已接收　参加 2026年IEEE声学、语音与信号处理国际会议（ICASSP）

通过多智能体系统与组合融合增强大语言模型的价值对齐

Yuanhong Wu⋆、*Djallel Bouneffouf*† 和 *D. Frank Hsu*⋆

⋆ 福特汉姆大学计算机与信息科学系，美国纽约州纽约市
† IBM研究院 T. J. Watson研究中心，美国纽约州约克镇高地

arXiv:2603.11126v1 [cs.MA] 11 Mar 2026

摘要

使大语言模型（LLM）与人类价值观保持一致，是确保其可信、安全部署所面临的核心挑战。尽管基于人类反馈的强化学习（RLHF）及其变体等现有方法提升了对齐效果，但它们往往依赖单个评估者或范围狭窄的奖励信号，因此难以充分体现伦理多元主义。本文提出基于组合融合分析的价值对齐系统（VAS-CFA），这是一种将多智能体融合对齐付诸实践的框架。该系统设置了多个道德智能体，每个智能体都经过微调，以代表一种独特的规范性视角；随后，系统采用 CFA，通过基于排序和基于分数的聚合方式融合各智能体的输出。这一设计利用智能体之间的认知多样性，缓解多个智能体间的冲突与冗余，从而生成更能体现人类价值观的回答。实证评估表明，在标准度量上，VAS-CFA 优于单智能体基线及既有聚合方法，说明多智能体融合为推进大语言模型的价值对齐提供了一种稳健有效的机制。

索引词—认知多样性、组合融合分析、大语言模型、多智能体系统、价值对齐

1. 引言

使大语言模型（LLM）与人类价值观对齐至关重要，因为基于广泛网络语料库预训练的模型可能会生成不真实、不安全或不符合用户意图的输出 [1, 2]。开发对齐方法旨在弥合这些差距。近年来，人们开发了许多技术，以更好地使大语言模型与人类价值观对齐，下面将对其中几种进行总结。

经典的对齐方法是基于人类反馈的强化学习（RLHF）。这一过程包括先对示范进行监督微调，再使用基于人类成对偏好训练的奖励模型进行策略优化，从而显著改善有毒或不安全的输出 [3]。为降低人工标注的高昂成本，基于人工智能反馈的强化学习（RLAIF） [4] 在很大程度上以人工智能评审者取代人工评审者，这些人工智能评审者依据书面宪章提供批评和偏好标签，从而以远少于人工评分的方式打造出无害但不回避问题的助手。其他方法则旨在简化复杂的优化过程。例如，直接偏好优化（DPO） [5]以针对偏好对的简单监督目标取代在线强化学习，但其表现仍可媲美甚至超过采用近端策略优化（PPO）的基于人类反馈的强化学习。

研究人员不再局限于成对偏好，而是开始利用更丰富的文本信号，直接引导模型行为。例如，Self-Refine [6]指示模型修改自身输出，无需额外更新权重即可提升人类偏好和任务表现分数。类似地，Reflexion [7] 通过让模型生成简短的反思，将反馈转化为口头自我监督，以指导模型后续尝试，从而在无需额外训练的情况下提升推理表现。Diverse AI Feedback（DAIF） [8] 整合了这些理念，将批评、改进和偏好纳入同一流程。它会将任务分配给信息量最大的反馈类型，表现优于单一反馈基线。这表明，异质反馈是提升对齐效果的有力手段。

越来越多的文献警告称，RLxF 可能导致目标狭隘，并忽视关键的伦理复杂性 [9]。为应对这一差距，相关方法正在不断涌现。User-Driven Value Alignment [10] 记录了用户识别、质疑和纠正偏见输出的策略，从而推动了用于自下而上的引导和集体审计的界面设计。在社区层面，STELA [11]与代表性不足群体开展审议式规范引导，并制定社区价值对齐规则集，将对齐标准定位为可迭代、可修订的标准。

本文提出的系统与以单个智能体为核心的既有对齐方法不同。基于CFA的价值对齐系统（VAS-CFA）汇集多个各不相同的道德智能体，每个智能体都旨在体现一种独特的规范性视角，并通过结合基于分数和基于排名的聚合方法的组合融合分析框架，融合它们发出的信号。VAS-CFA利用智能体之间的多样性，而非依赖单个评估者，并明确将排名纳入融合步骤，从而提供了一种新的

