

Contextual Value Alignment via Multilayer Combinatorial Fusion

Yuanhong Wu^{*}, Djallel Bouneffouf[†], and D. Frank Hsu^{*}

^{*} Dept. of Computer and Information Science, Fordham University, New York, NY, USA

[†] T. J. Watson Research Center, IBM Research, Yorktown Heights, NY, USA

Abstract

Aligning large language models (LLMs) with human values remains a major challenge, especially for trustworthy AI. While existing approaches such as RLHF, CAI, and their variants have achieved promising results, they often rely on a single-agent framework and a unified reward system. This limits their ability to capture ethical pluralism, adapt to diverse moral contexts, and reflect the dynamics of multi-agent moral reasoning.

In this work, we propose a framework that utilizes multilayer combinatorial fusion for contextual value alignment (MCF-CVA). At the first layer of the framework, it instantiates multiple moral agents, each fine-tuned to represent a distinctive value. Their outputs are then expanded combinatorially using both score- and rank-combinations as well as average and weighted aggregations. These combined models are then reduced to the same number of initial moral agents. This expansion and reduction (EAR) process continues for multi-layers until a stopping criterion is reached.

The MCF-CVA framework leverages cognitive diversity between agents to mitigate conflicts and redundancies across multiple agents, producing responses that better reflect contextual human values. The framework using the EAR algorithm is performed on the dual architecture of Euclidean score space and Kemeny rank space. Empirical evaluations demonstrated that the proposed framework outperforms single-agent baselines, multi-agent single-layer results, and previous aggregation approaches on standard metrics, showing that the MCF-CVA framework provides a robust and effective mechanism for advancing contextual value alignment in LLMs.

Introduction

The *value alignment problem*—ensuring that autonomous systems act in accordance with human moral principles—has emerged as one of the most pressing challenges in the development of trustworthy AI (Russell, Dewey, and Tegmark 2015; Ji et al. 2025). Traditional approaches have treated alignment primarily as a form of *preference inference*, attempting to derive human values from behavioral data or human feedback. However, as artificial agents become more embedded in social and moral contexts, it is increasingly evident that human morality is *pluralistic, contextual, and dynamic* (Gabriel 2020). A single normative model cannot adequately represent the diversity of moral reasoning found across individuals, cultures, and situations.

A wide range of approaches have been developed to align large language models with human values (Norhashim and Hahn 2024; Singh et al. 2025; Pappas and Vassilakopoulou 2025; Ji et al. 2025; Li et al. 2026). For example, reinforcement learning from human feedback (RLHF) trains a reward model on human preference data and then optimizes the policy against this learned reward (Ouyang et al. 2022). Constitutional AI (CAI) aligns models by enforcing a fixed set of normative principles, using those principles to guide self-critique and AI-generated feedback during training (Bai et al. 2022). However, these methods share a key limitation: they all treat value alignment as a single-agent, single-objective problem. This design cannot leverage the diversity between distinct moral perspectives, nor can it explicitly represent or reason about disagreements among them. Yet human values are pluralistic and context-dependent: different situations prioritize different values, and different individuals or communities weight these values in different ways. A single agent with a fixed objective is therefore ill-suited to capture this structured pluralism or to support context-sensitive value trade-offs.

To address this limitation, Wu, Bouneffouf, and Hsu (2026) proposed a single-layer multi-agent value alignment framework, Value Alignment System using Combinatorial Fusion Analysis (VAS-CFA). In this paper, we extend the VAS-CFA to a multi-layer value alignment framework, Multi-layer Combinatorial Fusion for Contextual Value Alignment (MCF-CVA). After instantiating a set of moderate size with diverse moral agents, each aligned with a specific moral value, we apply CFA to quantify the diversity between these agents and fuse their generated responses. MCF is then used to perform multi-layer combinatorial fusion using the expansion and reduction (EAR) algorithm (Zhong et al. 2019). The MCF-CVA framework yields responses that are more faithfully aligned with human values than other single layer results.

Combinatorial Fusion Analysis (CFA), proposed by Hsu, Chung, and Kristal (2006); Hsu, Kristal, and Schweikert (2024), is a robust model fusion paradigm that combines multiple scoring systems by systematically varying combination size and weighting schemes. A key feature of CFA is its notion of cognitive diversity, a diversity measure that leverages both scores and ranks to quantify the dissimilarity in how different scoring systems represent informa-

tion. Cognitive diversity has been shown to be highly informative for combining diverse models in practice (Hsu et al. 2019). Multilayer Combinatorial Fusion (MCF) extends CFA by stacking multiple layers of CFA combinations to form deeper fusion architectures, thereby more fully exploiting the potential of model fusion across hierarchical layers.

Moral Foundations Theory (MFT) is a widely used tool for modeling moral pluralism and systematic variation in moral judgments across individuals, cultures, and political groups. It was developed by Haidt and Joseph (2004) to explain how multiple “intuitive ethics” give rise to diverse but patterned moral concerns. In its original formulation, MFT proposed five core moral foundations—Care, Fairness, Loyalty, Authority, and Sanctity. This five-factor structure has been widely applied and shows good psychometric support (Graham et al. 2011; Graham, Haidt, and Nosek 2009). Later work proposed Liberty as a candidate sixth foundation. However, the Liberty foundation was introduced later as a candidate sixth foundation rather than part of the original core set, and its status as a distinct foundation remains under active discussion (Graham et al. 2013). Much empirical work—especially factor-analytic and cross-cultural validation studies of the Moral Foundations Questionnaire—continues to adopt and support the original five-foundation structure (De Buck and Pauwels 2023; Doğruyol, Alper, and Yilmaz 2019). In this paper, we therefore adopt the original five foundations as a well-established and widely used moral basis, and we utilize them by training five independent moral agents, each aligned with one foundation (Care, Fairness, Loyalty, Authority, Sanctity), which serve as the base experts that our MCF-CVA framework later combines.

Multilayer Combinatorial Fusion for Contextual Value Alignment (MCF-CVA) introduced in this paper is a framework that utilizes multi-agent moral reasoning through multi-layer iterative fusion and refinement. At the first layer, MCF-CVA instantiates multiple moral agents, each fine-tuned to represent a distinct moral foundation—such as authority, care, fairness, loyalty, and sanctity. Their outputs are combined using CFA with both score-based and rank-based fusion methods within a dual space of *Euclidean (score)* and *Kemeny (rank)* representations. The system then applies an Expansion-and-Reduction (EAR) algorithm to iteratively fuse, evaluate, and prune agents based on their diversity strength, repeating this process across layers until it stables.

Through this multilayer iterative process, MCF-CVA evolves toward moral decisions that are simultaneously contextually coherent and diversely informed. Our experiments demonstrate that MCF-CVA consistently outperforms single-agent and single-layer baselines across standard metrics, offering a principled approach for reconciling heterogeneous moral reasoning within large language models.

In summary, this work makes the following contributions:

1. A novel framework MCF-CVA was proposed to fuse multiple moral agents across multiple iterative reasoning layers by constructing multiple scoring systems corresponding to the multiple moral agents, respectively.

2. Cognitive diversity was used to measure diversity between the scoring systems (moral agents).
3. An integration of combinatorial fusion analysis (CFA) and the expansion and reduction (EAR) algorithm to facilitate adaptive multilayer aggregation in both Euclidean score space and Kemeny rank space.
4. It was demonstrated that the MCF-CVA framework improves moral coherence and is robust and effective over prior contextual alignment methods.

In the following sections, we include: Related Work, Multilayer Combinatorial Fusion, MCF-CVA Workflow, Results, Discussion, and Limitations followed by Conclusion. Six appendices are also included: (A) Rank-score function and cognitive diversity, (B) Kemeny rank space as the MCF-CVA architecture, (C) Figure skating judgement, (D) EAR algorithm and the MCF-CVA framework, (E) Performance summary w.r.t. question q_{2835} , and (F) A detailed example of response decomposition in MCF-CVA workflow.

Related work

The value alignment problem (Russell, Dewey, and Tegmark 2015) aims to design AI systems whose goals and behavior stay aligned with human values. For large language models, one common strategy for value alignment is to fine-tune a single policy with reinforcement learning from human feedback: for example, Wu et al. (2023) propose Fine-Grained RLHF, which learns reward models from detailed human annotations (toxicity, factuality, relevance, etc.) and then optimizes the model against this scalar reward signal to reduce harmful or low-quality generations. Another widely adopted algorithm is direct preference optimization (DPO) (Rafailov et al. 2023), where a single language model is trained to prefer human-preferred answers over rejected ones via a contrastive objective, thereby aligning the model to human preferences without an explicit reward model or RL loop.

Beyond single-policy approaches, some work uses multiple human or model sources in value alignment. Noothigattu et al. (2018) learn a probabilistic model of societal preferences from large-scale Moral Machine data and, at decision time, apply a swap-dominance voting rule to pick the action that best reflects the learned social welfare function, effectively aggregating many human “voters” into one ethical decision. Du et al. (2023) use a multi-agent debate framework where several language model instances independently generate answers and reasoning, then critique and debate each other over multiple rounds before converging on a final response. And Pang et al. (2024) introduce a self-alignment framework where a single LLM performs multi-agent role-playing simulations to model social interactions and consequences, then uses these simulated critiques to revise its own responses.

Dognin et al. (2025) propose contextual moral value alignment, training several agents each associated with different moral values and learning a context-dependent aggregator that selects or combines their responses based on a moral profile for the user. This contextual aggregation idea inspires our work. However, existing multi-source approaches aggregate signals without explicitly measuring or

exploiting diversity among moral agents. They do not quantify how differently agents represent moral information. Our MCF-CVA framework, as an extension to the VAS-CFA framework (Wu, Bouneffouf, and Hsu 2026), builds on this approach by quantifying diversity between moral agents and using multilayer combinatorial fusion to systematically exploit that diversity when fusing their outputs.

Unlike prior ensemble or alignment techniques, our framework explicitly quantifies and leverages diversity strength across moral agents, applying an Expansion–Reduction process that iteratively fuses and prunes models across both score and rank spaces.

Multilayer Combinatorial Fusion

Multilayer combinatorial fusion (MCF) employs the expansion and reduction (EAR) algorithm multiple times until it stables. The expansion step uses the CFA while the reduction step utilizes a sliding rule to select top combined models with large cognitive diversity. This section consists of: (a) Combinatorial fusion analysis, and (b) multilayer combinatorial fusion (MCF) with the EAR algorithm.

Combinatorial Fusion Analysis

Combinatorial Fusion Analysis (CFA) provides methods and workflows for combining multiple scoring systems (MSS) (e.g., multiple classifier systems, multiple expert systems, multiple language models, and multiple agent systems) in computational learning and modeling, informatics, and intelligent ML/AI systems (Hsu, Chung, and Kristal 2006; Hsu, Kristal, and Schweikert 2024). CFA characterizes a scoring system A with a score function s_A , a derived rank function r_A , and a function f_A that relates normalized score values to rank values.

