arXiv:2602.01208v1 [cs.CL] 1 Feb 2026

Chronos：学习推理链的时间动态以实现测试时扩展

张凯*，**廖佳怡***，**Chengpeng Li**、**谢子渊**、**李思航**†，**Xiang Wang**†中国科学技术大学 {kaizhang99, xzy101}@mail.ustc.edu.cn, {joyliao7777, sihang0520, xiangwang1223}@gmail.com

摘要

测试时扩展（TTS）已成为提升大语言模型（LLM）推理性能的一种有效范式。然而，现有方法——尤其是多数投票和启发式令牌级评分——将推理轨迹或令牌一视同仁，因此容易受到轨迹质量显著差异和局部逻辑失效的影响。本文提出**Chronos**，一种轻量、即插即用的时序推理评分器，它将每条轨迹建模为时间序列。具体而言，Chronos 学习捕捉令牌概率的轨迹特征，并据此分配质量分数，同时采用加权投票机制。在域内和域外基准测试上的大量评估表明，Chronos 在多种模型上持续带来显著提升，而计算开销可以忽略不计。值得注意的是，在使用 Qwen3-4B-Thinking-2507 的 HMMT25 测试中，Chronos@128 相较于 Pass@1 和 Maj@128 分别提升了 34.21% 和 22.70%，凸显了其有效性。

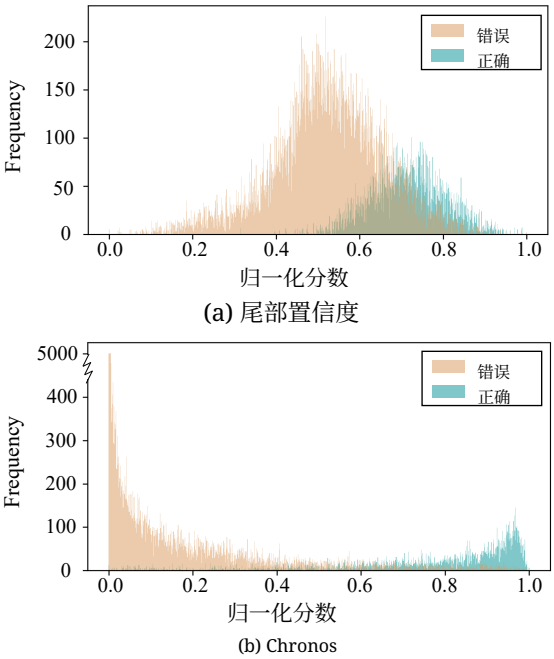


图 1：AIME25 上正确与错误轨迹的得分分布。(a) 尾部置信度和 (b) Chronos。所有实验均使用 DeepSeek-1.5B，每道题重复 16 次，每次采样 128 个样本。

1 引言

测试时扩展 (TTS) (Wu 等人，2025年；Brown 等人，2024年）已成为一种强大的范式，可与模型训练相辅相成，显著提升大语言模型（LLM）的性能（OpenAI，2025年；Yang等人，2025年；Gemini-Team，2025年；DeepSeek-AI，2024年），适用于广泛的复杂推理任务（Kamalloo 等人，2023年；Li 等人，2025b；Chen 等人，2025年）。这些性能提升的一个关键驱动因素是采用并行多样本聚合策略：生成多个相互独立的推理轨迹，并将其结合起来推断共识解答。

早期的自洽性方法通过多数投票来汇总最终答案（Wang 等人，2023年）。尽管这种方法有效，却存在一个根本局限：它将所有推理轨迹视为同等可靠，忽视了它们在质量上的显著差异。因此，错误或低质量轨迹可能会对汇总结果产生不成比例的影响，导致预测效果欠佳。这一局限引出了一个重要的研究问题：如何准确评估推理轨迹的质量？

近期研究（Fadeeva 等人，2024年；Kang 等人，2025年；Fu 等人，2025年）利用词元级分布统计来评估推理轨迹的质量。然而，这些方法很大程度上依赖预定义的启发式指标——例如词元级不确定性（Fadeeva 等人，2024年）或置信度分数（Fu 等人，2025年）——并通过均匀平均池化将其汇总为单一的轨迹质量估计。这种设计隐含地假设

*贡献相同。
†通讯作者。

所有词元均无序，且对推理轨迹的逻辑有效性贡献相同。在实践中，这种同质化可能掩盖中间步骤中的关键失误，妨碍检测局部推理错误，并显著削弱判别能力。如图1所示，尽管尾部置信度分数（Fu 等人，2025年）能够区分正确与错误的推理轨迹，但两者仍存在显著重叠，造成歧义，导致轨迹质量估计欠佳。

