CHESS: A Multi-Agent Framework for Automated Chinese Essay Scoring in Standardized Testing

Xinzhuo Hou
Department of Linguistics and Translation
Hong Kong, China
xinzhuhou2-c@my.cityu.edu.hk

Wanli Ouyang
Department of Information Engineering
Hong Kong, China
wlouyang@ie.cuhk.edu.hk

Taoyong Cui
Department of Computer Science and Engineering
Hong Kong, China
cty21@tsinghua.org.cn

Pheng-Ann Heng
Department of Computer Science and Engineering
Hong Kong, China
pheng@cse.cuhk.edu.hk

ABSTRACT

Automated essay scoring (AES) plays a central role in large-scale standardized testing, yet existing systems developed primarily for English face critical challenges when applied to Chinese. The logographic nature of Chinese characters, flexible word boundaries, context-dependent syntax, and deep cultural embedding of writing styles make holistic and interpretable evaluation particularly difficult. Recent neural approaches based on pre-trained language models achieve strong overall performance but remain opaque and struggle to provide trait-specific diagnostic feedback, which is essential in high-stakes educational contexts. We propose **CHESS** (Chinese Essay Scoring System), a multi-agent framework that decomposes essay evaluation into specialized agents for grammar, vocabulary, structure, and content, orchestrated by a coordinator agent. This design yields transparent, trait-level assessment while maintaining holistic scoring consistency. To further address cultural and linguistic subtleties, CHESS incorporates customized rubrics aligned with HSK scoring criteria and leverages data augmentation strategies including paraphrasing, error injection, and style transfer. Experiments on Chinese essay datasets demonstrate that CHESS achieves higher agreement with human raters than strong baselines, while providing interpretable and pedagogically meaningful feedback. Our results highlight the promise of multi-agent reasoning for advancing fair, robust, and interpretable AES in complex linguistic settings.

Corresponding authors: Taoyong Cui (cty21@tsinghua.org.cn, Pheng-Ann Heng (pheng@cse.cuhk.edu.hk)

Artifact Availability:

We will release code, prompts, and evaluation scripts; report model sizes, training budgets, and seeds; and include full metric definitions and splits. Ablation checklists and configuration files will accompany the artifact.

1 INTRODUCTION

Automated essay scoring (AES) has emerged as a vital application of natural language processing (NLP) and artificial intelligence, enabling scalable evaluation in educational testing, language learning, and large-scale assessments [1–4]. Decades of research in English AES have led to systems deployed in standardized exams such as GRE, TOEFL, and GMAT, demonstrating both cost-effectiveness and reliability. These systems not only alleviate the burden on human graders but also provide rapid feedback for learners, making AES one of the most practically influential tasks in applied NLP [5–7].

Despite such progress, extending AES to Chinese remains an open challenge [8–13]. Unlike alphabetic languages, Chinese is logographic, where characters convey meaning at multiple granularities and word boundaries are not explicitly marked. Its syntax is more flexible and context-dependent, while effective writing often relies on idiomatic expressions, rhetorical structures, and culturally embedded conventions [14]. These characteristics render standard feature-based or end-to-end neural scoring approaches—largely developed in English—less effective when transferred to Chinese [6]. Moreover, most existing AES systems are designed as monolithic predictors, producing holistic scores but offering little interpretability or diagnostic feedback, which is particularly problematic in high-stakes educational contexts such as the Hanyu Shuiping Kaoshi (HSK) [15, 16].

These limitations point to a gap: the lack of AES frameworks that can simultaneously capture the linguistic complexity of Chinese, provide trait-specific evaluation, and deliver interpretable feedback aligned with educational rubrics [17–20]. Addressing this gap requires moving beyond single-agent black-box models towards systems that explicitly model the multifaceted dimensions of essay quality, from grammar and vocabulary to structure and content relevance [21].

In this paper, we introduce **CHESS** (Chinese Essay Scoring System), a multi-agent framework designed for interpretable and culturally informed AES in Chinese. CHESS decomposes the evaluation process into four specialized agents: a *Grammar Agent* that detects syntactic correctness and fluency, a *Vocabulary Agent* that measures lexical diversity and appropriateness of word choice, a

Structure Agent that assesses coherence and discourse organization, and a *Content Agent* that evaluates topical relevance and argumentative quality. Their outputs are integrated by a central *Coordinator Agent*, which aggregates trait-level assessments into a holistic score. This design allows CHESS to mimic human raters' multi-dimensional judgment, while ensuring transparency and actionable feedback. Furthermore, the framework incorporates culturally aligned rubrics from the HSK standard and leverages augmentation strategies such as paraphrasing, controlled error injection, and style transfer to enhance robustness across writing styles and proficiency levels.

2 RELATED WORK

Automated essay scoring (AES) has a long history, beginning with early systems like PEG [22] that used handcrafted features and regression models to predict holistic essay scores [23]. Over time, research has evolved into hybrid and neural approaches that incorporate richer linguistic, discourse, and semantic representations [24]. AES systems are typically classified along several axes: (i) holistic vs. analytic/trait scoring, (ii) feature-based vs. end-to-end models, (iii) monolithic vs. modular or multi-agent paradigms, and (iv) single-language vs. cross-lingual or language-specific designs.

AES in English and General Domain. Classical AES systems such as E-rater [25], IEA [26], and PEG [27] rely on manually engineered features (e.g. lexical sophistication, syntactic complexity, error counts) and regression or classification models. Recent advances [28] leverage pre-trained large language models, fine-tuned or prompted to predict essay scores directly. However, many such models remain opaque and produce only a single holistic score, limiting interpretability and trait-level feedback.

Trait / Multi-Dimensional Scoring. To address the limitation of holistic scoring, prior work has proposed multi-trait and multi-task architectures. For example, the Hierarchical Multi-task Trait Scorer [3] (HMTS) has been applied to Chinese AES, modeling four traits (organization, topic, logic, language) and using inter-sequence attention to capture cross-trait dependencies. In the broader AES literature, such multi-task or modular designs [29] enable models to output trait-level scores in addition to the overall grade.