Let A be a scoring system on the dataset $D = \{d_1, \dots, d_n\}$. Let $s_A : D \rightarrow R$ be a score function. Rank function $r_A : D \rightarrow N$ is derived by sorting the score values and assigning an increasing rank value to the data item in D on the decreasing score values. Rank–score function (RSF) $f_A : N \rightarrow R$ is defined as $f_A(i) = s_A(r_A^{-1}(i)) = (s_A \circ r_A^{-1})(i)$. See Figure A1(a) in Appendix A for the relationship between these three functions for a scoring system A (Hsu, Chung, and Kristal 2006). For scoring systems A and B , cognitive diversity (CD) between A and B , $CD(A, B)$, is defined as the difference between f_A and f_B (Hsu et al. 2019): $CD(A, B) = d(f_A, f_B) = (\sum_{i=1}^n (f_A(i) - f_B(i))^2 / n)^{1/2}$. In addition, the diversity strength of the scoring system A , $DS(A)$, is the average of $CD(A, A')$ between A and all other scoring system A' under consideration. Cognitive diversity $CD(A, B)$ between scoring systems A and B measures the diversity between A and B by calculating the difference between rank–score functions f_A and f_B . It is analogous to, but different from, Pearson correlation (p), Spearman rho (ρ) correlation, and Kendall tau (τ) correlation in statistics. Figure A1(b) in Appendix A depicts their relationships and comparisons (Hsu et al. 2019; Hurley et al. 2020). Note that the score function is defined in Euclidean space, whereas the rank function—which handles rankings with or without

ties—operates in the Kemeny rank space. Further details on the Kemeny rank space are provided in Appendix B.

The four combination schemes (average score combination (ASC), weighted score combination by diversity strength (WSCDS), average rank combination (ARC), and the weighted rank combination by diversity strength (WR-CDS) are defined for each $t \in \{2, 3, 4, 5\}$ as below, respectively.

$$s_{ASC}(d_i) = \frac{\sum_{A_j \in \mathcal{A}'} s_{A_j}(d_i)}{t} \quad (1)$$

$$s_{WSCDS}(d_i) = \frac{\sum_{A_j \in \mathcal{A}'} s_{A_j}(d_i) \cdot DS(A_j)}{\sum_{A_j \in \mathcal{A}'} DS(A_j)} \quad (2)$$

$$r_{ARC}(d_i) = \frac{\sum_{A_j \in \mathcal{A}'} r_{A_j}(d_i)}{t} \quad (3)$$

$$r_{WR-CDS}(d_i) = \frac{\sum_{A_j \in \mathcal{A}'} r_{A_j}(d_i) \cdot \frac{1}{DS(A_j)}}{\sum_{A_j \in \mathcal{A}'} \frac{1}{DS(A_j)}} \quad (4)$$

where $\mathcal{A}'$ is any subset of the full scoring system set $\mathcal{A} = \{A_1, \dots, A_5\}$ containing at least 2 elements, i.e., $\mathcal{A}' \subseteq \mathcal{A}$, and $\|\mathcal{A}'\| = t$, with $t \in \{2, 3, 4, 5\}$, $s_{A_j}(d_i)$ and $r_{A_j}(d_i)$ are the score and rank assigned by scoring system A_j to the data item d_i , respectively; $DS(A_j)$ represents the diversity strength of scoring system A_j .

CFA has been applied to a wide variety of domain applications in scientific discovery and decision making, including more recent results in wireless network selection (Kustiawan, Liu, and Hsu 2017), material science (Tang et al. 2021), drug discovery (Jiang et al. 2023), sentiment analysis (Patten et al. 2025), cybersecurity (Alfatemi et al. 2025; Owusu et al. 2025), Bitcoin price prediction (Wu et al. 2025), and text classification (Xu et al. 2025). It was shown that rank combination tends to perform better than score combination w.r.t. larger cognitive diversity (Hsu and Taksa 2005). An example of figure skating judgment is included in Appendix C to illustrate score function, rank function, rank–score function of a scoring system A , as well as score and rank combinations, and cognitive diversity between two scoring systems, A and B .

MCF with the EAR algorithm

Multilayer combinatorial fusion employs the expansion and reduction (EAR) algorithm at each layer until it reaches to the layer where there is almost no diversity between models. The pseudocode for EAR algorithm is in Algorithm D1 (Appendix D). We first construct the set of models to be combined. With five models, the total number of nontrivial model subsets is $2^5 - 5 - 1 = 26$. Since each scoring system consists of both a score function and a rank function, we then extract the corresponding score and rank functions from each system separately. For each input type (scores or ranks), we aggregate them using either average combination or weighted combination by diversity strength, as defined in formulas (1)–(4).

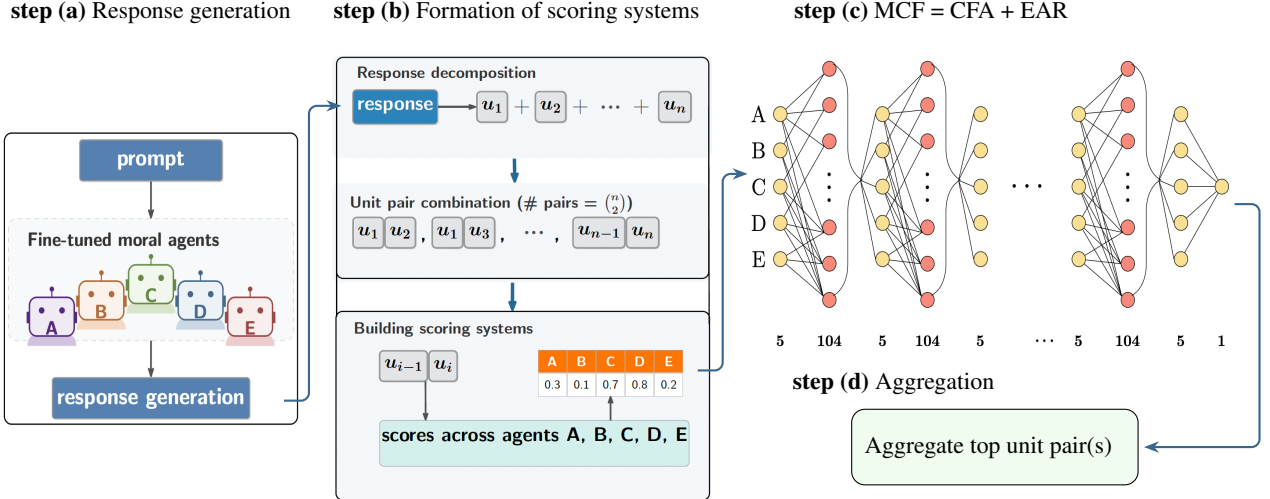


Figure 1: The workflow for the MCF-CVA framework.

Table 1: Number of questions on each layer for MCF.

#Layer	Number of questions
Layer 1	1552
Layer 2	5457
Layer 3	2988
Layer 4	1056
Layer 5	248
Layer 6	74
Total	11375

The expansion step employs CFA to score combination and rank combination as well as average combination and weighted combination of the initial (or previous) layers to obtain $f(t) = (2^t - 1 - t) \times 4$ models. The reduction step uses the sliding rule to reduce the $f(t)$ models to t models based on diversity strength at that layer. This process continues to the next layer until there is almost no diversity between models.

MCF-CVA workflow

We propose a new framework that integrates Multilayer Combinatorial Fusion into context value alignment (MCF-CVA). Figure 1 shows the main steps for the proposed framework.

In **step (a)** of our MCF-CVA workflow, we fine-tuned five value-specific agents—Authority (A), Care (B), Fairness (C), Loyalty (D), and Sanctity (E)—starting from the OpenAssistant/oasst-sft-4-pythia-12b-epoch-3.5 SFT checkpoint, using Direct Preference Optimization (DPO) with QLoRA (Dettmers et al. 2023) on a single NVIDIA A100-40GB. The Moral Integrity Corpus (MIC) (Ziems et al. 2022) is used for fine-tuning the individual agents. MIC dataset provides a large set of prompt-response pairs of 113.8K with human revised answers and rich ethical annotations. It uses fixed data splits of 91.0K/11.4K/11.4K

samples for train/validation/test. Each moral agent was trained independently from the same base model using different subset of the dataset for 1 epoch with $\beta = 0.1$, learning rate 1×10^{-5} , per-device batch size 2 and 8 gradient-accumulation steps. Specifically, each prompt-response pair is annotated with one or more moral value labels. We partition the dataset into five different subsets, each of which corresponds to a moral value. We then fine-tune the same base model independently on each subset, resulting in five moral agents, each aligned with a specific moral value. QLoRA loads the 12B base in 4-bit NF4 and trains only LoRA adapters, enabling practical single-GPU fine-tuning. Each fine-tuned agent in step (a) generates a response for each prompt in the test set.

Step (b) decomposes each agent’s output into *moral units*, where each unit expresses a single moral claim in the form of a phrase or short sentence. The motivation for this decomposition is that a response often contains multiple moral claims, reflecting moral pluralism. By breaking a complex moral response into distinct, stand-alone claims, we can better identify and manage moral conflict while also reducing redundancy. An additional advantage is that the final response length can be controlled by combining different numbers of moral units. This decomposition is performed using GPT-4.1-nano. We then pool all units together, and form all possible unit pairs by concatenating two units. A detailed example of response decomposition into units and the formation of moral unit pool is provided in Appendix F. The unit pairs will be considered as data items in multilayer combination fusion. To score each unit pair for its alignment within each of the five moral values, we train a “moral classifier.” Concretely, we encode the human-revised answer using SentenceTransformer (all-MiniLM-L6-v2) and train a logistic regression model for multi-label prediction. Given a pair of moral units, the classifier returns five scores, one per moral value, yielding the scoring table with one row at the end of step (b). We stack the tables together for all data

Table 2: Performance summary for the best single models and the best combined models for layer 1,2,3,4 and 5 on question q_{822} , respectively. (a: ASC; b: WSCDS; c: ASC&WSCDS; d: ARC, e: WRCDS; f: ARC&WRCDS; g: all four types of combination.)

#Layer	Model type	Best model	F1 BERTScore
Layer 1	single	B	0.8582
	combined	CE (c), ACE (c), CDE (c), ACDE (c)	0.8621
Layer 2	single	D, E	0.8594
	combined	AD (g), AE (g), DE (g), ACD (e), ACE (f), ADE(g), ACDE (f)	0.8659
Layer 3	single	A, C, D, E	0.8521
	combined	AB (f), DE (g)	0.8596
Layer 4	single	B, C, D, E	0.8521
	combined	AD (c, e), AE (c), BC (g), ADE (b)	0.8596
Layer 5	single	A, B, C, D	0.8596
	combined	CD (g), BCD (d)	0.8659

items (unit pairs) to form a big scoring table. Interpreting each column in that table as a value-specific scoring system, we obtain five scoring systems w.r.t. the five models A, B, C, D, and E for the five moral agents, respectively. It’s important to note that each prompt in the test set yields a set of five scoring systems and the number of data items may vary across prompts.

