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

Supplementary Material

Xinzhuo Hou, Wanli Ouyang, Taoyong Cui, Pheng-Ann Heng
The Chinese University of Hong Kong

Appendix A: Prompt Templates for CHESS Agents

This supplementary document provides the full prompt templates used in the CHESS system to evaluate essays. Each agent performs a specialized role corresponding to a trait in the HSK rubric, and their outputs are combined by a coordinator into the final holistic score.

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: subject-predicate agreement, function words, aspect markers,
       classifiers, and word order.
  - Identify common learner errors such as subject-predicate 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 0-5 (0 = unintelligible, 5 = error-free).

STEPS:
  1. Detect and categorize morpho-syntactic errors (subject-predicate 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 (0-5).

OUTPUT_SCHEMA (JSON):
{
  "Score": <int 0-5>,
  "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>
```

Listing 2 Prompt Template: Vocabulary Agent

```
ROLE:
  You are a Chinese vocabulary examiner specializing in lexical diversity, appropriateness, and
      register control for HSK writing.

GOAL:
  Evaluate the richness, precision, and appropriateness of word usage according to the HSK
      vocabulary rubric.

CONSTRAINTS:
  - Focus on lexical diversity, collocation, register control, and idiomatic usage.
  - Consider lexical sophistication indices (e.g., type-token ratio, rare-word coverage) and
      embedding-based semantic similarity to rubric exemplars.
  - Identify overuse of simple words or repetitive expressions.
  - Reward appropriate use of idioms and advanced vocabulary.
  - Score range: integer 0-5.

STEPS:
  1. Analyze lexical variety and sophistication (type-token ratio, rare-word coverage).
  2. Compute semantic similarity between essay vocabulary and rubric exemplar embeddings.
  3. Detect misused or awkward word choices and collocation errors.
  4. Assess appropriateness of vocabulary for topic and proficiency level.
  5. Provide examples of excellent and problematic usage.
  6. Assign Vocabulary Score (0-5).

OUTPUT_SCHEMA (JSON):
{
  "Score": <int 0-5>,
  "Highlights": [
    {"type": "good", "span": "<well-chosen word or phrase>", "note": "<why appropriate>"},
    {"type": "bad", "span": "<problematic word>", "note": "<reason or correction suggestion>"}
  ],
  "Explanation": "<summary comment in Chinese on vocabulary range, register, and accuracy>"
}

INPUT:
<Essay text here>
```

Listing 3 Prompt Template: Structure Agent

```
ROLE:
  You are a discourse-structure examiner for Chinese essays.

GOAL:
  Evaluate discourse-level organization, paragraph coherence, and logical flow following Chinese
      writing conventions (Qi–Cheng–Zhuan–He).

CONSTRAINTS:
  - Focus on global structure: introduction, development, turn, and conclusion.
  - Examine cohesion and progression across sentences and paragraphs using implicit similarity of
       meaning (embedding-level reasoning).
  - Identify abrupt topic shifts, missing transitions, or incoherent paragraphing.
  - Assess the balance and logical sequencing of ideas.
  - Score range: integer 0–5.

STEPS:
  1. Analyze overall discourse organization and identify each paragraph’ s role (Qi–Cheng–Zhuan–
      He).
  2. Evaluate semantic continuity between adjacent paragraphs (topic relevance and transition
      strength).
  3. Detect abrupt or incoherent transitions indicating weak rhetorical linkage.
  4. Comment on clarity, logical development, and global cohesion.
  5. Assign Structure Score (0–5).

OUTPUT_SCHEMA (JSON):
{
  "Score": <int 0–5>,
  "Issues": [
    {
      "category": "<missing-transition | abrupt-shift | poor-organization | weak-conclusion>",
      "span": "<sentence or paragraph excerpt>",
      "note": "<diagnostic comment>"
    }
  ],
  "Explanation": "<overall comment in Chinese summarizing discourse organization, cohesion, and
      logical flow>"
}

INPUT:
<Essay text here>
```

Listing 4 Prompt Template: Content Agent

```
ROLE:
  You are a content examiner evaluating topical relevance, argument completeness, and factual
       correctness of Chinese essays.

GOAL:
  Measure how well the essay addresses the assigned topic and expected rubric keypoints, using
       prompt-based retrieval augmentation to compare with reference materials and exemplar
       answers from the GraphRAG-CHESS knowledge base.

CONSTRAINTS:
  - Focus on topical relevance, logical argumentation, and factual accuracy.
  - Retrieve rubric keypoints and exemplar arguments through the GraphRAG namespace for
       comparison.
  - Evaluate argument coverage, reasoning depth, and factual grounding.
  - Penalize digressions, irrelevant narratives, or unsupported claims.
  - Score range: integer 0-5 (0 = off-topic or factually wrong, 5 = complete and well-grounded).

STEPS:
  1. Identify the essay's main claims and supporting ideas.
  2. Retrieve relevant rubric keypoints and exemplar arguments via GraphRAG retrieval.
  3. Compare the essay's content with these references to assess topical coverage and factual
       consistency.
  4. Compute reasoning depth based on how well arguments are supported by evidence or logic.
  5. Highlight missing or incorrect arguments, or unsupported claims.
  6. Summarize findings and assign a final Content Score (0-5).

