

Beyond Over-Simplification and Over-Personalization: Lexical Complexity Prediction for Leveled Reading

Huidan Xu and Ying Liu

Department of Chinese Language And Literature, Tsinghua University
xhd24@mails.tsinghua.edu.cn, yingliu@tsinghua.edu.cn

Abstract

Lexical Complexity Prediction (LCP) in real-world scenarios, particularly in language education, is often limited by the trade-off between overly simplified general models and unscalable personalized ones. To address this, we introduce LCP for Leveled Reading (LCP-LR), a task aiming to operationalize group-level personalization by identifying complex words relative to specific proficiency levels, thereby balancing pedagogical guidance and scalability. We construct the first large-scale benchmark dataset for this task, comprising 100,917 instance annotations across 1,401 leveled texts. Furthermore, we propose Dual-Prior Adversarial Learning (DPAL), a framework that incorporates word- and text-level difficulty priors alongside adversarial debiasing to mitigate the model's over-reliance on simplistic absolute difficulty shortcuts. Experimental results show that DPAL significantly outperforms both conventional pre-trained baselines and large language models, improving the Macro F_1 of the best baseline (XLNet) from 74.19% to 77.81%. Our evaluations also reveal that large language models underperform in zero- and few-shot settings without fine-tuning or Chain-of-Thought prompting. Overall, this work promotes the practical deployment of LCP and lays a crucial technical foundation for dynamic text adaptation within leveled reading.¹

1 Introduction

Reading is a cornerstone of language instruction and acquisition, and its efficiency is highly dependent on the alignment between text difficulty and reader proficiency. This has driven leveled reading, a pedagogical paradigm that categorizes reading materials by difficulty and matches readers with texts calibrated to their specific proficiency levels (Pondiscio and Mahnken, 2014). This approach

plays a key role in improving vocabulary acquisition, promoting general language comprehension, and mitigating cognitive load and reading anxiety (Albay, 2017; Alowais, 2021).

Lexical coverage, the proportion of words in a text known to the reader, is a critical dynamic index reflecting text-reader alignment. Empirical insights from language pedagogy and psycholinguistics demonstrate that readers must be familiar with at least 95% of the running words to achieve effective comprehension (Laufer, 1989; Nation, 2006; Laufer and Ravenhorst-Kalovski, 2010). Predicting complex words relative to specific proficiency levels and pinpointing targeted comprehension barriers emerge as a pivotal yet challenging task, essential for automated text adaptation and leveled material curation.

Identifying complex words is typically formulated as two distinct paradigms depending on downstream applications (Alfter, 2025): Complex Word Identification (CWI) for lexical simplification, and Lexical Complexity Prediction (LCP) for language learning. While CWI operates as a binary classification task to determine whether a token requires simplification (Shardlow, 2013), LCP quantifies vocabulary difficulty along a fine-grained spectrum to represent the comprehension challenges faced by a target audience (Shardlow et al., 2021; North et al., 2023). By providing finer-grained difficulty information, LCP has recently emerged as the dominant paradigm, with significant advancements in benchmark datasets (e.g., CompLex 2.0) and shared tasks such as SemEval-2021 and MLSP-2024@BEA (Shardlow et al., 2021, 2024a,b).

However, the practical deployment of LCP in real-world scenarios remains severely constrained. We argue that current research is hindered by two counterproductive trends: a "regressive" trend toward oversimplified static modeling, and an "aggressive" trend toward cost-prohibitive individual-centric personalization.

¹<https://github.com/xu97-varro/LCP-LR>

The Regressive Trend. To reduce engineering overhead, many approaches collapse LCP into a static, context-independent setting. A striking example is the MLSP-2024 training data set (n=300),² where a single token (the Catalan *decàleg*) appears in two contexts with nearly identical scores (0.65 vs. 0.625), while the remaining 299 words are strictly restricted to solitary contexts. This paradigm degrades LCP into a context-agnostic vocabulary lookup, masking the dynamic nature of lexical difficulty. In contrast, the SemEval-2021 CompLex 2.0 dataset demonstrates that nearly 40% of word types (2,194 / 5,617) span multiple contexts with variant complexity values.

The Aggressive Trend. Individual-centric approaches attempt to build bespoke adaptive models for each learner by gathering granular personal trait data (Bingel et al., 2018; Gooding and Tragut, 2022). Although aligned with the ultimate vision of tailored education, this individual-level personalization faces a double whammy: First, group-level consensus predominates over individual idiosyncrasies; learners with similar backgrounds (e.g., shared L1 and proficiency) exhibit high agreement in complexity judgments (Gooding and Kochmar, 2018; Gooding et al., 2021), rendering the marginal gains of individual modeling negligible. Second, collecting extensive individual user profiles incurs prohibitive engineering overhead and triggers critical ethical concerns regarding user privacy and data security, limiting its viable adoption in large-scale educational scenarios.

To promote the real-world utility of LCP while moving beyond these twin trends, we propose Lexical Complexity Prediction for Leveled Reading (LCP-LR) (Figure 1). This task formulates vocabulary difficulty as a level-conditioned function, estimating the relative complexity of individual words to identify tokens that pose comprehension barriers for a designated reader tier. Our main contributions are summarized as follows.

- **Promoting the real-world application of LCP.** We introduce LCP-LR, a level-conditioned relative complexity evaluation task that establishes a practical, scalable foundation for group-level adaptive reading instruction.
- **Constructing a high-density annotated dataset for LCP-LR.** Comprising 1,401 lev-

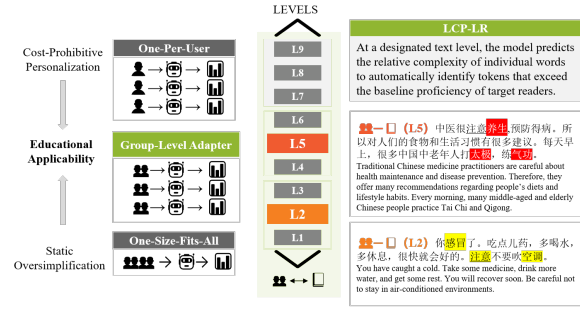


Figure 1: Illustration of the LCP-LR task. The model predicts level-conditioned relative lexical complexity for target readers. For instance, “注意” (pay attention to) is deemed familiar to L_5 learners yet flagged as a complex token exceeding the L_2 proficiency baseline.

eled texts and over 100,000 annotated tokens, this corpus is the largest fine-grained resource available to date for Chinese lexical complexity measurement.

- **Proposing the Dual-Prior Adversarial Learning (DPAL) framework.** By integrating word- and text-level difficulty priors with adversarial debiasing, DPAL effectively mitigates the model’s over-reliance on simplistic absolute difficulty shortcuts, achieving precise relative complexity estimation under target level constraints.
- **Conducting comprehensive evaluations across diverse discriminative PLMs and Large Language Models (LLMs).** Empirical results validate the effectiveness of our framework, and reveal its potential for contextual inference and cross-domain generalization, highlighting its practical value in real-world reading scenarios.

2 Related Work

Leveled Reading and Text Readability Assessment. Leveled reading matches texts with readers based on language proficiency, with text readability assessment serving as its core technology. Early research relied on surface features like word and sentence length to construct readability formulas like Flesch-Kincaid (Kincaid et al., 1975). Subsequently, machine learning approaches modeled readability evaluation as a classification task (Petersen and Ostendorf, 2009; Sung et al., 2015), expanding feature analysis from shallow metrics to deep semantic and discourse dimensions (Wu et al.,

²https://github.com/MLSP2024/MLSP_Data

2020; Cheng et al., 2020; Lee et al., 2021). Recently, pre-trained language models (PLMs) have further improved evaluation precision (Li and Wu, 2023; Tan et al., 2024), and large language models (LLMs) have been explored for this task as well (Wang et al., 2024; Paraschiv et al., 2025). Online platforms like Coh-Metrix,³ CRIE,⁴ and CTAP,⁵ provide technical support for leveled reading practices. In Chinese leveled reading, the *Chinese Graded Readers Standards for International Chinese Language Education* (2025) explicitly regulates the proportion of complex words across different Chinese (L2) reading levels for the first time. This regulatory shift imposes a more urgent and refined technical demand on lexical-granular complexity evaluation.

