马尔可夫常微分方程引导评分可评估语言模型离线推理轨迹的质量

Arghodeep Nandi[1], **OjasvaSaxena**[1],和 **Tanmoy Chakraborty**[*1,2, 3]

[1]电气工程系，印度理工学院德里分校，印度新德里[2]亚迪人工智能学院，印度理工学院德里分校，印度新德里[3]印度理工学院德里分校，阿联酋阿布扎比

arXiv:2603.01580v1 [cs.CL] 2 Mar 2026

摘要

生成式语言模型生成的推理轨迹正越来越多地用于各种任务，从数学问题求解到自动事实核查。然而，现有评估方法在很大程度上仍停留于机械层面，无法以一种能适用于多样且逐步退化的推理过程的方式，捕捉以人为中心的推理质量概念。我们提出 MarODE，这是一种离线评估框架，可为推理轨迹赋予质量分数。我们通过以人为中心的扰动和人类判断来评估其有效性，二者共同衡量评估指标的基本维度——优良性和可靠性。该方法以推理进程的马尔可夫表述和基于常微分方程的轨迹动态刻画为基础，从而能够高效评估推理质量。在大规模评估中，MarODE 在 Somers' D 相关性指标下的表现比现有基线方法高出 250% 以上。我们的结果凸显了理论驱动的评估框架的价值，因为推理轨迹正日益成为语言模型系统的核心。

1 引言

随着大语言模型（LLM）的兴起，人们越来越关注如何利用其生成能力，生成类人化的多步推理轨迹。随着这类轨迹日益普及，评估其质量已成为一项关键挑战。近期研究表明，这一挑战源于缺乏可靠的推理链自动评估方法，也缺乏能够系统比较不同任务中各种推理策略的通用框架（Hao 等人，2024）。推理轨迹通常以思维链或中间推理依据的形式呈现，研究表明，它们能够显著提升模型在算术、符号操作、多跳问答和逻辑推理等复杂任务上的表现（Wei 等人，2022；Kojima 等人，2022；Wang 等人，2022b）。这些轨迹使模型能够将复杂问题分解为更简单的子步骤，因此在数学、事实核验、编程和决策支持等广泛领域都能发挥作用。实证分析表明，显式生成中间推理步骤是带来这些改进的关键因素；即使推理轨迹存在噪声或不够完善，也能显著提升多步推理表现（Wang 等人，2023年）。既有研究表明，生成的推理依据并不总是与模型底层的决策过程一致，这引发了人们对可靠性和可解释性的担忧（Jacovi 和 Goldberg，2020；Yao 等人，2023年）。随着推理轨迹日益成为人类与智能系统交互的媒介，系统地界定和评估其质量，对于确保系统可靠、有效地部署仍至关重要。

现有的推理评估方法大致可分为两种范式。生成时或线上评估在解码过程中评估推理质量，通常利用模型的内在信号（例如词元概率、不确定性估计或中间验证器反馈）来引导或评分生成过程（Lee 和 Hockenmaier，2025年）。相比之下，事后或离线评估针对已完成的推理轨迹，将其视为文本材料，并依据步骤级有效性、连贯性、事实依据和答案一致性等外部标准评估其质量。本文聚焦于推理轨迹的事后评估，旨在独立于生成机制，刻画并衡量已完成推理过程的质量。

先前的研究往往难以捕捉推理评估的一个关键方面——自然语言推理本质上是以人为中心的，无法仅依靠纯粹的机械扰动或旨在检测此类扰动的指标来可靠评估（Chrysostomou 和 Aletras，2021年）。随着语言和推理风格的演变，评估指标必须保持可泛化性，而不能与特定的表层模式紧密耦合（Jacovi 和 Goldberg，2020）。为弥补这一差距，我们引入了以人为中心的扰动（第4.2节），以更好地反映推理中的现实变化。这些

*通讯作者：tanchak@iitd.ac.in

扰动分数作为真实值，用于衡量指标分数与退化程度的相关性。我们表明，既有基准（如 ROSCOE（Golovneva 等人，2023年））的性能在这些条件下会急剧下降。受此观察启发，我们提出 MarODE——一种旨在捕捉连贯性、质量和一致性的新型评估框架。我们的方法以基于马尔可夫链的原则性表述，以及使用常微分方程（ODE）建模的一致性概率为指导。

在不同数据分布下，MarODE 在对扰动的敏感度方面始终优于现有基线方法；在经专家评估的推理轨迹上，它与人工判断的一致性也更高，这表明其优良性和可靠性。由于高质量长篇推理轨迹有限，我们利用 LIAR（Wang，2017 年）和 PolitiFact（Põldvere 等人，2023年）等事实性数据集中的主张—判决—证据结构，构建用于扰动分析的推理轨迹。我们还在涵盖多个推理领域的多种数据集上，进一步评估其与人工判断的一致性。

2 结果

MarODE 的形式化定义。评估推理轨迹的质量，不能只看最终答案正确性或表层流畅度。既有研究一再表明，高质量推理的特征包括步骤间递进连贯、全局方向一致，以及明确以证据或前提为依据（Dalvi 等人，2021年）。这些原则是众多推理基准和评估框架的基础，包括 ROSCOE（Golovneva 等人，2023年）、EntailmentBank（Dalvi 等人，2021年）、ProofWriter（Tafjord 等人，2021年），以及用于数学和事实推理的基于验证器的方法（Cobbe 等人，2021年；Geva 等人，2021年）。受这些见解启发，我们将推理质量定义为：中间步骤在多大程度上连贯、不冗余、逻辑递进、以证据为依据，并朝着某个思路推进。我们对推理质量的定义独立于最终结论，强调思路流的方向，而非其终点。

MarODE 在统一的评估框架中直接将这些维度转化为可操作的指标。第一个组成部分——马尔可夫连贯性——通过奖励各步骤之间连续的逻辑转换来评估局部结构。这与近期量化推理链中的语义对齐和逻辑错误的研究相一致（Wang 等人，2026年）。第二个组成部分——基于常微分方程的质量建模——捕捉冗余和方向性进展等全局属性，并抑制循环或重复的推理序列（Chen 等人，2025年）。近期研究指出，语言模型可能生成流畅却不忠实的解释，无法反映其内在决策过程（Turpin 等人，2023年）。第三个组成部分——证据对齐——确保推理步骤与相关证据之间存在依据关联，这与基于验证器的范式及过程级有效性指标相一致。方法部分及附录A和B 详细介绍了 MarODE 的形式化定义。

通过将这些互补组成部分整合为一个统一分数，MarODE 能够全面衡量推理轨迹质量，同时捕捉结构连贯性和证据可靠性。这一设计使 MarODE 能够作为一种通用评估指标，适用于事实、逻辑和数学推理，同时以高质量人类推理的既有原则为依据。下文将给出完整的形式化定义。

$$\begin{aligned}\mathrm{MarODE}(r) = w_c\,\mathbb{E}_{\pi\sim\mathbf{P}}\left[\frac{1}{|\pi|-1}\sum_t \mathbb{I}(|i_{t+1}-i_t|=1)\right] \\ + w_q\left[\tfrac{1}{2}\left(1-\frac{1}{|\mathscr{P}|}\sum_{(i,j)\in\mathscr{P}}\rho_{ij}\right)+\tfrac{1}{2}\,p_K\right] \\ + w_e\left[\frac{1}{K+1}\sum_{i=0}^{K}\mathrm{clip}_{[0,1]}\Big(0.5+\lambda_e\,\Pr_i(\mathrm{entail})-\lambda_c\,\Pr_i(\mathrm{contradict})\Big)\right]\end{aligned} \tag{1}$$

其中 $\mathbf{P}$ 是步骤嵌入上的马尔可夫转移矩阵，$\pi=(i_0,i1,\ldots,i_L)$ 表示一条采样游走，$\mathscr{P}$ 表示上三角步骤对集合，ρ_{ij} 表示冗余惩罚，p_K 是 ODE 积分方向信念，而 $\Pr_i(\cdot)$ 是 R_i 与证据之间的自然语言推理（自然语言推理）概率。

我们给出了 MarODE 在多种数据集、模型和评估设置上的全面实证评估。我们首先考察其与以人为中心的扰动的一致性，随后通过受控消融实验分离各个组件的贡献。接着，我们在既有基准上评估其与专家人工判断的一致性。最后，我们分析其在不同上下文内监督下的鲁棒性，并突出其在不同样本数设置下的稳定性。

评估概述。我们在多个数据集、模型和不同样本数设置下，将 MarODE 与成熟的推理质量基线进行比较。我们从两个维度考察性能：优良性，即衡量该指标对通过以人为中心的扰动控制推理退化的敏感度；以及可靠性，即衡量其与人类专家判断的一致程度。

合成推理轨迹（第 4.1.1 节）由五种推理优化型大语言模型生成：DeepSeek-Qwen-14B、DeepSeek-LLa MA-8B、DeepSeek-Qwen-7B（Guo 等人， 2025年）、Qwen-3B-CoT（Bai 等人， 2023年）和 GPT-OSS-20B（Kumar 等人， 2025年）。这些模型在从 LIAR 和 PolitiFact 提取的 8,600 条事实性主张上，采用 1、2 和 4 样本提示，为每个模型生成 25,800 条结构化推理轨迹。所用提示详见附录D。评估对象包括合成扰动推理轨迹，以及涵盖事实、逻辑和数学推理的人工标注基准（第 4.1.2 节）。我们使用 Somers'D量化指标分数与扰动级别之间的相关性。我们将 MarODE 与 ROSCOE、ReCEval、局部与全局连贯性以及大语言模型评审进行比较，以便系统评估其在不同评估设置下的敏感度、显著性和一致性。详情见方法部分（第 4.4 节）。

以人为中心的扰动下的相关性分析。表 1 报告了不同数据集、提示设置和骨干模型上的 Somers' D 相关性，用于衡量与以人为中心的扰动的一致性。在 LIAR（Wang，2017 年）和 PolitiFact（Põldvere 等人，2023年）数据集、不同样例数以及不同模型系列中，MarODE 始终取得最强的相关性，优于所有基线方法。我们发现，ROSCOE（Golovneva 等，2023）各变体在由不同模型变体生成的数据集上表现出有限的一致性。语义对齐与相似度（ROSCOE-SA、ROSCOE-SS）产生中等程度的相关性（∼0.10–0.14），而逻辑推理（ROSCOE-LI）的相关性仍然较弱，逻辑一致性（ROSCOE- LC）则接近于零或为负，且有多项结果未达到统计显著性。聚合提升了鲁棒性和统计显著性（ROSCOE_MEAN ∼0.06–0.10），但表现仍明显低于 MarODE。其他基线方法，包括 LLM-AS-A-JUDGE（Kim 等人，2024）、局部与全局连贯性（Kotonya 和 Toni，2020）以及 RECEVAL（Prasad 等，2023），相关性始终较低（通常为 < 0.05），且对样例数增加的敏感度有限。相比之下，MarODE 组件在不同数据集、提示方案和模型上均表现出持续较高且具有统计显著性的相关性（∼0.24– 0.29）。与 ROSCOE_MEAN 相比，MarODE 的相关强度提升了 235%–279%（表 1），且这一优势在不同样例数和架构变体下保持稳定。值得注意的是，对于由不同模型系列生成的推理轨迹，指标的相对排序保持不变，这表明 MarODE 的改进具有系统性，而非特定于模型。

指标	通义千问-30亿参数思维链	DeepSeek-通义千问-7B	DeepSeek-通义千问-14B	Deepseek-LLaMA-8B	GPT-OSS-20B	LIAR (1样本)	LIAR (2样本)	LIAR (4样本)	PolitiFact (1样本)	PolitiFact (2样本)	PolitiFact (4样本)
ROSCOE-SA	0.1187	0.1053	0.1111	0.1117	0.1048	0.1004	0.0983	0.1075	0.1145	0.1152	**0.1281**
ROSCOE-SS	0.1294	0.1242	0.1301	0.1256	0.1284	0.1167	0.1146	**0.1386**	0.1264	0.1350	0.1331
ROSCOE-LI	0.0318	0.0163	0.0496	0.0172	0.0146	0.0253	0.0279	0.0286	0.0283	0.0256	0.0200
ROSCOE-LC	-0.0189	-0.0009	-0.0223	0.0228	-0.0173	-0.0073	-0.0076	-0.0184	0.0002	-0.0047	-0.0066
ROSCOE _MEAN	0.0819	0.0686	0.0958	0.0689	0.0622	0.0691	0.0728	0.0776	0.0793	0.0785	0.0742
LLM_作为_评判者	0.0436	0.0271	0.0421	0.0375	0.0330	0.0367	0.0282	0.0140	0.0516	0.0455	0.0407
局部 _and_全局	0.0458	0.0192	0.0516	0.0321	0.0316	0.0417	0.0397	0.0347	0.0296	0.0379	0.0330
ReCEval	0.0457	0.0049	0.0311	-0.0018	0.0072	0.0082	0.0084	0.0211	0.0253	0.0243	0.0177
MarODE	**0.2937**	**0.2371**	**0.2882**	**0.2921**	**0.2634**	**0.2618**	**0.2636**	**0.2604**	**0.2895**	**0.2840**	**0.2792**
MarODE_连贯性(α)	**0.2857**	0.1747	**0.2798**	0.2826	0.2014	0.2395	0.2357	0.2427	0.2165	0.2284	0.2296
MarODE_质量(β)	0.0286	0.0639	0.0331	0.0082	0.0467	0.0181	0.0342	0.0358	0.0374	0.0405	0.0410
MarODE_证据(γ)	0.2353	0.1955	0.2253	0.2367	0.2139	0.2075	0.2122	0.2082	0.2398	0.2341	0.2279
MarODE($\alpha\beta$)	**0.3272**	**0.2093**	**0.3279**	**0.3118**	**0.2309**	**0.2762**	**0.2743**	**0.2769**	**0.2583**	**0.2652**	**0.2639**
MarODE($\beta\gamma$)	0.2403	0.1992	0.2309	0.2392	0.2173	0.2103	0.2164	0.2129	0.2444	0.2379	0.2322
MarODE($\alpha\gamma$)	0.2857	**0.2303**	0.2791	**0.2857**	**0.2570**	**0.2550**	**0.2560**	**0.2527**	**0.2799**	**0.2763**	**0.2714**

