Diverse Expert Elicitation: Boosting Discriminative Performance in Fine-grained Pairwise Evaluation

Anonymous ACL submission

Abstract

Automatic evaluation is a central task in natural language assessment, particularly for methods leveraging LLMs aligned with human preferences. While existing LLM-based evaluators excel at distinguishing text pairs with quality gaps, these evaluators struggle to distinguish fine-grained text pairs with high similarity. Current multi-agent evaluation studies often rely on fixed, predefined roles and fail to fully activate the specialized knowledge within LLMs, limiting their effectiveness on complex evaluation samples, limiting their effectiveness on complex evaluation samples. To elicit diverse experts for high-quality text evaluation, we propose VMoE, a novel multi-agent framework. It employs an automated expert persona generation mechanism to activate diverse expertise, while using redundancy elimination to manage evaluation costs. Experiments on reconstructed fine-grained benchmarks across four backbone models demonstrate that VMoE achieves strong aggregate performance, improving accuracy and consistency by up to 4.4% and 5.8% respectively over the best single-agent baseline, while reducing token consumption by 39.4% compared to debate-based multi-agent methods.

1 Introduction

Evaluating text quality is one of the key challenges in natural language processing (NLP) (Chang et al., 2024), especially as LLMs produce increasingly high-quality outputs demanding more refined evaluation. Pairwise comparison aligns well with human judgment and remains robust to absolute scale biases, yet designing accurate and scalable pairwise evaluation remains an open problem.

The rapid rise of LLMs has reshaped text automatic evaluation (Chang et al., 2024). The LLM-as-a-judge (Chiang and Lee, 2023; Zheng et al., 2023) paradigm has demonstrated near-human discriminative ability in text quality assessment. Researchers have explored multiple strategies to improve LLM-based evaluation. While early work primarily focused on basic prompt engineering, subsequent research has increasingly utilized sophisticated techniques—such as chain-of-thought (CoT) prompting (Wei et al., 2022) and elaborate evaluation instructions (Liu et al., 2023)—to better elicit the reasoning capabilities of models. From these efforts, two mainstream paradigms have emerged: post-training optimization (Kim et al., 2023; Li et al., 2023; Zhu et al., 2023), which aligns model preferences with human standards via reinforcement learning, and the multi-agent framework (Du et al., 2024; Chan et al., 2023; Wu et al., 2024; Guo et al., 2024a), which leverages multi-role collaboration to enhance judgment robustness.

Despite these advances, pairwise evaluation remains challenging. Post-training methods are constrained by data quality and model scale, limiting cross-domain generalization without retraining (Wang et al., 2024a). Multi-agent frameworks face three key issues: (1) static prompts fail to activate diverse expert knowledge within LLMs, causing systematic bias (Park et al., 2024); (2) role design requires manual, task-specific engineering, limiting flexibility across dimensions; and (3) debate-based methods such as MAD (Du et al., 2024) incur prohibitive inference costs due to multi-turn interactions.

To address these challenges, we propose Virtual Mixture of Experts (VMoE), a novel framework centered on Diverse Expert Elicitation (DEE). Figure 1 illustrates a comparison between our method and existing approaches.

VMoE is motivated by a key observation: applying differentiated role guidance to a single LLM during inference can elicit diverse evaluation experts with broad professional knowledge—without any post-training. These experts provide less biased assessments and generalize across tasks through autonomous role generation. To control inference

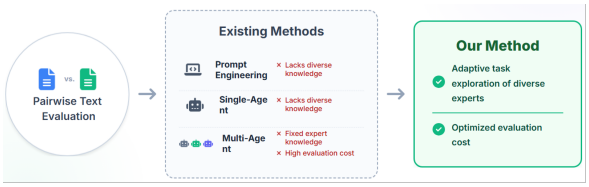


Figure 1: The motivation of our proposed VMoE framework. VMoE overcomes several critical limitations in previous methods, such as the limited evaluative diversity or excessive evaluation costs.

costs, we introduce an expert pool that filters redundant perspectives via Redundancy Elimination, achieving both diversity and efficiency.

Experiments on reconstructed SummEval and CSEE benchmarks show that VMoE achieves favorable results against strong single- and multi-agent baselines, with notable gains in aggregate consistency and inference efficiency.

The main contributions of this paper are as follows:

We propose an automated expert role generation strategy. It addresses the issues of insufficient expert diversity and evaluation bias in existing frameworks by automatically generating a specialized role pool for each evaluation dimension and refining these expert roles through dedicated behavioral grounding processes.

We design an effective inference mechanism based on redundancy elimination. To tackle the high inference overhead of multi-agent frameworks, we propose a semantic embedding-based redundancy elimination strategy that reduces the number of evaluative experts while preserving evaluation quality.

We systematically reconstruct open-source datasets. We release a fine-grained data processing pipeline that reconstructs existing text benchmarks into high-quality pairwise comparison formats while preserving dimension-level scores, thereby providing a reliable evaluation protocol for future research.

2 Related Work

Reinforcement Learning for LLM-as-a-judge Post-training optimization has emerged as a key paradigm for enhancing LLM evaluators. Supervised fine-tuning approaches such as Prometheus (Kim et al., 2024) and JudgeLM (Zhu et al., 2025) train models on human-annotated or synthetic judgment data to align evaluation preferences with human standards. More recently, reinforcement learning methods have been developed, including direct preference optimization (Wang et al., 2025; Rafailov et al., 2023), reward reasoning models (Guo et al., 2025), and unified frameworks that convert both verifiable and non-verifiable judgment tasks into a unified format with verifiable rewards, enabling GRPO-based (Shao et al., 2024) training. However, these methods are constrained by training data quality and struggle to generalize across domains without retraining (Wang et al., 2024a).

LLM-based Evaluation and Multi-Agent Systems Multi-agent collaboration frameworks have been shown to improve evaluation robustness through multi-role interaction. Representative works include Multi-Agent Debate (Du et al., 2024), ChatEval (Chan et al., 2023), and AutoGen (Wu et al., 2024), which leverage diverse agent perspectives to mitigate individual bias. Beyond evaluation-specific frameworks, general-purpose multi-agent systems such as AgentVerse (Chen et al., 2024) and MetaGPT (Hong et al., 2024) have explored structured collaboration mechanisms—AgentVerse facilitates dynamic team formation for emergent behaviors, while MetaGPT incorporates standard operating procedures into multi-agent workflows. Despite their effectiveness, these frameworks face challenges of high inference costs due to iterative multi-turn interactions, limited expert diversity from static role designs, and labor-intensive manual role configuration (Guo et al., 2024b).

Persona-driven Expert Elicitation and MoE Paradigms The Mixture-of-Experts architecture (Shazeer et al., 2017; Fedus et al., 2022) demonstrates that selective knowledge activation via specialized experts can enhance model efficiency. Recent work has further identified an optimal activation rate region where MoE consistently outperforms dense models (Li et al., 2026). Concurrently, research on persona-driven prompting shows that role assignment (Wei et al., 2022; Wang et al., 2024b) can effectively activate specific evaluation capabilities in LLMs. Our VMoE bridges these two lines of work by achieving expert diversity at inference time through virtual persona activation, without any post-training or requiring architectural modifications.

3 Problem Definition

In this study, we focus on the problem of multi-agent collaborative evaluation of fine-grained pairwise text pairs, where text pairs exhibit subtle quality differences that challenge standard evaluators.

3.1 Collection Definitions

Under the setting where multiple candidate responses are associated with a single question, we formulate the evaluation task as follows. Let the question set be denoted as

$$Q = \{q_1, q_2, \ldots, q_n\}$$

where each question q_i consists of textual information including the task description T, scoring rules R, and evaluation dimensions d. For each question q_i, there exists a corresponding candidate response set

$$A_i = \{a_{i,1}, a_{i,2}, \ldots, a_{i,j}\}$$

We further define the evaluation sample pairs as

$$P = \{(q_i, a_{i,u}, a_{i,v})\}$$

where $a_{i,u}$ and $a_{i,v}$ denote two distinct candidate responses associated with the same question q_i.

