邁向繁體中文 ModernBERT 的初步研究
Toward Traditional Chinese ModernBERT: A Preliminary Study

Yi-En Chen, Qiao-Ying He, Kuan-Yu Chen
Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology
yechen.research@gmail.com, qiaoyinghe.cs@gmail.com, kychen@mail.ntust.edu.tw

摘要

本研究使用多項最新技術，包含 RoPE、Flash Attention 等，並在結合大規模中文網路語料與百科全書資料的基礎上，預訓練一個專爲長文本設計的繁體中文編碼器模型。我們在閱讀理解、文本分類等任務上進行評估，結果顯示模型效能整體落後於既有中文基準。透過 pseudo-perplexity 分析，我們推測預訓練階段未能充分學習資料分布，並討論了超參數、收斂與資料品質等可能影響因素。雖然結果不理想，本研究仍提供中文語言模型發展的實驗經驗與改進方向。

Abstract

This study employs several state-of-the-art techniques, including RoPE and Flash Attention, and leverages large-scale Chinese web corpora and encyclopedic data to pre-train an encoder model specifically designed for long text in Traditional Chinese. We evaluate the model on tasks such as reading comprehension and text classification, and the results show that its overall performance lags behind existing Chinese benchmarks. Through pseudo-perplexity analysis, we infer that the pre-training phase did not sufficiently capture the data distribution, potentially due to factors such as hyperparameters, convergence, and data quality. Although the results are suboptimal, this study still offers valuable experimental insights and directions for improving Chinese language model development.

關鍵字： Transformer、繁體中文、長文本、預訓練語言模型

Keywords: Transformer, Traditional Chinese, Long Context, Pretrained Language Model

1 前言 Introduction

自 BERT (Devlin et al., 2019) 問世以來，僅編碼器（encoder-only）的 Transformer (Vaswani et al., 2017) 模型已成爲眾多自然語言處理（NLP）任務的核心架構。儘管近年來主流的大型語言模型（LLM）如 GPT (Radford et al., 2018, 2019)、Qwen (Bai et al., 2023; Yang et al., 2024) 等多採用僅解碼器（decoder-only）架構，但基於編碼器的模型在性能與計算效率的平衡上仍具備獨特優勢，並在文本分類、命名實體識別（NER）、資訊檢索（IR）等下游任務中獲得廣泛應用。

上下文長度對大型語言模型而言至關重要，擁有更長的上下文處理能力，使 LLM 能夠勝任更多樣化的任務。同樣地，對於 BERT、RoBERTa 等編碼器模型，上下文長度亦是關鍵因素。然而，這些基於 Transformer 的模型由於 self-attention 機制具有 $O(n^2)$ 的時間複雜度，限制了其處理長上下文的能力，導致可處理的上下文長度普遍較短。

雖然後續推出的編碼器模型在上下文長度方面有所改進，但主要集中於英文和簡體中文的優化，針對繁體中文的研究相對稀少。因此，本研究旨在填補此一研究缺口，爲繁體中文的長上下文編碼器模型發展貢獻力量。

2 研究方法 Methods

2.1 模型架構 Model Architecture

本研究使用了多項已經過廣泛測試的最新進展，以提升模型表現以及訓練效率:

- Rotary Positional Embedding (RoPE)：相較於傳統的絕對位置編碼，RoPE (Su et al., 2024) 能更有效捕捉長距離依賴，提升模型處理長文本的能力。
- Pre-Normalization：採用 Pre-Normalization (Xiong et al., 2020) 結構，有助於穩定訓練過程，促進深層模型的收斂。

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 133–139

- Alternating Global and Local Attention：交替使用全局與局部注意力機制 (Beltagy et al., 2020)。在局部注意力下，token 僅能與 sliding window 內的其他 token 互動；全局注意力則可與整個序列互動。我們每三層採用一次全局注意力。局部注意力的 RoPE theta 設爲 10,000，全局注意力的 RoPE theta 設爲 160,000。