In **step (c)**, we fuse the five scoring systems using MCF. The EAR (Expansion-and-Reduction) algorithm first expands the five systems into 104 scoring systems via four combination schemes (ASC, ARC, WSCDS, WRCDS), and then applies a sliding rule to reduce back to five systems (see Algorithm D1 in Appendix D). Because MCF combines models combinatorially, we enumerate model sets of sizes $k \in \{2, 3, 4, 5\}$. For each k , we construct all $\binom{5}{k}$ subsets from the five systems, yielding $10 + 10 + 5 + 1 = 26$ distinct model sets in total. For each set, we consider both score-based and rank-based combinations. In Algorithm D1, we extract the score and rank functions, respectively, from each scoring system (each system comprises a score function, a rank function, and a rank-score function). We then apply the four combination methods within each model set, producing four fused models per set. Hence, the expansion stage yields $4 \times 26 = 104$ combined models prior to reduction.

For reduction, we compute each model’s diversity strength across the 104 models. To reduce to five models, we rank all 104 models by diversity strength in descending order. We then apply a sliding-rule selection: starting at the top of ranking, we sweep downward until we obtain five scoring systems that have the highest diversity strength; these constitute the reduced set. We repeat the expansion–reduction cycle on this reduced set until a stopping criterion is met: either the maximum diversity strength falls below 0.05 or the number of layers reaches six. At the final layer, we apply average rank combination to fuse the five models into a single final model.

In **step (d)**, this single final model is used for unit aggregation. The number of unit pairs to be aggregated can be adjusted according to the desired response length. For a shorter answer, only the top-ranked unit pair is aggregated, whereas for a longer and more morally rich response, the

top two or three unit pairs can be combined. The aggregation is performed by GPT-4.1-nano. Note that sentence decomposition and unit aggregation used the same language model. Since the key part of our framework is the MCF process, we used a small off-the-shelf LLM without additional fine-tuning for both decomposition and aggregation. The resulting response is then evaluated against the ground-truth response using F1 BERTScore. The score obtained is reported as the performance of the MCF-CVA workflow for that question. This process is repeated for every question in the test set. The pseudocode for MCF-CVA is in Algorithm D2 (Appendix D).

We note that each data item corresponds to a unit pair, and aggregation over unit pairs yields the final response for a prompt. However, response decomposition typically produces many candidate units, making the choice of which two units to aggregate non-trivial, as units may be redundant or conflicting. MCF-CVA addresses this challenge by providing a systematic mechanism to leverage the diversity of scoring systems—through four distinct combinations—thereby promoting the more informative unit pair.

Results

In this section, we report results for the MCF-CVA framework: (a) Results evaluated on the Moral Integrity Corpus (MIC) dataset; (b) Ablation results for the framework to show MCF is the key for performance gain; (c) Results of applying MCF to value alignment classification task on the Commonsense dataset.

Results on MIC dataset

Table 1 summarizes the number of questions that stop at each layer. A question is said to stop at a layer if it satisfies the MCF stopping criterion and halts at that layer. Overall, 1,552 questions stop at Layer 1, while 9,823 questions continue beyond the first layer. About half of them (5,457) stop at Layer 2, 2,988 at Layer 3, and 1,056 at Layer 4. And even 248 questions go to Layer 5 while Layer 6 has 74 questions. These findings indicate that MCF-CVA adapts at the question level: different questions may continue at different

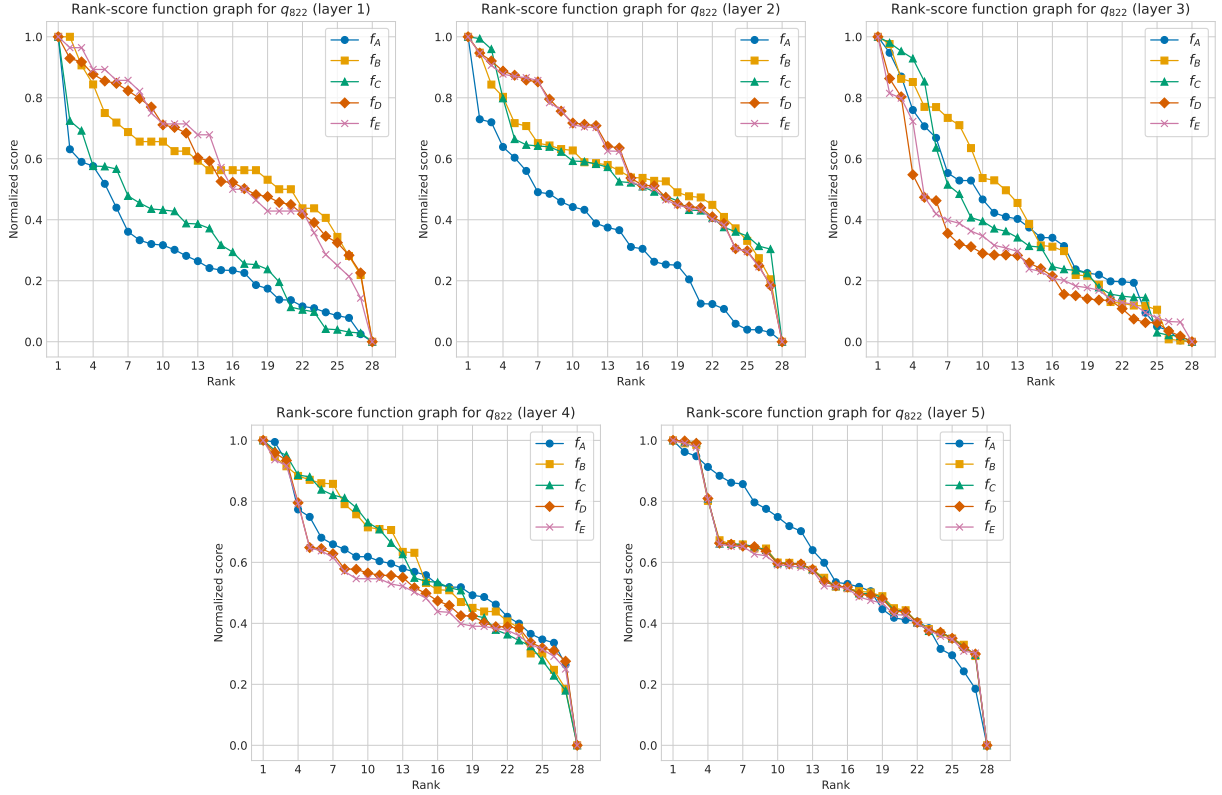


Figure 2: Rank-score function graph for 5 layers on question q_{822} .

depths, which can make MCF more effective than a single-layer CFA. We provide two illustrative examples q_{822} and q_{2835} . For each example, we plot the rank-score function graph and the performance of best combined models of each layer. The question q_{822} that stops at Layer 5 appears in Figure 2. The second example q_{2835} that stops at Layer 6 is included in Appendix E. From Figure 2, we observe how the rank-score functions evolve across five layers; their diversity decreases gradually through the five layers. Table 2 reports the best performance at each layer for question q_{822} , listing both the top individual model(s) and the best combined model(s). The best individual model at the beginning is B with an F1 BERTScore of 0.8582; through multilayer iteration the question stops at Layer 5, achieving 0.8659.

Then, we compare our MCF-CVA model performance against individual moral agents in Table 3. The first row reports test performance for the five single agents. The best agent is Loyalty with a score of 0.8663.

Next, we have some other multi-agent models. On the same dataset, prior multi-agent methods CVA-GS and CVA-GS-DYN report a BERT score of 0.8728 and 0.8754 (Dognin et al. 2025), and our previous model single-layer VAS-CVA using a single unit using CFA (Wu, Bouneffouf, and Hsu 2026) has a performance of 0.8849. Multi-agent debate improves value alignment through an iterative debate process. Following Du et al. (2023), we use the five single agents as debaters and implement the debate in 2 rounds.

We consider this debate as a 2-layer multi-agent model. In each round, every agent is given the responses of the other four agents as additional context and is asked to revise its own response accordingly. The multi-agent debate framework achieves a performance of 0.8698, which is slightly higher than that of the best individual agent.

We then evaluate the single-layer variant of the MCF-CVA framework. For each question in the test set, we use the model in the first layer (step (c) in Figure 1). The final performance is obtained by averaging over all questions. The single-layer MCF-CVA achieves an accuracy of 0.9007, outperforming all individual base models and prior multi-agent models. We next evaluate our most powerful model, multi-layer MCF-CVA. The F1 BERT score achieved is 0.9098, which is almost 1% performance gain compared to the single-layer counterpart. In sum, MCF-CVA outperforms individual agents, single-layer variant, and prior multi-agent performance.

Figure 3 provides a visual comparison of individual models, existing multi-agent fusion baselines, and the proposed MCF-CVA models. The results show a clear progression from individual agents to existing fusion methods and finally to the proposed MCF-CVA models. This visual pattern makes the improvement trend more direct than the table, showing that fusion methods generally improve over single agents, while MCF-CVA further advances the performance beyond existing fusion baselines.

One notable observation from Figure 3 is both Multiagent Debate and the proposed MCF-CVA model are multi-agent, multi-layer frameworks in the third group, but their performance differs substantially. We believe this large difference may be attributed to the underlying architecture.

Table 3: Performance comparison of individual agents, prior models and the proposed MCF-CVA models. Agent A, B, C, D, and E refer to authority, care, fairness, loyalty, and sanctity, respectively.

Method	Model	F1 BERTScore
Single model	A	0.8569
	B	0.8533
	C	0.8628
	D	0.8663
	E	0.8653
Dognin et al. (2025)	CVA-GS	0.8728
	CVA-GS-DYN	0.8754
Wu, Bouneffouf, and Hsu (2026)	VAS-CFA	0.8849
Multi-agent debate	—	0.8698
MCF-CVA	single-layer	0.9007
	multi-layer	0.9098

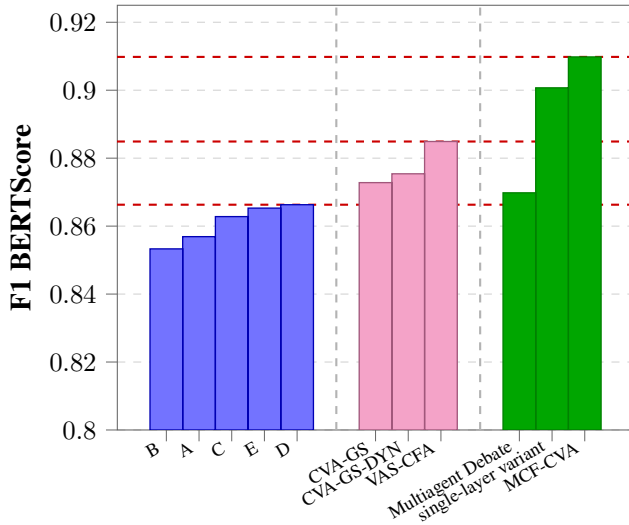


Figure 3: Comparison of MCF-CVA models with individual models (blue), other single-layer multi-agent fusion models (pink), and the multi-layer multi-agent debate model (green).