3.2 Fine-Grained Evaluation Boundary

In practical scenarios, standard LLM evaluators can reliably distinguish text pairs with substantial quality gaps. This work therefore focuses on the more challenging setting: fine-grained pairwise evaluation, where candidate pairs exhibit subtle quality differences bounded by a conceptual threshold ϵ—in practice, we instantiate ϵ as score differences within $[1, 3]$—making them difficult to separate through conventional single-pass assessment. For such pairs, we deploy the VMoE framework to perform deeper, multi-perspective discrimination.

3.3 Evaluation Objectives and Metrics

Our primary objective is to maximize the alignment between the preference predicted by the VMoE framework and the ground-truth human S^*_{gt}. To assess the discriminative performance, we employ Accuracy(acc) and Cohen's Kappa coefficient (κ) (Cohen, 1960) as the metrics for evaluation E_{eval}:

$$E_{eval} = \text{Metric}(S^*, S^*_{gt}) \quad (1)$$

where S^* denotes the final consensus judgment derived from the VMoE. By focusing DEE on fine-grained pairs where subtle differences pose the greatest challenge, we aim to improve discriminative precision on the hardest cases while maintaining computational efficiency.

4 Methodology

In this section, we present VMoE, a framework for diverse expert elicitation in fine-grained pairwise evaluation. It targets candidate pairs whose human-annotated score differences fall within the range of $[1, 3]$. By relying solely on the LLM's capabilities instead of external tools or code execution, VMoE elicits diverse expert personas via in-context prompting. Prior research has demonstrated that the quality of prompts is directly reflected in the quality of model-generated outputs (Deng et al., 2023; Xiang et al., 2025). These personas reflect the model's ability to address complex evaluation criteria by evaluating the text from diverse professional perspectives. Detailed descriptions of VMoE and the evaluation workflow are presented in Figure 2.

The VMoE framework operates in two stages: expert role generation and aggregated expert pool construction. The first stage automatically generates and refines a diverse set of expert personas tailored to the evaluation task. The second stage filters redundant experts via semantic clustering and performs collaborative evaluation through single-pass scoring and aggregation.

4.1 Building Experts

The goal of building experts is to elicit specialized evaluation knowledge from the LLM, tailored to the evaluation dimensions and scoring rubrics of a given task. Expert personas are automatically generated from the task input, then refined through concise self-introductions that reinforce their professional identities.

We solidify these roles through a pilot test on a randomly sampled text pair x_r; the resulting evaluative responses are recorded as reasoning traces for subsequent clustering.

The creation, refinement, and subsequent evaluation of expert personas (formalized in Appendix 1) rely fundamentally on the LLM's capabilities. The process begins by retrieving an evaluation query q from the problem set $\mathcal{Q}$. Given the evaluation query $q = (T, R, d)$ as defined in Section 3, the framework requires defining two hyperparameters: the initial expert count N and the post-aggregation

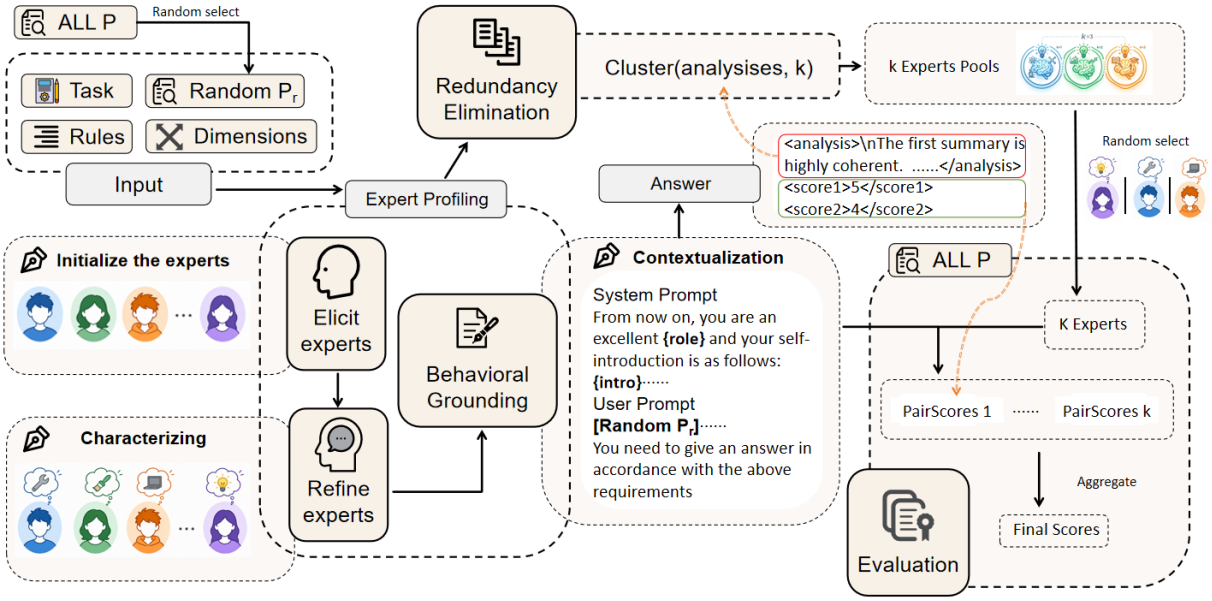


Figure 2: **The pipeline of the VMoE architecture and evaluation workflow.** Diverse experts are first elicited based on the problem definition q (task T, dimension d, and rules R) and a randomly sampled pair $P_r \in P$. Following the redundancy elimination strategy to form k optimized expert pools, the framework is then deployed to evaluate the complete text pair collection P

expert pool size k, with the latter representing the number of clusters utilized during the redundancy elimination phase.

By providing the query q and the initial count N to our structured prompt, we elicit diverse latent experts $e_i \in E$. Notably, these personas are generated in a single inference pass. We enforce output purity via:

```
Just give me roles or professions separated
by ||, without any other words.
```

Empirically, the generated expert personas demonstrate strong contextual relevance with significant individual variance across knowledge bases. In our experiments, MoE-based models tend to generate more diverse expert distributions.

Each initialized persona then generates a self-introduction via targeted prompts. We empirically optimized the introduction length to balance information density against noise, constraining self-introductions to approximately 50 words.

4.2 Redundancy Elimination

Compared to single-pass evaluations based on the CoT (Wei et al., 2022) method, alternative approaches such as single-agent refinement (Madaan et al., 2023), Multi-Agent Debate (Du et al., 2024), and majority voting (Wang et al., 2023) typically yield higher consistency and accuracy. However, these gains come at the cost of a multi-fold increase in computational overhead; iterative multi-turn methods, in particular, necessitate significant sacrifices in inference latency. To address these inefficiencies, we introduce the concept of an expert pool. By partitioning generated expert personas through clustering and accounting for the variance between reference and target samples, we randomly select a single representative from each pool for the evaluation task. This strategy is designed to approximate the results of a full-pool evaluation while reducing computational costs.

We perform Redundancy Elimination by embedding each expert's reasoning trace (the full evaluation analysis from the pilot test) via a frozen semantic encoder (`all-mpnet-base-v2` (Reimers and Gurevych, 2019)) and applying K-means (Arthur et al., 2007) to partition experts into k pools. This assumes that experts within the same cluster exhibit partially overlapping reasoning behaviors, making a random representative sufficient.

4.3 Implementation Steps

Upon constructing heterogeneous expert pools, the collaborative evaluation phase leverages complementary expert perspectives to discern subtle text-pair nuances.

