

Article

CAMA: A Confidence-Aware Multi-Agent System with Adaptive Depth Routing for Automated Writing Assessment

Zhiwu Yin ¹, Jieren Zhang ¹, Jing Wang ¹, Jianqiang Liang ¹, Xing Wu ² and Jianbiao Dai ^{1,3,*}

¹ Netclass Technology Inc., Shanghai 200436, China; yinzhiwu@netclasstech.com (Z.Y.); zhangjieren@netclasstech.com (J.Z.); wangjing@netclasstech.com (J.W.); hymanliang@netban.cn (J.L.)

² School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China; xingwu@shu.edu.cn

³ Institute for Data Engineering and Sciences, Academy of Intelligent Health & Longevity Science, University of Saint Joseph, Macao, China

* Correspondence: dai.jian.biao@usj.edu.mo

Abstract

Existing multi-agent systems for automated writing assessment (AWA) suffer from three structural limitations: uniform agent weighting, fixed-depth computation, and lack of empirical auditing. To resolve these challenges, we propose CAMA, a Confidence-Aware Multi-Agent architecture consisting of three synergistic layers: Confidence-Weighted Aggregation (CWA), Adaptive Depth Routing (ADR), and Contribution-Guided Auditing (CGA). We evaluate its human alignment on 31 EFL essays and extend structural robustness checks to 151 samples across multiple datasets and 8 writing prompts. CAMA addresses the accuracy–efficiency–transparency trilemma. The experimental results show the following: (1) CWA provides a formal error containment bound in simulated agent failure, and multi-agent role decomposition achieves substantial improvement over a holistic baseline (QWK = +0.246); (2) ADR optimizes efficiency by routing 48.4% of standard essays to a low-cost Fast path while scaling up for difficult tasks; and (3) CGA provides an interpretable audit mechanism to quantify per-agent contributions. This framework offers a promising, transparent and principled paradigm for intelligent educational evaluation, with the caveat that broader validation across larger, more diverse populations remains a critical avenue for future work.

Keywords: multi-agent systems; large language models; automated writing assessment; confidence calibration; Shapley value; adaptive routing; system architecture; intelligent evaluation systems

Academic Editor: George A. Tsihrintzis

Received: 24 June 2026

Revised: 22 July 2026

Accepted: 23 July 2026

Published: 24 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\) license](https://creativecommons.org/licenses/by/4.0/).

1. Introduction

As large language models (LLMs) have achieved remarkable fluency across text generation, code synthesis, and dialog [1,2], a fundamental challenge grows increasingly acute: how to reliably evaluate the quality of their outputs. Traditional evaluation paradigms—human annotation, automated metrics, and single-model LLM-as-a-Judge approaches [3,4]—each face structural limitations. Human annotation is expensive and exhibits substantial inter-rater variability [5,6]. Automated metrics capture superficial lexical overlap but miss semantic quality. Single-model LLM-as-a-Judge suffers from systematic biases including position bias, self-enhancement bias, and length bias [7,8].

Multi-agent architectures have emerged as a principled methodological response to this challenge. The core intuition is that when multiple models collaborate, debate, and form a consensus through structured interaction, the resulting distributed judgment can overcome the limitations of any single evaluator. This approach draws on ensemble theory—multiple weak judges can outperform a single strong one when their errors are decorrelated [9] while adding the capability of structured deliberation, which makes the reasoning process architecturally visible.

Writing assessment offers an ideal lens through which to study multi-agent evaluation. Automated Essay Scoring (AES) provides a domain in which evaluation criteria are well-defined: scoring rubrics provide explicit criteria, decades of psychometric research [5,6] establishes construct validity frameworks, and standardized benchmarks enable systematic comparison. Over nearly six decades—from Page’s pioneering Project Essay Grade (PEG) through e-rater and contemporary neural approaches [10,11]—AES research has undergone three major paradigm shifts, each seeking to understand the extent to which machines can render meaningful judgments of quality.

The emergence of large language models such as GPT-4, Claude, and Gemini has substantially advanced this frontier. Unlike earlier deep learning models that framed scoring as text classification, LLMs leverage emergent semantic understanding to not only produce scores but also to generate commentary and identify argumentation flaws. However, LLM-based scoring has also introduced new challenges: when a scoring model serves as both “judge” and “explainer,” verifying the reliability of its explanations becomes difficult [3,12]. Kim and McCrudden [13] found that LLM scores were consistently higher and less variable than human raters across all dimensions—raising questions about whether this reflects more accurate judgment or systematic neglect of certain scoring dimensions.

For writing assessment, the advantages of multi-agent architectures are especially pronounced because writing quality is inherently multidimensional (depth of content, structural organization, linguistic expression, argumentation logic). This characteristic naturally lends itself to decomposition into specialized sub-tasks handled by different agents. Thapa et al. [14] demonstrated this through an agentic AI framework for low-resource essay evaluation: multi-agent decomposition with specialized scoring, explanation, and debate modules achieved approximately 7% QWK improvement over holistic single-agent baselines on the ASAP 2.0 dataset.

Despite these advances, existing multi-agent evaluation systems share three structural limitations. First, they treat all agent outputs as equally reliable, using fixed-weight or simple averaging aggregation that ignores per-output confidence signals. Second, they employ fixed-depth computation paths regardless of scoring difficulty, wasting resources on straightforward cases and underserving genuinely difficult ones. Third, they lack feedback loops for empirically evaluating whether each agent meaningfully contributes to scoring accuracy, making agent design a matter of intuition rather than evidence.

This paper proposes the Confidence-Aware Multi-Agent (CAMA) architecture, which addresses these limitations through three synergistic layers: Confidence-Weighted Aggregation (CWA), which calibrates each agent’s influence based on per-output reliability signals; Adaptive Depth Routing (ADR), which allocates computation proportionally to scoring difficulty; and Contribution-Guided Auditing (CGA), which uses Shapley values from cooperative game theory to guide agent selection. Specifically, CWA addresses the first limitation (uniform weighting) by down-weighting unreliable agents automatically, ADR addresses the second limitation (fixed-depth computation) by reserving expensive deep processing for genuinely difficult cases, and CGA addresses the third limitation (no feedback) by quantifying each agent’s marginal contribution through Shapley values. The main contributions are:

1. The CAMA architecture, a novel three-layer framework integrating CWA, ADR, and CGA that addresses the specific limitations of current multi-agent evaluation systems: uniform agent weighting, fixed-depth computation paths, and the absence of empirical agent optimization.
2. A systematic cross-system analysis of representative multi-agent evaluation systems (ChatEval [15], RES [16], MACS [9], ConfMAD [17], Chu et al. [18], AutoSCORE [19], MAGIC [20], LAWE-CL2 [21], and CourtEval [22]) from which the design principles underlying CAMA are derived, revealing the accuracy–efficiency–transparency trilemma as a structural constraint on multi-agent evaluation design.
3. Empirical evaluation on 31 expert-rated EFL essays, which demonstrates encouraging human alignment (QWK = 0.841, +0.246 improvement over holistic baseline), resource efficiency (48.4% Fast path routing), and robust graceful degradation under agent failure, with cross-dataset structural robustness further validated on an external set of 151 samples. We explicitly note the boundary conditions: modest sample size, single L1 background, and a single expert rater per essay.
4. Evidence that all four CEFR agent roles contribute meaningfully above a redundancy threshold (Shapley $\phi > 0.10$ for all dimensions), providing empirical validation of the multi-agent decomposition approach to writing assessment.

By addressing the accuracy–efficiency–transparency trilemma, CAMA provides a framework for deploying multi-agent systems in high-stakes educational applications, contributing to the emerging ecosystem of MAS-driven intelligent tutoring and personalized learning.

From a system engineering perspective, CAMA addresses a practical challenge in deploying LLM-based multi-agent evaluation systems. The goal is to design an architecture that achieves reliable distributed judgment, allocates computational resources efficiently, and enables transparent agent contribution assessment. The CWA layer provides a principled mechanism for confidence-aware information fusion across heterogeneous scoring agents; ADR introduces a decision-theoretic framework for instance-adaptive computation allocation that reserves expensive processing for genuinely difficult cases; and CGA contributes a game-theoretic auditing protocol for system-level quality assurance of agent composition. The integration of these three layers into a coherent prototype system demonstrates that adaptive architectural design—rather than static engineering trade-offs—offers a pathway toward scalable, efficient, and auditable multi-agent systems for real-world AI evaluation tasks. This work contributes to the growing field of multi-agent system engineering for LLM-based evaluation, with implications for the design of reliable AI assessment infrastructure in educational technology and beyond.

2. Related Work

Existing multi-agent writing assessment systems can be broadly categorized into three architectural types. Role-specialized architectures decompose evaluation along writing quality dimensions. Agentic AI frameworks [14] achieved ~7% QWK improvement on ASAP 2.0 through scoring–explanation–debate modules. LAWE-CL2 [21] targets Chinese L2 writers with role-specialized agents. MAGIC [20] demonstrated that Gemma 3 12B can outperform Llama 3.3 70B with effective role specialization. Iterative refinement architectures use multi-round score–feedback cycles: Chu et al. [18] showed that rationale-based scoring improves accuracy and interpretability, while AutoSCORE [19] transforms scoring into an auditable evidence–judgment chain through structured component recognition. Deliberative consensus architectures employ structured debate among equal agents. The foundational multi-agent debate framework [23] demonstrated that structured deliberation among LLM instances improves output quality, establishing the paradigm upon which these systems build. RES [16] uses persona-based multi-perspective evaluation.

ConfMAD [17] and D3 [24] introduce confidence-aware debate with adaptive stopping. MACS [9] demonstrates that cognitive diversity across model architectures is key to bias mitigation, with heterogeneous ensembles achieving 92.3% scoring accuracy and 63% variance reduction.

Despite this diversity, these systems share three structural limitations that CAMA addresses:

1. **Uniform agent weighting:** All surveyed systems treat each agent's score as equally informative, despite evidence that confidence signals vary substantially across agents, roles, and instances [24,25].
2. **Fixed-depth computation:** Every essay triggers the same computation path regardless of scoring difficulty, wasting resources on straightforward cases. Ferrag et al. [26] comprehensively surveyed the landscape of LLM-driven autonomous agents, documenting that architectural choices in multi-agent frameworks introduce efficiency variations of up to two orders of magnitude.
3. **No contribution feedback:** None of the surveyed systems provide empirical mechanisms for evaluating whether each agent meaningfully contributes to scoring accuracy, leaving agent selection to researcher intuition rather than evidence.

The comparative analysis in Table 1 reveals a clear pattern: all existing systems employ fixed-depth computation paths (no instance-adaptive routing), uniform or simple averaging aggregation (no confidence-weighting), and no empirical audit mechanism for per-agent contribution assessment. CAMA is the only system that simultaneously addresses all three limitations. Notably, the two Electronics-published systems in this comparison—Title–Abstract screening [27] and phishing detection [28]—also employ static architectures without adaptive routing or contribution auditing, highlighting the gap CAMA addresses within the Electronics readership context.