Chinese Essay Scoring. Chinese AES (or Chinese L2 writing AES) poses unique challenges and has attracted increasing research attention in recent years. A prompt-independent and interpretable approach leverages 90 linguistic features with ordinal logistic regression, achieving reasonable accuracy while maintaining interpretability [30]. Yuan et al. [6] further proposed a hybrid deep learning model that combines handcrafted statistical features with latent semantic indexing (LSI) features for Chinese essay scoring. The "SmartWriting-Mandarin" system [31] targets Chinese-as-a-foreign-language learners and integrates preprocessing, feature extraction, and scoring modules tailored for typical non-native writing errors. In addition, recent surveys [32] emphasize persistent challenges in Chinese AES, such as limited corpora, difficulties in capturing grammatical error patterns, and the need to model stylistic or rhetorical nuances (e.g., idiomatic or classical usages).

Multi-Agent and Modular Scoring Systems. Recently, multi-agent architectures have started appearing in AES settings [33, 34]. For example, the CAFES framework [35] applies collaborative agents to multi-granular, multimodal essay scoring. The RES (Roundtable Essay Scoring) framework [34] uses multiple LLM-based evaluators that deliberate via a dialectical protocol to reach a final score, improving alignment with human judgments. These works illustrate the potential of agent-based decomposition to reconcile trait-level judgments and holistic consistency.

Chinese AES with LLM & Ranking Techniques. In the Chinese AES domain, recent work explores *rank-then-score* pipelines, where a pairwise (or listwise) ranker precedes a calibrated scorer. For example, Rank-Then-Score (RTS) [36] reports state-of-the-art results on HSK and English AES benchmarks.

Gaps and Opportunities. Despite this progress, existing Chinese AES systems often remain monolithic or feature-heavy, lack deep interpretability, and struggle with domain generalization when faced with new prompts or rhetoric. Agent-based or modular frameworks in the Chinese context remain underexplored. Moreover, incorporation of cultural and rhetorical knowledge (such as Qi-Cheng-Zhuan-He, idiomatic usage, register control) is rare. Our proposed CHESS system seeks to fill these gaps by offering a modular, trait-aware, culturally sensitive multi-agent architecture for Chinese AES.

3 CHESS FRAMEWORK

We propose **CHESS** (Chinese Essay Scoring System), a multi-agent framework designed to capture the multi-dimensional nature of essay quality while providing interpretable and pedagogically meaningful feedback. Unlike monolithic black-box AES models, CHESS explicitly decomposes the scoring process into modular agents, each responsible for a distinct trait, coordinated by an orchestrator to produce a holistic and consistent score. Figure 1 illustrates the overall architecture.

3.1 Design Principles

CHESS is guided by three key principles:

(1) **Trait-specific evaluation**: Human raters rely on separate rubrics for grammar, vocabulary, structure, and content. CHESS mirrors this practice by dedicating specialized agents to each trait.
(2) **Cultural alignment**: Scoring rubrics integrate criteria derived from HSK standards and Chinese writing pedagogy, ensuring sensitivity to idiomatic usage, discourse conventions, and rhetorical devices.
(3) **Interpretability and feedback**: Each agent outputs both a numeric sub-score and diagnostic rationales (e.g., error categories, lexical richness indicators), enabling transparent, actionable feedback rather than opaque holistic predictions.

3.2 Agent Modules

CHESS consists of four specialized evaluators:

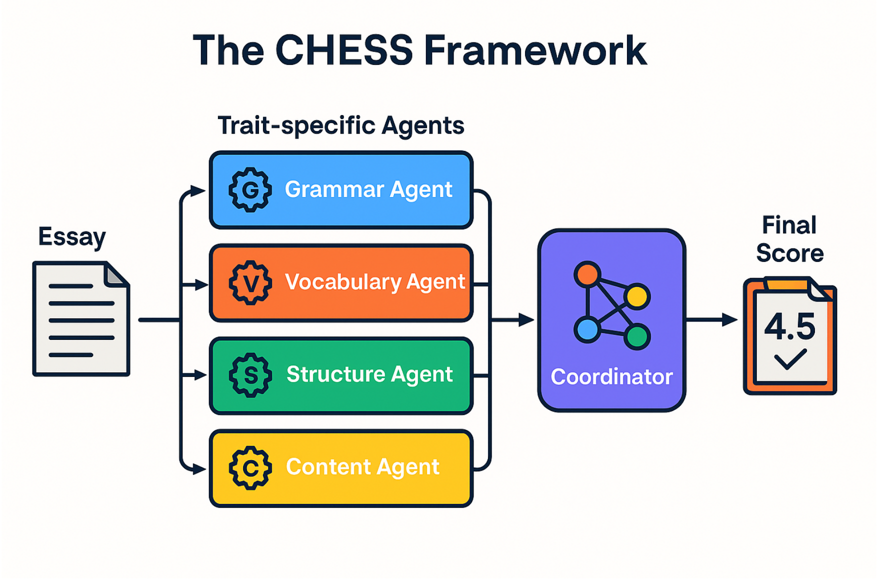

Figure 1: Overview of the CHESS framework. An input Chinese essay is evaluated in parallel by four prompt-based agents—*Grammar*, *Vocabulary*, *Structure*, and *Content*. Each agent produces a trait score and explanatory feedback. The *Coordinator* aggregates trait scores via weighted integration with consistency checks to yield the final holistic score and explanations. A training-side data augmentation pipeline (paraphrasing, error injection, style transfer) improves robustness of agents.

Grammar Agent. Analyzes morpho-syntactic correctness, word order, and clause-level coherence. Leveraging large language models fine-tuned on error-annotated corpora, it detects subject–predicate mismatches, misuse of function words, and sentence fragment errors. Outputs include error types and frequency, aligned with typical learner mistakes.