表 1. Somers' D 相关性用于衡量以人为中心的扰动分数与不同骨干模型及提示设置下评估指标之间的关联。列对应于由多种大语言模型生成的推理轨迹，以及在不同提示样例配置下的事实核查基准（LIAR 和 POLITIFACT）。数值越高，表明对推理质量中扰动引起的变化越敏感。每列中最高的三个相关性以粗体标出，而带下划线的数值表示相关性在统计学上不显著（$p \geq 0.05$）。结果包括 ROSCOE 变体、基于大语言模型和面向连贯性的基线方法、ReCEval，以及 MarODE 的组件级和成对消融实验，从而能够详细分析不同模型和提示方案下的优良性。

MarODE 的分块分析。我们对 MarODE 进行分块消融，以评估各组成部分——马尔可夫连贯性（α）、ODE引导的质量（β）和证据对齐（γ）在不同数据集和模型上的贡献及相互作用（表 1）。 马尔可夫连贯性 (α) 是主导因素，其独立相关性较高（∼0.24–0.29）。移除其他组成部分通常能提升性能：不包含证据支撑的成对组合，尤其是

指标	蕴含银行证明编写器_		GSM	STRATQA
ROSCOE-SA	-0.10337	0.17031	-0.04427	0.06290
ROSCOE-SS	-0.04991	0.15439	-0.03248	0.01250
ROSCOE-LI	0.03013	0.05386	-0.02203	0.06036
ROSCOE-LC	0.01167	0.09714	-0.05359	0.01563
ROSCOE _MEAN	-0.00730	0.21145	-0.03248	0.08869
LLM_作为_一名_评判者	0.07436	0.00000	0.00277	0.00449
局部 _and_全局 _连贯性	0.03185	-0.00670	0.01650	0.04894
ReCEval	-0.02211	-0.02183	**0.05052**	**0.25611**
MarODE	**0.16981**	**0.33896**	**0.18578**	**0.22563**
MarODE_连贯性(α)	0.03114	0.21407	**0.18086**	0.12542
MarODE_质量(β)	0.02810	**0.23998**	0.02121	0.22133
MarODE_证据(γ)	**0.16870**	0.13692	-	0.07443
MarODE($\alpha\beta$)	0.02293	**0.33857**	-	**0.27349**
MarODE($\beta\gamma$)	0.14841	0.23823	-	0.19516
MarODE($\alpha\gamma$)	**0.17732**	0.21572	-	0.11604

表 2. Somers D 相关性，用于衡量四个推理基准中人工（手动）判断与评估指标之间的一致性：ENTAILMENTBANK、PROOFWRITER、GSM 和 STRATEGYQA。数值越高，表示与人工评估的一致性越强。对于每个数据集，相关性最高的三项以**粗体**标出。带下划线的数值表示相关性在统计上不显著（$p \geq 0.05$）。结果涵盖 ROSCOE 变体、聚合方法和基于大语言模型的基线方法、ReCEval，以及 MarODE 及其组件级和成对消融实验，从而能够比较整体性能以及各个推理维度的贡献。

($\alpha\beta$)，始终优于（$\alpha\gamma$），总体上取得最强的成对相关性（∼0.21–0.33）。这表明证据信号可能降低一致性。证据对齐(γ) 的独立表现中等（∼0.20–0.24），但与其他模块组合时带来的提升有限。在若干情况下，与基于（α）的组合相比，加入（γ）仅带来微小甚至负向变化，表明其作用次要或取决于具体情境。*ODE*引导的质量(β) 的独立一致性较弱（大多为 < 0.06），但与（α）搭配时可发挥稳定作用，提升一致性和相关强度。然而，从完整聚合中移除（β）反而可能提高相关性，这表明它的作用并非普遍适用。总体而言，尽管 MarODE 依然表现强劲（∼0.24–0.29），消融实验显示，最佳性能往往来自以连贯性为核心的组合。这些结果表明，MarODE 的有效性主要由结构连贯性驱动，质量发挥辅助作用，而证据的贡献有限。

与人类判断进行评估。表2 报告了四个人工评估基准中指标分数与人类判断之间的 Somers D 相关性：ENTAILMENTBANK（Dalvi 等人，2021年）、PROOFWRITER（Tafjord 等人，2021年）、GSM8K（Cobbe 等人，2021年）和 STRATEGYQA（Geva 等人，2021年）。总体而言，先前的基线方法与人工评估的一致性较弱且不稳定，且往往不显著；在多个任务中，一些相关性接近于零或为负。ROSCOE 各组件的表现差异很大。尽管 ROSCOE-SA 和 ROSCOE-SS 在 PROOFWRITER 上取得了中等程度的正相关，但它们在其他数据集上的表现有所下降，包括在 ENTAILMENTBANK 和 GSM 上出现负相关，而且许多结果在统计上不显著。聚合方法（ROSCOE_MEAN）提高了在 PROOFWRITER 上的一致性，但在其他数据集上仍接近于零或为负。其他基线方法（LLM-AS-A-JUDGE、局部与全局连贯性）和 ReCEval 的一致性也较有限；ReCEval 仅在 STRATEGYQA 上表现出较强相关性。相比之下，MarODE 及其组件在各基准中始终与人类判断保持一致。尽管各个维度的表现因数据集而异，组合变体提高了一致性和统计显著性。聚合后的 MarODE 整体一致性最强，且相关性显著。

按样例数分析指标分数。图1 分析了上下文示例数量（1、2 和 4样本提示）变化对未扰动推理轨迹指标分数的影响。在所有评估指标中，改变样例数并未导致分数分布出现显著变化。小提琴图的形状、中位数和四分位距都相似，表明无论提供多少个样例，生成的推理轨迹的整体质量都保持稳定。

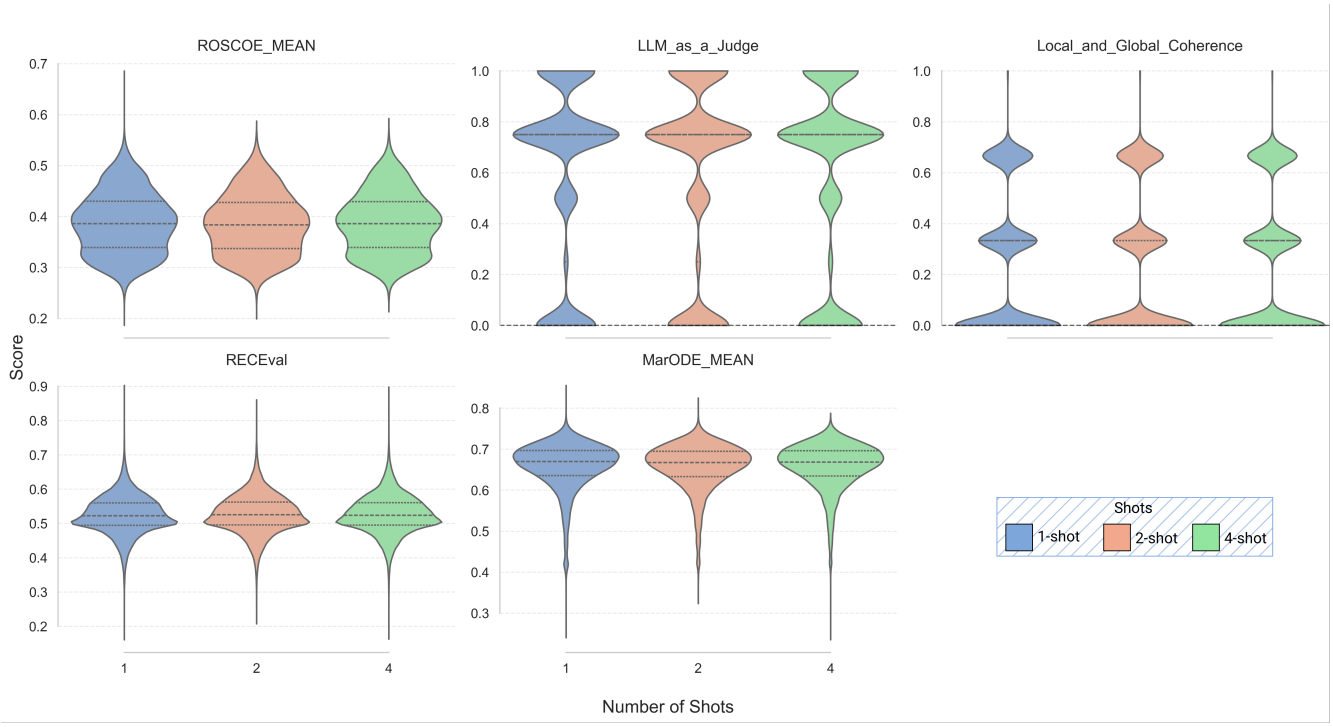


图1。 不同提示方案下推理轨迹质量分数的分布。 小提琴图展示了五种评估指标——ROSCOE、LLM-as-a-Judge、局部与全局连贯性、ReCEval 和 MarODE——在三种用于生成推理轨迹的提示设置（1、2 和 4样本）下的分数分布。对于给定的基线方法，分布的整体形状在不同样例数下基本保持一致，这表明该指标对提示变化具有相对稳定性。相比之下，不同基线方法的分布形状差异显著，反映了不同评估方法在尺度、敏感度和评分行为上的区别。

Wilcoxon 符号秩检验同样未能发现使用不同样例数生成的推理轨迹质量有任何显著变化（表3）。一些基线方法（如 LLM-as-a-Judge 和局部与全局连贯性）在不同样本数设置下，集中趋势的离散程度显著。其他基线方法（如 ROSCOE 和 ReCEval）虽然离散程度较低，但其分数分布存在不对称性。相比之下，MarODE 在所有提示样例配置下的分布都明显更集中、更对称，体现出更高的稳定性。

3 讨论

MarODE为评估生成文本中的推理质量提供了更广阔的视角。除了针对特定任务的评估，它还提供了一个通用框架，用于比较生成的解释的质量、区分较强和较弱的推理模式，并评估不同模型、提示策略和领域中的可靠性。当前推理评估领域的研究同样主张采用评估过程而非仅评估结果的指标（Qiu 等人，2025年；Zhou 等人，2025年）。通过明确建模结构连贯性、推进过程和证据支持，该方法能够在不同推理场景中提供更稳定、可解释且可迁移的评估，这与近期关于超越特定任务指标、开展整体推理评估的呼吁相一致（Murthy 等人，2025年）。随着大语言模型被用于生成多步推理，这一相关性日益重要，因为理解推理过程的质量与理解最终结果同样重要。

MarODE的优良性。评估指标的优良性是指该指标在多大程度上能够捕捉其旨在衡量的潜在现象中有意义的变化，尤其是在现实变动下（Yeh，2000年；Amidei等人，2018）。对于推理评估，指标应能有意义地反映现实变动下推理质量的变化——能够响应逻辑连贯性、证据使用和推理一致性方面的差异，同时不受表面或风格变化的影响（Lan等人，2025年；Lee和Hockenmaier，2025年）。在所有实验设置中，MarODE始终展现出作为评估指标的优良性，这一点体现在它与以人为中心的扰动之间具有较高且稳定的相关性（表1，图2）。与相关性较弱或在不同数据集、样例数和骨干模型上有所下降的先前基线方法不同，MarODE在各种条件下均保持稳定表现。相较于现有方法的提升幅度

