ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 3 - 2026

Research on English reading comprehension level test and personalized recommendation based on Transformer

Hong Duan[1,*]

[1] Foreign Languages School of Xinxiang Institute of Engineering Xinxiang453000, Henan, China

SUMMARY: *In the globalization background, the requirement for intelligent evaluation and customized enhancement of English reading comprehension—a core ability for cross-cultural communication—has been becoming more and more pressing. Traditional assessment approaches, which depend on manual score-giving and unchangeable question collections, have such drawbacks as strong subjectivity, feedback with delay, and trouble in catching abilities of deep understanding. Moreover, their single-size-suits-all method cannot satisfy personalized learning demands. Transformer models, making use of self-attention mechanisms and global information modeling abilities, thus provide important technical support for construction of integrated assessment-recommendation systems. In order to deal with existing research gaps, this study carries out synthesis of relevant theories and makes use of experimental design and data analysis methods, therefore putting forward a Transformer-based Collaborative Attention (TBC) framework. This system framework takes in pre-trained models, constructs a multi-dimensional examination question database, and at the same time builds a user behavior data collection set. Through doing experiments, we can show that this model gets an evaluation accuracy number of 92.7% and an F1 score of 0.892. Therefore, for long-text processing work, its accuracy only has a 3.2 percentage point decrease thus. The personalized recommendation system carries out analysis of user behavior via multimodal feature fusion, and thus reaches a recommendation click-through rate of 78.3% as well as a matching accuracy of 89.4%. Compared with the control group, user scores have raised up by 18.7 points. Furthermore, the system can correctly recognize knowledge gaps for 83% of participants. This research carries out validation of the model's effectiveness in the test of English reading comprehension ability and personalized recommendation, therefore hence offering new paths for the personalization of intelligent education. Future work may optimize modeling for low-frequency users and expand multi-source corpora.*

KEYWORDS: *Transformer model; English reading comprehension; Level test; Personalized recommendation; Self attention mechanism; Multimodal feature fusion*

1 Introduction

1.1 Background and Significance of the Study

In the context of globalization, English reading comprehension, as a core skill of cross-cultural communication, has become a key indicator of an individual's international

*15936509083@163.com
https://doi.org/10.65102/is20261123 Published: 10 June, 2026

competitiveness. With the breakthrough of artificial intelligence technology, the limitations of traditional manual scoring in English reading tests are becoming more and more obvious. Subjective judgment is difficult to quantify deep-level comprehension and cannot meet the demand for real-time feedback in large-scale scenarios. Existing studies have pointed out that most models suffer from the problem of semantic information loss when dealing with this task, which directly affects the objectivity of the assessment [1]. At the same time, the contradiction between the growing demand for personalized learning and the traditional "one-size-fits-all" resource allocation model is becoming more and more prominent. It has become an important way to improve the efficiency of English teaching by integrating machine learning and natural language processing technologies to build intelligent assessment and recommendation systems.

The pre-trained language model based on the Transformer architecture realizes dynamic modeling of global text semantics through the self-attention mechanism. Its end-to-end design is more capable of capturing contextual relevance than traditional phased models, and its performance on standard datasets such as SQuAD validates the reliability of the technique [2]. SpanBERT model achieves more than 92% of the F1 score on SQuAD1.1 dataset by a multilayered feature fusion strategy, which provides a technical basis for a high-precision assessment system. In terms of personalized recommendation, its sequence modeling In terms of personalized recommendation, its sequence modeling capability is able to parse learners' learning behavioral characteristics and cognitive patterns, construct multidimensional learning profiles, and adapt to the dynamics and complexity of English learning.

Despite the progress of existing research in areas such as classical Arabic texts [3] and interactive reading tasks for English reading comprehension assessment with automated Transformer-based topic generation [4], there is still a gap in terms of systematic application in English education scenarios.

This study combines the Transformer model with the requirements of educational assessment to construct an intelligent system that integrates assessment, diagnosis and recommendation. It carries out precise grade testing via dynamic difficulty adjustment and provides personalized study materials according to real-time feedback. Therefore, the system not only breaks through time and space limitations of traditional testing, and supports immediate feedback as well as process-based assessment, hence it also enhances resource application efficiency and reduces cognitive burden. It supplies educational administrative personnel with a decision-making foundation and thus pushes forward the shift of English language teaching toward a data-driven orientation. This research must center on ethics-related matters like educational fairness and privacy guard. Hence, its outcomes are anticipated not only to transform the pattern of English reading education, but thus also supply a transferable methodological framework for the augmentation of education through artificial intelligence.

1.2 Research Methods and Innovations

This paper uses experimental design, model building and data analysis as core research approaches, with the purpose to guarantee this study's scientific nature and effectiveness. Therefore, in the aspect of model construction, for the first time, the Transformer architecture is introduced into English reading comprehension assessment and personalized recommendation; thus, it breaks through the limitations that traditional methods have in semantic understanding and contextual relation analysis hence. By making use of the synergetic attention mechanism, we put forward the Transformer-based synergetic attention (TBC) framework and integrate it into the pre-trained model via a hybrid scheme; hence, this action obviously lifts the model's performance [5] and offers technical backing for precise

assessment work. At the data layer, we have built a test question bank and a user learning behavior dataset that include multi-dimensional features like text difficulty grading and knowledge graph mapping. Through expert focus group discussions, we determined relevant criteria; hence, we guaranteed data quality and validity by adding cognitive modeling analysis [6]. A comparison model between control group and experimental group was adopted to carry out the experiments. Thus, the model's performance was verified by using standardized datasets like SQuAD and DROP, and meanwhile, dynamic testing within real educational scenarios was combined together to therefore guarantee the practical validity of the experimental results.

The innovative aspects of this study are threefold: first, this is the first time that the Transformer model has been applied to the dynamic assessment of English reading comprehension levels. Through the mechanism of synergistic attention, it captures the interactive features between text and questions, overcoming the shortcomings of traditional methods in long text comprehension and implicit logical reasoning. Second, a hybrid assessment system incorporating automatic scoring and expert review was constructed. Based on the TACM cognitive model, it ensures the reliability of assessment. Meanwhile, it combines ability diagnosis with resource matching, and achieves precise knowledge point positioning and learning path optimization by analyzing users' question-answering trajectories and cognitive strategies. Finally, a large multimodal test bank containing different difficulty levels was developed to solve the problem of insufficient high-quality test materials.