Vocabulary Agent. Assesses lexical diversity, word appropriateness, and register control. This agent uses lexical sophistication indices (e.g., type–token ratio, rare word coverage) in combination with embedding-based semantic similarity to reference rubrics. It flags over-reliance on simple vocabulary and improper word collocations.

Structure Agent. Evaluates discourse-level organization, cohesion, and logical flow. Using sentence embedding similarity graphs and rhetorical structure modeling, it detects abrupt topic shifts and lack of coherence. It also scores paragraph organization following Chinese writing norms (qi–cheng–zhuan–he).

Content Agent. Measures topical relevance, argument quality, and factual correctness. This agent employs prompt-based retrieval augmentation to compare essays with reference materials or model answers, ensuring coverage of expected key points. It penalizes digressions, irrelevant narratives, or unsupported claims.

Prompt Design. Each agent in CHESS is realized by prompting the DeepSeek v3 model with rubric-aligned instructions, without any fine-tuning. To illustrate, the Grammar Agent is instantiated with the following template:

Analogous rubric-aligned prompts are designed for the Vocabulary, Structure, and Content agents, each tailored to their specific scoring dimension. Full templates are provided in the Appendix for reproducibility.

3.3 Coordinator and Score Aggregation

The **Coordinator Agent** integrates outputs from the four evaluators. It applies a learned weighting function—trained via multi-task regression—to reconcile sub-scores into a final holistic grade. To ensure fairness and robustness, the coordinator employs consistency checks: (i) verifying alignment with HSK rubrics, (ii) calibrating inter-agent disagreements, and (iii) normalizing across writing prompts to avoid topic bias.

3.4 Coordinator and Integration Details

While each agent produces a trait-specific score and feedback, the final holistic score requires reconciling potentially divergent judgments. We formalize this process in three steps: attention-based integration, cross-agent communication, and uncertainty-aware decision support.

Attention-based Integration. Given trait-level scores S_t and confidence estimates C_t from each agent $t \in \{g, v, s, c\}$ (grammar, vocabulary, structure, content), the coordinator computes attention weights as

$$\alpha_t = \frac{\exp(h_t^\top W_c u)}{\sum_{t'} \exp(h_{t'}^\top W_c u)},$$

where h_t is the hidden representation extracted from agent t, u encodes essay-level context features, and W_c is a learned parameter matrix. The final holistic score is

$$S_h = \sum_t \alpha_t \cdot S_t,$$

Listing 1 Prompt Template: Grammar Agent

```
ROLE:
  You are an experienced Chinese language examiner for
       HSK writing.

GOAL:
  Analyze morpho-syntactic correctness, clause-level
       coherence, and sentence completeness according to the
        official HSK rubric.

CONSTRAINTS:
  - Focus on morpho-syntax: –subjectpredicate agreement,
       function words, aspect markers, classifiers, and word
       order.
  - Identify common learner errors such as –subjectpredicate
        mismatch, function-word misuse, clause fragment,
       redundant particles, or missing subjects.
  - Provide concise diagnostic phrasing in Chinese, aligned
       with HSK grammar criteria.
  - Score range: integer –05 (0 = unintelligible, 5 = error-
       free).

STEPS:
  1. Detect and categorize morpho-syntactic errors (–
       subjectpredicate mismatch, function-word misuse,
       clause fragment, classifier error, word-order issue).
  2. Estimate error frequency and severity (rare / occasional
        / frequent).
  3. Summarize overall grammatical accuracy and fluency.
  4. Assign a final Grammar Score –(05).

OUTPUT_SCHEMA (JSON):
{
  "Score": <int –05>,
  "Errors": [
    {
      "category": "<subject-predicate-mismatch | function-
       word-misuse | clause-fragment | classifier-error | word-
       order-issue>",
      "span": "<offending sentence fragment>",
      "note": "<short diagnostic explanation>",
      "severity": "<rare | occasional | frequent>"
    }
  ],
  "Explanation": "<overall comment in Chinese
       summarizing major error types and their frequency>"
}

INPUT:
<Essay text here>
```

ensuring that more reliable agents (with higher C_t and context alignment) contribute more strongly to the overall decision.

Algorithm 1 GraphRAG-CHESS for Agent Scoring

```
Input: Essay $\mathcal{E}$, Prompt $\mathcal{P}$, Rubric graph $\mathcal{G}$
Output: $S_c$, structured evidence $\mathcal{R}_{\text{graph}}$

Build essay claim graph $\mathcal{G}_{\mathcal{E}}$ (NER, relation, discourse cues)
Query-expand via rubric keyphrases and neighbor nodes; retrieve $\hat{\mathcal{G}} \subset \mathcal{G}$
Graph-aware rerank (node semantic match + edge-type compatibility)
For each rubric item $r_j$: compute $s_j^{\text{cov}}$ by subgraph alignment; $s_j^{\text{con}}$ by path entailment
Compute $s^{\text{theme}}$ using cluster-weighted coverage
Aggregate $S_c$; emit $\mathcal{R}_{\text{graph}}$ with matched nodes and supporting paths
```

Cross-Agent Communication. Although agents are specialized, the coordinator allows limited cross-agent information flow during training via an inter-agent attention mechanism:

$$\tilde{h}_t = h_t + \sum_{t' \neq t} \beta_{t,t'} \cdot h_{t'},$$

where $\beta_{t,t'}$ reflects learned correlations between traits (e.g., structure often depends on grammar quality). This design enables CHESS to capture interdependencies without collapsing agents into a single monolithic model.

Uncertainty and Contestability. Each agent produces not only a score S_t but also an uncertainty estimate U_t (e.g., variance across dropout ensembles). The coordinator aggregates them into a system-level uncertainty U. When U exceeds a predefined threshold τ, the system triggers a *contestability flag*, recommending human review or secondary scoring. This mechanism ensures that CHESS remains reliable in high-stakes contexts by routing ambiguous cases to human raters rather than issuing overconfident predictions.