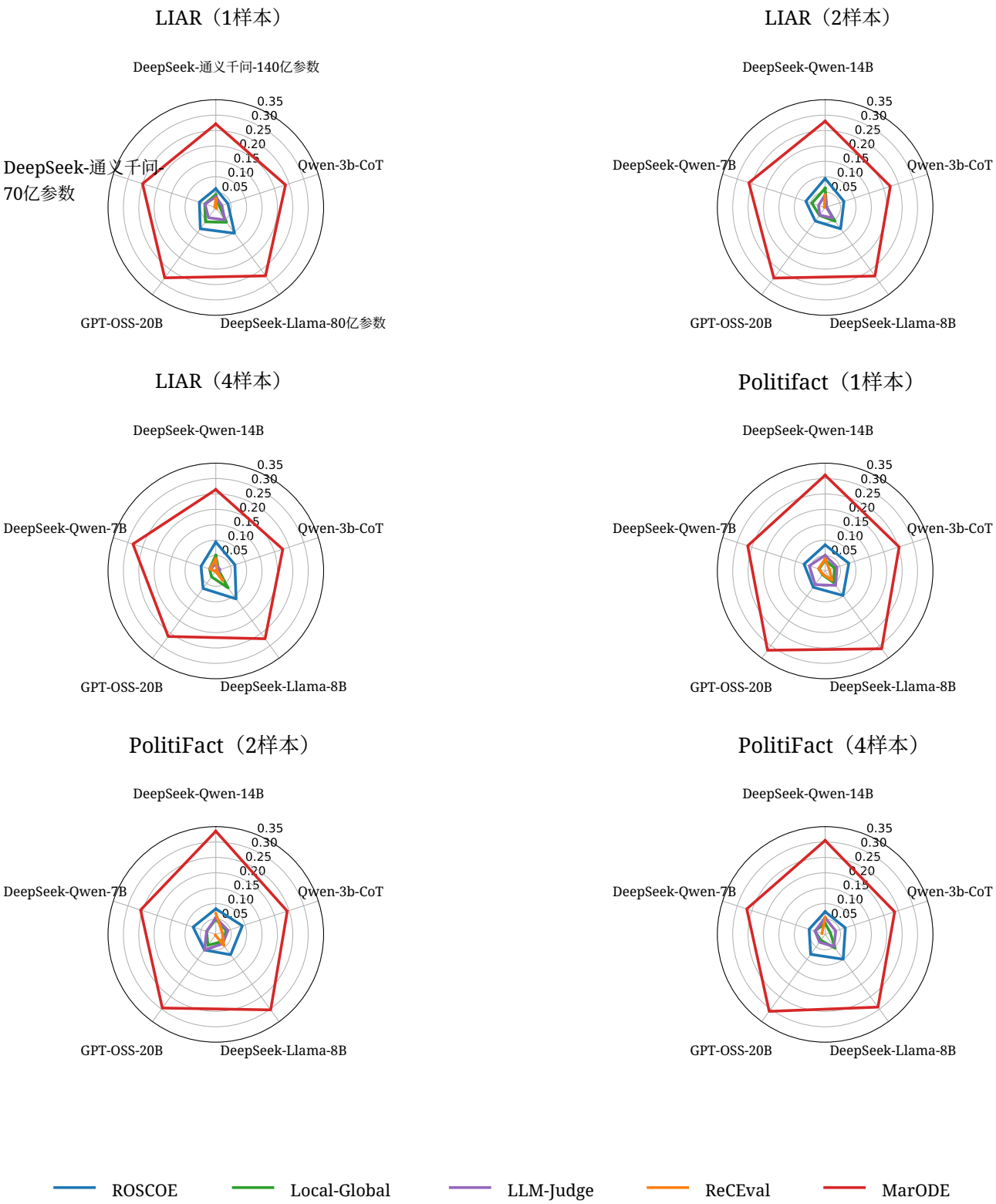


图2。 以雷达图展示各数据集和模型上的事后推理评估指标对比表现。 每个子图对应一个特定数据集（LIAR 和 PolitiFact，分别采用 1、2 和 4 样本提示），坐标轴代表五种骨干模型。不同颜色的线表示 ROSCOE、Local-Global、LLM-as-a-Judge、ReCEval 和我们提出的 MarODE 各指标与以人为中心的扰动之间的 Somers' *D* 相关性。MarODE 在不同数据集、样例数和模型系列中始终表现出最高的相关性，表明其对影响推理质量的扰动具有更强的敏感度。

基线方法	(4样本 – 1样本)		(2样本 – 1样本)		(4样本 – 2样本)	
	中位数差异	显著	中位数差异	显著	中位数差异	显著
ROSCOE-SA	-0.000769	TRUE	0.000777	TRUE	-0.001501	TRUE
ROSCOE-SS	0.001300	TRUE	-0.001400	TRUE	0.002700	TRUE
ROSCOE-LI	-0.000336	FALSE	-0.002918	TRUE	0.004092	TRUE
ROSCOE-LC	-0.001100	TRUE	-0.001100	TRUE	0.000200	TRUE
ROSCOE _MEAN	-0.000374	FALSE	-0.001979	TRUE	0.001659	TRUE
LLM_作为_一名_评判者	0.000000	否	0.000000	TRUE	0.000000	TRUE
局部 _and_全局 _连贯性	0.000000	TRUE	0.000000	否	0.000000	TRUE
ReCEval	0.001079	TRUE	0.002021	TRUE	-0.001183	TRUE
MarODE	-0.000479	TRUE	-0.001809	TRUE	0.001725	TRUE
MarODE_连贯性(α)	-0.002344	TRUE	-0.000930	TRUE	-0.001465	TRUE
MarODE_质量(β)	0.000188	TRUE	-0.000102	TRUE	0.000257	TRUE
MarODE_证据(γ)	-0.000658	假	-0.004585	TRUE	0.003825	TRUE
MarODE($\alpha\beta$)	-0.000588	TRUE	-0.000589	TRUE	-0.000115	FALSE
MarODE($\beta\gamma$)	-0.000095	FALSE	-0.002524	TRUE	0.002290	TRUE
MarODE($\alpha\gamma$)	-0.001165	TRUE	-0.002749	TRUE	0.002195	TRUE

表 3. 三种比较设置（4–1）、（2–1）和（4–2）下，配对 Wilcoxon 符号秩检验的中位数分数差异及统计显著性（$p < 0.05$）。每种设置均比较在不同提示样例条件下生成的推理轨迹；中位数差异为正或负，分别表示每对中前一种条件下的指标分数较高或较低。显著 列表示观察到的中位数差异是否具有统计可靠性。结果涵盖 ROSCOE 变体、聚合基线、基于 LLM 的评估器和连贯性评估器、ReCEval，以及 MarODE 及其组成部分消融实验和成对消融，从而能够细致分析其对推理质量变化的敏感度。

方法的相关性强度显著。与聚合的 ROSCOE 变体相比，MarODE 的相关强度提高了 235%–279%，表明它对影响推理质量而非表层文本属性的扰动具有更强的敏感度。重要的是，这种提升并不局限于特定数据集或模型家族；不同架构和提示方案下，指标的相对排序仍保持不变。这种一致性表明，MarODE 可能捕捉到了与推理轨迹质量相关的基本属性，而非利用特定数据集伪影。综上，这些结果表明，MarODE 满足事后推理评估的一项核心要求——在真实的人本变化条件下，它能够可靠地区分高质量和低质量的推理轨迹。

MarODE 的可靠性。 除了原始性能之外，可靠的推理指标还应在不同推理领域中与人工判断保持一致（Chen 等人, 2023年; DeYoung 等人, 2020）。人工评估基准上的结果（表 2）表明，在所有评估指标中，MarODE 与专家标注的一致性最强且最稳定。现有基线方法，包括 ROSCOE、LLM-as-a-Judge 和 Local–Global Coherence，表现出高度可变且依赖任务的行为，往往产生不显著或负相关。相比之下，MarODE 在涵盖符号推理、数学问题求解、多跳问答和基于策略的推理的四个基准测试中，均取得了统计显著的相关性。这种一致性表明，MarODE 捕捉的是对人工评估员有意义的推理属性，而非对特定推理风格过拟合。综合来看，这些发现支持将 MarODE 作为一种通用推理评估指标，其分数能够反映人类对推理质量的感知。指标分数的分布特性也为其可靠性提供了更多证据。如图 1 所示，MarODE 呈现平滑、单峰且近似对称的分布，而若干基线方法则表现出偏斜、多峰或重尾模式。此类不规则现象往往表明其对虚假线索敏感、存在启发式饱和，或评分行为不稳定（Fomicheva 和 Specia, 2019年）。从统计学视角来看，这种行为是隐式聚合多个弱相关推理因素的指标的典型特征（Longjohn 等人, 2025年）。当评估信号整合了多种底层属性时，分数分布往往更加平滑、规整，体现出稳定的聚合过程，而非脆弱的启发式方法（Soh 和 Zhao,2024）。先前研究还表明，分布结构连贯的指标比逐点度量或基于阈值的度量更贴近人工判断（Pillutla 等人, 2021年），而表现良好的评分函数能够产生一致且易于解释的输出（Rei 等人, 2023年）。在此背景下，MarODE 分数近似正态的形态表明，它对推理信号进行了均衡整合，而非由任何单一标准主导。相比之下，LLM-AS-A-JUDGE、局部与全局连贯性以及 ROSCOE 所呈现的不对称且不规则的分布，表明其聚合机制.....

可能对表层线索或特定推理风格更敏感。MarODE 更紧凑且更对称的表现，支持将其视为推理质量的稳定指标。

证据在推理轨迹评估中的重要性。我们的分析得出的一个重要洞见涉及显式证据对齐在推理评估中的作用。尽管证据支撑通常被认为是高质量推理的核心组成部分，但我们的结果呈现出更为细致的图景。证据对齐（γ）在基于扰动的评估和人工判断中展现出中等的单独表现。然而，将其与其他组成部分结合时，其贡献并不一致。在若干设置下，将证据对齐加入以连贯性为中心的公式会对相关强度产生边际影响，甚至负面影响（表1）。这一模式表明，显式证据信号有时会引入噪声，尤其是在证据相关性或归因存在歧义时。从经验上看，连贯性—质量配对（$\alpha\beta$）比包含全部三个组成部分的组合实现了更高的相关性。这些发现表明，通过马尔可夫转移捕捉的结构连贯性，相较于证据匹配，在感知推理质量中发挥着更为基础的作用。证据是一种次级的、依赖上下文的信号，只有在直接支撑中间步骤时，才会增强评估效果。这对指标设计具有重要蕴含，提醒我们应避免过度依赖证据信号。另一种解释与扰动本身有关。推理轨迹通常是在无法获取生成过程中使用的完整证据的情况下被单独评估的（Prasad 等人，2023年；Fang 等人，2024）。因此，仅应用于推理步骤的扰动可能无法充分捕捉依据关联性。因此，衡量证据敏感度可能需要直接针对事实一致性和证据归因的扰动，而不仅仅是局部、全局或表层的轨迹操纵。

推理质量的提示敏感性与稳定性。按示例数分析（第2节）表明，改变上下文示例的数量对生成的推理轨迹质量影响有限。如表3中的配对 Wilcoxon 符号秩检验所示，不同样本数设置之间未观察到一致或具有统计显著性的差异。在不同数据集和评估指标中，1、2 和 4样本提示的分数分布总体上十分相似，这表明增加示范样例并不会显著改变推理行为。这一模式说明，推理质量更多取决于模型和任务特征，而非提示长度。尽管若干基线指标在不同样本数设置下表现出波动或偏斜，MarODE 的表现仍相对稳定，其分数分布一致且对称。这种稳定性表明，MarODE 能够以受控的方式捕捉推理质量的逐渐变化，而不是放大提示诱发的波动。总体而言，这些发现对“增加上下文示例就能可靠提升推理质量”这一假设提出了质疑（Agarwal 等人，2024）。

图4中的信念轨迹分析进一步证明了这种稳定性。在始终支持、相互矛盾、振荡和逐渐漂移等多种推理动态情形下，MarODE 所依据的基于常微分方程的更新会使方向一致性平滑且连续地变化。这一行为与神经常微分方程领域的既有研究一致：研究表明，连续时间形式能够产生平滑的轨迹，并降低对离散化效应的敏感度（Chen 等人，2018）。相比之下，乘法更新往往会造成突变、坍缩至极端值或表现出高度波动，尤其是在翻转和极端确定性情形下。这种行为与信念与意见动态领域的研究发现一致：乘法规则容易受到噪声、排序效应和局部不一致性的影响，常常导致不稳定或过早确定（Su 和 Wu，2015）。关于稳定神经流的研究也进一步支持了基于常微分方程的表述：研究表明，施加适当约束的连续动力系统具有有界且可预测的行为，并能提高对扰动的敏感度（Massaroli 等人，2020）。

