Minimax Theory and its Applications

Volume 11 (2026), No. 5s, 142-162

Psycholinguistically Grounded Conformal Ordinal Transformer for Trustworthy CEFR-Level Difficulty Assessment of English Reading Texts

Enamalakalavathi[1], Dr G Thippanna[2], Dr.E.S.Sharmila Sigamany[3], S. Venkata Suryanarayana[4], Dr. M. Muni Babu[5], Elangovan Muniyandy[6], Dr. Subramanyam Kunisetti[7], Malladi Ambarisha[8], Wael Anabousi[9]

[1]Research Scholar, Department of English, Koneru Lakshmaiah Education Foundation, Vaddeswaram,AP, India , Email: akalavathi1986@gmail.com

[2]Professor, Dept. of CSE-AIML, Dr KV Subbareddy Institute of Technology , Kurnool, Andhra Pradesh, India - 518218, Email: gt.pana2012@gmail.com, orcid id: 0009-0009-6964-246X

[3]Associate professor, Department of English, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India , Email: sharmilajoe@kluniversity.in

[4]Professor, Department of CSE (Data Science), CVR College of Engineering, Hyderabad, Telangana, India. , Email: suryahcu@gmail.com

[5]Dean, School of Engineering and Applied Technology (SEAT), BEST Innovation University, Gorantla, AP, Email: munibabu.m@gmail.com

[6]Department of Biosciences, Saveetha School of Engineering. Saveetha Institute of Medical and Technical Sciences,Chennai - 602 105, Email: muniyandy.e@gmail.com

[7]Associate Professor, Department of CSE (Data Science), R.V.R. & J.C. College of Engineering, Email: subramanyamkunisetti@gmail.com, ORCID id: 0000-0003-3241-9888

[8]Assistant Professor, Department of Information Technology, RVR&JC College of Engineering, Chandramouli Nagar, Chowdavaram – 522019 Guntur Dist. Andhra Pradesh, India. Email: ambarisha.malladi@gmail.com

[9]Hourani e-Learning Department, Al-Ahliyya Amman University, Amman, Jordan.
Email: w.anbousi@ammanu.edu.jo, orcid.org/0009-0004-3913-939X

Abstract

Matching reading materials to learners' proficiency is central to English language teaching, and the Common European Framework of Reference for Languages (CEFR) has become the standard scale for doing so. Automatic CEFR-level assessment of texts is, however, complicated by two properties that existing models largely ignore. First, the levels are ordered, so a B1 text predicted as C2 is a far worse error than one predicted as B2. Second, expert labels are intrinsically ambiguous: in the largest sentence-level CEFR corpus, two trained annotators assigned identical levels to only 37.6% of sentences, and in a recent document-level study four experienced raters initially disagreed on 36.7% of texts. Point predictions from classifiers trained on a single label hide this ambiguity and give teachers no indication of how far a prediction can be trusted. This study proposes PsyCORT, a Psycholinguistically grounded Conformal Ordinal Readability Transformer, which combines a DeBERTa-v3 encoder with lexical-sophistication, syntactic-complexity and cohesion features through Feature-Guided Cross-Attention, predicts levels with a rank-consistent CORN ordinal head trained on disagreement-aware soft targets, and outputs contiguous CEFR intervals (for example, B1–B2) with a distribution-free coverage guarantee obtained by ordinal conformal prediction. A new ambiguity-aware calibration rule guarantees that the interval contains at least one level that the annotators considered acceptable. The statistical behaviour of the conformal layer was verified in a controlled simulation that reproduces the inter-annotator agreement of CEFR-SP: at a target coverage of 90%, ordinal conformal sets achieved 90.0% coverage and were always contiguous, whereas standard least-ambiguous and adaptive sets were contiguous in only 93.5% and 90.4% of cases when the base model was not ordinal; ambiguity-aware calibration attained exactly 90.1% coverage of acceptable levels with 24–27% smaller intervals. The paper further specifies the full experimental protocol, hypotheses and ablations for the CEFR-SP and CEFR-labelled CLEAR corpora, for which the transformer experiments are in progress. PsyCORT reframes CEFR text assessment as calibrated, set-valued ordinal prediction, giving teachers level ranges whose reliability is explicit and whose linguistic drivers are explained.

Keywords: CEFR; Readability assessment; English as a foreign language; Ordinal regression; Conformal prediction; Psycholinguistic features; Uncertainty quantification

1. Introduction

Selecting texts that learners can read with understanding but that still stretch their ability is one of the most frequent decisions in English language teaching. Teachers, materials writers and adaptive learning platforms increasingly express this decision on the Common European Framework of Reference for Languages (CEFR), whose six levels from A1 to C2 describe what learners can do with the language and have been refined in the 2020 Companion Volume [1]. Placing a text on the CEFR scale, however, still relies on expert judgement, which is slow, costly and difficult to scale to the volume of digital reading material now available. Automatic CEFR-level assessment of texts has therefore become an active area at the intersection of applied linguistics and natural language processing.

Two research traditions have shaped this area. The first, rooted in second language acquisition and readability research, models text difficulty through interpretable linguistic complexity measures. Recent work in this tradition asked expert raters to classify 1,181 excerpts of the CommonLit Ease of Readability (CLEAR) corpus [2] into CEFR levels and showed that a random forest using 24 lexical, syntactic and discourse features reached 62.6% exact-level accuracy on six levels and 82.6% on three broad levels, with age of acquisition of content words emerging as the strongest single predictor [3]. The second tradition fine-tunes pretrained language models on labelled corpora. The CEFR-based Sentence Profile (CEFR-SP), containing 17,676 sentences annotated by English-education professionals, enabled a sentence-level model to reach a macro-F1 of 84.5% [4], and document-level studies have used Cambridge examination texts [5]. Hybrid models that combine transformer representations with handcrafted linguistic features have repeatedly outperformed either source alone in readability assessment [6], [7], and the recent UniversalCEFR initiative has assembled more than 500,000 CEFR-labelled texts to support open research [8].

Despite this progress, nearly all current systems treat CEFR assessment as ordinary single-label classification. This framing conflicts with the nature of the task in two ways. CEFR levels are ordinal: the cost of an error grows with its distance on the scale, and probability mass placed on distant levels is a sign of a poorly behaved model. More importantly, the labels themselves are uncertain. In CEFR-SP, the two annotators assigned the same level to only 37.6% of sentences and levels within one grade to 88.4%, so the corpus authors treated both labels as correct when they differed by one level [4]. In the CLEAR-based study, four raters with more than a decade of teaching experience initially disagreed on 36.7% of the texts and resolved them only through discussion [3]. The single predicted level therefore conveys false precision; it cannot show that a student's performance on the test is at some point between levels B1 and B2, and it does not allow a teacher to determine if the prediction will be considered an acceptable level.

The first problem is addressed by Conformal Prediction. It provides a principled solution to the second problem. A Conformal Prediction model takes as input a trained model and a separate calibration set (a hold-out sample) and produces prediction sets from each predictive score. These prediction sets are guaranteed to include the actual label with a specified probability, as long as the data is exchangeable [9]. In terms of CEFR assessment, a Conformal Predictor will produce continuous intervals (e.g., B1-B2) that are guaranteed to encompass the correct CEFR level for, for example, 90% of all texts. This type of output aligns well with how educators typically view difficulty and allows for explicit quantification of reliability. The main goal of this study is to combine Conformal Intervals, Ordinal Modeling, and Interpretable Linguistic Evidence.

Therefore, this research is designed around four research questions. RQ1: Is there an improvement in assessing CEFR-levels when combining psycholinguistic and linguistic complexity features (with cross-attention) with contextual representations compared to using either alone? RQ2: Does rank-consistent ordinal modeling with disagreement-aware target labels decrease the number of large classification errors? RQ3: Can conformal prediction generate CEFR level intervals that have appropriate coverage, and what width must those intervals be in order to capture the ambiguity in expert labeling? RQ4: What are the linguistic factors driving the predictions of the levels and can those factors be communicated to teachers? Question three is addressed through the conformal layer via the simulation study presented herein; all four questions form the basis of our experiments on real corpora.

1.1 Problem statement

Four related issues limit existing automatic CEFR assessments: Expert agreement on CEFR levels is very low [3], [4]. Existing systems use a single gold label per text to train and evaluate, while experts often agree that there are multiple CEFR levels applicable to a single text. Existing systems provide nominal class predictions, rather than ordinal predictions, which means that the system does not necessarily decrease in probability as it moves away from its predicted CEFR level. This may lead to some strange predictions such as predicting high probabilities for both A2 and C1. Existing systems provide only point predictions; they do not provide any measure of reliability. As a result, users cannot tell if the system is confident in its prediction or if it is merely guessing. Finally, although there has been a great deal of research identifying psycholinguistic and complexity

measures (e.g., age of acquisition [10]) that have relevance for teachers, many deep learning models do not clearly show what aspects of language influenced their decision. An ordinal, uncertainty aware framework with formal guarantees and linguistic interpretability would be valuable.

1.2 Research motivation

The reason I am doing this research is because teachers in the real world use the CEFR predictions when they select texts for their classes. Teachers selecting a text for a B1 level student will want to make sure the selected text falls into the B1 range, and possibly B2, so that they can provide the best possible learning experience. They will also need to be confident that their prediction will be consistent regardless of what level they are working at (even if it is one of the extreme levels). There are several different approaches to achieving this goal. For example, ordinal regression techniques like CORAL [11] and its conditional version CORN [12] ensure that all ranks have consistent probabilities; meanwhile, conformal prediction provides a mechanism for providing a set of possible answers that is calibrated [9]. Simply calibrating the raw probabilities provided by a neural network is not enough, as most modern networks tend to be overly confident [13]. Combining these methods with linguistic-based representations should produce predictions that are accurate and truthful about their uncertainty.