Finally, we compare the performance of the MCF models with the base models for each question, as shown in Figure 4. Each pink point represents the performance of the final MCF model for a single question, and the dataset contains 11,375 questions in total. The black line represents the performance of the best base model on the same question. The questions are sorted in descending order according to

the best base-model performance. As shown in the figure, most pink points lie above the black line, indicating that the MCF models outperform the base models for the majority of questions, although a small number of cases fall below the line. This result demonstrates that MCF not only improves the average performance compared with the base models (as shown in Table 3) but also achieves better performance on most individual questions.

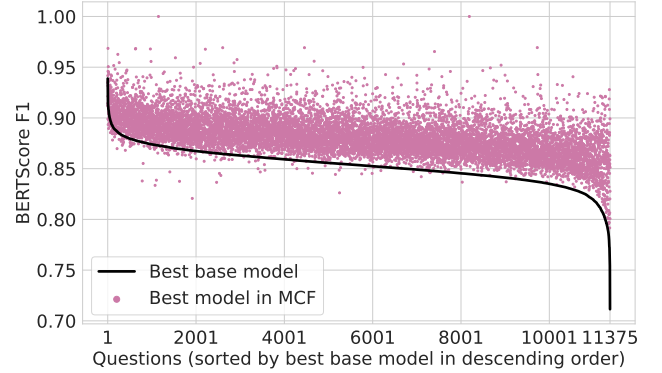


Figure 4: Performance comparison for best base models and best MCF models across all prompts in the test set.

Ablation study

Next, we conduct an ablation study to evaluate the contribution of each component in our framework. We also examine the effect of using different numbers of moral unit pairs. The results are reported in Table 4.

In the original MCF-CVA framework, we use the top 1 unit pair, which contains two moral units, for aggregation and evaluation. However, when the input context is more complex, two moral units may not be sufficient to capture the full moral meaning of the response. Therefore, we also test the setting of using the top 2 unit pairs. As shown in Table 4, using 2 unit pairs decreases the BERTScore from 0.9098 to 0.8936, a drop of about 1.6%. However, this result is still relatively strong compared with the previous baselines. One possible reason for this decrease is that the target responses in the MIC test dataset are usually short, suggesting that the moral context is relatively simple and may not require many moral units. Adding more unit pairs may introduce extra information or noise, which slightly reduces performance.

We then remove the MCF component from the framework. Without MCF, the framework cannot rank and select the most effective moral unit pairs. Therefore, we randomly select 1 or 2 unit pairs for aggregation and evaluation. When randomly selecting 1 unit pair, the BERTScore drops to 0.8827, which is a clear decrease compared with the full MCF-CVA result. When randomly selecting 2 unit pairs, the score further decreases to 0.8701. These results show that MCF plays an important role in identifying high-quality moral unit pairs and improving the final response quality.

Next, we remove the aggregation component. In this setting, the selected unit pairs are evaluated directly without be-

Table 4: Ablation study of the proposed framework by removing decomposition, MCF, or aggregation components.

Model	Moral Units	BERTScore
MCF-CVA	top 1 unit pair	0.9098
	top 2 unit pairs	0.8936
w/o MCF	random 1 unit pair	0.8827
	random 2 unit pair	0.8701
w/o aggre.	top 1 unit pair	0.9075
	top 2 unit pair	0.8916
w/o decomp. + MCF	–	0.8621

ing aggregated into a final response. The performance only slightly decreases from 0.9098 to 0.9075 when using the top 1 unit pair, and from 0.8936 to 0.8916 when using the top 2 unit pairs. This suggests that aggregation contributes to the final framework, but its effect is relatively small compared with the MCF component.

Finally, we remove both decomposition and MCF. Since MCF operates on decomposed moral units, removing decomposition also prevents the use of MCF. In this setting, we directly aggregate the original responses and evaluate the result. The BERTScore drops to 0.8621, which is lower than the best base models. This result indicates that directly combining full responses without moral-unit decomposition and MCF is not effective.

Overall, the ablation study shows that MCF is the key component driving the performance improvement of the proposed framework. Decomposition provides the necessary structure for MCF to operate on moral units, while aggregation further refines the selected units into a smooth final response.

Results on Commonsense dataset

Another common contextual value alignment task, beyond question answering, is value alignment classification: given a conversational context, the goal is to determine whether the response is aligned with human values. Our MCF-CVA framework can be adapted to this type of task. The shared core remains the use of MCF to fuse multiple scoring systems. In this subsection, we report the results on the Commonsense dataset.

The Commonsense Morality subset of the ETHICS benchmark contains scenario descriptions based on widely shared human moral intuitions regarding whether an action is clearly wrong (Hendrycks et al. 2020). The dataset includes both short (1–2 sentence) examples collected via Amazon Mechanical Turk and longer, multi-paragraph narratives curated from online forums. Labels were obtained through majority voting among multiple English-speaking annotators. The dataset is provided with a standard train/development split for supervised learning and a held-out test set for evaluation, making it suitable for training and assessing models’ ability to capture commonsense moral norms in open-world contextual scenarios.

The Commonsense dataset has two labels: 0 for not morally wrong (aligned) and 1 for clearly morally wrong (misaligned), so it is a binary classification problem. The goal is to predict whether an action “clearly should not have been done”. We apply MCF to combine five diverse base models: A) Linear SVM, B) XGBoost, C) Random Forest, D) RoBERTa-large, and E) DeBERTa-large. For the traditional ML models (A–C), we encode text using TF-IDF features, split the development set into train/validation (10% validation). For the language models (D–E), we fine-tune pre-trained transformers by tokenizing text with the model’s own tokenizer. The model diversity is strengthened by the different feature encodings (TF-IDF vs transformer tokenization). Finally, we run the MCF method iteratively until the stopping criterion is met.

The rank-score function graph for Commonsense dataset is shown in Figure 5. We can see the cognitive diversity between models in each layer shrinks across layers. The layer-1 rank-score function graph of model D (RoBERTa-large) and E (DeBERTa-large) are very close to each other while models A, B, and C cluster together. This indicates strong diversity between traditional ML models and language models, but the diversity within each group is relatively small.

Table 5: Best model performance comparison for each layer on Commonsense dataset.

	Model type	Model	Accuracy
Layer 1	single	E	0.8970
	combined	BCE(ASC)	0.9331
Layer 2	single	A	0.8988
	combined	ABE(ASC)	0.9243
Layer 3	single	D	0.9166
	combined	CDE(WSCDS)	0.9220
Hendrycks et al. (2020)	single	RoBERTa-large	0.9040
		BERT-large	0.8850

Table 5 presents the performance for the best base model and best combined model for each layer. We can see the best base model in the first layer is model E with accuracy of 0.8970. The best combined model in the first layer is BCE by average score combination, which achieves accuracy of 0.9331. It’s a big improvement. In the second layer, the performance for the best base model is 0.8988 while the performance is 0.9243 for the best combined model. The MCF stops at the third layer with the final accuracy of 0.9220. We compare our final performance on layer 3 with the result from Hendrycks et al. (2020). The final accuracy of 0.9220 exceeds the result reported in prior study, demonstrating the effectiveness of the MCF approach.

Combining all results together, multilayer combinatorial fusion can be effectively applied to both contextual value alignment tasks: question answering and alignment classification. In question answering, responses are split into units to construct scoring systems whereas in classification, the base models can directly output scores. The EAR algorithm is followed to iteratively expand and reduce the scoring sys-

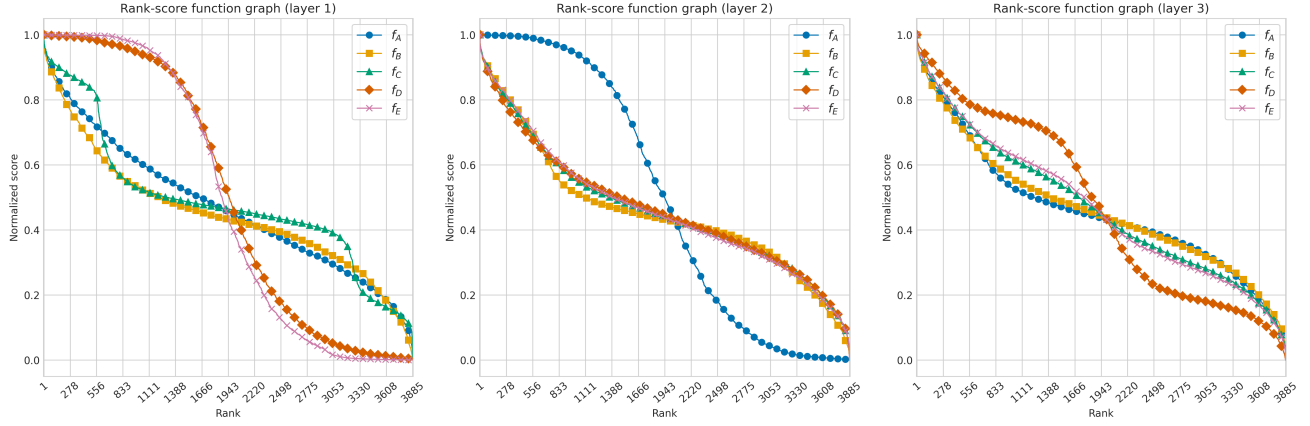


Figure 5: Rank-score function graph for Commonsense dataset.

tems until the stopping condition is reached.

Discussion

One important design choice in MCF-CVA is that the framework combines *moral units* rather than full agent responses. A complete response often contains several distinct moral claims. Directly aggregating whole responses may therefore mix together incompatible or repetitive meanings, making it harder to identify which parts of different agents are genuinely complementary. By decomposing a response into smaller stand-alone moral units, MCF-CVA creates a more structured representation of moral content, so that fusion operates on semantically cleaner building blocks. This design is also consistent with the role of decomposition in our workflow: it helps isolate conflict, reduce redundancy, and gives the framework direct control over the final response length through the number of selected unit pairs. In this sense, moral units provide an intermediate granularity that is more informative than tokens but more manageable than complete responses.

A related question is **why we use unit pairs rather than single units or larger groups such as triples**. The reason is both methodological and practical. In CFA, the scoring systems require a set of data items to compare and rank. If a single unit were treated as one data item, the pool would often be too limited to support robust fusion. At the other extreme, forming all possible triple-unit groups would substantially increase computational cost. More importantly, triple grouping would make response-length control less flexible: selecting one triple already yields three units, while selecting two triples yields six units, which may be too coarse for generating responses of different lengths. Unit pairs provide a useful middle ground. They expand the candidate space enough for meaningful scoring and ranking, but remain computationally feasible. They also offer finer control over final answer length, since longer responses can be produced by aggregating more pairs. Thus, the use of unit pairs is not arbitrary; it reflects a balance between representational adequacy, computational tractability, and controllable generation.