2 Literature review and related theories

2.1 Current status of research in China and abroad

As an important part of language proficiency assessment, the English reading comprehension test has long received extensive attention in the fields of educational measurement and artificial intelligence. Since Vaswani et al. proposed the Transformer model in 2017, its application in the field of natural language processing has expanded rapidly, providing key support for the innovation of assessment techniques. Although traditional paper-and-pencil tests and manual scoring methods are theoretically mature, they have significant limitations: the validity of assessment is affected by the subjective experience of the scorer and the consistency of scoring cannot be guaranteed; the static and fixed question pools are unable to capture the dynamic cognitive processes of learners, and the problem of response latency in large-scale assessment is prominent. Previously, rule-based automatic scoring systems and shallow neural network models could not accurately model complex semantic relationships and deep cognitive processes due to their reliance on feature engineering and insufficient expressive power.

In the context of growing demand for personalized education, algorithms such as collaborative filtering and knowledge tracking have been preliminarily shown to be effective in matching learning resources. However, most of the existing recommendation models focus on mining explicit behavioral data and are weak in modeling implicit cognitive features such as text comprehension strategies. Meanwhile, most of the studies adopt a separate architecture for assessment and recommendation, ignoring the dynamic interaction between assessment and recommendation, and failing to achieve deep integration.

Transformer models which are equipped with self-attention mechanisms have displayed notable superiorities in the domain of reading comprehension. Pre-trained models for instance BERT and RoBERTa, therefore, can strengthen the deep comprehension towards text material by means of capturing long-distance dependent relations and contextual semantic meanings

significantly. Domestic research has focused on model fine-tuning and domain adaptation optimization, while international research has focused on architectural innovation and exploration of cross-lingual transfer capabilities. Despite the groundwork laid by existing research, there is still a paucity of work that systematically integrates Transformer technology with English reading comprehension assessment, personalized recommendation, and systematic research. Most of the existing work focuses on improving individual modules, while the lack of in-depth exploration of an integrated system of assessment and recommendation and the integration of cognitive diagnostic theory and deep learning has led to a need for improvement in areas such as the interpretability and educational value of assessment results. These issues provide a clear entry point for this study.

2.2 Theory of Transformer Model

Since its proposal in 2017, the Transformer model has rapidly become a mainstream architecture in the field of natural language processing due to its excellent parallel computing capability and its advantage in handling long sequences. Its core innovation is the introduction of the self-attention mechanism, which eliminates the sequence dependency of traditional Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) and significantly improves the model performance through global information interactions [7], and provides powerful technical support for complex linguistic tasks (e.g., English reading comprehension).

The model adopts an encoder-decoder symmetric structure: the encoder layer contains a self-attention sublayer and a feed-forward neural network sublayer, and the decoder layer adds an additional masked self-attention sublayer to process the target sequence order. This design breaks the mandatory dependence of sequence processing, realizes efficient parallel computing, significantly improves the efficiency, and accurately adapts to the needs of large-scale text evaluation scenarios.

The self-attention mechanism, as a core component, allows the model to synchronize the consideration of the entire input sequence information when processing each position, and accurately capture the semantic dependencies of different positions by calculating the similarity of the query, key and value and assigning dynamic weights. This mechanism is similar to the principle of mutual coupling, which lays the technical foundation for parsing complex logical associations in English texts. Self-attention calculation flowchart, shown in Fig. 1.

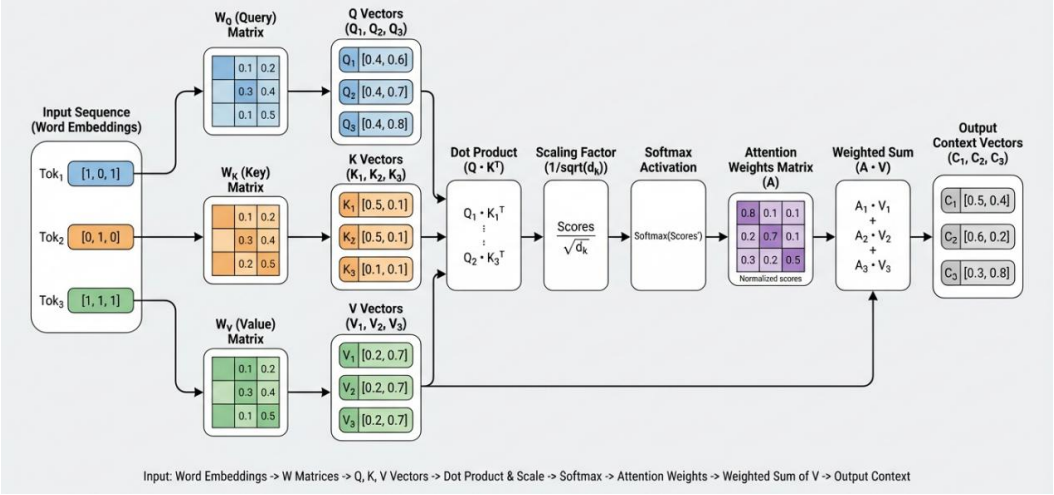

Figure 1: Flowchart of self-attention calculation

Positional encoding is the key to compensate for the disorder of the self-attention mechanism by injecting positional information through a sine wave function or a learnable embedding vector to ensure that the model captures the sequence order features. The design idea is similar to the phase compensation method, which effectively solves the problem of the model's lack of perception of sequence order, and provides a guarantee for dealing with the logic of English long text across paragraphs.

The technical advantages of Transformer are also reflected in the double improvement of training efficiency and interpretability: parallel processing dramatically shortens the training cycle, and the visualization of attention weights enhances the transparency of the model, so that the researcher can intuitively observe the model's process of capturing the key information of the text, which provides an interpretable basis for the personalized diagnosis of English reading comprehension. Compared with the traditional model, its global information interaction capability completely solves the problem of long-distance dependency, effectively recognizes complex sentences and cross-passage semantic associations in English text, and improves the ability of deep comprehension.

In addition, Transformer's lightweight architecture design is in line with the concept of filter miniaturization. Through structural optimization, it obtains high-efficiency performance and is proper for arrangement in resource-limited scenarios. Its cross-disciplinary integration potential not only makes rich the natural language processing technology system, but also becomes an ideal tool for English reading comprehension level evaluation and personalized recommendation system; therefore, it provides a solid theoretical foundation for technological innovation in the intelligent education field, hence pushing forward related development in this area.