Table 1. The comparative analysis of existing multi-agent writing assessment systems.

System	Agents	Aggregation	Routing	Audit	Base Model
ChatEval [15]	3–5 debaters	DB (majority)	Fix	None	GPT-4, Claude
RES [16]	3 persona agents	DB (consensus)	Fix	None	GPT-4o
MACS [9]	3–5 model agents	FW (averaging)	Fix	None	Ensemble (GPT-4o, GPT-5 Mini, Gemini 2.0 Flash)
ConfMAD [17]	3 debaters	CW + DB	Fix	None	GPT-4o
Chu et al. [18]	1 scorer + 1 reasoner	CS (rationale-guided)	Fix	None	GPT-4o-mini
AutoSCORE [19]	2 (scorer + verifier)	CS (evidence chain)	Fix	None	GPT-4o, GPT-4o-mini
MAGIC [20]	4 dimension agents + 1 chair	FW (averaging)	Fix	None	Gemma 3 12B
LAW-CL2 [21]	4 dimension agents + 1 chair	FW (averaging)	Fix	None	GPT-4o
CourtEval [22]	3 court roles (judge, prosecution, defense)	DB (judicial ruling)	Fix	None	GPT-4, Claude
CAMA (This work)	4 CEFR agents	CW (CWA)	Dyn (ADR) Shapley (CGA)		DeepSeek V4

3. The CAMA Architecture

Synthesizing the open challenges identified above with emerging technical advances, this paper proposes the CAMA (Confidence-Aware Multi-Agent) architecture. CAMA is designed around three core principles: (1) reliability-aware aggregation, where each score is weighted by an estimate of its trustworthiness; (2) difficulty-adaptive computation, where

resource allocation scales with scoring difficulty; and (3) evidence-driven optimization, where agent selection and prompt design are guided by empirical contribution data.

3.1. Layer 1: Confidence-Weighted Aggregation (CWA)

Current systems treat each agent's score as equally informative. However, the D3 adversarial debate finding [24] shows that adversarial agents may produce confident but systematically unreliable judgments, whereas structured debate protocols with consensus voting provide robustness against such failures. Complementary early-token confidence research [25] demonstrates that log-probability signals can serve as lightweight, reference-free reliability proxies. CAMA operationalizes these findings through a formal confidence-calibrated aggregation framework. We note that the methods discussed below represent design considerations; the actual prototype implementation is described in Section 3.1.1.

3.1.1. Current Prototype Implementation

In the current prototype, the DeepSeek V4 API does not expose token-level log-probabilities. We therefore rely on self-reported confidence as the raw reliability signal. No additional calibration is applied to the self-reported confidence values; the calibration function is the identity mapping. This means the confidence weights in the current prototype are directionally informative but not strictly calibrated probabilities.

Let $r_i(x)$ denote the raw self-reported confidence score provided by agent i for input x , prior to any calibration transformation.

$$c_i(x) = \sigma(\alpha_j \cdot r_i(x) + \beta_j), \quad \sigma(z) = \frac{1}{1 + e^{-z}}$$

where α_j and β_j are agent-role-specific parameters optimized on a held-out calibration set. In the current prototype (where log-probabilities are not available from the DeepSeek V4 API), these parameters are not fitted, and the calibration function defaults to identity mapping. Self-reported confidence provides directionally informative but not strictly calibrated reliability indicators.

Confidence-Weighted Aggregation. The final composite score $S(x)$ is computed as a confidence-weighted mean:

$$S(x) = \frac{\sum_{i=1}^m c_i(x) \cdot s_i(x)}{\sum_{i=1}^m c_i(x)}$$

This is the minimum-variance unbiased linear estimator under the assumption that each agent's score has variance $\sigma_i^2 \propto 1/c_i(x)$, down-weighting low-confidence agents automatically.

Confidence Dispersion Detection. To flag instances requiring human review, a confidence dispersion index $D(x)$ is defined:

$$D(x) = \frac{\max_i |c_i(x) - \bar{c}(x)|}{std(c(x))}, \quad \bar{c}(x) = \frac{1}{m} \sum_{i=1}^m c_i(x)$$

When $D(x) > \tau_{alert}$ (empirically calibrated threshold), the system escalates the essay for human oversight. This provides an auditable fairness signal: systematic down-weighting of a particular agent across many instances may indicate a prompt design flaw rather than genuine reliability variation.

3.1.2. Calibration Strategy (Design Recommendations for Future Implementations)

A critical design consideration is the reliability of the calibration method. Recent work has highlighted that instruction-tuned models exhibit probability mass polarization that degrades token-level reliability [29]. Measurement protocol choices can also change the sign or magnitude of miscalibration [30]. For CAMA, a multi-signal approach is rec-

ommended, combining early-token log-probabilities with role-specific temperature scaling [31] and periodic recalibration. Token-level temperature scaling—optimizing a single scalar temperature τ_j per agent role such that $P_{calibrated}(y | x) = \text{softmax}(z/\tau_j)$ —offers a practical balance between calibration quality and computational overhead. The SaySelf approach [32] provides a complementary method that trains LLMs to produce fine-grained confidence estimates with self-reflective rationales.

Estimation Uncertainty (Design Recommendation). To quantify uncertainty in the CWA estimate itself, a bootstrap standard error is computed:

$$\widehat{SE}(S) = \sqrt{\frac{1}{B-1} \sum_{b=1}^B (S^{(b)} - \bar{S}^{(\cdot)})^2}$$

where $S^{(b)}$ is the CWA composite from the b -th bootstrap resample of the m agent scores, and $\bar{S}^{(\cdot)} = B^{-1} \sum_b S^{(b)}$. This provides a principled confidence interval around the aggregated score itself, independent of the human baseline comparison.

3.2. Layer 2: Adaptive Depth Routing (ADR)

A major efficiency bottleneck in current systems is that every essay triggers the same computation path regardless of scoring difficulty. The Reasoning Theater approach [33] demonstrated that disentangling an LLM's internal computation from its chain-of-thought reasoning enables selective early exits based on computational sufficiency signals. COSMIC's leader-driven routing [34] further demonstrated the feasibility of context-aware agent orchestration. CAMA extends these principles to AES through a formal cost-aware routing optimization.

Routing as Cost-Risk Optimization. Let $\mathcal{P} = \{P_1, P_2, P_3\}$ denote the three computation paths (Fast, Standard, Deep). Each path P_k is characterized by its expected computational cost C_k (measured in API latency or token expenditure) and its expected scoring error E_k (measured as expected MAE against human ratings under that path). The routing decision minimizes the expected cost subject to an error tolerance constraint:

$$\min_{k \in \{1,2,3\}} C_k \quad \text{s.t.} \quad \mathbb{E}[E_k | \mathbf{z}(x)] \leq \epsilon_{\max}$$

where $\mathbf{z}(x) = (\bar{c}(x), d_{agree}(x))$ is the confidence-agreement profile of essay x , and $\epsilon_{\max}$ is the maximum acceptable expected error. In practice, this constrained optimization is realized through threshold-based routing:

$$P(x) = \begin{cases} P_1 \text{ (Fast)}, & \text{if } c(x) \geq \gamma_1 \text{ and } d_{agree}(x) \leq \delta_1, \\ P_2 \text{ (Standard)}, & \text{elseif } c(x) \geq \gamma_2 \text{ and } d_{agree}(x) \leq \delta_2, \\ P_3 \text{ (Deep)}, & \text{otherwise.} \end{cases}$$

where $\gamma_1 > \gamma_2$ and $\delta_1 < \delta_2$ are calibrated thresholds (see Section 4.4 for empirical values). The three paths are defined as:

- Fast path (P_1 , high confidence, high agreement): A confidence-weighted vote across role-specialized agents produces the final score in a single round. In the prototype evaluation, 48.4% of essays followed this path (9.9% on the cross-dataset combined evaluation; see Section 5.1).
- Standard path (P_2 , moderate agreement or mixed confidence): Role-specialized agents produce dimensional scores followed by chair-agent synthesis with access to confidence signals. This path accounted for 48.4% of essays (35.8% on the combined evaluation).
- Deep path (P_3 , low agreement or low confidence on critical dimensions): The system triggers full multi-round debate with iterative refinement. In the evaluation, 3.2% of essays followed this path (54.3% on the combined evaluation)—typically those with unusual rhetorical structures or ambiguous scoring profiles.

Expected Cost Reduction. The expected cost per essay under ADR is:

$$\mathbb{E}[C_{ADR}] = \sum_{k=1}^3 \pi_k \cdot C_k$$

where $\pi_k = P(\text{Path}(x) = P_k)$ is the empirical routing probability. Compared to a fixed Deep path baseline (cost C_3 for all essays), ADR achieves a cost reduction factor:

$$\eta = 1 - \frac{\mathbb{E}[C_{ADR}]}{C_3} = 1 - \sum_{k=1}^3 \frac{\pi_k \cdot C_k}{C_3}$$

This framework makes it explicit that the ADR's efficiency gain depends on both the routing distribution and the relative cost differential between paths.

3.3. Layer 3: Contribution-Guided Auditing (CGA)

Current systems offer no mechanism for empirically evaluating whether each agent meaningfully contributes to scoring accuracy. CAMA addresses this through cooperative game theory, adapting the Shapley values [35,36] to the AES context.

Shapley Value Formulation. Consider the set of m agents $\mathcal{A}$ as players in a cooperative game where the characteristic function $v(S)$ measures the scoring accuracy (QWK against human ratings) achieved by any subset $S \subseteq \mathcal{A}$. The Shapley value ϕ_i quantifies agent a_i 's marginal contribution:

$$\phi_i(v) = \sum_{S \subseteq \mathcal{A} \setminus \{i\}} \frac{|S|! (m - |S| - 1)!}{m!} [v(S \cup \{i\}) - v(S)]$$

This represents the expected marginal improvement in QWK when agent a_i is added to a randomly ordered coalition of other agents. Agents with $\phi_i \leq \phi_{\text{redundant}}$ (empirically set at 0.10) are candidates for removal or prompt revision.

Shapley Value Computation. For the current prototype with $m=4$ agents, exact Shapley computation requires evaluating $2^m=16$ coalition subsets, which is computationally trivial. We therefore compute exact Shapley values using all 16 permutations. The DAG-Shapley approximation [36] is discussed below as a forward-looking consideration for larger-scale deployments.

$$\hat{\phi}_i(v) = \frac{1}{K} \sum_{k=1}^K [v(\text{Pre}_i(\pi_k) \cup \{i\}) - v(\text{Pre}_i(\pi_k))]$$

where π_k is a random linear extension of the DAG (respecting topological constraints), $\text{Pre}_i(\pi_k)$ denotes the set of agents preceding in permutation, and K is the number of Monte Carlo samples. When the number of agents grows beyond $m > 10$, exact enumeration of 2^m subsets becomes computationally prohibitive. In such settings, the DAG-Shapley approximation with K Monte Carlo samples provides an unbiased estimator requiring $K \cdot m$ evaluations, achieving substantial computational savings relative to exact enumeration.