总之，这些结果表明，MarODE 的分布对称性和稳定性源于其对信念更新的连续建模。通过在推理步骤之间强制实现平滑过渡，MarODE 避免了离散乘法更新所特有的突变，因此在不同提示条件下表现出一致的评估行为。

扰动敏感性与指标泛化。 ROSCOE（Golovneva 等人，2023年）在其精选扰动下，在多个数据集上表现出完美相关性；然而，当扰动集发生变化时，这种表现无法泛化。在我们的实验中，改变扰动集会导致相关性大幅且不一致地下降，这表明 ROSCOE 和类似指标与其旨在检测的特定机械式扰动模式紧密耦合（Moradi 和 Samwald，2021年）。因此，在给定扰动集下的高相关性，可能反映的是对特定扰动的敏感性，而非通用推理评估。有效的推理指标应基本不受特定扰动算子的影响，而应捕捉推理轨迹的一般结构属性（Wang 等人，2022a）。MarODE 旨在实现这一目标，关注连贯性、方向性、冗余和证据依据，而非特定扰动线索。尽管这种通用表述得到的原始相关性得分较低，但与其他基线方法相比，MarODE 在检测多种以人为中心的扰动方面效果更佳，并且在不同分布设置下保持稳定的评估行为，因此更适用于通用推理评估。

总结。 MarODE 与以人为中心的扰动和专家判断的一致性最高，较先前指标将相关性提升了235%–279%。我们发现，连贯性是推理质量的主要驱动因素，ODE引导的质量则带来额外的协同作用，而证据对齐的作用有限。与基线方法相比，MarODE 还生成平滑且对称的分数分布，表明其评估行为稳定。总体而言，这些发现表明，推理质量在很大程度上取决于步骤推进是否连贯且一致，从而确立了 MarODE 作为评估推理轨迹的可靠通用框架。

4 方法

本文介绍如何构建、扰动和评估推理轨迹，以及这些要素如何汇聚于 MarODE。我们首先概述数据集和评估设置，随后说明如何通过连贯性、推进过程和证据对齐来建模推理质量。一个关键的设计选择是采用连续的 ODE 方法，逐步追踪推理的演变；与直接进行概率连乘相比，这种方法在较长的推理链中更加稳定，不易变得脆弱或过度敏感。综合这些选择，我们构建了一个透明且通用的框架，用于分析不同模型和设置下的推理质量。

4.1 数据集

本研究使用生成式与人工评估数据集评估推理轨迹质量，涵盖事实核验、逻辑推理和数学推理任务。数据集构建遵循统一流程，用于生成、提取和验证推理轨迹，从而确保不同模型和提示条件下的结果具有可比性。

4.1.1 生成的推理轨迹

我们使用两个广泛采用的事实核查基准构建大规模推理轨迹数据集：LIAR（Wang，2017 年）和 PolitiFact（Põldvere 等人，2023年）。这两个数据集共包含 8,600 条事实性主张，涉及带有真实标签的政治声明。我们为每条主张配备结构化提示，以引出逐步推理轨迹。我们使用五种专为推理优化的大语言模型生成推理轨迹：DeepSeek-Qwen-14B、DeepSeek-LLaMA-8B、DeepSeek-Qwen-7B（Guo 等人，2025年）、Qwen-3B-CoT（Bai 等人，2023年）和 GPT-OSS-20B（Kumar 等人，2025年）。对于每种模型，我们采用三种少样本提示配置，即 1样本、2样本和 4样本，因此每条主张对应三条不同的推理轨迹。对于每种模型，在所有提示设置下共生成 25,800 条推理轨迹。

令 $\mathscr{C} = \{c1, \ldots, c_N\}$ 表示事实性主张集合，并令 $\mathscr{P}_k(c)$表示针对主张 c 在 k-shot 配置下构造的提示，其中 $k \in \{1,2,4\}$。对于给定的语言模型 $m \in \mathscr{M}$，令 $p_m(y \mid \mathscr{P}_k(c))$表示以 $\mathscr{P}_k(c)$为条件时由m 诱导的输出序列条件概率分布。随后通过采样获得生成的回复

$$y_m^{(t)}(c,k) \sim p_m(\cdot \mid \mathscr{P}_k(c)), \tag{2}$$

其中 $t \in \{1, \ldots, T\}$ 表示重复解码尝试的索引，最多重试$T = 5$次。

$$\mathscr{E} : y \mapsto r, \tag{3}$$

每个生成的回复都会经过后处理，使用确定性提取算子隔离由明确推理标记界定的推理片段r。当且仅当推理轨迹 r 满足结构有效性谓词时，才视为有效

$$\mathscr{V}(r) = \mathbb{I}\big[r = \{R_0, R_1, \ldots, R_K, V_f\},\ K \geq 1\big], \tag{4}$$

其中 R_0表示初始推理步骤，$\{R1, \ldots, R_K\}$表示中间推理步骤，而Vf表示终止性的最终判定。

若 $\mathscr{V}(r) = 0$，则通过递增$t \leftarrow t + 1$重复生成，直到获得有效的推理轨迹或$t = T$。对于每个元组 (c,k,m)，保留的推理轨迹定义为

$$\boxed{r_m^*(c,k) = \arg\min_t \left\{ t \mid \mathscr{V}\left(\mathscr{E}\left(y_m^{(t)}(c,k)\right)\right) = 1 \right\}.} \tag{5}$$

用于生成这些轨迹的提示词详见附录 D。

4.1.2 人工评估的推理轨迹

除了模型生成的轨迹外，我们还纳入了一组经过精心筛选的推理轨迹，作为高质量的参考数据。这一子集取自四个成熟的推理基准：EntailmentBank（Dalvi 等人，2021年）、ProofWriter（Tafjord 等人，2021年）、GSM8K（Cobbe 等人，2021年）和 StrategyQA（Geva 等人，2021年）。我们从每个数据集中选取了 150 个实例，共计 600 条人工评估的推理轨迹。这些数据集涵盖演绎、逻辑、数值、常识和科学推理领域，从而能够在多样化的推理风格下进行评估。这些数据集覆盖了多种推理现象，包括多跳逻辑蕴涵、形式化证明构造、算术与符号推理以及常识问答。所有选定的轨迹都遵循明确的逐步推理格式，并清晰陈述结论，因此适合根据我们提出的方法对推理轨迹质量进行细粒度评估。我们使用三名专家人工评估员对这些数据集中的推理轨迹进行评估。评估涵盖多个维度，包括逻辑且连贯的推进、表达清晰度、无冗余性、与给定上下文的依据关联、支持最终判断的推理流程的连贯性，以及整体推理质量。

4.2 以人为中心的扰动

为评估推理轨迹在受控退化下的敏感度，我们对生成的推理轨迹施加一组以人为中心的扰动。这些扰动旨在有选择地破坏逻辑结构、语义连贯性或证据依据，同时保持表层流畅性，从而能够对推理质量进行细粒度评估。

令$r = \{R_0, R1, \ldots, R_K, V_f\}$表示由有序推理步骤 $\{R_i\}_{i=0}^{K}$和终止判定Vf组成的推理轨迹。扰动作用于步骤序列 $\{R_i\}_{i=0}^{K}$，除非明确施加扰动，否则Vf保持不变。

我们将扰动函数集（详细信息见表 4）定义为

$$\mathscr{F} = \{f_1, \ldots, f_{14}\},$$

其中，每个$f_j \in \mathscr{F}$都是一个部分算子

$$f_j : \{R_0, \ldots, R_K\} \to \{\tilde{R}_0, \ldots, \tilde{R}_K\},$$

返回一个成功指示器 $\sigma_j \in \{0, 1\}$。

为平衡扰动强度，轨迹被随机置换并分配一个目标扰动数量

$$p \in \{0, \ldots, 6\}.$$

对于每条轨迹，最多 p 个不同 t个扰动从 $\mathscr{F}$ 中均匀无放回抽样，并依次应用，直 il

$$\sum_j \sigma_j = p \quad \text{or} \quad \mathscr{F} \text{ is exhausted.}$$

扰动后的轨迹 ˜r 在保留步骤索引的情况下被重构。扰动分数定义为

$$s(p) = 1.0 - 0.1p, \qquad p \in \{0, \ldots, 6\},$$

得到 $s(p) \in [0.4, 1.0]$。每个数据实例存储 $(\tilde{r}, s(p), \{f_j\})$，用于下游评估。

对推理退化进行连续建模。与 ROSCOE 等先前基于扰动的评估框架主要采用二元标注来评估受扰动的推理轨迹不同，我们的方法将推理退化建模为一个渐进且结构化的过程。通过随机组合并分级施加以人为中心的扰动，针对局部步骤不一致、全局结构性破坏和表层噪声进行干预，我们对推理的渐进退化进行建模，反映人类推理中出现的部分逻辑崩溃、语义漂移和证据弱化（Dasgupta 等人，2022）。随机施加多个扰动算子，避免了固定转换模式带来的偏差，从而在保留表层流畅度的同时，对推理鲁棒性进行细粒度评估。这种设计支持对推理质量进行细致评估，而不局限于离散的故障检测。

4.3 MarODE：马尔可夫—常微分方程推理评估

我们提出 MarODE，这是一种用于评估推理轨迹质量的模块化指标。给定一条推理轨迹 $r = \{R_0, R_1, \ldots, R_K, V_f\}$, MarODE 通过三个组成部分计算质量：连贯性、质量和证据对齐，并将它们汇总为最终得分。图3 展示了 MarODE 的概览。

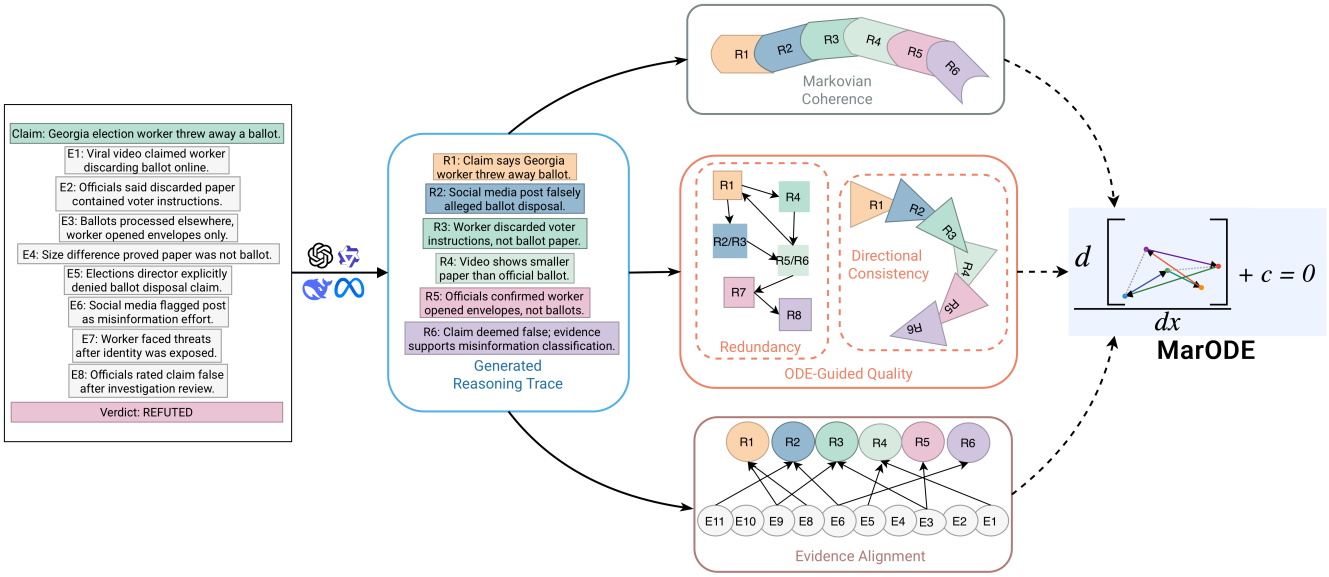


图3。 MarODE框架概览，该框架整合了三个互补的组成部分——马尔可夫连贯性、方向一致性和证据对齐，以捕捉推理轨迹完备性与可靠性的一般概念。(a) 马尔可夫连贯性 对推理的局部流动进行建模：高质量推理轨迹中的每个步骤都与前一步骤有意义地衔接，同时推动推理向后续步骤发展，并通常对信息重叠加以控制。(b) 方向一致性 确保推理沿着清晰、单向的方向推进，使每个步骤都朝结论迈进，而非反复摇摆或倒退。(c) 冗余，未能提供显著新信息的冗余步骤不仅会增加推理链的长度，还会损害其整体质量以及收敛到某个结论的能力。(d) 证据对齐 使推理步骤以支持性证据为依据，从而减少幻觉，避免其导致看似自信却错误的结论。

