2	0.9591849504729812	0.9614657473309609	0.9598301822708686	0.9614657473309609	0.1168164536356926	3.0897
4	0.9639877151221328	0.9652746886120996	0.9642676070302236	0.9652746886120996	0.1109251528 9783478	3.134
6	0.96665166	0.96719306	0.96687019	0.96719306	0.120304748	3.14

	36368039	04982206	72485794	04982206	41594696	27
8	0.96622835 21442985	0.96752669 03914591	0.96662462 75754931	0.96752669 03914591	0.133583903 3126831	3.06 76
10	0.96876239 78945292	0.96950066 72597865	0.96908259 48759366	0.96950066 72597865	0.154163137 07828522	3.05 11

数据集 B

Epoch	Eval Precision	Eval Recall	Eval F1	Eval Accuracy	Eval Loss	Eval Runtime
2	0.96407831 97007398	0.96525120 46485261	0.96390404 02066386	0.96525120 46485261	0.082221254 70638275	9.69 45
4	0.96121962 33374363	0.96908659 29705216	0.96293416 77537446	0.96908659 29705216	0.081010073 42338562	9.34 74
6	0.96147254 17305747	0.96737705 49886621	0.96339996 4029979	0.96737705 49886621	0.092464007 43722916	7.18 56
8	0.96124318 50099584	0.96758964 00226758	0.96360780 31217096	0.96758964 00226758	0.134863793 84994507	6.72 07
10	0.96141629 98781131	0.96671272 67573696	0.96350522 48092323	0.96671272 67573696	0.137260258 19778442	6.67 15

guwenbert-base

训练最终结果

数据集	Eval Precision	Eval Recall	Eval F1	Eval Accuracy	Eval Loss	Eval Runtime
数据集A	0.898053521 4526109	0.916260512 2324159	0.900511550 865519	0.916260512 2324159	0.304275244 474411	4.35 13
数据集B	0.950169384 2044238	0.961734693 8775511	0.954995042 3377727	0.961734693 8775511	0.138649314 64195251	8.13 59

数据集	Micro-Average Precision	Micro-Average Recall	Micro-Average F1	Macro-Average Precision	Macro-Average Recall	Macro-Average F1
数据集A	0.9162605 122324159	0.9162605 122324159	0.9162605 122324159	0.65320906 79952062	0.3968061 30054585	0.46904291 833949097
数	0.9959254	0.9959254	0.9959254	0.40479139	0.4110982	0.39143839

据集B	535147393	535147393	535147393	253470026	524602927	70521158

模型介绍和分析

guwenbert-base 是一个专门为古文设计的 BERT 模型。它利用大量古文语料进行训练，目标是提高古文文本的理解和处理能力。

训练过程分析

数据集 A

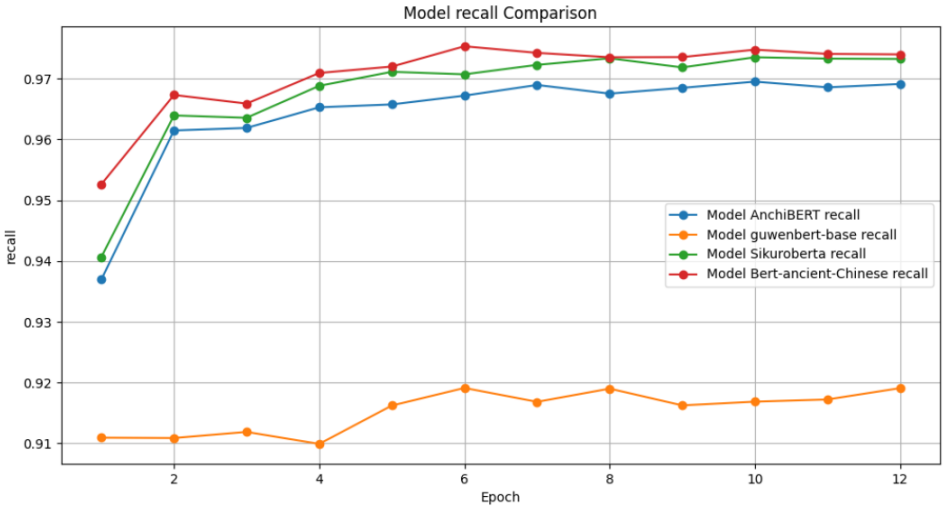
Epoch	Eval Precision	Eval Recall	Eval F1	Eval Accuracy	Eval Loss	Eval Runtime
2	0.889935668553033	0.9108930160142349	0.887439981938873	0.9108930160142349	0.3099377155303955	3.5996
4	0.893056266472675 9	0.9099199288256228	0.894712446268930 4	0.9099199288256228	0.3000724613666534 4	3.5699
6	0.907082207988853 5	0.9191225533807 83	0.895897794483824 3	0.9191225533807 83	0.2681475579738617	3.5973
8	0.902608138486159 8	0.9190113434163701	0.901291371786246 2	0.9190113434163701	0.2755137383937835 7	3.5942
10	0.899855753136015 2	0.9168705516014235	0.902567392103637 5	0.9168705516014235	0.2747617959976196 3	3.6135