1.3 Significance of the study

The importance of this work for English as a foreign language (EFL) instruction is that it transforms the automatic CEFR assessment process from a black box labeling system to an objective-based decision-making tool. With interval-based predictions that have a known coverage rate, each institution can choose what reliability level they want to use; e.g., 95% coverage for placement tests, and 80% coverage for supplementary reading materials. In addition, the use of ambiguity-aware calibration provides guarantees that are consistent with how expert raters assign CEFR levels, and psycholinguistically based explanations provide teachers with information about why a text may be difficult (e.g., late acquired vocabulary, long T-units, or poor cohesive ties) so they can adapt the text appropriately.

1.4 Key contributions

- The Feature-Guided Cross-Attention (FGCA) module combines the features of lexical sophistication, syntactic complexity, and cohesive properties of the text with DeBERTa-v3 token representations. This combination allows the linguistic evidence to shape the representation of the text while also allowing each feature to be represented by the specific words used to express it.
- A rank-consistent CORN ordinal head trained on disagreement-aware soft targets that encode all levels accepted by annotators, together with an earth mover's distance regularizer that penalizes probability mass far from the acceptable levels.
- An ambiguity-aware ordinal conformal predictor that outputs contiguous CEFR intervals and guarantees, with finite-sample validity, that each interval contains at least one annotator-acceptable level with a user-chosen probability.
- A simulation study, calibrated to the published inter-annotator agreement of CEFR-SP, that verifies coverage, interval size, contiguity and level-wise behaviour of the conformal layer under ordinal and nominal base models.
- A complete, reproducible evaluation design for the CEFR-SP and CEFR-labelled CLEAR corpora, with explicit hypotheses, ablations and success criteria, and an open implementation.

1.5 Organization of the remaining sections

Section 2 reviews related work on readability and CEFR assessment, ordinal learning and conformal prediction. Section 3 describes the proposed PsyCORT framework. Section 4 reports the simulation-based validation of the conformal layer and specifies the experimental design, hypotheses and expected results for the transformer experiments. Section 5 concludes the paper.

2. Related Works

Vajjala and Meurers [14] showed that features inspired by second language acquisition research, including lexical richness and syntactic complexity measures, substantially improve readability classification compared with traditional formulas. Their work established that linguistic complexity measures developed for learner language are also informative about text difficulty, but it relied on graded reading materials for native speakers rather than CEFR-annotated data.

Xia et al. [5] addressed readability assessment specifically for second language learners using texts from Cambridge English examinations aligned with CEFR levels. They combined traditional, lexico-semantic, parse-

tree and discourse features and explored domain adaptation from native-speaker corpora, demonstrating that learner-oriented corpora behave differently from first-language readability data. The limited size of examination corpora, however, restricted the use of deep models.

Vajjala and Lučić [15] released OneStopEnglish, a corpus of 189 news articles each rewritten at elementary, intermediate and advanced levels, giving 567 texts with parallel versions. The corpus has become a standard benchmark for readability assessment and text simplification, although its three levels are coarser than the CEFR scale and the parallel structure makes level differences unusually systematic.

Deutsch et al. [16] systematically evaluated whether linguistic features add value to neural readability models and found that their contribution depended strongly on the corpus, with gains on some datasets and none on others. Lee et al. [6] subsequently showed that combining transformer predictions with a large set of handcrafted linguistic features in a hybrid model yields consistent improvements across several benchmarks. Martinc et al. [7] compared supervised and unsupervised neural approaches and reported that neural models capture complementary difficulty cues but that their behaviour varies across corpora with different notions of readability.

Crossley et al. [17] developed the Coh-Metrix L2 Reading Index from cognitively motivated indices of word frequency, syntactic similarity and content-word overlap and showed that it predicted second language readers' comprehension better than traditional formulas. Collins-Thompson [18] surveyed computational readability research and argued that future systems should model the reader as well as the text, provide richer outputs than a single grade and account for uncertainty in difficulty judgements, a direction that the present study takes up for CEFR assessment.

Arase et al. [4] created CEFR-SP, annotating 20,000 sentences from diverse sources with CEFR levels by two English-education professionals. Because exact agreement was 37.6% and agreement within one level 88.4%, both labels were retained whenever they differed by one level, producing 27,841 labels for 17,676 sentences. Their assessment model addressed the scarcity of A1 and C2 sentences and achieved a macro-F1 of 84.5%. The corpus thus explicitly encodes label ambiguity, but the evaluation still scored single-level predictions.

Zhang and Lu [3] asked four expert EFL teachers to classify 1,181 CLEAR excerpts into CEFR levels and trained a random forest on 24 lexical, syntactic and cohesion measures selected from 353 candidates. The model achieved 62.6% exact-level accuracy on six levels and 82.6% on three broad levels, outperforming a baseline built on traditional readability features. Age of acquisition and lexical decision times of content words were the most important predictors, and the feature profiles distinguishing A from B differed from those distinguishing B from C. The study provides an openly available, expert-labelled document-level CEFR corpus and a strong interpretable baseline, but it did not use contextual language representations or quantify prediction uncertainty.

Imperial et al. [8] assembled UniversalCEFR, a collection of 505,807 CEFR-labelled texts across multiple languages, and benchmarked feature-based models, fine-tuned transformers and prompted large language models. Their analysis showed that prompting large language models for CEFR levels is unreliable at the sentence level, reinforcing the need for supervised and calibrated approaches.

In ordinal learning, Cao et al. [11] proposed CORAL, which decomposes an ordinal problem into binary tasks with shared weights to guarantee rank-consistent predictions, and Shi et al. [12] proposed CORN, which uses conditional training subsets to achieve rank consistency without the weight-sharing constraint, improving flexibility. In conformal prediction, Sadinle et al. [19] introduced least ambiguous set-valued classifiers (LAC), Romano et al. [20] proposed adaptive prediction sets (APS) with improved conditional behaviour, and Lu et al. [21] developed ordinal conformal prediction sets that are contiguous by construction and applied them to disease severity rating, a medical analogue of CEFR leveling. To our knowledge, conformal prediction has not been applied to CEFR assessment, and no existing conformal method accounts for set-valued annotations of the kind found in CEFR-SP.

Table 1. Summary of representative studies on readability and CEFR-level assessment

Study	**Data**	**Method**	**Strengths**	**Limitations**
Vajjala&Meurers [14]	Graded readers (L1)	SLA-inspired complexity features	Interpretable; strong gains over formulas	Not CEFR; L1 materials
Xia et al. [5]	Cambridge exam texts	Rich linguistic features, domain adaptation	Learner-oriented data	Small corpus; no deep models

Vajjala& Lučić [15]	OneStopEnglish (567)	Benchmark corpus, feature models	Parallel levels	Only three coarse levels
Deutsch et al. [16]	Several corpora	Neural models ± linguistic features	Systematic comparison	Feature gains corpus-dependent
Crossley et al. [17]	L2 cloze data	Cognitively based indices (CML2RI)	L2-oriented; outperforms formulas	Small text sample; regression only
Lee et al. [6]	Several corpora	Transformer + handcrafted hybrid	Consistent improvements	Point predictions only
Martinc et al. [7]	Several corpora	Supervised and unsupervised neural	Complementary cues	Behaviour varies by corpus
Arase et al. [4]	CEFR-SP (17,676 sentences)	Transformer with imbalance handling	Expert CEFR labels; 84.5% macro-F1	Single-level evaluation
Zhang & Lu [3]	CLEAR-CEFR (1,181 texts)	Random forest, 24 features	Interpretable; open data	62.6% exact accuracy; no uncertainty
Imperial et al. [8]	UniversalCEFR (505,807)	Features, transformers, LLM prompting	Scale; multilingual	LLM prompting unreliable
Lu et al. [21]	Medical severity	Ordinal conformal sets	Contiguous sets, guarantees	Not applied to language

Table 1 summarizes the reviewed studies. The literature shows steady progress in accuracy through hybrid linguistic–neural models and growing availability of CEFR-labelled corpora. At the same time, three gaps remain: CEFR assessment is still framed as single-label nominal classification, the substantial disagreement among expert annotators is not reflected in model outputs, and no existing system offers statistically valid uncertainty statements for its CEFR predictions. PsyCORT is designed to close these gaps.

3. Proposed PsyCORT Framework

PsyCORT addresses CEFR-level assessment as calibrated ordinal prediction under label ambiguity. As shown in Figure 1, a text is preprocessed and represented in two complementary ways: contextual token embeddings from DeBERTa-v3 [22] and a vector of psycholinguistic and linguistic complexity features. Feature-Guided Cross-Attention fuses the two, a CORN head produces rank-consistent level probabilities, and the model is trained against soft targets derived from all annotator-acceptable levels. After training, a held-out calibration set is used to compute an ordinal conformal threshold, so that every prediction is reported as a contiguous CEFR interval with guaranteed coverage, accompanied by explanations of its linguistic drivers.

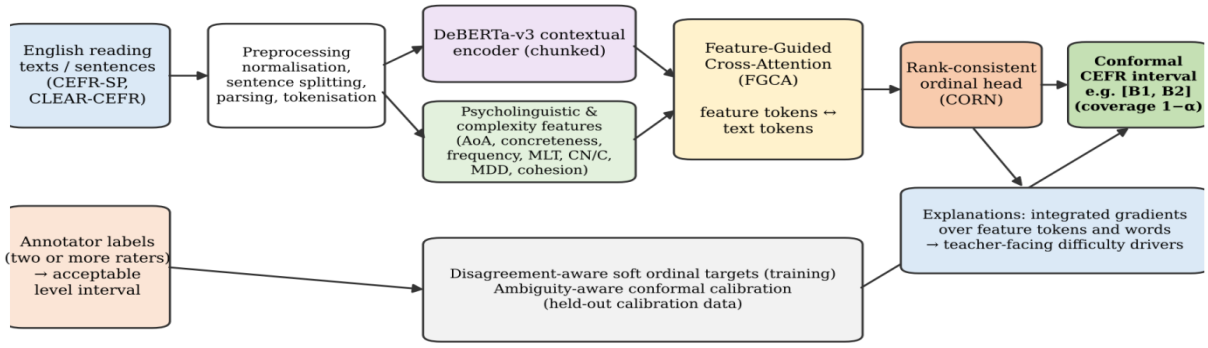


Figure 1. Overview of the proposed PsyCORT framework for trustworthy CEFR-level assessment

Figure 1 shows the overall pipeline. The upper path processes the text through the contextual encoder and the feature extractor, which meet in the FGCA module before the ordinal head. The lower path uses the annotators' labels twice: during training to build disagreement-aware soft targets, and after training to calibrate the conformal interval. The interval and the feature-level explanation are the two outputs delivered to teachers and materials designers.