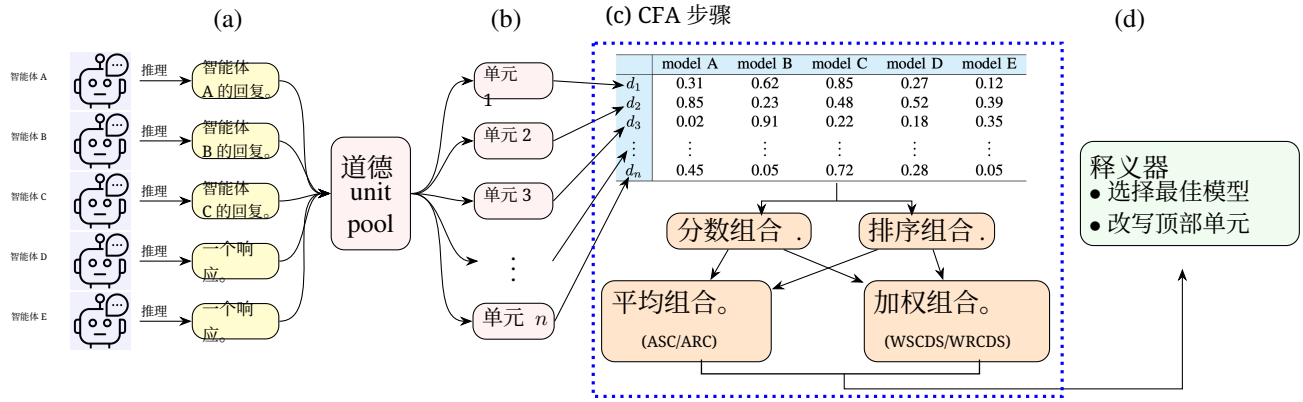

图 1：基于组合融合分析的价值对齐系统（VAS-CFA）示意图。

机制，引导大语言模型采取更能体现人类价值观的行为。

2. 组合融合分析（CFA）和 Kemeny 排名空间（K_N）

组合融合分析（CFA）为计算学习与建模、信息学以及智能 ML/AI 系统中多个评分系统（MSS）（例如多个分类器系统、多个专家系统、多个语言模型和多个智能体系统）的组合提供方法和工作流程 [12, 13]。CFA 将评分系统 A 描述为包含评分函数 s_A、由此导出的排名函数r_A，以及将归一化评分值与排名值关联起来的函数 [12–14]。

设 A 是数据集 $D = \{d_1, \ldots, d_n\}$上的一个评分系统。设 $s_A: D \to R$ 是一个评分函数。排名函数 $r_A: D \to N$ 通过对评分值排序，并根据递减的评分值为 D 中的数据项分配递增的排名值而导出。排名-评分函数（RSF） $f_A: N \to R$ 定义为 $f_A(i) = s_A(r_A^{-1}(i)) = (s_A \circ r_A^{-1})(i)$。对于评分系统 A 和 B， A 与 B之间的认知多样性（CD） $CD(A, B)$定义为 f_A与 f_B [12, 14, 15]之差： $CD(A, B) = d(f_A, f_B) = \left(\sum_{i=1}^{n}(f_A(i) - f_B(i))^2/(n-1)\right)^{1/2}$。此外，评分系统 A的多样性强度 $DS(A)$，是所考察的评分系统 A' 中， A 与所有其他评分系统之间的 $CD(A, A')$ 的平均值。

基于多样性强度的加权评分组合（WSCDS）和基于多样性强度的加权排名组合（WRCDS）分别为 t 在 {2, 3, 4, 5}中的定义如下

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

$$s_{WRCDS}(d_i) = \frac{\sum_{A_j \in \mathcal{A}'} r_{A_j}(d_i) \cdot \frac{1}{DS(A_j)}}{\sum_{A_j \in \mathcal{A}'} \frac{1}{DS(A_j)}}, d_i \in D$$