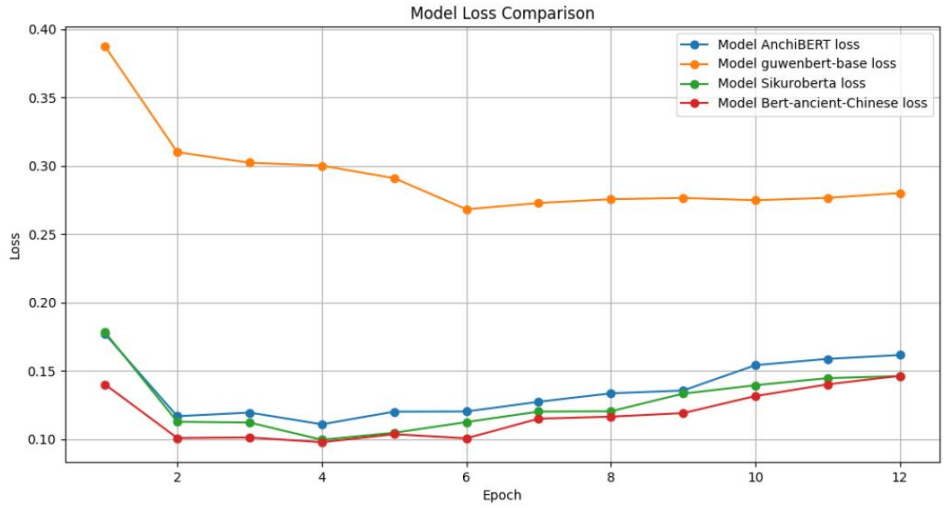
数据集 B

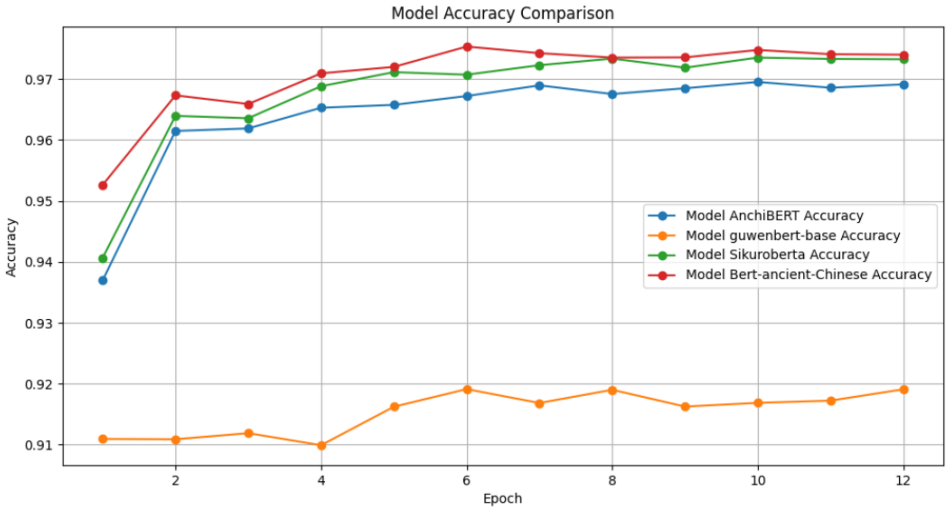
Epoch	Eval Precision	Eval Recall	Eval F1	Eval Accuracy	Eval Loss	Eval Runtime
2	0.945520060764842 2	0.9664912840136054	0.950324796561876 9	0.9664912840136054	0.1405127197504043 6	7.7432
4	0.946038592908899 8	0.9663407029478458	0.951331111574968 2	0.9663407029478458	0.1322239190340042	7.8783
6	0.950168255454545 35	0.9662078373015873	0.952695477062732 6	0.9662078373015873	0.1294545829296112	7.746
8	0.948499906885771	0.9653309240362812	0.953915135909169	0.9653309240362812	0.1329911351203918 5	7.5674
10	0.948868208728226 2	0.9633645124716553	0.954449811703410 5	0.9633645124716553	0.1366061121225357	7.5882