4.3.1 马尔可夫连贯性

一条优质的推理轨迹能够确保相邻轨迹元素之间的连接连贯。信息流应当渐进且连续，避免突兀或缺乏依据的跳跃。每个推理步骤R_i 都会嵌入到归一化的语义空间中，从而得到嵌入 $\{\mathbf{e}_i\}_{i=0}^{K}$。随后，利用余弦相似度构建转移矩阵 $\mathbf{P} \in \mathbb{R}^{K\times K}$，

$$P_{ij} = \frac{\exp\left(\cos(\mathbf{e}_i, \mathbf{e}_j)/\tau\right)}{\sum_{j'} \exp\left(\cos(\mathbf{e}_i, \mathbf{e}_{j'})/\tau\right)}, \tag{6}$$

其中， $\tau > 0$ 表示温度参数，且不考虑自转移。

这一表述受到先前研究的启发：这些研究通过在语义表示上进行随机游走来捕捉连贯性和相关性，尤其是（Ramage 等人，2009）提出的基于随机游走的语义相似度框架，以及（Petersen 等人，2015）提出的基于马尔可夫过程、针对话语实体的文档连贯性模型。我们通过在诱导出的马尔可夫链上模拟 256 次随机游走来估计推理连贯性。令 $\pi = (i_0, i1, \ldots, i_L)$ 表示一次采样游走。每次游走的得分根据单调转移进行奖励，其中 $i_{t+1} = i_t \pm 1$。为了获得质量更高的游走集合，我们确保起始状态位于推理步骤的前三分之一范围内。连贯性得分定义为

$$C(r) = \mathbb{E}_{\pi\sim\mathbf{P}}\left[\frac{1}{|\pi|-1}\sum_{t}\mathbb{I}\left(|i_{t+1} - i_t| = 1\right)\right], \tag{7}$$

通过蒙特卡洛平均进行近似。该值记为连贯性得分。

*4.3.2 ODE*引导的质量建模

通过连续时间表述，将冗余分析与方向一致性相结合，从而对全局推理质量进行建模。

冗余。每个推理步骤R_i会被分词为一组字母数字标记$\mathcal{T}(R_i)$。对于所有无序步骤对$(i,\ j)$，使用对称差计算冗余惩罚，

$$d_{ij} = \left|\mathcal{T}(R_i) \triangle \mathcal{T}(R_j)\right|.$$

惩罚按如下方式分配

$$\boxed{\rho_{ij} = \begin{cases} 1, & d_{ij} = 0, \\ \exp(-d_{ij}/4), & \text{otherwise.} \end{cases}}$$

冗余分数随后定义为

$$\boxed{R(r) = 1 - \frac{1}{|\mathscr{P}|} \sum_{(i,j)\in\mathscr{P}} \rho_{ij},} \tag{8}$$

其中$\mathscr{P}$表示上三角步骤对的集合。该表述对在各步骤间重复相同或极为相似词语的推理轨迹加以惩罚，鼓励每一步都补充新信息。因此，较高的冗余分数表明整条推理轨迹具有更高的词汇多样性、更少的重复。

通过常微分方程积分的方向一致性。方向一致性通过如下假设进行建模：推理轨迹元素沿统一方向趋向特定结论。该方向表示为潜在信念变量$p(t) \in (0,1)$的演化过程，初始化为$p(0) = 0.5$。对于每个推理步骤R_i，自然语言推理（NLI）模型会生成相对于R_{i-1}的蕴含概率与矛盾概率，从而得到归一化信号：

$$S_i = \mathrm{clip}\left(\frac{(e_i - c_i) + 1}{2}, 0, 1\right). \tag{9}$$

此处，e_i与c_i分别表示蕴含概率和矛盾概率。我们采用常微分方程（ODE）表述来计算潜在信念变量，而非使用经典概率论中的乘法概率链。受先前借助连续动力学刻画潜在状态轨迹的工作启发，我们将神经常微分方程的连续时间框架（Chen 等人， 2018年）以及神经微分方程在潜在序列动力学中的近期应用（Li 等人， 2025年）加以改造，通过对推理步信号进行常微分方程积分来定义方向一致性。以下设计选择的理据与推导见附录 A。

$$\boxed{\frac{dp}{dt} = \rho(S_i - 0.5)p(1-p),}$$

此处，$p \in [0,1]$表示不断演化的信念强度。我们通过连续动力学对信念演化进行建模，并使用稳定的龙格–库塔（RK4）方法对其在各推理步骤间进行数值积分（见附录 B）。

积分得到一个序列$\{p_i\}_{i=0}^{K}$，其中$p_{i+1} = \Phi(p_i, S_i)$，其中Φ表示数值 RK4 更新算子。最终的方向一致性得分定义为

$$\boxed{DC(r) = p_K,}$$

其中K是推理轨迹的长度。

质量聚合。整体质量得分计算如下：

$$\boxed{Q(r) = \tfrac{1}{2}R(r) + \tfrac{1}{2}DC(r),} \tag{10}$$

4.3.3 证据对齐

为评估证据依据关联，我们将每个推理步骤R_i与一组证据句$\mathscr{E}$进行对齐。对于每个步骤，根据嵌入空间中的余弦相似度选出最相似的证据句。这种证据对齐方法受到先前证据依据关联研究的启发；这些研究将推理步骤与支持句对齐，并利用自然语言推理信号优化对齐，相关方法见 ERASER（DeYoung 等，2020）以及 FEVER 风格的验证模型（Nie 等，2019年）。如果相似度超过阈值θ，则赋予一个基础对齐得分，并使用自然语言推理概率进行优化，

$$E_i = 0.5 + \lambda_e \Pr_i(\text{entail}) - \lambda_c \Pr_i(\text{contradict}), \tag{11}$$

截断到区间$[0,1]$。证据分数定义为

$$\boxed{E(r) = \frac{1}{K+1} \sum_{i=0}^{K} E_i,} \tag{12}$$

4.3.4 MarODE 最终聚合

最终的 MarODE 分数由这三个组成部分的凸组合计算得出，

$$\boxed{\text{MarODE}(r) = w_c\, C(r) + w_q\, Q(r) + w_e\, E(r)} \tag{13}$$

采用默认权重 $w_c = w_q = w_e = \frac{1}{3}$。MarODE 提供了一个统一指标，用于衡量推理连贯性、质量和证据依据。

4.4 实验设置

我们使用一种基于马尔可夫常微分方程的指标评估推理轨迹的质量，并将其与已建立的基线方法进行比较。评估依托基于合成扰动的标注，并以人工判断作为金标准，来评估所提出指标的优良性和可靠性。我们的分析涵盖多个数据集、提示方案、模型系列和评估指标。

4.4.1 可靠性与优良性的评估

我们在两种互补的设置下开展实验：

- **基于扰动的标注（生成的诊断数据集）**。我们通过对生成的推理轨迹施加受控的以人为中心的扰动（第4.2节），并根据扰动程度分配分级质量标签，构建带标签的数据集。随着扰动的数量和严重程度增加，推理轨迹的预期质量单调下降（Ribeiro 等人，2020年）。这种设置能够以可扩展且受控的方式评估指标对推理质量系统性退化的敏感度（Gardner 等人，2020年）。随后，我们使用所提出的指标和竞争性基线方法对扰动后的轨迹进行评分。我们使用 Somers' D 评估指标分数与基于扰动的质量标签之间的一致性，从而衡量指标对扰动引起的推理退化的敏感度。

- **人工标注（现有基准）**。我们使用来自成熟基准的、经过专家评估的推理轨迹，通过 Somers' D 相关性量化其与人类判断之间的一致性。我们准备了一组问题，以获得对每条推理轨迹质量的整体人类判断（图5）。这些专家标注作为外部金标准，用于评估所提出指标的优良性及其在现实世界推理评估设置中的有效性。这项评估通过将指标性能建立在专家对推理质量的人工评估之上，补充了基于合成扰动的分析（Chen 等人，2023年）。

4.4.2 上下文内监督与推理质量

为考察上下文监督对推理质量的影响，我们针对所有数据集、模型和评估指标，系统地进行了按示例数分析。在1样本、2样本和 4样本提示配置下生成推理轨迹，并在相应的原始轨迹上计算指标分数。我们通过对(4样本,1样本)、(2样本、1样本)和(4样本、2样本)配置进行成对比较，量化不同样例数设置下的趋势。我们采用 Wilcoxon 符号秩检验，评估不同样例数设置之间观察到的差异是否具有统计显著性（Dror 等人，2018）。这项分析使我们能够评估增加上下文示例数量是否会系统性地影响生成的推理轨迹质量（Brown 等人，2020），以及此类影响在不同评估指标下是否一致。我们还通过绘制不同样例数设置下生成的推理轨迹的小提琴图，进一步分析分数分布。随后，我们根据各分布的形状，考察样例数如何影响所有基线方法下生成的推理轨迹质量。

4.4.3 基线方法

我们将 MarODE 与多种成熟的基线方法进行比较，这些方法用于在离线设置下评估推理轨迹。下面简要介绍实验研究中使用的基线方法。详情见附录 C。**ROSCOE** (Golovneva 等人，2023年) 是一组可解释的无参考指标，旨在评估逐步推理的质量，重点关注语义对齐（ROSCOE-SA）、语义相似度（ROSCOE-SS）、逻辑推理（ROSCOE-LI）和语言连贯性（ROSCOE-LC），并采用理论驱动的错误类别和诊断数据集。 **ReCEval**(Prasad 等人，2023年) 从正确性和信息丰富度两个方面评估推理链，利用蕴含和基于信息的度量来量化各个推理步骤对最终答案的贡献。先前的工作（Kotonya 和 Toni，2020) 通过局部 和全局连贯性的概念评估解释质量，衡量各个推理步骤是否有充分依据，以及整个解释是否连贯。我们还纳入了一个**LLM-as-a-Judge** 基线方法：提示大语言模型直接为每条推理轨迹分配标量质量分数，该模型使用隐式学习的评估启发式进行判断。

对于基于蕴含的指标及相关的自然语言推理组件，我们在所有基线方法和 MarODE 中均使用 `deberta-xlarge-mnli`模型，因为它在自然语言推理方面表现出色。对于句子嵌入，我们在所有基线方法和 MarODE 中均使用`all-MiniLM-L6-v2` (Yin and Zhang，2024) 模型。选择该模型是因为它在语义相似度、检索和聚类任务中表现出色，同时能以较低延迟和高效推理保持高质量的嵌入。对于 LLM-as-a-Judge，我们使用 `prometheus-7b-v2.0`；与 GPT-4 等规模更大的开放模型相比，它的评估质量相当（Kim 等人，2024）。对于 ROSCOE 的语言一致性（ROSCOE-LC）组件，我们使用`gpt2-large` 和`tinybertcola` 等模型，以捕捉推理链中的词汇连贯性。所有基线方法均使用 Somers' D 相关性，在基于扰动的数据集和人工标注数据集上进行评估。

4.4.4 使用 ***Somers' D*** 进行元评估

为量化由受控扰动或专家人工评估得出的指标分数与真实标签之间的一致性，我们采用 Somers' D 相关性，沿用 ROSCOE 等先前研究的方法。Somers'D 是一种基于秩的序数关联度量，非常适用于涉及序数型真实值变量和连续型指标分数的元评估设置。Somers' D 作为一种针对序数型结果设计的非对称、基于秩的关联度量，已广泛应用于统计学和应用评估场景（Somers，1962；Newson，2002年）。

形式上，Somers' D 可表示为 Kendall 秩相关系数的归一化形式：

$$\boxed{D(Y \mid X) = \frac{\tau(X,Y)}{\tau(X,X)},}$$

其中，$\tau(X, Y)$表示序数型真实值变量X与指标分数Y之间的 Kendall 秩相关。由于$\tau(X, X)$量化了取值不相等的X值对的数量，因此 Somers' D 表示一致对与不一致对数量之差，并以真实标签不同的值对数量进行归一化。

数据和代码可用性

用于复现我们结果的源代码已公开发布于 `https://github.com/ojasiiitd/MarODE`。生成的合成数据集也已在同一 GitHub 仓库中开源。此外，我们还创建了一个 PyPI 软件包，以便于使用`MarODE`；安装和使用说明见附录 E。

作者贡献

A.N. 和 T.C. 构思了这一想法。A.N. 和 O.S. 完成了所有实验。A.N.、O.S. 和 T.C. 分析了结果并撰写了手稿。T.C. 负责监督整个项目。

利益冲突

作者不存在利益冲突。

参考文献

Agarwal, R., Singh, A., Zhang, L., Bohnet, B., Rosias, L., Chan, S., Zhang, B., Anand, A., Abbas, Z., Nova, A., Co-Reyes, J. D., Chu, E., Behbahani, F., Faust, A., and Larochelle, H. (2024). 多样本上下文学习。收录于 Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C.（编），神经信息处理系统进展，第37卷，第76930–76966页。Curran Associates, Inc.