为克服上述局限，我们提出 Chronos（Chronological Reasoning Scorer，时序推理评分器），这是一种用于高保真轨迹聚合的通用即插即用模块。与将推理轨迹视为无序 token 级统计量集合的启发式评分器（Fu et al.，2025）不同，我们认为，推理链的有效性本质上是一个时序过程。Chronos 将每个 token 映射到一个离散时间戳，并显式建模推理的方向性推进，使评分器能够捕捉序列依赖关系。我们将 Chronos 实现为一种轻量级的学习式时序评分模型，把推理轨迹视为时间序列。该模型接收 token 级概率序列（反映模型的内部信号），并为每条推理轨迹生成一个标量质量分数。具体而言，Chronos 采用多尺度卷积架构来处理内部时序信号。通过使用具有不同卷积核长度的并行卷积滤波器，模型能够同时捕捉推理过程中的局部波动和全局依赖。这些多尺度模块堆叠在深度残差框架中，以捕捉复杂的时间模式，从而实现对轨迹质量的高保真估计。

我们在 AIME（2000–2023）上训练 Chronos，并在多个推理基准测试——AIME25、HMMT25 和 GPQA-Diamond——以及多种模型规模上进行域内和域外评估，其中包括 DeepSeek-1.5B、Qwen3-4B 和 DeepSeek-8B。在所有设置下，Chronos 都能持续显著提升测试时扩展（TTS）性能，同时仅带来可忽略的计算开销。值得注意的是，与 Qwen3-4B-Thinking 搭配时，Chronos@128 在 HMMT25 上达到 74.38% 的准确率，实际上已接近该基准测试的饱和水平；相比之下，Pass@1 的准确率为 55.42%，Maj@128（多数投票）的准确率为 60.62%，而推理 FLOPs 仅增加 0.0005%。

2 相关工作

扩展测试时计算。大语言模型的近期进展（Wu 等，2025；Brown 等，2024；Snell 等，2025）着重强调了沿两种互补范式扩展测试时计算：序列精炼与并行探索。序列扩展旨在通过迭代精炼解答（Madaan 等，2023），或通过强化学习鼓励生成更长的思维链（CoT）轨迹（Wei 等，2022；Jaech 等，2024；DeepSeek-AI，2024；Yang 等，2025），来深化推理。相比之下，并行扩展则通过多轨迹采样（Wang 等，2023；Li 等，2025a）或结构化树搜索方法（Yao 等，2023；Xie 等，2024）拓宽搜索空间。尽管这些方法行之有效，但一个关键瓶颈在于如何综合并行输出。现有方法主要依赖预定义启发式规则（Huang 等，2024；Knappe 等，2024；Kang 等，2025；Fu 等，2025），难以区分正确解答与幻觉，因为它们要么一视同仁地处理所有推理步骤，要么采用基于规则的聚合方式（例如均匀均值池化），忽略了推理过程的时间动态。为解决这一局限，我们将推理建模为时序过程，并利用内部信号准确估计轨迹质量。

时间序列建模。序列数据建模是多个领域的核心，其发展历程从传统的基于距离的启发式方法（Berndt 和 Clifford，1994年；Lin 等人，2003年）显著演进至现代深度学习方法。为捕捉复杂的非线性依赖关系和长期相关性，研究人员探索了多种架构，包括基于多层感知机的混合器（Oreshkin 等人，2020年；Zeng 等人，2023年）、循环网络（Lai 等人，2018年；Flunkert 等人，2017年）、时序卷积网络（TCNs）（Bai 等人，2018年；Liu 等人，2022年）以及 Transformer（Li 等人，2019年；Zhou 等人，2021年）。受这些进展启发，我们将大语言模型推理链的评估重新定义为时间序列分类任务。我们将词元概率序列视为时间信号，并提出一种基于 InceptionTime 架构的轻量级聚合器（Szegedy 等人，2015年；Fawaz 等人，2020年）。通过利用多尺度卷积滤波器，我们的方法能够有效捕捉推理过程的时间演变，

实现高保真度的轨迹评分，同时几乎不增加计算开销。

3 方法

本节介绍**Chronos**，这是一种轻量、即插即用的时间顺序推理评分器，将推理轨迹建模为时间过程（参见图2）。我们首先在第3.1节介绍作为 Chronos 时间信号的令牌级统计量。随后在第3.2节介绍 Chronos 架构，详细说明其如何捕捉推理过程中的序列依赖。最后，在第3.3节介绍加权多数投票方案，该方案利用预测分数进行聚合。

3.1 令牌级统计量

基于 Transformer 的（Vaswani 等人，2017）大语言模型以自回归方式生成输出词元序列 $y = (y_1, \ldots, y_m)$，并以输入序列$x = (x_1, \ldots, x_n)$为条件。在每个解码步骤 t中，模型将前置上下文映射为一个 logit 向量$\mathbf{v}_t \in \mathbb{R}^{|\mathcal{V}|}$，其中 $\mathcal{V}$ 表示词表。这些 logit 经过 softmax 算子转换为下一词元上的概率分布： y_t:

$$P_t(\cdot|x, y_{<t}) = \text{Softmax}(\mathbf{v}_t) \in [0,1]^{|\mathcal{V}|}. \quad (1)$$