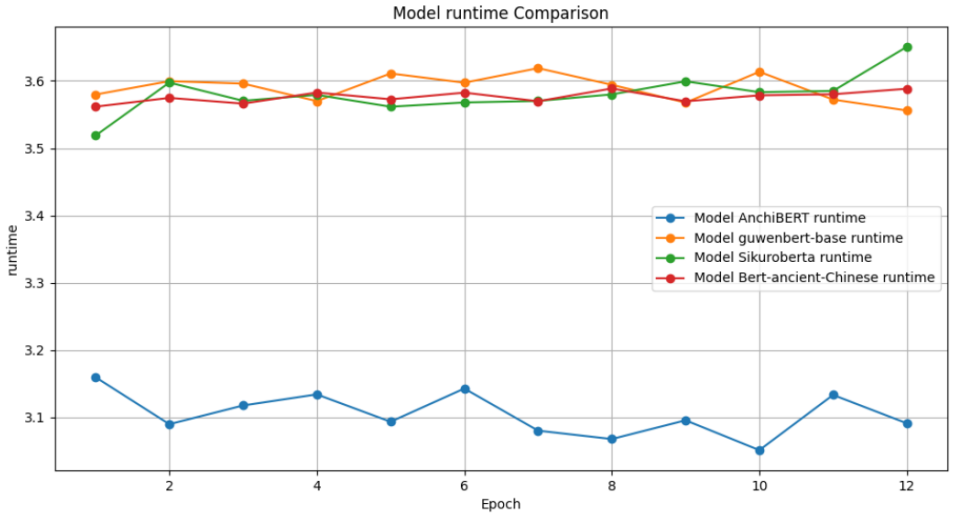
比较分析

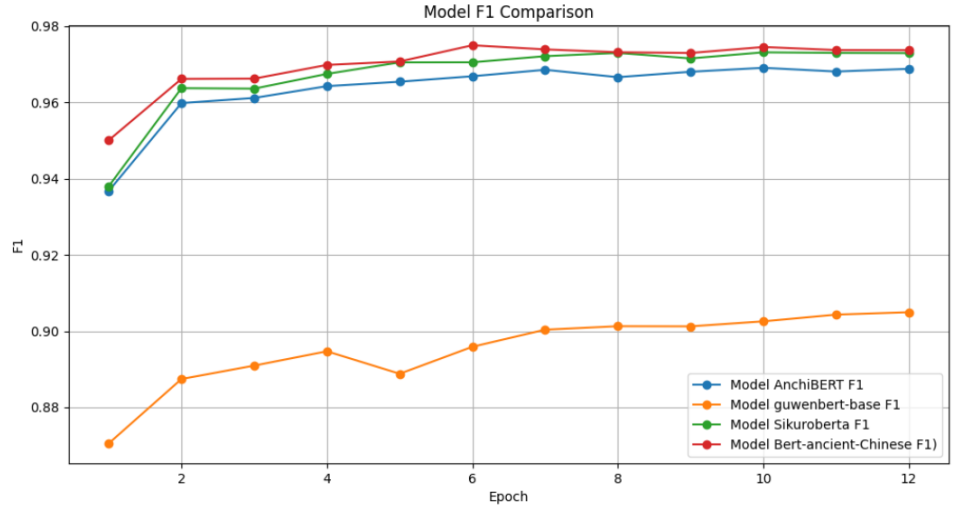
数据集 A

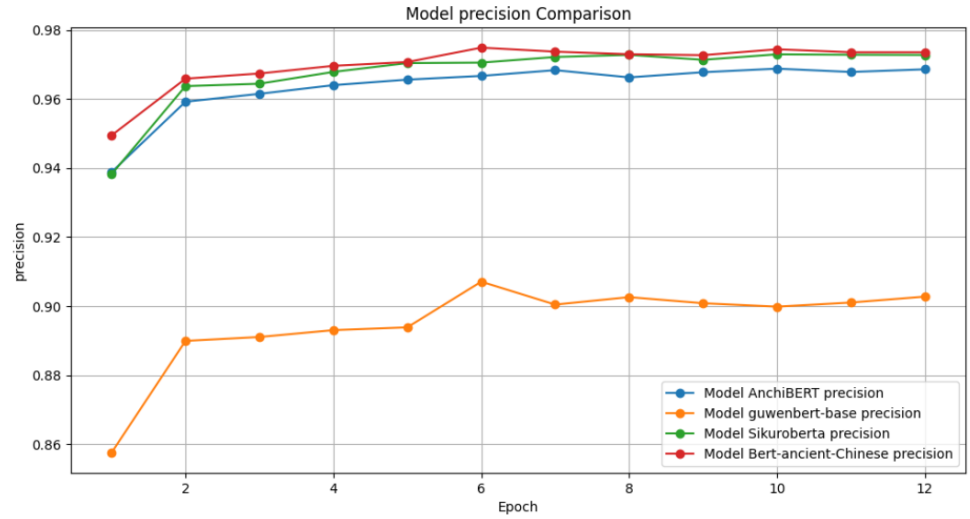

Model recall Comparison
recall
Epoch
Model AnchiBERT recall
Model guwenbert-base recall
Model Sikuroberta recall
Model Bert-ancient-Chinese recall


Model Loss Comparison
Loss
Epoch
Model AnchiBERT loss
Model guwenbert-base loss
Model Sikuroberta loss
Model Bert-ancient-Chinese loss


Model Accuracy Comparison
Accuracy
Epoch
Model AnchiBERT Accuracy
Model guwenbert-base Accuracy
Model Sikuroberta Accuracy
Model Bert-ancient-Chinese Accuracy


Model runtime Comparison
runtime
Epoch
Model AnchiBERT runtime
Model guwenbert-base runtime
Model Sikuroberta runtime
Model Bert-ancient-Chinese runtime


Model F1 Comparison
F1
Epoch
Model AnchiBERT F1
Model guwenbert-base F1
Model Sikuroberta F1
Model Bert-ancient-Chinese F1)


Model precision Comparison
precision
Epoch
Model AnchiBERT precision
Model guwenbert-base precision
Model Sikuroberta precision
Model Bert-ancient-Chinese precision