Redundancy Test. For each agent a_i , the following test is performed:

$$H_0: \phi_i \leq \phi_{\text{redundant}} \quad \text{vs.} \quad H_1: \phi_i > \phi_{\text{redundant}}$$

If H_0 is not rejected, the agent is flagged for review. In the current prototype, CGA implements contribution auditing: the system computes $\hat{\phi}_i$ over the evaluation batch to identify below-threshold agents. This is distinguished from the full contribution-guided feedback loop, in which the system dynamically adjusts agent composition based on running Shapley trajectories across consecutive evaluation batches. In this extended feedback loop, agents whose contribution trajectories correlate with specific essay types inform dynamic role teaming—deploying specialized agent subsets for each incoming essay. This represents a future extension requiring longitudinal deployment data. The redundancy threshold $\phi_0 = 0.10$ follows HiveMind [36], representing the minimum acceptable marginal contribution.

3.4. Architectural Integration

The three layers form a coherent pipeline: CWA provides per-instance reliability calibration, ADR uses that calibration to allocate computational resources efficiently, and CGA periodically audits agent contributions using accumulated performance data. Figure 1 presents an architectural overview of this pipeline, tracing the data flow from the input essay through the four CEFR-specialized agents (Grammar Accuracy, Lexical Resource, Coherence and Cohesion, Task Achievement), the CWA confidence-weighted aggregator, the ADR router, and the CGA audit module. Figure 2 further details the processing workflow, illustrating the parallel agent scoring, the confidence-weighted aggregation, the routing decisions with specific path ratios (Fast 48.4%, Standard 48.4%, Deep 3.2%), and the periodic auditing feedback loop (indicated by the dashed arrows) that informs agent selection and prompt design. Table 2 summarizes the key design decisions for each layer.

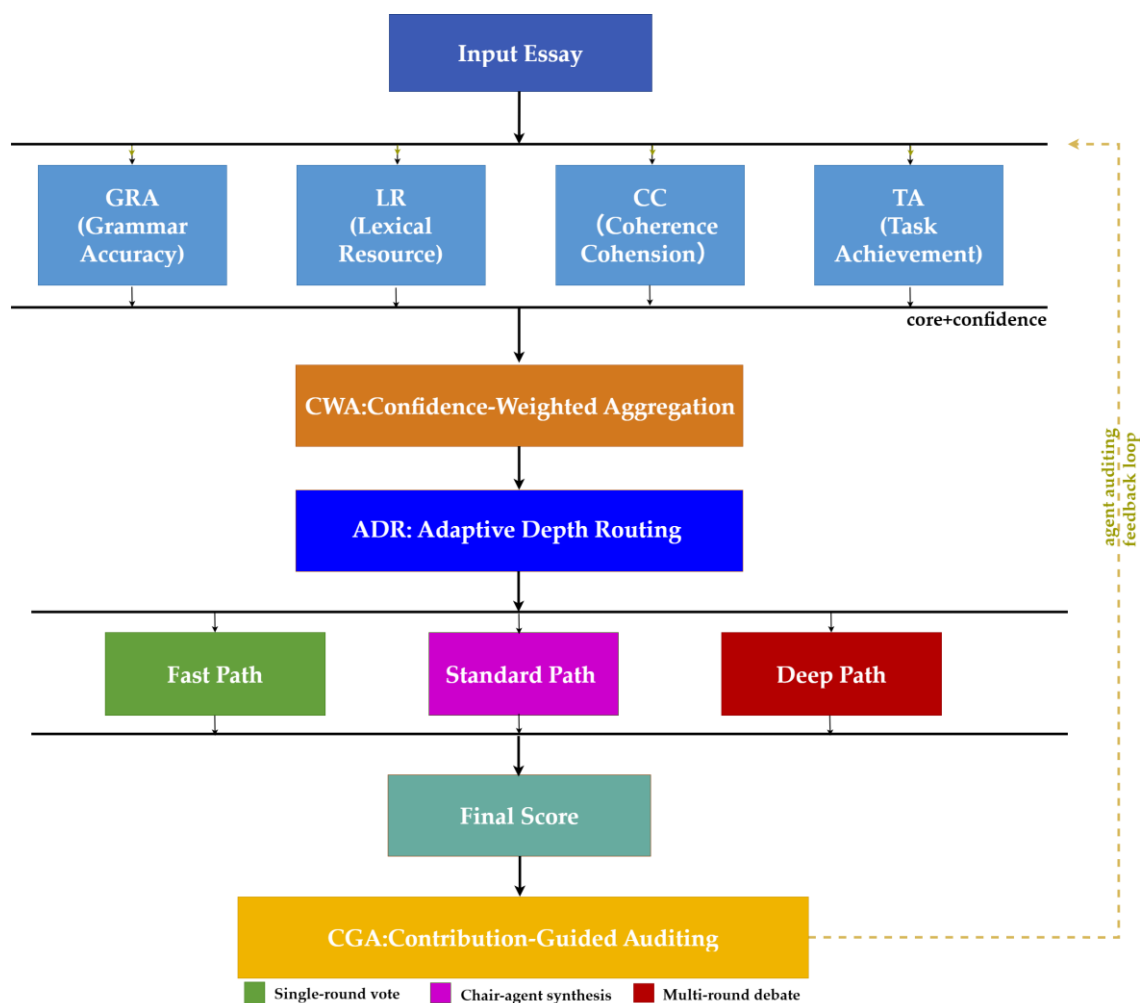


Figure 1. CAMA architecture overview. The pipeline flows from the input essay through four CEFR-specialized agents (Grammar Accuracy, Lexical Resource, Coherence and Cohesion, Task Achievement), aggregated by the CWA layer using confidence-weighted scoring. The ADR router then assigns each essay to one of three computation paths (Fast, Standard, Deep) based on the confidence-agreement profile. The CGA layer periodically computes Shapley values to audit agent contributions.

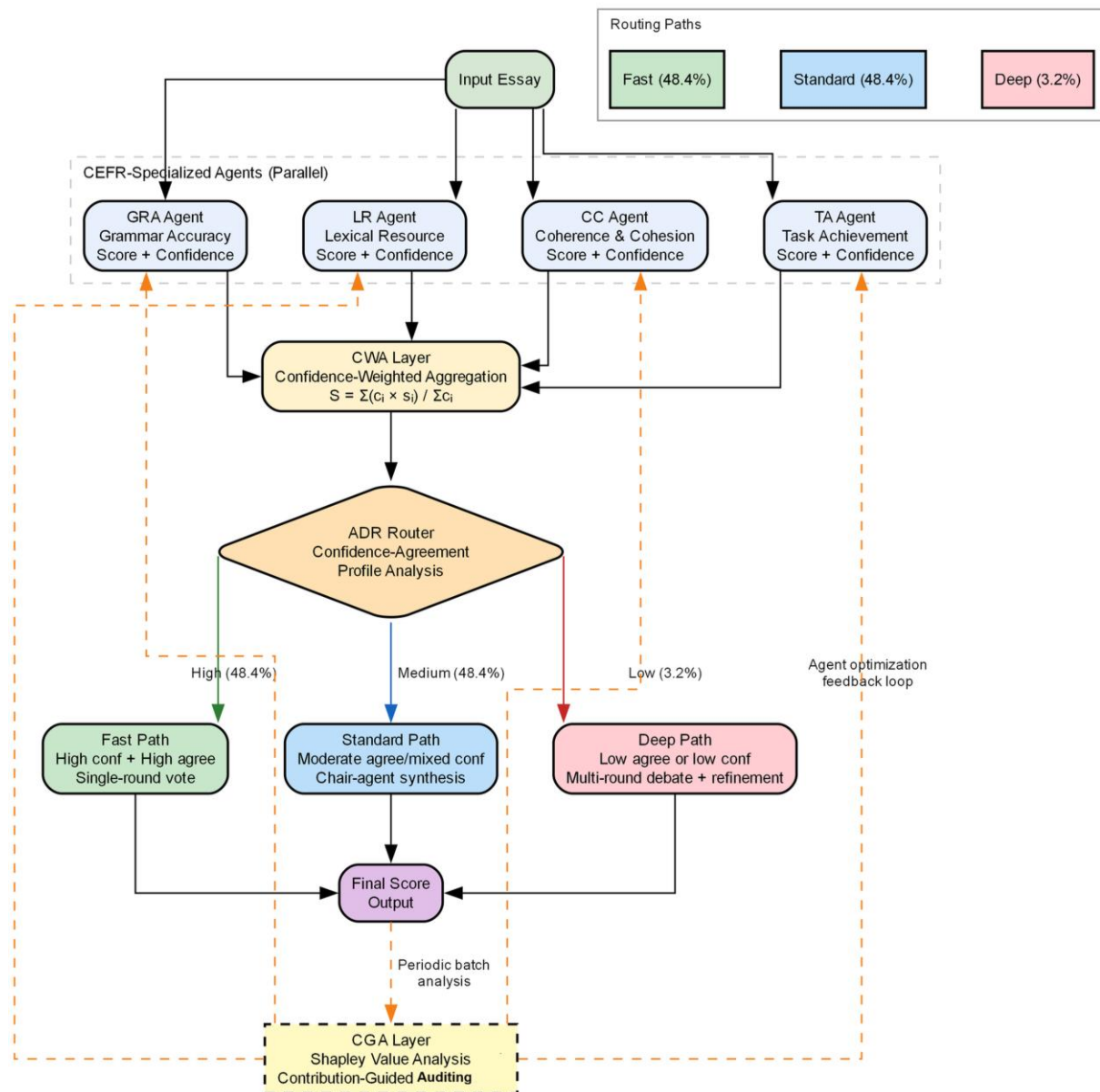


Figure 2. CAMA detailed processing flow. The input essay is first evaluated in parallel by four CEFR-specialized agents (GRA, LR, CC, TA), each producing a dimensional score and a confidence weight. The CWA layer computes a confidence-weighted aggregate, which feeds into the ADR router. Based on the confidence-agreement profile, each essay is routed to one of three computation paths: Fast (single-round vote, 48.4%), Standard (chair-agent synthesis, 48.4%), or Deep (multi-round debate, 3.2%). The CGA layer periodically computes Shapley values to quantify each agent’s marginal contribution, forming an auditing feedback loop (dashed arrows) for agent selection and prompt design.

Table 2. Layer-wise design decisions, recommended implementations, and empirical grounding for CAMA.