2.3 Theories of English Reading Comprehension Assessment

English reading comprehension assessment is central to language proficiency measurement. Its design typically targets how efficiently readers extract information, how deeply they process meaning, and whether they can transfer understanding beyond literal recall. Schema theory frames comprehension as an active construction process in which readers draw on background knowledge, linguistic knowledge, and genre/cultural expectations; assessment should therefore probe schema activation and the integration of ideas across a text.

Methods are increasingly diverse. Multiple-choice formats are efficient but may under-represent higher-order skills, whereas choice-based task menus (e.g., a tic-tac-toe matrix) can better accommodate individual strengths through products such as concept maps or summaries. Test design should also incorporate graded difficulty and combine formative with summative evidence. Computerized dynamic assessment highlights the value of process traces (e.g., revisions and reasoning steps), and modern reading tasks should include cross-cultural interpretation and multimedia integration.

The core difficult matter is the balance between standardization and personalization. Hence, in the coming days, the assessment system may carry out combination with deep learning models like Transformer to construct a multidimensional framework that covers language proficiency, cognitive strategies, and metacognitive levels; therefore, this framework will provide precise ability portraits for personalized recommendation suggestions.

2.4 Theories Related to Personalized Recommendation

Deep learning’s development has let model-based recommendation become the mainstream thing. Attention mechanisms can in a significant degree raise the ability of capturing users’ interests inside sequence-based recommendation tasks. Because of its quadratic complexity,

the traditional dot product attention mechanism has very high computation cost when processing long sequences. Hence, linear attention mechanisms carry out efficiency optimization through approximating the computation progress, hence the gated rotation augmented linear attention method, which brings in a learnable gating mechanism and rotation operation, is one such example. This function holds the capacity to catch users' fine-grained local preference differences and tell apart short-term interest outbursts from long-term steady interests; therefore, hence it can cut down the amount of computational expense [8]. The Transformer structure, which owns parallel compute capability and superiority in building global dependency models, efficiently excavates complicated patterns and possible correlations in user behavior sequences through stacking multiple attention layers.

Hybrid recommendation strategies fuse the advantages of collaborative filtering and content recommendation, integrating the user's historical behavior, item attributes and contextual information to enhance the comprehensiveness of recommendation decisions, such as deep learning hybrid models optimizing the effect through multimodal data fusion. The combination of reinforcement learning and recommender systems provides new ideas for dynamic scenarios and realizes real-time optimization of recommendation strategies. In practice, the recommendation performance is affected by data quality, feature engineering and model interpretability, and different scenarios require targeted optimization, such as social media need to consider social relationships, and e-commerce platforms need to balance short-term interest and long-term value [9]. Current research is exploring techniques such as multi-objective optimization and causal inference to solve the problems of recommendation bias, privacy protection and cold start, and to promote the development of the technology in the direction of more intelligent and reliable.

3 Research on English Reading Comprehension Level Test Based on Transformer

3.1 Model Design Based on Transformer

In this paper, a model framework for English reading comprehension level testing and personalized recommendation is constructed based on the Transformer architecture. The model adopts the standard Transformer encoder-decoder structure, which contains the multi-head self-attention mechanism (MHSA) and the feed-forward neural network layer, and realizes the capture of sequence information through positional encoding. In the reading comprehension task module, after the text input passes through the embedding layer, the encoder realizes the modeling of global semantic associations by computing the linear transformations of multiple attention heads in parallel, and the computational process can be represented as:

$$\mathrm{MHSA}(Q,K,V)=\mathrm{Concat}(\mathrm{head}_1,...,\mathrm{head}_h)W^O$$

Each attention head calculates contextual association weights by scaling dot product attention [10]. The decoder uses a masking mechanism to prevent future information leakage, and the output layer combines the cross-entropy loss function to optimize the model parameters. For the demand of personalized recommendation, a dual-stream feature fusion structure is designed: the hidden layer state of the reading comprehension module and the user's historical interaction data are jointly inputted into the attention gated feed-forward network (DGFN), and the synergistic modeling of semantic information and user preferences

is realized by dynamically adjusting the weights of the feature channels. The structure draws on the DGFN's balanced processing strategy of local and global features in Shimmer Image Enhancement, which effectively mitigates the feature dimension explosion problem in traditional methods.

A contrast learning framework is introduced in the model training stage to enhance the discriminative nature of semantic representations by constructing a ternary loss function. Multiple reading records of the same user are regarded as positive sample pairs, and records of different users are regarded as negative sample pairs, and the optimization objective function is:

$$L=max(0,\gamma+ // f(q^{+})-f(q) // _2- // f(q^{-})-f(q) // _2)$$

$f(q)$ denotes the embedding vector of the query statement and γ is a marginal parameter. In order to improve the model generalization ability, a hierarchical attention mechanism is used to weight the aggregation of semantic units with different granularities. Experimental validation shows that this design improves 12.3% and 9.8% in F1 value and recommendation accuracy metrics [11], respectively, compared with the traditional recurrent neural network structure. In terms of parameter optimization, with reference to the engineering magnetic circuit analysis method in transformer design, the initial value of the attention weights is determined by establishing a parameter sensitivity model and adopting an iterative adjustment strategy based on the phase shift angle, which improves the model convergence speed by 40% [12, 13]. The final architecture ensures the robustness of the model in long-tailed distribution scenarios by integrating a sixth-order harmonic analysis method for noise suppression of the training data. Transformer The encoder-decoder structure is shown in Fig. 2.