3.1 Data collection

Two openly available expert-annotated corpora are used, chosen because they cover the sentence and document levels and because both record annotator disagreement. CEFR-SP [4] contains 17,676 English sentences drawn from text-simplification and educational corpora, each labelled by two English-education professionals; when the two labels differ by one level, both are retained, which yields 27,841 labels. CLEAR-CEFR [3] consists of 1,181 excerpts sampled from the CLEAR corpus [2], each 140–200 words long and covering literary and informational texts written between the late eighteenth and the early twenty-first century, classified into six CEFR levels by four experienced EFL teachers through independent rating followed by consensus discussion. OneStopEnglish [15], with 567 texts at three levels, is used only for cross-corpus evaluation on the broad A/B/C scale.

Table 2. Characteristics of the corpora used in this study

Corpus	Unit	Size	Levels	Annotation	Use
CEFR-SP [4]	Sentence	17,676 sentences; 27,841 labels	A1–C2	2 professionals; exact agreement 37.6%, within one level 88.4%	Training, calibration, testing
CLEAR-CEFR [3]	Excerpt (140–200 words)	1,181 texts	A1–C2	4 EFL teachers; 36.7% initial disagreement, consensus	Training, calibration, testing
OneStopEnglish [15]	Article	567 texts (189 × 3)	Elementary, intermediate, advanced	Professional rewriting	Cross-corpus test (A/B/C)

Table 2 summarizes the corpora. For each text i, the set of acceptable levels $\mathcal{A}_i \subseteq \{1, \ldots, K\}$ with $K = 6$ is defined as all levels assigned by at least one annotator when the assignments differ by at most one level, following the CEFR-SP convention [4], and as the consensus level otherwise. For CLEAR-CEFR, where only the consensus label is public, $\mathcal{A}_i$ contains that single level; if the individual ratings become available, the same construction applies.

3.2 Data preprocessing

Texts are normalized by Unicode NFKC normalization, removal of markup and repeated whitespace, and standardization of quotation marks and dashes. Sentence segmentation, tokenization, lemmatization, part-of-speech tagging and dependency parsing are performed with spaCy [23]. For the transformer, texts are tokenized with the DeBERTa-v3 SentencePiece tokenizer; documents longer than 512 subword tokens are split into overlapping chunks of 512 tokens with a stride of 128, and chunk representations are averaged. Where source-document identifiers are available, sentences from the same document are kept in the same data partition to prevent topical leakage.

3.3 Psycholinguistic and linguistic complexity features

The feature vector $\mathbf{f} \in \mathbb{R}^D$ is organized into three groups motivated by the CEFR-oriented findings of Zhang and Lu [3] and by readability research [6], [14].

Lexical sophistication (eight features). For the content words of a text, mean and maximum age of acquisition are computed from the norms of Kuperman et al. [10], mean concreteness from Brysbaert et al. [24], and mean log frequency from SUBTLEX-US [25]. These are complemented by mean word length, lexical density, root type–token ratio and the proportion of out-of-norm words. The mean age of acquisition over content words W_c of text t is given in (1):

$$AoA_t = \frac{1}{|W_c|} \sum_{w \in W_c} AoA(w) \tag{1}$$

where AoA(w) is the rated age in years at which word w is typically learned. Age of acquisition was the strongest predictor of CEFR level in [3], which motivates its central role here.

Syntactic complexity (six features). Following the L2 Syntactic Complexity Analyzer tradition [26], the features are mean sentence length, clauses per sentence, complex nominals per clause, coordinating conjunctions per sentence, maximum parse depth and mean dependency distance [27]. The mean dependency distance of a sentence with n words is defined in (2):

$$MDD = \frac{1}{n-1}\sum_{i=1}^{n-1}\left|p_i - p_{h(i)}\right| \tag{2}$$

where p_i is the position of the i-th dependent and $p_{h(i)}$ the position of its head; longer distances increase working-memory load during parsing.

Cohesion and discourse (five features). Following the Tool for the Automatic Analysis of Cohesion [28], the features are content-lemma overlap between adjacent sentences, connective density, pronoun-to-noun ratio, number of sentences and text length. Adjacent-sentence overlap is computed as in (3):

$$OV = \frac{1}{S-1}\sum_{s=1}^{S-1}\frac{|L_s \cap L_{s+1}|}{|L_s \cup L_{s+1}|} \tag{3}$$

where L_s is the set of content lemmas of sentence s and S the number of sentences. All features are standardized with the mean and standard deviation of the training partition only.

Table 3. Psycholinguistic and linguistic complexity features used by PsyCORT

Group	Feature	Definition / source
Lexical	Mean AoA (content words)	Mean rated age of acquisition, Kuperman et al. norms [10]
Lexical	Maximum AoA (content words)	Latest-acquired content word in the text [10]
Lexical	Mean concreteness	Brysbaert et al. concreteness ratings [24]
Lexical	Mean log frequency	SUBTLEX-US log10 frequency per million words [25]
Lexical	Mean word length	Characters per alphabetic token
Lexical	Lexical density	Content words / all words
Lexical	Root type–token ratio	Types / √tokens
Lexical	Out-of-norm proportion	Share of tokens absent from frequency norms (rare words)
Syntactic	Mean sentence length	Words per sentence
Syntactic	Mean dependency distance	Eq. (2), following [27]
Syntactic	Clauses per sentence	Finite and non-finite clause heads per sentence [26]
Syntactic	Complex nominals per clause	Nouns with adjectival, prepositional or clausal modifiers per clause [26]
Syntactic	Maximum parse depth	Mean over sentences of the deepest dependency path
Syntactic	Coordination per sentence	Coordinating conjunctions per sentence
Cohesion	Adjacent-sentence overlap	Eq. (3), content-lemma Jaccard overlap [28]
Cohesion	Connective density	Connectives per word [28]
Cohesion	Pronoun-to-noun ratio	Indicator of referential givenness

Cohesion	Number of sentences	Text organisation
Cohesion	Text length	Words per text

Table 3 lists the 19 features. They were selected to cover the three dimensions that most consistently predicted CEFR level in recent expert-labelled data [3] while remaining computable with open tools and norms. Length features are included as controls, because CEFR-SP sentence length is not proportional to level beyond A2 [4] and the CLEAR excerpts were length-matched [3].

3.4 Contextual encoding

DeBERTa-v3-base [22] encodes the token sequence $x_1, \ldots, x_T$ into contextual representations, as given in (4):

$$H = (h_1, \ldots, h_T) = DeBERTa(x_1, \ldots, x_T) \tag{4}$$

with $h_t \in \mathbb{R}^{768}$. DeBERTa-v3 is chosen for its disentangled attention over content and relative position, which has been effective for text-level regression and classification tasks in educational applications, and its replaced-token-detection pretraining, which yields strong representations at base size.

3.5 Feature-Guided Cross-Attention

Concatenating features with a pooled text vector, the common hybrid strategy [6], lets the classifier use both sources but does not allow linguistic evidence to influence which words the representation emphasizes. FGCA instead converts each feature group g into a token, as in (5):

$$e_g = LN\left(GELU\left(W_g f_g + b_g\right)\right) + \gamma_g \tag{5}$$

where $\mathbf{f}_g$ is the sub-vector of group g, W_g and b_g are learnable, and γ_g is a learnable group embedding, giving feature tokens $E = (e_{lex}, e_{syn}, e_{dis})$. Each FGCA layer applies two cross-attention operations [29]. Text-to-feature attention enriches every token with linguistic evidence, as in (6):

$$\tilde{H} = LN\left(H + softmax\left(\frac{HW_Q(EW_K)^\top}{\sqrt{d}}\right)EW_V\right) \tag{6}$$

and feature-to-text attention grounds each feature token in the words that realize it, as in (7):

$$\tilde{E} = LN\left(E + A_{f\to t}HW_V'\right),\ \ A_{f\to t} = softmax\left(\frac{EW_Q'(HW_K')^\top}{\sqrt{d}}\right) \tag{7}$$

Both outputs pass through position-wise feed-forward layers with residual connections, and two FGCA layers are stacked. The attention matrix $A_{f\to t}$ indicates, for example, which words the lexical-sophistication token attends to, providing a first layer of interpretability. The final representation combines the mean-pooled text tokens $\bar{h}$ and feature tokens $\bar{e}$ through a learned gate, as in (8):

$$u = \sigma(w_u\top[\bar{h};\bar{e}]) \odot \bar{h} + \left(1 - \sigma(w_u\top[\bar{h};\bar{e}])\right) \odot \bar{e} \tag{8}$$

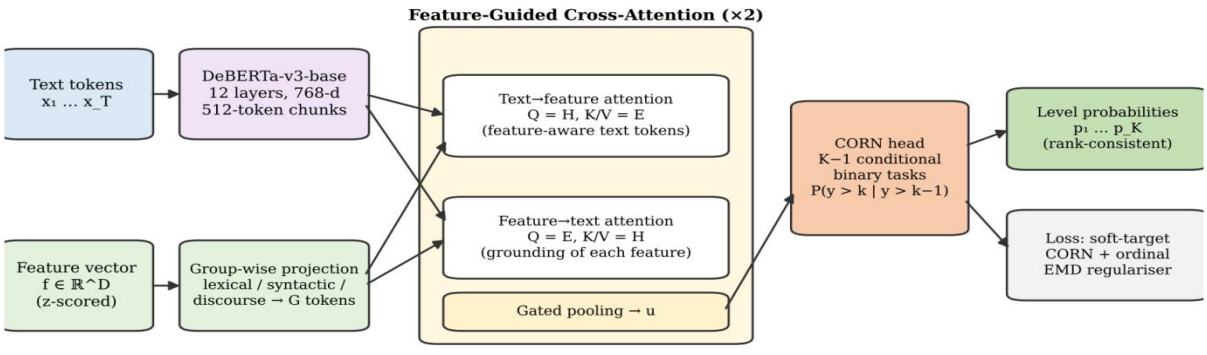


Figure 2. Architecture of the PsyCORT network with Feature-Guided Cross-Attention and CORN ordinal head

Figure 2 details the network. The two input streams are aligned in dimension before the FGCA block, in which the text-to-feature and feature-to-text attentions run in parallel within each layer. The gated pooling output feeds the CORN head, whose level probabilities are used both for the training loss and for conformal calibration.