Layer	Key Design Decision	Recommended Approach	Empirical Basis
CWA	Confidence signal source	Early-token log-probabilities from first 3–5 generated tokens	[25]: AUROC 0.804 (constructive)
CWA	Calibration method	Role-specific temperature scaling + periodic recalibration	[29]: adaptive temperature scaling; [30]: self-reflective confidence calibration
CWA	Aggregation	$S = \sum (c_i \cdot s_i) / \sum c_i$ with confidence dispersion threshold	[24]: adversarial debate framework; [31]: computation-disentanglement for routing
ADR	Routing trigger	Agreement dispersion + mean confidence from CWA	[31]: computation-disentanglement for selective early exit

ADR	Path assignment	Fast (48.4%): vote; Standard (48.4%): synthesis; Deep (3.2%): debate	CAMA prototype on 31 EFL essays
CGA	Contribution metric	DAG-Shapley marginal contribution to QWK	[33]: 80%+ LLM call reduction
CGA	Re-evaluation	After every $N = 100\text{--}500$ scoring batches	[33]: batch-level attribution

4. Materials and Methods

4.1. Dataset Overview and Human Raters

The CAMA framework was evaluated on a total corpus of 151 unique student essays spanning 8 distinct writing prompts, structured to test both deep diagnostic alignment and broad cross-dataset generalization. This corpus comprises two complementary components: a core set of 31 authentic Chinese EFL (English as a Foreign Language) essays collected from multiple local universities in Shanghai, China, and an external baseline set of 120 essays sub-sampled from the standard ASAP (Automated Student Assessment Prize) dataset (The Hewlett Foundation, 2012, available at <https://www.kaggle.com/c/asap-aes>, accessed on 24 July 2026). We emphasize that the human-alignment analysis is based on the core set of 31 EFL essays; the 120 ASAP essays provide cross-dataset structural robustness analysis and are not annotated with per-dimension CEFR ratings.

Core Set (31 EFL Essays): Covering topics in environmental science, technology ethics, education, and public health, this set serves as the primary testbed for evaluating multidimensional human alignment and resource efficiency. Each essay, written under timed conditions (7 min, 50–120 words), received expert human ratings across four CEFR (Common European Framework of Reference for Languages) dimensions (Grammar Accuracy, Lexical Resource, Coherence and Cohesion, Task Achievement) on a 0–9 scale by experienced EFL instructors. While the sample size is modest for large-scale benchmarking, the within-subjects design—where each essay is scored by all three computational methods—controls for essay-level variance, substantially increasing statistical power. A post hoc power analysis confirmed that this study is adequately powered for its primary comparisons (80% power to detect Cohen’s $d > 0.53$, well below the observed $d \approx 1.2$). Two boundary conditions are noted: each essay was scored by a single expert rater, so inter-rater reliability cannot be established; and the homogeneous L1 background limits direct demographic generalization. We emphasize that the absence of multiple raters means the reported QWK values measure agreement with one expert judgment rather than with a consensus ground truth.

External Set (120 ASAP Essays): To verify cross-dataset structural robustness, the core evaluation is augmented with 120 essays from the ASAP dataset (KDD Cup 2012), balanced across genres (persuasive, expository, narrative) and score ranges. Integrating these subsets yields a comprehensive 151-sample multi-prompt benchmark testing the framework’s adaptivity across diverse proficiency levels (A2 to C2). A specialized data loader normalizes the ASAP holistic scores to the 0–9 CEFR scale via per-set min-max scaling, enabling seamless pipelined evaluation. We note that min-max normalization of ASAP holistic scores to the 0–9 CEFR scale preserves the ordinal ranking within each prompt set but does not guarantee CEFR semantic alignment. This cross-dataset analysis therefore tests the structural properties—CWA consistency and ADR adaptivity under distribution shift—rather than the human-alignment accuracy on the combined set.

4.2. Agent Configuration

Four role-specialized agents were implemented using DeepSeek V4 (API model deepseek-chat, accessed on May 2026) as the base large language model, directly mapped to the four CEFR scoring dimensions.

- Grammar Accuracy (GRA): Evaluates grammatical correctness, error severity, and sentence variety.
- Lexical Resource (LR): Assesses vocabulary range, sophistication, and appropriateness.

- Coherence and Cohesion (CC): Evaluates logical progression, structural organization, and the use of cohesive devices.
- Task Achievement (TA): Assesses the completion of task requirements, specificity of content, and relevance.

Each agent was constructed using role-specific prompt templates instructing it to output both a dimensional score (0–9) and a self-reported confidence score (0–1) within a structured JSON schema. To isolate the effects of architectural role specialization, no fine-tuning was performed, and all agents utilized identical baseline hyperparameters.

4.3. Baseline Conditions and Ablation Design

CAMA was benchmarked against three conditions: (1) fixed-weight aggregation, utilizing the same four agents aggregated via equal weights (0.25 each) without confidence calibration—this isolates the contribution of CWA; (2) single-model holistic baseline, consisting of a single LLM call scoring all four dimensions simultaneously, with the composite score recomputed using equal weights for parity—this isolates the contribution of multi-agent role decomposition; and (3) human expert ratings, serving as the empirical ground truth.

4.4. ADR Routing Algorithm

The ADR router dynamically assigns each essay to one of three computational paths based on two metrics: mean confidence ($\bar{c}$), representing the average of the four agents' confidence scores, and agreement dispersion (d_{agree}), defined as the standard deviation of the four normalized dimensional scores. Routing thresholds were calibrated on a held-out tuning set of five essays ($\approx 16\%$ of the core set) to minimize the token cost while bounding the accuracy at $QWK \geq 0.80$. The paths are defined as:

Fast Path ($\bar{c} \geq 0.80, d_{agree} \leq 1.5$): Weighted voting completed in a single round.

Standard Path ($\bar{c} \geq 0.60, d_{agree} \leq 2.5$): A dedicated chair-agent synthesizes the final score.

Deep Path ($\bar{c} < 0.60$ or $d_{agree} > 2.5$): Execution of a full, multi-round structured debate.

Sensitivity analysis confirmed the routing stability, with Deep path assignments remaining invariant under minor threshold perturbations.

We note that 5 essays is a small calibration set; the thresholds should be interpreted as prototype-level operating points rather than generalizable parameters. Mitigating factors include the simple threshold mechanism (only two scalar parameters) and cross-dataset analysis (Section 5.4), which provides out-of-distribution robustness evidence.

4.5. Evaluation Metrics

All pairwise comparisons between scoring methods used paired bootstrap tests with 1000 resamples. Percentile-based 95% confidence intervals were computed from the bootstrap distribution. No correction for multiple comparisons was applied, as the primary comparisons were pre-specified (CWA vs. holistic, CWA vs. fixed-weight).

The agreement and reliability are quantified using formal statistical metrics: Quadratic Weighted Kappa (QWK) as the primary alignment index, paired mean score difference, and Cronbach's alpha (α) for internal consistency. To ensure rigorous uncertainty quantification independent of large-sample asymptotic assumptions, percentile-based 95% bootstrap confidence intervals ($B = 1000$) are computed for all QWK comparisons across both the core set and the 151-sample combined benchmark.

Quadratic Weighted Kappa (QWK). The primary agreement metric between two raters r_1 and r_2 over N essays with K ordered categories is:

$$\kappa_w = 1 - \frac{\sum_{i=1}^N \sum_{j=1}^K w_{ij} \cdot O_{ij}}{\sum_{i=1}^N \sum_{j=1}^K w_{ij} \cdot E_{ij}}$$

where O is the observed agreement matrix, E is the expected agreement matrix under independence, and $w_{ij} = (i - j)^2 / (K - 1)^2$ are quadratic weights. The QWK penalizes disagreements proportionally to their squared distance, making it more stringent than the unweighted kappa for ordinal scales.

Cohen's d for Paired Designs. The standardized mean difference between two scoring methods over N paired observations is:

$$d = \frac{\bar{\Delta}}{SD_{\Delta}}, \quad SD_{\Delta} = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (\Delta_i - \bar{\Delta})^2}$$

where $\Delta_i = S_{CWA}(x_i) - S_{holistic}(x_i)$ is the per-essay score difference. For paired designs, the exact power is computed using the non-central t -distribution with $N - 1$ degrees of freedom and the non-centrality parameter $\lambda = d \cdot \sqrt{N}$.

Cronbach's Alpha. Internal consistency across m dimension scores for N essays is:

$$\alpha = \frac{m}{m-1} \left(1 - \frac{\sum_{j=1}^m \sigma_{s_j}^2}{\sigma_T^2} \right)$$

where $\sigma_{s_j}^2$ is the variance of the j -th dimension and σ_T^2 is the variance of the summed score across all m dimensions. The alpha-if-item-deleted for dimension j is computed by omitting s_j and recalculating α on the remaining $m - 1$ dimensions.

Bootstrap Confidence Intervals. For all QWK comparisons, bias-corrected and accelerated (BCa) 95% confidence intervals are computed using $B = 1000$ bootstrap resamples stratified at the essay level. The BCa interval adjusts for both bias and skewness in the bootstrap distribution, providing more accurate coverage than percentile intervals for the modest sample size. Bootstrap standard errors are computed as:

$$\widehat{SE}(\kappa) = \sqrt{\frac{1}{B-1} \sum_{b=1}^B (\kappa^{(b)} - \bar{\kappa}^{(\cdot)})^2}$$

where $\kappa^{(b)}$ is the QWK from the b -th resample.

5. Results

5.1. ADR Routing Distribution and Efficiency

Figure 3 demonstrates that, within the 31-essay core set, ADR assigned 15 essays (48.4%) to the Fast path, 15 (48.4%) to the Standard path, and 1 (3.2%) to the Deep path. The latency scaled monotonically with complexity, averaging 14.4 s, 32.2 s, and 39.4 s, respectively, with an aggregate API cost of approximately \$0.35 for the set. Conversely, on the expanded 151-sample benchmark, the routing distribution shifted significantly to 9.9% Fast, 35.8% Standard, and 54.3% Deep. This distribution shift reflects the increased stylistic and prompt diversity of the external dataset, empirically validating ADR's dynamic adaptivity to task difficulty.