Lexical Complexity Prediction. Current LCP primarily serves as a preprocessing module for lexical simplification, identifying complex tokens that require replacement or explanation. The task has evolved from early binary classification (Paetzold and Specia, 2016) to continuous-scale regression (Shardlow et al., 2021), a shift that better reflects cognitive reality. By leveraging frequency, surface features, psycholinguistic attributes, and deep semantics, researchers have developed diverse systems across shared tasks like CWI 2018 (Yimam et al., 2018), SemEval-2021 (Shardlow et al., 2021), and MLSP-2024@BEA (Shardlow et al., 2024a). Additionally, several studies have explored personalized LCP to capture individual learner differences (Gooding and Tragut, 2022). Recent work (Kelious et al., 2025) shows that structured prompts and calibration improve LLM performance in LCP, though systematic biases remain. Despite these notable advances, the practical deployment of LCP in downstream applications remains severely limited, underscoring the urgent need for LCP research that addresses real-world requirements.

3 Task Definition

We introduce LCP-LR, a task that evaluates word difficulty relative to specific reader proficiency levels, enabling automated identification of complex words in graded texts for group-adaptive reading.

Formally, given a textual context containing a target word w along with its corresponding text level L (signifying that the text is curated for readers

at proficiency L), the model is required to determine whether w poses a comprehension barrier to these target readers, as well as the degree of this barrier. The core of this task lies in formulating lexical complexity as an interaction function between a word’s intrinsic attributes and the reader’s proficiency, thereby dynamically predicting the perceived difficulty for the target group.

Distinct from conventional LCP setups, our task prioritizes real-world pedagogical application. It focuses squarely on group-based adaptive scoping within leveled reading contexts, rather than relying on context-free absolute difficulty judgments or attempting to model highly subjective individual preferences.

4 The LCP-CLR Dataset

To support LCP-LR, we construct LCP-CLR, a densely annotated corpus of 1,401 graded Chinese texts across seven proficiency levels.

Large-scale assessment of reader knowledge outside laboratory settings remains challenging. Existing studies rely on proxy indicators such as specialized vocabularies (Elhadad and Sutaria, 2007), reading grade levels (Agrawal and Carpuat, 2023), or domains (Asthana et al., 2024). Building on this work, we introduce two critical resources from international Chinese language education for fine-grained characterization of reader vocabulary knowledge and reading ability: (I) a graded textbook corpus, and (II) a graded vocabulary syllabus (see Table 1).

Stage	Level	Texts	Syllabus Words
Elementary	1	141	500
	2	146	772
	3	139	973
Intermediate	4	133	1,000
	5	136	1,071
	6	120	1,140
Advanced	7 (7–9)	119	5,636
Total		934	11,092

Table 1: Distribution statistics of the graded textbook corpus and the vocabulary syllabus. For the vocabulary syllabus, Levels 7–9 are merged as they are undifferentiated in the official framework.

³<https://www.cometrix.com>

⁴<http://www.chinesereadability.net/CRIE/?LANG=CHT>

⁵<https://ctap.blcuicall.org>

4.1 Graded Textbook Corpus

The graded textbook corpus comprises over 900 texts from 50 International Chinese Language Education textbooks (e.g., *Learn Chinese with Me, Experiencing Chinese*), excluding classical poetry and songs. Text difficulty inherently aligns with target reader proficiency, providing leveled authentic data for complexity assessment.

For level annotation, we used the textbooks' original designated levels as the baseline and incorporated multi-source judgments from ACTR (Yang et al., 2025), LDU-TGP (Cheng et al., 2023), and the Hansi Chinese Leveled Reading System⁶ for calibration (see Appendix A.1). This multi-source reference strategy addresses two issues: inconsistent difficulty gradients across chapters and missing level labels in some textbooks. Final levels were manually verified by a senior instructor.

Ultimately, all texts were mapped into a seven-level taxonomy: Levels 1–3 correspond to the elementary stage, Levels 4–6 to the intermediate stage, and Level 7 to the advanced stage, aligning with the *Chinese Proficiency Grading Standards for International Chinese Language Education*.

4.2 Graded Vocabulary Syllabus

The vocabulary syllabus of the *Chinese Proficiency Grading Standards for International Chinese Language Education* explicitly defines the vocabulary sets that Chinese L2 learners are expected to master at each stage, serving as an authoritative baseline for lexical complexity. The syllabus comprises 11,092 vocabulary items, categorized into three stages and nine levels (with Levels 7–9 remaining undifferentiated). However, the syllabus does not systematically distinguish between different senses of the same word, annotating only 126 homographs, of which 89 span multiple proficiency levels. For these cross-level homographs, we determined their target levels during annotation by jointly analyzing part-of-speech and contextual usage (see Appendix A.2 for further details).

4.3 Lexical Complexity Annotation

The textbook corpus contains 7,048 syllabus words and 5,562 out-of-vocabulary (OOV) words.⁷ For the OOV words, five graduate students majoring in linguistics performed knowledge-intensive manual annotation based on the three-stage, nine-level

⁶<https://hansi.tracengine.com>

⁷<https://github.com/lancopku/pkuseg-python>

Annotated Text Examples from LCP-CLR

你_E感冒_C了_E。吃_E点_R药_R，多_E喝_E水_E，多_E休息_E，很_E快_E就_E会_E好_E的_E。注意_C不要_R吹_R空调_C。(L2)

Translation: You have caught a cold. Take some medicine, drink more water, and get some rest. You will recover soon. Be careful not to stay in air-conditioned environments.

中医_E很_E注意_E养生_D，预防_E得_E病_E。所以_E对_E人们_E的_E食物_E和_E生活_E习惯_E有_E很_E多_E建议_E。每_E天_E早上_E，很_E多_E中国_E中老年_E人_E打_E太极_D，练_E气功_D。(L5)

Translation: Traditional Chinese medicine practitioners are careful about health maintenance and disease prevention. Therefore, they offer many recommendations regarding people's diets and lifestyle habits. Every morning, many middle-aged and elderly Chinese people practice Tai Chi and Qigong.

Table 2: Examples from the LCP-CLR dataset. Subscripts represent four complexity levels colored by difficulty: Easy (E), Readable (R), Challenging (C), and Difficult (D).

criteria of the vocabulary syllabus. The annotators comprehensively evaluated word frequency, word-form information productivity, semantic complexity, and the level-analogy rules proposed by Zhang et al. (2019). The annotation process adopted a blind independent dual-scoring mechanism followed by a third-party arbitration phase; notably, 1,491 proper nouns were excluded from the prediction targets (see Appendix A.3 for further details).

Based on the mapping between text levels and lexical levels, we categorized lexical complexity into four distinct classes: **Easy** ($L_{\text{word}} < L_{\text{text}}$), **Readable** ($L_{\text{word}} = L_{\text{text}}$), **Challenging** ($L_{\text{word}} > L_{\text{text}}$ but within the same stage), and **Difficult** ($L_{\text{word}} > L_{\text{text}}$ and spanning across stages). All tokens were annotated accordingly.