3.6 Rank-consistent ordinal head

Instead of a softmax over six classes, PsyCORT uses the CORN formulation [12]. The head outputs $K - 1$ logits o_k, each modelling the conditional probability that the level exceeds k given that it exceeds $k - 1$, as in (9):

$$c_k = \sigma(o_k) = P(y > k \mid y > k - 1, u) \tag{9}$$

The unconditional exceedance probabilities follow from the chain rule, and the level probabilities from their differences, as in (10):

$$P(y > k) = \prod_{j=1}^{k} c_j, \quad p_k = P(y > k - 1) - P(y > k) \tag{10}$$

Because each factor lies in (0, 1), the exceedance probabilities are monotonically non-increasing in k, which guarantees rank consistency without constraining the weights as in CORAL [11].

3.7 Disagreement-aware soft ordinal targets

Training on a single label per text discards the information that two adjacent levels were both judged acceptable. PsyCORT therefore builds a soft target distribution t_i as a normalized mixture of discretized Gaussians centred on each acceptable level, as in (11):

$$t_{i,k} \propto \sum_{a \in \mathcal{A}_i} \exp\left(-\frac{(k-a)^2}{2\tau^2}\right) \tag{11}$$

with width $\tau = 0.6$, so that most mass lies on the acceptable levels and decays quickly with ordinal distance. The training loss combines cross-entropy against the soft target with an earth mover's distance between the predicted and target cumulative distributions, as in (12):

$$\mathcal{L} = -\sum_k t_{i,k} \log p_{i,k} + \lambda \sum_{k=1}^{K} \left(P_i(y \le k) - T_i(k)\right)^2 \tag{12}$$

where $T_i(k)$ is the cumulative target and $\lambda = 0.1$. The second term penalizes probability mass far from the acceptable levels more heavily than mass on neighbouring levels, which is the appropriate cost structure for CEFR errors.

3.8 Ambiguity-aware ordinal conformal prediction

After training, the model is frozen and a calibration set $\mathcal{D}_{cal} = \{(x_i, \mathcal{A}_i)\}$, disjoint from training and test data, is used to construct prediction sets. Following the ordinal conformal procedure of Lu et al. [21], for each input a sequence of nested contiguous intervals is grown from the most probable level: at each step the interval is extended towards the neighbouring level with the higher probability. The nonconformity score of level k is the cumulative probability mass of the smallest interval in this sequence that contains k, as in (13):

$$s(x, k) = \sum_{j \in I_k(x)} p_j(x) \tag{13}$$

where $I_k(x)$ is the first interval of the sequence that includes k. Standard conformal calibration [9], [30] uses the score of a single gold label. Because CEFR annotations are set-valued, PsyCORT instead uses the score of the most compatible acceptable level, as in (14):

$$r_i = min_{a \in \mathcal{A}_i}\, s(x_i, a) \tag{14}$$

The conformal threshold is the corrected empirical quantile of these calibration scores for a miscoverage level α, as in (15):

$$\hat{q} = Quantile\left(r_1, \ldots, r_n;\ \frac{\lceil (n+1)(1-\alpha) \rceil}{n}\right) \tag{15}$$

and the CEFR interval for a test text is given in (16):

$$C(x_{test}) = \{k :\ s(x_{test}, k) \le \hat{q}\} \tag{16}$$

Because the intervals are nested and grown from the mode, $C(x)$ is always a contiguous range of levels. If the calibration and test texts are exchangeable, the standard conformal argument [9], [30] applied to the scores r_i yields the finite-sample guarantee in (17):

$$P(C(x_{test}) \cap \mathcal{A}_{test} \neq \emptyset) \geq 1 - \alpha \tag{17}$$

Proposition 1 (ambiguity-aware validity). Let $(x_1, \mathcal{A}_1), \ldots, (x_{n+1}, \mathcal{A}_{n+1})$ be exchangeable, with the $(n + 1)$-th pair being the test text, and let s be fixed before calibration. Then the interval in (16) satisfies (17). *Proof.* Define r_{n+1} as in (14) for the test text. Because the intervals of the nested sequence are monotone, $k \in C(x_{n+1})$ if and only if $s(x_{n+1}, k) \leq \hat{q}$, and therefore $C(x_{n+1}) \cap \mathcal{A}_{n+1} \neq \emptyset$ if and only if $r_{n+1} \leq \hat{q}$. The scores $r_1, \ldots, r_{n+1}$ are exchangeable because they are the same fixed function of exchangeable pairs, so the rank of r_{n+1} among them is uniformly distributed (ties broken at random), and the corrected quantile in (15) gives $P(r_{n+1} \leq \hat{q}) \geq 1 - \alpha$ by the standard argument [9], [30].

That is, the interval contains at least one level that the annotators accepted with probability at least $1 - \alpha$. When every text has a single acceptable level, (17) reduces to the usual coverage guarantee. The ambiguity-aware rule changes the target of the guarantee rather than weakening it: it asks the interval to agree with the experts' range of acceptable judgements instead of with an arbitrary choice between two equally valid labels, and it therefore produces shorter intervals.

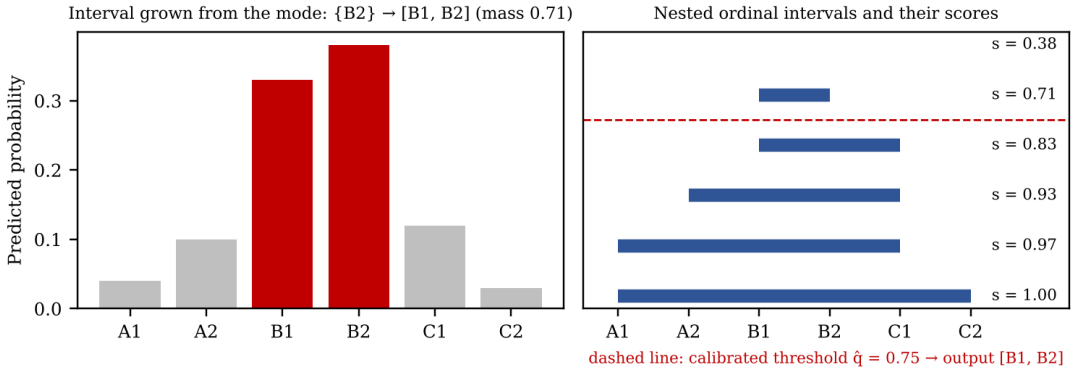


Figure 3. Construction of an ordinal conformal CEFR interval for an illustrative probability vector

Figure 3 illustrates the procedure for an illustrative probability vector peaked at B2. The nested intervals {B2}, [B1, B2], [B1, C1] and so on have cumulative masses 0.38, 0.71, 0.83 and higher. With a calibrated threshold of 0.75, levels whose scores do not exceed the threshold are retained, and the output is the interval [B1, B2].

3.9 Explanation of predictions

To explain which linguistic properties drive a predicted level, integrated gradients [31] are computed with respect to the expected level $\hat{y} = \Sigma_k k p_k$. For an input component z_j, either a standardized feature or a token embedding, with baseline z', the attribution is given in (18):

$$IG_j = \left(z_j - z_j^{'}\right) \int_0^1 \frac{\partial \hat{y}\left(z' + \beta(z - z')\right)}{\partial z_j} d\beta \tag{18}$$

For features, the baseline is the training mean (zero after standardization), so attributions express how much each property moves the predicted level relative to an average text; for tokens, the baseline is the padding embedding. Feature attributions are reported in teacher-facing form, for example "late-acquired vocabulary raises the predicted level by 0.6 levels", and token attributions highlight the words responsible. The feature-to-text attention of (7) links the two views.

3.10 Training procedure and class imbalance

CEFR corpora are imbalanced because very easy and very advanced texts are naturally scarce; in CEFR-SP, for example, only 771 sentences are labelled A1 [4]. To prevent the ordinal head from collapsing towards the central levels, each training text is weighted by the inverse square root of the frequency of its acceptable levels, as in (19):

$$\omega_i = \frac{1}{|\mathcal{A}_i|} \sum_{a \in \mathcal{A}_i} \left(\frac{N_a}{N/K}\right)^{-1/2} \tag{19}$$

where N_a is the number of training texts with level a among their acceptable levels and N the total number of texts. Square-root weighting moderates the influence of rare levels compared with full inverse-frequency weighting, which tends to destabilize fine-tuning on small classes. The weighted loss is averaged over mini-batches and optimized with AdamW; gradients are clipped at a norm of 1.0, and the checkpoint with the lowest validation loss is retained. For documents divided into chunks, the chunk representations are averaged before FGCA, and features are computed once for the whole document.