Based on the k expert pools $\{P_1, P_2, \ldots, P_k\}$ generated in the previous stage, we randomly sample one representative expert $\hat{e}_j \in P_j$ from each

Model	Method	Content		Language		Structure		Sum	
		Acc.	Kap.	Acc.	Kap.	Acc.	Kap.	Acc.	Kap.
Qwen3-8B	CoT	0.6436	0.2622	0.5065	0.1382	0.5185	0.2076	0.6706	0.2371
	RPP	0.6486	**0.2820**	0.4964	0.1235	0.5175	0.1909	0.6666	0.2413
	ChatEval	0.6326	0.2674	**0.5615**	**0.1549**	0.5175	0.1826	0.6866	0.2449
	VMoE-CE	0.6346$^{+0.3}$	0.2777$^{+3.9}$	0.5315$^{-5.3}$	0.1242$^{-19.8}$	**0.5255**$^{+1.5}$	**0.2092**$^{+14.6}$	0.6866$^{0.0}$	**0.2521**$^{+2.9}$
	VMoE	**0.6506**	0.2764	0.5525	0.1466	0.5145	0.1839	**0.6866**	0.2435
Qwen-Plus	CoT	0.6966	**0.3403**	0.5295	0.1308	0.5735	0.2536	0.7357	0.2317
	RPP	0.7047	0.3149	0.5435	0.1497	0.5725	**0.2547**	0.7317	0.2103
	ChatEval	**0.7587**	0.2806	**0.7397**	**0.2656**	**0.6736**	0.2277	**0.8118**	0.2918
	VMoE-CE	0.7407$^{-2.4}$	0.3217$^{+14.6}$	0.6746$^{-8.8}$	0.2464$^{-7.2}$	0.5775$^{-14.3}$	0.2502$^{+9.9}$	0.7827$^{-3.6}$	0.2435$^{-16.6}$
	VMoE	0.7067	0.3377	0.5795	0.1787	0.6296	0.2492	0.7527	**0.3011**
GPT-4o-mini	CoT	0.6556	0.2745	0.6156	0.2397	0.5725	0.2746	0.7057	0.2371
	RPP	0.6006	0.2299	0.6116	0.2095	0.5725	0.2811	0.7027	0.2413
	ChatEval	**0.6796**	0.2843	**0.6626**	**0.2867**	**0.6066**	0.3045	0.7447	0.2905
	VMoE-CE	0.6686$^{-1.6}$	0.2809$^{-1.2}$	0.6506$^{-1.8}$	0.2459$^{-14.2}$	0.5855$^{-3.5}$	**0.3266**$^{+7.3}$	**0.7517**$^{+0.9}$	**0.3058**$^{+5.3}$
	VMoE	0.6856	**0.3079**	0.6506	0.1973	0.5995	0.2825	0.7457	0.2997
Gemini-2.5-Pro	CoT	0.6641	**0.3085**	0.5538	0.2024	0.6040	0.3173	0.6992	0.2887
	RPP	0.6896	0.3014	0.5855	0.2082	0.6226	0.3169	0.7507	0.2767
	ChatEval	**0.7127**	0.3028	0.5935	0.2043	**0.6426**	0.3278	0.7687	0.2959
	VMoE-CE	0.6996$^{-1.8}$	0.2883$^{-4.8}$	**0.6006**$^{+1.2}$	**0.2268**$^{+11.0}$	0.6366$^{-0.9}$	**0.3319**$^{+1.3}$	**0.7807**$^{+1.6}$	**0.3084**$^{+4.2}$
	VMoE	0.7037	0.3079	0.5935	0.2043	0.6366	0.3266	0.7767	0.3058

Table 1: Accuracy (Acc.) and Kappa correlation coefficient (Kap.) of different methods on Reconstructed CSEE dataset.

pool to constitute the final evaluation committee $\hat{\mathcal{S}}$. To mitigate potential individual biases in any single expert and to enhance evaluation consistency, the system employs an arithmetic mean to aggregate the scores provided by the k experts. By sampling experts, scoring independently, and aggregating results, this single-pass workflow delivers robust collective decisions without the overhead of multi-turn iterations.

5 Experiments

5.1 Dataset

The evaluation datasets utilized in our experiments are fine-grained, pairwise text assessment sets reconstructed from two publicly available benchmarks: **SummEval** (Fabbri et al., 2021) and **CSEE** (Xiao et al., 2025).

SummEval This dataset comprises 100 news articles sourced from CNN/DailyMail, where each article is paired with 16 candidate summaries generated by diverse systems. Human annotators evaluated these summaries across four distinct dimensions: *Coherence*, *Consistency*, *Fluency*, and *Relevance*, using a Likert scale ranging from 1 to 5.

CSEE The Chinese Student English Essay (CSEE) dataset was collected from English writing examinations involving 13,372 students across 29 high schools in China. These essays were graded by experienced instructors following the official National College Entrance Examination (Gaokao) scoring guidelines. Scoring was conducted across three primary dimensions—*Content*, *Language*, and *Structure*—with the aggregate score serving as the final grade.

Pairwise Reconstruction We constructed an evaluation graph with individual samples as nodes, establishing edges where ground-truth score differences satisfied $1 \leq \Delta \leq 3$. The lower bound excludes tied pairs ($\Delta = 0$), ensuring genuine quality gaps; the upper bound retains only fine-grained pairs below threshold ϵ, maintaining discriminative difficulty. Optimal matching via the Blossom Algorithm (Edmonds, 1965) produced the final pairwise set; for CSEE, we selected the top 10% of student responses per prompt. Detailed pair statistics and a discussion of API-related data attrition are provided in Appendix C.

5.2 Implementation Details

To evaluate the effectiveness of our DEE method across varying scales and architectures, we se-

Model	Method	Coherence		Consistency		Fluency		Relevance		Sum	
		Acc.	Kap.	Acc.	Kap.	Acc.	Kap.	Acc.	Kap.	Acc.	Kap.
Qwen3-8B	CoT	0.6684	0.4391	0.3422	0.1978	0.3548	0.2463	0.5448	0.3079	0.7365	0.5477
	RPP	0.6541	0.3974	**0.3476**	**0.2028**	**0.3709**	0.2662	0.5824	**0.3680**	0.7329	0.5336
	ChatEval	0.6296	0.3589	0.1913	0.1794	0.3271	0.2237	0.5493	0.2198	0.7037	0.4202
	VMoE-CE	0.6790$^{+7.8}$	0.4602$^{+28.2}$	0.2901$^{+51.6}$	0.1964$^{+9.5}$	0.3333$^{+1.9}$	0.2422$^{+8.3}$	0.5061$^{-7.9}$	0.1921$^{-12.6}$	0.7160$^{+1.7}$	0.4341$^{+3.3}$
	VMoE	**0.7043**	**0.4776**	0.2777	0.1961	0.3476	**0.2712**	**0.6021**	0.3443	**0.7688**	**0.5498**
Qwen-Plus	CoT	0.7145	0.4948	**0.3913**	**0.2253**	0.3985	0.2369	0.6319	0.3780	0.7845	0.6100
	RPP	0.6642	0.4198	0.3590	0.1784	**0.4236**	0.2335	0.6319	0.3647	0.7648	0.5443
	ChatEval	0.6451	0.3304	0.2293	0.1935	0.2777	0.1466	0.6433	0.3712	0.7168	0.4351
	VMoE-CE	0.6792$^{+5.3}$	0.4225$^{+27.9}$	0.2741$^{+19.5}$	0.1900$^{-1.8}$	0.3154$^{+13.6}$	0.1934$^{+31.9}$	**0.6541**$^{+1.7}$	**0.4153**$^{+11.9}$	0.7437$^{+3.8}$	0.4944$^{+13.6}$
	VMoE	**0.7293**	**0.5053**	0.3082	0.1641	0.3655	**0.2379**	0.6505	0.3663	**0.8100**	**0.6325**
GPT-4o-mini	CoT	0.6804	0.5592	0.2585	0.1655	0.3267	0.2816	0.6588	**0.4307**	0.7360	0.5101
	RPP	0.6840	0.5411	**0.2926**	0.1731	**0.3411**	0.2776	0.6337	0.3896	0.7235	0.5052
	ChatEval	0.6302	0.5698	0.2226	0.1783	0.3004	**0.2953**	0.6197	0.3779	0.6932	0.4127
	VMoE-CE	0.6541$^{+3.8}$	0.5636$^{-1.1}$	0.2652$^{+19.1}$	**0.1803**$^{+1.1}$	0.2831$^{-5.8}$	0.2893$^{-2.0}$	0.6200$^{0.0}$	0.3744$^{-0.9}$	0.7186$^{+3.7}$	0.4752$^{+15.1}$
	VMoE	**0.7096**	**0.5993**	0.2311	0.1749	0.3387	0.2840	**0.6577**	0.4177	**0.7634**	**0.5396**
Gemini-2.5-Pro	CoT	0.7145	0.5592	**0.4201**	0.1655	**0.6104**	0.2816	0.6355	0.4012	0.7863	0.6205
	RPP	0.7127	0.5411	0.4147	0.1731	0.5655	0.2776	0.6301	0.4002	0.7899	0.6458
	ChatEval	0.7576	0.5698	0.2585	0.1783	0.4093	**0.2953**	0.6517	0.4061	0.7935	0.6028
	VMoE-CE	0.7419$^{-2.1}$	0.5636$^{-1.1}$	0.3458$^{+33.8}$	**0.1803**$^{+1.1}$	0.4498$^{+9.9}$	0.2893$^{-2.0}$	**0.6577**$^{+0.9}$	**0.4293**$^{+5.7}$	0.8207$^{+3.4}$	0.6498$^{+7.8}$
	VMoE	**0.7652**	**0.5993**	0.3297	0.1749	0.4641	0.2840	0.6469	0.4162	**0.8225**	**0.6537**