这种词元级分布反映了模型对下一词元预测的置信度。

不同于先前的方法通过均匀池化将词元级信号压缩为单个标量，Chronos 保留推理轨迹的词元级概率，并将其作为时序序列$\mathbf{s}$，从而保留细粒度的序列依赖关系：

$$\mathbf{s} = (s_1, s_2, \ldots, s_L), \quad (2)$$

$$s_t = -\frac{1}{k}\sum_{i=1}^{k} \log P_t(i \mid x, y_{<t}) \quad (3)$$

其中， s_t 表示解码步骤 t中排名前k 的候选词元的平均负对数概率。 s_t 的值越高，分布越集中，模型置信度也越高；值越低则表示不确定性增加。这种时序表示保留了推理过程的时间结构，使 Chronos 能够捕捉到被均匀池化策略掩盖的局部置信度波动。

3.2 模型架构

受时间信号处理领域先前工作的启发（Szegedy 等人，2015；Fawaz 等人，2020），这些工作表明，轻量级卷积架构能够有效捕捉时间模式。Chronos 对整个推理过程中 token 级统计量的动态演变进行建模，从而能够准确评估轨迹质量。

3.2.1 多尺度特征提取

核心是多尺度卷积块，旨在以不同窗口处理时间信号$\mathbf{s}$ 。受关键推理步骤和最终答案通常集中在轨迹最后阶段这一观察的启发，Chronos 仅关注最后L_{tail} 个词元。形式上，我们将输入定义为时间序列$\mathbf{s} \in \mathbb{R}^{1\times L_{tail}}$，其中通道维度为 1，序列长度为 L_{tail}。为增强模型容量，我们首先应用形状为 1×1 的 N_{Proj} 个滤波器，将其投影到$\mathbf{z}$:

$$\mathbf{z} = \text{Conv}(\mathbf{s}, 1, N_{Proj}) \in \mathbb{R}^{N_{Proj}\times L_{tail}}, \quad (4)$$

其中，Conv(·) 表示使用形状为 1×1的 N_{Proj} 滤波器执行的卷积操作。随后，扩展后的信号将通过卷积核长度为 l 的并行卷积滤波器进行处理:

$$\mathbf{h}_l = \text{ReLU}(\text{Conv}(\mathbf{z}, l, N_{Conv})) \in \mathbb{R}^{N_{Conv}\times L_{tail}}. \quad (5)$$

该过程涉及使用形状为 $1 \times l$ 的 N_{Conv} 滤波器执行卷积操作。具体而言，较短的滤波器能够感知推理步骤中的局部波动和即时不一致性，而较长的滤波器则能够捕捉推理轨迹中的更广泛趋势和长期依赖关系。这些并行卷积的输出，包括 $\mathbf{z}$ ——涵盖局部细节和全局上下文——沿通道维度进行拼接，以形成丰富的多尺度表示 $\mathbf{o} \in \mathbb{R}^{(N_{Proj}+k\cdot N_{Conv})\times L_{tail}}$:

$$\mathbf{o} = \text{Concat}(\mathbf{z}, \mathbf{h}_{l_1}, \mathbf{h}_{l_2}, \ldots, \mathbf{h}_{l_k}), \quad (6)$$

其中， k 是不同卷积核长度的数量，$\mathbf{h}_{l_i}$ 表示由卷积核长度为 l_i 的卷积生成的输出特征图。

3.2.2 深度残差架构

为便于训练能够捕获长上下文时间模式的深度网络，我们按顺序堆叠 N_{Blk} 个多尺度卷积块，并为每个块集成残差连接，以缓解梯度消失问题。形式上，令 $\mathbf{M}_i(\cdot)$ 表示

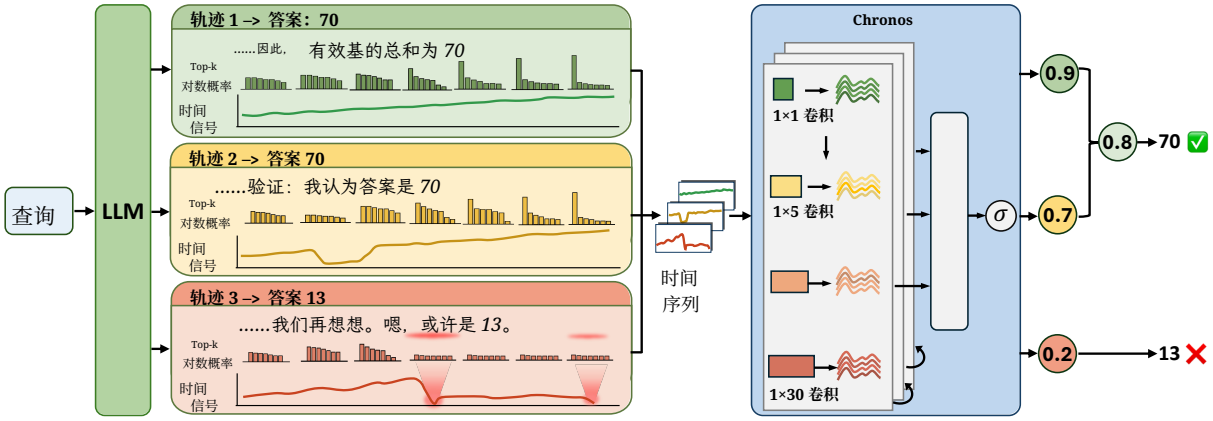


图 2：**Chronos** 的框架。它由三个阶段组成：(1) **多轨迹采样**：给定一个输入查询，我们采样多条相互独立的推理轨迹，并提取其词元级概率分布，将其视为时序信号。(2) **时序推理评分**：Chronos 使用多尺度卷积处理这些时序信号，以显式捕捉序列依赖，并为每条轨迹生成一个标量质量分数。(3) **加权多数投票**：利用预测的轨迹分数为候选答案加权，再进行汇总以确定最终输出。