- Unpadding：將 batch 內所有序列的 padding token 移除 (Zeng et al., 2022) 並串接成一個長序列計算，減少不必要的運算，提升訓練效率。

- Flash Attention：使用 Flash Attention (Dao et al., 2022) 以降低記憶體使用量，提升訓練速度。

2.2 中文分詞器 Tokenizer

本研究使用 benchang1110/Qwen2.5-Taiwan-1.5B-Instruct [1] 所提供的分詞器做爲基礎，此分詞器基於 Qwen2.5 (Qwen Team, 2024) 模型，並透過 tokenizer swapping，將簡體中文的 token 替換爲相對應的繁體中文 token，以優化對繁體中文的處理能力。

且爲了符合 BERT 模型的相容性，我們在此額外加入了 `[PAD]`、`[UNK]`、`[SEP]`、`[CLS]`、`[MASK]` 等特殊 token，並將模型配置中的 `pad_token` 明確指定爲 `[PAD]`，以取代原有的 `<|endoftext|>`，使模型的可以沿用以前 BERT 的預訓練任務。

2.3 訓練資料 Training Data

本研究的預訓練資料主要由以下兩部分構成:

- **FineWeb2**: FineWeb2 [2] 是一個大規模、開放且多語言的資料集，專門爲訓練高品質的大型語言模型 (LLM) 而設計。其資料主要源自於對 Common Crawl 存檔的大規模處理與精煉。該資料集涵蓋了從 2013 年到 2024 年 4 月的 96 個 Common Crawl 數據快照，並透過一個包含過濾、去重和語言辨識的複雜流程進行處理，旨在爲開發能夠理解和生成多種語言文字的大型語言模型提供一個強大且可擴展的高品質資源。我們使用的是 FineWeb2 的中文部分 (cmn_Hani) 的一個子集，約包含 4 千萬個樣本，其中繁體中文的樣本比例較高。

- 中文維基百科資料: 爲了增強模型在中文領域的知識與理解能力，我們額外納入了完整的中文維基百科資料 [3]。此資料集包含了百科全書中的所有條目，內容涵蓋歷史、文化、科技、地理等多個領域，具有結構化、高品質、事實性強的特點。使用此資料有助於模型學習中文世界豐富的背景知識和正規的書面語表達方式。其中約包含 140 萬個樣本。

2.4 資料預處理 Data Preprocessing

爲了確保訓練資料的品質與針對性，我們對原始資料進行了以下預處理步驟：

- 內容過濾：我們移除了訓練資料中的非必要章節，包括但不限於參考資料、外部連結、延伸閱讀等章節。這些章節通常包含大量的 URL 連結、引用格式等非自然語言內容，對模型學習語言理解能力的幫助有限，反而可能引入雜訊。

- 地區與語言變體篩選：考慮到中文存在不同的地區變體（如台灣、中國大陸、香港等），且不同地區在用詞、表達習慣上存在差異，我們在資料篩選時以保留臺灣正體的部分爲主，並賦予其較高的優先度。

經過上述預處理後，最終用於預訓練的資料集共包含約 35.7B 個 tokens，在保持多樣性的同時，也具備了更高的針對性與品質。

2.5 預訓練任務 Pretraining Task

本研究採用遮罩語言模型 **(Masked Language Modeling, MLM)** 任務，並結合 n-gram masking 策略，在中文語料上進行模型預訓練。整體訓練過程分爲以下兩個的階段：

1. 第一階段 (短文本學習): 模型以最大序列長度 1024 的文本進行訓練。此階段的核心目標是讓模型掌握中文的基礎語法結構與核心語義知識。

2. 第二階段 (長文本建模): 在繼承第一階段學習到的權重基礎上，我們將最大序列長度擴展至 8192 進行接續訓練。此階段旨在顯著提升模型對長距離文本依賴關係的捕捉與建模能力。