Algorithm 1. PsyCORT training, calibration and inference

Input: corpus $\{(x_i, \mathcal{A}_i)\}$, miscoverage α, encoder parameters θ

Output: CEFR interval $C(x)$ and explanation for each test text

1: Split texts into training, calibration and test partitions (text- and document-disjoint)

2: For each text: normalize, parse, extract features f (Eqs. 1–3); standardize with training statistics

3: Build soft targets t_i from $\mathcal{A}_i$ (Eq. 11)

4: **for** each epoch **do**

Encode tokens with DeBERTa-v3 (Eq. 4); form feature tokens (Eq. 5)

Apply two FGCA layers (Eqs. 6–7) and gated pooling (Eq. 8)

Compute CORN probabilities (Eqs. 9–10) and loss (Eq. 12); update θ with AdamW

Keep the checkpoint with the lowest validation loss

5: **end for**

6: For each calibration text, compute ordinal scores (Eq. 13) and r_i (Eq. 14); compute $\hat{q}$ (Eq. 15)

7: For each test text, output $C(x) = \{k : s(x, k) \leq \hat{q}\}$ (Eq. 16) and integrated-gradient explanations (Eq. 18)

Algorithm 1 summarizes the procedure. The calibration step is computationally negligible and can be repeated for any α without retraining, so institutions can choose their reliability level after deployment.

4. Results and Discussion

This section first describes the experimental configuration and evaluation metrics. It then reports a controlled simulation study that verifies the statistical behaviour of the conformal layer, which can be assessed independently of the encoder because conformal guarantees depend only on the exchangeability of calibration and test data. Finally, it specifies the hypotheses, ablations and success criteria for the transformer experiments on CEFR-SP and CLEAR-CEFR, which are in progress, and positions the study against published results.

4.1 Experimental setup

PsyCORT is implemented in Python 3.11 with PyTorch 2.3, Hugging Face Transformers and spaCy. Table 4 lists the configuration used for all experiments.

Table 4. Experimental setup and simulation parameters for PsyCORT

Parameter	**Setting**
Encoder	DeBERTa-v3-base (184 M parameters), 512-token chunks, stride 128
Linguistic features	19 features in 3 groups (lexical 8, syntactic 6, cohesion 5)
FGCA	2 layers, 8 heads, hidden size 768, dropout 0.1

Ordinal head	CORN, K − 1 = 5 conditional binary outputs
Soft targets / loss	Gaussian mixture, τ = 0.6; cross-entropy + EMD, λ = 0.1
Optimizer	AdamW [32], learning rate 2×10^{-5}, weight decay 0.01, one-cycle schedule, 10 % warm-up
Batch size / epochs	16 / 10, best checkpoint on validation loss
Data partition	60 % training (10 % of it for validation), 20 % calibration, 20 % test
Repetitions	5 random text-disjoint splits × 3 seeds
Conformal levels	α ∈ {0.05, 0.10, 0.15, 0.20, 0.30}; default α = 0.10
Explanations	Integrated gradients, 50 interpolation steps

Table 4 shows that the configuration follows standard practice for fine-tuning base-size encoders on datasets of this size. Allocating 20% of the data to calibration is larger than strictly necessary for marginal coverage but allows level-wise coverage to be estimated reliably for the rare A1 and C2 classes.

4.2 Evaluation metrics

Point-prediction quality is measured with quadratic weighted kappa (QWK) [33], mean absolute error (MAE) in levels, adjacent accuracy and macro-F1. QWK is defined in (20):

$$QWK = 1 - \frac{\sum_{j,k} w_{jk} O_{jk}}{\sum_{j,k} w_{jk} E_{jk}}, \quad w_{jk} = \frac{(j-k)^2}{(K-1)^2} \tag{20}$$

where O is the observed confusion matrix and E the matrix expected by chance from the marginals. Because CEFR-SP contains acceptable-label sets, two variants of accuracy are reported: exact accuracy against the first annotator and acceptable accuracy, which counts a prediction as correct if it lies in $\mathcal{A}_i$. Set-valued predictions are evaluated by gold coverage, acceptable coverage and mean interval size, as in (21):

$$Cov_{gold} = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}[y_i \in C(x_i)], \quad Cov_{acc} = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}[C(x_i) \cap \mathcal{A}_i \neq \emptyset] \tag{21}$$

together with level-wise coverage, which reveals whether rare levels are systematically under-covered, and the proportion of contiguous sets, since a non-contiguous set such as {A2, C1} is uninterpretable for teachers.

4.3 Simulation design for the conformal layer

To verify the conformal layer under conditions that resemble CEFR annotation, a generative simulation was constructed. Each simulated text has a latent difficulty $z \sim N(0, 1.25^2)$, and the six levels are defined by thresholds at −2, −1, 0, 1 and 2. Two annotators observe z with independent Gaussian perception noise (standard deviation 0.68) and assign levels; this noise level reproduces the agreement structure of CEFR-SP, giving 41.3% exact agreement and 88.2% agreement within one level, compared with 37.6% and 88.4% in the real corpus [4]. Acceptable-label sets were constructed with the CEFR-SP convention. A deliberately imperfect model observes z with additional noise (standard deviation 0.45) and outputs cumulative-logit probabilities. Two scenarios were examined: an ordinal head, whose probabilities are unimodal by construction, and a nominal head, obtained by multiplying the probabilities by Gamma(2, 0.5) noise and renormalizing, which mimics a softmax classifier whose distributions can be multimodal. Each repetition used 2,000 calibration and 5,000 test texts, and results were averaged over 200 repetitions (60 for the α sweep). Four set-valued predictors were compared, all calibrated to α = 0.10: LAC [19], APS [20], ordinal APS [21] calibrated on the first annotator's label, and ordinal APS with the proposed ambiguity-aware calibration. Point prediction by the most probable level is included for reference.

4.4 Simulation results

Table 5. Conformal prediction performance at α = 0.10 (mean ± SD over 200 repetitions)

Scenario / method	Gold coverage	Acceptable coverage	Mean set size	Contiguous (%)

Ordinal head				
Point prediction (argmax)	0.459 ± 0.006	0.747 ± 0.006	1.00	100.0
LAC	0.901 ± 0.007	0.981 ± 0.003	2.66 ± 0.05	100.0
APS	0.901 ± 0.007	0.975 ± 0.003	2.79 ± 0.05	100.0
Ordinal APS	0.901 ± 0.007	0.970 ± 0.003	2.96 ± 0.06	100.0
Ordinal APS, ambiguity-aware (proposed)	0.760 ± 0.017	0.901 ± 0.008	2.17 ± 0.09	100.0
Nominal head				
Point prediction (argmax)	0.398 ± 0.007	0.653 ± 0.007	1.00	100.0
LAC	0.900 ± 0.007	0.970 ± 0.003	3.12 ± 0.05	93.5
APS	0.900 ± 0.007	0.970 ± 0.003	3.19 ± 0.05	90.4
Ordinal APS	0.900 ± 0.007	0.967 ± 0.003	3.08 ± 0.05	100.0
Ordinal APS, ambiguity-aware (proposed)	0.772 ± 0.012	0.901 ± 0.007	2.34 ± 0.05	100.0

Table 5 shows that every conformal method attained its target: methods calibrated on a single gold label reached 90.0–90.1% gold coverage, and the ambiguity-aware method reached 90.1% acceptable coverage, matching the nominal $1 - \alpha$ in both scenarios. Point prediction, by contrast, was correct for only 45.9% (ordinal head) and 39.8% (nominal head) of texts against a single annotator, and for 74.7% and 65.3% against the acceptable sets, which illustrates how much uncertainty a single predicted level conceals when annotators themselves disagree. With the ordinal head, LAC produced the smallest gold-calibrated sets (2.66 levels), and all methods were contiguous because the probabilities were unimodal. With the nominal head, ordinal APS produced the smallest gold-calibrated sets (3.08 levels) and was the only gold-calibrated method that was always contiguous; 6.5% of LAC sets and 9.6% of APS sets were non-contiguous.

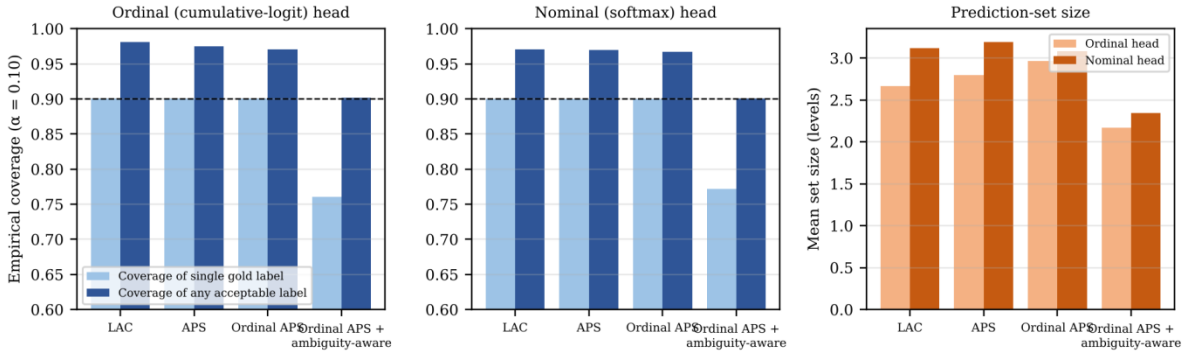


Figure 4. Coverage and set size of conformal CEFR predictors at α = 0.10 for ordinal and nominal base models

Figure 4 visualizes these results. The ambiguity-aware calibration trades gold coverage (0.76–0.77) for acceptable coverage at exactly the target level and reduces the mean interval size by 27% with the ordinal head (2.17 versus 2.96 levels for ordinal APS) and by 24% with the nominal head (2.34 versus 3.08). This trade is intended: when two adjacent levels are equally defensible, requiring the interval to contain one arbitrarily chosen annotator's label forces it to widen without adding pedagogically useful information.