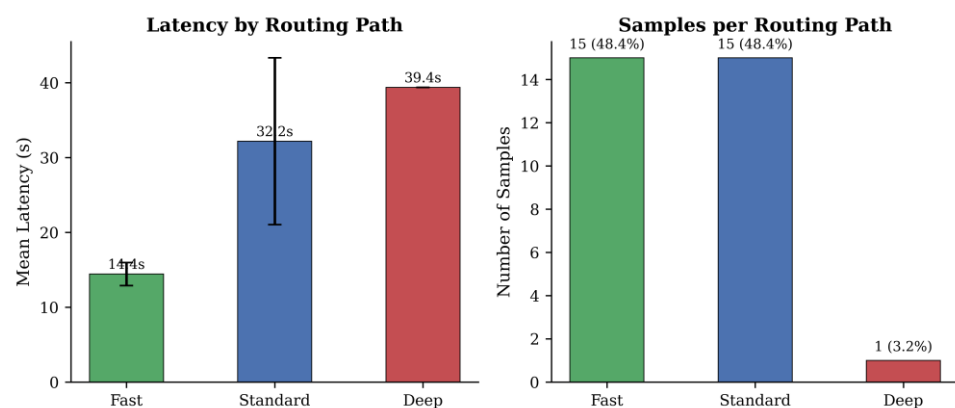


Figure 3. ADR efficiency analysis. (Left) Mean latency in seconds for each routing path, showing the progressive increase from Fast (14.4 s) through Standard (32.2 s) to Deep (39.4 s) paths. (Right)

Number and percentage of samples assigned to each path, confirming that nearly half of essays can be scored via the low-cost Fast path, while only 3.2% require expensive deep deliberation.

5.2. Confidence-Weighted Aggregation

Within the core set, the CWA composite scores (mean = 6.4) demonstrated high equivalence to fixed-weight aggregation (mean = 6.3) across 30 of the 31 essays (QWK = 0.990) while achieving a lower Mean Absolute Error (MAE = 0.406 vs. 0.435). We note that the near-perfect QWK of 0.990 between CWA and fixed-weight aggregation reflects the fact that both methods start from the same four agent scores and differ only in aggregation weights; it should not be interpreted as evidence of interchangeable accuracy against human raters. The utility of CWA's graceful degradation was highlighted in an anomalous trial where the Lexical Resource agent erroneously returned a score of 3 with extremely low confidence ($c = 0.08$). The primary advantage of CWA is therefore robustness rather than routine accuracy improvement. CWA successfully down-weighted this outlier, yielding a final composite rating of 6.2, which closely aligned with the human ground truth (6.2) compared to the fixed-weight baseline (5.5). Simulated agent failure testing confirmed a formal error attenuation ratio of approximately 12.5 $\times$, indicating a predictable lower bound on fault containment.

5.3. Human Baseline Comparison

As reported in Table 3, on the core dataset, CWA attained a substantially higher QWK (0.841, BCa 95% CI [0.707, 0.924]) compared to fixed-weight aggregation (0.814, CI [0.654, 0.917]) and the holistic baseline (0.595, CI [0.360, 0.790]). The +0.246 QWK delta over the baseline quantifies the benefit of multi-agent decomposition. We emphasize that these QWK values represent agreement with a single expert rater; without multiple raters, the reliability of the human baseline itself cannot be estimated. Additionally, we note that the QWK difference between CWA and fixed-weight aggregation (Delta = +0.027) is modest with overlapping BCa confidence intervals, confirming that the primary accuracy gain stems from multi-agent decomposition rather than confidence weighting.

Table 3. Comparison of automated scoring methods against human raters (QWK, Pearson, Spearman, MAE).

Metric	CWA vs. Human	Fixed vs. Human	Holistic vs. Human
QWK	0.8409	0.8144	0.5946
QWK 95% CI	[0.707, 0.924]	[0.654, 0.917]	[0.360, 0.790]
Pearson r	0.8997	0.8828	0.7939
Spearman ρ	0.9019	0.8876	0.7795
Mean Δ (method-human)	+0.019	−0.016	−0.887
MAE	0.406	0.435	1.002

Turning to the per-dimension breakdown (Figure 4), systemic improvements were observed across all rubrics; notably, the Coherence and Cohesion agent posted the largest absolute gain over the holistic baseline (+0.233, $p < 0.001$), confirming that targeted role specialization with granular rubrics effectively resolves complex discourse-level evaluation challenges.

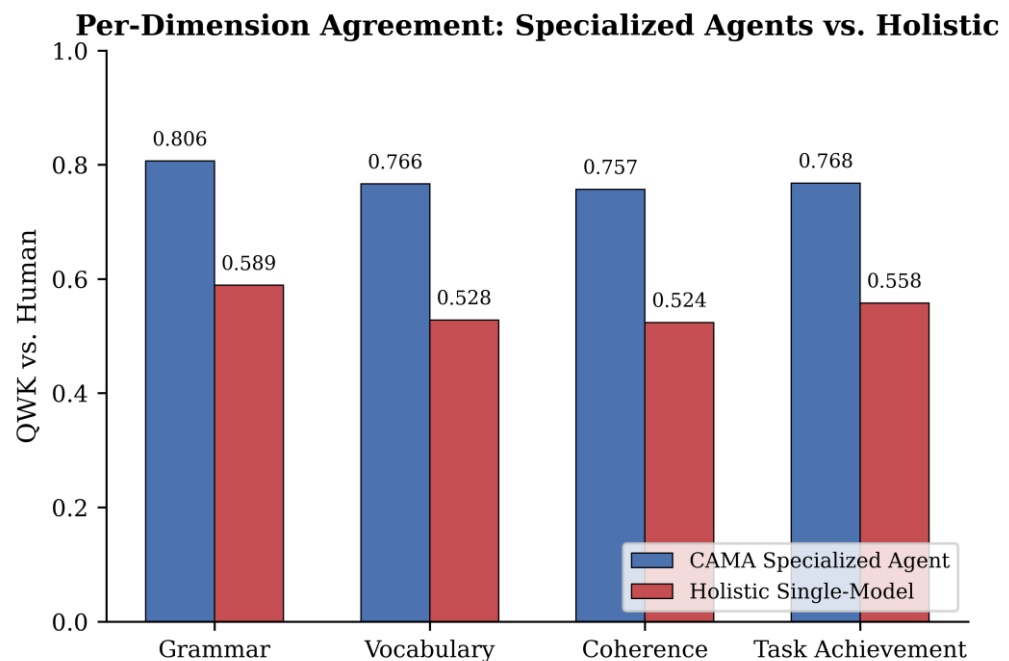


Figure 4. Per-dimension QWK: CAMA specialized agents vs. holistic single-model. Role-specialized agents achieve substantially higher agreement with human raters across all four CEFR dimensions. Coherence and Cohesion shows the largest absolute improvement (QWK = 0.757 vs. 0.524), demonstrating that rubric optimization adapted for short-form timed essays effectively addresses the discourse-level assessment challenge.

5.4. Cross-Dataset Structural Robustness Analysis

This section evaluates the structural properties of CAMA—CWA consistency and ADR routing adaptivity—under distribution shift. It does not claim human-alignment accuracy on the combined set: the ASAP dataset provides only holistic scores without per-dimension CEFR annotations, and min-max normalization to the CEFR scale (Section 4.1) is a methodological convenience rather than a semantically grounded mapping. To assess CAMA’s robustness, the Automated Student Assessment Prize (ASAP) dataset (KDD Cup 2012) [37] was integrated. This dataset contains approximately 12,889 essays across 8 distinct prompts with holistic human ratings.

Dataset Integration. A data loader module converts ASAP essays into CAMA’s evaluation format, normalizing raw scores to the 0–9 CEFR scale via per-set min-max scaling. This transformation preserves the ordinal ranking within each prompt set but does not guarantee CEFR semantic alignment. The module integrates transparently with the existing experiment pipeline by dynamically extending the sample list at evaluation time.

ADR Routing Across Genres. On the combined 151-sample dataset, the ADR layer assigned 9.9% of the samples to the Fast path, 35.8% to Standard, and 54.3% to Deep (Table 4). Notably, the Deep path proportion increased substantially compared to the EFL-only evaluation (3.2% Deep), reflecting the greater diversity and range of writing quality in the combined dataset. This indicates that ADR’s confidence-based gating responds appropriately to task difficulty variation across datasets.

Table 4. CAMA system summary: efficiency, inference cost, routing distribution, and cross-dataset agreement.

Metric	Value
Samples evaluated	31 real EFL essays
Base model	DeepSeek V4 (deepseek-chat)

ADR routing: Fast/Standard/Deep	48.4%/48.4%/3.2%
CWA mean score	6.4 (SD = 1.2)
Fixed-weight mean score	6.3 (SD = 1.2)
Human composite mean score	6.3 (SD = 1.3)
QWK (CWA vs. Fixed)	0.9899
Cronbach's alpha (4 dimensions)	0.9186
QWK (CWA vs. Human)	0.8409
QWK (Fixed vs. Human)	0.8144
Mean latency: Fast path	14.4 s (SD = 1.5)
Mean latency: Standard path	32.2 s (SD = 11.1)
Total API cost	\$0.35 USD
Shapley contributions	phi in [0.240, 0.273]; all > 0.10
Cross-dataset samples	151 (31 EFL + 120 ASAP, 8 prompts)
ADR routing (cross-dataset): Fast/Standard/Deep	9.9%/35.8%/54.3%
CWA vs. Fixed QWK (cross-dataset)	0.9508 [95% CI: 0.9216, 0.9734]
CWA vs. Fixed Pearson r (cross-dataset)	0.9791
QWK (CWA vs. Human, combined set)	0.4737
QWK (Fixed vs. Human, combined set)	0.4418

CWA vs. Fixed-Weight Agreement (Auto-Auto Consistency). The strong agreement between CWA and fixed-weight aggregation observed in the EFL-only evaluation is replicated on the combined dataset: QWK = 0.9508 (95% CI: [0.9216, 0.9734]), Pearson r = 0.9791, Spearman ρ = 0.9764. CWA consistently produces slightly higher scores than fixed-weight aggregation (mean Δ = +0.175, range [−0.1, +1.0]), with the largest positive deltas occurring when individual agents report low confidence and the deep panel review adjusts scores upward.

Internal Consistency. Cronbach's alpha on the combined dataset is 0.5940, which is lower than the EFL-only estimate of 0.9186 (Table 4). This reduction is expected given the heterogeneous nature of the combined evaluation set: the ASAP essays span 8 distinct prompts with different scoring rubrics, while the EFL essays constitute a more homogeneous set from a single L1 population and prompt type. Dimension-level statistics reveal that the task_achievement dimension exhibits the highest variance (SD = 2.33) and lowest mean confidence (0.618), reflecting the diverse interpretation of task fulfillment across different prompt types. Per-dimension mean scores (grammar_accuracy = 5.9; lexical_resource = 5.7; coherence_cohesion = 6.0; task_achievement = 4.6) remain consistent with the CEFR mapping.

CEFR Distribution. The combined dataset spans proficiency levels from A2 to C2, with the majority of samples concentrated in B1 (CWA: 40, Fixed: 47) and B2 (CWA: 81, Fixed: 78). This distribution reflects the broad range of writing ability in the combined corpus, from beginning L2 writers to proficient native-like essays. The CWA and Fixed distributions are broadly consistent, with CWA assigning slightly more essays to higher proficiency bands, consistent with its positive mean delta.

