arXiv:2602.07186v1 [cs.MA] 6 Feb 2026

方差的价值：通过不确定性驱动的策略优化缓解多智能体系统中的辩论崩溃

Luoxi Tang [1] **Yuqiao Meng** [1] **Joseph Costa** [1] **Yingxue Zhang** [1] **Muchao Ye** [2] **Zhaohan Xi** [1]

摘要

多智能体辩论（MAD）系统通过迭代审议提升大语言模型的推理能力，但仍容易发生辩论崩溃，即智能体基于错误推理做出的最终决策受到影响的一类失败。现有方法缺乏检测或预防此类失败的原则性机制。为弥补这一不足，我们首先提出一种分层指标，用于量化三个层面的行为不确定性：智能体内（个体推理不确定性）、智能体间（交互不确定性）以及系统级（输出不确定性）。对多个基准测试的实证分析表明，我们提出的不确定性量化方法能够可靠地指示系统故障，这验证了将其用作系统故障诊断指标的有效性。随后，我们提出一种缓解策略，通过制定不确定性驱动的策略优化，在动态辩论环境中惩罚自相矛盾、同伴冲突和低置信度输出。实验表明，我们提出的不确定性驱动的缓解措施能够可靠地校准多智能体系统，持续提高决策准确率，同时减少系统分歧。

1. 引言

大语言模型（LLM）已进入协作智能时代，多智能体辩论（MAD）系统得到深入发展，通过迭代不断改进其推理能力（Irving 等人，2018；Du 等，2023）。不同于对特定实例错误高度敏感的单智能体生成，MAD 系统利用多样化视角进行交叉验证，建立了新的范式，

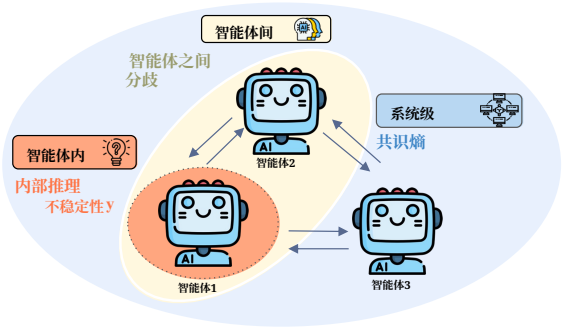


图 *1.* 三个层级的不确定性量化示意图：智能体内（单个大语言模型智能体推理过程中的不确定性）、智能体间（交互过程中一对智能体之间的不确定性）以及系统级（整个多智能体系统决策中的不确定性）。

从而在数学推理（Cobbe 等人，2021）、事实性问答（Lin 等人，2022）和常识生成（Talmor 等人，2019；Lazaridou & Baroni，2020）等广泛任务中展现出更高的有效性和鲁棒性。

然而，多智能体辩论并非最终解决方案。近期研究揭示了其容易发生辩论崩-溃的脆弱性。具体而言，研究人员发现，智能体会受到自信但已受损的少数派影响（Khan 等，2024），或陷入相互冲突的论点所形成的无休止循环，无法达成共识（Cemri 等人，2025；Foerster 等人，2017）。值得注意的是，对这些失败的行为分析仍未得到充分探索。现有方法通常将辩论崩溃视为结果层面的错误，而非动态的程序性失败（Du 等，2023；Perez 等人，2022），因而缺乏识别辩论何时以及如何开始偏离正轨所需的细粒度指标。当前的缓解策略通常依赖外部裁判智能体或僵化的交接式交互，却未能解决辩论智能体本身固有的不稳定性和鲁棒性局限（Chan 等，2023；Perez 等人，2022）。

这些局限表明存在两项研究空白：（1）缺乏成熟的框架来指示辩论崩溃；（2）缺少能够从根本上提升多智能体辩论鲁棒性的缓解措施，而非借助辅助组件（例如引入裁判智能体）暂时弥补失败。

本研究。为弥补这两方面的不足，本文从两个层面开展

[1]宾汉姆顿大学计算机科学系，美国纽约州宾汉姆顿 [2]爱荷华大学，美国爱荷华州爱荷华城。通信联系人：Zhaohan Xi <zxi1@binghamton.edu>。

论文集收录于 43^{rd} 国际机器学习会议，韩国首尔。PMLR 306，2026年。版权所有 © 2026，作者所有。

研究：

I. 辩论崩溃的不确定性驱动指标

从根本上说，人工智能系统的失败往往伴随着可观察到的行为指标（Kadavath 等，2022；Kuhn 等，2023）。正如困惑的人可能会犹豫、结巴或自相矛盾，难以完成任务的人工智能系统也常表现出输出不稳定、内部状态不一致的现象，尤其是在置信度与正确性不匹配时（Xiong 等，2023；Tomani 等，2024）。在多智能体场景中，这类现象会呈现出更复杂的维度。先前研究表明，基于智能体的审议可能遭遇协调失效、振荡行为，以及由信念不一致或不可靠的推理过程引发的失效模式（Du 等，2023；Cemri 等人，2025）。例如，频繁修改立场、偏离同伴模型或产生低置信度输出的智能体，可能会损害整个系统的鲁棒性。这些行为波动并非单纯的噪声，而是系统可靠性的内在信号。直观而言，对此类信号进行量化，有助于在辩论过程最终得出错误决策之前诊断其鲁棒性。

相应地，我们提出了一个分层框架（图 1），用于量化三个行为层面的不确定性：**（i）智能体内**，我们通过追踪单个智能体在辩论过程中改变立场的频率来衡量自一致性，这一做法基于已有证据：频繁的信念修正与不可靠的推理相关（Wang 等，2022）。**（ii）智能体间**，我们量化同伴分歧，因为智能体之间持续的冲突往往反映的是歧义或幻觉事实，而非富有成效的审议（Huang 等，2024；Cemri 等人，2025）。**（iii）系统级**，我们评估最终共识的熵，将其作为集体输出的整体置信度指标，并与既有的基于不确定性的诊断方法相一致（Kuhn 等，2023；Tomani 等人，2024）。

利用不确定性驱动的量化方法，我们在标准推理基准测试上进行了广泛的实证分析（Cobbe 等人，2021；Lin 等人，2022；Talmor 等人，2019）。实验表明，与正确预测相比，错误预测始终表现出显著更高的智能体内立场变化率、智能体间分歧以及系统级输出熵。这说明辩论崩溃并非随机现象；相反，它会留下独特的行为特征，从而将失败案例与成功的智能体审议区分开来。这些结果验证了我们提出的指标是评估多智能体辩论系统鲁棒性的可靠诊断工具。

II. 不确定性驱动的崩溃缓解

随后，我们使用不确定性衍生指标来指导我们的缓解方案设计。我们提出不确定性驱动策略优化（UDPO），将多智能体辩论建模为动态交互过程，并引入不确定性驱动奖励函数，对自我矛盾（内部稳定性）、同伴冲突（交互稳定性）和低置信度输出（系统稳定性）进行惩罚。为考虑智能体呈现不同的不确定性层级，我们进一步设计非对称优化目标，在分配智能体特定惩罚的同时纳入共享任务奖励，以确保解题正确性。

最后，我们证明，基于 UDPO 的缓解措施显著增强了多智能体辩论系统的韧性。实验结果表明，采用这一目标进行微调的多智能体辩论，在崩溃场景中恢复时的准确率比基线多智能体微调方法高出 25% 以上（Yu等人，2022；He et al.，2023）。即使多智能体辩论中存在显式攻击（Huang 等，2024）（即智能体遭到攻陷），我们的方法也能有效减少由幻觉引发的不确定性，并提高多智能体辩论的准确率，这表明它为实现稳健的多智能体智能提供了一种可行的解决方案。

总而言之，本文作出了三项贡献：(1) 我们提出了用于分析多智能体辩论行为的层次化不确定性指标，以诊断辩论崩溃。(2) 我们通过实证和统计分析证明，辩论崩溃与这些不确定性指标呈正相关。(3) 我们提出了一种不确定性驱动的策略优化（UDPO）策略，以缓解辩论崩溃。通过广泛评估，我们表明 UDPO 能有效减少辩论崩溃，并在多智能体共识过程中抵御攻击。代码已发布于：https: // anonymous.4open.science/r/UDPO-5BE2/。

2. 多智能体辩论中的不确定性量化

本节详细介绍如何将不确定性驱动的量化作为辩论崩溃的指标。

设定。我们研究多智能体辩论（MAD）系统，其中一组智能体 $\{\mathcal{A}_i\}_{i=1}^{N}$ 经过 T 轮审议来回答问题 q。我们遵循 Du 等人（2023）和 Chan等人（2023）中的代表性设计：智能体迭代交换论点、相互批评对方的回答，并在多轮辩论中完善推理。

我们提出从三个互补层面量化不确定性：（i）智能体内——每个智能体在辩论过程中改变自身想法的程度；（ii）智能体间——智能体彼此意见不一致的程度；（iii）系统级——对最终预测的确信程度。所有量均针对每个问题 q 定义，之后再在数据集上进行汇总。

智能体内不确定性。令 $a_t^{(i)} \in \mathcal{Y}$ 为.....的答案

智能体 i 在第 t 轮之后的答案，其中 $t=0$ 表示辩论开始前的初始回答。我们衡量智能体在审议过程中改变立场的频率：

$$F(q)=\frac{1}{N(T-1)}\sum_{i=1}^{N}\sum_{t=1}^{T-1}\mathbf{1}\big[a_t^{(i)}\neq a_{t+1}^{(i)}\big]$$

我们将其称为立场翻转率。我们还会衡量智能体最终的答案是否与初始答案不同：

$$M(q)=\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\big[a_0^{(i)}\neq a_T^{(i)}\big]$$

即信念修正率。智能体内不确定性为$U_{\text{intra}}(q)=\lambda F(q)+(1-\lambda)M(q)$，其中 $\lambda\in[0,1]$。较大的U_{intra}意味着智能体经常修正其答案，表明其在这个问题上的推理不稳定。

智能体间不确定性。在每一轮 t 中，我们通过成对冲突来衡量智能体之间意见分歧的程度：

$$C_t(q)=\frac{2}{N(N-1)}\sum_{1\le i<j\le N}\mathbf{1}\big[a_t^{(i)}\neq a_t^{(j)}\big]$$

并对各轮次进行汇总：$U_{\text{inter}}(q)=\frac{1}{T+1}\sum_{t=0}^{T}C_t(q)$。较高的 U_{inter} 意味着智能体在辩论过程中表现出波动，表明该问题在决策方面存在较大分歧。