Table 2: Accuracy (Acc.) and Kappa correlation coefficient (Kap.) of different methods on Reconstructed SummEval dataset.

lect a diverse range of models: Qwen3-8B (Yang et al., 2025), Qwen-Plus (Alibaba Cloud, API-only), Gemini-2.5-Pro (Comanici et al., 2025), and GPT-4o-mini (OpenAI, 2024). To ensure maximum reproducibility, the decoding temperature is strictly set to 0 for all experiments. Following G-Eval (Liu et al., 2023), which samples 20 evaluation instances, we set the initial expert count to $N = 10$ as a fixed hyperparameter without further tuning, balancing diversity of perspectives against computational overhead.

Qwen3-8B is deployed on two NVIDIA A100 (40G) GPUs; closed-source models use official APIs (Gemini-2.5-Pro: $1.25 (input) / $10.00 (output) per 1M tokens). Token consumption details are in Appendix C.

5.3 Prompt

VMoE's efficacy stems from prompt-guided generation with consistent prompt structure across all phases; generalized templates optimized for diverse tasks are provided in Appendix F.

For fairness, we conduct robustness checks by replicating ChatEval (Chan et al., 2023) templates with our expert personas substituted for fixed roles, isolating the effect of expert elicitation from prompt engineering. This control experiment ensures that observed gains stem from expert quality rather than prompt design. Full templates are in Appendix F.

5.4 Baselines

We benchmark against three categories: prompt engineering, single-agent role-playing, and multi-agent consensus.

- **Zero-shot CoT** (Kojima et al., 2022): A standard prompt engineering baseline that excludes specific persona settings. It guides the model through logical reasoning using the "Step-by-step" instruction to establish a performance floor.
- **Role-Play Prompting (RPP)** (Kong et al., 2024): This method represents the state-of-the-art in single-agent role-playing paradigms. Unlike G-Eval, which focuses primarily on task decomposition, RPP emphasizes the depth of expert persona shaping to enhance evaluative precision.
- **ChatEval** (Chan et al., 2023): A classic multi-agent evaluation framework. Its central mechanism involves a "multi-turn debate" among a set of pre-defined agents to reach a collective consensus on the given text pairs.

To ensure a fair comparison, we meticulously optimized the prompts for all baselines for the pairwise assessment task, ensuring that each method achieves its optimal performance.

5.5 Main results

We conducted experiments with two prompt templates: VMoE and VMoE-CE. VMoE-CE uses adapted ChatEval (Chan et al., 2023) templates by replacing static roles with VMoE-generated dynamic personas, validating that gains are robust across prompt architectures. Templates are in Appendix F.

We utilize two metrics to quantify the alignment between model assessments and human ground truth: Acc. and Kap. While Acc. reflects the direct proportion of consistent preferences, Kap. provides a more robust measure of inter-rater reliability by accounting for the probability of random agreement. Together, these metrics ensure a comprehensive evaluation of both precision and stability.

5.5.1 Evaluation Result on CSEE

Table 1 presents the primary experimental results of the four base models on the reconstructed CSEE dataset. The VMoE and VMoE-CE series demonstrate a consistent advantage in overall evaluation consistency. By introducing dynamically generated expert personas into the ChatEval framework, VMoE-CE directly surpasses the original ChatEval across most models and dimensions, with stable performance in the Structure category. Further, the VMoE series maintains a consistent lead over single-agent baselines such as CoT and RPP across all tested models. These results provide evidence that our virtual expert elicitation strategy, which is built on redundancy elimination, can improve the reliability of pairwise assessments with lower overhead costs than fixed-role or multi-turn debate methods.

5.5.2 Evaluation Result on SummEval

Table 2 presents the experimental results on the reconstructed SummEval dataset, showing that the VMoE series achieves notable gains in both accuracy and consistency on summary quality assessment. VMoE holds a consistent and substantial improvement in the Coherence dimension, where it reaches the highest Accuracy and Kappa values across all four base models.

VMoE-CE outperforms ChatEval in Sum Kappa on every model by solely replacing expert input, supporting that adaptive expert pools provide targeted insights as a modular enhancement for multi-agent frameworks.

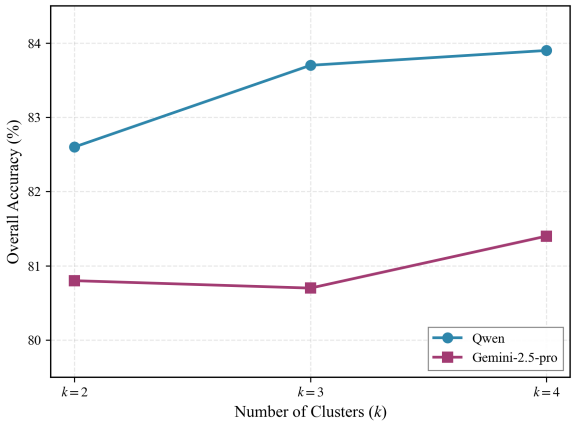


Figure 3: Impact of cluster number k on evaluation performance across different base models.

Although VMoE uses fewer experts than VMoE-CE, it achieves better Sum results and outperforms single-agent baselines, suggesting that redundancy-aware filtering is essential: simply adding experts is insufficient without eliminating repetitive reasoning.

5.6 Analytical and Ablation Study

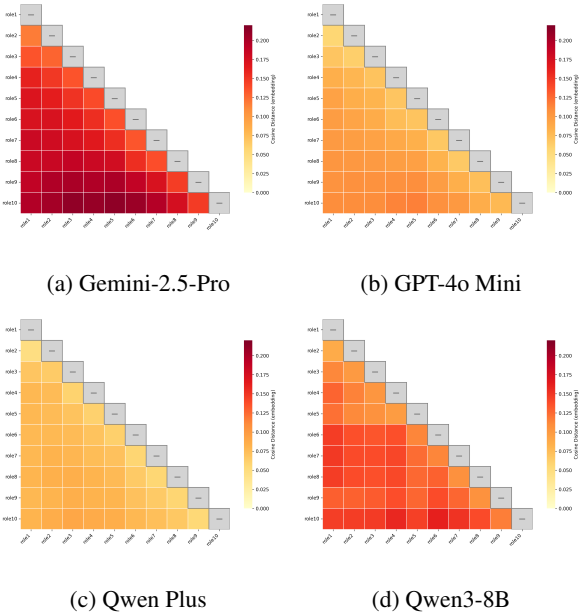
We further conducted analytical and ablation studies on 95 randomly sampled text pairs to manage computational cost.