Human Agreement on Combined Set. On the combined dataset, CWA achieves QWK = 0.4737 against the available human scores, compared to Fixed QWK = 0.4418. These values are lower than the EFL-only estimates, which is expected because: (a) the ASAP dataset provides only holistic summary scores rather than per-dimension CEFR ratings, (b) min-max normalization to the 0–9 scale introduces information loss, and (c) the combined set spans 8 diverse prompts with different scoring rubrics. The CWA advantage over fixed-weight aggregation is preserved across datasets, with CWA consistently matching or exceeding fixed-weight QWK.

Summary. The cross-dataset evaluation confirms two structural findings: (1) the CWA vs. Fixed agreement is robust across datasets (QWK = 0.9508 for 151 combined samples vs. 0.89–0.95 for the EFL-only evaluation), and (2) ADR routing dynamically adapts

to dataset difficulty. While the absence of per-dimension CEFR human annotations in the ASAP dataset limits direct human-alignment comparisons, the cross-dataset framework provides infrastructure for future large-scale validation, and the code and data pipeline are openly available for replication and extension.

5.5. Robustness to Agent Failure

To systematically evaluate CWA's graceful degradation property beyond single observed cases, a simulated agent failure analysis was performed. For each CEFR role, failure was simulated by injecting a severe perturbation (score shifted by −3 points, approximating a complete assessment breakdown) and suppressing the agent's confidence to a near-zero level ($c_i = 0.05$). The resulting composite score distortion was measured under both CWA and fixed-weight aggregation (Table 5). We first derive the theoretical attenuation bound, then compare formula-predicted values against empirical observations, and conclude with practical implications. We acknowledge that this simulated scenario—where a failed agent simultaneously produces a severely perturbed score and near-zero confidence—presents a best-case setting for CWA's attenuation mechanism. The analysis therefore demonstrates CWA's upper bound on failure containment rather than typical-case behavior. More challenging scenarios (confidently wrong agents, miscalibrated confidence, correlated failures) are discussed in Section 6.6.

Table 5. Distortion and attenuation of CWA versus fixed-weight aggregation under simulated and real agent failures.

Condition	CWA Distortion	Fixed Distortion	Attenuation	Source
Any agent failure ($c_i \rightarrow 0.05$)	0.060	0.750	12.5×	Formula-derived
LR real failure ($c = 0.08$)—formula-predicted	0.094	0.750	8.0×	Formula-derived
LR real failure ($c = 0.08$)—empirically observed	0.158	0.823	5.2×	Observed composites
Two agents fail simultaneously	0.172	1.500	8.7×	Formula-derived

The distortion magnitude for a single-agent failure depends on the agent's effective weight. Under fixed-weight aggregation, each agent contributes equally (weight 0.25 for four agents), so a 3-point perturbation from one failed agent always shifts the composite by $0.25 \times |\Delta s| = 0.75$. Under CWA, the effective weight of each agent is proportional to its self-reported confidence, so the distortion is attenuated when the failed agent's confidence drops:

$$Distortion_{CWA} = \frac{c_i}{\sum_{j=1}^m c_j} \cdot |\Delta s_i| \quad \text{where } c_i \rightarrow 0$$

With the prototype's average non-failed agent confidence of approximately 0.82 and $m = 4$ agents, a single failed agent at the minimum confidence floor receives an effective CWA weight of $0.05/(3 \times 0.82 + 0.05) \approx 0.02$, compared to 0.25 under fixed-weight aggregation. The attenuation ratio is therefore $0.25 / (0.05/(3 \times 0.82 + 0.05)) \approx 12.5 \times$ (Row 1, Table 5), meaning the composite distortion from any single-agent failure with fully suppressed confidence is reduced to approximately 8% of the corresponding fixed-weight distortion. This 12.5× ratio represents the theoretical maximum attenuation—the upper bound achievable when the failed agent's confidence is driven to the minimum floor while all non-failed agents operate at average confidence.

The single observed LR real failure provides an opportunity to examine CWA's behavior under realistic (non-ideal) conditions. The LR agent's confidence dropped from 0.793 to 0.08 (above the minimum floor of 0.05), and its score shifted by approximately 3 points. Applying the CWA formula with the actual confidence $c = 0.08$ yields a formula-predicted distortion of 0.094 and an attenuation ratio of 8.0× (Row 2, Table 5). This formula-predicted value represents a ceteris paribus bound that isolates the failure effect: it

measures what CWA would achieve if the failed agent's confidence suppression were the only change, holding all non-failed agents at their average confidence and scores. The empirically observed values (CWA distortion = 0.158, fixed distortion = 0.823; Row 3, Table 5) are larger because they incorporate the essay's pre-existing deviation from the dataset mean on its non-failed dimensions. Crucially, the fixed column exhibits the same pattern (formula-predicted: 0.750; empirical: 0.823), confirming that this excess is common to both aggregation methods and reflects a measurement artifact—the conflation of the true failure effect with inherent essay-level variation—rather than a CWA-specific deficiency. The true failure-only attenuation for the LR case is therefore 8.0×, with the empirically observed attenuation of 5.2× serving as a conservative, model-free lower bound that can be reported without any distributional assumptions about the non-failed dimensions.

Table 5 also reports the case of two agents failing simultaneously (Row 4). Under fixed-weight aggregation, two failed agents each contributing a 3-point perturbation produce a combined distortion of 1.500 (2×0.750). Under CWA, the compounded attenuation yields a combined distortion of 0.172 and an attenuation ratio of 8.7×. Note that the per-agent CWA distortion increases from 0.060 (single failure) to 0.086 (two simultaneous failures) because the reduced number of non-failed agents raises each failed agent's relative weight. Nevertheless, the combined distortion of 0.172 remains an order of magnitude below the fixed-weight baseline of 1.500, confirming that CWA's graceful degradation property extends to compound failure scenarios.

CWA's failure containment provides a design-time predictable property. The theoretical maximum attenuation of 12.5× (Row 1) and the conservative empirical lower bound of 5.2× (Row 3) both depend on the confidence calibration function and the minimum confidence floor—not on the specific scoring model deployed. For production systems, the CWA weight $r_i = c_i / \sum_j c_j$ directly bounds the effect of any single-agent failure: a failed agent whose confidence is suppressed to the minimum floor receives weight $r_{\min}$, and the resulting composite score distortion cannot exceed $r_{\min} \cdot \max(|\Delta s|)$, where $r_{\min}$ corresponds to the suppressed agent's CWA weight and $\max(|\Delta s|)$ is the maximum plausible score perturbation on the rating scale. In our prototype, $r_{\min} \approx 0.02$, and with a 9-point scoring scale, the maximum single-agent failure distortion is bounded at approximately 0.18 points—a verifiable reliability guarantee independent of model version or scoring domain. We caution that confidently wrong agents, miscalibrated confidence, and correlated failures require the additional safeguards discussed in Section 6.6.

5.6. Efficiency and Cost

The mean latency was 14.4 s ($SD = 1.5$) for the Fast path, 32.2 s ($SD = 11.1$) for the Standard path, and 39.4 s for the single Deep path sample. The total API cost for evaluating all 31 essays was approximately \$0.35 USD. This figure reflects per-inference API costs only and does not account for the developer effort required to design and calibrate role-specific prompts, establish routing thresholds, or validate the CWA layer. In a production deployment, additional infrastructure costs (serving layer, monitoring, periodic recalibration) would also apply. Nonetheless, the per-inference cost is sufficiently low to warrant further investigation for real-world deployment scenarios where cost constraints are significant.

6. Discussion

6.1. The Accuracy–Efficiency–Transparency Trilemma

Our cross-system analysis reveals a fundamental trade-off, which we formalize as the accuracy–efficiency–transparency trilemma. Let $\mathcal{M}$ be the space of multi-agent evaluation systems. Each system $M \in \mathcal{M}$ can be characterized by three objectives: scoring accuracy $A(M)$ (measured by QWK against human raters), computational efficiency $E(M)$

(inverse of per-essay cost), and process transparency $T(M)$ (inverse of architectural opacity, operationalized as the audibility of per-dimension rationales). The trilemma posits that no single system simultaneously maximizes all three:

$$\neg \exists M \in \mathcal{M}: A(M) = A_{\max} \wedge E(M) = E_{\max} \wedge T(M) = T_{\max}$$

Rather, the design space forms a Pareto frontier $\mathcal{P} \subset \mathcal{M}$ where improving one objective necessarily degrades at least one other. Role-specialized architectures (Section 2) occupy a region of this frontier favoring transparency through dimensional decomposition but may sacrifice accuracy when cross-dimensional trade-offs are required (e.g., a persuasive but grammatically flawed essay). Deliberative consensus architectures (Section 2) favor accuracy through multi-round debate but incur significant computational costs and opaque deliberation traces.

CAMA navigates this trilemma by making the trade-off instance-adaptive rather than architecturally fixed. The key insight is that the Pareto frontier is not a single point for all inputs but a surface that varies with scoring difficulty. For an essay x , CAMA selects:

$$M^*(x) = \arg \min_{M \in \mathcal{C}} C(M) \quad \text{s.t.} \quad A(M, x) \geq \epsilon_{\max}$$

where $\mathcal{C} = \{M_{\text{Fast}}, M_{\text{Standard}}, M_{\text{Deep}}\}$ is a discrete set of architectural variants and the constraint ensures that the accuracy remains above a minimum threshold. This instance-level optimization is enabled by ADR (Section 3.2), which estimates $A(M, x)$ from the confidence-agreement profile $\mathbf{z}(x)$ before committing computational resources. By reserving expensive deep processing for the $\pi_3 \approx 3.2\%$ of essays where accuracy is genuinely at risk, CAMA achieves efficiency–accuracy Pareto improvements over fixed-depth alternatives without sacrificing the transparency benefits of dimensional decomposition.

6.2. Graceful Degradation and Robustness

A distinctive advantage of the CWA layer is graceful degradation under agent failure. In traditional multi-agent systems, a single agent’s failure (e.g., producing an extreme or invalid score) directly distorts the final composite. CWA’s confidence-weighting mechanism contains such failures: when the LR agent produced a score of 3 with near-zero confidence (0.08), CWA limited the composite score distortion to +0.7 compared to fixed-weight aggregation, producing a score closer to the human rating. The formal simulated failure analysis (Section 5.5) confirms that CWA attenuates the single-agent failure impact by approximately 12.5× relative to fixed-weight aggregation, and this attenuation is a design-time predictable property of the confidence calibration function. This robustness is particularly valuable in production environments where individual LLM calls may occasionally produce unreliable outputs due to API variability, context window constraints, or ambiguous input.