i 个卷积块，$\mathbf{o}_i$ 表示$\mathbf{M}_i$的输出。前向过程定义如下：

$$\mathbf{o}_i = \mathbf{M}_i(\mathbf{o}_{i-1}),\ \hat{y} = \sigma(\text{MLP}(\sum_{i=0}^{N_{Blk}} \mathbf{o}_i)), \quad (7)$$

其中，$\mathbf{o}_0 = \mathbf{z}$ 是扩展后的输入嵌入，$\sigma(\cdot)$ 是 Sigmoid 函数，$\hat{y}$ 表示最终预测分数。对于数据集 $\{\tau_i\}_{i=1}^{N}$ 中给定的轨迹 τ_i，我们根据最终答案的正确性赋予真实标签 $y_i \in \{0, 1\}$，其中 $y_i = 1$ 表示预测正确，$y_i = 0$ 表示预测错误。模型通过最小化预测分数 $\hat{y}_i$ 与标签 y_i 之间的二元交叉熵（BCE）损失进行训练：

$$\mathcal{L} = -\sum_{i=1}^{N} [y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)]. \quad (8)$$

在实践中，我们采用集成策略，由多个训练良好的模型独立评估轨迹，并以其平均值作为最终得分。如图1所示，通过将推理轨迹建模为时间序列，Chronos 有效减少了轨迹之间的重叠歧义，并显著提高了轨迹质量预测的准确率。

3.3 分数加权多数投票

为了进一步提升聚合性能，我们实现了基于分数的筛选（Fu 等，2025年），并结合加权多数投票。形式上，给定问题 Q，我们采样 N_{smp} 推理轨迹 $\{\tau_i\}_{i=1}^{N_{smp}}$ 并预测其对应的分数 $\{s_i\}_{i=1}^{N_{smp}}$。我们按 s_i 降序排列轨迹，并确定索引子集 $\mathcal{I}_{top}$：

$$\mathcal{I}_{top} = \{i \mid \text{rank}(\hat{y}_i) \leq \lfloor \eta \cdot N_{sam} \rfloor\}, \quad (9)$$

其中，η 为保留比例。随后，通过对筛选后的子集进行加权多数投票，确定最终答案 $\hat{a}$：

$$\hat{a} = \arg\max_{a \in \mathcal{A}} \sum_{i \in \mathcal{I}_{top}} \hat{y}_i \cdot \mathbb{I}(a_i = a), \quad (10)$$

其中，$\mathcal{A}$ 表示唯一候选答案的集合，Q表示从轨迹 τ_i中提取的最终答案，$\mathbb{I}(\cdot)$为指示函数：若 a_i 与候选答案a匹配，则取值为 1，否则取值为 0。该方法根据预测得分，仅保留排名前 η 的轨迹，从而优先选择高质量的推理轨迹，确保最终共识仅基于最可靠的轨迹。

4 实验

我们开展了大量实验，以证明 Chronos 的有效性。具体而言，我们的实验旨在回答以下研究问题：

- **研究问题1：** Chronos 与现有 TTS 方法相比表现如何？
- **研究问题2：** Chronos 引入了多少计算开销？随着测试时计算规模的扩大，准确率如何变化？

• **RQ3：** 将 Chronos 应用于不同的采样模型时，其泛化表现如何?

4.1 实验设置

模型。 我们在多种规模的模型上评估 Chronos，包括： **DeepSeek-1.5B**[1]，**Qwen3-4B**[2] 和 **DeepSeek-8B**[3]。我们选择这些模型，是因为它们在推理任务中已证明具有良好的效果。完整的生成设置和提示模板见附录A。

数据集。 为防止数据泄漏，我们仅使用 2000 年至 2023 年的 AIME 档案训练 Chronos，确保训练数据的难度与评估基准测试相当。我们在 AIME25、HMMT25（2月）和 GPQA-Diamond 上评估 Chronos；这些都是广泛用于评估前沿大语言模型推理能力的基准测试。值得注意的是，AIME 和 HMMT 包含竞赛数学题，而 GPQA-Diamond 则包含科学推理问题，因此能够严格评估 Chronos 的跨领域泛化能力。更多数据集详情见附录B。