系统级不确定性。最后，我们通过汇总三个互补信号来量化系统级不确定性：(1) 归一化熵 $H(q)$，用于衡量所有轮次中的输出多样性；(2) 最终分歧 $D(q)\in[0,1]$，用于反映终止时缺乏共识的程度；以及 (3) 留一法不稳定性 $L(q)$，用于捕捉系统对单个智能体的敏感性。系统级不确定性定义为
as:

$$U_{\text{sys}}(q)=\frac{1}{3}\Big[\tilde{H}(q)+D(q)+L(q)\Big]$$

这三个组成部分的形式化定义均见附录D。

推理依据。我们的直觉是，正确预测通常在这些层面上都表现出较低的不确定性（智能体保持稳定、彼此意见一致，并有把握地达成一致），而错误预测往往至少在一个组成部分上表现出较高的不确定性。接下来，我们将在多种推理任务中通过实证验证这一直觉。

3. 不确定性量化评估

本节旨在通过实证评估不确定性量化的实用性。

3.1. 评估设置

MAD系统。我们采用一种具有代表性的多智能体辩论（MAD）设置，遵循 Du 等人（2023）和 Chan 等人（2023）的方法，该设置已用于许多现有的 MAD 系统（Liang 等人，2024；Khan 等，2024；Wu 等，2024）。在这一设置中，智能体采用扁平的角色结构：它们首先独立回答问题，然后在多个讨论轮次中反复观察、批评并完善彼此的回答，最后通过多数投票作出决定。

数据集。我们在三个不同领域中被广泛使用的基准测试上进行评估：(i) GSM8K（Cobbe 等人，2021）：多步数学推理；(ii) TruthfulQA（Lin 等人，2022）：关于误解的事实问答；(iii) CommonsenseQA（Talmor 等人，2019）：常识知识。

模型。我们使用大语言模型实现多智能体（$N=5$）、多轮次（$T=5$）辩论：Llama-4-Scout、Qwen-3-8B、Gemma-3n-E4B、Mixtral-8x7B 和 Phi-4。此处我们选择开源大语言模型，以便后续微调。我们还在§5中评估了不同数量的智能体组（$N=3$ 或 10）。

辩论崩溃设置。我们考虑两种条件：**（1）自然崩溃（无攻击）：**我们直接评估自然的MAD系统。**（2）被攻陷的智能体（有攻击）：**遵循 AUTOTRANSFORM 攻击（Huang 等，2024），其中一个智能体被攻陷，并故意引入隐蔽错误，以巧妙地误导其他智能体。

3.2. 评估结果

本节详细介绍我们的评估结果。由于篇幅有限，补充结果见附录 B。

不确定性分布。图2比较了推理失败与成功时的不确定性得分。在所有不确定性类型中，失败情况的不确定性始终高于成功情况。例如，失败时的系统级均值不确定性（U_{sys}）为 0.67，而成功时为 0.09。假设检验所得的较小 p值（例如 $p<0.001$）和较大的效应量（Cohen's $d>0.8$）（Cohen，2013）进一步支持了这一差异，表明失败与成功推理的不确定性分布之间存在统计显著且具有实际意义的差异。

相关性分析。图3展示了相关矩阵。三种不确定性度量均与准确率呈显著负相关（$r<0$, $p<0.001$），表明较高的不确定性可以可靠地预测较低的准确率。有趣的是，这些不确定性度量彼此之间呈显著正相关（r 接近 1），但其分布存在显著差异（$p<0.001$）。这表明它们提供了一致的整体信号，同时捕捉了不确定性的互补方面。

不确定性指标在筛选对抗攻击方面的效用。我们还考虑一种涉及攻击的多智能体辩论设置，其中部分智能体可能已遭到攻陷。遵循

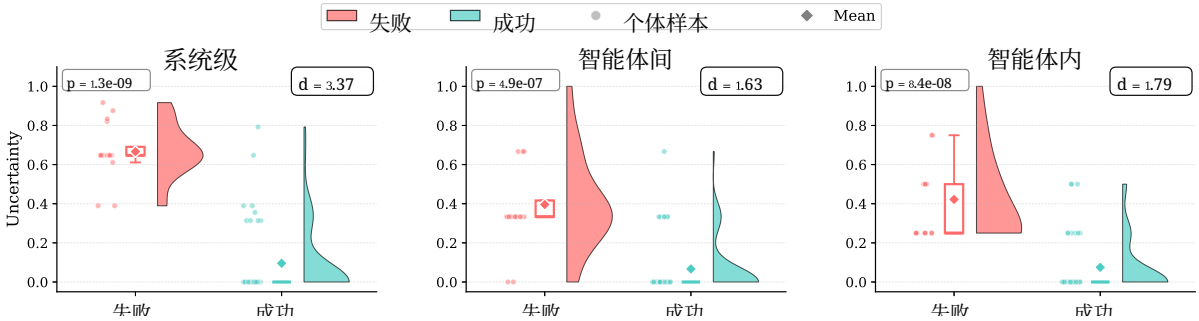


图2. GSM8K上失败与成功推理的不确定性分布。我们进行双样本 t检验，零假设（H_0）为两个分布相同。较小的 p值（例如 $p < 0.05$）表明拒绝 H_0的证据更强，说明两个分布在统计上存在差异。我们还计算Cohen's d（Cohen，2013），以量化两个分布均值之间的标准化差异（$d > 0.8$ 表示差异较大）。其他数据集的结果见附录 B.1。

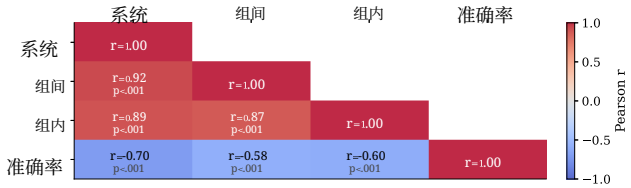


图3. 基于皮尔逊相关系数 r（维基百科贡献者，2026）的相关矩阵（TruthfulQA）。我们还进行与图2相同的双样本 t 检验。所有不确定性类型均与准确率呈负相关（$r < 0$, $p < 0.001$）。其他数据集的结果见附录 B.2。

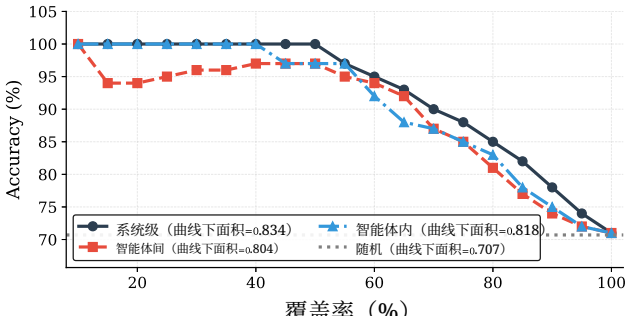


图 4. TruthfulQA 上的不确定性驱动的数据选择。更多结果请参阅附录 B.3。

AUTOTRANSFORM（Huang 等，2024），我们注入隐蔽的推理错误，这些错误表面上看似合理，却会导致错误结论。随后，我们进行不确定性驱动的数据选择分析，以评估不确定性指标能否筛选对抗性推理并减轻攻击影响。具体而言，我们根据每种不确定性度量（U_{intra}、U_{inter} 和 U_{sys}）对所有推理轨迹进行排序，并保留不确定性最低的前 k%。图 4 展示了准确率随 k 变化的情况。我们观察到一个明显趋势：低不确定性子集（例如 $k = 50\%$）的准确率显著更高（最高可达 100%），这表明不确定性指标能够区分正常推理与经对抗性扰动的推理。随着 k 增大，纳入更多高不确定性（且往往失败的）样本后，准确率急剧下降。所有数据集均呈现相同趋势（图 8）。这些结果证实，不确定性是筛选可靠推理的有效标准，也为我们提出的不确定性驱动的回放机制提供了依据。该机制会在训练期间优先处理高不确定性的推理轨迹，以减少失败（附录 A.3）。

由于篇幅限制，我们将不确定性分布和相关性的进一步分析留至附录 B。

摘要。 我们的实验揭示了不确定性对 MAD 系统的作用。**首先**，三种不确定性度量（U_{intra}、U_{inter} 和 U_{sys}）均与系统故障呈正相关，这证实了不确定性在多个粒度上都能提供有意义的信号。

其次，尽管这些度量之间高度相关，但它们的分布存在显著差异，反映了多智能体审议中的不同失效模式：智能体可能会犯下孤立的错误，这些错误随后会在交互中得到纠正（或被放大）；而一个自信的智能体仍可能受到其他参与审议的智能体影响。这凸显了采用多维视角的必要性，即考虑全部三个不确定性层级，而不是将它们合并为单一分数。**第**三，不确定性指标在对抗性扰动下仍然有效，这表明无论失败是自然发生还是由攻击诱发，它们都能捕捉协作推理质量的稳健信号。

洞见。 我们的三级不确定性量化（$U_{\text{intra}}, U_{\text{inter}}, U_{\text{sys}}$）可作为**MAD系统故障的可靠指标**，为后续缓解方案设计提供可行的信号。

4. 不确定性驱动的缓解

本节利用不确定性指标作为辩论崩溃的指示信号（§3），设计了一套缓解框架，以提升

系统鲁棒性。

4.1. 概述

鉴于辩论崩溃发生于智能体交互过程中，我们在动态辩论环境中制定缓解措施，在该环境中，智能体反复交换批评意见并修订其响应。在此过程中，我们从三个层面量化不确定性： U_{intra} 刻画智能体推理轨迹中的不稳定性， U_{inter} 衡量智能体之间的分歧， U_{sys} 反映对最终输出缺乏置信度。我们进一步设计了一个**面向智能体的非对称策略优化目标**，以应对智能体不同的不确定性水平。最小化该目标将推动辩论动态趋向稳健共识。

4.2. 不确定性驱动的奖励设计

鉴于多智能体辩论呈现于动态的交互式环境中，我们设计了一种由不确定性指标（§2）驱动的强化式奖励。我们将每个智能体视为一个策略，在多轮辩论中生成响应。由此产生的多智能体辩论轨迹，$\tau = (a_0^{(1)}、\ldots、a_0^{(N)})\ldots\ldots、(a_T^{(1)}、\ldots\ldots、a_T^{(N)})$，由 N个智能体经过 T 轮次构成，是定义奖励函数的基本单位。