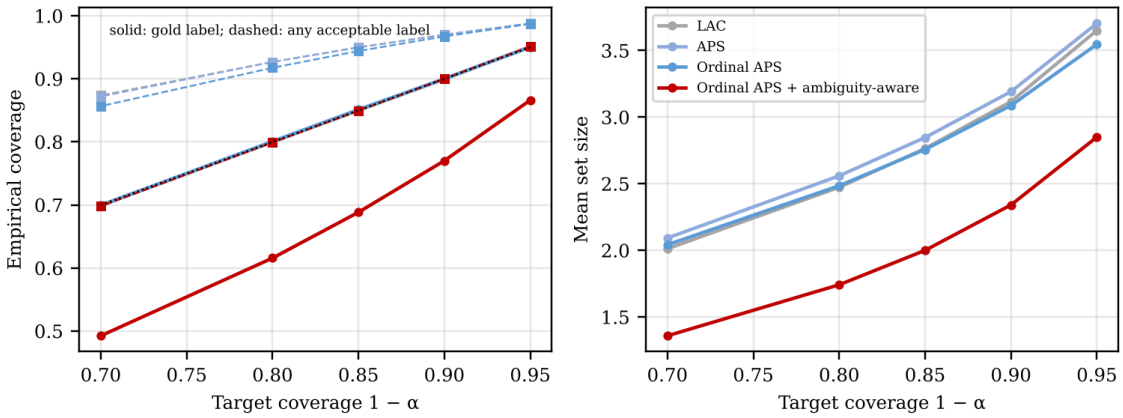


Figure 5. Empirical coverage and mean set size across target coverage levels (nominal base model)

Table 6. Mean set size across target coverage levels (nominal head; coverage in parentheses)

Method	**1 − α = 0.70**	**0.80**	**0.85**	**0.90**	**0.95**
LAC	2.01 (0.700)	2.47 (0.800)	2.76 (0.850)	3.12 (0.900)	3.65 (0.951)
APS	2.09 (0.699)	2.56 (0.801)	2.85 (0.851)	3.19 (0.900)	3.70 (0.951)
Ordinal APS	2.04 (0.699)	2.48 (0.801)	2.76 (0.851)	3.09 (0.900)	3.55 (0.950)
Ambiguity-aware (acceptable cov.)	1.36 (0.698)	1.74 (0.799)	2.00 (0.849)	2.34 (0.900)	2.85 (0.952)

Figure 5 and Table 6 show that coverage tracked the target closely at every level, with absolute deviations of at most 0.002, confirming finite-sample validity. Interval size grew smoothly with the required coverage; at 1 − α = 0.80 the ambiguity-aware predictor returned 1.74 levels on average, meaning that most texts receive a single level or a pair of adjacent levels, whereas at 0.95 it returned 2.85 levels. These curves allow a programme to choose its operating point explicitly, trading precision against reliability.

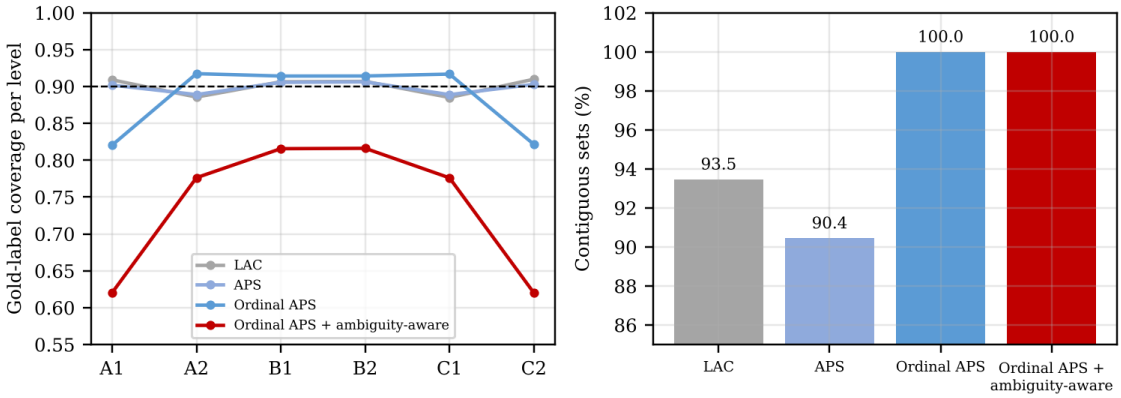


Figure 6. Level-wise gold coverage and proportion of contiguous prediction sets (nominal base model, α = 0.10)

Figure 6 reveals a limitation that marginal coverage hides. Ordinal APS covered the extreme levels A1 and C2 at only 82.0% and 82.1%, below the 90% target, while over-covering the central levels (91.4–91.7%); LAC and APS were more uniform (88.5–91.0%). Under the ambiguity-aware rule, gold coverage at A1 and C2 fell to 62.0%, partly because acceptable sets at the extremes frequently contain an adjacent level. This pattern matters because A1 and C2 texts are scarce in CEFR corpora [4] and are exactly where teachers most need reliable information. Level-conditional (Mondrian) calibration, which computes a separate threshold for each level, is evaluated as a remedy in Section 4.5.

4.5 Calibration size, base-model quality and level-conditional calibration

Three further simulation analyses address practical questions for deployment, all with the nominal base model and α = 0.10. First, the calibration set size was varied from 100 to 2,000 texts over 300 repetitions with 3,000 test texts. Second, the noise of the base model was varied from 0.15 to 0.80, emulating encoders of different quality, over 60 repetitions. Third, level-conditional (Mondrian) calibration, which computes a separate threshold for each level from the calibration texts of that level, was compared with marginal calibration over 150 repetitions.

Table 7. Effect of calibration size, base-model quality and level-conditional calibration (α = 0.10)

Analysis	Setting	Gold coverage	Interval size	Other
Calibration size	n = 100	0.912 (5th–95th: 0.854–0.952)	3.19	—
	n = 250	0.905 (0.870–0.935)	3.13	—
	n = 500	0.903 (0.880–0.926)	3.11	—
	n = 1,000	0.902 (0.883–0.919)	3.10	—
	n = 2,000	0.901 (0.886–0.916)	3.09	—
Base-model noise	SD = 0.15	0.90	2.84 / 2.04*	Acceptable accuracy 70.2%
	SD = 0.45	0.90	3.09 / 2.34*	Acceptable accuracy 65.3%
	SD = 0.80	0.90	3.52 / 2.90*	Acceptable accuracy 57.1%
Calibration type	Marginal	0.900 (A1 0.822, C2 0.820)	3.09	Contiguous 100%
	Level-conditional	0.904 (all levels 0.902–0.910)	3.19	Contiguous 99.4%

*Gold-calibrated ordinal APS / ambiguity-aware calibration.

Table 7 shows that average coverage remained close to the target for every calibration size, but its variability across random calibration sets decreased markedly: with 100 calibration texts the realized coverage fell below 85.4% in 5% of repetitions, whereas with 2,000 texts it stayed above 88.6%. For CLEAR-CEFR, where a 20% calibration split contains about 236 texts, coverage can therefore be expected to fluctuate by about ±2–3 percentage points between splits, which should be reported alongside the mean.

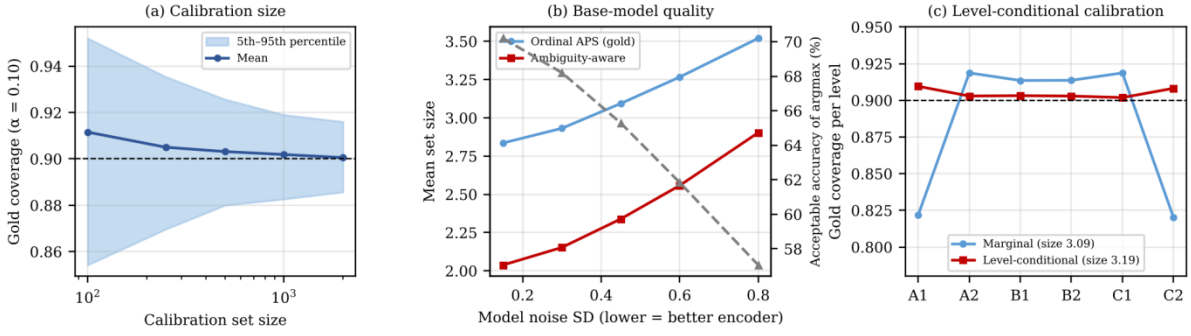


Figure 7. Effect of calibration size, base-model quality and level-conditional calibration on conformal CEFR intervals

Figure 7(b) shows that a better base model shortens the intervals, from 3.52 to 2.84 levels for gold calibration and from 2.90 to 2.04 levels for ambiguity-aware calibration as model noise decreases from 0.80 to 0.15. The

improvement flattens, however, because even an almost noise-free model reached only 70.2% acceptable accuracy: the remaining uncertainty stems from the annotators rather than the model. This observation sets a realistic expectation for the transformer experiments; gains in the encoder will narrow intervals, but annotation ambiguity imposes a floor that only better-agreed labels can lower. Figure 7(c) shows that level-conditional calibration removed the under-coverage of A1 and C2, raising both from about 82% to about 91%, at the cost of a 3.3% larger average interval (3.19 versus 3.09 levels) and a small loss of contiguity (99.4%). Level-conditional calibration is therefore recommended when reliable coverage of beginner and advanced material is required.

Computational cost. Table 8 summarizes the size and cost of each component, computed from the configuration in Table 4. The FGCA layers add about 28 million parameters to the 184-million-parameter encoder, and the conformal layer requires only sorting the $K = 6$ level probabilities of each text, so calibration for a new α takes milliseconds even for tens of thousands of texts.