guwenbert-base

guwenbert-base 在各方面表现较差。

Bert-ancient-Chinese

Bert-ancient-Chinese 在整个训练中效果最好。

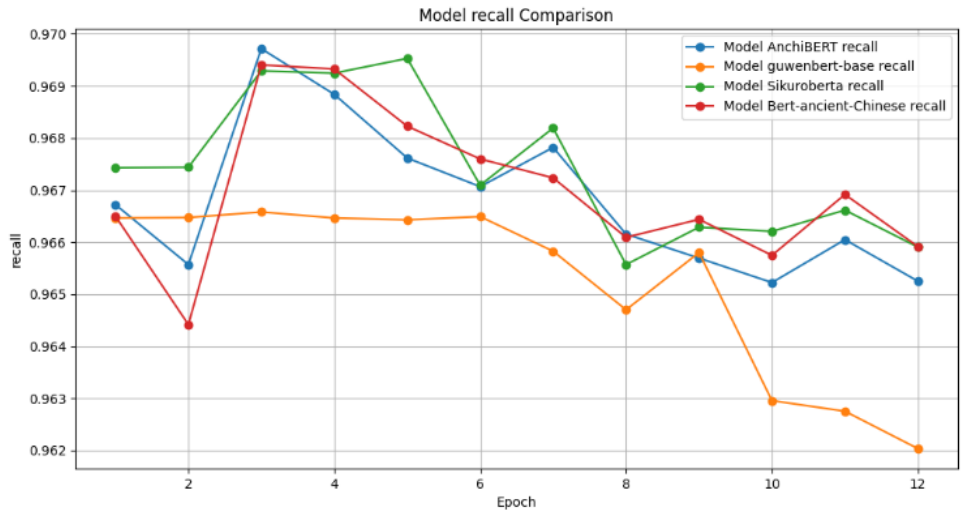
AnchiBERT

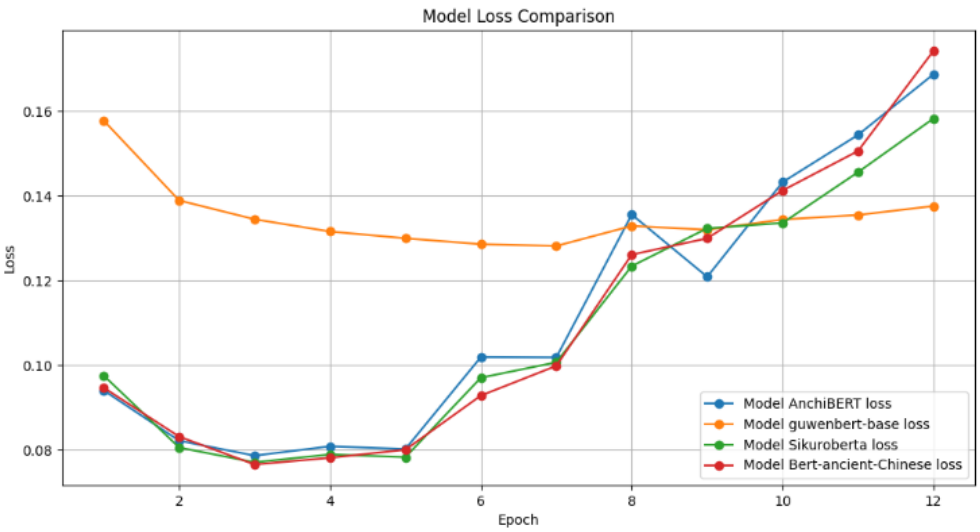
AnchiBERT 在整个训练中不如 SikuRoberta 和 Bert-ancient-Chinese，但是时间最短。