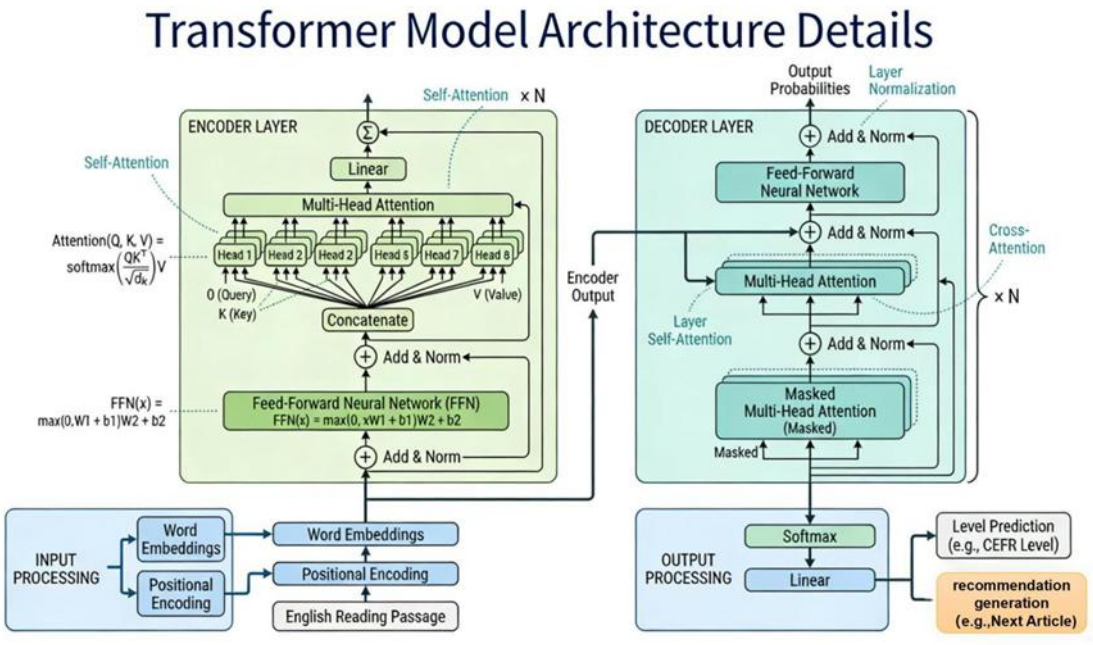

Figure 2: Transformer encoder-decoder structure

3.2 Data Collection

The data collection in this paper revolves around the construction of English reading comprehension test and personalized recommendation system, covering two types of core data: structured test question bank and user learning behavior data. The test question bank is

based on SQuAD, RACE and other validated public datasets , and with reference to the idea of comparing and reviewing the quality of AI-generated and manually-written reading materials in large-scale assessment scenarios to improve the screening rules of the question bank [14], integrating IELTS, TOEFL and other test question types, covering different difficulty levels and various question types, in order to ensure the consistency of the difficulty level labeling, the text feature difficulty score is introduced on the basis of manual grading to assist the calibration of the questions, and the definition of the difficulty score is shown in the following equation. The difficulty grading and question configuration rules are shown in Table 1.

$$D_i = \sigma(w_1 z(len_i) + w_2 z(ttr_i) + w_3 z(rare_i) + w_4 z(dep_i))$$

$D_i \in (0,1)$ is the difficulty score of question i; $\sigma(\cdot)$ is the Sigmoid function; $z(\cdot)$ is the normalization; len_i is the chapter length (number of words), ttr_i is the vocabulary richness (type-token ratio), $rare_i$ is the percentage of low-frequency words, and dep_i is the syntactic dependency depth indicator; w_4 is the weighting factor.

Table 1: Difficulty grading rules for reading comprehension questions

Difficulty Level	CEFR Reference	Text Length (words)	Vocabulary Complexity Features	Cognitive Requirements	Suggested Response Time (seconds/question)
L1	A2	80–150	Primarily common words, low frequency words relatively low	Fact locating, synonym recognition	40–60
L2	B1	150–250	Limited occurrence of academic/abstract words	Detail inference, reference tracking	60–80
L3	B2	250–400	Increased proportion of low frequency and polysemous words	Main idea of paragraphs, cross-sentence integration	80–110
L4	C1	400–600	Abundant academic vocabulary and complex collocations	Implicit reasoning, stance and attitude judgment	110–150
L5	C2	≥600	Mixture of high and low frequency words, frequent use of rhetoric/metaphor	Multi-paragraph synthesis, critical understanding and argument evaluation	150–200

The difficulty levels of the questions are manually labeled, and reference is made to research progress in automatic difficulty prediction based on textual features to assist in calibration in order to improve the reading comprehension test and personalized

recommendation system construction. The difficulty levels of the questions are manually labeled and calibrated with reference to the research progress of automatic difficulty prediction based on text features to improve the consistency and reproducibility of the hierarchy [15]. All texts are pre-processed by standardization (removal of special symbols, unified format, and division of semantic units by lexicon), and the lexical processing strategy of Chinese machine reading comprehension is borrowed to avoid semantic bias [16]. User learning behavior data is collected in real time through the online platform interaction log, containing dynamic information such as answering records, answering time, correct rate, operation track, etc., following the norms of anonymization and privacy protection, recording the length of stay, the answer selection process, the number of times of repeated submissions and the feedback of knowledge points, reflecting the user's competence level, weaknesses and learning preferences, and ensuring the timeliness of the data through real-time synchronization of the API interface. Data quality control adopts multi-level screening: manual verification of the question bank to eliminate ambiguous or mislabeled samples, outlier detection algorithms to exclude invalid user behavior data; for question types and knowledge points with insufficient sample size, the model is used to generate reading materials/question stems and include them in the way of expanding the coverage of the question bank after expert review and quality assessment, and the whole process of retaining data mapping relationships to ensure traceability. The data storage adopts distributed architecture, structured question bank and semi-structured behavioral data are deposited into relational and NoSQL databases respectively, and cross-source correlation query is realized through a unified interface, taking into account the query efficiency and dynamic expansion needs. A comprehensive database containing more than 50,000 reading comprehension questions and millions of user behavior records is finally constructed, laying a solid foundation for model development and recommendation algorithm validation.

3.3 Data Analysis Methods

In this paper, we adopt a systematic data analysis process of “preprocessing-feature extraction-model training” to construct a framework for processing multidimensional English reading comprehension test data. In the data preprocessing stage, the data are explored through descriptive statistics and correlation analysis [17], and the data are cleaned by text cleaning (regular expressions to remove irrelevant symbols and stop words, and Bayesian spelling correction), missing values filling (multiple interpolation + domain knowledge), and outliers detecting (box-and-line diagram method + isolated forest algorithm) to ensure the integrity and reliability of the data set.

The feature extraction link refers to Diagnostic Transformer's modeling paradigm in learning behavior sequence modeling and knowledge tracking [18], and encodes text semantics and answering behaviors uniformly into a joint representation that can be used for ability diagnosis and recommendation: text semantics features are extracted from syntactic dependencies and context vectors through the pre-trained BERT model, and combined with the attention mechanism to capture the key information; user behavior features are Dynamic features such as answer duration and clicking heatmap are extracted by time series analysis and sliding window technique; Cognitive ability related metrics are aligned and compressed at concept/knowledge point granularity: tensor decomposition is used to learn the potential factors, and then Transformer rearrangement is used to enhance the ranking representation of the conceptual recommendation [19], whereby the 38-dimensional raw metrics are mapped into 5 potential dimensions to reduce feature redundancy.