3.5 Graph-based Knowledge Integration (GraphRAG-CHESS)

Although CHESS decomposes essay evaluation into modular agents, each agent still operates primarily on surface text representations. To enhance the system's capacity for higher-order reasoning and factual grounding, we introduce GraphRAG [37], a graph-based retrieval-augmented reasoning module that extends all four agents. The GraphRAG knowledge corpus is constructed from rubric-aligned keypoints, annotated exemplars, and short factual background passages for each writing prompt, forming a controllable and auditable repository of pedagogical evidence.

Motivation. Conventional retrieval-augmented scoring treats reference materials as independent passages. However, Chinese argumentative writing often exhibits relational patterns *Qi–Cheng–Zhuan–He* structure, cause–effect reasoning, and topic–comment progression—that require understanding inter-sentence links. GraphRAG encodes these relations explicitly as nodes and edges, allowing the model to evaluate not only topic relevance but also discourse connectivity and reasoning depth.

Graph-based Retrieval and Alignment. Given a student essay $\mathcal{E}$, we extract its claim graph $\mathcal{G}_{\mathcal{E}}$ and perform subgraph retrieval over the prompt's reference graph $\mathcal{G}$. The similarity between essay and reference structures is computed as

$$\mathrm{sim}(\mathcal{G}_{\mathcal{E}}, \mathcal{G}) = \lambda_v \cdot \mathrm{NodeSim} + \lambda_e \cdot \mathrm{EdgeMatch},$$

where node similarity reflects semantic correspondence of key concepts, and edge matching measures structural coherence. The retrieved subgraph provides explicit evidence for topical coverage and argument completeness.

Integration with the Content Agent. The Agent fuses textual and graph-level signals to compute the final content trait score:

$$S_c = \alpha \cdot S_{\text{text}} + \beta \cdot S_{\text{graph}},$$

where S_{text} is derived from prompt-based LLM evaluation and S_{graph} quantifies graph structural alignment. This design enables context-aware reasoning that captures both semantic and relational adequacy.

Interpretability and Feedback. GraphRAG-CHESS produces structured feedback highlighting matched rubric keypoints, missing reasoning links, and supporting evidence paths. This visualizable graph evidence provides learners with pedagogically meaningful guidance and allows instructors to trace how each argument contributes to the final score.

Efficiency and Safety. All reference graphs are precomputed offline, ensuring minimal inference overhead. Node texts are anonymized and truncated to avoid data leakage from exemplar essays. Graph matching is implemented with lightweight vector indices and optimized nearest-neighbor search, achieving sub-second evaluation per essay. Each rubric graph contains 350 nodes and about 700 edges on average, processed within 0.7 s per essay.

GraphRAG Knowledge Corpus Construction. To support graph-based retrieval and reasoning, we construct a compact, rubric-aligned knowledge corpus for each writing prompt, organized as a *prompt namespace*. Each namespace contains three layers: (i) **Rubric & Keypoints**, extracted from official HSK rubrics and rewritten into itemized content requirements; (ii) **Exemplars**, high-, medium-, and low-quality essays annotated with span-level roles (*claim*, *reason*, *evidence*); and (iii) **Background Snippets**, short (2–3 sentence) teacher-curated notes providing factual or cultural context. All materials are normalized, segmented into 200–500 character passages, and assigned namespace and role tags. Personal information is automatically redacted.

The corpus is then converted into a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where nodes represent rubric keypoints, exemplar spans, and background snippets, and edges capture rhetorical or logical relations such as supports, contradicts, and elaborates. Edges are induced from annotated span roles and discourse cues, and community detection (Louvain) produces thematic clusters used by the theme-coverage score. Hybrid BM25 [38] and dense indices are built per namespace for efficient node retrieval, and a graph-aware reranker evaluates both node similarity and edge-type compatibility. To ensure fairness and auditability, we cap snippet length, anonymize exemplar identifiers, and log retrieval hits for periodic bias inspection.

This corpus design operationalizes the rubric-aligned and evidence-grounded principle of CHESS, enabling structured and interpretable retrieval for evaluation under the GraphRAG-CHESS framework.

3.6 Data Augmentation and Robustness

To enhance generalization, CHESS incorporates multiple data augmentation strategies:

- **Paraphrasing**: generates semantically equivalent variants of essays for robustness against surface variation.
- **Error Injection**: introduces controlled grammatical and lexical errors to improve sensitivity of trait-specific agents.
- **Style Transfer**: simulates essays of different proficiency levels to balance training distribution.

3.7 Advantages over Existing AES Systems

Compared to holistic black-box AES models, CHESS offers:

- **Transparency**: trait-level scores and rationales.
- **Cultural validity**: alignment with Chinese rhetorical traditions and HSK rubrics.
- **Flexibility**: modularity allows independent upgrading of agents (e.g., replacing vocabulary model with a newer embedding).
- **Pedagogical utility**: diagnostic feedback can support formative assessment, beyond final scoring.

4 DATA: HSK-ENHANCED AND SPLITS

To evaluate CHESS in realistic high-stakes scenarios, we curate a dataset aligned with the Chinese Proficiency Test (HSK), which serves as the official standard for Chinese language assessment. The dataset integrates existing public AES corpora with HSK-aligned essays collected from examination materials and mock test platforms, covering a wide range of prompts, topics, and proficiency levels (HSK 3–6). Each essay is scored by certified raters following official HSK rubrics, ensuring consistency across grammar, vocabulary, structure, and content dimensions.

4.1 HSK-Enhanced Corpus

The HSK-enhanced corpus differs from generic Chinese AES datasets in three aspects:

1. **Rubric Alignment:** Scores are mapped to HSK scoring standards, capturing fine-grained linguistic and cultural criteria beyond holistic ratings.
2. **Trait Annotations:** In addition to overall scores, essays include trait-level annotations (e.g., morpho-syntactic correctness, lexical diversity, discourse coherence, topical relevance) to support multi-agent training.
3. **Augmentation:** We expand the dataset using paraphrasing, error injection, and style-transfer techniques to simulate diverse learner errors and enrich robustness.