我们首先定义一种鼓励智能体内稳定性的奖励。直观而言，除非有力证据促使智能体修正其信念，否则它在整个辩论过程中都应保持立场一致：

$$r_{\text{intra}}(\tau) = 1 - \frac{1}{N(T-1)} \sum_{i=1}^{N} \sum_{t=0}^{T-1} \mathbf{1}[a_t^{(i)} \neq a_{t+1}^{(i)}] \quad (1)$$

当智能体表现出稳定的推理轨迹时，该奖励达到最大值。仅有稳定性并不足够：我们将任务监督纳入训练目标，确保只有当稳定性与正确性一致时，才会强化这种稳定性。

接下来，我们定义一种鼓励智能体间达成一致的奖励：

$$r_{\text{inter}}(\tau) = 1 - \frac{1}{T+1} \sum_{t=0}^{T} \frac{2}{N(N-1)} \sum_{i<j} \mathbf{1}[a_t^{(i)} \neq a_t^{(j)}] \quad (2)$$

这一表述捕捉了整个辩论轨迹中的一致性，并通过奖励早期收敛而非后期共识来体现这一点。

最后，我们定义一种系统级奖励，将熵、分歧和留一不稳定性纳入其中。令$p(y \mid \tau)$表示由最终轮回答构成的答案经验分布：

$$r_{\text{sys}}(\tau) = 1 - \frac{1}{3}\big[\tilde{H}(\tau) + D(\tau) + L(\tau)\big] \quad (3)$$

~与公式 2 类似， $H(\tau)$是归一化熵， $D(\tau) \in \{0,1\}$ 表示残余分歧，而 $L(\tau)$衡量留一不稳定性，即缺少某个智能体会导致最终决策改变的智能体比例。该奖励倾向于奖励能够抵御个别智能体扰动的高置信度预测。请注意，各组成部分的正式定义见附录D。

4.3. 不确定性驱动策略优化的非对称目标（UDPO）

我们方法的一个关键洞见是，**并非所有智能体对辩论崩溃的影响都相同**。有些智能体表现出更高的不确定性，或更容易受到对抗性影响，而另一些则发挥稳定作用。因此，我们采用一种**非对称优化** 策略，根据不同智能体各自的不确定性分布，对其施加不同强度的训练。

给定一条辩论轨迹 τ，智能体 i 的总奖励为：

$$r_{\text{UDPO}}^{(i)}(\tau) = \underbrace{\alpha^{(i)} r_{\text{intra}} + \beta^{(i)} r_{\text{inter}} + \gamma^{(i)} r_{\text{sys}}}_{\text{uncertainty-driven reward}} + \underbrace{\lambda^{(i)} r_{\text{task}}}_{\text{task reward}} \quad (4)$$

其中， $r_{\text{task}}(\tau) = \mathbf{1}[\hat{y}(\tau) = y^*]$ 表示最终多数投票预测是否与真实值相符。至关重要的是，系数 $\alpha^{(i)}$、 $\beta^{(i)}$、 $\gamma^{(i)}$ 和 $\lambda^{(i)}$ 是**针对智能体量身定制**的：历史不确定性较高的智能体会获得更强的不确定性驱动奖励，以促进其稳定；而已经稳定的智能体则会获得相对较高的任务奖励。这种非对称加权既能避免对表现良好的智能体施加过度正则化，又能将优化精力集中在最需要的地方。

裁剪相对更新目标。为了在这一奖励下优化智能体策略，我们采用裁剪相对更新目标，通过限制策略变化的幅度并将更新锚定于参考策略来稳定学习。令 $\pi_\theta^{(i)}$ 表示智能体 i 的当前策略， $\pi_{\theta_{\text{ref}}}^{(i)}$ 表示固定的参考策略（例如滞后快照或基础模型）。我们定义轨迹级似然比：

$$\rho_\theta^{(i)}(\tau) = \exp\big(\log \pi_\theta^{(i)}(\tau) - \log \pi_{\theta_{\text{ref}}}^{(i)}(\tau)\big) \quad (5)$$

智能体 i 的优化目标为：

$$\mathcal{L}_{\text{UDPO}}^{(i)}(\theta) = \mathbb{E}_\tau\big[\ell_{\text{clip}}^{(i)}\big] - \eta^{(i)} \mathbb{E}_\tau\big[\ell_{\text{anchor}}^{(i)}\big] \quad (6)$$

其中，期望是对 $\tau \sim \pi_{\theta_{\text{ref}}}$求取的， ϵ 控制最大相对更新幅度，而 $\eta^{(i)}$ 是针对智能体定制的锚定权重。容易出现不稳定性的智能体会获得更强的锚定（$\eta^{(i)}$ 更大），以防止更新出现异常。

表1. **不同数据集、多智能体辩论系统和基线比较** 在自然崩溃（无攻击）设置下的主要结果。我们报告准确率（%）和不确定性指标。最佳结果以**粗体**文本标出。 U_{in}、 U_{ir}、 U_{s} 分别表示 U_{intra}、 U_{inter}、 U_{sys} 。不同智能体选择或不同辩论轮次数量（T）下的更多结果见 C.1 和 C.2。

数据集	N 智能体 在多智能体辩论中	标准 MAD				MAPPO				RMAAC				UDPO（我们的方法）			
		Acc	U_{in}	U_{ir}	U_{s}	Acc	U_{in}	U_{ir}	U_{s}	Acc	U_{in}	U_{ir}	U_{s}	Acc	U_{in}	U_{ir}	U_{s}
GSM8K	N=3	51.2	.231	.268	.372	64.8	.218	.185	.278	66.3	.205	.172	.258	**84.6**	**.068**	**.052**	**.078**
	N=5	68.4	.228	.172	.285	73.6	.185	.145	.232	75.8	.168	.132	.215	**92.3**	**.065**	**.038**	**.058**
	N=10	65.7	.235	.188	.298	70.2	.192	.158	.248	72.1	.175	.145	.228	**89.8**	**.071**	**.055**	**.072**
TruthfulQA	N=3	62.4	.225	.178	.275	68.5	.192	.152	.238	70.2	.185	.148	.225	**85.2**	**.072**	**.058**	**.082**
	N=5	71.8	.218	.142	.235	74.2	.158	.122	.198	76.5	.152	.115	.188	**88.7**	**.068**	**.048**	**.068**
	N=10	73.5	.222	.128	.218	76.8	.165	.108	.182	78.2	.158	.102	.172	**91.4**	**.075**	**.035**	**.055**
CSQA	N=3	68.2	.205	.162	.258	72.5	.172	.138	.222	74.1	.168	.132	.212	**86.8**	**.062**	**.048**	**.072**
	N=5	75.4	.198	.135	.225	78.2	.148	.112	.188	79.8	.142	.108	.178	**91.5**	**.058**	**.032**	**.052**
	N=10	73.8	.202	.148	.238	76.5	.155	.125	.202	77.2	.148	.118	.192	**88.2**	**.064**	**.045**	**.065**

裁剪代理目标和锚定项如下：

$$\ell_{\text{clip}}^{(i)} = \min\big(\rho^{(i)}\hat{A}^{(i)},\ \text{clip}(\rho^{(i)}, 1-\epsilon, 1+\epsilon)\hat{A}^{(i)}\big) \quad (7)$$

$$\ell_{\text{anchor}}^{(i)} = \text{KL}\big(\pi_{\theta}^{(i)}(\cdot|\tau)\,\|\,\pi_{\theta_{\text{ref}}}^{(i)}(\cdot|\tau)\big) \quad (8)$$

其中，我们将 $\rho^{(i)} = \rho_{\theta}^{(i)}(\tau)$ 和 $\hat{}A^{(i)} = \hat{}A^{(i)}(\tau)$ 简写为相应形式。

优势估计根据智能体特定奖励计算：

$$\hat{A}^{(i)}(\tau) = r^{(i)}(\tau) - b \quad (9)$$

其中， b是用于降低方差的批次基线。

针对智能体的超参数选择。系数 $\alpha^{(i)}$、 $\beta^{(i)}$、 $\gamma^{(i)}$、 $\lambda^{(i)}$ 和 $\eta^{(i)}$ 根据各智能体在预热阶段测得的经验不确定性概况进行选择。具体而言，我们计算每个智能体在其推理轨迹中的平均 U_{intra}、U_{inter} 以及对 U_{sys} 的贡献：不确定性较高的智能体会获得相应更大的系数。这种针对智能体的校准确保非对称优化能够适应辩论集成中不同大语言模型的异质行为。参数选择过程的详细信息见附录 A.2。

5. 缓解措施评估

本节通过四个研究问题评估基于 UDPO 的缓解措施：（研究问题 1）UDPO 是否能提升整体性能？（研究问题 2）各损失分量分别有何贡献？（研究问题 3）我们的方法在遭受攻击时能否保持有效？（研究问题 4）我们的方法在何种情况下最有帮助？

5.1. 实验设置

数据集和模型。 我们使用与第 3 节相同的数据集和大语言模型。

基线方法。 我们与三种方法进行比较。标准 MAD 是未经训练的普通多智能体辩论（Du 等人，2023；Liang 等，2024）。MAPPO（Yu等人，2022）将近端策略优化扩展到采用集中式价值函数的合作多智能体场景；我们将其适配到多智能体辩论中，把每个智能体的答案选择视为一种策略，并针对共识进行优化。RMAAC（He等人，2023）是一种通过对抗性扰动对状态不确定性进行建模的鲁棒多智能体演员-评论家方法；我们通过将答案分布视为状态、并针对扰动训练智能体来对其进行适配。

5.2. 主要结果（RQ1）

观察结果。 表1比较了所有方法在各数据集–模型组合上的表现。UDPO始终取得最佳的整体性能。例如，相较于标准 MAD，UDPO在 N=5 设置下的提升最高可达25个百分点（GSM8K上为68.4% $\rightarrow$ 92.3%）。与MAPPO等具有代表性的多智能体强化学习基线相比，UDPO的准确率最高提升20个百分点（GSM8K上为73.6% $\rightarrow$ 92.3%）。UDPO还显著降低了不确定性：在GSM8K上，与标准 MAD相比，其不确定性降低了约80%。这与我们的实证发现一致：不确定性是辩论崩溃的可靠信号，明确对其进行控制能够提升多智能体辩论的性能。

解释。 这些提升主要源于各训练目标明确优化的内容。**MAPPO**本质上是采用集中式价值函数 （评论家）的 PPO，该函数在训练期间可以获取全局信息，而每个智能体在执行时仅依据自身的局部观测（即集中训练、分散执行）（Amato, 2024年；Wen et al., 2021）。这种集中式评论家有助于智能体在共享奖励下进行协调，但不会直接编码行为稳定性（例如，避免在不同轮次间反复改变立场，或解决持续分歧）。因此，智能体仍可能达成脆弱的共识，甚至对错误的推理达成共识。另一方面，**RMAAC** 从不同角度提升鲁棒性：它将多智能体强化学习建模为马尔可夫博弈，并通过均衡学习，学习能够抵御最坏情况（可能具有

表 2. **消融**　**研究**通过移除各个奖励分量进行。红色标出[illegible]消融中受影响最大的指标　　n.