Annotation statistics reveal a typical long-tail distribution: Easy dominates (87.6%), followed by Readable (7.6%), Difficult (3.8%), and Challenging (1.0%). This reflects the sparsity of complex words in natural texts, consistent with the "comprehensible input" principle. Despite comprising less than 5% combined, Challenging and Difficult words are the primary sources of comprehension barriers in actual reading.

4.4 Dataset Split and Statistics

The LCP-CLR dataset uses paragraphs as the basic units, encompassing 1,401 paragraphs,⁸ 9,265

⁸Paragraphs containing fewer than three sentences are excluded to prevent truncated paragraphs from deviating from

unique word types, and 100,917 tokens, as exemplified in Table 2. Notably, 19.4% (1,801) of the word types exhibit varying complexity labels across different contexts. To prevent data leakage, the experimental dataset is partitioned at the paragraph level. Table 3 summarizes the detailed statistics of the resulting training, validation, and test splits.

Our dataset is characterized by three prominent features: (1) **Cost-Effective Construction**: Leveraging existing authoritative resources enables highly efficient annotation, successfully bypassing the prohibitive costs of from-scratch manual labeling. (2) **Dense, Token-Level Coverage**: Unlike selective keyword tagging, our dataset encapsulates every token in the running text at scale, effectively alleviating the data bottleneck for relative lexical analysis. (3) **Real-World Application Alignment**: The annotated difficulty directly mirrors the cognitive load across diverse text proficiencies, precisely tailoring LCP systems to the pragmatic demands of leveled reading.

Metric	Training	Validation	Test
Paragraphs	980	210	211
Tokens	70,879	15,059	14,979
Easy	62,010	13,207	13,162
Readable	5,497	1,069	1,100
Challenging	723	159	167
Difficult	2,649	624	550

Table 3: Statistics of the dataset splits.

5 Experiments

Utilizing the LCP-CLR dataset, we evaluate the performance and enhancement strategies of multiple pre-trained models along with open- and closed-source LLMs on the LCP-LR task.

5.1 Models

Pre-trained Language Models. To systematically assess the impact of architecture and parameter scale on performance, we baseline ten pre-trained models: BERT-base (Devlin et al., 2019), RoBERTa-wwm-ext (Cui et al., 2020), MacBERT-base (Cui et al., 2021), Chinese-ELECTRA-180g-base (Cui et al., 2020), XLNet-base (Yang et al., 2019), RoBERTa-wwm-ext-large (Cui et al., 2020), MacBERT-large (Cui et al., 2021), Chinese-ELECTRA-180g-large (Cui et al., 2020), XLNet-

mid (Yang et al., 2019), and Chinese-ModernBERT (Zhao et al., 2025). All models use a unified hyperparameter set and a weighted cross-entropy loss to address class imbalance (see Appendix B.1).

Built upon the best-performing baseline, we propose **Dual-Prior Adversarial Learning (DPAL)** (Figure 2), a framework that integrates multi-task learning with adversarial debiasing to balance absolute difficulty priors and relative complexity prediction. The core idea is twofold. First, multi-task learning injects absolute difficulty priors from both the text and word levels; however, this may induce the model to rely on simplistic shortcuts (e.g., "higher level $\rightarrow$ higher difficulty"). Second, adversarial debiasing via gradient reversal layers (GRL) suppresses the shared encoder’s ability to encode absolute difficulty, thereby reducing its interference with relative difficulty judgments. To avoid direct conflict with the auxiliary tasks, the adversarial tasks target coarse-grained proficiency bands rather than exact levels. By driving the encoder to update toward maximizing band classification loss, the GRL effectively mitigates the influence of absolute difficulty information on the shared representation, enabling the model to preserve fine-grained level knowledge while avoiding over-reliance on level-based heuristics for difficulty reasoning.

The joint training objective is:

$$\begin{aligned} \mathcal{L} = & \mathcal{L}_{\text{main}} + \lambda_{\text{text}} \mathcal{L}_{\text{text}} + \lambda_{\text{word}} \mathcal{L}_{\text{word}} \\ & + \lambda_{\text{adv_text}} \mathcal{L}_{\text{adv_text}} + \lambda_{\text{adv_word}} \mathcal{L}_{\text{adv_word}} \end{aligned} \quad (1)$$

where $\mathcal{L}_{\text{main}}$ is the main LCP loss, $\mathcal{L}_{\text{text}}$ and $\mathcal{L}_{\text{word}}$ are auxiliary losses for text-level and word-level proficiency prediction (7 classes each), and $\mathcal{L}_{\text{adv_text}}$ and $\mathcal{L}_{\text{adv_word}}$ are adversarial losses for text-level and word-level proficiency band prediction (3 classes each).⁹ We adopt a dynamic adversarial coefficient α (increasing linearly from 0 to 0.1 over the first three epochs) to balance early training stability and late-stage debiasing.

Large Language Models. We evaluate four large language models (LLMs): two proprietary models, GPT-4o-mini (OpenAI, 2024) and Qwen3-Max (Bai et al., 2023); and two open-source models, Qwen2.5-7B-Instruct (Bai et al., 2023) and Llama3-Chinese-8B-Instruct (FlagAlpha, 2024). Prompting covers zero-shot, few-shot, and chain-of-thought (CoT) settings. The open-source models are instruction-tuned using LoRA (Hu et al., 2022).

⁹The loss weights are set to $\lambda_{\text{text}} = 0.1$, $\lambda_{\text{word}} = 0.1$, $\lambda_{\text{adv_text}} = 0.05$, and $\lambda_{\text{adv_word}} = 0.05$.

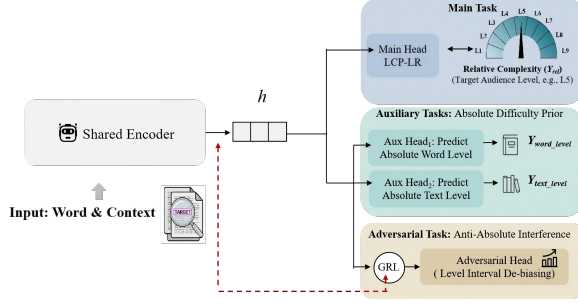


Figure 2: The DPAL architecture. The shared encoder h feeds into five components: a main head for LCP-LR, two auxiliary heads for word- and text-level proficiency prediction (grades 1–7), and two GRL-equipped adversarial heads for word- and text-level proficiency band prediction (bands: beginner, intermediate, advanced). For visual clarity, the two adversarial heads are consolidated in the figure; their individual loss terms are detailed in Eq. (1).

Appendix B.2 details experimental configurations and prompt templates.

Evaluation Metrics. We employ Accuracy (Acc.), Macro-F1 (M-F1), and Weighted-F1 (W-F1) as our primary evaluation metrics. We also report Precision (P), Recall (R), and F1-score (F1) for each of the four complexity categories.

5.2 Experimental Results of PLMs

Table 4 compares the performance of different PLMs. XLNet-base achieves the best overall performance, outperforming other models across all three primary metrics: Accuracy (94.72%), Macro-F1 (74.19%), and Weighted-F1 (94.65%). This suggests that the bidirectional autoregressive pre-training paradigm is better suited for this task. Regarding model scale, larger variants do not consistently outperform their base counterparts; for instance, MacBERT-large and XLNet-mid both underperform relative to their respective base models, indicating that the impact of parameter scaling varies with model architecture and pre-training objectives.

Per-category performance varies significantly, as shown in Figure 3. While the F1-scores for Easy generally exceed 97%, they drop to 59%–68% for Readable, 70%–78% for Difficult, and hit the lowest bound (40%–55%) for Challenging. This performance discrepancy is likely attributable to both data imbalance and the intrinsic separability of individual categories.