4.2 Data Splits

We design splits to balance difficulty and generalization:

- **Train:** 70% of essays, stratified across HSK levels and prompts, for model training.

Algorithm 2 CHESS Inference Pipeline: From Essay Input to Holistic Score

```
Input: Essay $\mathcal{E}$, Prompt $\mathcal{P}$, Trait Rubrics $\mathcal{R} = \{\mathcal{R}_g, \mathcal{R}_v, \mathcal{R}_s, \mathcal{R}_c\}$
  Output: Holistic score $S_h$, trait scores $S_t = \{S_g, S_v, S_s, S_c\}$,
      feedback $F = \{F_g, F_v, F_s, F_c\}$, system confidence $U$
Preprocess $\mathcal{E}$: normalize characters, segment sentences, detect paragraphs; extract prompt features from $\mathcal{P}$.
Initialize working state $\Phi \leftarrow \emptyset$.
Parallel Trait Scoring
    Grammar Agent $\Rightarrow (S_g, F_g, C_g) \leftarrow \mathrm{Grammar}(\mathcal{E}, \mathcal{R}_g)$
    Vocabulary Agent $\Rightarrow (S_v, F_v, C_v) \leftarrow \mathrm{Vocab}(\mathcal{E}, \mathcal{R}_v)$
    Structure Agent $\Rightarrow (S_s, F_s, C_s) \leftarrow \mathrm{Structure}(\mathcal{E}, \mathcal{R}_s)$
    Content Agent $\Rightarrow (S_c, F_c, C_c) \leftarrow \mathrm{Content}(\mathcal{E}, \mathcal{P}, \mathcal{R}_c)$
Collect scores $S_t \leftarrow \{S_g, S_v, S_s, S_c\}$, feedback $F \leftarrow \{F_g, F_v, F_s, F_c\}$, confidences $C \leftarrow \{C_g, C_v, C_s, C_c\}$.
Disagreement Check
    $\delta \leftarrow \mathrm{Var}(S_t)$
if $\delta > \tau$ then
    $F \leftarrow F \cup \mathrm{Explain}(\mathrm{Grammar}, \mathrm{Vocab}, \mathrm{Structure}, \mathrm{Content})$

    $S_t \leftarrow \mathrm{ResolveDiscrepancy}(S_t, C, F)$ {e.g., lightweight iterative refinement or rule-based tie-breaks}
end if
Coordinator Integration
    $w \leftarrow \mathrm{AttentionWeights}(S_t, C, \Phi)$ {confidence- and context-aware weights}
    $z \leftarrow \sum_{t \in \{g,v,s,c\}} w_t \cdot \mathrm{Feat}_t(\mathcal{E}, F_t)$ {feature fusion}
    $S_h, U \leftarrow \mathrm{Integrator}(z, S_t, C)$ {holistic score and system-level confidence}
    $S_t \leftarrow \mathrm{Calibrate}(S_t, \mathcal{R})$; $S_h \leftarrow \mathrm{Calibrate}(S_h, \mathcal{R})$
Finalize Output
    return $S_h, S_t, F, U$
```

- **Validation:** 10% of essays, used for hyperparameter tuning and early stopping.
- **Test-In-Distribution (Test-ID):** 10% of essays drawn from the same prompt distribution as training, for in-distribution evaluation.
- **Test-Out-of-Distribution (Test-OOD):** 10% of essays from unseen prompts and topics, to measure robustness and generalization.

4.3 Dataset Statistics

On average, each essay contains approximately 280–350 Chinese characters (roughly 180–220 words in English equivalence). Across HSK levels, the dataset includes over 20,000 essays with balanced representation of writing proficiency. The HSK-Enhanced dataset further comprises 41 prompts (levels 4–6) with 8.3K original essays and rubric-aligned trait annotations. Each prompt includes an average of 20 keypoints across four traits, annotated by two certified examiners with inter-rater reliability of κ=0.82 and holistic QWK of 0.85. Data augmentation using paraphrasing, controlled error injection, and style transfer expands the training set by approximately 90% while preserving the overall score distribution. Comprehensive corpus statistics and reliability analyses are provided in Appendix A.

5 EXPERIMENTS

5.1 Experimental Setup

We evaluate CHESS on the HSK-enhanced dataset introduced in Section 4. All experiments are conducted on NVIDIA A800 GPUs using PyTorch.

To isolate the contribution of the multi-agent design, we additionally compare with a vanilla DeepSeek baseline prompted to produce only a holistic score without trait decomposition. For each run, we report results averaged over three random seeds.

5.2 Baselines

We compare CHESS against a range of representative AES systems:

- **LSTM-AES** [39]: A recurrent neural network model for essay scoring.
- **BERT-AES** [40]: A fine-tuned pre-trained language model achieving strong AES performance.
- **Prompt-based AES (T5)** [41]: A T5-based approach that leverages prompt conditioning for holistic scoring.
- **Rank-Then-Score (RTS)** [36]: A two-stage large language model framework that first performs pairwise ranking of essays and then calibrates absolute scores, representing the state-of-the-art in AES calibration accuracy.
- **DeepSeek v3 (vanilla) [42]**: The backbone model prompted to produce holistic scores, without trait-level decomposition.
- **CHESS (ours)**: A prompt-based multi-agent framework built on DeepSeek v3, with specialized agents for grammar, vocabulary, structure, and content.

These baselines represent both classical AES models and modern large language model approaches.

5.3 Evaluation Metrics

We follow common AES evaluation protocols:

- **Quadratic Weighted Kappa (QWK)**: Agreement between system predictions and human raters, robust to ordinal scoring scales.
- **Pearson Correlation (PCC)**: Correlation between predicted and true scores.
- **Mean Absolute Error (MAE)**: Absolute difference between system and human scores.
- **Trait-level Accuracy**: Agreement on grammar, vocabulary, structure, and content sub-scores with human annotations.