Another central design choice is **the set of moral values**

used to instantiate the base agents. In this work, we adopt the five canonical moral foundations—Care, Fairness, Loyalty, Authority, and Sanctity—because they provide a well-established and widely used basis for modeling moral pluralism, and we already notes that the original five-foundation structure remains the most stable starting point compared with later additions such as Liberty. At the same time, our framework does not assume that more moral values are always better. The goal of MCF-CVA is not simply to maximize the number of base agents, but to select agents with sufficient diversity. Additional values such as Liberty or Honesty could be incorporated in future extensions, but only if they contribute meaningfully different moral perspectives. If newly added values overlap too strongly with existing ones, their orthogonality is limited, and the benefit of adding them may be small. This suggests an important future direction: instead of asking how many moral values can be included, we should ask which subsets of moral values are most diverse, most complementary, and most useful for contextual alignment. More broadly, this raises a deeper question for value alignment research: whether effective pluralistic alignment requires a large inventory of moral values, or whether a smaller but carefully chosen set of sufficiently distinct foundations is already enough to capture most of the practical gains.

One question of great interest in computational learning and ensemble methods is: **How many base models (or agents, scoring systems) one should use?** Although it may depend on the domain application, an ideal number of base models is $4 \leq n \leq 6$ as long as the base models are relatively good (close to the target) and diverse. Wang and Wang (2001) demonstrated that the distance between n combined models (scoring systems) and the target sharply decreases at n between 3 and 4. Adding more base models may still improve the performance. But the principle of diminishing return kicks in.

One more issue is **tie ranking (or tie scoring)**. If a scoring system A on n data items has ties w.r.t. score function values s_A , then the rank function r_A is not a permutation of the n elements $\{1, 2, 3, \dots\}$. In this case, a ranking with

ties is a node in Kemeny rank space K_n (Figure B4, Appendix B2), which is a superset of the Bubble sort Cayley graph space B_n (Appendix B1). In such a case, a rank function with ties is represented as a matrix instead of an array (Emond and Mason 2002). We note that tie rankings do occur even if it's not often. In current work, we resolved tie ranking by taking average of the rank numbers. Hence, CFA and MCF are performed in the Bubble sort Cayley graph space (Appendix B1).

Limitations

While the results of MCF-CVA are encouraging, several limitations remain and point to important directions for future work. First, the core iterative mechanism of our framework, namely the Expansion-and-Reduction (EAR) algorithm, is currently supported primarily by empirical evidence rather than a full theoretical analysis. In the present work, the multilayer process stops when either the maximum diversity strength falls below 0.05 or the number of layers reaches six. Empirically, we observe that diversity usually decreases as the iterations proceed, and this trend can be visualized through the rank-score function graphs across layers. However, we do not yet have a closed-form characterization of the EAR dynamics, nor a formal analysis of how diversity strength evolves over iterations or under what conditions convergence behavior can be guaranteed.

A second limitation concerns evaluation. As an initial study of multilayer combinatorial fusion for value alignment in LLMs, this paper uses F1 BERTScore as the main automatic metric for comparing generated responses with reference answers. This provides a useful first view of response quality, but it does not fully capture all aspects of value alignment. Future work should broaden the evaluation framework by incorporating additional automatic metrics such as ROUGE-L and BLEURT. More importantly, human evaluation will be essential for assessing dimensions that automatic metrics cannot fully measure. Expanding evaluation along these directions would provide a more comprehensive assessment of the strengths of MCF-CVA while helping identify where the framework can be further improved.

Conclusion

This paper proposed the MCF-CVA framework, which integrates multi-agent moral reasoning with multilayer combinatorial fusion. By leveraging cognitive diversity among value-specific agents and applying the EAR algorithm across multiple layers, MCF-CVA dynamically refines moral judgments through both score and rank aggregation in Euclidean and Kemeny spaces.

Our results demonstrate that MCF-CVA achieves superior performance compared to single-agent and multi-agent baselines. The multilayer fusion process enables the model to evolve toward contextually coherent and morally pluralistic decisions while maintaining diversity strength across layers. Empirical evaluations show consistent improvements in alignment quality, indicating that multilayer aggregation offers a scalable pathway toward more context-sensitive and trustworthy AI systems.

Beyond its empirical results, MCF-CVA contributes conceptually to the study of *computational moral diversity*. It formalizes how moral agents with different normative perspectives can be combined into a single coherent decision-making system, providing a mathematical and algorithmic foundation for pluralistic alignment. The dual-space formulation further bridges moral reasoning and combinatorial optimization, linking ethics, information fusion, and multi-agent learning under a unified framework.

Future work will extend MCF-CVA along several directions. First, incorporating additional moral dimensions and dynamic weighting of moral foundations could enhance adaptability to novel cultural or situational contexts. Second, human-in-the-loop evaluation of moral coherence will help assess whether multilayer fusion aligns with human moral intuitions beyond static benchmarks. Finally, we aim to explore applications of MCF-CVA in collaborative agent systems and autonomous decision-making, where reconciling diverse moral reasoning is essential for collective intelligence.

Overall, MCF-CVA represents a principled step toward the operationalization of moral pluralism in large language models and offers a robust foundation for future research in multi-agent value alignment.

References

- Akbari, S.; and Escobedo, A. R. 2023. Beyond kemeny rank aggregation: A parameterizable-penalty framework for robust ranking aggregation with ties. *Omega*, 119: 102893.
- Alfatemi, A.; Rahouti, M.; Hsu, D. F.; Schweikert, C.; Ghani, N.; Solyman, A.; and Assaqt, M. I. S. 2025. Identifying distributed denial of service attacks through multi-model deep learning fusion and combinatorial analysis. *Journal of Network and Systems Management*, 33(1): 8.
- Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. 2022. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.
- Cocho, G.; Rodríguez, R. F.; Sánchez, S.; Flores, J.; Pineda, C.; and Gershenson, C. 2019. Rank-frequency distribution of natural languages: A difference of probabilities approach. *Physica A: Statistical Mechanics and its Applications*, 532: 121795.
- De Buck, A.; and Pauwels, L. J. 2023. Moral foundations questionnaire and moral foundations sacredness scale: Assessing the factorial structure of the Dutch translations. *Psychologica Belgica*, 63(1): 92.
- Dettmers, T.; Pagnoni, A.; Holtzman, A.; and Zettlemoyer, L. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36: 10088–10115.
- Diaconis, P. 1988. *Group representations in probability and statistics*, volume 11 of *Lecture Notes–Monograph Series*. Institute of Mathematical Statistics.
- Dognin, P.; Rios, J.; Luss, R.; Sattigeri, P.; Liu, M.; Padhi, I.; Riemer, M.; Nagireddy, M.; Varshney, K.; and Bouneffouf,

- D. 2025. Contextual value alignment. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.
- Doğruyol, B.; Alper, S.; and Yilmaz, O. 2019. The five-factor model of the moral foundations theory is stable across WEIRD and non-WEIRD cultures. *Personality and Individual Differences*, 151: 109547.
- Du, Y.; Li, S.; Torralba, A.; Tenenbaum, J. B.; and Mordatch, I. 2023. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*.
- Emond, E. J.; and Mason, D. W. 2002. A new rank correlation coefficient with application to the consensus ranking problem. *Journal of Multi-Criteria Decision Analysis*, 11(1): 17–28.
- Gabriel, I. 2020. Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3): 411–437.
- Good, I. 1975. The number of orderings of n candidates when ties are permitted. *The Fibonacci Quarterly*, 13(1): 11–18.
- Graham, J.; Haidt, J.; Koleva, S.; Motyl, M.; Iyer, R.; Wojcik, S. P.; and Ditto, P. H. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, 55–130. Elsevier.
- Graham, J.; Haidt, J.; and Nosek, B. A. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5): 1029.
- Graham, J.; Nosek, B. A.; Haidt, J.; Iyer, R.; Koleva, S.; and Ditto, P. H. 2011. Mapping the moral domain. *Journal of personality and social psychology*, 101(2): 366.
- Gross, O. A. 1962. Preferential arrangements. *The American Mathematical Monthly*, 69(1): 4–8.
- Haidt, J.; and Joseph, C. 2004. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4): 55–66.
- Hendrycks, D.; Burns, C.; Basart, S.; Critch, A.; Li, J.; Song, D.; and Steinhardt, J. 2020. Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*.
- Hsu, D. F.; Chung, Y.-S.; and Kristal, B. S. 2006. Combinatorial fusion analysis: methods and practices of combining multiple scoring systems. In *Advanced Data Mining Technologies in Bioinformatics*, 32–62.
- Hsu, D. F.; Kristal, B. S.; Hao, Y.; and Schweikert, C. 2019. Cognitive diversity: A measurement of dissimilarity between multiple scoring systems. *Journal of Interconnection Networks*, 19(01): 1940001.
- Hsu, D. F.; Kristal, B. S.; and Schweikert, C. 2024. Combinatorial fusion analysis. *Computer*, 57(09): 96–100.
- Hsu, D. F.; Shapiro, J.; and Taksa, I. 2002. Methods of data fusion in information retrieval: Rank vs. score combination. *DIMACS Technical Report*, 58: 662–667.
- Hsu, D. F.; and Taksa, I. 2005. Comparing rank and score combination methods for data fusion in information retrieval. *Information Retrieval*, 8(3): 449–480.
- Hurley, L.; Kristal, B. S.; Sirimulla, S.; Schweikert, C.; and Hsu, D. F. 2020. Multi-layer combinatorial fusion using cognitive diversity. *IEEE Access*, 9: 3919–3935.
- Ji, J.; Qiu, T.; Chen, B.; Zhou, J.; Zhang, B.; Hong, D.; Lou, H.; Wang, K.; Duan, Y.; He, Z.; et al. 2025. Ai alignment: A contemporary survey. *ACM Computing Surveys*, 58(5): 1–38.
- Jiang, N.; Quazi, M.; Schweikert, C.; Hsu, D.; Oprea, T.; and Sirimulla, S. 2023. Enhancing ADMET property models performance through combinatorial fusion analysis. *ChemRxiv*; doi:10.26434/chemrxiv-2023-dh70x.
- Kemeny, J. G.; and Snell, J. L. 1962a. Mathematical Models in the Social Sciences, chapter Preference Rankings: An Axiomatic Approach.
- Kemeny, J. G.; and Snell, L. 1962b. Preference ranking: an axiomatic approach. *Mathematical models in the social sciences*, 9–23.
- Kuncheva, L. I. 2004. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley. ISBN 0-471-21078-1.
- Kustiawan, I.; Liu, C.-Y.; and Hsu, D. F. 2017. Vertical handoff decision using fuzzification and combinatorial fusion. *IEEE Communications Letters*, 21(9): 2089–2092.
- Li, X.; Jiang, Q.; Jiang, L.; Zhang, S.; and Hu, S. 2026. The landscape of AI alignment: A comprehensive review of theories and methods. *International Journal of Pattern Recognition and Artificial Intelligence*, 40(03): 2539001.
- Marmor, Y. N.; Gadrich, T.; and Bashkansky, E. 2021. Accuracy of multi-experts’ prioritization under Mallows’ model of errors creation. *Quality Engineering*, 33(2): 286–299.
- Noothigattu, R.; Gaikwad, S.; Awad, E.; Dsouza, S.; Rahwan, I.; Ravikumar, P.; and Procaccia, A. 2018. A voting-based system for ethical decision making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
- Norhashim, H.; and Hahn, J. 2024. Measuring human-ai value alignment in large language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 1063–1073.
- Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.
- Owusu, E.; Rahouti, M.; Verma, D.; Xin, Y.; Hsu, D. F.; and Schweikert, C. 2025. Generalizable Multi-Model Fusion for Multi-Class DoS Detection Using Cognitive Diversity and Rank-Score Analysis. *ACM Transactions on Privacy and Security*, 28(3): 1–37.
- Pang, X.; Tang, S.; Ye, R.; Xiong, Y.; Zhang, B.; Wang, Y.; and Chen, S. 2024. Self-alignment of large language models via multi-agent social simulation. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.
- Pappas, I. O.; and Vassilakopoulou, P. 2025. Human Value Alignment in AI. In *Handbook of Human-Centered Artificial Intelligence*, 1–33. Springer.