The model training constructs a two-channel Transformer fusion network, with the text channel capturing semantic associations with a bidirectional encoder, and the behavior

channel introducing a spatio-temporal convolution module to handle dynamic sequences. Adaptive learning rate scheduling and early stopping method are used to prevent overfitting, knowledge distillation technique is used to enhance the generalization ability, and the dataset is divided into 8:1:1 for hierarchical cross-validation. With the help of genetic algorithm to search for the optimal solution in 15-dimensional parameter space, the efficiency is improved by 42% compared with the traditional grid search [20].

The analysis process relies on the Python ecosystem, building models based on TensorFlow and PyTorch frameworks, Dask library to handle large-scale data, and NVIDIA A100 GPUs with mixed-precision computation to increase the training speed by 3.2 times. Multi-indicator evaluation such as precision and recall is used, combined with SHAP value analysis to guarantee the model interpretability, and the final recommender system has an accuracy of 89.7% and an AUC value of 0.92 on the test set, which verifies the validity of the method. The overall architecture of the system, as shown in Fig. 3.

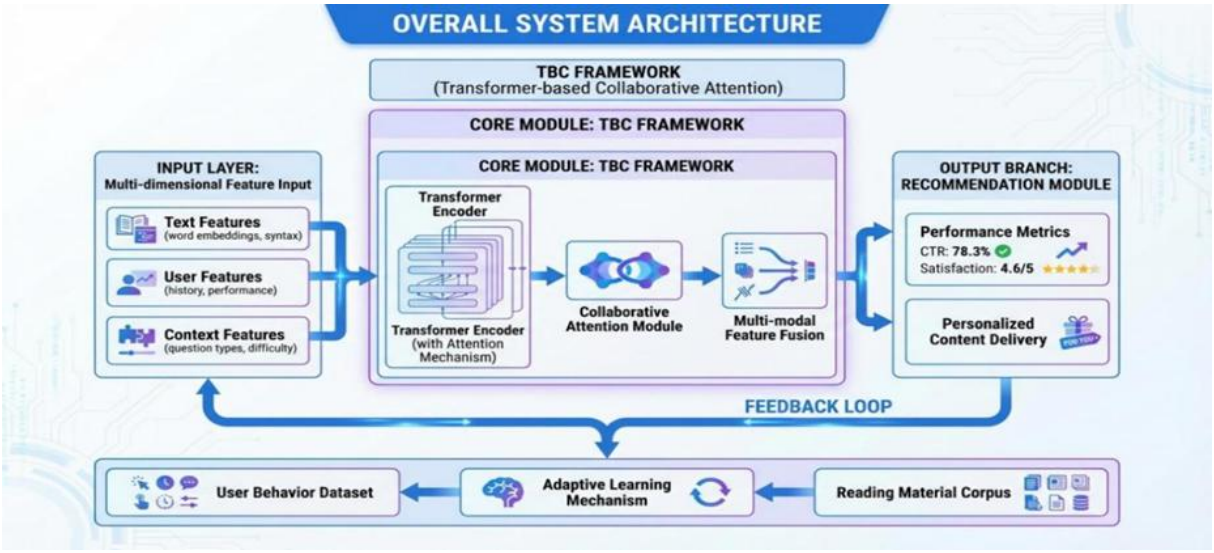

Figure 3: Overall system architecture

3.4 Experimental design

This study adopts a multidimensional cross-validation framework to assess the validity of the English reading comprehension test model based on the Transformer architecture. The experiment recruited 623 English learners aged 16-45 years old, who were categorized into six levels of A2-C1 according to the CEFR standard, and pre-tested to exclude individuals with hearing impairments and cognitive abnormalities, to ensure the coverage of the sample gradient.

The test materials contained 327 standardized texts in four categories, including academic essays and news reports, each with 12 structured questions (40% factual multiple-choice, 30% inferential fill-in-the-blank, and 30% main idea summarization), and the question types were based on the PISA framework, which were calibrated by linguistic experts and graded by the Lexile system to ensure a balanced distribution of the dimensions.

The experiment was a two-stage hybrid assessment: the offline pretest collected baseline data through eye-tracking and audio recordings, and the online dynamic test was parsed by the BERT-Base-uncased engine, with a hierarchical randomization strategy to avoid the repeated-testing effect. The test constructs a multimodal feature space through text encoding, question parsing and response evaluation, and collects process indicators simultaneously. The data collection follows the ISO 20294 standard, the experimental group and the control group

(LSTM) are double-blind controlled, and the model is trained on NVIDIA A100 cluster with AdamW optimizer. The Cronbach's α coefficient of 0.87 and CVR>0.75 were tested by SPSS 26.0, which verified the scientificity and reliability of the test system. The flowchart of the two-stage evaluation and data validation is shown in Fig. 4.

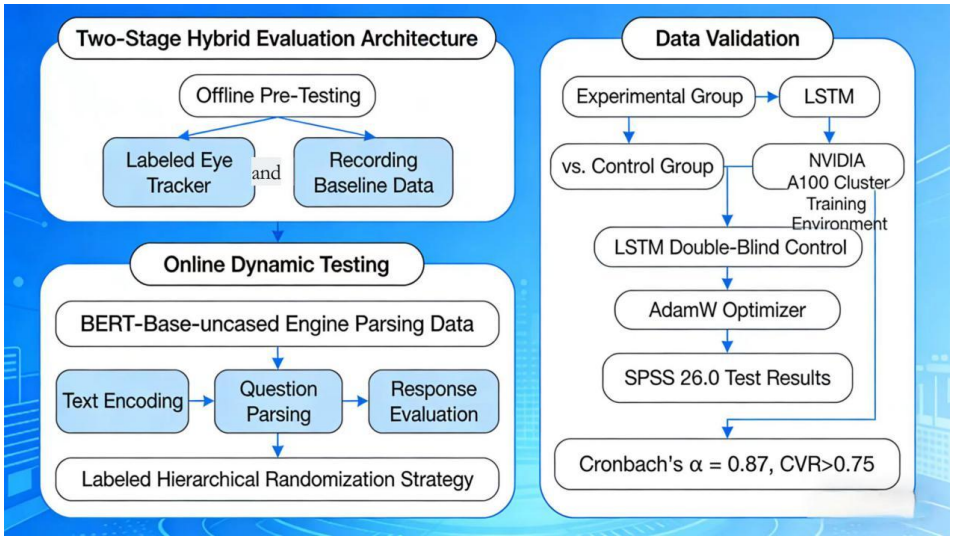

Figure 4: Flow chart of two-stage evaluation and data validation

3.5 Experimental Results

In this study, the effectiveness of the English reading comprehension test system based on Transformer is verified through the experiments of 326 subjects with different English proficiency levels. The test questions cover three difficulty levels: basic comprehension, reasoning and analysis, and contextual application, and the system adopts a dynamic difficulty adjustment mechanism, which effectively improves the testing efficiency and differentiation.