5.6.1 Effect of Expert Pool Size

To examine how the number of clusters k in the redundancy elimination stage affects performance, we tested three configurations ($k = 2, 3, 4$) on both Qwen and Gemini-2.5-Pro. As shown in Figure 3, the two models respond differently to changes in k. Qwen-Plus benefits from larger k, suggesting that more clusters activate diverse perspectives; Gemini-2.5-Pro shows marginal gains. Given that each additional cluster raises cost by 50%, we adopt $k = 3$ across all experiments to balance performance and overhead.

5.6.2 Effectiveness of Diverse Experts

Following prior work (Lo et al., 2025), we conduct experiments on the generated outputs, as illustrated in Figure 4. The results reveal significant embedding distances between the various automatically generated roles (Role 1 to Role 10). This provides evidence that our DEE mechanism does not merely



(a) Gemini-2.5-Pro (b) GPT-4o Mini

(c) Qwen Plus (d) Qwen3-8B

Figure 4: Cosine similarity heatmaps across four models.

produce redundant or homogeneous descriptions, but instead successfully activates distinct specialized pathways within the model. The heatmap further suggests this differentiation, as both the dense-architecture Qwen series and the MoE-based Gemini-2.5-Pro exhibit sharp color contrasts between experts. These visual patterns reflect a clear divergence in analytical perspectives, regardless of the underlying model architecture.

Similar to the ORPP (Wu et al., 2025), we utilize T-SNE (Van der Maaten and Hinton, 2008) to visualize generation outcomes, which are presented in Appendix E. This illustrates more dispersed expert distributions on the MoE-based Gemini-2.5-Pro., while further demonstrating how the multi-agent framework explores varied specialized trajectories.

5.6.3 Impact of Redundancy Elimination

We conducted a comparative experiment on 95 text pairs to evaluate the impact of our redundancy elimination strategy, using the average score of all 10 experts as the reference baseline. The results, as detailed in Appendix D (Table 4), demonstrate that the cluster-based method achieves the best balance between performance and stability. Specifically, it achieves a combined MSE of 19.136, outperforming both *rand3* and *first3*. This boost in precision is driven by the systematic partitioning of the latent expert space through clustering, which effectively captures complementary perspectives that random sampling tends to overlook. Further, *cluster3* improves evaluation robustness, bringing the STD down to 0.437, compared to 0.502 for *rand3*.

5.6.4 Impact of Self-Introduction Refinement

The ablation study regarding expert persona initialization is detailed in Appendix D, where the empirical results demonstrate that removing this refinement step generally degrades performance, thereby highlighting the importance of this refinement step.

6 Conclusion

We propose VMoE, a multi-agent framework that enhances the accuracy and consistency of pairwise text evaluation by eliciting diverse experts through automated role generation and redundancy elimination. Experiments on reconstructed SummEval and CSEE benchmarks across four backbone models demonstrate consistent improvements over single-agent and comparable performance to multi-agent baselines at lower inference costs than debate-based methods. We release a data processing pipeline for constructing fine-grained pairwise benchmarks to facilitate future research.

Limitations

The randomness introduced by expert pool sampling enhances generalization yet occasionally compromises evaluation stability on boundary cases. VMoE currently operates as a single-pass pipeline that lacks iterative refinement, which could be addressed by multiple rounds of problem analysis or a supervisory judge role for final assessment. Our validation scope remains limited to three multi-agent baselines and four backbone models, and comparative studies with additional frameworks and fine-tuned evaluators are needed to verify VMoE's modularity across diverse architectures. Finally, the framework does not retain historical evaluation insights, which could otherwise reduce input overhead for recurring assessment tasks.

Ethical Considerations

VMoE employs LLMs as evaluators, which may inherit biases from pretraining data and should complement rather than replace human judgment in high-stakes settings. Our experiments use publicly available datasets under their original licenses; no individual-level data is released. While multi-agent evaluation increases inference overhead, our redundancy elimination strategy reduces token usage by

39.4% compared to debate-based methods. We release a data processing pipeline and document all prompts to facilitate reproducibility.

References

David Arthur, Sergei Vassilvitskii, and 1 others. 2007. k-means++: The advantages of careful seeding. In *Soda*, volume 7, pages 1027–1035.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2024. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *The Twelfth International Conference on Learning Representations*.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Yihe Deng, Weitong Zhang, Zixiang Chen, and Quanquan Gu. 2023. Rephrase and respond: Let large language models ask better questions for themselves. *arXiv preprint arXiv:2311.04205*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first international conference on machine learning*.

Jack Edmonds. 1965. Paths, trees, and flowers. *Canadian Journal of Mathematics*, 17:449–467.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.

Jiaxin Guo, Zewen Chi, Li Dong, Qingxiu Dong, Xun Wu, Shaohan Huang, and Furu Wei. 2025. Reward reasoning model. *arXiv preprint arXiv:2505.14674*.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024a. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024b. Large language model based multi-agents: A survey of progress and challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 8048–8057. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. MetaGPT: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and 1 others. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. Better zero-shot reasoning with role-play prompting. In *Proceedings of the 2024*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 4099–4113.

Houyi Li, Ka Man Lo, Shijie Xuyang, Ziqi Wang, Wenzhen Zheng, Haocheng Zhang, Zhao Li, Shuigeng Zhou, Xiangyu Zhang, and Daxin Jiang. 2026. Mixture-of-experts can surpass dense LLMs under strictly equal resource. In *The Fourteenth International Conference on Learning Representations*.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2023. Generative judge for evaluating alignment. *arXiv preprint arXiv:2310.05470*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 2511–2522.

Ka Man Lo, Zeyu Huang, Zihan Qiu, Zili Wang, and Jie Fu. 2025. A closer look into mixture-of-experts in large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4427–4447.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

OpenAI. 2024. Gpt-4o mini: Advancing cost-efficient intelligence. Accessed: 2024-07-18.

Junsoo Park, Seungyeon Jwa, Ren Meiying, Daeyoung Kim, and Sanghyuk Choi. 2024. Offsetbias: Leveraging debiased data for tuning evaluators. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1043–1067.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Noam Shazeer, *Azalia Mirhoseini, *Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

PeiFeng Wang, Austin Xu, Yilun Zhou, Caiming Xiong, and Shafiq Joty. 2025. Direct judgement preference optimization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1979–2009, Suzhou, China. Association for Computational Linguistics.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024a. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2024b. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 257–279, Mexico City, Mexico. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, and 1 others. 2024. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First conference on language modeling*.

Yue Wu, Zhiqing Sun, Huizhuo Yuan, Kaixuan Ji, Yiming Yang, and Quanquan Gu. 2025. Self-play preference optimization for language model alignment. In *The Thirteenth International Conference on Learning Representations*.

Jinyu Xiang, Jiayi Zhang, Zhaoyang Yu, Fengwei Teng, Jinhao Tu, Xinbing Liang, Sirui Hong, Chenglin Wu, and Yuyu Luo. 2025. Self-supervised prompt optimization. *arXiv preprint arXiv:2502.06855*.