Table 5 presents the performance of DPAL using the best-performing baseline (XLNet-base) as

Model	Acc.	M-F1	W-F1
BERT-base	94.02	70.15	93.81
RoBERTa-wwm-ext	94.32	71.32	94.05
MacBERT-base	94.51	71.90	94.26
ELECTRA-base	93.92	69.65	93.69
XLNet-base	94.72	74.19	94.65
RoBERTa-wwm-ext-large	94.70	73.63	94.42
MacBERT-large	94.04	68.94	93.84
ELECTRA-large	94.43	71.80	94.16
XLNet-mid	94.64	74.04	94.46
Chinese-ModernBERT	94.34	69.49	94.05

Table 4: Performance comparison of different PLMs on the test set (%). Best results are highlighted in **bold**.

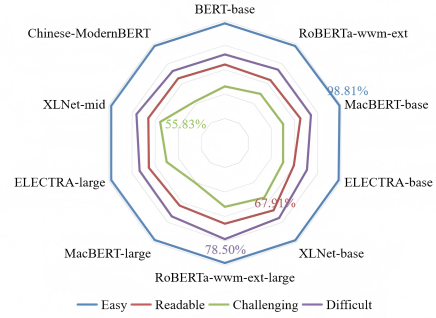


Figure 3: F1-score comparison of different PLMs across four complexity categories (%). Only the peak performance for each category is annotated.

its backbone. DPAL improves classification on Readable and Challenging while maintaining performance on Easy and Difficult. The gain is most pronounced on the highly imbalanced Challenging: compared to XLNet-base, recall increases by 13.77%, precision by 6.03%, and F1 by 10.59%. A slight precision drop (-1.23%) is observed for Difficult, suggesting a mild over-prediction tendency. Misclassifications occur mainly between adjacent difficulty levels, with a bias toward Readable (see Appendix C.1 for the confusion matrix).

We additionally evaluate JUST BLUE, the top-ranking ensemble baseline from SemEval-2021 Task 1 (Bani Yaseen et al., 2021). On our test set, JUST BLUE achieves 94.19% accuracy and 71.10% Macro-F1, outperforming most standard PLMs but still below DPAL. Notably, DPAL surpasses JUST BLUE by 17.60% in F1 on Challenging, validating the effectiveness of our framework.

Ablation Study. To scrutinize the individual contribution of each core component within DPAL, we conduct a systematic ablation study on the test set, as summarized in Table 6 (see Appendix C.2 for precision/recall breakdown by category). The primary insights are detailed below:

Model	Acc.	M-F1	W-F1	Easy			Readable			Challenging			Difficult		
				P	R	F1	P	R	F1	P	R	F1	P	R	F1
XLNet-base	94.72	74.19	94.65	97.97	98.40	98.18	67.91	67.91	67.91	62.60	49.10	55.03	77.13	74.18	75.63
JUST BLUE	94.19	71.10	93.98	97.05	98.57	97.81	71.44	60.73	65.65	48.77	47.31	48.02	75.49	70.55	72.93
DPAL (ours)	95.15	77.81	95.14	98.28	98.34	98.31	71.00	71.00	71.00	68.63	62.87	65.62	75.90	76.73	76.31
Δ (vs. XLNet)	+0.43	+3.62	+0.49	+0.31	-0.06	+0.13	+3.09	+3.09	+3.09	+6.03	+13.77	+10.59	-1.23	+2.55	+0.68

Table 5: Performance comparison among XLNet-base, JUST BLUE, and our DPAL framework on the test set (%). Best results are highlighted in **bold**. Δ denotes the absolute improvement of DPAL over XLNet-base.

Setting	Model Variant	Acc.	M-F1	W-F1	Easy-F1	Readable-F1	Challenging-F1	Difficult-F1
Dual Aux.	DPAL (Full)	95.15	77.81	95.14	98.31	71.00	65.62	76.31
	with Static α	95.29	77.45	95.22	98.47	70.86	64.94	75.52
	w/o Text Adv.	95.19	77.03	95.11	98.36	70.19	62.42	77.16
	w/o Word Adv.	95.00	75.95	94.95	98.35	69.14	60.98	75.31
	w/o Total Adv.	<u>93.98</u>	74.10	94.26	<u>97.77</u>	66.97	54.89	76.76
Single Aux.	Word Aux. + Adv.	94.79	74.27	94.73	98.26	67.60	<u>54.32</u>	76.90
	Text Aux. + Adv.	94.27	<u>73.73</u>	<u>94.18</u>	97.78	<u>66.07</u>	55.00	76.05
	Text Aux. Only	94.90	75.73	94.82	98.14	68.64	56.68	79.45
	Word Aux. Only	94.77	75.71	94.75	98.20	68.58	61.40	<u>74.66</u>

Table 6: Ablation study of our DPAL framework on the test set (%). Best results in each column are highlighted in **bold**, and the worst results are underlined.

Crucial Role of Adversarial De-biasing. Removing the full adversarial mechanism (*w/o Total Adv.*) drops Macro-F1 from 77.81% to 74.10% and yields the lowest score on Easy, demonstrating that absolute difficulty priors impair relative complexity generalization without adversarial constraints. Specifically, the performance drop from removing lexical de-biasing (*w/o Word Adv.*, $\downarrow 1.86\%$) is substantially harsher than that from removing textual de-biasing (*w/o Text Adv.*, $\downarrow 0.78\%$), highlighting that word-level de-biasing is more critical to our task. Furthermore, the full model equipped with a dynamic coefficient strategy outperforms its static counterpart (*with Static α*) by 0.36% in Macro-F1, validating the efficacy of progressive de-biasing schedules.

Complementarity and Synergy of Multi-Granular Priors. Word- and text-level difficulty priors exhibit distinct functional specializations: the lexical prior (*Word Aux. Only*) excels at Challenging (61.40% F1), whereas the textual prior (*Text Aux. Only*) is highly beneficial for Difficult (79.45% F1) but at the expense of severely compromising Challenging (dropping to a mere 56.68% F1). Moreover, enforcing adversarial constraints into single-source settings triggers over-correction due to informational scarcity, leading to a pervasive performance decay across single-auxiliary variants upon adding adversarial heads. This boundary phenomenon indicates that the true utility of adversar-

ial de-biasing requires a dual-source collaborative environment to be fully unlocked.

Global Optimization and Balance. The full DPAL framework achieves the best global balance with a peak Macro-F1 of 77.81%. Performance on Easy remains highly robust across all variants ($\approx 98\%$), confirming that the core challenge lies in distinguishing mid-to-high difficulty levels. By integrating multi-granular priors with dynamic adversarial debiasing, DPAL (Full) achieves the optimal trade-off on these bottleneck categories.

5.3 Experimental Results of LLMs

To ensure comparability and efficiency, all LLM experiments use stratified sampling to construct training/validation/test sets (20k/2k/2k). Table 7 reports the performance of four LLMs under three prompting settings, along with fine-tuned open-source models. Oversampling fine-tuning experiments are described in Appendix C.3.