方法	GSM8K				TruthfulQA				CSQA			
	Acc	U_{intra}	U_{inter}	U_{sys}	Acc	U_{intra}	U_{inter}	U_{sys}	Acc	U_{intra}	U_{inter}	U_{sys}
完整 UDPO	**92.3**	**.065**	**.038**	**.058**	**88.7**	**.068**	**.048**	**.068**	**91.5**	**.058**	**.032**	**.052**
不含 $\mathcal{L}_{\text{intra}}$	86.5	.142	.052	.085	83.2	.138	.062	.092	85.8	.145	.048	.082
不含 $\mathcal{L}_{\text{inter}}$	88.2	.078	.098	.095	85.4	.082	.105	.098	87.2	.072	.095	.088
不含 $\mathcal{L}_{\text{sys}}$	84.8	.082	.058	.128	81.5	.085	.068	.135	83.6	.078	.055	.125
标准 MAD	68.4	.228	.172	.285	71.8	.218	.142	.235	75.4	.198	.135	.225

表 3. **存在多个受损智能体时的攻击结果**。 我们使用 AUTOTRANSFORM（Huang 等，2024）向 $N{=}5$ 辩论中的 $m \in \{1,2,3\}$ 智能体注入隐蔽的语义错误。我们报告准确率（%）和不确定性指标。每一行中的最佳结果（固定数据集和 m）以**粗体**标出。U_{in}、U_{ir}、U_{s} 分别表示 U_{intra}、U_{inter} 和 U_{sys}。

数据集	# 受损	标准 MAD				MAPPO				RMAAC				UDPO（本文方法）			
		Acc	U_{in}	U_{ir}	U_{s}	Acc	U_{in}	U_{ir}	U_{s}	Acc	U_{in}	U_{ir}	U_{s}	Acc	U_{in}	U_{ir}	U_{s}
GSM8K	1	54.9	.286	.255	.522	62.7	.303	.202	.338	65.4	.232	.188	.315	**79.6**	**.112**	**.084**	**.125**
	2	41.8	.412	.408	.548	50.6	.318	.275	.468	54.2	.295	.255	.432	**68.9**	**.172**	**.138**	**.198**
	3	9.4	.662	.525	.614	36.7	.402	.355	.592	40.5	.372	.332	.548	**56.2**	**.258**	**.212**	**.288**
TruthfulQA	1	58.7	.247	.212	.362	63.5	.225	.178	.305	66.1	.212	.165	.285	**82.4**	**.108**	**.076**	**.118**
	2	45.9	.385	.298	.434	52.4	.298	.245	.418	56.3	.275	.228	.451	**71.8**	**.165**	**.118**	**.182**
	3	23.8	.593	.582	.576	51.2	.372	.315	.528	53.1	.342	.295	.470	**70.7**	**.242**	**.185**	**.255**
CSQA	1	63.2	.310	.205	.338	58.0	.215	.172	.288	69.5	.205	.162	.275	**85.1**	**.098**	**.072**	**.108**
	2	51.6	.335	.275	.442	56.9	.275	.228	.378	59.8	.258	.215	.355	**76.4**	**.148**	**.112**	**.162**
	3	29.7	.512	.552	.573	44.8	.342	.295	.472	48.6	.318	.278	.438	**73.9**	**.215**	**.168**	**.235**

对抗性）的状态扰动的策略。尽管 RMAAC 强调对受损观测的鲁棒性，但它并非旨在从细粒度的智能体行为层面诊断哪些辩论行为出现问题（自相矛盾、同伴冲突还是低置信度聚合）。

相比之下，**UDPO** 将辩论崩溃直接分解为三种可解释的失效模式，并分别对其施加惩罚：智能体内反复摇摆（$\mathcal{L}_{\text{intra}}$）、智能体间未解决的分歧（$\mathcal{L}_{\text{inter}}$）以及低置信度最终结果（$\mathcal{L}_{\text{sys}}$）。UDPO 通过直接针对不确定性指标进行优化，更可靠地推动辩论达成稳定、可验证的一致意见（即使存在攻击），而非依赖集中式信用分配（MAPPO）或最坏情况扰动鲁棒性（RMAAC）。

5.3. 消融研究（RQ2）

观察结果。 表 2 展示了消融结果。移除任何一个组件都会降低准确率并增加不确定性。其影响具有针对性：移除 $\mathcal{L}_{\text{intra}}$ 会导致 U_{intra} 大约翻倍，而对其他指标的影响较小；$\mathcal{L}_{\text{inter}}$和 $\mathcal{L}_{\text{sys}}$也有类似情况。在这三个组件中，移除 $\mathcal{L}_{\text{sys}}$ 造成的准确率下降最大，其次是 $\mathcal{L}_{\text{intra}}$ 和 $\mathcal{L}_{\text{inter}}$。这一排序在所有三个模型中均一致。

启示。 每个损失分量都针对预期的失效模式，验证了我们框架设计的合理性。排名（$\mathcal{L}_{\text{sys}} > \mathcal{L}_{\text{intra}} > \mathcal{L}_{\text{inter}}$）表明，系统级稳定性对最终准确率最为重要（Guo 等，2017），因为它汇总了所有智能体和轮次的信号，捕捉局部指标无法体现的全局模式。然而，正如我们在 §3 中所展示的，这三者缺一不可，因为它们代表了多智能体辩论的不同方面。完整模型的表现显著优于任何消融模型，这表明这些失效模式彼此互补，而非冗余。

5.4. 对攻击的鲁棒性（RQ3）

观察结果。与压力测试一致，表 3 展示了受损智能体数量不同时施加攻击的结果。当仅有一个智能体受到攻击时，标准的多智能体辩论已经出现明显性能下降，而 UDPO 仍保持较高的准确率，且不确定性仅小幅上升，这表明它能够吸收局部篡改，而不会破坏辩论的稳定性。当受损智能体增至两个和三个时，所有方法的表现均有所下降，但趋势各不相同。具体而言，基线方法的 U_{s} 和 U_{ir} 急剧上升，表明攻击会愈发从局部不一致传播为群体级分歧，并最终导致系统级犹豫不决。相比之下，UDPO 的性能下降更为平缓，即使攻击率较高，仍能保持更大的性能优势，这表明其稳定性控制能够推迟受损内容主导集体结果的临界点。

解释。 关键机制并不只是 UDPO 提高了准确率，而是它改变了辩论在对抗压力下失效的方式。在标准 MAD（以及以协调为核心的基线方法）中，遭受攻击的智能体可以注入

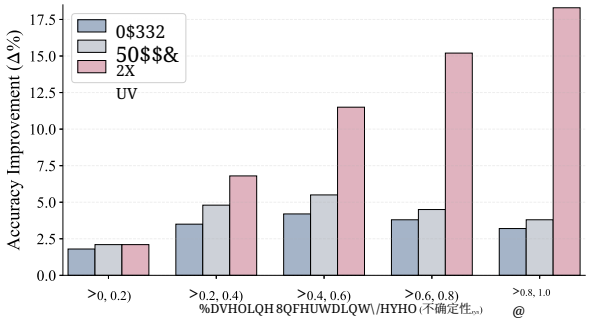


图 *5*。 **准确率提升（Δ%），相较于标准 MAD，按不确定性水平划分。** 结果为跨数据集和模型的平均值。随着不确定性升高，我们的方法取得的收益不断增加，而 MAPPO 和 RMAAC 则趋于停滞或性能下降。基于 U_{intra} 和 U_{inter} 的难度分层见附录C.4。

貌似合理但存在细微缺陷的论据，其他智能体可能会对此过度拟合，从而形成脆弱共识或持续性冲突。随后，系统会在错误一致与未解决的分歧之间摇摆，表现为 U_{in} 和U_{ir} 不断上升，并最终导致 U_{s}居高不下。UDPO 通过惩罚自相矛盾、未解决的同伴冲突，以及阻碍共识的低置信度聚合，明确地对这些趋势进行正则化。要达成共识，群体必须减少不稳定性信号，而非仅仅在表面答案上达成一致。即使面对遭到攻击的轨迹，恶意内容如今也必须既能说服他人，又能经受交叉质询而保持稳定，这要困难得多。因此，即使受损智能体占据主导比例（超过半数），UDPO 仍能实现更强的鲁棒性，并延缓系统崩溃。

附录 C.3 报告了 UDPO 在攻击设置下缓解效果的统计分析（类似于 §3）。

5.5. 分析：UDPO 在何种情况下最有帮助？（研究问题 4）

为评估缓解措施的敏感性，我们根据标准 MAD 下计算的基线系统级不确定性 U_{sys}，将问题划分为五个难度层级。这样，我们便能确定哪些不确定性（难度）层级最能从我们的方法中获益。

观察结果。 图 5 显示，在五个不确定性层级中，我们的方法相较于标准 MAD 均提升了准确率，且呈现明显的单调趋势：增益从不确定性水平最低时的 +2.1%（$U_{\text{sys}} < 0.2$）上升至最高时的 +18.3%（$U_{\text{sys}} \geq 0.8$）。相比之下，MAPPO 的增益较小，在中等不确定性时达到峰值 +4.2%，在难题上则降至 +3.2%。RMAAC 的表现趋势相似，峰值为 +5.5%，随后降至 +3.8%。随着难度增加，这两种基线方法都无法维持效果。

洞见。 这些不同的模式揭示了各方法处理不确定性的根本差异。MAPPO 的集中式评论家有助于在智能体之间存在中等分歧时实现协调，但无法阻止难题中出现的级联式不稳定性（Wang 等，2024；Estornell 和 Liu，2024）。RMAAC 的对抗鲁棒性在扰动有界时有所帮助，但当辩论动态变得高度混乱时便难以应对。我们的方法明确惩罚反复摇摆、分歧和低置信度，而这些恰恰是难题中的主导行为。不断扩大的收益证实，将不确定性分解为可解释的组成部分，比将其视为协调失败（MAPPO）或对抗噪声（RMAAC）更有效。这与已有发现一致：推理一致性对复杂问题至关重要（Wang 等，2022）。