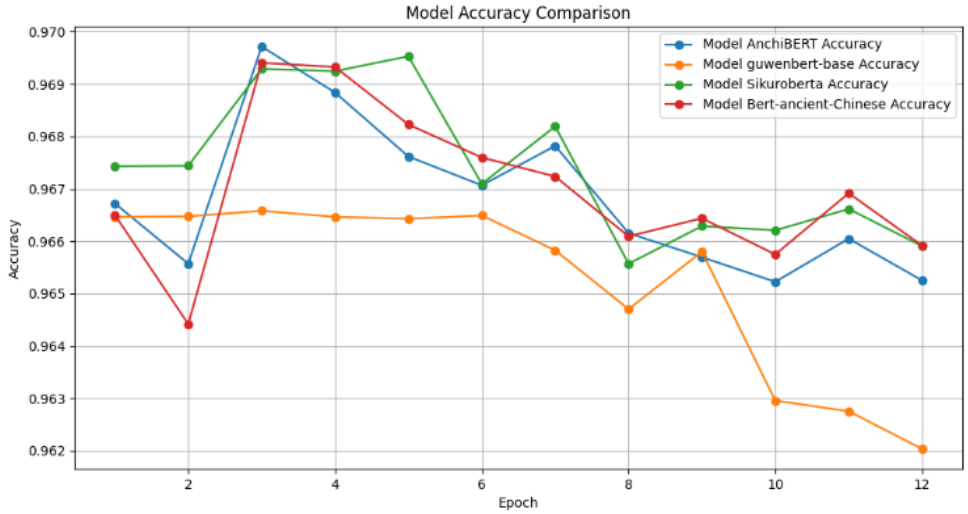
SikuRoberta

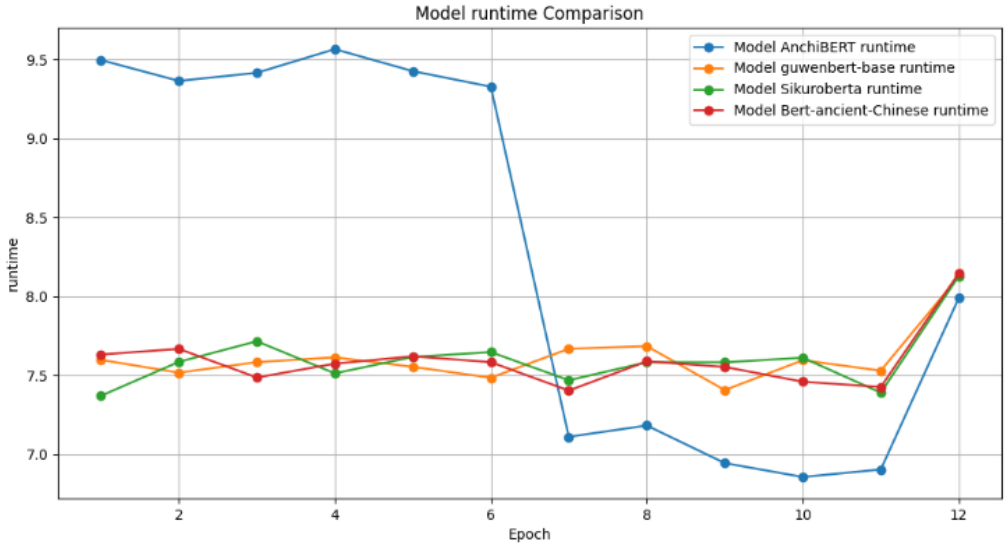
SikuRoberta 各项指标稍逊于 Bert-ancient-Chinese，训练效果相对优秀。