Closed-source LLMs generally outperform open-source models. CoT prompting substantially boosts performance (e.g., Qwen3-Max Macro-F1 rises from 11.71% to 35.15%), indicating that structured reasoning is critical for fine-grained judgments. Fine-tuning improves open-source models, but both Qwen2.5-7B and Llama3-Chinese-8B achieve zero recall on the Challenging category, merely predicting majority classes. This reveals a fundamental limitation of data imbalance on stan-

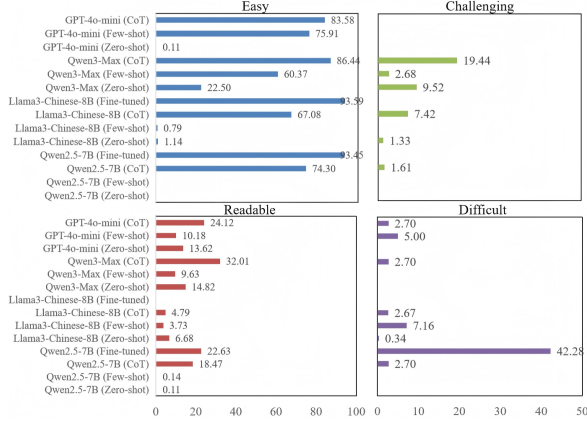


Figure 4: F1-score comparison of different LLMs across four complexity categories (%).

dard LLM fine-tuning. In contrast, DPAL achieves 94.95% accuracy and 79.21% Macro-F1 on the new 2k test set, substantially outperforming all LLMs and demonstrating the advantage of task-specific lightweight architectures.

As shown in Figure 4, most LLMs perform well on Easy, but drop sharply on Readable. Performance on Challenging and Difficult is even more severe. This trend highlights a fundamental limitation of current LLMs on fine-grained difficulty prediction: as category difficulty increases and sample sparsity intensifies, performance deteriorates sharply, posing severe challenges for intermediate and minority classes.

Model	Strategy	Acc.	M-F1	W-F1
DPAL (ours)	-	94.95	79.21	94.99
Qwen2.5-7B	Zero-shot	5.90	2.81	0.82
	Few-shot	7.30	3.41	1.09
	CoT	57.30	24.28	66.82
	Fine-tuned	84.00	39.59	85.39
Llama3-8B	Zero-shot	4.67	2.37	1.52
	Few-shot	4.05	2.92	1.23
	CoT	48.69	20.49	59.44
	Fine-tuned	87.95	23.40	82.31
Qwen3-Max	Zero-shot	18.30	11.71	20.98
	Few-shot	41.00	18.17	53.83
	CoT	74.35	35.15	78.63
GPT-4o-mini	Zero-shot	7.35	3.43	1.09
	Few-shot	58.43	22.77	67.66
	CoT	69.65	27.60	75.37

Table 7: Performance comparison of LLMs on the proposed dataset (%).

6 Discussion

We further investigate three aspects of our model’s behavior: failure modes, sensitivity to contextual informativeness, and generalization capabilities.

Error Analysis. To gain a deeper understanding of the model’s failure modes, we conducted two complementary error analyses.

Confusion matrix analysis. The results (Appendix C.1, Figure 5) show that 77.6% of errors occur between adjacent difficulty levels, with the most prominent confusions being Readable vs. Easy (18.2% of Readable instances misclassified as Easy) and Challenging vs. Readable (24.6% of Challenging instances misclassified as Readable). This indicates that the model effectively perceives difficulty as a continuous spectrum but struggles to identify sharp boundaries between adjacent levels. Furthermore, the errors exhibit a clear asymmetry: the model tends to underestimate high-difficulty words as lower-difficulty ones, suggesting a conservative decision strategy on boundary cases.

Linguistic feature analysis. We examined factors such as word frequency and lexical category. Results show that the average word frequency (based on the BCC corpus) of error samples is significantly lower than that of correctly classified samples, indicating that low-frequency words pose a major challenge. We further observed that a considerable proportion of errors involve idiomatic expressions. Using a Chinese idiom list to automatically annotate idioms in the test set, we found that the error rate for idioms is approximately 20.18%, significantly higher than the overall error rate of 4.85%. We attribute this to the semantic opacity of idioms, which makes it difficult for the model to infer their meaning from context alone, thereby leading to frequent misclassification.

The Modulating Role of Contextual Informativeness in Lexical Difficulty Perception. Contextual informativeness is a key modulator of lexical difficulty perception. For the Chinese word “敷衍” (to be perfunctory), a low-informativeness context (“所有人都看得出来，他在敷衍”; “Everyone could see that he was being perfunctory”) provides no cues for meaning inference, while a high-informativeness context with contrastive cues (“虽然时间紧、任务重，但他每个细节都认真对待，绝不敷衍”; “Despite the tight schedule, he treated every detail seriously and was never perfunctory”) enables inference. Unlike sentence-level modeling, our paragraph-level approach cap-

tures richer long-range contextual information. Our model predicts “敷衍” as Challenging in the low-informativeness context and Readable in the high-informativeness context, demonstrating its emerging capacity for sense inference based on contextual cues. Future work may explicitly incorporate contextual information density into the prediction framework to further enhance robustness.

Generalization Capabilities. We further evaluated the model’s generalization capabilities from two perspectives:

Cross-domain generalization to real-world Chinese texts. We tested our model on an adapted excerpt from *Romance of the Three Kingdoms* (Level 5).¹⁰ The predictions of our model show that Challenging and Difficult words account for approximately 5.3% of the text. Our model accurately identifies complex terms such as “三顾茅庐” (three visits to the thatched cottage), while maintaining stable recognition of intermediate vocabulary like “真诚” (sincerely) and “邀请” (invite) (see Table 8). This case demonstrates strong generalization beyond the training distribution, indicating practical applicability in real-world Chinese reading scenarios.

Generalizability across frameworks and languages. DPAL is designed as a general framework whose core mechanism does not depend on any specific language or proficiency framework. As long as grade-labeled resources (graded vocabulary lists and corpora) are available, the dataset construction approach and training framework can be directly transferred. For example, under the CEFR framework, the English Vocabulary Profile (EVP) and corresponding graded corpora (such as Universal-CEFR) can support LCP-LR data annotation and model training for English. EVP provides more detailed sense-level proficiency information than the HSK vocabulary list used in this work, potentially supporting finer-grained annotation and difficulty modeling, while multilingual corpora such as UniversalCEFR provide a foundation for cross-lingual extension. Although the current work is primarily validated within the Chinese HSK framework, our approach opens a clear path for cross-framework and cross-lingual extension.

7 Conclusion

To facilitate the practical deployment of LCP, we introduce the LCP-LR task. We construct LCP-

Romance of the Three Kingdoms (Excerpt)

这_E个_E故事_E就是_E“三顾茅庐”_D。“三顾茅庐”_D的字面_D意思_E是_E三_E次_E去_E草屋_C拜访_R。现在_E它_E是_E一_E个_E成语_R了_E，意思_E是_E一_E个_E人_E为_E了_E完成_E一_E件_E重要_E的_E事情_E，非常_E真诚_R地_E，一_E次_E又_E一_E次_E地_E邀请_R别人_E。

Translation: This story is “Three Visits to the Thatched Cottage.” The literal meaning of this idiom is to go to a thatched cottage three times to pay a visit. Now it signifies a person sincerely and repeatedly inviting others to accomplish an important task.

Table 8: Annotation of a sample paragraph from *Romance of the Three Kingdoms* using our DPAL model. Labels: Easy (E), Readable (R), Challenging (C), and Difficult (D).

CLR, the first densely annotated full-token dataset for Chinese leveled reading, comprising 1,401 texts and 100,917 annotated instances. Experimental results show that our DPAL framework significantly outperforms existing baselines and LLMs, demonstrating the advantage of lightweight, task-specific architectures. Further error analyses and case studies provide deeper insights into the model’s limitations and generalization potential. Collectively, this work provides a robust technical foundation for dynamic text difficulty adaptation, advancing the integration of LCP into intelligent education.

Limitations

This study has three main limitations.