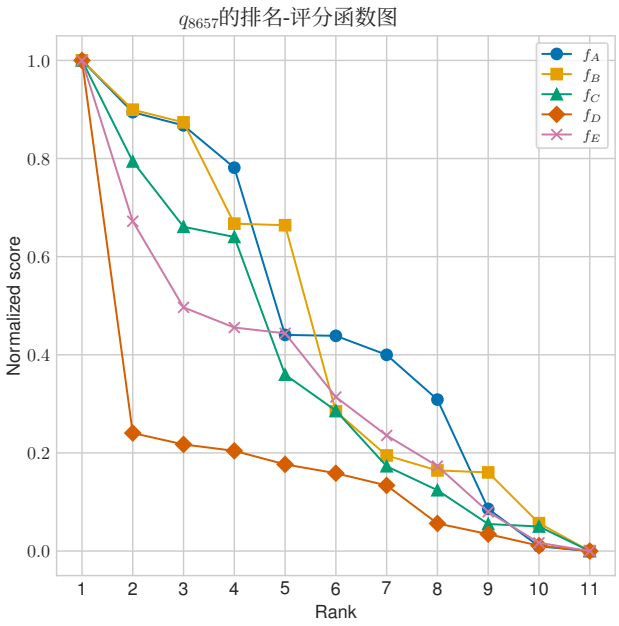

图 2：关于问题 q_{8657}:f_A 的排名-评分函数图，f_B，f_C，f_D 其中 f_E 分别代表智能体 A、B、C、D 和 E 在权威、关怀、公平、忠诚和神圣性方面的取值。

其中， $\mathcal{A}'$ 是完整评分系统集合 $\mathcal{A} = \{A_1, \cdots, A_5\}$ 的任意子集，且至少包含 2 个元素，即 $\mathcal{A}' \subseteq \mathcal{A}$，并且 $\|\mathcal{A}'\| = t$，其中 $t \in \{2, 3, 4, 5\}$, $s_{A_j}(d_i)$ 和 $r_{A_j}(d_i)$分别是评分系统 A_j 对数据项 d_i赋予的分数和排名； $DS(A_j)$ 表示评分系统 A_j的多样性强度。

CFA 已广泛应用于科学发现和决策等多个领域，包括近期在药物发现和材料科学方面取得的成果 [16, 17]。研究表明 [18]，随着认知多样性的增加，秩组合往往表现更好。

CFA架构包含连续的欧几里得空间和离散的秩空间，分别对应于

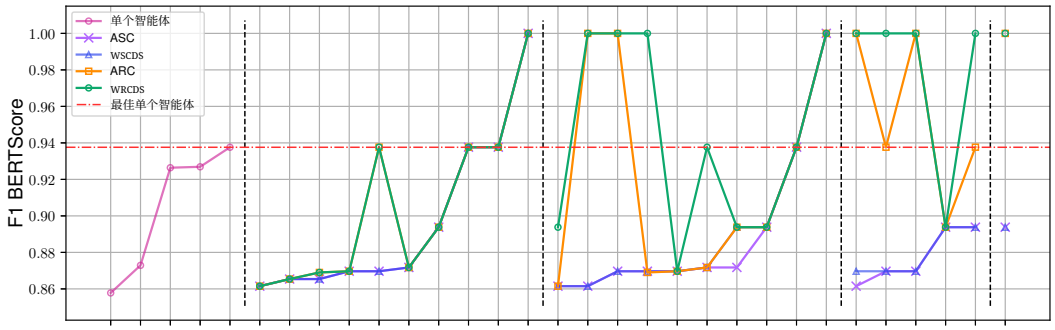

图3：问题 q_{8657} 在四种 CFA 组合类型（ASC、WSCDS、ARC、WRCDS）下 26 种组合的 F1 BERTScore（每个模型组内的 ASC 按非递减顺序排列）。