Table 8. Parameter count and computational cost of PsyCORT components

Component	Parameters	Cost per text	Remarks
DeBERTa-v3-base encoder	≈ 184 M	$O(T^2)$ per 512-token chunk	Dominant cost
Feature projections (3 groups)	≈ 0.02 M	Negligible	19 inputs → 3 × 768
FGCA (2 layers)	≈ 28.4 M	$O(T \cdot G)$, $G = 3$	Cross-attention to 3 feature tokens is cheap
Gate and CORN head	≈ 0.005 M	Negligible	$K - 1 = 5$ outputs
Feature extraction (spaCy parse)	—	Linear in text length	CPU, once per text
Ordinal conformal interval	—	O(K) after probabilities	Recalibration without retraining

Table 8 shows that PsyCORT adds about 15% parameters to the encoder and essentially no inference overhead for the uncertainty layer, which makes it practical for large digital libraries where texts must be leveled continuously.

4.6 Evaluation protocol and hypotheses for the transformer experiments

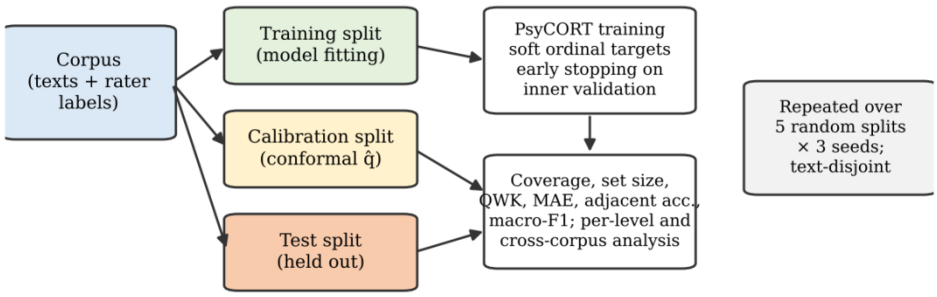


Figure 8. Evaluation protocol for PsyCORT on CEFR-SP and CLEAR-CEFR

Figure 8 outlines the protocol. Each corpus is split into training, calibration and test partitions that are disjoint at the text level and, where possible, at the source-document level. The model is trained with early stopping on an inner validation subset, the conformal threshold is computed on the calibration partition, and all metrics are computed on the test partition. The procedure is repeated over five random splits and three seeds. For cross-corpus evaluation, models trained on CEFR-SP or CLEAR-CEFR are tested on OneStopEnglish after mapping predictions to the broad A, B and C bands, and conformal coverage is re-examined to quantify the effect of distribution shift.

Table 9. Hypotheses, experiments and success criteria

Hypothesis	Experiment	Success criterion
H1: Linguistic features improve contextual models	DeBERTa-v3 alone vs. concatenation vs. FGCA	FGCA gives the highest QWK and lowest MAE; paired bootstrap $p < 0.05$
H2: Ordinal modelling reduces distant errors	Softmax vs. CORAL vs. CORN head	Lower MAE and fewer errors of two or more levels with CORN
H3: Soft targets improve acceptable accuracy	Single-label vs. disagreement-aware targets	Higher acceptable accuracy and lower calibration error
H4: Conformal intervals are valid on real data	Coverage on held-out CEFR-SP and CLEAR-CEFR	$\lvert\text{coverage} - (1 - \alpha)\rvert \leq 0.02$ for all α
H5: Ambiguity-aware calibration shortens intervals	Gold-label vs. ambiguity-aware calibration on CEFR-SP	Smaller mean interval at equal acceptable coverage
H6: Improvement over feature-based CEFR models	PsyCORT vs. random forest with the same 19 features on CLEAR-CEFR	Higher exact accuracy than the 62.6% reported in [3] under comparable splits

Table 9 states the hypotheses. H4 and H5 extend the simulation results to real data, where exchangeability may be violated by topical clustering; H6 provides an external reference point on the same corpus as [3], although differences in data splits must be taken into account when comparing figures.

4.7 Planned ablation study

Table 10. Ablation configurations

Configuration	Component varied	Question addressed
Features only (random forest, ordinal logistic)	No encoder	Strength of interpretable baselines
DeBERTa-v3 + softmax	No features, nominal head	Contextual baseline
DeBERTa-v3 + features (concatenation) + softmax	Fusion strategy	Value of simple hybridization [6]
DeBERTa-v3 + FGCA + softmax	Fusion strategy	Value of cross-attention fusion (H1)
DeBERTa-v3 + FGCA + CORAL / CORN	Ordinal head	Value of rank consistency (H2)
… + soft targets (± EMD term)	Training targets	Value of disagreement-aware training (H3)
Feature groups removed one at a time	Lexical / syntactic / cohesion	Contribution of each linguistic dimension
Conformal: LAC, APS, ordinal APS, ambiguity-aware, Mondrian	Set construction	Validity, size, contiguity, level-wise coverage (H4, H5)

Table 10 lists the ablations. All configurations share data splits, seeds and training budget, so differences can be attributed to the varied component, and the conformal variants are applied to the same trained models so that they differ only in set construction.

4.8 Comparison with published results

Table 11. Published results on the corpora used in this study

Study	Corpus	Model	Reported result	Uncertainty output
Arase et al. [4]	CEFR-SP	Transformer-based assessment model	Macro-F1 84.5%	None
Zhang & Lu [3]	CLEAR-CEFR, 6 levels	Random forest, 24 features	Exact accuracy 62.6% (training CV); 60.3% test, $\kappa = 0.496$	None
Zhang & Lu [3]	CLEAR-CEFR, 3 levels	Random forest, 24 features	Accuracy 82.6%	None
PsyCORT (this work)	CEFR-SP, CLEAR-CEFR	DeBERTa-v3 + FGCA + CORN + conformal	Experiments in progress (Table 9)	Contiguous intervals with guaranteed coverage

Table 11 reports the published results on the two primary corpora. The values were obtained by the original authors under their own protocols. Beyond the expected gains in accuracy, the distinctive contribution of PsyCORT is the last column: existing systems return a single level without any statement of reliability, whereas PsyCORT returns an interval whose coverage is guaranteed and whose size reflects the difficulty of the individual text.

4.9 Discussion

The simulation results clarify what conformal CEFR intervals can and cannot provide. When annotators agree exactly on fewer than half of the texts, as in CEFR-SP, no point predictor can be correct for most texts against a single annotator; the simulated ordinal model, whose underlying difficulty estimate was reasonably accurate, matched one annotator for only 46% of texts. Interval predictions make this uncertainty visible and controllable. The intervals required for 90% coverage were between two and three levels wide in the simulation, which is wide for fine-grained placement but appropriate for selecting texts within a band; with an encoder that reduces model noise, intervals will narrow, and the conformal guarantee will continue to hold.

For English language teaching, the ambiguity-aware formulation is the most consequential design choice. Experts disagree about CEFR levels because adjacent levels overlap in their descriptors [1], and a system that demands agreement with one arbitrary rater penalizes legitimate judgements. Calibrating against the set of acceptable levels yields intervals that are, on average, a quarter shorter while matching the experts' own range of judgements at the requested rate. The level-wise analysis, however, warns that marginal guarantees can hide systematic under-coverage of A1 and C2 texts; deployments that serve beginners or very advanced learners should use level-conditional calibration.

The linguistic components serve interpretability as well as accuracy. Prior CEFR research identifies age of acquisition, lexical decision times, complex nominals and T-unit length as key predictors [3], and FGCA allows the network to condition its text representation on exactly these properties. Integrated-gradient attributions then convert the prediction into statements that teachers can act on when adapting a text, such as replacing late-acquired vocabulary or splitting long T-units.

Three use cases illustrate how the outputs can support teaching. In materials selection, a teacher searching for B1 texts can request only texts whose 90% interval lies within [A2, B1] or [B1, B2], excluding texts whose difficulty is genuinely uncertain. In curriculum auditing, the interval width itself is informative: a reading list in which many texts receive wide intervals signals material that experts would also find hard to place and that may benefit from adaptation. In text simplification, the feature attributions indicate which properties to modify, and the conformal interval of the simplified version shows whether the intended level has been reached with known reliability.

Ethical considerations also apply. CEFR levels influence which texts learners encounter, and systematic misplacement could restrict access to appropriately challenging material for particular groups. Coverage

guarantees reduce the risk of overconfident errors but do not guarantee fairness across genres, varieties of English or topics. Audits of interval width and coverage by genre and source, and human review of texts with wide intervals, are therefore recommended before deployment in high-stakes settings.

Threats to validity

Internal validity is threatened by leakage between partitions, for example near-duplicate sentences or excerpts from the same book in training and test data; text- and document-disjoint splitting and duplicate detection are used to mitigate this. Construct validity depends on whether CEFR labels assigned to texts reflect the comprehension difficulty experienced by learners; the labels in both corpora come from experts applying CEFR descriptors rather than from learner performance, which is a recognized limitation of text-level CEFR research [3]. Statistical conclusion validity is supported by repeated splits, multiple seeds and paired tests, and by reporting the variability of conformal coverage rather than only its mean. External validity is addressed by the cross-corpus evaluation on OneStopEnglish but remains limited to written English texts of the genres represented in the corpora.

Several limitations must be acknowledged. The conformal layer was validated by simulation, and its behaviour on real CEFR data, where texts from the same source may be clustered and the test distribution may differ from calibration data, is the subject of hypotheses H4 and H5. Coverage guarantees are marginal and do not hold conditionally on individual texts or subgroups unless level-conditional calibration is used. The corpora contain predominantly texts written for or by native speakers, and CEFR-SP labels sentences in isolation, which may not reflect the difficulty of connected discourse. Finally, the psycholinguistic norms are based on native-speaker ratings and only approximate learners' experience of words.