- Patten, S.; Chen, P.-Y.; Schweikert, C.; and Hsu, D. F. 2025. Enhancing Sentiment Classification with Machine Learning and Combinatorial Fusion. In *2025 IEEE Conference on Pervasive and Intelligent Computing (PICom)*, 74–81. IEEE.
- Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36: 53728–53741.
- Russell, S.; Dewey, D.; and Tegmark, M. 2015. Research priorities for robust and beneficial artificial intelligence. *AI Magazine*, 36(4): 105–114.
- Singh, S.; Kumar, A.; Jha, A.; Jacob, N.; and Mondal, M. S. 2025. AI Alignment: Ensuring AI objectives match human values. *Int. J. Sci. Res. Eng. Manag*, 9(04): 1–9.
- Tang, Y.; Li, Z.; Nellikkal, M. A. N.; Eramian, H.; Chan, E. M.; Norquist, A. J.; Hsu, D. F.; and Schrier, J. 2021. Improving data and prediction quality of high-throughput perovskite synthesis with model fusion. *Journal of Chemical Information and Modeling*, 61(4): 1593–1602.
- Wang, R.; and Wang, S. 2001. How does consensus scoring work for virtual library screening? An idealized computer experiment. *Journal of Chemical Information and Computer Sciences*, 41(5): 1422–1426.
- Wu, Y.; Bouneffouf, D.; and Hsu, D. F. 2026. Enhancing Value Alignment of LLMs with Multi-Agent System and Combinatorial Fusion. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1756–1760. IEEE.
- Wu, Y.; Ye, W.; Xu, J.; and Hsu, D. F. 2025. Bitcoin Price Prediction using Machine Learning and Combinatorial Fusion Analysis. In *2025 IEEE Conference on Artificial Intelligence (CAI)*, 61–68. IEEE.
- Wu, Z.; Hu, Y.; Shi, W.; Dziri, N.; Suhr, A.; Ammanabrolu, P.; Smith, N. A.; Ostendorf, M.; and Hajishirzi, H. 2023. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36: 59008–59033.
- Xu, J.; LaFleur, M. L.; Schweikert, C.; and Hsu, D. F. 2025. Enhancing SDG-Text Classification with Combinatorial Fusion Analysis and Generative AI. In *2025 IEEE Conference on Pervasive and Intelligent Computing (PICom)*, 58–65. IEEE.
- Zhang, J. 2004. Binary choice, subset choice, random utility, and ranking: A unified perspective using the permutahedron. *Journal of Mathematical Psychology*, 48(2): 107–134.
- Zhong, X.; Hurley, L.; Sirimulla, S.; Schweikert, C.; and Hsu, D. 2019. Combining multiple ranking systems on the generalized permutation rank space. In *2019 IEEE 5th International Conference on Big Data Intelligence and Computing (DATACom)*, 123–129. IEEE.
- Zhou, Z.-H. 2012. Ensemble Methods: Foundations and Algorithms.
- Ziems, C.; Yu, J.; Wang, Y.-C.; Halevy, A.; and Yang, D. 2022. The Moral Integrity Corpus: A Benchmark for Ethical Dialogue Systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3755–3773.
- Zipf, G. K. 2016. *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio books.

Contextual Value Alignment via Multilayer Combinatorial Fusion

Appendix A: Rank-score Function and Cognitive Diversity

Under a scoring system A, score function s_A , its derived rank function r_A , and rank-score function f_A as well as their relationships are shown in Figure A1(a). We note that the rank-score function f_A of a scoring system A has two distinctive properties. It not only characterizes the scoring (or ranking) behaviour of the scoring system A but also provides a projection (and relation) between the rank function values and the score function values. (Hsu, Shapiro, and Taksa 2002; Hsu, Chung, and Kristal 2006). The notion of rank-score function is analogous to but different from the power law phenomena and rank-frequency distribution (Zipf 2016; Cocho et al. 2019). Figure A1(b) compares cognitive diversity (CD), defined as difference between two rank-score functions f_A and f_B , to three statistical data correlations: (1) Pearson correlation (p) between score function s_A and s_B , (2) Spearman's rho correlation (ρ) and (3) Kendall's tau correlation (τ) between r_A and r_B (Hsu et al. 2019). Moreover, CD is also shown to be analogous to but different from several diversity measures in computational learning and ensemble methods (Kuncheva 2004; Zhou 2012; Hsu et al. 2019; Hurley et al. 2020).

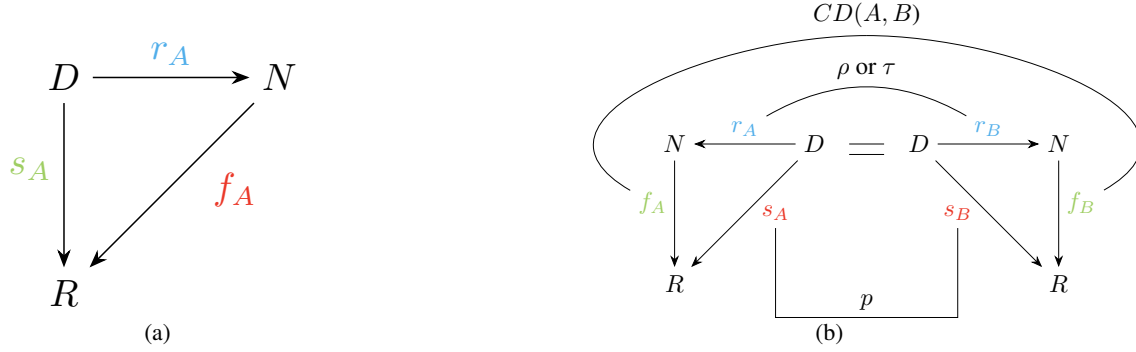


Figure A1: (a) Relationship among rank function r_A , score function s_A and rank-score function f_A (Hsu, Chung, and Kristal 2006); (b) Cognitive diversity CD vs. Pearson's correlation (p), Spearman's correlation (ρ) and Kendall's correlation (τ) (Hsu et al. 2019).

Appendix B: Kemeny Rank Space as the Architecture for the MCF-CVA Framework

B1: Bubble sort Cayley graph space and the permutahedron

In the symmetric group S_n , let T_n be the subset of all $n - 1$ adjacent transpositions $\{(1\ 2), (2\ 3), \dots, (n - 1\ n)\}$. The Cayley graph $\text{Cay}(S_n, T_n)$ is constructed as the graph with S_n as the vertex set and $\{(\pi, \pi \circ t)\}$ as the edge set, where $\pi \in S_n$ and $t \in T_n$ (Diaconis 1988). This graph $B_n = \text{Cay}(S_n, T_n)$ is also called Bubble Sort Cayley Graph because any pair of vertices A and B are connected by a path of distance equal to the number of adjacent swaps by the sorting algorithm: bubble sort. The graph B_n has many combinatorial and graphical properties. For example, it is $(n - 1)$ -regular because the generating set T_n has $n - 1$ elements; it is bipartite and has diameter $n(n - 1)/2$. Moreover, graph B_{n+1} can be recursively constructed by $n + 1$ copies of B_n . Figure B1 illustrates the graph B_3 and Figure B2 shows the two representations of the graph of B_4 .

If we treat each permutation (or rankings without ties) as a vector, the set of all permutations on the n elements constitutes the convex hull of a permutahedron P_n (Zhang 2004). Considering each permutation of the set of data items D , $|D| = n$, as a preference chain of these n items, the geodesic distance metric (GDM) of dissimilarity (or diversity) between two preference chains is proportional to the angle and hence the length of the geodesic arc connecting the two corresponding points on the surface of the n -dimensional sphere. Marmor, Gadrich, and Bashkansky (2021) showed that this GDM satisfies the requirements of a distance metric, including the triangle inequality, and is related to Kendall's τ correlation coefficient of similarity. Examples of P_3 and P_4 are illustrated in Figure B3.

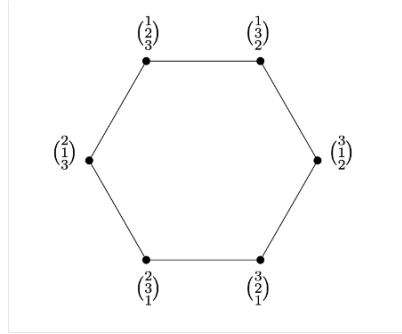


Figure B1: B_3 (Zhong et al. 2019).

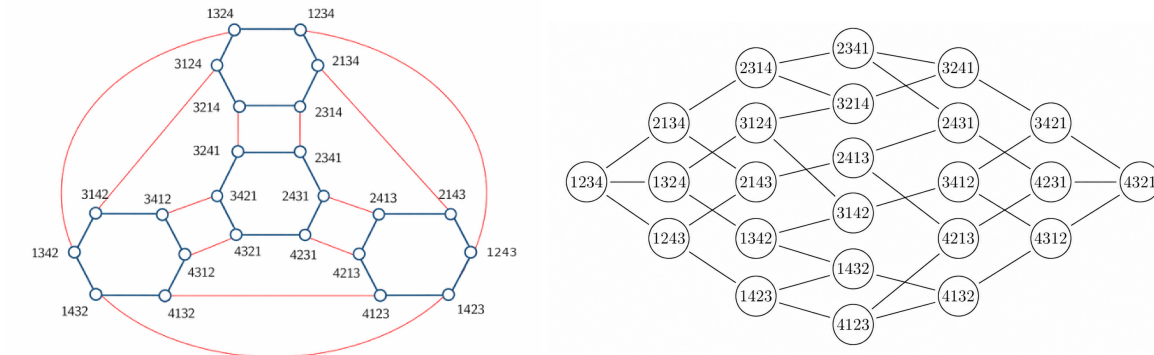


Figure B2: Left: B_4 (recursive); Right: B_4 (Cayley graph structure with generating set $\{(12), (23), (34)\}$) (Zhong et al. 2019).