6. 相关工作

多智能体辩论。 多智能体辩论（MAD）通过利用多样化的观点和迭代改进，已成为提升大语言模型推理能力的有效方法（Du 等，2023；Liang 等，2024；Chan 等，2023）。近期研究探索了多种辩论架构，包括结构化论证（Khan 等，2024）、基于角色的讨论（Wu 等，2024）以及评审介导的审议（Koutcheme 等，2024）。尽管这些方法展现出优于单智能体基线方法的准确率，但它们主要聚焦于架构创新，缺乏检测或预防辩论崩溃的原则性机制。我们的工作通过提供用于诊断系统可靠性的不确定性指标，以及能够明确缓解失效模式的训练目标，对这一研究方向作出了补充。

大语言模型中的不确定性量化。 针对大语言模型的不确定性估计已从多个角度得到研究，包括词元级熵（Kadavath 等，2022）、语义等价聚类（Kuhn 等，2023）和自一致性度量（Wang 等，2022）。近期研究还探讨了置信度校准（Tian 等，2023）和语言化不确定性（Xiong 等，2023）。然而，这些方法聚焦于单智能体场景，无法捕捉多智能体系统中丰富的行为动态。我们的分层框架将不确定性量化扩展至多智能体辩论（MAD），将系统不确定性分解为组内、组间和系统级组件，每个组件都对应不同的失效特征。

7. 结论

本文提出了一种面向多智能体辩论系统的分层不确定性量化框架，用于捕捉智能体内、智能体间和系统层面的行为不稳定性。我们发现，在多个推理基准测试中，这些不确定性信号始终能够区分成功的辩论与失败的辩论。在此分析基础上，我们设计了 UDPO——一种不确定性驱动的策略

优化框架，其采用非对称训练目标，旨在促进稳定推理、智能体一致以及高置信度输出。实证结果表明，该方法显著提高了准确率，同时大幅降低了行为波动性，从而构建出更稳健的多智能体系统，使其更加可靠且校准得更好。

影响声明

本文通过引入不确定性驱动的诊断方法和优化目标，推动了多智能体学习的方法学发展，旨在提升多智能体辩论系统的鲁棒性。所提出的框架旨在支持可靠且可解释的协作推理研究，而非用于在现实决策场景中的直接部署。与许多通用机器学习方法一样，本文研究的技术可以适用于广泛的应用，其社会影响将取决于最终的使用方式。我们鼓励未来的研究在实际场景中应用多智能体学习系统时，进一步探讨负责任的部署实践和人工监督。

参考文献

Amato, C. 关于合作多智能体强化学习中集中式训练与分散式执行的介绍。*arXiv* 预印本 *arXiv:2409.03052*，2024年。

Cemri, M.、Pan, M. Z.、Yang, S.、Agrawal, L. A.、Chopra, B.、Tiwari, R.、Keutzer, K.、Parameswaran, A.、Klein, D.、Ramchandran, K. 等。多智能体大语言模型系统为何会失败？*arXiv* 预印本 *arXiv:2503.13657*，2025年。

Chan, C.-M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., Fu, J., 和 Liu, Z. ChatEval：通过多智能体辩论改进基于大语言模型的评估器。*arXiv* 预印本 *arXiv:2308.07201*，2023年。

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., 等。训练验证器以解决数学文字题。*arXiv* 预印本 *arXiv:2110.14168*，2021年。

Cohen, J. 《行为科学统计功效分析》。劳特利奇出版社，2013年。

Du, Y., Li, S., Torralba, A., Tenenbaum, J.B., 和 Mordatch, I. 通过多智能体辩论提升语言模型的事实准确性和推理能力。载于第*41*届国际机器学习会议，2023年。

Estornell, A. and Liu, Y. 多大语言模型辩论：框架、原则与干预措施。神经信息处理系统进展, 37:28938–28964, 2024年。

Foerster, J. N., Chen, R. Y., Al-Shedivat, M., Whiteson, S., Abbeel, P., and Mordatch, I. 通过对手学习意识进行学习。*arXiv* 预印本 *arXiv:1709.04326*，2017年。Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. 关于现代神经网络的校准。发表于国际机器学习会议，第 1321–1330 页。PMLR，2017年。

He, S., Han, S., Su, S., Han, S., Zou, S., and Miao, F. 状态不确定性下的鲁棒多智能体强化学习。*arXiv* 预印本 *arXiv:2307.16212*，2023年。

Huang, J.-t., Zhou, J., Jin, T., Zhou, X., Chen, Z., Wang, W., Yuan, Y., Lyu, M. R., and Sap, M. 故障智能体下基于大语言模型的多智能体协作的韧性。*arXiv* 预印本 *arXiv:2408.00989*，2024年。

Irving, G., Christiano, P., and Amodei, D. 通过辩论实现人工智能安全。*arXiv* 预印本 *arXiv:1805.00899*，2018年。

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., et al. 语言模型（大多）知道自己知道什么。*arXiv* 预印本 *arXiv:2207.05221*，2022年。

Khan, A., Hughes, J., Valentine, D., Ruis, L., Sachan, K., Radhakrishnan, A., Grefenstette, E., Bowman, S. R., Rockt¨aschel, T. 和 Perez, E. 与更具说服力的大语言模型辩论会带来更真实的答案。*arXiv* 预印本 *arXiv:2402.06782*，2024年。

Koutcheme, C., Dainese, N., Sarsa, S., Hellas, A., Leinonen, J. 和 Denny, P. 开源语言模型能够提供反馈：评估大语言模型通过 GPT-4 作为评判者帮助学生的能力。载于*2024*年计算机科学教育创新与技术会议论文集第*1*卷，第52–58页。2024年。

Kuhn, L., Gal, Y. 和 Farquhar, S. 语义不确定性：用于自然语言生成中不确定性估计的语言不变性。*arXiv* 预印本 *arXiv:2302.09664*，2023年。

Lazaridou, A. 和 Baroni, M. 深度学习时代涌现的多智能体通信。(2020). *URLhttps://arxiv. org/abs/2006.02419*,2020.

Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Shi, S., and Tu, Z. 通过多智能体辩论鼓励大语言模型进行发散性思维。载于*2024*年自然语言处理经验方法会议论文集，第17889–17904页，2024年。

Lin, S., Hilton, J., and Evans, O. TruthfulQA：衡量模型如何模仿人类谬误。载于计算语言学协会第*60*届年会论文集（第*1*卷：长文），第3214–3252页，2022年。

Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N.和Irving, G. 使用语言模型对语言模型进行红队测试。*arXiv*预印本 *arXiv:2202.03286*，2022年。

Talmor, A., Herzig, J., Lourie, N.和Berant, J. CommonsenseQA：一项针对常识知识的问答挑战。载于*2019*年北美计算语言学协会分会会议论文集：人类语言技术，第*1*卷（长文与短文），第4149–4158页，2019年。

Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., Finn, C.和Manning, C. D. 只需请求校准：从经人类反馈微调的语言模型中引出校准后的置信度分数的策略。*arXiv*预印本 *arXiv:2305.14975*，2023年。Tomani, C., Chaudhuri, K., Evtimov, I., Cremers, D.和Ibrahim, M. 大语言模型中的基于不确定性的弃答可提升安全性并减少幻觉。*arXiv*预印本 *arXiv:2404.10960*，2024年。Wang, Q., Wang, Z., Su, Y., Tong, H.和Song, Y. 重新思考大语言模型推理的边界：多智能体讨论是关键吗？*arXiv*预印本 *arXiv:2402.18272*，2024年。Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A.和Zhou, D. 自一致性提升语言模型中的思维链推理。*arXiv*预印本 *arXiv:2203.11171*，2022年。Wen, G., Fu, J., Dai, P.和Zhou, J. DTDE：一种新的合作式多智能体强化学习框架。*TheInnovation*，2(4)，2021年。维基百科贡献者。皮尔逊相关系数——维基百科，自由的百科全书。https://en.wikipedia.org/wiki/ Pearson_correlation_coefficient，2026年。Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J.等。AutoGen：通过多智能体对话赋能下一代大语言模型应用。载于第一届语言建模会议，2024年。Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J.和Hooi, B. 大语言模型能否表达其不确定性？对大语言模型中置信度 elicitation的实证评估。*arXiv*预印本 *arXiv:2306.13063*，2023年。

Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Bayen, A. 和 Wu, Y. 合作式多智能体博弈中近端策略优化的惊人效果。神经信息处理系统进展，35:24611–24624，2022。

A. 训练细节

本附录详细介绍训练流程、超参数选择以及不确定性驱动的回放机制。

A.1. 实验设置

我们使用 AdamW 优化器对所有模型进行微调，共训练 10 个 epoch，学习率为 1×10^{-5}。训练使用 8 块 NVIDIA A100 GPU，每次更新的批次大小为 32 条轨迹。前 500 步采用线性预热调度，随后进行余弦衰减。梯度裁剪的最大范数设为 1.0。

A.2. 非对称训练与智能体定制超参数选择

如第 4.3 节所述，公式 (4) 中的奖励系数是**针对智能体定制的**：不同智能体会根据各自的不确定性分布获得不同的系数值。这种非对称设计考虑到异构大语言模型集成中的智能体具有不同的稳定性特征。我们根据每个智能体的不确定性水平校准训练强度。

两阶段校准流程。

阶段 *1*：不确定性分析。 在训练前，我们让每个智能体在训练数据的留出子集上（占轨迹的 10%）经历一个预热阶段。对于每个智能体 i，我们计算：

- 智能体内平均不确定性： $\bar{U}_{(i)\,\text{组内}}$
- 对智能体间不确定性的平均贡献： $\bar{U}_{(i)\,\text{组间}}$
- 对留一不稳定性的平均贡献： $\bar{L}_{(i)} = \frac{1}{|\mathcal{D}_{\text{预热阶段}}|} \sum_{\tau} \mathbf{1}[\hat{y}(\tau) \neq \hat{y}(\tau^{-i})]$