OUTPUT_SCHEMA (JSON):
{
  "Score": <int 0-5>,
  "Matched_Keypoints": [
    {"keypoint": "<rubric-item-id>", "match": "<essay span>", "similarity": "<float 0-1>"}
  ],
  "Missing_Keypoints": ["<rubric-item-id>"],
  "Explanation": "<summary in Chinese describing topical coverage, argument quality, and factual
       correctness>"
}

INPUT:
{
  "Essay": "<essay text here>",
  "Prompt": "<official HSK prompt>",
  "Rubric_Graph": "<GraphRAG namespace ID>"
}
```

Appendix B: Dataset Details and Augmentation Examples

The HSK-Enhanced dataset extends the original HSK corpus through controlled augmentation strategies designed to improve data diversity and coverage of linguistic phenomena. Three augmentation modes are used: (1) paraphrasing for lexical and syntactic diversity, (2) error injection for robustness in grammar and usage, and (3) style transfer for variation in register and tone. All augmented samples preserve semantic meaning and remain aligned with rubric keypoints to ensure fair scoring.

Each augmented subset is tagged with its generation source, transformation type, and confidence score. During training, these tags are used as auxiliary supervision signals to encourage trait-level robustness. Error-injected data helps the Grammar Agent recognize common learner mistakes, while style-transferred data exposes the Vocabulary and Structure Agents to richer syntactic and pragmatic variations.

Table 1: Examples of Augmentation Strategies in the HSK-Enhanced Dataset. Each original essay sentence is paired with its augmented version and the applied transformation type.

Type	Original Sentence	Augmented Version
Paraphrase	我认为这个活动非常有意义。	我觉得这次活动特别有价值。
Error Injection	他们昨天去了图书馆学习。 (swap aspect marker and word order to simulate learner error)	他们昨天去图书馆学习了。
Style Transfer	这场演讲让我印象深刻。	这次分享会真让我感到震撼。
Error Injection (vocabulary misuse)	我希望可以提高我的中文水平。 (replace “提高”with “增加”to simulate lexical misuse)	我想可以增加我的中文水平。
Paraphrase (syntactic reordering)	如果有机会，我想再去一次北京。	我想如果有机会的话，再去一趟北京。

Appendix C: Knowledge Sources for GraphRAG-CHESS

The retrieval-augmented component of CHESS, termed **GraphRAG-CHESS**, is built upon a curated corpus that integrates rubric-aligned pedagogical materials, exemplar essays, and background reference texts. These sources form the node set V and edge relations E of the retrieval graph, which are indexed for hybrid retrieval (BM25 + dense encoder). The major categories of source materials are summarized below.

Each node in the graph thus corresponds to either a rubric descriptor, exemplar span, or background snippet. Edges encode rhetorical and logical relations (supports, elaborates, contradicts, defines), and Louvain community detection is used to cluster semantically related regions. Hybrid retrieval employs both sparse (BM25) and dense (multilingual MiniLM) indices, while a graph-aware reranker evaluates edge-type compatibility and contextual coherence.

All sources are either publicly available or reproduced under educational fair-use, and no proprietary test materials or personally identifiable information are included.

Table 2: Knowledge sources and corresponding functions in GraphRAG-CHESS.

Source Type	**Description and Role in GraphRAG-CHESS**
HSK Rubric Manuals	Official HSK writing rubrics and level descriptors (HSK 3–6) published by the Chinese Testing International (CTI). Used to extract *rubric keypoints* (e.g., ”clarity of theme”, ”accuracy of grammar”, ”appropriate word choice”) as base nodes in the graph.
Exemplar Essays	1,200 high- and mid-scoring sample compositions collected from publicly released HSK writing samples, annotated with trait-level comments by certified examiners. Used to construct *exemplar spans* and to attach supporting evidence edges (”supports”, ”elaborates”).
Learner Essays (HSK-Enhanced)	The same corpus used for model training, included here to supply realistic learner language expressions. Only de-identified text segments are retained. Edges connect common learner errors or expressions to corresponding rubric keypoints (”contradicts”, ”violates”).
Pedagogical Guides and Linguistic References	Excerpts from Chinese language teaching materials such as 《汉语水平考试写作指导》(2021), 《HSK 标准教程》(Books 3–6), and reference grammar compendia. Provide background snippets for contextual retrieval (”supports” or ”defines”).
Topic Background Corpus	3.4k short encyclopedia-style entries from open Chinese corpora (e.g., Baidu Baike, WikiZh) for essay topics like ” 环境保护”, ” 网络安全”, and ” 文化交流”. Used to supply factual grounding and content elaboration edges (”provides_evidence”).
Rubric–Essay Mapping Annotations	A small expert-labeled set (≈ 500 essays) aligning sentences to rubric keypoints and scoring dimensions. Serves as supervised signal for community detection and graph edge type learning.

Appendix D: Coordinator Network Implementation

D.1 Overview

The Coordinator module in CHESS is a lightweight neural network that learns to combine the four trait-level scores (*Grammar*, *Vocabulary*, *Structure*, and *Content*) into a calibrated holistic score. While the individual agents operate purely via prompt-based inference, the Coordinator is trained via multi-task regression to model inter-trait dependencies and scoring reliability. This design preserves interpretability while improving holistic accuracy and stability.

D.2 Input Representation

Each essay yields trait-level outputs

$$S_t \in [0, 5], \quad C_t \in [0, 1], \quad h_t \in \mathbb{R}^d, \quad t \in \{g, v, s, c\},$$

where S_t is the numerical score, C_t the confidence estimate, and h_t a dense embedding representing the agent's textual rationale (typically obtained by mean-pooling a sentence embedding of the agent' s explanatory feedback). In addition, a prompt-level context vector u is extracted from essay metadata (e.g., length, HSK level, topic ID, keypoint coverage).

The input feature is constructed as

$$x = [S_g, S_v, S_s, S_c,\, C_g, C_v, C_s, C_c,\, u],$$

which serves as the Coordinator' s input.

D.3 Architecture

The Coordinator adopts a shallow multi-layer perceptron (MLP) with an attention-based fusion head. Given trait embeddings h_t, it computes attention weights

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

and outputs the holistic score

$$S_h = \sum_t \alpha_t S_t.$$

The hidden feature z is obtained as

$$z = \mathrm{ReLU}(W_1 x + b_1), \quad z' = \mathrm{Dropout}(0.1)(\mathrm{ReLU}(W_2 z + b_2)).$$

A confidence head predicts system-level uncertainty $\hat{U}$ through a small regression branch:

$$\hat{U} = \sigma(W_u z' + b_u),$$

where σ denotes a sigmoid activation.

D.4 Training Objective

The Coordinator is trained on a development set with available holistic human scores S^* using a multi-objective loss:

$$\mathcal{L} = \lambda_1 \, \text{Huber}(S_h, S^*) + \lambda_2 \, \text{KL}(\alpha \, \| \, \alpha^{\text{prior}}) + \lambda_3 \, \text{CalibLoss}(S_h),$$

where $\alpha^{\text{prior}} = [0.25, 0.25, 0.25, 0.25]$ or rubric-based priors (Content=0.30, Grammar=0.25, Structure=0.25, Vocabulary=0.20). The calibration loss enforces monotonic alignment between predicted holistic scores and observed grade ranks across prompts.

All parameters are optimized using AdamW (lr $= 5 \times 10^{-4}$, weight decay 10^{-4}), with early stopping based on validation QWK (Quadratic Weighted Kappa). Temperature scaling is applied post-training to normalize score ranges to [0,5].

D.5 Uncertainty and Contestability Routing

During inference, the Coordinator performs Monte Carlo dropout ($p = 0.1$, $N = 10$ forward passes) to estimate prediction variance $\hat{U}$. If $\hat{U} > \tau$ (95th percentile threshold) or trait disagreement $\max_t |S_t - \bar{S}| > \delta$ (typically $\delta = 1.0$), a *contestability flag* is raised, recommending human re-evaluation or secondary scoring. This mechanism prevents overconfident predictions and ensures reliability under ambiguous or off-topic essays.

D.6 Pseudocode