Changrong Xiao, Wenxing Ma, Qingping Song, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Qi Fu. 2025. Human-ai collaborative essay scoring: A dual-process framework with llms. In *Proceedings of the 15th international learning analytics and knowledge conference*, pages 293–305.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2023. Judgelm: Fine-tuned large language models are scalable judges. *arXiv preprint arXiv:2310.17631*.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2025. JudgeLM: Fine-tuned large language models are scalable judges. In *The Thirteenth International Conference on Learning Representations*.

A Algorithm: VMoE Framework

Algorithm 1 formalizes the end-to-end VMoE pipeline described in Section 4.3. It comprises three phases: (1) Building Experts (Role Elicitation, Self-Introduction Refinement, and Reference Pair Scoring), (2) Redundancy Elimination via K-means clustering of embeddings derived from expert reasoning traces (Arthur et al., 2007), and (3) Batch Evaluation with arithmetic mean aggregation of expert scores.

Algorithm 1 VMoE Framework: From Expert Elicitation to Batch Evaluation

Input: Sample set $\mathcal{X}$, evaluation query $q \in Q$ (where $q = \{T, R, d\}$), initial count N, and pool size k
Output: Final score set $\mathcal{S}^* = \{s_1^*, s_2^*, \ldots, s_M^*\}$

// Phase 1: Building Experts
$E \leftarrow$ *GenerateExperts*(q, N) ▷ Elicit N diverse latent experts based on q
for each $e_i \in E$ **do**
  $e_i' \leftarrow$ *SelfIntro*(e_i) ▷ Refine experts with specialized self-introductions
end for
$E' \leftarrow \{e_i' \mid e_i \in E\}$
Sample $x_r \sim \mathcal{X}$ ▷ Draw an anchor sample for context-specific elicitation
for each $e_i' \in E'$ **do**
  $e_i'' \leftarrow$ *Test*(e_i', x_r) ▷ Generate reasoning traces
end for
$E'' \leftarrow \{e_i'' \mid e_i' \in E'\}$
// Phase 2: Redundancy Elimination via Clustering
$\{P_1, \ldots, P_k\} \leftarrow$ *Cluster*(E'', k) ▷ Partition into k clusters by reasoning trace embeddings
$\hat{\mathcal{S}} \leftarrow \{$*RandomSelect*$(P_j) \mid j = 1, \ldots, k\}$ ▷ Select representative experts from k pools
// Phase 3: Batch Evaluation across All Samples
$\mathcal{S}^* \leftarrow \emptyset$
for each sample $x_i \in \mathcal{X}$ **do**
  $\{s_{i,1}, s_{i,2}, \ldots, s_{i,k}\} \leftarrow$ *Evaluate*$(x_i, \hat{\mathcal{S}}, q)$ ▷ Individual scoring guided by q
  $\bar{s}_i \leftarrow \frac{1}{k}\sum_{j=1}^{k} s_{i,j}$ ▷ Aggregate scores via arithmetic mean
  $\mathcal{S}^* \leftarrow \mathcal{S}^* \cup \{\bar{s}_i\}$
end for
return $\mathcal{S}^*$

B Dataset Construction Details

The SummEval reconstruction yields 583 fine-grained pairwise comparisons from 100 articles, while the CSEE reconstruction produces 6,620 pairs from 13,240 unique student essays. During evaluation, 5 SummEval prompts (IDs 22, 60, 64, 80, 99) were excluded as their article content triggered API safety filters, resulting in the loss of approximately 28 pairs. The remaining pairs were used for all experiments reported in the main text.

Method	Calls	Input	Output	Total	vs. ChatEval
VMoE (Eval Only)	7.6	7,460	4,886	12,346	43.3%
VMoE (Full)	16.0	9,666	7,632	17,297	60.6%
ChatEval (Baseline)	20.2	25,863	2,660	28,523	100.0%
VMoE-CE	7.6	7,462	1,064	8,526	29.9%

Table 3: Token cost comparison. VMoE (Full) includes the amortized cost of persona pool construction (Elicit experts and Refine experts) at a rate of 10 pairs per prompt. VMoE-CE represents a variant using concise chat-style prompts without detailed analysis.

C Token Consumption

To evaluate the computational efficiency of our proposed framework, we compare the token consumption of VMoE against the competitive multi-agent baseline, ChatEval. Table 3 details the average token usage per text pair.

VMoE substantially reduces computational overhead compared to ChatEval, a competitive multi-turn debate baseline. Even when accounting for the full pipeline—including persona pool construction—our framework (VMoE Full) achieves a 39.4% reduction in total token usage. This gain in efficiency is primarily due to the removal of the repetitive reasoning cycles in debate-based protocols.

The cost advantage is even more pronounced in the evaluation phase alone, where savings reach 56.7%. While automated expert generation involves an initial investment, these costs are effectively amortized as they are spread across multiple evaluation tasks. As task volume increases, the marginal cost per evaluation drops significantly—a trend observed in our experiments on the CSEE dataset. Consequently, VMoE shows high scalability, making it well-suited for large-scale benchmarking.

D Detailed Ablation Study of Expert Refinement

D.1 Redundancy Elimination Ablation

The ablation study regarding expert persona initialization is presented in detail below. As shown in Table 5 and Table 6, removing the expert self-introduction step generally degrades evaluation performance on both CSEE and SummEval datasets across most backbone models, with Gemini-2.5-Pro exhibiting the most substantial decline (e.g., Sum Kappa drops from 0.3058 to 0.1115 on CSEE), thereby highlighting the importance of this refinement step to the VMoE framework.

Strategy	Coherence	Consistency	Fluency	Relevance	Overall
MSE (Mean Squared Error, lower is better)					
cluster3	**1.381**	4.221	**2.611**	1.439	**19.136**
rand3	1.423	4.258	2.737	1.463	19.724
first3	1.457	**4.186**	2.774	**1.387**	19.416
STD (Standard Deviation, lower is more stable)					
cluster3	**0.130**	**0.180**	**0.239**	**0.175**	**0.437**
rand3	0.164	0.211	0.277	0.196	0.502
first3†	-	-	-	-	-

Table 4: Comparison of different expert selection strategies based on 95 text pairs. The reference baseline is the mean score of all 10 experts. †Note that first3 shows zero STD due to its deterministic nature.

E T-SNE

Figure 5 presents a T-SNE visualization of generation outcomes across four backbone models on the SummEval dataset.

F Prompt Templates

This appendix documents the prompt templates and formal algorithms of the VMoE framework. Section F.1 presents the three-stage prompt design (Role Elicitation, Self-Introduction Refinement, and Reference Pair Scoring). Appendix 1 formalizes the expert pool construction and redundancy elimination procedures as pseudo-code. All prompt templates below correspond to the Relevance dimension on the SummEval dataset; templates for other dimensions and the CSEE dataset follow the same structure with dimension-specific task descriptions and scoring rubrics. We replace the **{colored}** slots with real text before querying the LLMs.

F.1 Prompt Templates

F.1.1 Role Elicitation Prompt

This prompt elicits a diverse set of expert personas from the LLM in a single inference pass. The output is constrained to a pipe-separated list.