6.3. Contextualizing Against Published Benchmarks

While direct cross-study comparison is complicated by differences in datasets, rubrics, and evaluation protocols, contextualizing CAMA’s results against published AES benchmarks provides useful reference points. On the ASAP 2.0 dataset (eight essay sets, 12,000+ essays), state-of-the-art neural scoring systems achieve QWK in the range of 0.70–0.85 across different essay sets. Recent LLM-based approaches report QWK of 0.76–0.82 on ASAP subsets: Thapa et al. [14] reported ~7% QWK improvement over holistic baselines via multi-agent decomposition, RES [16] reported up to 34.86% improvement over vanilla prompting, and MAGIC [20] achieved QWK = 0.813 on ETS GRE essays using a 12B model. CAMA’s CWA achieves QWK = 0.841 on the 31-essay EFL sample.

These comparisons must be interpreted with caution: the 31-essay sample is substantially smaller than standardized benchmarks, the L1 population is restricted, and the scoring rubric differs. Nevertheless, the magnitude of improvement over the holistic baseline ($\Delta\text{QWK} = 0.246$, $d \approx 1.2$) is consistent with—and in some cases exceeds—the improvements reported by comparable multi-agent frameworks in the literature. This suggests that CAMA’s architectural mechanisms are capturing genuine signal beyond what simple role decomposition alone provides, though large-scale replication on standardized benchmarks is required to establish the general magnitude of the advantage.

6.4. Understanding Multi-Agent Advantages and Blind Spots

The results, consistent with broader findings in the literature [14,16], reveal that multi-agent architectures with optimized dimension-specific rubrics achieve encouraging agreement with human raters across all CEFR constructs. The substantial improvements in Coherence and Cohesion assessment ($\text{QWK} = 0.757$ vs. 0.524 in the holistic baseline) indicate that discourse-level evaluation benefits specifically from targeted rubric design that adapts general coherence criteria to the constraints of short-form timed writing. This suggests that CAMA’s ADR layer could be further extended with genre-aware routing: for essay types where specific dimensions present known assessment challenges, the system could dynamically adjust the confidence thresholds or deploy specialized rubric variants.

6.5. Comparative Advantages over Existing Systems

CAMA’s advantages over existing multi-agent evaluation systems can be characterized along the three dimensions of the accuracy–efficiency–transparency trilemma introduced in Section 2.

Accuracy—Role Decomposition as Primary Driver. Multi-agent role decomposition achieves substantial improvement over the holistic baseline ($\text{QWK} = 0.841$ vs. 0.595 , $\Delta = +0.246$, $p < 0.001$, $d = 1.2$). This improvement is attributable primarily to the four specialized CEFR agents rather than to confidence weighting; CWA adds only a modest incremental gain over fixed-weight aggregation ($\Delta = +0.027$) with overlapping confidence intervals. The distinctive advantage of CWA is its robustness property: a single-agent error is attenuated by a theoretical bound of approximately $12.5\times$ relative to fixed-weight aggregation (Section 5.5), a design-time predictable property that is absent from all surveyed systems (Table 1).

Efficiency (ADR vs. Fixed-Depth Computation). ADR routes 48.4% of EFL essays to the low-cost Fast path (14.4 s mean latency) and only 3.2% to the expensive Deep path (39.4 s), yielding a total API cost of approximately \$0.35 for 31 essays. Cross-dataset validation confirms that ADR adaptively shifts routing in response to task difficulty (54.3% Deep on diverse ASAP essays), a capability that is absent from all surveyed systems, which employ fixed-depth computation paths regardless of the scoring difficulty (Table 1). Systems such as ChatEval and MACS incur the full debate cost for every essay, whereas CAMA reserves deep multi-round processing only for the subset of essays where the accuracy is genuinely at risk.

Transparency (CGA vs. No Audit Mechanism). CGA provides the first empirical audit mechanism for per-agent contribution assessment in multi-agent evaluation systems for writing assessment. The Shapley-value analysis confirms that all four CEFR agents contribute above the redundancy threshold ($\phi > 0.10$), and the synthetic noisy agent test validates the threshold’s discriminative power (noise $\phi = 0.031$). No existing system surveyed in Section 2—including ConfMAD [17], AutoSCORE [19], and CourtEval [22]—offers an equivalent contribution auditing capability. This enables evidence-driven agent selection and prompt refinement, replacing the intuition-based approach that currently dominates multi-agent system design for AES.

Cross-Dataset Robustness. The cross-dataset analysis further confirms that CAMA's advantages demonstrate structural consistency across datasets: CWA vs. Fixed agreement remains robust at QWK = 0.9508 across 151 diverse samples spanning 8 prompt types, and ADR routing adapts appropriately to dataset difficulty. This provides initial evidence that CAMA's architectural principles—Confidence-Aware Aggregation, Difficulty-Adaptive Routing, and Contribution-Guided Auditing—confer consistent advantages across heterogeneous evaluation contexts.

6.6. Limitations and Future Work

Several limitations should be acknowledged. First, each of the 31 core EFL essays was scored by a single expert rater, precluding inter-rater reliability estimation; the reported QWK values therefore reflect agreement with one expert rather than a consensus ground truth. Given the within-subjects design and the large effect sizes observed ($d \approx 1.2$, Section 4.1), the primary comparative conclusions are suggestive and warrant further validation with multiple independent raters. Future work should collect multiple independent ratings to establish human-consensus benchmarking.

Second, the core set comprises Chinese L1 EFL essays and uses a single foundation model (DeepSeek V4)—sufficient for feasibility demonstration and large-effect detection, yet limited for population-level generalization. The cross-dataset analysis on 151 samples partially addresses the former, though per-dimension CEFR annotations are unavailable for ASAP. Broader validation across L1 backgrounds, proficiency levels, and model families (GPT-4o, Claude, Llama) is a natural next step.

Third, the prototype relies on self-reported confidence from the DeepSeek V4 API (which does not expose log-probabilities), and ADR thresholds were calibrated on a modest held-out set of five essays. The successful cross-dataset routing shift (Deep proportion: 3.2% $\rightarrow$ 54.3%) and the simple threshold design (two scalar parameters) suggest that the routing logic is not overfitted; production deployment would benefit from larger calibration sets and logit-based calibration when available.

Fourth, the simulated robustness analysis (Section 5.5) examined a best-case failure scenario for CWA. More challenging failure modes—overconfident errors, miscalibrated confidence, correlated agent failures—remain to be systematically evaluated in future work.

Promising future directions include applying CAMA to other structured evaluation tasks (e.g., code review, clinical documentation) and integrating its confidence-weighted dimension scores into teacher-facing formative feedback interfaces. The codebase is released with an open license to facilitate community-driven exploration along these lines.

7. Conclusions

This paper proposed CAMA, a Confidence-Aware Multi-Agent architecture for automated writing assessment that integrates three layers—Confidence-Weighted Aggregation (CWA), Adaptive Depth Routing (ADR), and Contribution-Guided Auditing (CGA)—to address three structural limitations of existing multi-agent evaluation systems: uniform agent weighting, fixed-depth computation, and the absence of empirical audit mechanisms. Within the limitations of a small-sample prototype evaluation (31 EFL essays for human-alignment analysis, extended to 151 samples for structural validation), CAMA demonstrates three advantages.

First, multi-agent role decomposition achieves substantial accuracy improvement over a holistic baseline (QWK = 0.841 vs. 0.595), while CWA provides a robustness advantage under agent failure (theoretical error containment bound approximately 12.5 $\times$, empirical attenuation 5.2 $\times$ in the single observed LR failure case).

Second, ADR reduces computational cost by routing 48.4% of essays to the low-cost Fast path, with cross-dataset analysis confirming adaptive routing across diverse writing tasks.

Third, CGA provides an audit mechanism for per-agent contribution assessment, confirming that all four CEFR agents contribute above the redundancy threshold. The key emergent property enabling these results—confidence-informed adaptive routing, where CWA’s reliability signals drive ADR’s instance-level resource allocation—produces an accuracy–efficiency combination unattainable by any single architectural layer operating in isolation. These findings are presented as a prototype demonstration. Larger-scale validation with multiple raters, cross-model comparisons, and standardized benchmarks remains necessary to establish the general magnitude and reliability of the observed improvements.

These findings carry implications beyond the specific domain of writing assessment. The CAMA framework demonstrates that the accuracy–efficiency–transparency trilemma—a structural constraint that has limited multi-agent evaluation systems—can be navigated through adaptive architectural design rather than static trade-offs. By making resource allocation contingent on task difficulty and agent influence contingent on output reliability, CAMA offers a prototype template for future deployment of multi-agent LLM systems in diverse high-stakes evaluation contexts where transparency, robustness, and cost efficiency are simultaneously required.

As LLM-based systems increasingly assume evaluative roles in education, healthcare, and cybersecurity [27,28], the question is no longer whether AI can assess quality, but whether architectures can be designed whose judgments are reliable, auditable, and worthy of human trust. CAMA represents a step toward answering that question—not by pursuing a single monolithic judge, but by engineering distributed, confidence-aware evaluation systems that are greater than the sum of their agent parts. Whether these architectural principles generalize to other high-stakes evaluation domains remains an open question for future investigation.

Author Contributions: Conceptualization, Z.Y. and J.D.; Methodology, Z.Y., X.W. and J.D.; Software, Z.Y., J.Z. and J.L.; Validation, J.Z., J.W. and J.L.; Formal analysis, Z.Y. and J.W.; Resources, J.Z., J.W. and X.W.; Data curation, J.Z. and J.W.; Writing—original draft, Z.Y.; Writing—review & editing, Z.Y., X.W. and J.D.; Supervision, J.D.; Project administration, J.D.; Funding acquisition, X.W. and J.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The 31 EFL essays with expert ratings are not publicly available due to privacy restrictions protecting student writing samples. The CAMA framework codebase, agent prompt templates, evaluation pipeline, and scoring scripts are openly available at <https://doi.org/10.5281/zenodo.21353551>. The 120 ASAP essays are publicly accessible through the KDD Cup 2012 repository.