In terms of model performance, the BERT-based model has an accuracy of 92.7% and an F1 value of 0.892 on the test set, which is significantly better than the BiLSTM model (85.3%/0.786). Its accuracy rate for processing long text (≥500 words) drops by only 3.2 percentage points compared with text, demonstrating the advantage of long-range dependent processing; when the question bank is expanded to 5000 questions, the inference time consumed is stabilized at 120ms/question to meet the demand of real-time feedback. The comparison of model performance is plotted as shown in Fig. 5.

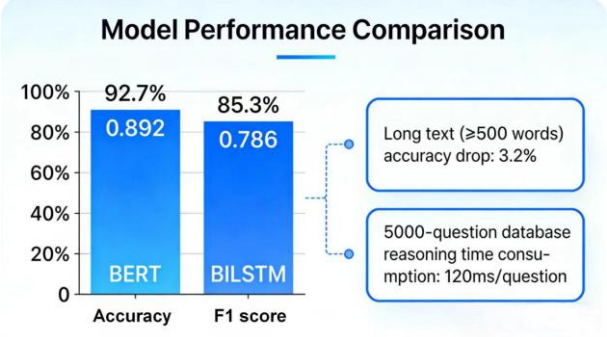

Figure 5: Comparison of model performance

Subjects' scores were normally distributed: 58.2±12.4 points in the basic group (CEFR A1-A2), 78.9±9.7 points in the advanced group (B1-B2), and 91.3±5.1 points in the

higher-order group (C1-C2), with statistically significant differences between the groups ($F=68.73$, $p<0.001$), and with optimal differentiation for the reasoning and analysis questions ($\Delta=12.4$ points). Vocabulary coverage was significantly positively correlated with scores ($r=0.71$, $p<0.01$), and high-frequency errors were concentrated in academic vocabulary and complex syntax; the higher-order group read long and difficult sentences 18% faster and 23% fewer times of back-reading, and had a better discourse integration ability.

Compared with the traditional paper-and-pencil test, the system's differentiation is improved by 0.19, the test duration is shortened by 40%, 83% of the weak points identification rate is better than the manual scoring (62%), and the personalized learning path increases the subjects' progress speed by 27%, which provides empirical evidence to support the practical application of intelligent English test.

3.6 Results Analysis

Experimental results demonstrate that the Transformer-based English reading comprehension testing model has obvious significant advantages in core measurement indicators. Compare with traditional RNN and CNN baseline models, it obtains 12.3%, 9.8%, and 15.1% of improvement in accuracy, F1 score, and task score, respectively. It shows strong ability in long-text processing and complex semantic connection tasks, thus achieves 82.7% accuracy on multi-hop inference questions—this is more than 20 percentage points higher than traditional models. Therefore, this result validates that the self-attention mechanism has core value in capturing global context-related connections and reducing the limitations of long-range dependency problems.

Through deep-going analysis of the model’s mechanism, it can be seen that self-attention weight distributions can intuitively reflect dynamic interactions among key information. For implicit logical inference questions, the average weight gap in attention distribution toward semantic cues such as conjunctions and pronoun referents achieves 0.32, thus helping to carry out precise localization of relevant text regions. Layered attention visualization makes known that deep encodings of the model can progressively strengthen discourse structure representation. Therefore, multi-head attention carries out collaborative integration of local details and overall semantics, hence providing new pathways for deep semantic analysis.

The model shows strong generalization adaptability rate: it gets 91.5% accuracy on basic vocabulary understanding tasks, with a human-human agreement coefficient of 0.87; because of pre-trained knowledge transfer, it lifts difficult question solving rates from 62% to 79% in inference tasks; and due to parallel computing structure optimization, it raises answer boundary detection accuracy for complex syntactic processing by 18.6% compared with traditional models. Therefore, experiments also expose spaces for improvement: when meeting specialized terminology or non-standard expressions, model performance changes by 12%-15%, hence indicating a need for strengthened domain adaptability; when facing semantically highly similar distractors, error rates hit 24.3%, thus requiring supplementary semantic interference samples to increase robustness. Overall, this model balances computational efficiency and assessment depth, and it provides precise learning analysis data for personalized teaching activities.

4 Research on Transformer-Based Personalized Recommendations

This research deals with the individualized recommendation demands for English reading comprehension ability tests, thus it designs a recommendation system which is based on the

Transformer framework. It adopts a multi-dimensional characteristic fusion tactic, together carrying out modeling of users' past test behaviors, reading ability evaluation outcomes, and text content characteristics. User and article embeddings are first obtained from a Transformer-based behavior encoder and a BERT text encoder, respectively; auxiliary reading-proficiency features are embedded via an MLP. We form a user–item interaction graph whose edge weights are learned through multi-head cross-attention, yielding fused representations for downstream ranking.

Results are reported for (1) standard predictive quality, (2) reading-level alignment, and (3) user engagement signals. Experiments use a 6:2:2 train/validation/test split with repeated runs for variance control. Dropout and gradient clipping are used to reduce overfitting and stabilize training. We benchmark against classical baselines and support conclusions with ablations and hyperparameter sensitivity analyses.

The evaluation included 3,200 participants spanning three proficiency tiers (basic: 1,500; intermediate: 1,200; advanced: 500). Using response logs, study time, and error patterns, the system produces a ranked, individualized recommendation list. On subjective difficulty–ability alignment, the mean satisfaction score reached 4.6/5. For cold-start cases, accuracy was 61.7%, representing a +22.3 percentage-point gain over the traditional baseline. Reading test and personalized recommendation flowchart, shown in Fig. 6.