评估。我们采用 Pass@1 和多数投票作为基础评估指标。Pass@1 衡量单条推理链的准确率，而 Maj@K 则通过选择出现频率最高的答案，汇总 K 条并行推理路径的结果，并将所有轨迹视为同等可靠。此外，我们还纳入 DeepConf（Fu 等人，2025年）作为具有代表性的加权投票基线。DeepConf 根据从词元级概率分布中估算的统计置信度为轨迹分配权重，这与标准多数投票采用的均匀加权不同。按照官方实现，我们报告在不同 top-η 过滤设置下，均值、底部和尾部置信度变体所取得的最佳性能。

实现细节。对于 AIME（2000–2023）数据集，我们为每道题目采样 32 条完整的推理轨迹，并按 8:1:1 的比例随机划分为训练集、验证集和测试集。在实验设置中，我们将输入序列长度固定为 $L_{tail} = 2048$，将多尺度卷积块的数量设为 $N_{Blk} = 3$，并将保留比例设为 $\eta = 0.1$。我们对其余组件进行超参数搜索：投影维度 N_{Proj} 从 {8, 16} 中选取，滤波器数量 N_{Conv} 在 {4, 8, 16} 中搜索，不同卷积核长度的数量为 $k = 3$，多尺度卷积滤波器长度集合 l 从 {{10, 20, 40}, {20, 40, 80}, {40, 80, 160}} 中选取。

训练 Chronos 后，我们选择在测试集上取得最高曲线下面积的超参数配置，作为最终评估的评分器。在评估阶段，我们首先为每个问题生成一个包含 512 条完整轨迹的候选池。在每项实验中，我们从该候选池中为每个问题抽取 128 条轨迹，以应用分数加权多数投票方法。

4.2 主要结果（研究问题1）

表格1 总结了 Chronos 与 Pass@1、多数投票以及置信度加权基线 DeepConf 在三个模型和三个基准测试上的性能对比。

我们在下文重点介绍主要观察结果：

优于多数投票。如表 1 所示，在所有实验设置中，Chronos 的表现始终优于标准多数投票（Maj@128）。虽然与 Pass@1 相比，多数投票能有效降低方差，但它将所有推理路径一视同仁，因此当正确答案占少数时往往会失效。Chronos 则通过分配能够反映质量的分数来解决这一问题。例如，在 DeepSeek-1.5B 模型上，Chronos 将平均准确率从 29.38%（Maj@128）提升至 37.09%，绝对提升幅度高达 7.71%。同样，在更大的 DeepSeek-8B 模型上，Chronos 将平均准确率从 73.54% 提升至 78.68%。这些结果表明，识别并加权高质量轨迹，显著优于简单的基于频率的聚合。

相对于均匀池化的令牌级统计量的优势。Chronos 在全部九种模型-数据集组合上始终优于 DeepConf@128。DeepConf 通过均匀均值池化聚合令牌级统计量，将推理过程视为令牌概率的集合。相比之下，Chronos 将每条推理轨迹建模为一个时间过程，并利用内部时间动态来评估正确性。这种持续的性能差距——例如，

[1]DeepSeek-1.5B：https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B。[2]Qwen3-4B：https://huggingface.co/Qwen/Qwen3-4B-Thinking-2507。[3]DeepSeek-8B：https://huggingface.co/deepseek-ai/DeepSeek-R1-0528-Qwen3-8B。

表 1：性能比较。报告准确率（%）。所有实验均重复 16 次。排名前两位的结果以**粗体**和下划线标出。

模型	数据集	Pass@1	Maj@128	DeepConf@128	Chronos@128
DeepSeek-1.5B	AIME25	24.58	37.50	<u>40.21</u>	**44.17**
	HMMT25	12.08	16.67	<u>25.21</u>	**27.92**
	GPQA-D	21.67	33.96	<u>36.88</u>	**39.17**
	平均值	19.44	29.38	<u>34.10</u>	**37.09**
Qwen3-4B	AIME25	78.67	**86.67**	85.83	**86.67**
	HMMT25	55.42	60.62	<u>67.08</u>	**74.38**
	GPQA-D	59.79	<u>70.83</u>	68.96	**71.25**
	平均值	63.96	72.71	<u>73.96</u>	**77.43**
DeepSeek-8B	AIME25	75.21	83.96	<u>86.46</u>	**88.75**
	HMMT25	60.42	68.75	<u>73.13</u>	**76.04**
	GPQA-D	59.17	67.92	<u>70.21</u>	**71.25**
	平均值	64.93	73.54	<u>76.60</u>	**78.68**

在 DeepSeek-1.5B 上从 34.10%（DeepConf）提升至 37.09%（Chronos）——支持了我们的假设：逻辑一致性的时间演化提供了区分正确推理与幻觉的关键判别信号，而这是均匀池化无法捕捉的。

在具有挑战性的基准测试上取得显著提升。性能提升在复杂推理任务上尤为显著。在使用 Qwen3-4B 的 HMMT25 基准测试中，Chronos 取得了高达 74.38% 的准确率，而 Maj@128 为 60.62%，DeepConf 为 67.08%。相较于多数投票，这 13.76% 的绝对提升凸显了 Chronos 即使在正确解于样本池中统计上代表性不足时仍能识别它们的能力。此外，Chronos 在域外 GPQA-D 基准测试上展现出强大的泛化能力，在所有模型规模上均持续优于基线方法，证明了其在数学领域之外的稳健性。