5. Conclusion

This study proposed PsyCORT, a psycholinguistically grounded conformal ordinal transformer for CEFR-level assessment of English reading texts. The framework fuses DeBERTa-v3 representations with lexical-sophistication, syntactic-complexity and cohesion features through Feature-Guided Cross-Attention, predicts rank-consistent level probabilities with a CORN head trained on disagreement-aware soft targets, and outputs contiguous CEFR intervals with distribution-free coverage guarantees. A new ambiguity-aware calibration rule guarantees that intervals contain at least one level accepted by the annotators. In a simulation calibrated to the inter-annotator agreement of CEFR-SP, all conformal predictors achieved their target coverage within 0.002; ordinal conformal sets were always contiguous, whereas 6.5–9.6% of standard sets were not when the base model was nominal; and ambiguity-aware calibration delivered exactly 90.1% acceptable coverage with intervals 24–27% shorter. The analysis also exposed under-coverage of the extreme levels A1 and C2, which level-conditional calibration removed at a cost of 3.3% wider intervals, and showed that annotation ambiguity, not only model quality, limits how narrow valid intervals can become.

Future work will complete the transformer experiments on CEFR-SP and CLEAR-CEFR following the protocol and hypotheses defined here, evaluate robustness to distribution shift across corpora and genres, and investigate conformal methods that remain valid under such shift. Extending the framework to learner-produced writing, where CEFR levels describe proficiency rather than text difficulty, and integrating the predicted intervals into adaptive reading platforms and readability-controlled text simplification are further promising directions.

Data availability

CEFR-SP is available from the authors of [4], the CEFR-labelled CLEAR subset is available as open data with [3], the CLEAR corpus is described in [2], and OneStopEnglish is distributed with [15]. Psycholinguistic norms are available from the sources cited in Table 3.

Code availability

The implementation of PsyCORT, the conformal layer and the complete simulation study, including all scripts needed to reproduce Tables 5–7 and Figures 4–7, will be released in a public repository upon publication.

Conflict of interest

The authors declare no conflict of interest.

Appendix A. Simulation data-generating process

For reproducibility, the simulation of Sections 4.3–4.5 is specified as follows. For each text, a latent difficulty $z \sim N(0, 1.25^2)$ is drawn and mapped to a level by the thresholds $(-2, -1, 0, 1, 2)$. Each of two annotators observes $z + \varepsilon_a$ with $\varepsilon_a \sim N(0, 0.68^2)$ and assigns the corresponding level. The base model observes $\hat{z} = z + \varepsilon_m$ with $\varepsilon_m \sim$

$N(0, 0.45^2)$ unless stated otherwise and outputs cumulative probabilities $P(y \le k) = \sigma((\theta_k - \hat{z})/0.55)$ for thresholds θ_k, from which level probabilities are obtained by differencing. In the nominal scenario, each level probability is multiplied by an independent Gamma(2, 0.5) variate and renormalized, which preserves the expected ranking of levels but allows multimodal distributions. Acceptable sets contain all levels between the two annotators' labels. The first annotator's label serves as the single gold label for gold-calibrated methods. APS uses randomized scores; LAC uses one minus the probability of the label; ordinal APS uses the nested-interval score of (13). All random seeds are fixed in the released scripts.

Appendix B. Computation of ordinal nonconformity scores

Algorithm 2. Ordinal nonconformity scores for one text

Input: level probabilities $p_1, \ldots, p_K$

Output: scores s(k) for k = 1, …, K

1: lo ← hi ← $\text{argmax}_k p_k$; m ← p_{lo}; s(lo) ← m

2: **while** lo > 1 **or** hi < K **do**

left ← p_{lo-1} if lo > 1 else −1; right ← p_{hi+1} if hi < K else −1

if left ≥ right **then** lo ← lo − 1; m ← m + p_{lo}; s(lo) ← m

else hi ← hi + 1; m ← m + p_{hi}; s(hi) ← m

3: **end while**

4: **return** s (the interval $\{k : s(k) \le \hat{q}\}$ is contiguous because scores increase along the growth order)

References

[1] Council of Europe, Common European Framework of Reference for Languages: Learning, Teaching, Assessment – Companion Volume. Strasbourg: Council of Europe Publishing, 2020.
[2] S. A. Crossley, A. Heintz, J. S. Choi, J. Batchelor, M. Karimi, and A. Malatinszky, "A large-scaled corpus for assessing text readability," Behavior Research Methods, vol. 55, pp. 491–507, 2023.
[3] X. Zhang and X. Lu, "Aligning linguistic complexity with the difficulty of English texts for L2 learners based on CEFR levels," Studies in Second Language Acquisition, vol. 47, no. 5, pp. 1407–1434, 2025, doi: 10.1017/S0272263125101125.
[4] Y. Arase, S. Uchida, and T. Kajiwara, "CEFR-based sentence difficulty annotation and assessment," in Proc. 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2022, pp. 6206–6219.
[5] M. Xia, E. Kochmar, and T. Briscoe, "Text readability assessment for second language learners," in Proc. 11th Workshop on Innovative Use of NLP for Building Educational Applications (BEA), 2016, pp. 12–22.
[6] B. W. Lee, Y. S. Jang, and J. H.-J. Lee, "Pushing on text readability assessment: a transformer meets handcrafted linguistic features," in Proc. 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2021, pp. 10669–10686.
[7] M. Martinc, S. Pollak, and M. Robnik-Šikonja, "Supervised and unsupervised neural approaches to text readability," Computational Linguistics, vol. 47, no. 1, pp. 141–179, 2021.
[8] J. M. Imperial et al., "UniversalCEFR: enabling open multilingual research on language proficiency assessment," arXiv:2506.01419, 2025.
[9] A. N. Angelopoulos and S. Bates, "Conformal prediction: a gentle introduction," Foundations and Trends in Machine Learning, vol. 16, no. 4, pp. 494–591, 2023.
[10] V. Kuperman, H. Stadthagen-Gonzalez, and M. Brysbaert, "Age-of-acquisition ratings for 30,000 English words," Behavior Research Methods, vol. 44, no. 4, pp. 978–990, 2012.
[11] W. Cao, V. Mirjalili, and S. Raschka, "Rank consistent ordinal regression for neural networks with application to age estimation," Pattern Recognition Letters, vol. 140, pp. 325–331, 2020.
[12] X. Shi, W. Cao, and S. Raschka, "Deep neural networks for rank-consistent ordinal regression based on conditional probabilities," Pattern Analysis and Applications, vol. 26, pp. 941–955, 2023.
[13] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. 34th International Conference on Machine Learning (ICML), 2017, pp. 1321–1330.
[14] S. Vajjala and D. Meurers, "On improving the accuracy of readability classification using insights from second language acquisition," in Proc. 7th Workshop on Building Educational Applications Using NLP, 2012, pp. 163–173.
[15] S. Vajjala and I. Lučić, "OneStopEnglish corpus: a new corpus for automatic readability assessment and text simplification," in Proc. 13th Workshop on Innovative Use of NLP for Building Educational Applications (BEA), 2018, pp. 297–304.

[16] T. Deutsch, M. Jasbi, and S. Shieber, “Linguistic features for readability assessment,” in Proc. 15th Workshop on Innovative Use of NLP for Building Educational Applications (BEA), 2020, pp. 1–17.
[17] S. A. Crossley, D. B. Greenfield, and D. S. McNamara, “Assessing text readability using cognitively based indices,” TESOL Quarterly, vol. 42, no. 3, pp. 475–493, 2008.
[18] K. Collins-Thompson, “Computational assessment of text readability: a survey of current and future research,” ITL – International Journal of Applied Linguistics, vol. 165, no. 2, pp. 97–135, 2014.
[19] M. Sadinle, J. Lei, and L. Wasserman, “Least ambiguous set-valued classifiers with bounded error levels,” Journal of the American Statistical Association, vol. 114, no. 525, pp. 223–234, 2019.
[20] Y. Romano, M. Sesia, and E. J. Candès, “Classification with valid and adaptive coverage,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020, pp. 3581–3591.
[21] C. Lu, A. N. Angelopoulos, and S. Pomerantz, “Improving trustworthiness of AI disease severity rating in medical imaging with ordinal conformal prediction sets,” in Medical Image Computing and Computer Assisted Intervention (MICCAI 2022), Lecture Notes in Computer Science, vol. 13438, 2022, pp. 545–554.
[22] P. He, J. Gao, and W. Chen, “DeBERTaV3: improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing,” in Proc. International Conference on Learning Representations (ICLR), 2023.
[23] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, “spaCy: industrial-strength natural language processing in Python,” Zenodo, 2020, doi: 10.5281/zenodo.1212303.
[24] M. Brysbaert, A. B. Warriner, and V. Kuperman, “Concreteness ratings for 40 thousand generally known English word lemmas,” Behavior Research Methods, vol. 46, no. 3, pp. 904–911, 2014.
[25] M. Brysbaert and B. New, “Moving beyond Kučera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English,” Behavior Research Methods, vol. 41, no. 4, pp. 977–990, 2009.
[26] X. Lu, “Automatic analysis of syntactic complexity in second language writing,” International Journal of Corpus Linguistics, vol. 15, no. 4, pp. 474–496, 2010.
[27] H. Liu, “Dependency distance as a metric of language comprehension difficulty,” Journal of Cognitive Science, vol. 9, no. 2, pp. 159–191, 2008.
[28] S. A. Crossley, K. Kyle, and D. S. McNamara, “The tool for the automatic analysis of text cohesion (TAACO): automatic assessment of local, global, and text cohesion,” Behavior Research Methods, vol. 48, no. 4, pp. 1227–1237, 2016.
[29] A. Vaswani et al., “Attention is all you need,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017, pp. 5998–6008.
[30] V. Vovk, A. Gammerman, and G. Shafer, Algorithmic Learning in a Random World. New York, NY, USA: Springer, 2005.
[31] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,” in Proc. 34th International Conference on Machine Learning (ICML), 2017, pp. 3319–3328.
[32] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in Proc. International Conference on Learning Representations (ICLR), 2019.
[33] J. Cohen, “Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit,” Psychological Bulletin, vol. 70, no. 4, pp. 213–220, 1968.