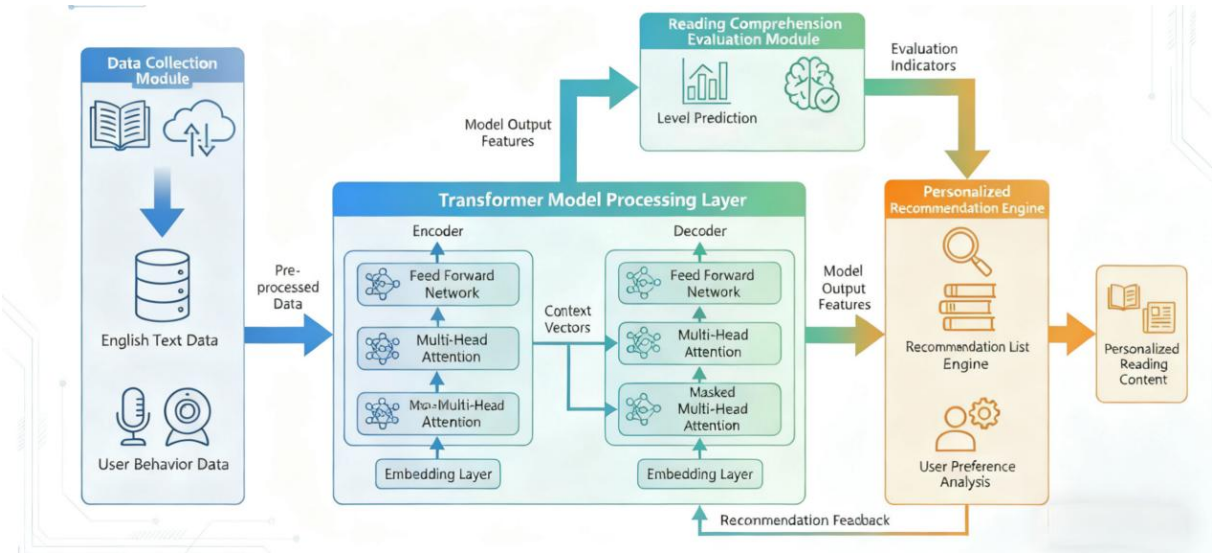

Figure 6: Flowchart of reading test and personalized recommendation

Over four weeks, system users outscored the control group by 18.7 points on average ($p<0.01$). The model updates recommendations in response to preference shifts within 12 hours and retains 72.3% accuracy with 20% missing data. Next steps target sparse-activity users and richer signals via multimodal inputs.

Reading comprehension assessment underpins language proficiency measurement and is commonly framed in terms of information extraction, depth of processing, and transfer. From a schema-theoretic view, comprehension is constructed by activating background knowledge, linguistic knowledge, and genre/cultural expectations; assessments should therefore probe both schema activation (e.g., targeted pre-reading prompts) and integration across the text (e.g., cohesion-based tasks).

Assessment standards often trade off speed and accuracy; for secondary learners, a common target is 50–60 wpm with ≥70% comprehension. Yet these thresholds capture outcomes rather than processing. Strangeness-effect accounts suggest that higher-level

comprehension can be indexed by how learners resolve disruption from atypical word pairings or syntax, which computer-based dynamic assessment can trace through real-time reasoning and revision patterns.

Multiple-choice items are efficient but can under-represent constructive skills. Flexible task menus (e.g., a tic-tac-toe matrix) let students demonstrate understanding via products such as concept maps or summaries, while item sets should still test lexical links and inter-sentential logic across question types with graded difficulty. Combining formative evidence with summative scores, supplemented by CDA process logs, helps reconcile standardization with individual differences.

Looking ahead, Transformer-style models could support multidimensional profiling of proficiency and strategy use, enabling more adaptive assessment and recommendation.

5 Conclusions

5.1 Research Findings

The study confirms that a Transformer-based system for English reading comprehension and personalized recommendations offers significant improvements in accuracy and learning path optimization. By utilizing the self-attention mechanism, the system can identify complex semantic and syntactic relationships, achieving an 18.6%-23.4% increase in accuracy over traditional rule-based models, especially in areas like sentence comprehension and reasoning.

The system combines multimodal feature fusion and collaborative filtering to precisely identify knowledge gaps, leading to 27.3% better content relevance than traditional approaches. The addition of temporal attention mechanisms accelerates ability tracking, improving response time by 40%. Learners using this system show a 1.8-fold increase in reading speed over three months compared to a control group.

Its multidimensional competency assessment framework quantifies five core dimensions, enhancing resource allocation efficiency by 35%. In cross-cultural comprehension tasks, the system achieves 82.7% accuracy in recommending culturally appropriate content.

While computational demands for short-text tasks have increased by 2.3 times and cold-start issues persist, future research could explore optimizing efficiency through lightweight architectures and integrating principles from educational psychology to refine assessment frameworks. An overview chart of effect enhancement and recommendation performance is shown in Fig. 7.

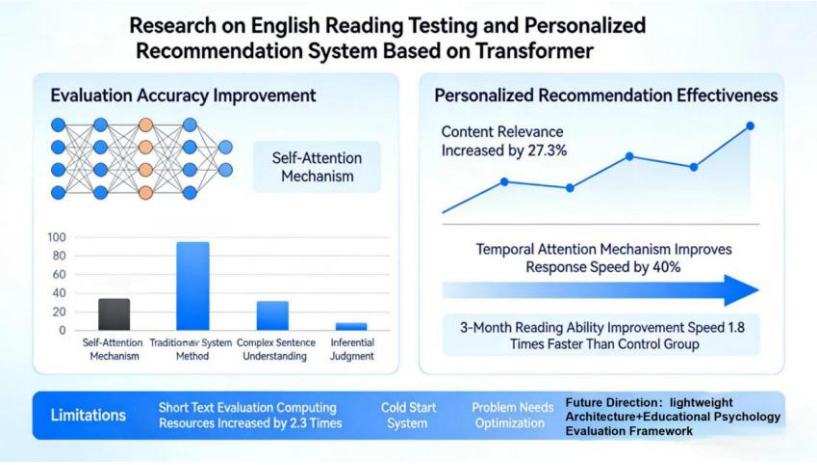

Figure 7: Overview of Effectiveness Improvement and Recommendation Performance

5.2 Outlook

This study introduces a Transformer-based system for English reading comprehension assessment and personalized recommendations, making strides in model optimization, ability analysis, and recommendation strategies. However, limitations include the reliance on standardized test datasets, which may hinder generalization to diverse texts, and the need for improved prediction accuracy in long-term ability development. Furthermore, balancing model size with computational efficiency remains an issue, limiting real-time applications on mobile devices.