4.3 扩展实验（研究问题2）

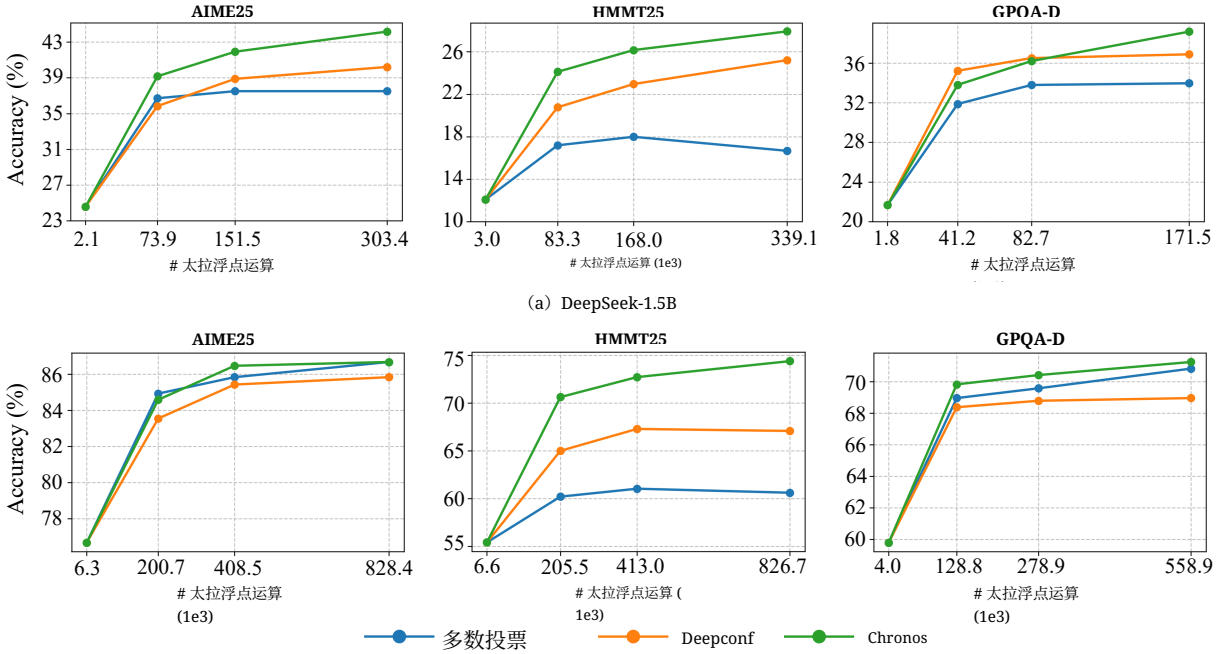
我们分析 Chronos 的计算效率及其在测试时计算预算增加时的稳定性。分析重点关注两个方面：（1）Chronos 评分器引入的额外计算开销；（2）性能随采样轨迹数量变化的趋势。如图3所示，我们考察计算成本与准确率之间的权衡，并揭示了有关 Chronos 效率和可扩展性的两个关键发现：

对测试时扩展具有很强的适应性。Chronos 展现出非凡的适应能力能够适应测试时扩展策略。随着采样轨迹数量的增加，Chronos 的性能持续提升。值得注意的是，在评估的计算预算范围内（每个问题从 1 条到 128 条轨迹），它并未表现出明显的收益递减（Chen 等人，2024年），这表明随着搜索空间扩大，其性能仍有很大的提升潜力。这种稳健的扩展行为表明，Chronos 能有效利用额外的计算资源来甄别高质量推理路径，并为正确答案赋予更高权重。

计算开销微乎其微。重要的是，相较于生成式推理过程，Chronos 几乎不会引入额外的计算开销。Chronos 处理单通道时间信号，并聚焦于关键尾部片段的固定窗口（即最后L_{tail}个词元），而非整个序列。平均而言，Chronos 对一批 30 个查询进行一次前向传播需要 3.9 BFLOPs。相比之下，同一批查询的 DeepSeek-1.5B 推理约消耗 2,000 TFLOPs。因此，Chronos 仅使推理 FLOPs 总量增加 0.0005%，这确保了观察到的扩展趋势几乎完全由轨迹采样而非评分器开销驱动。