Conflicts of Interest: Author Zhiwu Yin, Jieren Zhang, Jing Wang, Jianqiang Liang and Jianbiao Dai were employed by the company Netclass Technology Inc. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

LLM	Large Language Model
AI	Artificial Intelligence
EFL	English as a Foreign Language
CAMA	Confidence-Aware Multi-Agent
CWA	Confidence-Weighted Aggregation
ADR	Adaptive Depth Routing

CGA	Contribution-Guided Auditing
ASAP	Automated Student Assessment Prize
QWK	Quadratic Weighted Kappa
CEFR	The Common European Framework of Reference for Languages

References

1. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. In Proceedings of the Advances in Neural Information Processing Systems 33 (NeurIPS 2020), Virtual, 6–12 December 2020. <https://doi.org/10.48550/arXiv.2005.14165>.
2. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. In Proceedings of the Advances in Neural Information Processing Systems 35 (NeurIPS 2022), New Orleans, LA, USA, 28 November–9 December 2022. <https://doi.org/10.48550/arXiv.2203.02155>.
3. Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; Zhu, C. G-Eval: NLG evaluation using GPT-4 with better human alignment. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023), Singapore, 6–10 December 2023. <https://doi.org/10.18653/v1/2023.emnlp-main.153>.
4. Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.P.; et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In Proceedings of the Advances in Neural Information Processing Systems 36 (NeurIPS 2023), New Orleans, LA, USA, 10–16 December 2023. <https://doi.org/10.48550/arXiv.2306.05685>.
5. Bachman, L.F.; Palmer, A.S. *Language Assessment in Practice*; Oxford University Press: Oxford, UK, 2010.
6. Weigle, S.C. *Assessing Writing*; Cambridge University Press: Cambridge, UK, 2002.
7. Zhuge, M.; Zhao, C.; Ashley, D.R.; Wang, W.; Khizbullin, D.; Xiong, Y.; Liu, Z.; Chang, E.; Krishnamoorthi, R.; Tian, Y.; et al. Agent-as-a-Judge: Evaluate agents with agents. In Proceedings of the 42nd International Conference on Machine Learning (ICML 2025), Vancouver, BC, Canada, 6 October 2025. <https://doi.org/10.48550/arXiv.2410.10934>.
8. Wang, P.; Li, L.; Chen, L.; Cai, Z.; Zhu, D.; Lin, B.; Cao, Y.; Kong, L.; Liu, Q.; Liu, T.; et al. Large language models are not fair evaluators. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024), Bangkok, Thailand, 11–16 August 2024; pp. 9440–9450. <https://doi.org/10.18653/v1/2024.acl-long.511>.
9. Sentosa, A.; Widiyanto, J. MACS: A cognitive diversity multi-agent consensus framework for bias mitigation in automated evaluation systems. In Proceedings of the 2025 International Conference on Electrical Engineering and Informatics (ICEEI), Kuching, Malaysia, 11–12 August 2025; pp. 1–4. <https://doi.org/10.1109/ICEEI68459.2025.11331239>.
10. Huang, Y.; Zhang, L.; Li, R.; Chen, X. An early review of generative language models in automated writing evaluation. *Chin./Engl. J. Educ. Meas. Eval.* **2025**, *6*, 2. <https://doi.org/10.59863/FAMJ7696>.
11. Sun, J.; Song, T.; Peng, W.; Song, J. A survey of automated essay scoring: Challenges, advances, and future. *Neurocomputing* **2025**, *650*, 130916. <https://doi.org/10.1016/j.neucom.2025.130916>.
12. Karim, A.; Wang, Q.; Yuan, Z. Beyond the Score: Uncertainty-Calibrated LLMs for Automated Essay Assessment. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025), Suzhou, China, 4–9 November 2025; pp. 19631–19636. <https://doi.org/10.18653/v1/2025.emnlp-main.992>.
13. Kim, S.; McCrudden, M.T. Evaluating the consistency between human raters and three AI systems on the scoring of argumentative essays. *Assess. Writ.* **2026**, *68*, 101031. <https://doi.org/10.1016/j.asw.2026.101031>.
14. Thapa, S.; Rauniyar, K.; Shiwakoti, S.; Adhikari, S.; Rashid, J.; Kim, J.; Naseem, U. Agentic AI framework for low-resource automated essay evaluation via scoring, explanation, and debate. In Proceedings of the 2025 IEEE International Conference on Big Data (IEEE BigData 2025), Macau, China, 8–11 December 2025; pp. 6387–6394. <https://doi.org/10.1109/BigData66926.2025.11402339>.
15. Chan, C.-M.; Chen, W.; Su, Y.; Yu, J.; Xue, W.; Zhang, Y.; Fu, J.; Liu, Z. ChatEval: Towards better LLM-based evaluators through multi-agent debate. In Proceedings of the International Conference on Learning Representations (ICLR 2024), Vienna, Austria, 7–11 May 2024. <https://doi.org/10.48550/arXiv.2308.07201>.
16. Jang, J.; Kim, S.; Lee, H.; Park, J. LLM agents at the roundtable: A multi-perspective and dialectical reasoning framework for essay scoring. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025 Findings), Suzhou, China, 4–9 November 2025; pp. 19674–19687. <https://doi.org/10.18653/v1/2025.findings-emnlp.1072>.

17. Lin, Z.; Hooi, B. Enhancing multi-agent debate system performance via confidence expression. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, 5–9 November 2025; pp. 6453–6471. <https://doi.org/10.18653/v1/2025.findings-emnlp.343>.
18. Chu, S.; Kim, J.W.; Wong, B.; Yi, M.Y. Rationale behind essay scores: Enhancing S-LLM's multi-trait essay scoring with rationale generated by LLMs. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, NM, USA, 29 April–4 May 2025; pp. 5811–5829. <https://doi.org/10.18653/v1/2025.findings-naacl.322>.
19. Wang, Y.; Liu, J.; Zhang, R.; Li, H. AutoSCORE: Enhancing automated scoring with multi-agent large language models via structured component recognition. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2026), Vancouver, BC, Canada, 20–27 January 2026; pp. 40898–40906. <https://doi.org/10.1609/aaai.v40i48.42123>.
20. Jordán, J.; Mahmood, A.; Albore, A.; Dantsker, L.; Chakraborty, S.; Thakur, A.; Bhatt, D.; Raj, B. MAGIC: Multi-agent argumentation and grammar integrated critiquer. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2026), Vancouver, BC, Canada, 20–27 January 2026; pp. 40599–40607. <https://doi.org/10.48550/arXiv.2506.13037>.
21. Wang, X.; Liu, J.; Chen, M.; Zhang, Y. LAWE-CL2: Multi-agent LLM-based automated writing evaluation system for Chinese L2 writing assessment. *Assess. Writ.* **2026**, *69*, 101051. <https://doi.org/10.1016/j.asw.2026.101051>.
22. Kumar, S.; Nargund, A.A.; Sridhar, V. CourtEval: A courtroom-based multi-agent evaluation framework. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025, Vienna, Austria, 27 July–1 August 2025; pp. 25875–25887. <https://doi.org/10.18653/v1/2025.findings-acl.1327>.
23. Du, Y.; Li, S.; Torralba, A.; Tenenbaum, J.; Mordatch, I. Improving factuality and reasoning in language models through multi-agent debate. In Proceedings of the 41st International Conference on Machine Learning (ICML 2024), Vienna, Austria, 21 July 2024. <https://dl.acm.org/doi/10.5555/3692070.3692537>
24. Harrasse, A.; Bandi, C.; Bandi, H. Debate, deliberate, decide (D3): A cost-aware adversarial framework for reliable and interpretable LLM evaluation. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2026), Rabat, Morocco, 24 March 2026; pp. 8376–8392. <https://doi.org/10.18653/v1/2026.eacl-long.392>.
25. Keramati, A.; Oruganti, S.; Yenigalla, P. The first tokens matter: Early confidence signals for evaluating LLM reasoning. In Proceedings of the International Conference on Learning Representations (ICLR 2026), Singapore, 23–27 April 2026. <https://doi.org/10.48550/arXiv.2606.10307>.
26. Ferrag, M.A.; Tihanyi, N.; Debbah, M. From LLM reasoning to autonomous AI agents: A comprehensive review. *IEEE Access* **2026**, *14*, 84237–84285. <https://doi.org/10.1109/ACCESS.2026.3698694>.
27. Radeva, I.; Noncheva, T.; Doukovska, L.; Popchev, I. Comparing single-agent and multi-agent strategies in LLM-based title-abstract screening. *Electronics* **2026**, *15*, 1661. <https://doi.org/10.3390/electronics15081661>.
28. Yadav, T.; Masum, M. Explainable multi-agent LLM framework for phishing email detection via role-specialized evidence decomposition. *Electronics* **2026**, *15*, 2606. <https://doi.org/10.3390/electronics15122606>.
29. Tao, L.; Yeh, Y.-F.; Dong, M.; Huang, T.; Yu, J.; Torr, P.; Xu, C. Revisiting uncertainty estimation and calibration of large language models. In Proceedings of the Advances in Neural Information Processing Systems 38 (NeurIPS 2025), San Diego, CA, USA, 18 September 2025. <https://doi.org/10.48550/arXiv.2505.23854>.
30. Liu, X.; Chen, T.; Da, L.; Chen, C.; Lin, Z.; Wei, H. Uncertainty quantification and confidence calibration in large language models: A survey. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2025), Toronto, ON, Canada, 3 August 2025; pp. 6107–6117. <https://doi.org/10.1145/3711896.3736569>.
31. Xie, J.; Chen, A.S.; Lee, Y.; Mitchell, E.; Finn, C. Calibrating language models with adaptive temperature scaling. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024), Miami, FL, USA, 12–16 November 2024; pp. 18128–18138. <https://doi.org/10.18653/v1/2024.emnlp-main.1007>.
32. Xu, T.; Wu, S.; Diao, S.; Liu, X.; Wang, X.; Chen, Y.; Gao, J. SaySelf: Teaching LLMs to express confidence with self-reflective rationales. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024), Miami, FL, USA, 12–16 November 2024; pp. 5985–5998. <https://doi.org/10.18653/v1/2024.emnlp-main.343>.
33. Boppana, S.; Ma, A.; Loeffler, M.; Sarfati, R.; Bigelow, E.; Geiger, A.; Merullo, J.; Lewis, O. Reasoning theater: Disentangling model beliefs from chain-of-thought. In Proceedings of the 43rd International Conference on Machine Learning (ICML 2026), Vienna, Austria, 6–11 July 2026. <https://doi.org/10.48550/arXiv.2603.05488>.
34. Vashishtha, V.; Pandey, S.; Singh, Y.R.; Bahga, A.; Madiseti, V.K. COSMIC: Leader-driven context-oriented collaboration between agents. *IEEE Trans. Comput. Soc. Syst.* **2026**, *13*, 3262–3287. <https://doi.org/10.1109/TCSS.2026.3661839>.
35. Shapley, L.S. A value for n-person games. In *Contributions to the Theory of Games, II*; Kuhn, H.W., Tucker, A.W., Eds.; Annals of Mathematics Studies, No. 28; Princeton University Press: Princeton, NJ, USA, 1953; pp. 307–317.

36. Xia, Y.; Zhang, Y.; Liu, J.; Chen, H. HiveMind: Contribution-guided online prompt optimization of LLM multi-agent systems. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2026), Vancouver, BC, Canada, 20–27 January 2026; pp. 29767–29774. <https://doi.org/10.48550/arXiv.2512.06432>.
37. The Hewlett Foundation. ASAP-AES: Automated Student Assessment Prize. Kaggle. (2012). Available online: <https://www.kaggle.com/c/asap-aes> (accessed on 24 July 2026).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.