评分函数向量和排名函数向量。当评分函数为一一函数时，包含 n 个元素的数据集 D 上的排名函数是 n 个正整数 $[1,n]$ =(1，2，3，.....，n)的一个完整排列。在这种情况下，集合 $[1,n]$上的完整排列构成一个阶为 n 的对称群，其运算为复合运算（ [19]和 [20]上确界），记作 S_n。如果将每个排列视为图中的一个顶点（该图有 $n!$ 个顶点），并将两个顶点之间的邻接关系定义为交换一对相邻元素，则得到的图称为冒泡排序凯莱图 B_n。（该凯莱图由一组顶点生成，这组顶点由所有相邻元素对之间的对换组成。）图 B_n 是一个$(n-1)$-正则图，连通度为$n-1$，并且 B_n 可由 n 个 B_{n-1} [20,21]的副本递归构造。

当评分函数不是一一映射时，得出的排名会出现并列。在这种情况下，排名不是一个完整的排列。Kemeny 和 Snell 提出了一种包含并列排名的度量 [22]。采用该度量的 Kemeny 排名空间 K_n 的顶点数大于 $n!$（这是 B_n的顶点数）。与 K_n 相关的研究包括排名聚合、多层组合融合和优先级排名 [21,23,24]。

3. VAS-CFA工作流程

我们提出了一种将组合融合分析整合到多智能体聚合中的新框架。我们将其称为基于组合融合分析的价值对齐系统（VAS-CFA）。图 1 展示了所提出系统的主要步骤。在VAS-CFA工作流程的第一步中，我们以OpenAssistant/oasst-sft-4-pythia-12b-epoch-3.5 SFT 检查点为起点，使用带有 QLoRA [25]的直接偏好优化（直接偏好优化），在单个 NVIDIA A100-40GB 上对五个特定价值智能体进行微调——权威（A）、关怀（B）、公平（C）、忠诚（D）和神圣性（E）。道德诚信语料库（MIC） [26] 用于对各个智能体进行微调。MIC 数据集公开可用，包含 113.8K 个提示-回答对，这些回答经过人工修订并带有丰富的伦理标注。它采用固定的数据集划分，训练集/验证集/测试集分别包含 91.0K/11.4K/11.4K 个样本。每个道德智能体都以相同的基础模型为起点独立训练 1 个 epoch，使用 $\beta = 0.1$、学习率 1×10^{-5}、每设备批量大小 2 和 8 个梯度累积步数。QLoRA 以 4 位 NF4 格式加载 12B 基础模型，仅训练 LoRA适配器，从而实现实用的单 GPU 微调；我们选择直接偏好优化，是因为它无需奖励模型或在线强化学习即可优化人类偏好。与基于PPO的基于人类反馈的强化学习相比，这简化了流程，并提升了训练稳定性和样本效率。

对五个道德智能体分别进行微调，使其各自与一种不同的道德价值对齐后，对于每个测试提示，我们都会得到每个智能体生成的一个回答（图1(a)）。一种天真的做法是汇总这五个回答，采用 Dognin 等人的价值融合策略，利用上下文信息组合输出。然而，直接汇总可能导致语义冲突：不同智能体可能表达彼此不兼容的道德承诺，从而产生淡化或不连贯的回答，并削弱价值对齐。受此启发，我们使用 GPT-4.1 nano 将每个智能体的输出拆解为道德单元，其中每个单元表达一个道德主张（图1(b)）。这一设计也得到以下观察的进一步支持：经人工修订的标准答案通常简短，只包含一两个道德观点，而对话模型的输出则更长，且涉及多个主题。将回答拆分为单元的一个例子是：智能体 B（关怀）给出的回答是“应优先培养智力，以确保你的孩子健康、繁荣地成长”，该回答被拆解为三个单元：“应优先培养智力”“确保你的孩子健康成长很重要”以及“确保你的孩子繁荣成长很重要”。

随后，我们将从五个智能体回答中提取的所有道德单元汇集起来。为了评估每个单元与五种道德价值中每一种的对齐程度，我们训练了一个“道德分类器”。具体而言，我们使用SentenceTransformer（all-MiniLM-L6-v2）对人工修订答案进行编码，并拟合逻辑回归模型以进行多标签预测。给定一个道德单元，分类器会返回五个分数，每种道德价值对应一个分数，从而生成 CFA 步骤中使用的评分表（图1(c)）。将每一列视为一种特定价值的评分系统后，我们便得到了分别对应五个道德智能体的模型 A、B、C、D 和 E 的五套评分系统。