詳細的超參數設置如表 1所示，模型訓練使用 8 張 NVIDIA H100 GPU，總訓練時間約爲 7 天。

[1] https://huggingface.co/benchang1110/Qwen2.5-Taiwan-1.5B-Instruct

[2] https://huggingface.co/datasets/HuggingFaceFW/fineweb-2

[3] https://dumps.wikimedia.org/zhwiki/20250520/

	Phase 1	Phase 2
Max Sequence Length	1024	8192
Batch Size	512	512
Training Steps	233000	40000
Learning Rate	8e-4	3e-4
LR Schedule	WSD	WSD
Weight Decay	1e-5	1e-5
Optimizer	AdamW	AdamW
Betas	(0.9, 0.999)	(0.9, 0.999)
Epsilon	1e-8	1e-8

表 1. 訓練參數設定

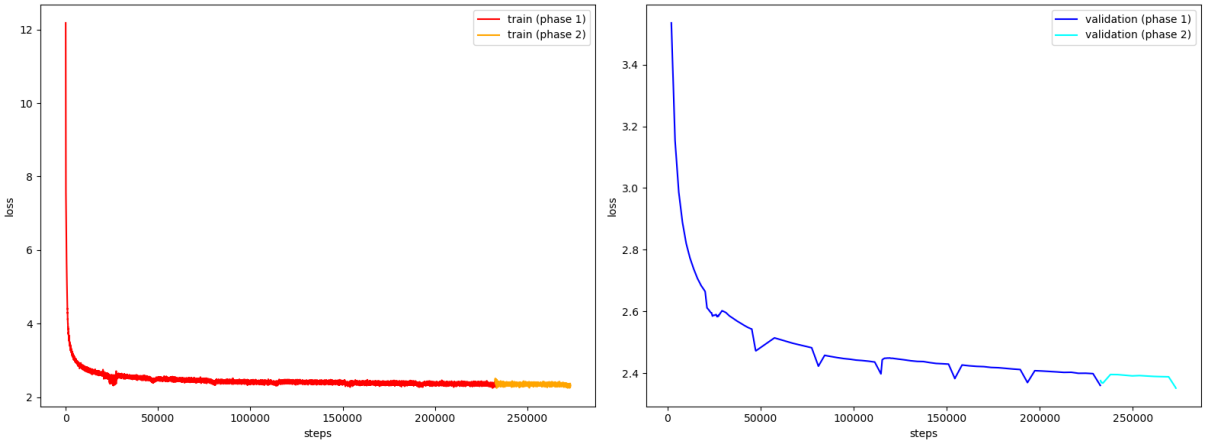

圖 1: 預訓練過程中的損失函數變化。左圖爲訓練損失，右圖爲驗證損失。

3 評估 evaluation

我們將模型在多個下游任務上進行實驗，以驗證其有效性，並將結果與其他研究論文中報告的中文模型結果進行比較 (Cui et al., 2021) 。實驗任務包括：

- 閱讀理解任務：CMRC2018 (Cui et al., 2019)、DRCD (Shao et al., 2019)
- 單句分類任務：ChnSentiCorp (Tan and Zhang, 2008)、THUCNews (Li and Sun, 2007)、TNEWS (Xu et al., 2020)
- 句子對分類任務：XNLI (Conneau et al., 2018)、LCQMC (Liu et al., 2018)、BQ (Chen et al., 2018)、OCNLI (Hu et al., 2020)