5.4 Main Results

Table 1 reports results on both in-distribution (Test-ID) and out-of-distribution (Test-OOD) settings. CHESS consistently outperforms baselines in QWK and PCC, with especially strong gains on Test-OOD prompts, demonstrating robustness to unseen topics. Trait-level evaluations further show that CHESS achieves higher alignment with human raters across all four scoring dimensions, validating the effectiveness of the multi-agent design.

5.5 Human Expert Comparison

To contextualize system performance, we compare CHESS with human expert raters. Specifically, we compute inter-rater agreement between pairs of certified HSK examiners on a subset of 1,000 essays covering all proficiency levels and prompts. Human agreement, measured by QWK, is approximately 0.84 for holistic scores and ranges from 0.78–0.83 for trait-level scores.

As shown in Table 2, CHESS approaches human-level agreement: achieving holistic QWK of 0.84 and trait-level QWKs between 0.78–0.84. These results indicate that CHESS not only outperforms strong neural baselines but also operates within the range of professional human graders.

5.6 Evaluation on Real Exam Essays

To further validate the practical value of CHESS, we evaluate the system on a held-out set of 500 essays sampled directly from official HSK examinations across levels 4–6. Each essay was scored independently by two certified human examiners following the official HSK rubrics, and disagreements were adjudicated by a senior linguistics expert.

As shown in Table 3, CHESS achieves holistic QWK of 0.83 and trait-level average QWK of 0.78 when compared with the adjudicated human scores. These results closely approach human–human agreement (0.84 holistic, 0.80 trait-level), and substantially outperform baseline AES models (best baseline: 0.75 holistic QWK). This demonstrates that CHESS generalizes well to real exam data and provides performance comparable to professional human graders.

5.7 Ablation Studies

To better understand the contribution of each component, we conducted a series of ablation experiments. Removing individual agents leads to measurable degradations in trait-level accuracy. For example, excluding the Grammar Agent decreases OOD QWK from 0.82 to 0.79 and increases MAE by 0.03, while removing the Structure Agent reduces QWK by 0.04 and raises MAE by 0.05, highlighting their importance for syntactic reliability and discourse coherence. The Content Agent also plays a key role in topical relevance: removing it lowers content trait agreement by 0.06 and holistic QWK by 0.03. The Vocabulary Agent has a smaller but still noticeable effect, with its removal reducing OOD QWK from 0.82 to 0.80. When data augmentation strategies (paraphrasing, error injection, style transfer) are removed, Test-OOD performance drops by 0.03 QWK and MAE increases by 0.02, confirming the importance of augmentation for generalization. Finally, replacing the Coordinator with a simple unweighted averaging scheme reduces holistic QWK by 3–5 points (from 0.84 to 0.79) and increases MAE by 0.09, underscoring the importance of adaptive score aggregation.

We further compare the coordinator's learned attention-based aggregation with a simple averaging baseline. As shown in Table 4, attention-based integration consistently improves agreement with human raters, demonstrating the benefit of modeling trait-specific reliability and essay-level context.

We further isolate the contribution of the **GraphRAG** mechanism within the Content Agent. Starting from a passage-level RAG baseline, progressively incorporating graph retrieval, edge-type compatibility, and path entailment reasoning leads to steady improvements on content QWK and MAE (Table 5). These results confirm that explicit modeling of rhetorical and logical relations enhances content evaluation quality beyond text-only retrieval.

5.8 Qualitative Analysis

We further analyze essays sampled across different proficiency levels (low-, mid-, and high-scoring). CHESS provides interpretable feedback, such as highlighting morpho-syntactic errors, detecting vocabulary repetition, and identifying discourse-level inconsistencies. Human raters confirm that such feedback aligns with HSK rubrics and offers pedagogically meaningful guidance beyond a holistic score.

5.9 Summary

Overall Performance. CHESS targets holistic QWK $\geq$ 0.82 and improves trait QWKs by ≈3–8 points over single-agent baselines, with reduced cross-prompt variance.

Interpretability and Feedback. Agent-level rationales and localized evidence provide transparent explanations. Teacher/student studies indicate higher specificity and cultural sensitivity.

6 DISCUSSION

The experimental results demonstrate that CHESS achieves both higher accuracy and stronger interpretability compared with existing AES systems. By decomposing essay scoring into multiple specialized agents, CHESS can provide trait-level feedback that aligns closely with HSK rubrics, offering practical value for both high-stakes testing and language learning support. Importantly, the framework shows robustness under out-of-distribution prompts, indicating that modular reasoning and data augmentation enhance generalization beyond training topics.

6.1 Interpretability and Pedagogical Value

One of the most significant outcomes is the ability of CHESS to deliver fine-grained, human-readable feedback. Unlike black-box neural models, CHESS highlights grammar errors, vocabulary repetition, discourse issues, and content relevance explicitly. Such interpretability not only increases trust from educators and policymakers but also provides learners with actionable guidance for improvement.

6.2 Limitations

Despite its advantages, several limitations remain. First, while the HSK-enhanced dataset covers a broad range of prompts and proficiency levels, the scope is still narrower than real-world essay contexts such as creative writing or domain-specific writing. Second, agent specialization relies on rubric-aligned annotations, which can be expensive to obtain at scale. Third, while the coordinator

Model	Test-ID (Seen Prompts)			Test-OOD (Unseen Prompts)		
	QWK ↑	PCC ↑	MAE ↓	QWK ↑	PCC ↑	MAE ↓
LSTM-AES	0.68	0.65	0.74	0.61	0.59	0.83
BERT-AES	0.77	0.74	0.61	0.70	0.68	0.72
Prompt-based AES (T5)	0.75	0.73	0.58	0.72	0.70	0.69
DeepSeek v3 (vanilla)	0.79	0.77	0.61	0.74	0.72	0.67
Rank-Then-Score	0.80	0.78	0.63	0.70	0.72	0.68
CHESS (ours)	**0.85**	**0.82**	**0.49**	**0.82**	**0.80**	**0.53**

Table 1: Main results on the HSK-enhanced dataset. We report performance on in-distribution (Test-ID) and out-of-distribution (Test-OOD) prompts. Metrics include Quadratic Weighted Kappa (QWK), Pearson Correlation Coefficient (PCC), and Mean Absolute Error (MAE). CHESS consistently outperforms strong baselines, especially under OOD conditions.