4.4 跨模型泛化测试（RQ3）

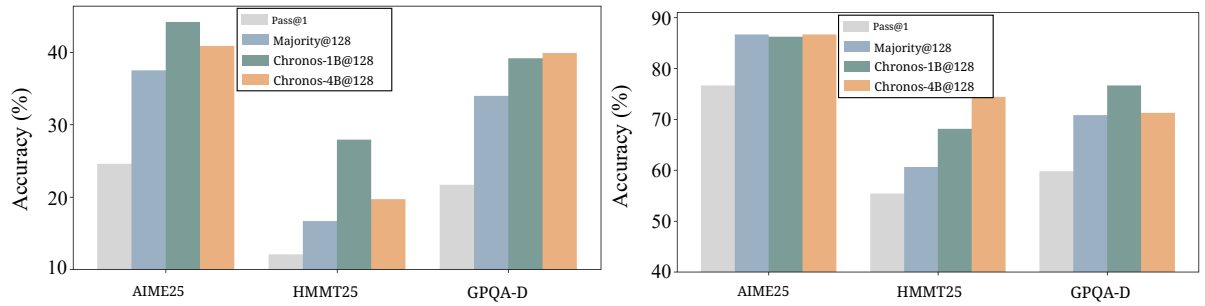
本节将考察 Chronos 是否捕捉到了模型无关的推理有效性时间模式，还是主要过拟合于特定采样模型的分布特征。为评估跨模型泛化能力，我们将进行交叉评估，使用两个独立训练的评分器：



(a) DeepSeek-1.5B

(b) Qwen3-4B

图3：在AIME25、HMMT25和GPQA-D基准测试上的测试时扩展性能。x轴表示推理计算预算（#太拉浮点运算），分别对应每个问题1、32、64、128条采样轨迹。



(a) 在DeepSeek-1.5B上的评估

(b) 在Qwen3-4B上的评估

图4：Chronos的跨模型泛化测试。

Chronos-1B 仅使用由 DeepSeek-1.5B 生成的轨迹进行训练，而**Chronos-4B**则使用来自 Qwen3-4B 的轨迹进行训练。随后，将每个评分器应用于由另一个模型生成的测试轨迹。

如图 4所示，Chronos 展现出强大的跨模型泛化能力。例如，将 Chronos-4B 用于为 DeepSeek-1.5B 生成的轨迹评分时，其性能依然极具竞争力，显著优于 Majority@128。诚然，由于不同大语言模型具有独特的内在推理逻辑和词元级概率分布，与原生域内评分器相比，受分布偏移影响，性能略有下降。然而，结果表明，有效推理链的时间信号在不同模型间总体上具有较强的鲁棒性，而 Chronos 成功捕捉到了底层不变量。它有效学习这些模型无关模式，使训练良好的评分器无需重新训练即可接入不同的大语言模型，凸显了其作为测试时扩展通用模块的潜力。

4.5 进一步分析

我们进一步开展超参数分析并可视化分数分布，以更好地理解 Chronos 的行为。受篇幅限制，更多案例研究将在附录 B 中介绍。

超参数实验。 我们考察 Chronos 对投影维度的敏感性N_{Proj}、卷积滤波器数量 N_{Conv}和卷积核长度 l，如图 5所示。在 DeepSeek-1.5B 轨迹上，较低的投影维度（$N_{Proj} = 8$）与较多的滤波器（$N_{Conv} = 12$）相结合时可获得最佳性

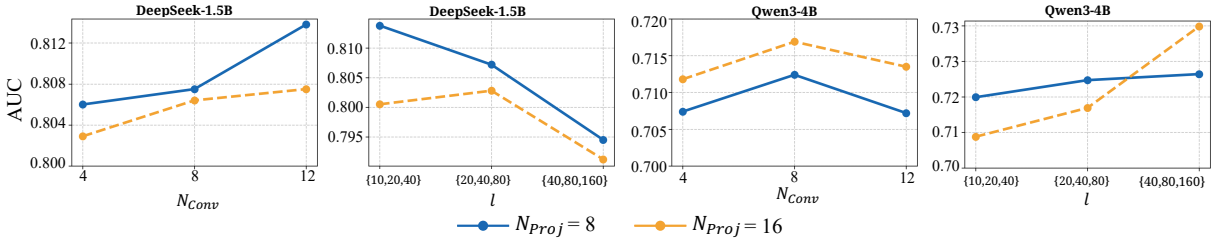
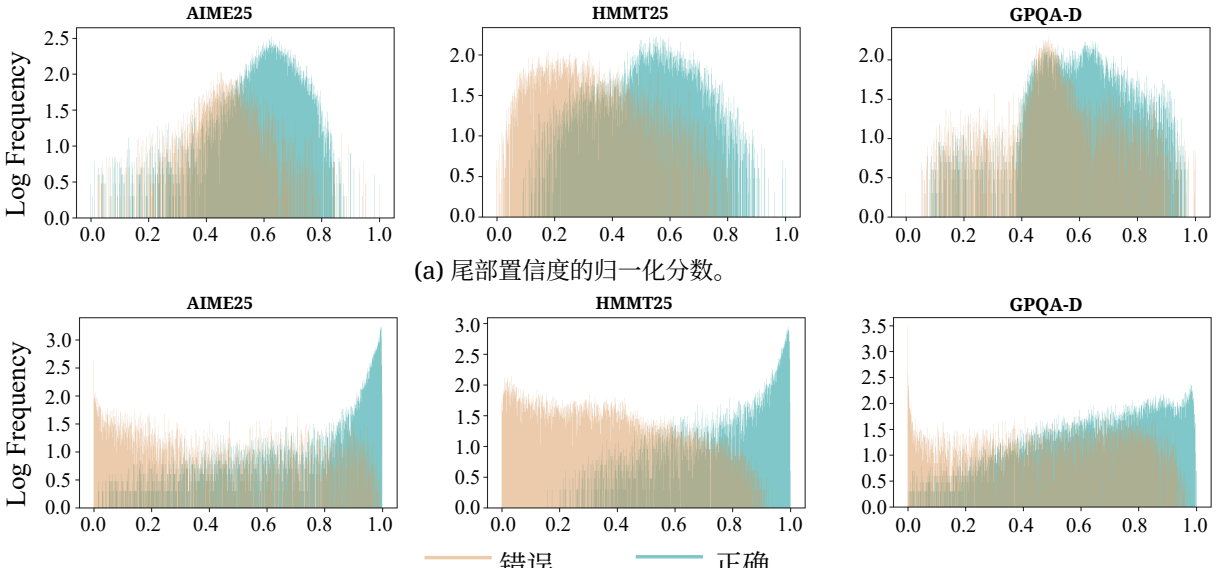