```
# Training
for batch in loader:
    x = concat(S, C, u)
    z = ReLU(W1 @ x + b1)
    z = Dropout(0.1)(ReLU(W2 @ z + b2))
    alpha = softmax([h_t.T @ Wc @ u for t in traits])
    Sh = sum(alpha_t * S_t for t in traits)
    loss = huber(Sh, S_star) + kl(alpha, alpha_prior) + calib(Sh)
    loss.backward(); opt.step()

# Inference
with mc_dropout:
    Sh_list = []
    for _ in range(10):
        Sh_list.append(forward(S, C, u, h))
    Sh_mean = mean(Sh_list)
    U = var(Sh_list)
if U > tau or range(S) > delta:
    flag_contest = True
return Sh_mean, alpha, U, flag_contest
```

Listing 1: Coordinator Training and Inference Pipeline

Appendix E: GraphRAG Construction Details.

We adopt the official Microsoft GraphRAG pipeline [?] (`https://microsoft.github.io/graphrag`) to construct a rubric-aligned knowledge graph for each writing prompt. All rubric descriptors, exemplar essays, and background notes are first segmented into standardized text units (200–500 characters) and processed through entity and relation extraction to identify salient concepts and rhetorical links. These units are then transformed into a directed graph $G = (V, E)$, where nodes correspond to rubric keypoints, exemplar spans, and background snippets, and edges encode discourse or logical relations such as *supports*, *contradicts*, and *elaborates.* Hierarchical community detection (Leiden/Louvain) clusters semantically related nodes, and each cluster is summarized to capture higher-level themes. During inference, CHESS performs subgraph retrieval via hybrid BM25 + dense-vector indices and graph-aware reranking that jointly optimizes node similarity and edge-type compatibility. This implementation follows the GraphRAG modular pipeline—text unitization → entity/relation extraction → graph construction → community summarization → graph-based retrieval—providing interpretable, efficient, and auditable reasoning for content evaluation.