Amidei, J., Piwek, P., and Willis, A. (2018). 自动问题生成中的评估方法，2013—2018年。收录于 Krahmer, E., Gatt, A., and Goudbeek, M.（编），第*11*届国际自然语言生成会议论文集，第307–317页，蒂尔堡大学，荷兰。计算语言学协会。

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F.等（2023年）。通义千问技术报告。*arXiv*预印本 *arXiv:2309.16609*。

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). 语言模型是少样本学习者。载于 Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. 和 Lin, H.（编），神经信息处理系统进展，第 33 卷，第 1877–1901 页。Curran Associates, Inc.

Chen, H., Brahman, F., Ren, X., Ji, Y., Choi, Y., 和 Swayamdipta, S.（2023年）。REV：自由文本理由的信息论评估。载于 Rogers, A., Boyd-Graber, J. 和 Okazaki, N.（编），第*61*届计算语言学协会年会论文集（第*1*卷：长论文），第 2007–2030 页，多伦多，加拿大。计算语言学协会。

Chen, Q., Qin, L., Liu, J., Peng, D., Guan, J., Wang, P., Hu, M., Zhou, Y., Gao, T., and Che, W.（2025年）。迈向推理时代：推理大语言模型长思维链综述。*arXiv* 预印本 *arXiv:2503.09567*。

Chen, R. T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K.（2018）。神经常微分方程。载于 Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. 编，神经信息处理系统进展，第 31 卷。Curran Associates, Inc.

Chrysostomou, G. and Aletras, N.（2021年）。利用任务特定信息提升文本分类中基于注意力的解释的忠实性。载于计算语言学协会第*59*届年会暨自然语言处理国际联合会议第*11*届会议论文集（第*1*卷：长论文），第477–488页。

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

Daheim, N., Macina, J., Kapur, M., Gurevych, I., 和 Sachan, M. (2024)。大语言模型导师对学生推理错误进行逐步验证和补救。载于 Al-Onaizan, Y., Bansal, M., 和 Chen, Y.-N. 编，*2024*年自然语言处理实证方法会议论文集，第 8386–8411 页，迈阿密，佛罗里达，美国。计算语言学协会。

Dalvi, B., Jansen, P., Tafjord, O., Xie, Z., Smith, H., Pipatanangkura, L., 和 Clark, P. (2021年)。用蕴含树解释答案。载于 Moens, M.-F., Huang, X., Specia, L., 和 Yih, S. W.-t. 编，*2021*年自然语言处理实证方法会议论文集，第 7358–7370 页，线上及多米尼加共和国蓬塔卡纳。计算语言学协会。

Dasgupta, I., Lampinen, A.K., Chan, S. C., Sheahan, H. R., Creswell, A., Kumaran, D., McClelland, J. L., andHill, F. (2022). 语言模型在推理任务中表现出类似人类的内容效应。*arXiv* 预印本 *arXiv:2207.07051*。

DeYoung, J., Jain, S., Rajani, N. F., Lehman, E., Xiong, C., Socher, R., and Wallace, B. C. (2020). ERASER：用于评估合理化自然语言处理模型的基准测试。载于 Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J.（编），第*58*届计算语言学协会年会论文集，第 4443–4458 页，线上。计算语言学协会。

Dror, R., Baumer, G., Shlomov, S., and Reichart, R. (2018). 自然语言处理中统计显著性检验指南。载于 Gurevych, I. and Miyao, Y.（编），第*56*届计算语言学协会年会论文集（第*1*卷：长论文），第 1383–1392 页，墨尔本，澳大利亚。计算语言学协会。

Fang, J., Meng, Z., and MacDonald, C. (2024). TRACE 证据：为检索增强生成构建知识支撑的推理链。载于 Al-Onaizan, Y., Bansal, M., and Chen, Y.-N.（编），计算语言学协会年会论文集：*EMNLP 2024*，第 8472–8494 页，迈阿密，佛罗里达州，美国。计算语言学协会。

Fomicheva, M. and Specia, L.（2019年）。将机器翻译评估指标推向极端：超越与人工判断的相关性。《计算语言学》，45(3):515–558。

Gardner, M., Artzi, Y., Basmov, V., Berant, J., Bogin, B., Chen, S., Dasigi, P., Dua, D., Elazar, Y., Gottumukkala, A., Gupta, N., Hajishirzi, H., Ilharco, G., Khashabi, D., Lin, K., Liu, J., Liu, N. F., Mulcaire, P., Ning, Q., Singh, S., Smith, N. A., Subramanian, S., Tsarfaty, R., Wallace, E., Zhang, A., and Zhou, B.（2020）。通过对比集评估模型的局部决策边界。载于 Cohn, T., He, Y., and Liu, Y.（编），计算语言学协会年会论文集：*EMNLP 2020*，第 1307–1323 页，线上。计算语言学协会。

Geva, M., Khashabi, D., Segal, E., Khot, T., Roth, D., 和 Berant, J.（2021年）。亚里士多德用过笔记本电脑吗？一种采用隐式推理策略的问答基准。《计算语言学协会会刊》，9：346–361。

Golovneva, O., Chen, M. P., Poff, S., Corredor, M., Zettlemoyer, L., Fazel-Zarandi, M., and Celikyilmaz, A. (2023年). ROSCOE：一套用于评估逐步推理的指标。载于第十一届国际学习表征会议。

Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Xu, H., Ding, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Chen, J., Yuan, J., Tu, J., Qiu, J., Li, J., Cai, J. L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., You, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Zhou, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R. J., Jin, R. L., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S. S., Zhou, S., Wu, S., Yun, T., Pei, T., Sun, T., Wang, T., Zeng, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Xiao, W. L., An, W., Liu, X., Wang, X., Chen, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yang, X., Li, X., Su, X., Lin, X., Li, X. Q., Jin, X., Shen, X., Chen, X., Sun, X., Wang, X., Song, X., Zhou, X., Wang, X., Shan, X., Li, Y. K., Wang, Y. Q., Wei, Y. X., Zhang, Y., Xu, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Yu, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Ou, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Xiong, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Zhu, Y. X., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Tang, Y., Zha, Y., Yan, Y., Ren, Z. Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., Zhang, Z., Hao, Z., Ma, Z., Yan, Z., Wu, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Pan, Z., Huang, Z., Xu, Z., Zhang, Z., and Zhang, Z. (2025年). DeepSeek-R1 通过强化学习激励大语言模型进行推理。《自然》，645(8081):633–638。

Hao, S., Gu, Y., Luo, H., Liu, T., Shao, X., Wang, X., Xie, S., Ma, H., Samavedhi, A., Gao, Q., Wang, Z. 和 Hu, Z.（2024）。大语言模型推理器：针对大语言模型逐步推理的新评估、工具库与分析。发表于第一届语言建模会议。

Hasani, R., Lechner, M., Amini, A., Liebenwein, L., Ray, A., Tschaikowski, M., Teschl, G. 和 Rus, D.（2022）。闭式连续时间神经网络。《自然·机器智能》，4(11):992–1003。

Jacovi, A. 和 Goldberg, Y.（2020）。迈向忠实且可解释的自然语言处理系统：我们应如何定义和评估忠实性？发表于 Jurafsky, D., Chai, J., Schluter, N. 和 Tetreault, J.（编），第*58*届计算语言学协会年会论文集，第4198–4205页，线上。计算语言学协会。

Jin, Q.、Dhingra, B.、Liu, Z.、Cohen, W.与Lu, X.（2019年）。Pubmedqa：生物医学研究问答数据集。载于*2019*年自然语言处理实证方法会议暨第*9*届自然语言处理国际联合会议论文集（*EMNLP-IJCNLP*），第2567–2577页。

Kidger, P.、Morrill, J.、Foster, J.与Lyons, T.（2020）。面向不规则时间序列的神经控制微分方程。见Larochelle, H.、Ranzato, M.、Hadsell, R.、Balcan, M.与Lin, H.编，神经信息处理系统进展，第33卷，第6696–6707页。Curran Associates, Inc.

Kim, S.、Suk, J.、Longpre, S.、Lin, B. Y.、Shin, J.、Welleck, S.、Neubig, G.、Lee, M.、Lee, K.与Seo, M.（2024）。Prometheus 2：专门用于评估其他语言模型的开源语言模型。载于*2024*自然语言处理实证方法会议论文集，第4334–4353页。

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y. 和 Iwasawa, Y.（2022）。大语言模型是零样本推理器。神经信息处理系统进展，35:22199–22213。

Kotonya, N. 和 Toni, F.（2020）。面向公共卫生主张的可解释自动事实核查。载于 Webber, B.、Cohn, T.、He, Y. 和 Liu, Y.（编），*2020* 年自然语言处理实证方法会议论文集（*EMNLP*），第 7740–7754 页，线上。计算语言学协会。

Kumar, D.、Yadav, D. 和 Patel, Y.（2025）。GPT-OSS-20B：对 OpenAI 开放权重混合专家模型的全面以部署为中心的分析。*arXiv* 预印本 *arXiv:2508.16700*。

Lan, T.、Zhou, Y.-H.、Ma, Z.-A.、Sun, F.、Sun, R.-Q.、Luo, J.、Tu, R.-C.、Huang, H.、Xu, C.、Wu, Z. 等（2025）。文本、视觉和语音生成自动评估方法综述。*arXiv* 预印本 *arXiv:2506.10019*。

Lee, J. 和 Hockenmaier, J.（2025年）。评估逐步推理轨迹：一项综述。载于 Christodoulopoulos, C.、Chakraborty, T.、Rose, C. 和 Peng, V.（编），计算语言学协会 *EMNLP 2025* 论文集，第 1789–1814 页，苏州，中国。计算语言学协会。

Li, Q.、Geng, J.、Chen, Z.、Zhu, D.、Wang, Y.、Ma, C.、Lyu, C.与Karray, F.（2025年）。Hd-ndes：用于大语言模型幻觉检测的神经微分方程。见计算语言学协会第*63*届年会论文集（第*1*卷：长论文），第6173–6186页。

Longjohn, R.、Gopalan, G.与Casleton, E.（2025年）。机器学习基准测试中聚合性能指标的统计不确定性量化。*arXiv*预印本 *arXiv:2501.04234*。Massaroli, S.、Poli, M.、Bin, M.、Park, J.、Yamashita, A.与Asama, H.（2020）。稳定神经流。*arXiv*预印本 *arXiv:2003.08063*。Moradi, M.与Samwald, M.（2021年）。评估神经语言模型对输入扰动的鲁棒性。载于Moens, M.-F.、Huang, X.、Specia, L.与Yih, S. W.-t.编，*2021*年自然语言处理实证方法会议论文集，第1558–1570页，线上及多米尼加共和国蓬塔卡纳。计算语言学协会。Murthy, A. B.、Sanneman, L.与Mink, J.（2025年）。用于对大语言模型关于行动、变化与规划之推理进行整体评估的指标。见神经信息处理系统大会*2025*年研讨会：评估不断演进的大语言模型生命周期：基准、涌现能力与扩展。

Newson, R. B.（2002年）。“非参数”统计学背后的参数：肯德尔秩相关系数、Somers’ D 和中位数差异。《*Stata*》期刊，2：45–64。

Nie, Y.、Chen, H. 和 Bansal, M.（2019年）。结合事实抽取与验证的神经语义匹配网络。载于人工智能促进协会人工智能会议论文集，第33卷，第6859–6866页。

Petersen, C.、Lioma, C.、Simonsen, J. G. 和 Larsen, B.（2015）。利用话语实体，通过熵和基于图的建模研究文档连贯性：在信息检索中的应用。载于*2015*年信息检索理论国际会议论文集，第191–200页。

Pillutla, K.、Swayamdipta, S.、Zellers, R.、Thickstun, J.、Welleck, S.、Choi, Y. 和 Harchaoui, Z.（2021年）。Mauve：利用散度前沿衡量神经文本与人类文本之间的差距。载于 Ranzato, M.、Beygelzimer, A.、Dauphin, Y.、Liang, P. 和 Vaughan, J. W.（编），神经信息处理系统进展，第34卷，第4816–4828页。Curran Associates, Inc.