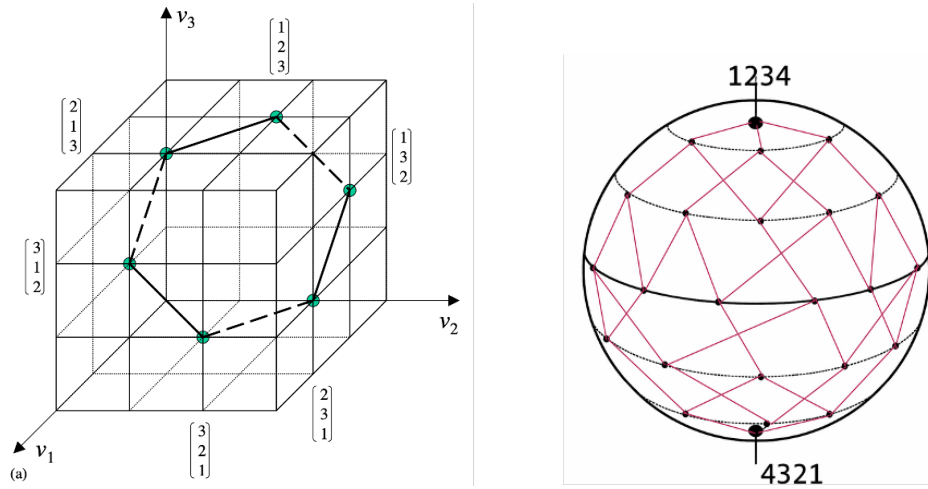


Figure B3: Left: P_3 (Zhang 2004); Right: P_4 (Marmor, Gadrich, and Bashkansky 2021).

B2: Kemeny rank space

When the score function s_A of a scoring system A is not 1-1 function, the derived rank function is not a permutation. This is the case when “tie ranking” occurs. Although there are several methods to resolve tie ranking situation such as taking the average rank, Kendall’s τ correlation distance does not apply to tie rankings. Kemeny and Snell (1962a) proposed a distance metric d_K that can handle ties. The d_K metric was updated by Emond and Mason (2002) to a new rank correlation τ_x between two weak orders (rankings with ties) as inner product of their score function values due to efficient computation.

Since the Kemeny rank space K_n consists of all rankings (ties or no ties) of the dataset items $\{d_1, d_2, \dots, d_n\}$, the vertex set of the Kemeny rank space K_n has more vertices than the bubble-sort Cayley graph space B_n , which is $n!$. The number of vertices $f(n)$ of the Kemeny rank space K_n has been calculated from two perspectives. Good (1975) showed that $f(n) = \sum_{b=1}^n b! S(n, b)$, where $S(n, b)$ is the Stirling number of the second kind. On the other hand, $f(n)$ was obtained by Gross (1962) as a recursive formula $f(n) = 1 + \sum_{i=1}^n \binom{n}{i} f(n-i)$ in terms of preference arrangements w.r.t. n objects allowing indifferences. The Kemeny rank space K_3 is illustrated in Figure B4.

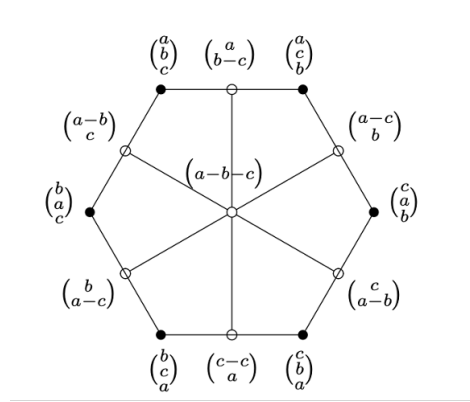


Figure B4: Kemeny rank space K_3 (Kemeny and Snell 1962b).

The Kemeny rank space K_n has been used in information retrieval, multi-criteria decision making, and computational social science (Akbari and Escobedo 2023). More recently, focus has been on rank aggregation on the Kemeny space, in which a derived median ranking is pursued in terms of optimizing computational complexity (Akbari and Escobedo 2023). In our application, we perform multi-agent multilayer combinatorial fusion (MCF) on the Kemeny rank space using both average and weighted combinations of multi-agents’ rankings.

Appendix C: Figure Skating Judgement

In this section, we use an example in figure skating judgement to illustrate (a) score combination v.s. rank combination and (b) rank-score function and comparison of rank-score function graphs w.r.t. judges J_1 , J_2 , and J_3 (Jiang et al. 2023). In the judgement of figure skating competition, suppose we would like to rank a group of 8 skaters by 3 judges. The set of skaters $D = \{d_1, d_2, \dots, d_8\}$, and the set of judges is $\{J_1, J_2, J_3\}$. The following tables C1, C2, C3 and C4 illustrate the score functions s_{J_1} , s_{J_2} , and s_{J_3} and their score combinations; the rank functions r_{J_1} , r_{J_2} , and r_{J_3} and their rank combinations; the normalized score values; and the rank-score functions f_{J_1} , f_{J_2} , and f_{J_3} , respectively. Figure C1 illustrates the rank-score function graphs. The cognitive diversity between scoring systems A and B is the area between these two graphs.

Note that this example demonstrates that judge J_3 manipulates the score on skater d_8 . Table C2 indicates how to mitigate the problem while Figure C1 shows how to detect the bias by exhibiting rank-score function graphs of judges J_1 , J_2 , and J_3 .

Table C1: Score function and score combination.

	s_{J_1}	s_{J_2}	s_{J_3}	s_{SC}	r_{SC}
d_1	9.6	9.7	9.8	29.1	2
d_2	9.8	9.2	9.9	28.9	3
d_3	9.7	9.9	10	29.6	1
d_4	9.5	9.3	9.7	28.5	6
d_5	9.9	9.4	9.5	28.8	4
d_6	9.4	9.6	9.6	28.6	5
d_7	9.3	9.5	9.4	28.2	7
d_8	10	10	7	27.0	8

Table C2: Rank function and rank combination.

	r_{J_1}	r_{J_2}	r_{J_3}	s_{RC}	r_{RC}
d_1	5	3	3	11	3
d_2	3	8	2	13	4
d_3	4	2	1	7	1
d_4	6	7	4	17	7
d_5	2	6	6	14	5
d_6	7	4	5	16	6
d_7	8	5	7	20	8
d_8	1	1	8	10	2

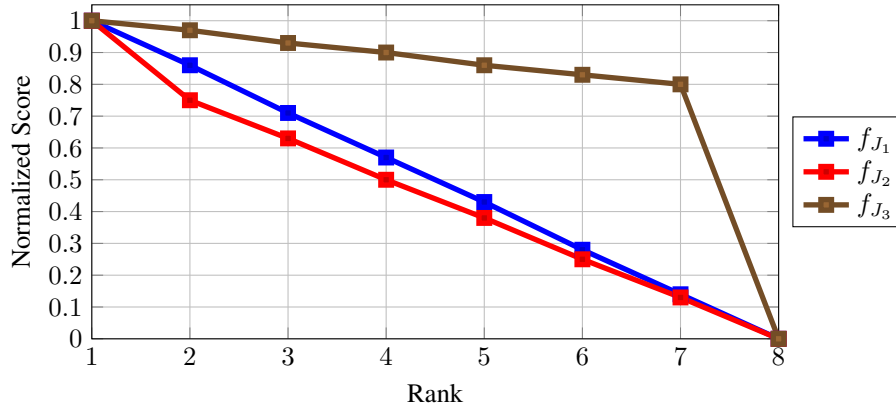
Table C3: Normalized score values.

	s_{J_1}	s_{J_2}	s_{J_3}
d_1	0.43	0.63	0.93
d_2	0.71	0.00	0.97
d_3	0.57	0.75	1.00
d_4	0.28	0.13	0.90
d_5	0.86	0.25	0.83
d_6	0.14	0.50	0.86
d_7	0.00	0.38	0.80
d_8	1.00	1.00	0.00

Table C4: Rank-score functions f_{J_1} , f_{J_2} , f_{J_3} .

Rank	f_{J_1}	f_{J_2}	f_{J_3}
1	1.00	1.00	1.00
2	0.86	0.75	0.97
3	0.71	0.63	0.93
4	0.57	0.50	0.90
5	0.43	0.38	0.86
6	0.28	0.25	0.83
7	0.14	0.13	0.80
8	0.00	0.00	0.00

Figure C1: Rank-score function graphs.



Appendix D: EAR Algorithm and the MCF-CVA Framework

Algorithm D1 Expansion-and-Reduction (EAR)

Input: t scoring systems $\mathcal{M} = \{M_1, \dots, M_t\}$ and set of data items $D = \{d_1, d_2, \dots, d_n\}$.

Output: New t scoring systems $\mathcal{M}' = \{M'_1, \dots, M'_t\}$ for next layer.

```

1: for  $k = 2, 3, \dots, t$  do                                ▷  $k$  models to combine: 2-com, 3-com,  $\dots$ ,  $t$ -com.
2:   for all  $I \in \text{Choose}(\{1, \dots, t\}, k)$  do              ▷ Select which  $k$  scoring systems for combination.
3:      $\{\mathcal{S}_i : i \in I\} \leftarrow$  Get score function for each system  $\{\mathcal{M}_i : i \in I\}$ .
4:      $S_I^{\text{ASC}} \leftarrow$  Do average score combination for  $\{\mathcal{S}_i : i \in I\}$  using formula (1).
5:     Compute cognitive diversity for every pair from  $\{\mathcal{S}_i : i \in I\}$ .
6:      $DS \leftarrow$  Compute diversity strength for each system.
7:      $S_I^{\text{WSCDS}} \leftarrow$  Do weighted score combination by DS using formula (2).
8:      $\{\mathcal{R}_i : i \in I\} \leftarrow$  Get rank function for each system  $\{\mathcal{M}_i : i \in I\}$ .
9:      $S_I^{\text{ARC}} \leftarrow$  Do average rank combination for  $\{\mathcal{R}_i : i \in I\}$  using formula (3).
10:     $S_I^{\text{WRCDS}} \leftarrow$  Do weighted rank combination by DS using formula (4).
11:  end for
12: end for
13:  $\mathcal{S}_{n \times (4(2^t - t - 1))} \leftarrow$  Horizontally stack  $\{S_I^{\text{ASC}}, S_I^{\text{WSCDS}}, S_I^{\text{SARC}}, S_I^{\text{WRCDS}}\}$  together.  ▷ Each fusion type has  $2^t - t - 1$ 
    combs.
14:  $DS_{1 \times (4(2^t - t - 1))} \leftarrow$  Compute diversity strength for each column of  $\mathcal{S}_{n \times (4(2^t - t - 1))}$ .
15:  $\{\mathcal{M}'_1, \dots, \mathcal{M}'_t\} \leftarrow$  Obtain top  $t$  systems from  $\mathcal{S}_{n \times (4(2^t - t - 1))}$  based on  $DS_{1 \times (4(2^t - t - 1))}$ .  ▷ Sort all systems by DS in
    descending order and take the top  $t$  systems.
16: Return  $\{\mathcal{M}'_1, \dots, \mathcal{M}'_t\}$ 

```

Algorithm D2 Multilayer combinatorial fusion for contextual value alignment (MCF-CVA)

Input: Five fine-tuned moral agents $\mathcal{A} = \{A_1, \dots, A_5\}$; test question set $\mathcal{Q}$.