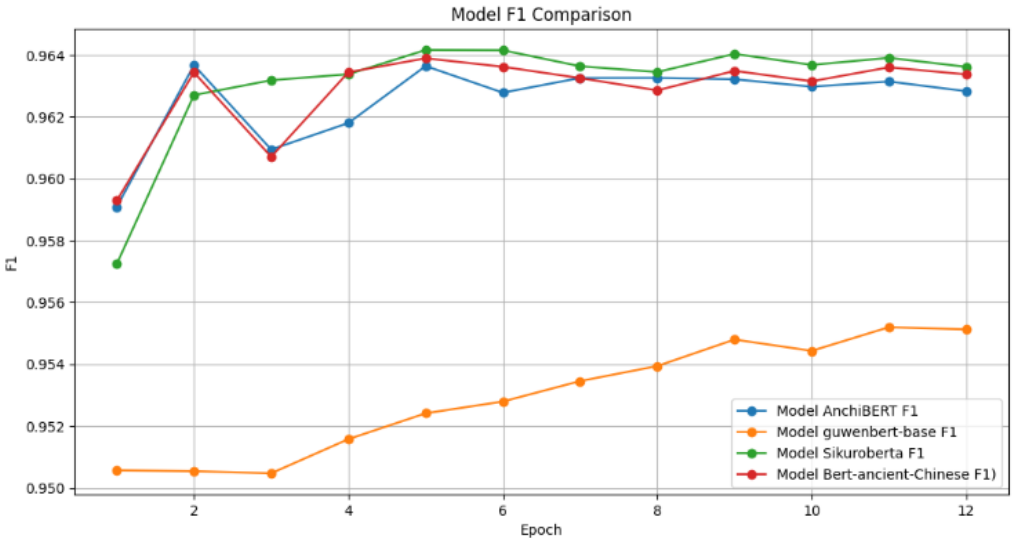
数据集 B

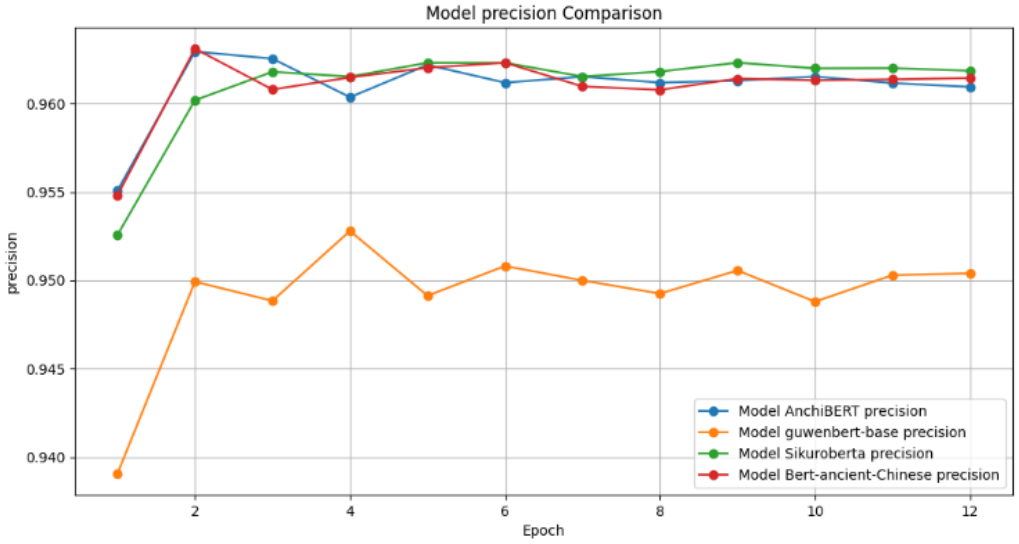





Model Accuracy Comparison
Model AnchiBERT Accuracy
Model guwenbert-base Accuracy
Model Sikuroberta Accuracy
Model Bert-ancient-Chinese Accuracy
Accuracy
Epoch


Model runtime Comparison
Model AnchiBERT runtime
Model guwenbert-base runtime
Model Sikuroberta runtime
Model Bert-ancient-Chinese runtime
runtime
Epoch


Model F1 Comparison
Model AnchiBERT F1
Model guwenbert-base F1
Model Sikuroberta F1
Model Bert-ancient-Chinese F1)
F1
Epoch



guwenbert-base

guwenbert-base 在各方面表现不如其他三个模型。

Bert-ancient-Chinese

Bert-ancient-Chinese 在 precision 和 f1 的表现与 AnchiBERT 和 SikuRoberta 类似，在后续 epoch 中相较于 SikuRoberta 更加平稳，表现相对不错。

AnchiBERT

anchbert 在其他方面表现不错，但在 runtime 中不太稳定。

SikuRoberta

SikuRoberta 在 precision 和 f1 的表现与 AnchiBERT 和 SikuRoberta 类似，表现相对不错。

综合比较

Bert-ancient-Chinese 和 SikuRoberta 表现都相对出色，Bert-ancient-Chinese 在后续 epoch 中更加平稳。

实验整体分析

数据集 A

模型	Micro-Average Precision	Micro-Average Recall	Micro-Average F1	Macro-Average Precision	Macro-Average Recall	Macro-Average F1
Bert-ancient-Chinese	0.971712538226299 7	0.971712538226299 7	0.971712538226299 7	0.867328927657945 3	0.847813052694254 4	0.857396230870138 4
SikuR	0.9683438	0.9683438	0.9683438	0.8362269	0.8354067	0.8356550