阶段 *2*：系数校准。 我们根据不确定性分布设置针对智能体定制的系数，使其与不确定性分布成正比：

$$\alpha^{(i)} = \alpha_{\text{base}} \cdot \left(1 + \kappa \cdot \bar{U}^{(i)}_{\text{intra}}\right) \tag{10}$$

$$\beta^{(i)} = \beta_{\text{base}} \cdot \left(1 + \kappa \cdot \bar{U}^{(i)}_{\text{inter}}\right) \tag{11}$$

$$\gamma^{(i)} = \gamma_{\text{base}} \cdot \left(1 + \kappa \cdot \bar{L}^{(i)}\right) \tag{12}$$

$$\lambda^{(i)} = \lambda_{\text{base}} / \left(1 + \kappa \cdot \bar{U}^{(i)}_{\text{sys}}\right) \tag{13}$$

$$\eta^{(i)} = \eta_{\text{base}} \cdot \left(1 + \kappa \cdot \bar{U}^{(i)}_{\text{sys}}\right) \tag{14}$$

其中， $\kappa > 0$ 是控制不对称程度的缩放因子。不确定性较高的智能体会获得更强的均衡奖励（ $\alpha^{(i)}$、 $\beta^{(i)}$、 $\gamma^{(i)}$更大）、更弱的任务奖励（ $\lambda^{(i)}$更小）以及更强的锚定（ $\eta^{(i)}$更大）。

基础超参数。 默认使用 $\kappa = 1.5$ 。基础系数 α_{base}、 β_{base}、 γ_{base} 根据上述 Cohen 的 d归一化方法设置，其中 $\lambda_{\text{base}} = 1$ 和 $\eta_{\text{base}} = 0.01$。

设计依据。 这种不对称设计有两个目的：(1) 防止对已经稳定的智能体进行过度正则化，以免损害其任务表现；(2) 将训练投入集中在对辩论崩溃贡献最大的智能体上，同时提高样本效率。

A.3. 不确定性驱动的重放

第 3 节中的选择性预测分析表明，高不确定性轨迹对应着本质上更难的问题。我们不会在推理时丢弃这类样本，而是在训练过程中利用这一信号，更频繁地重放高不确定性轨迹。

我们维护一个回放缓冲区 $\mathcal{B}$，用于存储近期的辩论轨迹。每个训练轮次结束后，我们会计算一个综合不确定性分数：

$$U(\tau) = \alpha \cdot (1 - r_{\text{intra}}(\tau)) + \beta \cdot (1 - r_{\text{inter}}(\tau)) + \gamma \cdot (1 - r_{\text{sys}}(\tau)), \tag{15}$$

并以概率 $p(\tau) \propto U(\tau)^{\eta}$采样轨迹，其中 $\eta \geq 0$控制优先级强度（默认为 $\eta = 1$）。

我们会定期使用更新后的策略重新评估已存储的问题，以刷新回放缓冲区，并应用重要性采样校正，确保优化过程无偏。

A.4. UDPO算法

下面详细介绍UDPO算法:

算法 1 不确定性驱动的策略优化（UDPO）

需要： 智能体数量 N；辩论轮次 T；奖励权重（$\alpha, \beta, \gamma, \lambda$）；裁剪参数 ϵ；参考正则化系数 η；训练迭代次数 K。

初始化智能体策略参数 θ
初始化参考策略参数：$\theta_{\text{ref}} \leftarrow \theta$
对于 $k = 1$ **到** K **执行**
  采样一批任务 $\{x_m\}_{m=1}^{M}$
  对于所有 x_m **do**
    使用策略 π_θ 进行多智能体辩论，共进行 T 轮次
    收集轨迹：
      $\tau_m = \{(a_0^{(1)} \ldots, a^{(N)\,0}) \ldots, (a_T^{(1)} \ldots, a_T^{(N)})\}$
    计算基于不确定性的奖励：
      $r_{\text{intra}}(\tau_m)$ $r_{\text{inter}}(\tau_m)$ $r_{\text{sys}}(\tau_m)$
    计算任务奖励 $r_{\text{task}}(\tau_m)$
    根据公式 (4) 计算总奖励 $r(\tau_m)$
  结束循环
  计算基线 $b \leftarrow \frac{1}{M} \sum_{m=1}^{M} r(\tau_m)$
  计算优势 $\hat{A}(\tau_m) \leftarrow r(\tau_m) - b$
  计算似然比：
    $\rho_\theta(\tau_m) \leftarrow \text{指数}\Big(\text{对数}\ \pi_\theta(\tau_m) - \text{对数}\ \pi_{\theta_{\text{ref}}}(\tau_m)\Big)$
  通过最大化 $\mathcal{L}_{\text{UDPO}}(\theta)$（公式 6）来更新策略参数 θ
  定期更新参考策略：$\theta_{\text{ref}} \leftarrow \theta$
结束循环
返回优化后的策略参数 θ

B. 不确定性量化的补充结果

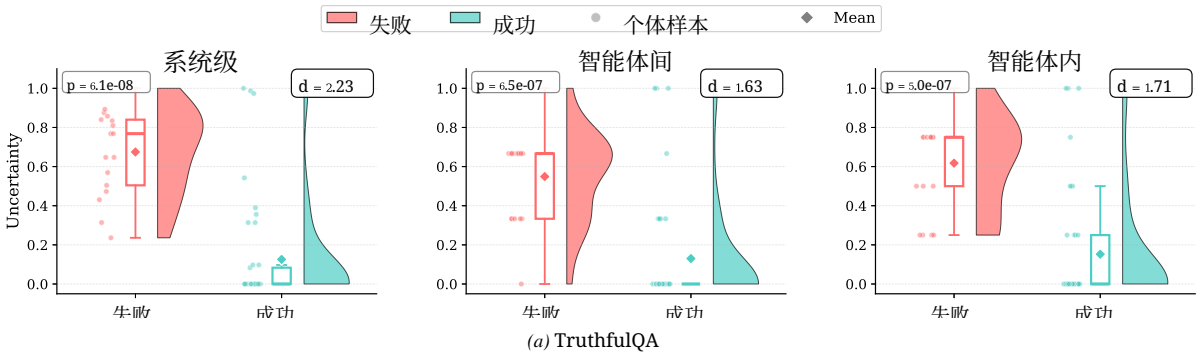
本附录展示了有关不确定性量化的补充实验结果，以补充第 3 节中的主要发现。这些结果侧重于验证不确定性指标作为辩论崩溃诊断工具的有效性，评估对象为**未采用缓解措施的自然 MAD 系统**。

B.1. 不确定性分布

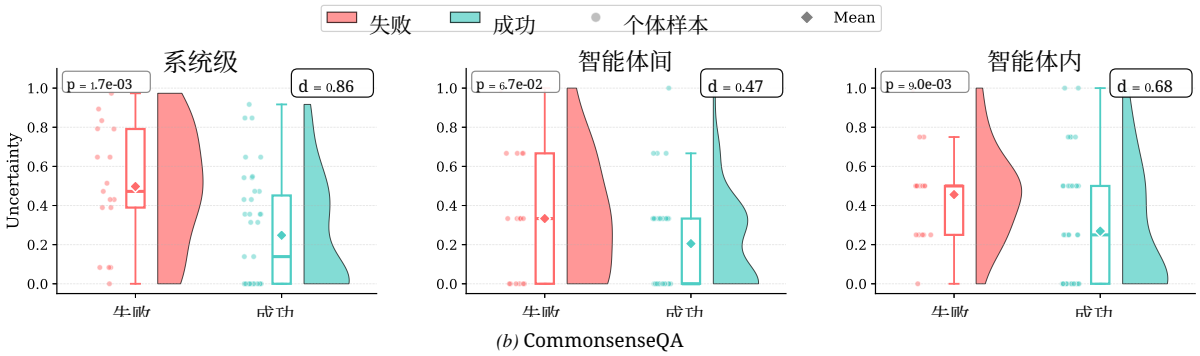
图6将不确定性分布分析（正文中的图2）扩展至 TruthfulQA 和 CommonsenseQA。与 GSM8K 的结果一致，在所有三项指标上，错误预测的不确定性都显著更高。效应量（Cohen's d）仍然很大，证实我们的不确定性量化能够在不同任务领域中可靠地区分推理失败与成功。

B.2. 相关性分析

图7展示了GSM8K和CommonsenseQA的相关矩阵，补充了正文中的TruthfulQA结果。在不同数据集中，所有不确定性类型均与准确率显著负相关，同时彼此之间呈正相关。这一模式验证了我们的三级不确定性框架能够捕捉互补且一致的系统可靠性信号。

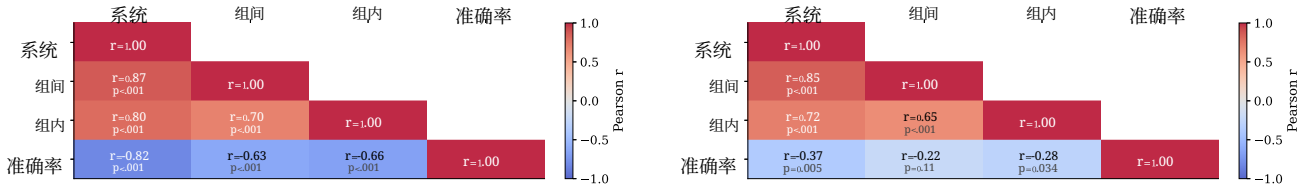


(a) TruthfulQA



(b) CommonsenseQA

图*6*。 **不确定性分布（额外数据集）**。 补充正文中的图2。与 GSM8K 一致，在所有指标上，错误预测都表现出更高的不确定性。



(a) GSM8K *(b)* CommonsenseQA

图*7*。 **相关矩阵（额外数据集）**。 补充正文中的图3。所有不确定性类型均与准确率呈负相关。

B.3. 选择性预测

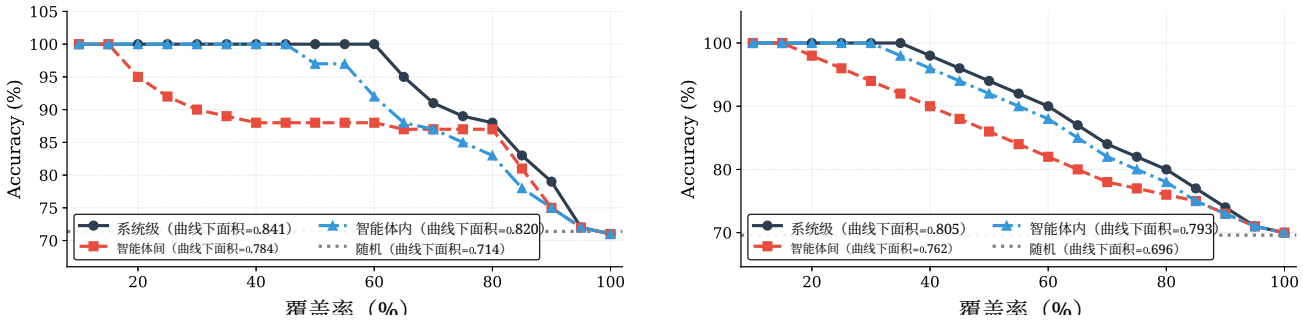
图 8展示了不确定性驱动的数据选择在GSM8K和CommonsenseQA上的效用。结果与TruthfulQA的结果相呼应：仅保留低不确定性的预测可显著提高准确率，证实我们的不确定性指标能够有效筛除不可靠的推理轨迹。