Backbone	Method	ACC				Kappa			
		Cont.	Lang.	Stru.	Sum	Cont.	Lang.	Stru.	Sum
Qwen	w/o intro	0.7037	0.5735	0.5625	0.7347	0.3231	0.1572	0.2284	0.2841
	ours	**0.7067**$^{+0.4\%}$	**0.5795**$^{+1.0\%}$	**0.5775**$^{+2.7\%}$	**0.7527**$^{+2.5\%}$	**0.3377**$^{+.015}$	**0.1787**$^{+.022}$	**0.2492**$^{+.021}$	**0.2970**$^{+.013}$
Qwen3-8B	**w/o intro**	**0.6606**	**0.5535**	**0.5325**	**0.7017**	0.2602	**0.1530**	**0.2069**	**0.2545**
	ours	0.6506$^{-1.5\%}$	0.5525$^{-0.2\%}$	0.5145$^{-3.4\%}$	0.6866$^{-2.1\%}$	**0.2764**$^{+.016}$	0.1466$^{-.006}$	0.1839$^{-.023}$	0.2435$^{-.011}$
GPT-4o-mini	w/o intro	0.6716	**0.6626**	0.5945	0.7377	0.2885	**0.2553**	0.2918	0.2657
	ours	**0.6856**$^{+2.1\%}$	0.6506$^{-1.8\%}$	**0.5995**$^{+0.8\%}$	**0.7457**$^{+1.1\%}$	**0.3079**$^{+.019}$	0.1973$^{-.058}$	**0.3266**$^{+.035}$	**0.3058**$^{+.040}$
Gemini-2.5-Pro	w/o intro	0.6634	0.5665	0.5980	0.7191	0.2237	**0.2061**	0.2826	0.1115
	ours	**0.7037**$^{+6.1\%}$	**0.5845**$^{+3.2\%}$	**0.6366**$^{+6.5\%}$	**0.7767**$^{+8.0\%}$	**0.3079**$^{+.084}$	0.1973$^{-.009}$	**0.3266**$^{+.044}$	**0.3058**$^{+.194}$

Table 5: Ablation results on CSEE. Δ denotes the performance gain from Expert refinement.

Backbone	Method	ACC					Kappa				
		Coh.	Con.	Flu.	Rel.	Sum	Coh.	Con.	Flu.	Rel.	Sum
Qwen	w/o intro	**0.7365**	0.2921	0.3584	0.6487	0.8028	**0.5217**	**0.1682**	**0.2712**	**0.4111**	0.6039
	ours	0.7293$^{-1.0\%}$	**0.3082**$^{+5.5\%}$	**0.3655**$^{+2.0\%}$	**0.6505**$^{+0.3\%}$	**0.8100**$^{+0.9\%}$	0.5053$^{-.016}$	0.1641$^{-.004}$	0.2379$^{-.033}$	0.3663$^{-.045}$	**0.6325**$^{+.029}$
Qwen3-8B	w/o intro	0.7025	0.2598	**0.3548**	**0.6057**	**0.7759**	0.4517	0.1589	0.2514	0.3388	**0.5751**
	ours	**0.7043**$^{+0.3\%}$	**0.2777**$^{+6.9\%}$	0.3476$^{-2.0\%}$	0.6021$^{-0.6\%}$	0.7688$^{-0.9\%}$	**0.4776**$^{+.026}$	**0.1961**$^{+.037}$	**0.2712**$^{+.020}$	**0.3443**$^{+.006}$	0.5498$^{-.025}$
GPT-4o-mini	w/o intro	0.6989	**0.2401**	0.3189	0.6523	0.7455	0.4528	**0.1699**	0.1880	0.3998	0.5042
	ours	**0.7096**$^{+1.5\%}$	0.2311$^{-3.7\%}$	**0.3387**$^{+6.2\%}$	**0.6577**$^{+0.8\%}$	**0.7634**$^{+2.4\%}$	**0.4803**$^{+.028}$	0.1603$^{-.010}$	**0.2370**$^{+.049}$	**0.4177**$^{+.018}$	**0.5396**$^{+.035}$
Gemini-2.5-Pro	w/o intro	0.7387	**0.3441**	**0.5117**	0.6396	0.8036	0.5943	0.1738	0.2797	**0.4234**	0.6147
	ours	**0.7652**$^{+3.6\%}$	0.3297$^{-4.2\%}$	0.4641$^{-9.3\%}$	**0.6469**$^{+1.1\%}$	**0.8225**$^{+2.4\%}$	**0.5993**$^{+.005}$	**0.1749**$^{+.001}$	**0.2840**$^{+.004}$	0.4162$^{-.007}$	**0.6537**$^{+.039}$

Table 6: Ablation results on SummEval. Δ denotes the performance gain from Expert refinement.

F.1.2 Self-Introduction Refinement Prompt

For each role generated in the Role Elicitation phase, this prompt guides the LLM to produce a concise self-introduction (~50 words, enclosed in `<intro>` tags). The placeholder `{role}` is filled with each generated role name in turn.

F.1.3 Reference Pair Scoring Prompt

Each expert evaluates a randomly sampled reference pair to establish its scoring style. The system prompt (Table 9) injects the expert's role and self-introduction, while the user prompt (Table 10) supplies the article and summaries. The output enforces a structured format with reasoning in `<analysis>` tags and scores in `<score1>` / `<score2>` tags. The resulting reasoning traces are embedded for clustering in the Redundancy Elimination phase.

G Case Study

This appendix presents a representative case study on the SummEval dataset (Gemini-2.5-Pro, Pair 7, Relevance dimension, $N = 10$, $k = 3$). The prompts used are documented in Appendix F; here we focus on the outputs.

G.1 Source Article

Naoki Ogane says that Chelsea have made an offer for Yoshinori Muto. The 22-year-old forward has one goal in 11 games for Japan. Muto admits that it is an 'honour' to receive an offer from the Blues. Chelsea have signed a £200m sponsorship deal with Yokohama Rubber. Muto graduated from university with an economics degree two weeks ago. He would become the first Japanese player to sign for Chelsea. Chelsea have shown interest and have made an offer to recruit forward Yoshinori Muto from Fc Tokyo. Chelsea offered Yoshinori Muto from fc Tokyo a spot on the team with Vitesse Arnhem. The team claims their interest in Muto has nothing to do with the #200 million sponsorship deal with

Prompt: Role Pool Initialization

```
You need to provide 10 roles or professions most likely to accomplish the task based on the
following tasks and requirements.
[### task ###]
In this task you will evaluate the quality of summaries written for a news article.
To correctly solve this task, follow these steps:
1. Carefully read the news article, be aware of the information it contains.
2. Read the proposed summary.
3. Rate summary on a scale from 1 (worst) to 5 (best) by its relevance.
[### requirements ###]
You focus on reviewing the relevance of the summary submitted according to the guidelines. And
you need to refer to the evaluation criteria of the relevance of the story.
Relevance: The rating measures how well the summary captures the key points of the article.
Consider whether all and only the important aspects are contained in the summary.
Just give me roles or professions separated by ||, without any other words.
```

Table 7: Role elicitation prompt template.

Prompt: Role Self-Introduction

```
From now on, you are an excellent {role}. You need to complete the task according to the task and
rules below.
[### task ###]
Here are two summaries for a article, you need to give scores in the sections according rule below.
[### rule ###]
Relevance: The rating measures how well the summary captures the key points of the article.
Consider whether all and only the important aspects are contained in the summary.
You are required to provide a self-introduction for the HR team: a brief introduction that
highlights your capabilities, explains your advantages in completing the task, and conveys your
confidence.
Remember: Your self-introduction should be around 50 words in length.
The self-introduction is enclosed within <intro> </intro>.
You need to give a self-introduction in accordance with the above requirements.
```

Table 8: Self-introduction refinement prompt template. We replace the `{role}` slot with each generated role name.

the Japanese Company Yokohama Rubber. Muto, who just graduated from the University of Tokyo with an economics degree a few weeks ago, may be the first Japanese player for Chelsea if he decides to move to West London. Muto says he is honored to receive the offer, but he hasn't made a decision yet. 22 year old Tokyo forward Yoshinori Muto may soon be playing for Chelsea, who formally made an offer for the star. He has gotten eleven caps for Japan since starting in 2014. If he makes the move, he will be the first Japanese player ever to play for Chelsea. A student who recently graduated from university has been doing well after signing a contract. He has scored several times and many sponsors have enjoyed working with him. Chelsea have made Yoshinori Muto an offer. He would be the first Japanese player to ever represent Chelsea. Muto has won awards in the J-league. The degree earned from Keio University by Yoshinori Muto was in economics. In his debut season, Muto scored 13 times with FC Tokyo. A leading tyre manufacturer is Yokohama Rubber. Yoshino Muto, who earned an economics degree from Keio University, scored thirteen goals in his debut season with FC Tokyo. The offer that Chelsea has made for Muto is not connected in any way to the 200 million dollar sponsorship deal they signed with Yokahama Rubber, a japanese company that manufactures automotive tires, according to FC Tokyo club president Naoki Ogane. Muto graduated with a economics degree. Muto had scored 1 times when he debuted. Yokohama Rubber manufactures tyres. Chelsea offers tokyo player yoshinori muto a deal to join the dutch club. Muto has earned an economics degree from kei university. Muto has not made a decision yet. A leader in world tire manufactering is becoming Chelsea's shirt sponsor.