First, limited annotation granularity. Due to the linguistic reality that most words in reading texts are simple and common, approximately 80% of word types in our dataset are associated with only a single label, indicating that our work has not fully escaped the simplified “one word, one complexity” paradigm. Moreover, although we have annotated certain homographs at the usage level, large-scale fine-grained sense-level annotation remains limited, hindering the model’s ability to capture difficulty variations of polysemous words across contexts. Before constructing large-scale sense-level resources, two questions warrant further investigation: (1) Is the difficulty variance across different senses statistically significant? (2) Given the disambiguating function of context, can contextual semantic cues provide sufficient evidence to affect the perceived difficulty of different senses? Balancing annotation costs and practical benefits, future work may build sense-level resources for key polysemous words.

Second, class imbalance. Class imbalance re-

¹⁰<https://duchinese.net>

mains a fundamental constraint on model performance. Future work may introduce external data augmentation or active learning strategies to selectively collect more minority-class samples, thereby further enhancing generalization in real-world reading scenarios.

Third, balancing group-level adaptation and personalization. LCP-LR is designed for the standard graded reading scenario, where learners are assumed to read texts matched to their current proficiency. Accordingly, we adopt text level as a proxy for reader proficiency and vocabulary syllabi as a proxy for readers’ lexical knowledge. While this strategy supports effective group-level adaptation, it remains a coarse-grained approximation and may introduce mismatches between reader and text levels in practical settings. Large-scale estimation of individual readers’ true knowledge levels is inherently difficult. Striking a better balance between group-level adaptation and personalized modeling represents a promising direction for future work.

Ethical Considerations

This work adheres to the ACL Ethics Policy. All textbooks used are publicly available publications, and their use complies with the principles of academic fair use. The vocabulary proficiency syllabi are standard industry documents. Out-of-vocabulary (OOV) items were manually annotated by five graduate students in linguistics; all annotators were informed of the annotation purpose, data usage, and guidelines, and participated voluntarily. Annotators were fairly compensated for their time and effort and consented to the use of their annotations for academic research. No personally identifiable information will be disclosed upon dataset release.

All pre-trained models employed in this work are publicly available under permissive licenses (e.g., Apache 2.0, MIT).

Our model is designed to assist lexical difficulty assessment in leveled reading. The predictions should not be treated as absolute indicators of difficulty; users should consider specific pedagogical contexts and learners’ actual proficiency. The model has limitations in identifying minority-class samples, which users should be aware of.

Acknowledgments

This work is partially supported by the National Social Science Fund of China under Grant No.

23BYY198, and the Tsinghua University Graduate Education Reform Project 2026 under Grant No. 202604J023. The computations were performed on the High Performance Computing Platform of Tsinghua University. We also thank the anonymous reviewers for their constructive feedback.

References

- Sweta Agrawal and Marine Carpuat. 2023. [Controlling pre-trained language models for grade-specific text simplification](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12807–12819, Singapore. Association for Computational Linguistics.
- Mustafa Albay. 2017. The benefits of graded reading. *International Journal of Social Sciences & Educational Studies*, 3(4):177–180.
- David Alfter. 2025. [The need for truly graded lexical complexity prediction](#). In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 326–333, Vienna, Austria. Association for Computational Linguistics.
- Aisha Alowais. 2021. The effects of leveled reading on second language learners. *International Journal of Research in Education and Science*, 7(4):1281–1299.
- Sumit Asthana, Hannah Rashkin, Elizabeth Clark, Fantine Huot, and Mirella Lapata. 2024. [Evaluating LLMs for targeted concept simplification for domain-specific texts](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6208–6226, Miami, Florida, USA. Association for Computational Linguistics.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. [Qwen technical report](#). *arXiv preprint arXiv:2309.16609*.
- Tuqa Bani Yaseen, Qusai Ismail, Sarah Al-Omari, Es-lam Al-Sobh, and Malak Abdullah. 2021. [JUST-BLUE at SemEval-2021 task 1: Predicting lexical complexity using BERT and RoBERTa pre-trained language models](#). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 661–666, Online. Association for Computational Linguistics.
- Joachim Bingel, Gustavo Paetzold, and Anders Søgaard. 2018. [Lexi: A tool for adaptive, personalized text simplification](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 245–258, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Yong Cheng, Jun Dong, and Shuhua Jin. 2023. [Construction and application of the new standard-based cfl text reading difficulty grading system](#). *Chinese Teaching in the World*, 37(1):98–110.

- Yong Cheng, Dekuan Xu, and Jun Dong. 2020. Automatic grading of chinese text reading difficulty based on multiple linguistic features and deep features. *Journal of Chinese Information Processing*, 34(4):101–110.
- Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. [Revisiting pre-trained models for Chinese natural language processing](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668, Online. 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.
- 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.
- Noemie Elhadad and Komal Sutaria. 2007. [Mining a lexicon of technical terms and lay equivalents](#). In *Biological, translational, and clinical language processing*, pages 49–56, Prague, Czech Republic. Association for Computational Linguistics.
- FlagAlpha. 2024. [Llama3-Chinese-8B-Instruct: A chinese instruction-tuned model based on Meta-Llama-3-8B](#). Hugging Face.
- Sian Gooding and Ekaterina Kochmar. 2018. [CAMB at CWI shared task 2018: Complex word identification with ensemble-based voting](#). In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 184–194, New Orleans, Louisiana. Association for Computational Linguistics.
- Sian Gooding, Ekaterina Kochmar, Seid Muhie Yimam, and Chris Biemann. 2021. [Word complexity is in the eye of the beholder](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4439–4449, Online. Association for Computational Linguistics.
- Sian Gooding and Manuel Tragut. 2022. [One size does not fit all: The case for personalised word complexity models](#). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 353–365, Seattle, United States. Association for Computational Linguistics.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [Lora: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations (ICLR)*.
- Abdelhak Kelious, Mathieu Constant, and Christophe Coeur. 2025. [How do large language models evaluate lexical complexity?](#) In *Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*, pages 348–361, Suzhou, China. Association for Computational Linguistics.
- J. Peter Kincaid, Robert P. Jr. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Institute for Simulation and Training, University of Central Florida.
- Batia Laufer. 1989. What percentage of text-lexis is essential for comprehension? In *Vocabulary*, pages 246–251. Routledge.
- Batia Laufer and Geke C Ravenhorst-Kalovski. 2010. Lexical threshold revisited: Lexical text coverage, learners’ vocabulary size and reading comprehension. *Reading in a Foreign Language*, 22(1):15–30.
- Bruce W. Lee, Yoo Sung Jang, and Jason Lee. 2021. [Pushing on text readability assessment: A transformer meets handcrafted linguistic features](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10669–10686, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Wenbiao Li and Yunfang Wu. 2023. Chinese readability assessment based on deep neural networks. *Journal of Chinese Information Processing*, 37(2):158–168.
- I Nation. 2006. How large a vocabulary is needed for reading and listening? *Canadian modern language review*, 63(1):59–82.
- Kai North, Marcos Zampieri, and Matthew Shardlow. 2023. Lexical complexity prediction: An overview. *ACM Computing Surveys*, 55(9):1–42.
- OpenAI. 2024. [GPT-4o Mini: Advancing cost-efficient intelligence](#). OpenAI Blog.
- Gustavo Paetzold and Lucia Specia. 2016. [Inferring psycholinguistic properties of words](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 435–440, San Diego, California. Association for Computational Linguistics.
- Andrei Paraschiv, Mihai Dascalu, and Marina Solnyshkina. 2025. Assessing the readability of russian textbooks using large language models. *Information*, 16(12):1071.
- Sarah E Petersen and Mari Ostendorf. 2009. A machine learning approach to reading level assessment. *Computer speech & language*, 23(1):89–106.
- Robert Pondiscio and Kevin Mahnken. 2014. Leveled reading: The making of a literacy myth. *The Thomas B. Fordham Institute*.