C. 缓解的补充结果

本附录展示了基于 UDPO 的缓解措施的补充实验结果，包括对抗鲁棒性分析以及按难度分层的性能细分。

C.1. 同质模型集成

正文中的表1采用了来自不同大语言模型系列的异质集成。此处我们评估同质集成，其中所有 $N = 5$智能体都使用相同的基础模型。这样可以将模型容量的影响与集成多样性区分开来。



(a) GSM8K *(b)* CommonsenseQA

图 *8.* **选择性预测（额外数据集）**。补充正文中的图 4。不确定性引导的弃权始终能提升准确率。

表 *4.* **同质 MAD 结果**（由 $N = 5$ 个相同模型类型的智能体组成）。我们报告准确率（%）和不确定性指标。与表1中的异质集成相比，同质集成的基线性能略低，但 UDPO 带来的相对提升相近。

数据集	模型 (5 个智能体)	标准 MAD				MAPPO				RMAAC				UDPO（我们的方法）			
		Acc	U_{in}	U_{ir}	U_{s}	Acc	U_{in}	U_{ir}	U_{s}	Acc	U_{in}	U_{ir}	U_{s}	Acc	U_{in}	U_{ir}	U_{s}
GSM8K	Llama-4	62.5	.242	.185	.298	68.2	.198	.158	.248	70.4	.182	.145	.228	**88.6**	**.072**	**.045**	**.068**
	Qwen-3	65.8	.235	.178	.288	71.5	.188	.148	.235	73.2	.172	.138	.218	**90.2**	**.068**	**.042**	**.062**
	Gemma-3n	58.2	.258	.198	.318	64.5	.215	.168	.265	66.8	.198	.155	.245	**85.4**	**.078**	**.052**	**.075**
	Mixtral	64.2	.238	.182	.295	70.8	.192	.152	.242	72.5	.178	.142	.225	**89.5**	**.070**	**.044**	**.065**
	Phi-4	60.5	.248	.192	.308	66.8	.205	.162	.255	68.5	.188	.148	.235	**86.8**	**.075**	**.048**	**.072**
TruthfulQA	Llama-4	66.2	.228	.155	.248	70.5	.185	.132	.212	72.8	.168	.125	.198	**86.5**	**.075**	**.052**	**.075**
	Qwen-3	69.5	.218	.145	.238	73.8	.172	.122	.198	75.5	.158	.115	.185	**88.2**	**.070**	**.048**	**.070**
	Gemma-3n	62.8	.245	.172	.272	67.2	.198	.148	.232	69.5	.182	.138	.215	**83.8**	**.082**	**.058**	**.082**
	Mixtral	68.2	.222	.148	.242	72.5	.178	.125	.205	74.2	.162	.118	.192	**87.5**	**.072**	**.050**	**.072**
	Phi-4	64.5	.235	.162	.258	68.8	.192	.138	.218	70.8	.175	.128	.202	**85.2**	**.078**	**.055**	**.078**
CSQA	Llama-4	70.2	.212	.148	.238	74.5	.168	.125	.202	76.2	.155	.118	.188	**88.5**	**.065**	**.042**	**.062**
	Qwen-3	73.5	.202	.138	.225	77.2	.158	.115	.188	78.8	.145	.108	.175	**90.2**	**.060**	**.038**	**.058**
	Gemma-3n	66.8	.228	.165	.258	71.2	.185	.138	.218	73.5	.168	.128	.202	**85.8**	**.072**	**.048**	**.070**
	Mixtral	72.2	.208	.142	.232	76.5	.162	.118	.195	78.2	.148	.112	.182	**89.5**	**.062**	**.040**	**.060**
	Phi-4	68.5	.218	.155	.245	72.8	.175	.128	.208	74.5	.162	.122	.195	**87.2**	**.068**	**.045**	**.065**

观察结果。同质集成的基线准确率通常低于异质集成（表1），因为它们无法受益于结合不同模型架构所带来的多样性。不过，UDPO 在所有模型类型上都带来了稳定的提升，准确率比标准 MAD 高出 18–24 个百分点。值得注意的是，Qwen-3 在所有数据集上的表现均为最佳，而 Gemma-3n 从 UDPO 中获得的相对提升最大，这表明我们的方法对基线不确定性较高的模型尤其有效。

C.2. 辩论轮次的影响

主要实验采用 $T = 5$ 个辩论轮次。这里，我们通过测试异质集成 ($N = 5$) 在 $T \in \{3, 5, 10\}$ 下的表现，来评估对这一选择的敏感性。

观察结果。 增加辩论轮次能够提升所有方法的表现，但基线方法的收益会逐渐递减。标准 MAD 每增加 5 轮，准确率约提升 2–3%；而 UDPO 提升 1.5–2%。重要的是，无论 T 如何变化，UDPO 始终比标准 MAD 高出 16–20 个百分点。不确定性指标随着轮次增加而下降，表明延长审议时间能让智能体更加可靠地达成一致。不过，计算成本会随 T 线性增长，因此 $T = 5$ 是兼顾性能与效率的实用默认选择。</output>

表5. **辩论轮次的影响**在使用 $N = 5$ 个智能体时对性能的影响。增加轮次通常能提升基线方法的表现，但收益会逐渐递减；UDPO 在所有设置下均保持显著优势。

数据集	轮次	标准 MAD				MAPPO				RMAAC				UDPO（我们的方法）			
	T	Acc	U_{in}	U_{ir}	U_{s}	Acc	U_{in}	U_{ir}	U_{s}	Acc	U_{in}	U_{ir}	U_{s}	Acc	U_{in}	U_{ir}	U_{s}
GSM8K	T=3	64.2	.248	.195	.312	69.5	.205	.165	.258	71.2	.188	.152	.238	**89.5**	**.072**	**.048**	**.068**
	T=5	68.4	.228	.172	.285	73.6	.185	.145	.232	75.8	.168	.132	.215	**92.3**	**.065**	**.038**	**.058**
	T=10	70.5	.215	.158	.268	75.2	.172	.132	.218	77.5	.158	.122	.202	**93.8**	**.058**	**.032**	**.052**
TruthfulQA	T=3	67.5	.238	.165	.262	70.8	.195	.142	.225	72.5	.178	.132	.208	**85.8**	**.078**	**.058**	**.078**
	T=5	71.8	.218	.142	.235	74.2	.158	.122	.198	76.5	.152	.115	.188	**88.7**	**.068**	**.048**	**.068**
	T=10	73.2	.202	.128	.218	76.5	.145	.108	.182	78.2	.138	.102	.172	**90.5**	**.062**	**.042**	**.060**
CSQA	T=3	71.2	.218	.158	.252	74.5	.175	.132	.212	76.2	.162	.125	.198	**88.2**	**.068**	**.042**	**.062**
	T=5	75.4	.198	.135	.225	78.2	.148	.112	.188	79.8	.142	.108	.178	**91.5**	**.058**	**.032**	**.052**
	T=10	77.2	.182	.122	.208	80.5	.135	.098	.172	81.8	.128	.095	.165	**93.2**	**.052**	**.028**	**.045**

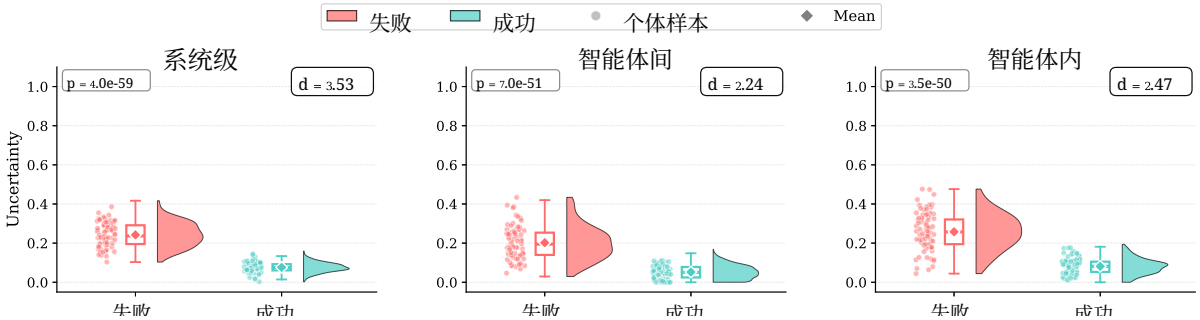


图 9. **对抗攻击下的不确定性分布（GSM8K）**。即使只有一个恶意智能体，错误预测在所有指标上的不确定性也显著更高。效应量（Cohen's d）仍然很大。TruthfulQA 和 CommonsenseQA 上的结果呈现出类似模式。

C.3. 对抗鲁棒性：攻击下的缓解措施

本节评估了是否 我们的不确定性指标和 UDPO 缓解措施在一个智能体具有恶意时是否仍然有效 s.