Output: The F1 BERTScore at final layer, MCFperf and the last layer ℓ^* at which the stopping criterion is met.

```
for all  $q \in \mathcal{Q}$  do                                     ▷ Per question.
  for each moral agent  $A_i \in \mathcal{A}$  do
     $r_i \leftarrow$  Generate a response for agent  $A_i$  on question  $q$ .
     $U_i \leftarrow$  Partition  $r_i$  into moral units using GPT-4.1-nano.
  end for
   $U \leftarrow$  Pool all moral units  $U_1 \cup U_2 \cup \dots \cup U_5$ .
   $\langle d_1, \dots, d_n \rangle \leftarrow$  Form all possible unit pairs  $\{(u_i, u_j) \mid u_i, u_j \in U, i < j\}$  as data items.
   $ST \leftarrow$  SentenceTransformer(all-MiniLM-L6-v2).
  for  $k \leftarrow$  Each data item (unit pair) 1 to  $n$  do
    Encode each data item to tokens for  $d_k$  using ST.                                     ▷ Features for each pair.
     $\hat{p}_k \leftarrow$  Train Logistic Regression and get predictions for  $d_k$ .                 ▷ Pre-trained multi-label classifier.
  end for
   $\{\mathcal{S}_1, \dots, \mathcal{S}_5\} \leftarrow$  Build five scoring systems  $\{\hat{p}_k\}_{k=1}^n$ .                 ▷ Each aligned with one moral value.
  Initialize  $\ell = 0$ .                                     ▷ Layer 0.
  while StoppingCriteria( $\{\mathcal{M}_1, \dots, \mathcal{M}_5\}$ ) do       ▷ Maximum diversity strength  $\geq 0.05$  and number of layer  $\leq 6$ .
     $\ell \leftarrow \ell + 1$ .                                   ▷ Layer increases.
     $\{\mathcal{M}_1, \dots, \mathcal{M}_5\} \leftarrow$  Expansion-and-Reduction( $\{\mathcal{M}_1, \dots, \mathcal{M}_5\}$ ).       ▷ From Algorithm D1.
  end while
   $\ell^*(q) \leftarrow \ell$ .                                   ▷ Store the number of layer for question  $q$ .
  Average rank combination of  $\{\mathcal{M}_1, \dots, \mathcal{M}_5\}$ .       ▷ Get final ranking of data items (unit pair).
  Obtain best data item(s).                               ▷ Sort ranks ascendingly to get the top data item(s).
  MCFperf( $q$ )  $\leftarrow$  Aggregating the data item(s).
end for
Return  $\langle \text{MCFperf}, \ell^* \rangle$ 
```

Appendix E: Performance Summary w.r.t. Question q_{2835}

Table E1: Performance summary for the best single models and the best combined models for layer 1,2,3,4, 5 and 6 on question q_{2835} , respectively (a: ASC; b: WSCDS; c: ASC&WSCDS; d: ARC, e: WRCDS; f: ARC&WRCDS; g: all four types of combination.)

#Layer	Model type	Best model	F1 BERTScore
Layer 1	single	C	0.8663
	combined	AC (d), ACE (d), ABCDE (d)	0.8756
Layer 2	single	C, E	0.8628
	combined	AC (f), CD (g), ACD (d)	0.8698
Layer 3	single	A	0.8698
	combined	ABC (d), ACE (f)	0.8708
Layer 4	single	A	0.8698
	combined	AB (f), ABD (g), BDE (a)	0.8755
Layer 5	single	B	0.8698
	combined	CE (f), DE (d), ADE (a)	0.8755
Layer 6	single	A	0.8698
	combined	AE (g), ACE (f), ADE (d)	0.8797

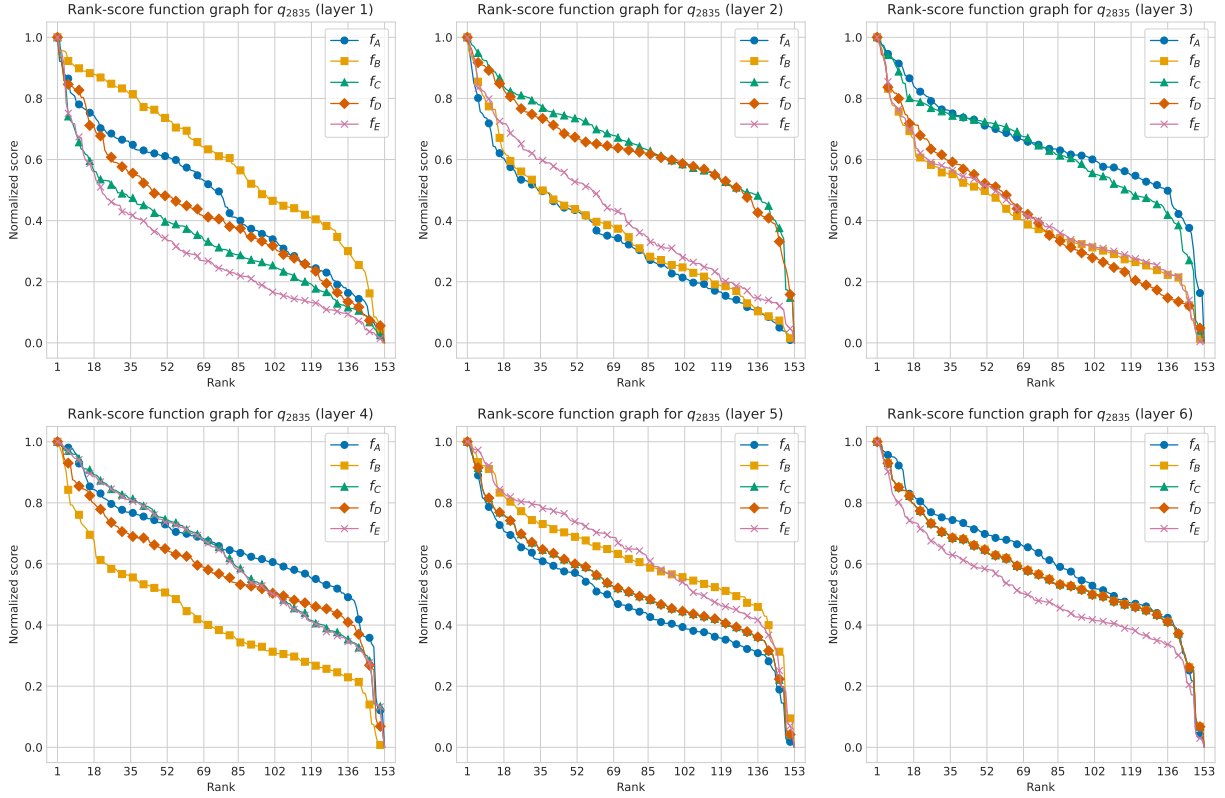


Figure E1: Rank-score function graph for 6 layers w.r.t. question q_{2835} .

Appendix F: A Detailed Example of Response Decomposition in MCF-CVA Workflow

Question: If we never started to globally recycle in the 90s, how bad would it be right now and would the world look any different?

Moral agent responses

Moral agent A (Authority): If global recycling had never begun in the 1990s, governments and institutions would have faced greater difficulty maintaining environmental order and public standards. The lack of coordinated recycling policies would weaken regulatory authority, leading to more uncontrolled waste and environmental degradation.

Moral agent B (Care): Recycling has been beneficial to people and other living beings, reducing pollution and helping protect the places they depend on. Without recycling, there would likely be much more pollution today, causing greater harm to communities, wildlife, and ecosystems.

Moral agent C (Fairness): If we never started to globally recycle in the 90s, the world would be worse off in how environmental costs and benefits are shared. The amount of waste and pollution would be much higher, and there would be fewer efforts to conserve our shared environment.

Moral agent D (Loyalty): If global recycling had never started in the 1990s, societies would show less collective responsibility toward protecting shared environments. Communities would be less united in caring for common resources, weakening cooperation across nations and groups.

Moral agent E (Sanctity): If global recycling had never begun in the 1990s, the natural world would be treated with less respect and care. Pollution and waste would accumulate more heavily, degrading the purity of land and water. This neglect would reflect a failure to protect the integrity of the environment.

A1. Without global recycling, governments and institutions would have weaker control over environmental regulation and waste management.
A2. The absence of coordinated recycling policies would undermine public order and environmental standards enforced by authorities.

B1. Recycling has been beneficial to people and other living beings.
B2. Recycling reduces pollution and helps protect the environments.
B3. If recycling had not been widely adopted, pollution levels today would likely be much higher.
B4. Higher pollution levels would cause greater harm to human communities, wildlife, and ecosystems.

C1. The absence of global recycling since the 1990s would make the distribution of environmental costs and benefits more unfair.
C2. Without global recycling, the amount of waste and pollution would be much higher.
C3. Global recycling programs help conserve the shared environment by reducing waste and reusing materials.

D1. Global recycling encourages collective responsibility toward protecting shared environmental resources.
D2. Without recycling, cooperation and solidarity among communities and nations in environmental protection would be weaker.

E1. Without recycling, pollution and waste would accumulate, degrading the purity of land and water.
E2. Recycling reflects respect for the natural world and its inherent cleanliness and integrity.
E3. Neglecting recycling represents a failure to protect the sanctity of the environment from contamination.

Unit pair pool: $\binom{14}{2} = 91$ unit pairs for the base units { A1, A2, B1, B2, B3, B4, C1, C2, C3, D1, D2, E1, E2, E3 }.

Best unit pair (B3, C3): If recycling had not been widely adopted, pollution levels today would likely be much higher. Global recycling programs help conserve the shared environment by reducing waste and reusing materials.

Aggregation: If we had never started global recycling in the 90s, pollution levels today would likely be much higher, because recycling programs help conserve our shared environment by reducing waste and reusing materials.

Ground truth: There could be a lot of pollution. Recycling helps conserve the environment.

Figure F1: A detailed example of response decomposition in MCF-CVA workflow.