	Holistic QWK	Avg. Trait QWK
Human–Human Agreement	0.84	0.80
Best Baseline (Rank-Then-Score)	0.77	0.73
CHESS (ours)	**0.84**	**0.81**

Table 2: Comparison of CHESS and baselines with human inter-rater agreement. Trait QWK is averaged across grammar, vocabulary, structure, and content dimensions.

	Holistic QWK	Avg. Trait QWK
Human–Human Agreement	0.84	0.80
DeepSeek v3 (vanilla)	0.75	0.70
CHESS (ours)	**0.83**	**0.78**

Table 3: Performance on real HSK exam essays. CHESS approaches human-level agreement and surpasses baseline AES systems.

Coordinator Method	Holistic QWK ↑	PCC ↑	MAE ↓
Simple Averaging	0.79	0.75	0.58
Attention-based (ours)	**0.84**	**0.82**	**0.49**

Table 4: Comparison of coordinator strategies. Our attention-based integration outperforms simple averaging across all metrics, demonstrating the importance of learned aggregation.

effectively integrates agent outputs, it may still propagate systematic biases if all agents share similar weaknesses. Although prompt-based specialization is efficient and interpretable, it may be less stable than fine-tuned models in edge cases. Future work could explore hybrid approaches, where lightweight fine-tuning (e.g., LoRA adapters) is used to complement prompt engineering for traits that are particularly difficult to capture (e.g., discourse coherence).

Method	Content QWK ↑	MAE ↓
Passage RAG only	0.806	0.599
+ Graph retrieval (no edges)	0.812	0.583
+ Edge-type compatibility	0.828	0.496
+ Path entailment verifier	0.834	0.492
+ Theme coverage (cluster-weighted)	**0.843**	**0.485**

Table 5: Ablations of GraphRAG components on HSK dataset. Each step incrementally adds graph reasoning modules, showing consistent gains in content trait accuracy and reduced error.

6.3 Future Directions and Broader Implications

Future research can address these limitations along three lines. (1) **Dataset expansion:** Incorporating essays from diverse domains and proficiency contexts would strengthen robustness and fairness. (2) **Adaptive agent design:** Leveraging self-improving or reinforcement-learned agents could enhance flexibility in handling novel error types. (3) **Cross-linguistic generalization:** Extending the multi-agent AES paradigm to other non-alphabetic or morphologically complex languages (e.g., Japanese, Korean, Arabic) would further test the scalability of this approach.

The CHESS framework exemplifies how multi-agent LLM systems can move beyond single-task black boxes toward interpretable, domain-sensitive evaluation. This direction opens possibilities not only in language assessment but also in other domains where fairness, explainability, and cultural alignment are critical, such as medical report analysis and legal document review. *More broadly, our work highlights the potential of multi-agent reasoning as a general paradigm for extending automated evaluation to linguistically and culturally diverse contexts.*

ETHICAL STATEMENT

We address fairness across proficiency levels and backgrounds, provide contestability mechanisms and explanations, and restrict use to formative/low-stakes settings unless validated. Data are anonymized with consent; raters are compensated. Cultural corpora are curated to avoid harmful stereotypes.

REFERENCES

[1] You-Jin Jong, Yong-Jin Kim, and Ok-Chol Ri. Improving performance of automated essay scoring by using back-translation essays and adjusted scores. *Mathematical Problems in Engineering*, 2022(1):6906587, 2022.

[2] Wei Song, Kai Zhang, Ruiji Fu, Lizhen Liu, Ting Liu, and Miaomiao Cheng. Multi-stage pre-training for automated chinese essay scoring. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 6723–6733, 2020.

[3] Yaqiong He, Feng Jiang, Xiaomin Chu, and Peifeng Li. Automated chinese essay scoring from multiple traits. In *Proceedings of the 29th international conference on computational linguistics*, pages 3007–3016, 2022.

[4] Leyi Qian, Yali Zhao, and Yan Cheng. Evaluating china's automated essay scoring system iwrite. *Journal of Educational Computing Research*, 58(4):771–790, 2020.

[5] Abejide Olu Ade-Ibijola, Ibiba Wakama, and Juliet Chioma Amadi. An expert system for automated essay scoring (aes) in computing using shallow nlp techniques for inferencing. *International Journal of Computer Applications*, 51(10):37–45, 2012.

[6] Shuai Yuan, Tingting He, Huan Huang, Rui Hou, and Meng Wang. Automated chinese essay scoring based on deep learning. *Computers, Materials & Continua*, 65(1), 2020.

[7] Haiyue Feng, Sixuan Du, Gaoxia Zhu, Yan Zou, Poh Boon Phua, Yuhong Feng, Haoming Zhong, Zhiqi Shen, and Siyuan Liu. Leveraging large language models for automated chinese essay scoring. In *International Conference on Artificial Intelligence in Education*, pages 454–467. Springer, 2024.

[8] Vivekanandan Kumar and David Boulanger. Explainable automated essay scoring: Deep learning really has pedagogical value. In *Frontiers in education*, volume 5, page 572367. Frontiers Media SA, 2020.

[9] Vivekanandan S Kumar and David Boulanger. Automated essay scoring and the deep learning black box: How are rubric scores determined? *International Journal of Artificial Intelligence in Education*, 31(3):538–584, 2021.

[10] Jill Burstein. The e-rater scoring engine: Automated essay scoring with natural language processing. *Automated essay scoring: A cross-disciplinary perspective*, 113121, 2003.

[11] Atsushi Mizumoto and Masaki Eguchi. Exploring the potential of using an ai language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2):100050, 2023.

[12] Vanessa De Wilde and Orphée De Clercq. Challenges and opportunities of automated essay scoring for low-proficient l2 english writers. *Assessing Writing*, 66:100982, 2025.