- Matthew Shardlow. 2013. [A comparison of techniques to automatically identify complex words](#). In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 103–109, Sofia, Bulgaria. Association for Computational Linguistics.
- Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Hülsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Pérez Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, and 3 others. 2024a. [The BEA 2024 shared task on the multilingual lexical simplification pipeline](#). In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 571–589, Mexico City, Mexico. Association for Computational Linguistics.
- Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. [SemEval-2021 task 1: Lexical complexity prediction](#). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1–16, Online. Association for Computational Linguistics.
- Matthew Shardlow, Kai North, and Marcos Zampieri. 2024b. [Multilingual resources for lexical complexity prediction: A review](#). In *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024*, pages 51–59, Torino, Italia. ELRA and ICCL.
- Yao-Ting Sung, Ju-Ling Chen, Ji-Her Cha, Hou-Chiang Tseng, Tao-Hsing Chang, and Kuo-En Chang. 2015. Constructing and validating readability models: the method of integrating multilevel linguistic features with machine learning. *Behavior research methods*, 47(2):340–354.
- Keren Tan, Yunshi Lan, Yang Zhang, and Anqi Ding. 2024. Chinese text readability grading via multilevel linguistic feature fusion. *Journal of Chinese Information Processing*, 38(5):41–52.
- Ziyang Wang, Sanwoo Lee, Hsiu-Yuan Huang, and Yunfang Wu. 2024. [FPT: Feature prompt tuning for few-shot readability assessment](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 280–295, Mexico City, Mexico. Association for Computational Linguistics.
- Siyuan Wu, Dong Yu, and Xin Jiang. 2020. [Development of linguistic features system for chinese text readability assessment and its validity verification](#). *Chinese Teaching in the World*, 34(1):81–97.
- Lijiao Yang, Huidan Xu, and Peiyan Song. 2025. [A study on the feature system for readability assessment of chinese texts: From a language cognition-based interpretability perspective](#). *Applied Linguistics*, 34(1):100–113.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.
- Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. [A report on the complex word identification shared task 2018](#). In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, Louisiana. Association for Computational Linguistics.
- Yinbing Zhang, Jihua Song, Weiming Peng, Dongdong Guo, and Jin Zhang. 2019. Analogy analysis of vocabulary level based on new hsk vocabulary syllabus. *Journal of Chinese Information Processing*, 33(9):39–49.
- Zeyu Zhao, Ningtao Wang, Xing Fu, and Yu Cheng. 2025. [Chinese modernbert with whole-word masking](#). *arXiv preprint*, abs/2510.12285.

A Dataset Construction: Human Annotation

A.1 Textbook Grading

The grading of textbook texts followed a consensus-based voting mechanism. The original publisher-designated levels served as the baseline, which were cross-validated by three independent systems: ACTR, LDU-TGP, and the Hansi Chinese Leveled Reading System. A text was assigned a level if at least two systems produced results consistent with the publisher’s designation. Instances where all three systems disagreed or showed substantial deviation from the original level were flagged as “pending review.” These cases were then manually audited by a senior instructor of Teaching Chinese to Speakers of Other Languages (TCSOL), who made the final determination by synthesizing linguistic features, thematic content, and target reader proficiency.

A.2 Homograph Annotation

Of the 89 cross-level homographs in the vocabulary syllabus, 87 appeared in our corpus, totaling 8,382 instances. Among these, 73 words (1,459 instances) exhibited potential complexity variation depending on context. A two-tier annotation strategy was adopted:

- **POS-distinguishable (58 words, 1,095 instances, 75.1 %):** Difficulty was determined by part-of-speech. The pkuseg toolkit

(v0.0.25, default model) was used for POS tagging, with levels assigned automatically.

- **POS-indistinguishable (15 words, 364 instances, 24.9%):** Two linguistics graduate students performed independent manual annotation. Inter-annotator agreement reached 97.5%, with remaining discrepancies resolved through team discussion.

A.3 Out-of-Vocabulary (OOV) Annotation

The corpus contains 7,048 syllabus words and 5,562 OOV words. A team of five graduate students in linguistics and TCSOL conducted the annotation following a three-stage, nine-level framework based on the *Chinese Proficiency Grading Standards for International Chinese Language Education*. Annotators received a one-week training program covering syllabus interpretation, guideline instruction, and pilot exercises. The annotation process lasted approximately three weeks.

Each OOV word was assessed along four dimensions:

- **Word frequency:** Based on the BCC Corpus word frequency list.¹¹
- **Morphological structure:** Morpheme composition, derivational capacity, and combinatorial patterns.
- **Semantic complexity:** Abstractness (referencing the Chinese Concreteness Norms),¹² number of senses, and semantic transparency.
- **Level inference rules:** Following Zhang et al. (2019), inferring the level of unknown words from morphemes of known-level words.

Note that 1,491 proper nouns were excluded from prediction targets but retained as contextual information.

The annotation workflow proceeded in three phases:

1. **Independent annotation:** Each word annotated by two annotators.
2. **Cross-validation:** Annotations compared; consistent ones accepted, divergent ones marked “pending adjudication.”

3. **Adjudication:** Authors reviewed pending cases and made final determinations.

To assess reliability, 10% of annotations were randomly sampled and re-annotated by two independent researchers. Cohen’s Kappa exceeded 0.85, indicating high reliability. All disputed items were resolved through adjudication.

B Experimental Setup

B.1 Pre-trained Language Models

We evaluated multiple pre-trained language models under a unified hyperparameter configuration (Table 9). For batch size configuration, we used a training batch size of 16 and a validation batch size of 32 for base-scale models (≤ 110 M parameters), with a gradient accumulation step of 1. For large-scale models (≥ 250 M parameters), we used a training batch size of 8 and a validation batch size of 16, with a gradient accumulation step of 2.

Parameter	Value
Max Seq Len	256
Epochs	5
Learning Rate	2e-5
Optimizer	AdamW
Weight Decay	0.01
Warm-up Steps	10%
Loss Function	Weighted CE

Table 9: Unified hyperparameter settings.

Class weights were calculated using inverse frequency weighting:

$$w_c = \frac{N}{C \times n_c} \quad (2)$$

where N is the total number of samples, C is the number of classes, and n_c is the sample count for class c . The resulting weights are shown in Table 10.

Category	Sample Proportion	Weight
Easy	86.2%	0.3
Readable	8.3%	3.5
Challenging	1.3%	20.0
Difficult	4.2%	7.0

Table 10: Class weight settings computed via inverse frequency weighting.

¹¹<https://bcc.blcu.edu.cn>

¹²https://github.com/xiezp1976/chinese_concreteness

	Easy	Readable	Challenging	Difficult
Easy	0.983	0.014	0.001	0.002
Readable	0.182	0.710	0.019	0.089
Challenging	0.048	0.246	0.629	0.078
Difficult	0.033	0.167	0.033	0.767

Figure 5: Confusion matrix of the proposed DPAL framework on the test set. Values are normalized by row to represent percentages (%). Misclassifications occur mainly between adjacent difficulty levels, with a bias toward Readable.