System Prompt: Reference Example Scoring

```
From now on, you are an excellent {role} and your self-introduction is as follows: {intro}.
You can complete the task assigned to you correctly based on rule as well as your own knowledge.
[### task ###]
Here are two summaries for a article, you need to give scores in the sections according rule below.
[### rule ###]
Relevance: The rating measures how well the summary captures the key points of the article.
Consider whether all and only the important aspects are contained in the summary.
You need to provide the rating within the range of 0-5.
```

Table 9: Reference scoring system prompt. We replace the {role} and {intro} slots with each expert's role name and self-introduction.

User Prompt: Reference Example Scoring

```
[### article ###]
{article}
[### first summary ###]
{summary1}
[### second summary ###]
{summary2}
To correctly complete this task, please follow these steps:
1. Analyze the task requirements and scoring criteria.
2. Assess article and summaries.
3. Analyze above and give reasonable scores.
The analysis process is enclosed within <analysis> </analysis> and score is enclosed within
<score1> </score1> and <score2> </score2> tags, respectively, i.e., <analysis> reasoning process
here </analysis> <score1> score1 here </score1><score2> score2 here </score2>.
Only numbers are allowed inside the <score> tag.
You need to give an answer in accordance with the above requirements.
```

Table 10: Reference scoring user prompt. We replace the {article}, {summary1}, and {summary2} slots with the reference article and summaries.

Muto graduated with an Economics degree from Keio. Muto scored 13 goals in his first season with FC Tokyo. Yokohama Rubber makes car tires.

G.2 Candidate Summaries

Summary A (Text 1):

chelsea have made an offer for fc tokyo 's yoshinori muto . japan international will join chelsea 's vitesse arnhem on loan next season . muto claims chelsea 's interest in muto is not connected to the # 200million deal they signed with yokohama rubber in february .

Summary B (Text 2):

chelsea have made an offer for fc tokyo 's 22-year-old forward yoshinori muto . the japan international will join chelsea 's dutch partner club vitesse arnhem on loan next season if he completes a move to stamford bridge this summer . muto claims chelsea 's interest in muto is not connected to the # 200million sponsorship deal they signed with japanese company yokohama rubber in february .

G.3 Workflow and Final Results

Table 11 summarizes the key outputs at each phase of the VMoE workflow for this case.

The workflow proceeds as follows. First, 10 expert roles are elicited in a single inference pass (Role Elicitation). Each role then generates a brief self-introduction that reinforces its professional identity (Self-Introduction). A randomly sampled reference pair is evaluated by all 10 experts to establish reasoning traces (Reference Scoring). The reasoning traces are embedded via `all-mpnet-base-v2` (Reimers and Gurevych, 2019) and clustered into $k = 3$ groups; one representative is randomly selected from each cluster (Redundancy Elimination). Finally, the three representatives independently score the target pair using the same template as the Reference Scoring phase, and scores are aggregated by arithmetic mean (Collaborative Evaluation). Table 12 reports the final evaluation results across all five dimensions.

Phase	Key Output
Role Elicitation (D.1)	**Role list**: Editor, Journalist, Librarian, Research Analyst, Fact-Checker, Media Analyst, Copy Editor, Intelligence Analyst, Paralegal, Content Strategist
Self-Introduction (D.2)	**Self-introductions**: Each role generates a ∼50-word self-introduction enclosed in `<intro>` tags
Reference Scoring (D.3)	**Reference scoring**: Each expert evaluates a reference pair, producing structured `<analysis>` and `<score>` output; the reasoning traces are embedded for clustering
Redundancy Elimination	**Clustering**: 10 experts partitioned into $k = 3$ pools; representatives selected: Editor, Content Strategist, Fact-Checker
Collaborative Evaluation (D.3)	**Target scoring**: Three representatives independently score the target pair; scores aggregated via arithmetic mean

Table 11: VMoE workflow summary for the case study. Phase references (D.X) point to the corresponding prompt templates in Appendix F.

Dimension	GT (A / B)	VMoE (A / B)	Rank Match
Coherence	3.33 / 4.00	3.00 / 4.00	✓
Consistency	4.67 / 5.00	2.00 / 4.00	✓
Fluency	5.00 / 5.00	3.00 / 3.00	✓
Relevance	3.00 / 4.33	3.00 / 4.00	✓
Total	16.00 / 18.33	11.00 / 15.00	✓

Table 12: Final evaluation results for the case study (SummEval, Gemini-2.5-Pro, Pair 7). All five dimensions correctly identify Summary B as superior to Summary A, consistent with ground-truth human annotations.

Across all five evaluation dimensions, VMoE correctly identifies Summary B as the higher-quality candidate, consistent with the ground-truth human annotations. While the absolute scores differ from human judgments due to known calibration discrepancies in LLM-based evaluation, the rank order is perfectly preserved, demonstrating the discriminative reliability of the multi-expert aggregation framework.

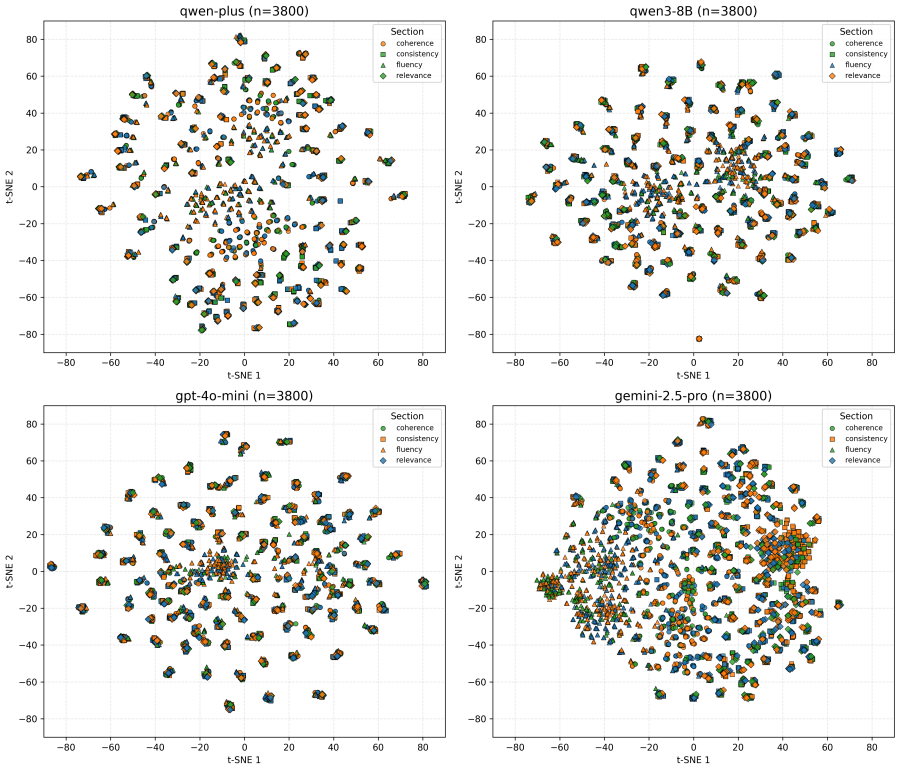


Figure 5: T-SNE of evaluation performance across different base models on the SummEval dataset.