Prasad, A.、Saha, S.、Zhou, X. 和 Bansal, M.（2023年）。ReCEval：通过正确性和信息量评估推理链。载于*2023*年自然语言处理实证方法会议。

Põldvere, N.、Uddin, Z. 和 Thomas, A.（2023年）。politifact-oslo 语料库：用于假新闻分析与检测的新数据集。*Information*，14（12）。

Qiu, P.、Wu, C.、Liu, S.、Fan, Y.、Zhao, W.、Chen, Z.、Gu, H.、Peng, C.、Zhang, Y.、Wang, Y. 和 Xie, W.（2025年）。量化大语言模型在临床病例上的推理能力。《自然·通讯》，16（1）：9799。

Ramage, D.、Rafferty, A. N. 和 Manning, C. D.（2009）。用于文本语义相似度的随机游走。载于*2009*年基于图的自然语言处理方法研讨会论文集，TextGraphs-4，第 23–31 页，美国。计算语言学协会。

Rei, R., Guerreiro, N. M., Treviso, M., Coheur, L., Lavie, A., 和 Martins, A.（2023年）。内情：迈向对机器翻译神经评估指标的更深入理解。载于 Rogers, A., Boyd-Graber, J., 和 Okazaki, N. 编，第61届计算语言学协会年会论文集（第2卷：短论文），第1089–1105页，多伦多，加拿大。计算语言学协会。

Ribeiro, M. T., Wu, T., Guestrin, C., 和 Singh, S.（2020）。超越准确率：使用 CheckList 对自然语言处理模型进行行为测试。载于 Jurafsky, D., Chai, J., Schluter, N., 和 Tetreault, J. 编，第58届计算语言学协会年会论文集，第4902–4912页，线上。计算语言学协会。

Soh, Y. J. 和 Zhao, J.（2024）。迈向评分器混合：现有自动评估指标的统计分析。*arXiv* 预印本 *arXiv:2410.10030*。

Somers, R. H.（1962）。一种用于序数变量的非对称关联度量。《美国社会学评论》，27:799。

Su, Q. 和 Wu, Y.-C.（2015）。关于高斯置信传播的收敛条件。信号处理汇刊，63(5):1144–1155。

Tafjord, O.、Dalvi, B. 与 Clark, P.（2021年）。ProofWriter：在自然语言上生成蕴含、证明与溯因陈述。见 Zong, C.、Xia, F.、Li, W. 与 Navigli, R. 编，计算语言学协会论文集：*ACL-IJCNLP 2021*，第3621–3634页，线上。计算语言学协会。

Turpin, M.、Michael, J.、Perez, E. 与 Bowman, S. R.（2023年）。语言模型并不总是说出它们所想的：思维链提示中的不忠实解释。见第三十七届神经信息处理系统会议。

Wang, B., Li, Z., Huang, X., Huang, X., 和 Dong, Y. (2026 年). 思维链如同一面透镜：评估人类偏好与大语言模型之间的结构化推理对齐。*arXiv* 预印本 *arXiv:2511.06168*。

Wang, B., Min, S., Deng, X., Shen, J., Wu, Y., Zettlemoyer, L., 和 Sun, H. (2023年). 探索思维链提示：关于关键因素的实证研究。载于 Rogers, A., Boyd-Graber, J., 和 Okazaki, N. 编，第*61*届计算语言学协会年会论文集（第*1*卷：长论文），第 2717–2739 页，多伦多，加拿大。计算语言学协会。

Wang, W. Y. (2017 年). “骗子，骗子，裤子着火”：用于假新闻检测的新基准数据集。载于 Barzilay, R. 和 Kan, M.-Y. 编，第*55*届计算语言学协会年会论文集（第*2*卷：短论文），第 422–426 页，温哥华，加拿大。计算语言学协会。

Wang, X., Wang, H., and Yang, D.（2022a）。衡量并提升自然语言处理模型的鲁棒性：一项综述。载于 Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V. 编，《2022 年计算语言学协会北美分会年会论文集：人类语言技术》，第 4569–4586 页，美国西雅图。计算语言学协会。

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D.（2022b）。自洽性提升了语言模型中的思维链推理。*arXiv* 预印本 *arXiv:2203.11171*。

Wei, J., Wang, X., Schuurmans, D., Bosma, M., brian ichter, Xia, F., Chi, E. H., Le, Q. V., and Zhou, D.（2022）。思维链提示能够引出大语言模型的推理能力。载于 Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. 编，神经信息处理系统进展。

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K. R. 和 Cao, Y.（2023）。ReAct：在语言模型中协同推理与行动。载于第十一届国际学习表征会议。

Yeh, A.（2000）。针对结果差异的统计显著性进行更准确的检验。载于*COLING 2000* 第*2*卷：第*18*届国际计算语言学会议。

Yin, C. 和 Zhang, Z.（2024）。微调后基于 all-minilm-l6-v2 模型研究“语义相同、结构不同”的句子相似度。载于*2024*年第二届图像、算法与人工智能国际会议（*ICIAAI 2024*）论文集，第677–684页。亚特兰蒂斯出版社。

Zhou, S., Xie, W., Li, J., Zhan, Z., Song, M., Yang, H., Espinoza, C., Welton, L., Mai, X., Jin, Y., Xu, Z., Chung, Y.-H., Xing, Y., Tsai, M.-H., Schaffer, E., Shi, Y., Liu, N., Liu, Z. 和 Zhang, R.（2025年）。自动化评估大语言模型的专家级医学推理。*npj*数字医学，9(1):34。

附录

方向一致性动力学的推导

令 $\mathscr{R}_i$ 表示序列中的第i个推理步骤。我们将连续推理步骤之间的局部蕴含概率和矛盾概率定义为

$$P_i^e = \mathbb{P}(\mathscr{R}_i \models \mathscr{R}_{i-1}), \qquad P_i^c = \mathbb{P}(\mathscr{R}_i \perp \mathscr{R}_{i-1}), \tag{14}$$

其中，$\models$和 $\perp$分别表示蕴含和矛盾。

累积蕴含　　蕴含和矛盾。截至第i 步的累积蕴含概率定义为　　s

$$E_i = \mathbb{P}(\mathscr{R}_i \models \mathscr{R}_{i-1} \models \ldots \models \mathscr{R}_0) = \prod_{k=1}^{i} P_k^e, \tag{15}$$

类似地，累积矛盾概率为

$$C_i = \mathbb{P}(\mathscr{R}_i \perp \mathscr{R}_{i-1} \perp \cdots \perp \mathscr{R}_0) = \prod_{k=1}^{i} P_k^c. \tag{16}$$

优势比表述。 我们将步骤 $\mathscr{R}_i$ 的局部优势比定义为

$$\Lambda(\mathscr{R}_i) = \frac{P_i^e}{P_i^c}. \tag{17}$$

截至步骤 i 的累积优势比为

$$O_i = \frac{E_i}{C_i} = \frac{P_i^e P_{i-1}^e \ldots P_1^e}{P_i^c P_{i-1}^c \ldots P_1^c} = \Lambda(\mathscr{R}_i)\, O_{i-1}. \tag{18}$$

取对数后得到加法形式

$$\log O_i = \log O_{i-1} + \log \Lambda(\mathscr{R}_i). \tag{19}$$

令 $L_i = \log O_i$，可得

$$L_i - L_{i-1} = \log \Lambda(\mathscr{R}_i) = \log P_i^e - \log P_i^c. \tag{20}$$

基于自然语言推理的近似 。 我们使用自然语言推理（NLI）分数来近似对数优势增量。Le　　t

$$S_i^{\text{raw}} = \text{Entail}_{\text{NLI}}(\mathscr{R}_i, \mathscr{R}_{i-1}) - \text{Contradict}_{\text{NLI}}(\mathscr{R}_i, \mathscr{R}_{i-1}), \tag{21}$$

并将其归一化到单位区间：

$$S_i = \max\left(0, \min\left(1, \frac{S_i^{\text{raw}} + 1}{2}\right)\right). \tag{22}$$

然后我们近似

$$\log \Lambda(\mathscr{R}_i) \approx c\left(S_i - \frac{1}{2}\right), \tag{23}$$

其中c是一个扩展常数。该近似保留了$\log P_i^e - \log P_i^c$的定性行为，并利用自然语言推理信号刻画蕴含相对于矛盾的相对主导性。

连续形式。　　将推理索引i 解释为连续变量，我们将更新方程重写为 a　　s

$$\frac{\partial L}{\partial i} = r\left(S_i - \frac{1}{2}\right), \tag{24}$$

其中r 是变化率常数。

方向一致性 cy 概率。我们将方向一致性（蕴含一致性）概率定义为

$$\pi_i = \sigma(L_i) = \frac{1}{1+e^{-L_i}} = \frac{E_i}{E_i + C_i}, \tag{25}$$

其中 $\sigma(\cdot)$表示sigmoid函数。

将 π_i 对i求导，得到

$$\frac{\partial \pi_i}{\partial i} = \frac{d\sigma(L_i)}{dL_i} \cdot \frac{\partial L_i}{\partial i} = \pi_i(1-\pi_i) \cdot \frac{\partial L_i}{\partial i}. \tag{26}$$

由式（24）代入，我们得到本方法所使用的控制方程：

$$\boxed{\frac{\partial \pi_i}{\partial i} = \pi_i(1-\pi_i)\, r \left(S_i - \frac{1}{2} \right)} \tag{27}$$

该方程刻画了方向一致性如何沿推理序列演化，其中 π_i 表示截至步骤i 的序列保持一致蕴含方向的概率。

方向一致性分析。为考察不同推理动态如何影响方向一致性的更新，我们模拟了七种自然语言推理（自然语言推理）信号情景，范围从始终支持到始终矛盾、振荡、翻转、逐渐漂移以及极端确定性，如图4所示。对于每一组蕴含、矛盾和中性概率序列，我们使用两种互补方法计算方向一致性轨迹：（i）通过四阶龙格–库塔方法求解的逻辑斯谛常微分方程，以及（ii）累积蕴含与矛盾似然的乘法贝叶斯式更新。所得图示展示了各情景下方向一致性如何逐步演化，突出了平滑的常微分方程动态与更尖锐的乘法更新之间的对比。基于常微分方程的表述会产生更平滑、更稳定的轨迹，尤其适用于更长或噪声更大的推理链，而乘法更新往往会放大微小的局部变化，并可能迅速变得极端。诸如神经控制微分方程之类的扩展进一步将这种连续建模视角推广到动态信号（Kidger等人，2020）。连续时间建模在近期机器学习研究中被强调为一种相较传统离散更新方法更能捕捉更平滑、更稳定时间动态的途径（Hasani等人，2022）。这一比较说明了为何基于连续常微分方程的方法更适合追踪推理进程。

扰动名称	范围	描述	针对的失效模式
倒数第二歧义	局部	在倒数第二个推理步骤中引入歧义	判决理由
时间混淆	局部	改变或错置时间指代	时间连贯性
无依据的结论	全局	引入了前面步骤未能支持的结论	证据支撑
无关扩展	局部	添加相关但流畅的内容	逻辑焦点
重复	局部	逐字重复推理步骤	冗余
描述不足	局部	省略了关键的解释性细节	步骤完整性
反义词插入	局部	向关键主张注入语义对立	语义一致性
定义冗余	局部	重复定义而不增加信息	简洁性
模态逻辑混淆	局部	混淆必然性、可能性或确定性	逻辑模态
排序	全局	重新排列推理步骤	因果结构
量词滥用	局部	歪曲全称或存在性断言	定量有效性
随机断字	表层	插入多余的断字伪影	表层噪声
关键概念替换	局部	替换核心实体或概念	概念完整性
最终判定插入	全局	将过早裁决注入推理过程	推理闭合

表 4. 本研究中使用的推理轨迹扰动函数汇总。每种扰动都针对多步推理中的一种特定失效模式，涵盖局部步骤级不一致性、全局结构性破坏和表层噪声。扰动会依次无放回地应用于每条推理轨迹，每条轨迹最多成功应用六种扰动，从而在可控的同时，以多样化的方式降低推理质量。