C.3.1. 攻击设置

遵循 Huang 等人（2024）的做法，我们使用 AUTOTRANSFORM 将一名辩手转换为恶意智能体，使其引入隐蔽的语义错误——这些回复表面上逻辑合理，实则得出错误结论。

C.3.2. 不确定性仍能区分正确与错误

在遭受攻击时，错误预测在所有三项指标上的不确定性仍显著高于正确预测。图 9 展示了 GSM8K 上的这一模式，效应量仍然很大（Cohen's $d = 2.24$–3.53）。系统级不确定性（$d = 3.53$）的区分效果最为显著，因为它汇总了来自组内和组间的信号。我们在 TruthfulQA 和 CommonsenseQA 上也观察到了类似模式，这证实即使有一个智能体积极破坏辩论，我们的不确定性分解仍保有诊断能力。

C.3.3. 相关性依然很强

图 10 展示了 GSM8K 在攻击下不确定性指标与准确率之间的相关结构。三种不确定性类型与准确率均保持较强的负相关（$r = -0.77$ 至 -0.89，均为 $p < 0.001$）。不确定性指标之间也呈正相关（$r = 0.57$–0.83），反映出它们对辩论不稳定性的敏感性相似。表 6 总结了这些相关性，并将其与不同数据集在自然设定下的相关性进行了比较。

这一解释很直接：当恶意智能体引入错误时，其他智能体会遇到与自身推理相冲突的论点，却无法轻易识别不一致的来源。这种困惑表现为不确定性升高，而我们的指标能够将其捕捉到。不确定性在遭受攻击时仍能持续预测失败，这一事实表明，我们的

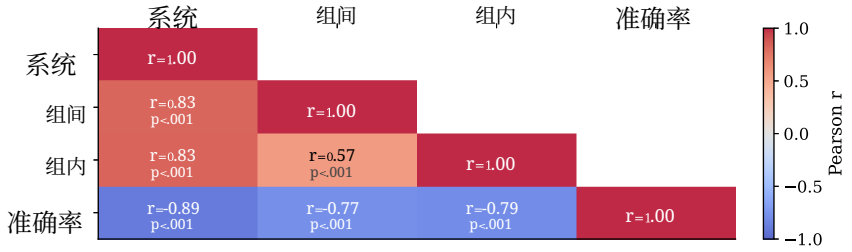


图*10.* **对抗攻击下的相关矩阵（GSM8K）**。不确定性指标与准确率仍保持强负相关，且彼此之间呈正相关。该模式与自然设定一致，表明对抗性扰动并未破坏不确定性-准确率关系。

表*6.* 攻击下与准确率的相关性（通义千问）。“自然”展示无攻击时的相关性，以供比较。

指标	攻击	自然	p 值
GSM8K			
U_{intra}	-0.79	-0.70	< 0.001
U_{inter}	-0.77	-0.68	< 0.001
U_{sys}	-0.89	-0.82	< 0.001
TruthfulQA			
U_{intra}	-0.72	-0.66	< 0.001
U_{inter}	-0.75	-0.71	< 0.001
U_{sys}	-0.84	-0.79	< 0.001
CommonsenseQA			
U_{intra}	-0.68	-0.62	< 0.001
U_{inter}	-0.71	-0.65	< 0.001
U_{sys}	-0.81	-0.76	< 0.001

表 *7.* 攻击下的性能（通义千问）。“Drop” =相较于自然设置的变化。

方法	GSM8K	Drop	TQA	Drop
标准 MAD	35.2%	-34.6pp	28.4%	-44.1个百分点
MAPPO	41.7%	-32.5个百分点	36.1%	-39.3个百分点
RMAAC	52.3%	-23.8个百分点	44.8%	-31.4个百分点
UDPO	77.2%	-5.2个百分点	66.4%	-14.8个百分点

框架测量 衡量的是辩论质量的某种根本属性，而非自然崩溃所特有的伪影 se.

值得注意的是，所有数据集在攻击下的相关性都略强于自然条件下（例如，在 GSM8K 上，U_{sys} 的 $r = -0.89$ 与 -0.82 相比）。这可能反映出对抗性输入会产生更极端的不确定性值，从而强化不确定性与准确率之间的关系。

C.3.4. UDPO 最稳健

表 7 比较了不同方法的性能下降情况。UDPO 在攻击下仍保持最高准确率（GSM8K 上为 77.2%，TruthfulQA 上为 66.4%），且相较于自然设定，其性能下降幅度最小（分别为 -5.2 个百分点和 -14.8 个百分点）。相比之下，标准 MAD 的准确率下降超过 34 个百分点，而即便是为鲁棒性而设计的 RMAAC，准确率也下降了 23.8 个百分点。

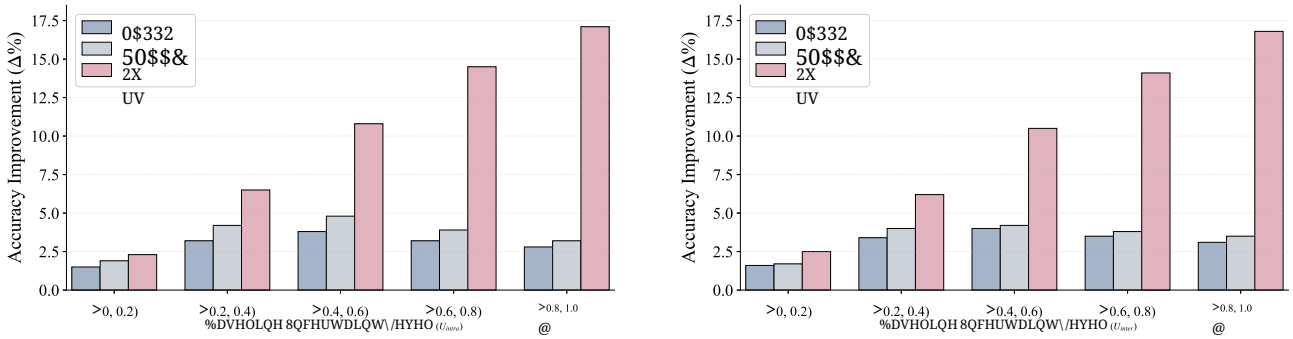
这种鲁棒性源于显式的不确定性分解：U_{intra} 检测恶意输入所致的推理不一致，U_{inter} 在隐蔽错误与正确推理相冲突时标记分歧，而 U_{sys} 则将这些信号整合为一种自然的故障检测机制。标准 MAD 缺乏任何防御；MAPPO 的集中式协调可能被操纵，从而形成错误共识；RMAAC 假设的是随机噪声，而非对抗意图。

C.4. 按问题难度进行分析

正文中的图 5 展示了按 U_{sys}分层的准确率提升。这里我们采用 U_{intra}和 U_{inter} 作为分层标准，给出类似的分析，以证明研究结果的一致性。

观察结果。这三种不确定性指标呈现的趋势均一致：

• **按.....分层** $U_{\text{组内}}$（图 11a）：UDPO 从 +2.3%（最低水平）提升至 +17.1%（最高水平）。MAPPO 从+1.5%提升至 +2.8%，并在中等不确定性下达到峰值 +3.8%。RMAAC 从 +1.9%提升至 +3.2%，并达到峰值 +4.8%。



(a) 按 U_{intra} 分层　　*(b)* 按 U_{inter} 分层

图 *11.* **准确率提升 (Δ%)，按不确定性水平划分** 并采用其他分层指标。补充正文中的图6。结果为所有数据集和模型的平均值。结果呈现一致的模式：UDPO 的增益单调递增，而 MAPPO 和 RMAAC 在高不确定性下趋于平稳或下降。

• **按.....分层** $U_{\text{组间}}$ （图 11b）：UDPO 从 +2.5%提升至 +16.8%。MAPPO 从 +1.6%提升至 +3.1%，并达到峰值+4.0%。RMAAC 从 +1.7%提升至 +3.5%，并达到峰值 +4.2%。

启示。 三种分层指标均得出一致结果，证实了我们的发现具有稳健性。无论采用哪种不确定性指标来界定“难度”，UDPO 在较难问题上都能带来更大的增益，而两种基线方法的效果则趋于平稳或下降。MAPPO 和 RMAAC 均呈现倒 U 型模式：在中等难度下效果最好，在难度两端的有效性则有所下降。这证实了 UDPO 从根本上解决了辩论崩溃问题，而非利用特定指标所产生的伪象。

D. 不确定性量化详情

本附录为构成系统级不确定性 U_{sys} 的三个组成部分提供正式定义。

D.1. 组件定义

令 $\{a_T^{(1)}, \ldots, a_T^{(N)}\}$ 表示来自 N 个智能体的最终轮回答，并令 $p(y \mid \tau)$ 表示由这些回答产生的候选答案经验分布。

归一化熵。 熵这一组成部分刻画了答案分布的离散程度：

$$\tilde{H}(\tau) = -\frac{1}{\log K} \sum_{y \in \mathcal{Y}} p(y \mid \tau) \log p(y \mid \tau) \tag{16}$$

其中 $K = |\mathcal{Y}|$ 是不同候选答案的数量。通过 $\log K$ 进行归一化可确保 $\tilde{H}(\tau) \in [0,1]$，其中 $\tilde{H}(\tau) = 0$ 表示一致同意，而 $\tilde{H}(\tau) = 1$ 表示所有候选答案上的均匀分布。

分歧指标。 分歧分量用于判断智能体是否达成共识：

$$D(\tau) = \mathbf{1}\Big[\max_{y \in \mathcal{Y}} p(y \mid \tau) < 1\Big] \tag{17}$$

只要有任何智能体的答案与多数不同，该二元指标就为 1；只有当所有智能体给出相同答案时，该指标才为 0。尽管它与熵有关，分歧仍体现了一种独特的失效模式：即使少数意见占比很小，也可能表明存在潜在的推理冲突。

留一不稳定性。该指标衡量最终决策对单个智能体的敏感性：

$$L(\tau) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\Big[\hat{y}(\tau) \neq \hat{y}(\tau^{-i})\Big] \tag{18}$$

其中，$\hat{y}(\tau)$ 表示使用全部 N 个智能体进行多数投票所得的预测，$\hat{y}(\tau^{-i})$ 表示移除智能体 i 后进行多数投票所得的预测。该指标用于识别脆弱共识：如果移除单个智能体就会改变结果，那么集体决策就缺乏鲁棒性。

D.2. 系统级不确定性与奖励

系统级不确定性结合了以下三个信号：

$$U_{\text{sys}}(\tau) = \frac{1}{3}\Big[\tilde{H}(\tau) + D(\tau) + L(\tau)\Big] \tag{19}$$

每个组成部分都刻画了集体不可靠性的一个不同方面：熵衡量答案分散度，分歧检测少数派异议，而留一不稳定性则识别脆弱多数。

UDPO（第 4.2 节）中使用的相应奖励会直接反转这种不确定性：

$$r_{\text{sys}}(\tau) = 1 - U_{\text{sys}}(\tau) = 1 - \frac{1}{3}\Big[\tilde{H}(\tau) + D(\tau) + L(\tau)\Big] \tag{20}$$

当熵较低（预测置信度高）、不存在分歧（一致同意），且留一不稳定性为零（稳健共识）时，该奖励达到最大值。