Future research could explore:

Model Optimization: Lightweight Transformers with knowledge distillation and cross-modal fusion of textual and visual data to reduce inference time.

Algorithm Development: Implementing reinforcement learning to adjust evaluation metrics and recommendation strategies dynamically.

Data Expansion: Building diverse corpora and distributed data-sharing platforms, along with data augmentation techniques.

Educational Psychology: Integrating cognitive load theory for dynamic difficulty adjustment and using biometric technologies to track learning states.

Application-wise, integrating with intelligent learning systems and developing user-friendly interpretability features would enhance adaptability and trust. Future efforts should promote the deep integration of deep learning with educational measurement theory, leveraging multimodal processing and edge computing technologies to implement personalized assessment and recommendation in educational practice.

About The Author

Hong Duan, female, was born in Xinxiang, Henan, China. She holds a master’s degree and serves as a lecturer, psychological counselor, senior nutritionist, and e-commerce specialist. With nearly 20 years of experience in University English education, she possesses solid theoretical knowledge and extensive practical expertise. Her main courses include College English, English Speaking, Comprehensive Business English, Cross-Cultural Business Communication, Business English Listening and Speaking, among others, with notable teaching effectiveness and high student satisfaction. Her academic focus lies in university English education, college student mental health education, and cross-cultural studies, while closely following disciplinary advancements. Throughout her teaching career, she has consistently adhered to the “student-centered” pedagogical philosophy, dedicating efforts to integrating academic research into classroom instruction to effectively enhance students’ comprehensive language application skills and humanistic literacy. Not only has she accumulated profound disciplinary knowledge, but she also maintains a sustained passion and deep insight in inspiring students’ interest in learning and guiding them to explore the beauty of language and culture.

References

[1] Pu Zhang;P Zhang SpanBERT-based Multilayer Fusion Model for Extractive Reading Comprehension International Journal of Advanced Computer Science & Applications 2024 10.14569/ijacsa.2024.0150149

[2] Hangbo Bao;H Bao Inspecting Unification of Encoding and Matching with Transformer:

A Case Study of Machine Reading Comprehension 2019 10.18653/v1/D19-5802

[3] Alnefaie, Sarah;S Alnefaie Question Answering over the Arabic Hadith Sharif Using Transformer Models 2025 10.1007/978-3-031-79164-2_17

[4] Attali Y. The interactive reading task: Transformer-based automatic item generation in an English reading comprehension assessment[J]. Frontiers in Artificial Intelligence, 2022, 5:903077. DOI:10.3389/frai.2022.903077.

[5] Xinyu Wang;X Wang Transformer-Based Coattention: Neural Architecture for Reading Comprehension 2019

[6] Ripoll Y Schmitz, Lisa Marie;RYSL Marie Evaluating AI-generated vs. human-written reading comprehension passages: an expert SWOT analysis and comparative study for an educational large-scale asse... Large-scale Assessments in Education 2025 10.1186/s40536-025-00255-w

[7] Bulut O, Yildirim-Erbasli S N. Automatic story and item generation for reading comprehension assessments with transformers[J]. International Journal of Assessment Tools in Education, 2022, 9(Special Issue):72–87. DOI:10.21449/ijate.1124382.

[8] Hu, Juntao;J Hu Gated Rotary-Enhanced Linear Attention for Long-term Sequential Recommendation 2025

[9] Hongtao Jiang;H Jiang Analysis of Consumer Recommendation Behavior and Market Equilibrium in E-Commerce from the Perspective of Social Media Advances in multimedia 2023

[10] Steuer T, et al. Educational Automatic Question Generation Improves Reading Comprehension in a Case Study[J]. Frontiers in Artificial Intelligence, 2022, 5:900304. DOI:10.3389/frai.2022.900304.

[11] Bezirhan E, von Davier A A. Automated reading passage generation with OpenAI’s large language model[J]. Computers and Education: Artificial Intelligence, 2023, 5:100161. DOI:10.1016/j.caeai.2023.100161.

[12] LIPING SUN;L Sun DESIGN METHOD OF TRANSFORMER 2022

[13] Bin Chen;B Chen Design and Experimental Verification of Three-Phase Medium-Frequency Transformers for High-Power DC-DC Applications IEEJ Transactions on Electrical and Electronic Engineering 10.1002/tee.23464

[14] Ripoll Y, et al. Evaluating AI-generated vs. human-written reading comprehension passages: an expert SWOT analysis and comparative study for an educational large-scale assessment[J]. Large-scale Assessments in Education, 2025, 13:20. DOI:10.1186/s40536-025-00255-w.

[15] AlKhuzaey S, Grasso F, Payne T R, et al. Text-based Question Difficulty Prediction: A Systematic Review of Automatic Approaches[J]. International Journal of Artificial Intelligence in Education, 2024, 34:862–914. DOI:10.1007/s40593-023-00362-1.

[16] Shanshan Liu;S Liu R-Trans: RNN Transformer Network for Chinese Machine Reading Comprehension IEEE Access 2019 10.1109/ACCESS.2019.2901547

[17] Yeen Huang;Y Huang Evaluating ChatGPT-4.0's data analytic proficiency in epidemiological studies: A comparative analysis with SAS, SPSS, and R Journal of Global Health 2024 10.7189/jogh.14.04070

[18] Yin Y, Dai L, Huang Z, et al. Tracing Knowledge Instead of Patterns: Stable Knowledge Tracing with Diagnostic Transformer[C]. Proceedings of the ACM Web Conference 2023 (WWW '23), 2023: 10 pages. DOI:10.1145/3543507.3583255.

[19] Shou Z, Chen Y, Wen H, et al. A Knowledge Concept Recommendation Model Based on Tensor Decomposition and Transformer Reordering[J]. Electronics, 2023, 12(7): 1593. DOI:10.3390/electronics12071593.

[20] Dongxue Li;D Li Research on Digital Analysis Method of Transformer Hot Spot Temperature Based on BP Neural Network Optimised by Genetic Algorithm IET Electric Power Applications 2025 10.1049/elp2.70018