由於訓練語料中繁體中文比例較高，所有資料集在實驗前皆使用 OpenCC [4] 進行簡繁轉換。

[4] https://github.com/BYVoid/OpenCC

3.1 閱讀理解任務 Machine Reading Comprehension

閱讀理解任務會提供模型一組 (文本, 問題) 的文字對，模型需根據文本內容生成問題的答案，答案通常爲文本中的一段文字 (span)。我們使用 DRCD 與 CMRC2018 兩個中文閱讀理解資料集進行評估。結果展示在表 2. 中，其中 EM (Exact Match) 表示完全匹配率，F1 Score 表示 F1 分數。從結果可見，本研究模型在 DRCD 資料集上的表現與早期 BERT 模型 (如 BERT-base) 相當，但在 CMRC2018 資料集上則有顯著的效能落差，尤其在測試集 (Test set) 的表現遠不如其他基準模型，這可能暗示模型在特定領域的泛化能力較弱。

3.2 單句分類任務 Single Sentence Classification

單句分類任務會提供模型一段句子，要求模型根據該句子進行分類。我們在三個中文資料集上進行驗證：

- ChnSentiCorp：中文情感分析資料集，包含正向 (positive) 與負向 (negative) 兩個類別。

	CMRC2018				DRCD			
	Dev		Test		Dev		Test	
Model	EM	F1	EM	F1	EM	F1	EM	F1
BERT-base	65.5	84.5	70.0	87.0	83.1	89.9	82.2	89.2
BERT-wwm	66.3	85.6	70.5	87.4	84.3	90.5	82.8	89.7
BERT-wwm-ext	67.1	85.7	71.4	87.7	85.0	91.2	83.6	90.4
RoBERTa-wwm-ext	67.4	87.2	72.6	89.4	86.6	92.5	85.6	92.0
ELECTRA-base	68.4	84.8	73.1	87.1	87.5	92.5	86.9	91.8
MacBERT-base	68.5	87.9	73.2	89.5	89.4	94.3	89.5	93.8
Ours	55.2	81.3	32.2	68.1	83.4	92.5	82.6	91.9

表 2. 各模型於 CMRC2018 及 DRCD 資料集之閱讀理解任務表現，(單位：%)。

- THUCNews：新聞分類資料集，本研究使用其子集，任務爲判斷輸入新聞文本的主題類別。
- TNEWS：中文短新聞分類資料集，包含 15 個類別，任務爲對輸入文本進行主題分類。

表 3. 展示了各模型在單句分類任務上的準確率 (Accuracy) 表現。在 ChnSentiCorp 與 TNEWS 任務上，本模型表現與基準模型相近；然而，在 THUCNews 新聞分類任務上，效能則有明顯落後，顯示模型在處理長文本或特定主題分類時可能存在不足。

3.3 句子對分類任務 Sentence Pair Classification

句子對分類任務會提供模型一組 (文本, 文本) 的文字對，要求模型根據該句子進行分類。我們在四個中文資料集上進行驗證：

- XNLI: 跨語言自然語言推理資料集，本研究使用中文部分，任務爲判斷句子對是否具備「蘊涵、矛盾或中立」關係（textual entailment）。
- LCQMC: 大規模中文問答語料庫，用於評估中文問答對語意相似度的自然語言處理資料集。
- BQ Corpus: 中文語義等價判斷（Sentence Semantic Equivalence Identification, SSEI）資料集，用於識別句子對是否語義相同。
- OCNLI: 原生中文自然語言推理資料集，爲首個非翻譯、專爲中文語境設計的大型 NLI 資料集。

表 4 展示了各模型在句子對分類任務上的準確率 (Accuracy) 表現。相較於其他任務，本模型在句子對分類的整體表現較不理想。雖然在 XNLI 與 LCQMC 任務上僅略遜於基準模型，但在 BQ Corpus 與 OCNLI 資料集上出現了更大幅度的效能下降，尤其是在 OCNLI 這個專爲中文原生語境設計的資料集上，表現最不理想，這可能反映出模型在捕捉語意相似度與自然語言推論的細微差異方面仍有待加強。