以五个评分系统A、B、C、D和E代表五个基础道德智能体，我们使用第2节所述的CFA生成$\sum_{i=1}^{5}\binom{5}{i} = 26$种组合。在26种组合的每一种中，以多样性强度为权重的CFA框架产生四种组合类型：（1）平均分数组合（ASC），（2）基于多样性强度的加权评分组合（WSCDS），（3）平均秩组合（ARC），以及（4）基于多样性强度的加权秩组合（WRCDS）。

接下来，我们将这26个候选者在四种CFA组合类型（1）-（4）之一中与人工修订答案进行比较，并在每种配置中保留单个最佳单元，从而每个提示得到四个单元。在本研究中，我们仅采用顶部单元，并将其传递给释义器；该释义器被提示在保留该单元道德内容的同时回答用户的问题（图1(d)）。这一步是必要的，因为道德单元编码的是道德主张，而不一定是完整答案。如果选择了多个单元，聚合模块则会将它们融合成一个连贯的答案。

4. 结果

本节报告 VAS-CFA框架的结果。

图 2 绘制了测试数据集中问题q_{8657}对应的排名-分数函数图，该数据集包含 11.38 万个提示-回答对。对于每个问题，五种特定价值评分系统会生成五条曲线，我们将这些曲线叠加，以直观呈现它们之间的差异。结果表明，在测试数据集的大多数问题中，这五个智能体都表现出显著的认知多样性。

图 3 中问题 q_{8657} 对应的 26 种组合结果表明，排名组合（ARC/WRCDS）始终优于分数组合（ASC/WSCDS），这是因为道德智能体之间存在认知多样性，如图 2 [18] 所示。此外，将多样性强度（DS）作为权重，有助于形成理想的个体智能体非线性组合。

接下来，我们使用 ROUGE-L指标评估 VAS-CFA。ROUGE-L 衡量系统输出与参考文本之间的最长公共子序列，在允许非连续匹配的同时捕捉句子级重叠；我们报告标准的 F_1 分数，该分数以人工修订答案为基准，综合衡量精确率和召回率。表1(a)汇总了三组模型的结果。第 (i) 组由五个基础模型组成。第 (ii) 组列出了四种组合（ASC、ARC、WSCDS、WRCDS）各自的结果。第 (iii) 组包括原始聚合、CVA-GS 和 CVA-GS-DYN [27]。

表 1：性能汇总（ROUGE-L 和 F_1BERTScore）。

模型	(a) F_1 ROUGE-L	(b) F_1 BERTScore
(i). 个体道德主体		
A	0.0925	0.8569
B	0.0821	0.8533
C	0.1249	0.8628
D	0.1376	0.8663
E	0.1343	0.8653
(ii). 使用 CFA 的融合方法		
VAS-CFA：ASC	0.1594	0.8831
VAS-CFA：ARC	0.1691	0.8847
VAS-CFA：WSCDS	0.1598	0.8832
VAS-CFA：WRCDS	**0.1692**	**0.8849**
(iii). 不使用 CFA 的融合方法		
原始聚合 *	0.1318	0.8654
CVA-GS [27]	0.1120	0.8728
CVA-GS-DYN [27]	0.1450	0.8754

注：智能体 A、B、C、D 和 E 分别代表权威、关怀、公平、忠诚和神圣性。 * 原始聚合 = 是对道德智能体给出的原始五个回答进行聚合。

最后，表1(b)总结了三组模型 (i)、(ii) 和 (iii) 的 F_1 BERTScore 结果，与表1(a)中 F1 ROUGE-L 分数的呈现方式相同。

在表1的两种情况下，即 (a) F_1 ROUGE-L 和 (b) F_1 BERTScore，VAS-CFA 的结果均优于单个道德智能体的结果，也超过了此前由 CVA-GS 和 CVA-GS-DYN 得出的多智能体结果。此外，由于智能体之间存在认知多样性 [18]，排名组合（ARC/WRCDS）优于分数组合（ASC/WSCDS）。