B 方向一致性动力学的数值积分

为了稳定地对方向一致性常微分方程进行积分

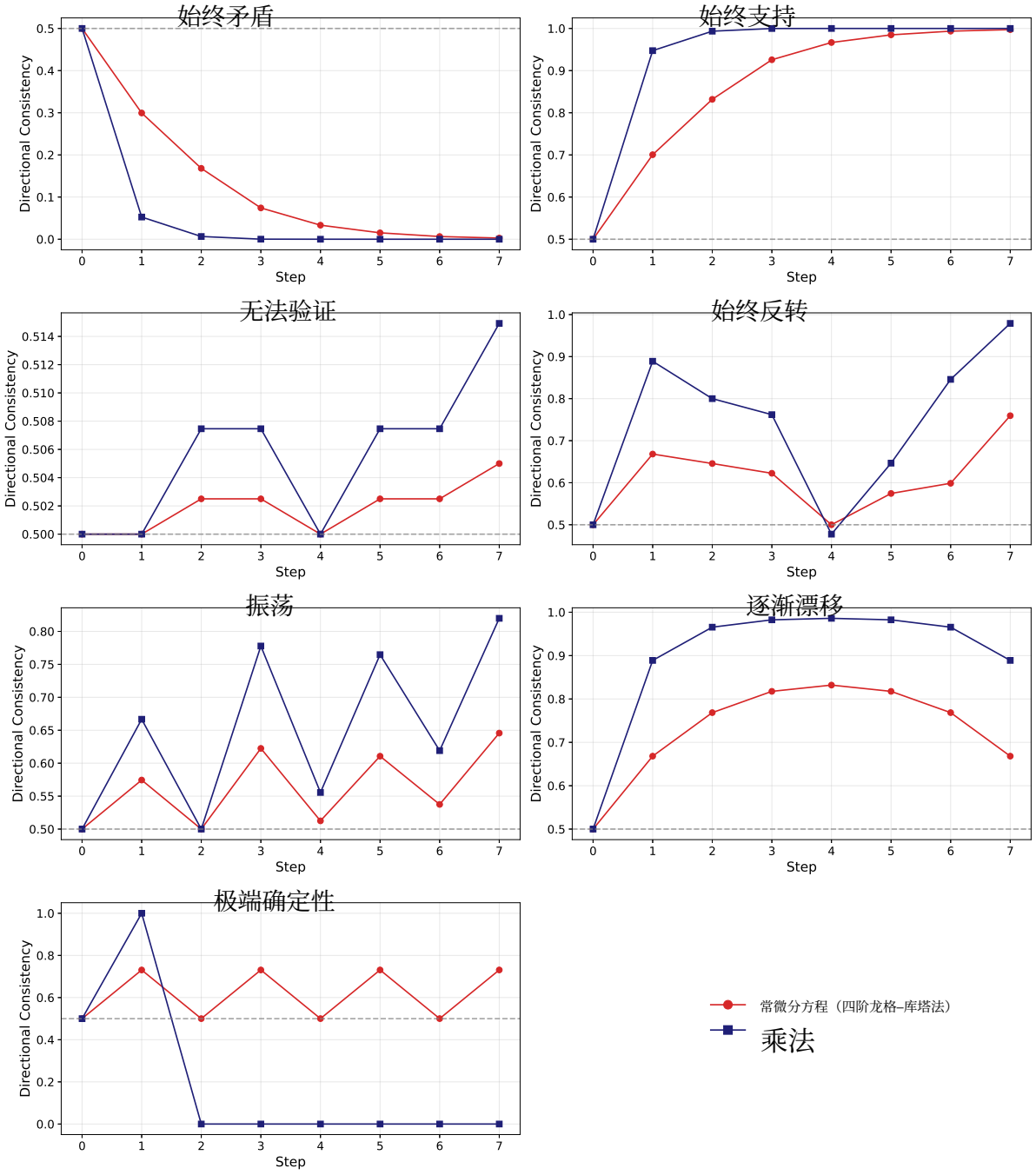


图4. 七种推理情景下的方向一致性。每个子图展示了两种更新方法下信念的演变：通过龙格–库塔法求解的逻辑斯蒂常微分方程（蓝色圆圈），以及乘法贝叶斯式更新（橙色方块）。这些情景包括始终支持、始终矛盾、无法验证、反转、振荡、逐渐漂移和极端确定性。该图突显了不同推理动态在稳定性和波动性方面的差异。

Reasoning Trace Human Evaluation Questionnaire	
Instruction: Please read the reasoning trace carefully and answer the following questions. For each question, select one value from **1 (Strongly Disagree)** to **5 (Strongly Agree)** by filling in one circle.	
Q1: Do the reasoning steps follow a logical and consistent progression without contradictions?	○ 1 ○ 2 ○ 3 ○ 4 ○ 5
Q2: Are the reasoning steps clearly expressed and easy to understand?	○ 1 ○ 2 ○ 3 ○ 4 ○ 5
Q3: Are the reasoning steps clearly expressed and easy to understand?	○ 1 ○ 2 ○ 3 ○ 4 ○ 5
Q4: Are the reasoning steps well-grounded in the provided evidence or context, without introducing unsupported claims?	○ 1 ○ 2 ○ 3 ○ 4 ○ 5
Q5: Does the final verdict logically follow from the reasoning steps?	○ 1 ○ 2 ○ 3 ○ 4 ○ 5
Q6: Considering all aspects (flow, clarity, novelty, evidence, and verdict), how would you rate the overall quality of this reasoning trace?	○ 1 ○ 2 ○ 3 ○ 4 ○ 5

图5. 用于评估推理轨迹质量的人工评估问卷。标注员使用五点李克特量表对每个问题进行评分（1 =非常不同意， 5 = 非常同意）。

$$\frac{dp}{dt} = \rho(S_i - 0.5)p(1-p),$$

我们采用经典的四阶龙格–库塔法（RK4），它在非线性动力学的数值精度与计算效率之间实现了平衡。

给定当前信念状态 p_i 和步信号 S_i，中间斜率计算如下：

$$\begin{aligned} k_1 &= f(p_i, S_i), \\ k_2 &= f\left(p_i + \tfrac{\Delta t}{2} k_1, S_i\right), \\ k_3 &= f\left(p_i + \tfrac{\Delta t}{2} k_2, S_i\right), \\ k_4 &= f\left(p_i + \Delta t\, k_3, S_i\right), \end{aligned}$$

其中

$$f(p, S_i) = \rho(S_i - 0.5)p(1-p).$$

随后更新信念状态：

$$p_{i+1} = p_i + \frac{\Delta t}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right),$$

然后将结果裁剪至单位区间。该过程在所有推理步骤中重复执行，最终得到方向一致性得分 $DC(r) = p_K$。

四阶龙格–库塔法因其四阶准确率和稳定性特性而被广泛用于积分非线性动力系统，因此适合对离散推理步骤中的平滑信念轨迹进行建模。

C 基线方法

ROSCOE（推理与逐步连贯性评估）。ROSCOE（Golovneva 等，2023年）是一个模块化框架，无需标准参考答案即可诊断思维链推理中的多种失败模式。它将推理质量分解为可解释的组件，用于刻画语义对齐、逻辑有效性和语言连贯性。ROSCOE-SA 和 ROSCOE-SS 计算相邻推理步骤之间，以及推理步骤与结论之间基于嵌入的语义相似度，以

Prompt: LLM as a Judge

Evaluate the reasoning trace using the scoring rubric below.
Assume the evidence to be true.
Scoring Rubric
1.00 (Outstanding)
0.75 (Good)
0.50 (Adequate)
0.25 (Poor)
0.00 (Critically Flawed)

CLAIM: {claim}
LABEL: {label}
EVIDENCE: {evidence_str}
REASONING TRACE: {reasoning_trace}

Provide ONLY a numerical score between 0.00 and 1.00 with exactly two decimal places. Do not include any text, explanation, or formatting.

Score:

图 6. LLM-as-a-Judge 评估范式中使用的提示模板。评估器模型会被要求评估给定推理轨迹的质量，并在连续尺度上输出单个标量分数，而不生成解释性文本。这种形式有助于实现可扩展的定性评估，但依赖于模型隐含的评估启发式。

评估主题一致性和语义演进。ROSCOE-LI 利用自然语言推理模型，对推理步骤之间的蕴含和矛盾关系进行量化，从而能够检测逻辑断裂或缺乏支持的过渡。ROSCOE-LC 使用预训练语言可接受性模型评估表层流畅度和句法连贯性。这些组件分数可以汇总为整体质量度量，同时保持对不同推理错误类别的诊断透明度。

RecEval（推理链评估）。RecEval（Prasad 等人，2023年）通过联合建模中间步骤的正确性与信息丰富度来评估推理轨迹。它使用推理步骤与最终结论之间基于蕴含的验证来估计逻辑正确性，确保中间论断得到支持。信息丰富度则衡量每个步骤在解决任务时提供新颖且不冗余信息的程度。通过整合这两个维度，RecEval 会惩罚存在逻辑缺陷或不必要重复的推理链，并将推理质量视为一个基于贡献的过程，而非单纯的语义相似度。

局部与全局连贯性指标。Local–Global Coherence（Kotonya 和 Toni，2020）通过区分逐步论证与整体叙事一致性来评估解释质量。局部连贯性评估各个推理步骤是否得到前文语境的支持，通常通过相邻步骤之间的语义相似度或蕴含关系进行判断。全局连贯性衡量完整的推理轨迹是否构成逻辑一致且目标导向的解释，并支持最终结论。这种分解有助于检测局部推理缺口和整体不一致，但通常依赖离散的成对比较，而不对累积推理动态进行建模。

LLM-as-a-Judge。 该*LLM-as-a-Judge* 范式将推理轨迹评估定义为一项由大语言模型直接执行的定性评分任务。给定一条生成的推理轨迹，评估器模型会被提示为其分配一个标量质量分数。该方法不依赖明确的符号指标，而是利用现代大语言模型隐含的推理和评估能力，大规模近似人类判断。

如图 6所示，评估提示将输出限制为单个连续数值分数，从而提升一致性，并便于对大型轨迹集合进行聚合。尽管这一范式在实践中行之有效，但仍容易受到提示措辞、评估模型偏差和领域迁移的影响；此外，由于依赖潜在评估启发式，而非明确的推理标准，其透明度也有限。

D 推理轨迹生成的提示设计

为生成结构化且以证据为依据的推理轨迹，我们采用标准化提示模板，明确约束模型响应的逻辑流程和输出格式（图 7）。该提示将模型设定为一名分析员，要求其重构逐步论证，将事实主张及其相关证据文本与给定的真实判定标签联系起来，同时不得依赖外部知识或假设。

每个提示实例由三个部分组成：(i) 一项事实主张，(ii) 从事实核查基准中检索到的一组证据片段，以及 (iii) 相应的真实性标签。模型收到的指令是生成一条推理轨迹：以重述初始主张（R_0）开头，接着给出长度可变、以证据为依据的一系列推理步骤（从 R_1 到 R_K），最后以终止性的最终判定作结，并明确重述给定标签。

为确保提取过程具有确定性且结构有效，推理片段使用显式标记 <Rstart>和 <Rend>进行界定。提示中嵌入的少样本示例展示了证据归因、矛盾消解，以及朝最终判定推进的逻辑过程。中间步骤的数量不受限制，因此轨迹长度可根据主张复杂度灵活调整。

E MarODE 安装

1. 安装软件包：

```
pip install marode
```

2. 导入并初始化评估器：

```
from marode.evaluator import MarODEEvaluator, EvaluatorConfig, get_device

config = EvaluatorConfig() device =get_device(-1) # -1 表示 CPU,
0 表示 GPU evaluator =MarODEEvaluator(config, device)
```

3. 准备一条包含推理轨迹和证据的条目：

```
条目 ={ "id": "示例_1", "推理_轨迹": "R0: ...\nR1: ...\nR2: ...", "证据
_文本": ["证据 1", "证据 2"]}
```

4. 为该条目评分，并仅打印分数：

```
已评分_条目 = 评估器.score_条目(entry) 分数 = 已
评分_条目["ourmetric"]for k, v in scores.items():
print(f"{k}: {v:.4f}")
```

-

Prompt: Reasoning Trace Generation

You are an analyst. Your task is to generate a clear, logically structured reasoning trace that explains and justifies a given verdict for a factual claim. You will be provided with a factual claim, a set of evidence texts, and a verdict label (Supported, Refuted, Partially Supported, Misleading, or Unverifiable). Your task is to reconstruct the reasoning that connects the provided evidence to the final verdict, without relying on any external knowledge or assumptions. The number of reasoning steps (R1…Rk) can vary based on complexity.

Output Format (use exactly):
<Rstart>
R0: [Interpret or restate the claim in clearer terms]
R1: [First reasoning step grounded in evidence, building toward the given verdict]
R2: [Second reasoning step grounded in evidence, continuing the logical path]
...
Rk: [Final reasoning step that solidifies the justification for the verdict]
Final Verdict: [Restate the provided verdict label here]
<Rend>

Here are some examples for the task:
Example 1
Input
Claim: " Georgia election worker threw away a ballot"
Label: "REFUTED"
Evidence Text: [
    "Viral video claimed worker discarding ballot online.",
    "Officials said discarded paper contained voter instructions.",
    "Ballots processed elsewhere, worker opened envelopes only.",
    "Size difference proved paper was not ballot.",
    "Elections director explicitly denied ballot disposal claim.",
    "Social media flagged post as misinformation effort.",
    "Worker faced threats after identity was exposed.",
    "Officials rated claim false after investigation review."
]
Output
<Rstart>
R0: Claim says Georgia worker threw away ballot.
R2: Social media post falsely alleged ballot disposal.
R3: Worker discarded voter instructions, not ballot paper.
R4: Video shows smaller paper than official ballot.
R5: Officials confirmed worker opened envelopes, not ballots.
R6: Claim deemed false; evidence supports misinformation classification.
<Rend>

Task: {{ test claim, label and evidence texts }}

Please provide the reasoning traces. Ensure that the reasoning output is between <Rstart> and <Rend>.

图 7. 用于推理轨迹生成的标准化提示模板。提示中提供事实主张、相关证据文本和真值裁决标签，并将模型输出限制为由显式标记界定的结构化多步推理格式。少样本示例引导证据支撑和逻辑推进，最终得出裁决，从而实现对生成的推理轨迹的确定性提取和结构验证。