4 討論 Discussion

雖然模型在多個下游任務微調後，表現仍顯著落後於現有基準，我們推測主要原因來自預訓練階段未能充分學習語料的潛在分佈。

爲了量化模型對訓練語料的擬合程度，我們引入了僞困惑度（Pseudo-Perplexity, PPPL）(Salazar et al., 2020) 作爲評估指標。對於自回歸語言模型，較低的困惑度（PPL）通常意味著模型能更好地建模語料；而 PPPL 則可對遮罩語言模型（MLM）進行近似評估。對於一個語料集 $\mathcal{W}$，其 PPPL 的計算方式如下：

$$\mathrm{PPPL}(\mathcal{W}) := \exp\left(-\frac{1}{N}\sum_{\mathbf{W}\in\mathcal{W}}\mathrm{PLL}(\mathbf{W})\right)$$

其中 N 是語料集的總詞數，而 $\mathrm{PLL}(\mathbf{W})$ 是單一樣本 $\mathbf{W}$ 的僞對數似然率（Pseudo-Log-Likelihood），其定義爲：

$$\mathrm{PLL}(\mathbf{W}) := \sum_{t=1}^{|\mathbf{W}|}\log P_{\mathrm{MLM}}(W_t \mid \mathbf{W}_{\setminus t};\Theta) \quad (1)$$

此處模型需預測每個詞 w_t 在其上下文 $\mathbf{w}_{\setminus t}$ 中的機率。

我們從訓練語料中隨機採樣文本，並在本模型及幾個基準模型上計算 pseudo-perplexity。結果如表 5. 顯示，部分基準模型對這些文本表現出更低的 pseudo-perplexity 值，這表明本模型在預訓練階段可能未能充分捕捉資料分布。基於此，我們提出以下幾個可能原因：

	ChnSentiCorp		THUCNews		TNEWS
Model	Dev	Test	Dev	Test	Dev
BERT-base	94.7	95.0	97.7	97.8	56.3
BERT-wwm	95.1	95.4	98.0	97.8	56.5
BERT-wwm-ext	95.4	95.3	97.7	97.7	57.0
RoBERTa-wwm-ext	95.0	95.6	98.3	97.8	57.4
ELECTRA-base	93.8	94.5	98.1	97.8	56.1
MacBERT-base	95.2	95.6	98.2	97.7	57.4
Ours	94.4	95.0	93.0	93.8	56.2

表 3. 各模型於 ChnSentiCorp、THUCNews 及 TNEWS 資料集之單句分類任務表現，(單位：%)。

	XNLI		LCQMC		BQ Corpus		OCNLI
Model	Dev	Test	Dev	Test	Dev	Test	Dev
BERT-base	77.8	77.8	89.4	86.9	86.0	84.8	86.0
BERT-wwm	79.0	78.2	89.4	87.0	86.1	85.2	86.1
BERT-wwm-ext	79.4	78.7	89.6	87.1	86.4	85.3	86.4
RoBERTa-wwm-ext	80.0	78.8	89.0	86.4	86.0	85.0	86.0
ELECTRA-base	77.9	78.4	90.2	87.6	84.8	84.5	84.8
MacBERT-base	80.3	79.3	89.5	87.0	86.0	85.2	86.0
Ours	77.1	77.1	86.0	86.6	82.0	79.6	70.0

表 4. 各模型於 XNLI、LCQMC、BQ Corpus 及 OCNLI 資料集之句子對分類任務表現，(單位：%)。

Model	pppl
BERT-base	2.49
BERT-wwm	2.73
BERT-wwm-ext	3.48
MacBERT-base	13.39
Ours	5.60

表 5. 各模型於本模型訓練集之 pseudo-perplexity。

4.1 超參數設置不理想

雖然大部分超參數與原始論文保持一致，但由於訓練語料不同，原論文的超參數對中文語料可能不完全適用，可能導致模型未達最佳效果。

4.2 模型尚未完全收斂

若模型在建模能力上仍不理想，可能原因之一是模型尚未完全收斂。我們推測第一階段短文本訓練可能過早結束，模型尚未充分學習資料分布。

4.3 訓練資料品質

資料中可能存在雜訊或標註不均衡，這會影響模型學習到準確的語言分布，進而影響下游任務表現。

除了預訓練因素，下游任務的資料分布亦可能影響模型表現。如同 2.3節所述，訓練語料以繁體中文爲主，因此模型的語言分布更接近繁體中文。然而下游任務資料集多爲簡體中文，即便經過繁簡轉換，語法、詞彙偏好與用字習慣仍存在差異，進而導致本模型在這些任務上的表現不如其他基準模型。