5. 结论

本文介绍了基于组合融合分析的价值对齐系统（VAS-CFA），这是一种通过实现多智能体聚合来推进价值对齐的框架。不同于依赖单一评估者或狭义定义的奖励信号的既有方法，VAS-CFA 实例化了多个道德智能体，每个智能体都与一种不同的规范性视角对齐，并通过组合融合分析（CFA）整合其输出。该设计明确利用认知多样性，同时采用基于排名和基于评分的融合，以减少冗余、解决冲突，并生成更加连贯且关注价值的回应。实验结果表明，VAS-CFA 具有稳健性，并且在标准评估指标上始终优于单智能体模型和现有聚合基线。这些发现凸显了多智能体融合作为一种强大机制的潜力：它能够体现多元价值观并提升对齐质量。

6. 参考文献

[1] Dario Amodei、Chris Olah、Jacob Steinhardt、Paul Christiano、John Schulman 和 Dan Man´e，《人工智能安全中的具体问题》，*arXiv* 预印本 *arXiv:1606.06565*，2016 年。[2] Yufei Wang、Wanjun Zhong、Liangyou Li、Fei Mi、Xingshan Zeng、Wenyong Huang、Lifeng Shang、Xin Jiang 和 Qun Liu，《使大语言模型与人类对齐：一项综述》，*arXiv* 预印本 *arXiv:2307.12966*，2023 年。[3] Long Ouyang、Jeffrey Wu、Xu Jiang、Diogo Almeida、Carroll Wainwright、Pamela Mishkin、Chong Zhang、Sandhini Agar- wal、Katarina Slama、Alex Ray 等，《通过人类反馈训练语言模型遵循指令》，《神经信息处理系统进展》，第 35 卷，第 27730–27744 页，2022 年。[4]Yuntao Bai、Saurav Kadavath、Sandipan Kundu、Amanda Askell、Jackson Kernion、Andy Jones、Anna Chen、Anna Goldie、Azalia Mirhoseini、Cameron McKinnon 等，《宪法式人工智能：通过人工智能反馈实现无害性》，*arXiv* 预印本 *arXiv:2212.08073*，2022 年。[5] Rafael Rafailov、Archit Sharma、Eric Mitchell、Christopher D Manning、Stefano Ermon 和 Chelsea Finn，《直接偏好优化：你的语言模型其实是一个奖励模型》，《神经信息处理系统进展》，第 36 卷，第 53728–53741 页，2023 年。[6] Aman Madaan、Niket Tandon、Prakhar Gupta、Skyler Hallinan、Luyu Gao、Sarah Wiegreffe、Uri Alon、Nouha Dziri、S hrimai Prabhumoye、Yiming Yang 等，《Self-Refine：通过自反馈进行迭代改进》，《神经信息处理系统进展》，第 36 卷，第 46534–46594 页，2023 年。[7] Noah Shinn、Federico Cassano、Ashwin Gopinath、Karthik Narasimhan 和 Shunyu Yao，《Reflexion：采用语言强化学习的语言智能体》，《神经信息处理系统进展》，第 36 卷，第 8634–8652 页，2023 年。[8] Tianshu Yu、Ting-En Lin、Yuchuan Wu、Min Yang、Fei Huang 和 Yongbin Li，《用于大语言模型对齐的多样化人工智能反馈》，《计算语言学协会汇刊》，第 13 卷，第 392–407 页，2025 年。[9] Adam Dahlgren Lindstr¨om、Leila Methnani、Lea Krause、Pet- ter Ericson、 Í˜nigo Mart´ınez de Rituerto de Troya、Dimitri Coelho Mollo 和 Roel Dobbe，《有帮助、无害、诚实？通过基于人类反馈的强化学习实现人工智能对齐与安全的社会技术限制：Ad lindstr¨om 等》，《伦理学与信息技术》，第 27 卷，28，2025 年 2 期，第 页。[10] Xianzhe Fan、Qing Xiao、Xuhui Zhou、Jiaxin Pei、Maarten Sap、Zhicong Lu 和 Hong Shen，《用户驱动的价值对齐：了解用户对人工智能伴侣中偏见和歧视性陈述的看法及应对策略》，载于*2025* 年计算机系统中的人为因素 *CHI* 会议论文集，2025 年，第 1–19 页。[11]Stevie Bergman、Nahema Marchal、John Mellor、Shakir Mohamed、Iason Gabriel 和 William Isaac，《STELA：以社区为中心的人工智能对齐规范引 elicitation 方法》，《科学报告》，第 14 卷，第 1 期，第 6616 页，2024 年。[12] D Frank Hsu、Yun-Sheng Chung 和 Bruce S Kristal，《组合融合分析：组合多个评分系统的方法与实践》，载于《生物信息学中的高级数据挖掘技术》