[13] Yigal Attali. Validity and reliability of automated essay scoring. In *Handbook of automated essay evaluation*, pages 181–198. Routledge, 2013.

[14] Yiran Hou. Implications of aes system of pigai for self-regulated learning. *Theory and Practice in Language Studies*, 10(3):261–268, 2020.

[15] Yaman Kumar Singla, Swapnil Parekh, Somesh Singh, Junyi Jessy Li, Rajiv Ratn Shah, and Changyou Chen. Aes systems are both overstable and oversensitive: Explaining why and proposing defenses. *arXiv preprint arXiv:2109.11728*, 2021.

[16] Tsegaye Misikir Tashu, Chandresh Kumar Maurya, and Tomas Horvath. Deep learning architecture for automatic essay scoring. *arXiv preprint arXiv:2206.08232*, 2022.

[17] Chun Then Lim, Chih How Bong, Wee Sian Wong, and Nung Kion Lee. A comprehensive review of automated essay scoring (aes) research and development. *Pertanika Journal of Science & Technology*, 29(3):1875–1899, 2021.

[18] Rui Xiao, Wenbin Guo, Yunchun Zhang, Xiaoyan Ma, and Jiaqi Jiang. Machine learning-based automated essay scoring system for chinese proficiency test (hsk). In *Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval*, pages 18–23, 2020.

[19] Xingyuan Peng, Dengfeng Ke, Zhenbiao Chen, and Bo Xu. Automated chinese essay scoring using vector space models. In *2010 4th International universal communication symposium*, pages 149–153. IEEE, 2010.

[20] Haojin Li and Tao Dai. Explore deep learning for chinese essay automated scoring. In *Journal of Physics: Conference Series*, volume 1631, page 012036. IOP Publishing, 2020.

[21] Kaixun Yang, Mladen Raković, Zhiping Liang, Lixiang Yan, Zijie Zeng, Yizhou Fan, Dragan Gašević, and Guanliang Chen. Modifying ai, enhancing essays: How active engagement with generative ai boosts writing quality. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 568–578, 2025.

[22] Semire Dıklı. Automated essay scoring. *Turkish Online Journal of Distance Education*, 7(1):49–62, 2006.

[23] Ronan Cummins and Marek Rei. Neural multi-task learning in automated assessment. *arXiv preprint arXiv:1801.06830*, 2018.

[24] VV Ramalingam, A Pandian, Prateek Chetry, and Himanshu Nigam. Automated essay grading using machine learning algorithm. In *Journal of Physics: Conference Series*, volume 1000, page 012030. IOP Publishing, 2018.

[25] Yigal Attali and Jill Burstein. Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment*, 4(3), 2006.

[26] Peter W Foltz, Darrell Laham, and Thomas K Landauer. Automated essay scoring: Applications to educational technology. In *Edmedia+ innovate learning*, pages 939–944. Association for the Advancement of Computing in Education (AACE), 1999.

[27] Mohamed Abdellatif Hussein, Hesham Hassan, and Mohammad Nassef. Automated language essay scoring systems: A literature review. *PeerJ Computer Science*, 5:e208, 2019.

[28] Austin Pack, Alex Barrett, and Juan Escalante. Large language models and automated essay scoring of english language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, 6:100234, 2024.

[29] Masaki Uto, Itsuki Aomi, Emiko Tsutsumi, and Maomi Ueno. Integration of prediction scores from various automated essay scoring models using item response theory. *IEEE Transactions on Learning Technologies*, 16(6):983–1000, 2023.

[30] Chao Han, Xiaolei Lu, and Shirong Chen. Modeling rater judgments of interpreting quality: Ordinal logistic regression using neural-based evaluation metrics, acoustic fluency measures, and computational linguistic indices. *Research Methods in Applied Linguistics*, 4(1):100194, 2025.

[31] Tao-Hsing Chang and Yao-Ting Sung. Smartwriting-mandarin: An automated essay scoring system for chinese foreign language learners. In *The Routledge International Handbook of Automated Essay Evaluation*, pages 78–90. Routledge, 2024.

[32] Hongwu Yang, Yanshan He, Xiaolong Bu, Hongwen Xu, and Weitong Guo. Automatic essay evaluation technologies in chinese writing—a systematic literature review. *Applied Sciences*, 13(19):10737, 2023.

[33] Ahmed M ElMassry, Nazar Zaki, Negmeldin AlSheikh, and Mohammed Mediani. A systematic review of pretrained models in automated essay scoring. *IEEE Access*, 2025.

[34] Jinhee Jang, Ayoung Moon, Minkyoung Jung, YoungBin Kim Lee, et al. Llm agents at the roundtable: A multi-perspective and dialectical reasoning framework for essay scoring. *arXiv preprint arXiv:2509.14834*, 2025.

[35] Jiamin Su, Yibo Yan, Zhuoran Gao, Han Zhang, Xiang Liu, and Xuming Hu. Cafes: A collaborative multi-agent framework for multi-granular multimodal essay scoring. *arXiv preprint arXiv:2505.13965*, 2025.

[36] Yida Cai, Kun Liang, Sanwoo Lee, Qinghan Wang, and Yunfang Wu. Rank-then-score: Enhancing large language models for automated essay scoring. *arXiv preprint arXiv:2504.05736*, 2025.

[37] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

[38] Stephen E Robertson and Steve Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*, pages 232–241. Springer, 1994.

[39] Huang Chimingyang. An automatic system for essay questions scoring based on lstm and word embedding. In *2020 5th International Conference on Information Science, Computer Technology and Transportation (ISCTT)*, pages 355–364. IEEE, 2020.

[40] Ridha Hussein Chassab, Lailatul Qadri Zakaria, and Sabrina Tiun. An optimized bert fine-tuned model using an artificial bee colony algorithm for automatic essay score prediction. *PeerJ Computer Science*, 10:e2191, 2024.

[41] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[42] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.