5 結論 Conclusion

在這篇論文中，我們嘗試使用文語料，對現有的 state-of-the-art 語言模型架構進行預訓練，並在多個下游任務上進行評估。雖然最終的實驗結果尚未能超越既有的中文基準模型，但這些結果爲我們提供了寶貴的觀察。

我們的分析指出，模型效能不佳可能與語料分布、語言特性、預訓練設定以及詞彙表設計等因素有關。這些推論顯示，將現有方法直接應用於中文語料並非一件簡單的工作，仍需更多針對中文語言特性的調整與研究。

在未來的工作中，我們計畫針對語料來源與品質進行優化，並嘗試更合適的預訓練策略與模型設定。我們相信這些經驗將成爲我們後續研究的重要基礎，幫助我們在中文語言模型的探索上持續改進。

6 致謝 Acknowledgment

This work was supported by the National Science and Technology Council (NSTC) of Taiwan under Grants NSTC 112-2636-E-011-002, NSTC 112-2628-E-011-008-MY3, and NSTC 113-2640-B-002-005. Additional support was provided by the "Empower Vocational Education Research Center" at the National Taiwan University of Science and Technology (NTUST) through the Featured Areas Research Center Program, as part of the Higher Education Sprout Project funded by the Ministry of Education (MOE), Taiwan. The authors also thank the National Center for High-Performance Computing, National Applied Research Laboratories (NARLabs), Taiwan, for providing essential computational and storage resources.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and et al. 2023. Qwen technical report. Technical report, arXiv preprint arXiv:2309.16609.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

Jing Chen, Qingcai Chen, Xin Liu, Haijun Yang, Daohe Lu, and Buzhou Tang. 2018. The BQ corpus: A large-scale domain-specific Chinese corpus for sentence semantic equivalence identification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4946–4951, Brussels, Belgium. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504—3514.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019. A span-extraction dataset for Chinese machine reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5883–5889, Hong Kong, China. Association for Computational Linguistics.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: fast and memory-efficient exact attention with io-awareness. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Hai Hu, Kyle Richardson, Liang Xu, Lu Li, Sandra Kuebler, and Lawrence S. Moss. 2020. Ocnli: Original chinese natural language inference.

Jingyang Li and Maosong Sun. 2007. Scalable term selection for text categorization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 774–782, Prague, Czech Republic. Association for Computational Linguistics.

Xin Liu, Qingcai Chen, Chong Deng, Huajun Zeng, Jing Chen, Dongfang Li, and Buzhou Tang. 2018. LCQMC:a large-scale Chinese question matching corpus. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1952–1962, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. Technical report, OpenAI.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. OpenAI Blog.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Chih Chieh Shao, Trois Liu, Yuting Lai, Yiying Tseng, and Sam Tsai. 2019. Drcd: a chinese machine reading comprehension dataset.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomput.*, 568(C).

Songbo Tan and Jin Zhang. 2008. An empirical study of sentiment analysis for chinese documents. *Expert Syst. Appl.*, 34(4):2622—2629.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. 2020. On layer normalization in the transformer architecture. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online). International Committee on Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, and et al. 2024. Qwen2 technical report. Technical report, arXiv preprint arXiv:2407.10671.

Jinle Zeng, Min Li, Zhihua Wu, Jiaqi Liu, Yuang Liu, Dianhai Yu, and Yanjun Ma. 2022. Boosting distributed training performance of the unpadded bert model.