[13] D Frank Hsu、Bruce S Kristal和Christina Schweikert，“组合融合分析”，《*Computer*》，第57卷，第09期，第96–100页，2024年。[14] D Frank Hsu、Bruce S Kristal和Christina Schweikert，“秩分数特征（RSC）函数与认知多样性”，载于国际脑信息学会议。施普林格，2010年，第42–54页。[15] D Frank Hsu、Bruce S Kristal、Yuhan Hao和Christina Schweikert，“认知多样性：多个评分系统之间相异性的度量”，《互连网络期刊》，第19卷，第01期，第1940001页，2019年。[16] Yuanqing Tang、Zhi Li、Mansoor Ani Najeeb Nellikkal、Hamed Eramian、Emory M Chan、Alexander J Norquist、D Frank Hsu和Joshua Schrier，“利用模型融合提高高通量钙钛矿合成的数据与预测质量”，《化学信息与建模期刊》，第61卷，第4期，第1593–1602页，2021年。[17] Jinn-Moon Yang、Yen-Fu Chen、Tsai-Wei Shen、Bruce S Kristal和D Frank Hsu，“用于提高虚拟筛选富集效果的共识评分标准”，《化学信息与建模期刊》，第45卷，第4期，第1134–1146页，2005年。[18] D Frank Hsu和Isak Taksa，“信息检索中数据融合的秩与分数组合方法比较”，《信息检索》，第8卷，第3期，第449–480页，2005年。[19] Persi Diaconis，“概率与统计学中的群表示”，《讲义-专著系列》，第11卷，1988年。[20] N Jiang、M Quazi、C Schweikert、DF Hsu、T Oprea和Suman Sirimulla，“通过组合融合分析提升ADMET性质模型性能”，*ChemRxiv*，2023年。[21] X Zhong、L Hurley、S Siri mulla、C Schweikert和DF Hsu，“在广义置换秩空间上组合多个排名系统”，载于*2019*年*IEEE*第五届大数据智能与计算国际会议（*DATACOM*）。IEEE，2019年，第123–129页。[22] John G Kemeny和J Laurie Snell，“社会科学中的数学模型，偏好排序一章：一种公理化方法”，1962年。[23] Sina Akbari和Adolfo R Escobedo，“超越Kemeny秩聚合：一种面向含并列的稳健排名聚合的可参数化惩罚框架”，《*Omega*》，第119卷，102893，2023年，页。[24] Wade D Cook和Lawrence M Seiford，“优先级排序与共识形成”，《管理科学》，第24卷，第16期，第1721–1732页，1978年。[25] Tim Dettmers、Artidoro Pagnoni、Ari Holtzman和Luke Zettlemoyer，“QLoRA：量化大语言模型的高效微调”，《神经信息处理系统进展》，第36卷，第10088–10115页，2023年。[26] Caleb Ziems、Jane Yu、Yi-Chia Wang、Alon Halevy和Diyi Yang，“道德诚信语料库：伦理对话系统的基准”，载于计算语言学协会第*60*届年会论文集（第*1*卷：长文），2022年，第3755–3773页。[27] Pierre Dognin、Jesus Rios、Ronny Luss、Prasanna Sattigeri、Miao Liu、Inkit Padhi、Matthew Riemer、Manish Nagireddy、Kush Varshney和Djallel Bouneffouf，“情境价值对齐”，载于*ICASSP 2025-2025 IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2025年，第1–5页。

’第32–62页。2006年。