obert a	45565749 3	45565749 3	45565749 3	47526488 6	66296890 5	520943984
Anchi BERT	0.9646167 813455657	0.9646167 813455657	0.9646167 813455657	0.8273717 706255621	0.8038283 103729272	0.8153108 407512697
guwe nbert -base	0.9162605 122324159	0.9162605 122324159	0.9162605 122324159	0.6532090 679952062	0.3968061 30054585	0.4690429 183394909 7

数据集 B

模型	Micro-Average Precision	Micro-Average Recall	Micro-Average F1	Macro-Average Precision	Macro-Average Recall	Macro-Average F1
Bert-ancient-Chinese	0.9629373 504784688	0.9629373 504784688	0.9629373 504784688	0.4463562 233568977	0.3469447 379134225	0.3858195 247664353 3
SikuRoberta	0.9640774 5215311	0.9640774 5215311	0.9640774 5215311	0.6133806 357771348	0.3840230 414658615	0.4477773 397719211 3
Anchi BERT	0.9632177 033492823	0.9632177 033492823	0.9632177 033492823	0.4179623 603844156	0.3378474 057022941	0.3702823 300980698
guwe nbert -base	0.9600029 90430622	0.9600029 90430622	0.9600029 90430622	0.3253865 76678924	0.2004047 49296543	0.2373501 789184325

Micro-Average 指标反映了所有类别的总体性能，这意味着在考虑样本不平衡的情况下，Bert-ancient-Chinese 在数据集 A 中的各项指标都是最高的，说明在整体上取得了最好的平衡性能。 在数据集 B 中指标相差不大。

Macro-Average 指标展示了模型对所有类别同等对待的性能，Bert-ancient-Chinese 在数据集 A 中的各项指标都是最高的，表明其对于少数类别的识别能力较强。 在数据集 B 中指标相差不大。