B.2 Large Language Models

Fine-tuning Settings for Open-source Models. We fine-tuned Qwen2.5-7B-Instruct (Bai et al., 2023) on 20,000 samples using LoRA (Hu et al., 2022) with $r = 16$, $\alpha = 32$, and dropout=0.1, applied to the projection layers (q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj). The model was loaded with 4-bit NF4 quantization using float16 precision. Inputs followed the Qwen chat template, including system prompts, user inputs, and assistant responses. Training used a batch size of 2 with 8 gradient accumulation steps (effective batch size 16), maximum sequence length of 384, learning rate of $2e-4$, 100 warm-up steps, cosine annealing, and 5 epochs, selecting the checkpoint with the lowest validation loss. The same configuration was applied to Llama3-Chinese-8B-Instruct (FlagAlpha, 2024), with inputs formatted to comply with the Llama3 chat template.

Prompt Engineering. The complete zero-shot prompt template is presented in Table 11. The few-shot and chain-of-thought templates follow the same base format but prepend labeled examples and step-by-step reasoning, respectively (see Tables 12 and 13 for examples).

All experiments were conducted on a single NVIDIA GeForce RTX 4090 GPU (24GB VRAM). The total computational budget for all experiments was approximately 150 GPU hours.

C Supplementary Experimental Results

C.1 Confusion Matrix of DPAL

Figure 5 presents the confusion matrix of DPAL on the test set.

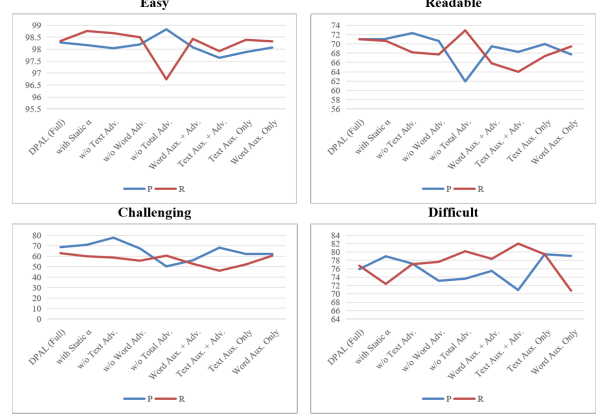


Figure 6: Precision and Recall Comparison of Ablation Models across Categories (%).

C.2 Ablation Study

Figure 6 presents the precision and recall of ablation variants across difficulty categories.

The analysis reveals three main findings. First, the Easy category remains highly stable across all variants. Second, the Readable and Challenging categories exhibit a consistent complementarity:

- **Lexical prior ensures recall.** Variants without lexical auxiliary tasks (Text Aux. + Adv.) show the lowest recall for Readable (64.00%) and Challenging (46.11%), indicating that absolute lexical difficulty priors are essential for coverage and recall.

- **Lexical adversarial training controls precision.** Variants retaining lexical adversarial debiasing (w/o Text Adv.) achieve high precision on Readable (72.32%) and Challenging (77.67%), whereas removing all adversarial components (w/o Total Adv.) causes precision to drop sharply to 61.93% and 50.25%. This reveals a delicate complementarity: lexical priors secure recall, while adversarial training refines precision.

Third, for the Difficult category, the T-Aux-only variant achieves the highest precision (79.45%), but introducing adversarial training increases recall (from 79.45% to 82.00%) at the cost of precision (dropping to 70.91%). The full model achieves a balanced trade-off. This suggests that absolute text difficulty priors are highly effective for the Difficult category, whereas introducing lexical or adversarial signals may introduce interference, implying that different difficulty categories may require differentiated modeling strategies.

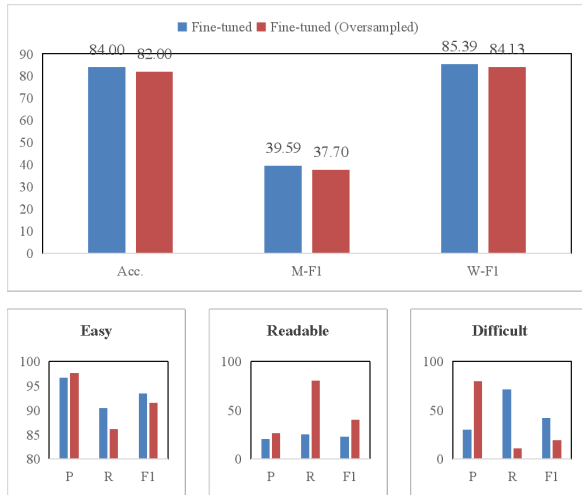


Figure 7: Results of the Oversampling Experiment (%).

C.3 Fine-tuning Performance with Oversampling

To mitigate the class imbalance in our dataset, we performed oversampling while preserving the original training set distribution. Specifically, we extracted all available Challenging samples (443 instances) and a subset of Difficult samples (253 instances) from the original data pool, increasing their totals in the training set to 647 and 1,000, respectively. The validation and test sets remained unchanged. Using Qwen2.5-7B as a representative model, Figure 7 reports the fine-tuning results.

The oversampled model achieved an overall accuracy of 82.00% and a macro-F1 of 37.70%, slightly lower than the original sampling. By category, the Easy F1 decreased marginally (91.54% vs. 93.45%), the Readable F1 improved (40.00% vs. 22.63%), while the Difficult F1 dropped substantially (19.28% vs. 42.28%), and the Challenging class remained largely unrecognizable.

Oversampling did not yield overall improvement and even degraded performance on certain categories, particularly Difficult. Two factors may explain this: (1) The Challenging class still lacks sufficient absolute sample size, and the duplicated samples introduced by oversampling provide no new semantic information; (2) The semantic boundary between Challenging and Readable is inherently ambiguous, which mere sample augmentation cannot resolve. These findings suggest that the constraints of data imbalance on model performance stem not only from sample quantity but also from class separability. Future work may explore data augmentation or active learning strategies to se-

lectively collect more discriminative samples for minority classes.

Zero-shot Prompt Template
You are an expert in International Chinese Language Education. Text levels range from 1 to 7 (1 = easiest, 7 = hardest), indicating the overall linguistic complexity of the context in which the target word appears and the proficiency level of the target reader. Note that the same word may pose different levels of difficulty to learners at different proficiency levels. Target word: {target_word} Context: {context} Text level: {text_level} Please determine the comprehension difficulty of the word for learners at Level {text_level}. Difficulty definitions: - Easy: The word is far below the learner's current level and poses no comprehension barrier. - Readable: The word falls within the learner's expected vocabulary range and can be understood without assistance. - Challenging: The word is slightly above the learner's current level and may require contextual inference or minimal explanation. - Difficult: The word is far above the learner's current level and poses a significant comprehension barrier even with contextual support. Output only one word: Easy / Readable / Challenging / Difficult

Table 11: Zero-shot prompting template.

Few-shot Prompt Template (Example)
[Example 1] Target word: scenic spot Context: China has a vast territory with beautiful landscapes. Across this expansive land, thousands of scenic spots are scattered. Each scenic spot has its own historical anecdotes, and each scenic area has its own legendary stories. Text level: 7 Answer: Easy <i>Three additional examples (Readable, Challenging, Difficult) follow the same format.</i>

Table 12: Few-shot prompting template (example).

Chain-of-Thought Prompt Template (Example)
[Example 1] Target word: scenic spot Context: China has a vast territory with beautiful landscapes. Across this expansive land, thousands of scenic spots are scattered. Each scenic spot has its own historical anecdotes, and each scenic area has its own legendary stories. Text level: 7 Reasoning: 1. Word complexity assessment: "scenic spot" is an intermediate-level word. 2. Text difficulty assessment: Level 7 is for advanced learners. 3. Matching analysis: Intermediate word + advanced text -> word difficulty below text level. 4. Overall judgment: For advanced learners, this word poses no comprehension barrier. Answer: Easy <i>Three additional examples (Readable, Challenging, Difficult) follow the same format with corresponding reasoning.</i>

Table 13: Chain-of-Thought prompting template (example).