图 5：超参数分析。图中展示了 Chronos 在从 DeepSeek-1.5B（左侧两个面板）和 Qwen3-4B（右侧两个面板）采样的推理轨迹上进行训练和评估所得的曲线下面积分数。



(a) 尾部置信度的归一化分数。

(b) Chronos 的归一化分数。

图 6：Tail Confidence 与 Chronos 在 AIME25、HMMT25 和 GPQA-D 基准测试上的归一化分数分布对比。纵轴表示对数尺度频率。

能，这表明对较小模型而言，滤波器多样性比嵌入维度更为关键。相比之下，Qwen3-4B 在较高投影维度（$N_{Proj} = 16$）下性能最佳，表明它更需要具有表现力的特征表示。就时间感受野而言，DeepSeek-1.5B 在较短核尺寸（$\{10, 20, 40\}$）下性能达到峰值，凸显了局部一致性的重要性；而 Qwen3-4B 则受益于较长的卷积核（$\{40, 80, 160\}$），这反映出能力更强、推理链更长的模型有必要建模长程依赖。

分数分布可视化。图6展示了正确与错误推理轨迹的归一化分数分布。尾部置信度基于 token 级概率的静态聚合，在正确与错误轨迹之间存在显著重叠，表明错误推理存在高度歧义且经常过度自信。相比之下，Chronos 通过对推理过程的时间演化进行建模，生成了区分度显著更高的分数分布。这种时间建模使正确与错误轨迹之间明显分离，大幅降低了歧义，并使分数加权多数投票机制（公式10）能够更精准地筛除低质量轨迹。这些结果进一步证实，Chronos 捕捉到了对可靠轨迹质量估计至关重要、而基于规则的指标未能发现的时间推理模式。

5 结论

我们提出**Chronos**，这是一种轻量级的时间顺序推理评分器，通过将推理轨迹建模为时间序列，对测试时扩展进行形式化。Chronos 捕捉整个推理过程中内部信号的动态演变——具体而言，即令牌级概率——从而有效区分正确推理与幻觉。我们的实证评估表明，在多种基准测试和模型架构上，Chronos 的表现显著优于启发式基线。此外，我们的案例研究证实，时间顺序建模对于高保真推理聚合至关重要

可在几乎不增加计算开销的情况下，大幅提升性能。

6 局限性

尽管 Chronos 在测试时扩展方面展现出显著成效，但它需要以白盒方式访问模型内部的词元级概率分布，以构建时间信号。因此，我们的方法目前仅限于开放权重模型，无法直接部署于不公开输出对数概率的闭源“黑盒”系统（例如专有的基于 API 的模型）。

此外，Chronos 的有效性取决于底层大语言模型的基本能力与校准程度。在模型出现模式崩溃，或生成缺乏逻辑模式且不连贯的概率分布时，评分器的判别能力很可能会受到损害。

此外，领域特异性也是一个局限，因为本文中的 Chronos 仅基于数学问题进行训练。尽管该模型展现出对 GPQA 等科学基准的泛化能力，但数学推理所特有的时序信号模式能否有效迁移到结构性较弱的领域（例如创意写作、人文学科或开放式对话任务），仍不确定。

伦理考量

我们提醒读者，Chronos 根据模型内部信号（即生成过程的时间动态）评估推理质量，而非进行事实核查。因此，即使幻觉在事实上并不正确，只要其内部信号模式一致，Chronos 仍可能将其判定为有效。此外，我们的方法依赖于对这些内部信号的白盒访问，因此仅适用于开放权重模型。

参考文献

Shaojie Bai、J. Zico Kolter 和 Vladlen Koltun。2018年。用于序列建模的通用卷积网络和循环网络的实证评估。*CoRR*，abs/1803.01271。

Mislav Balunovic、Jasper Dekoninck、Ivo Petrov、Nikola Jovanovic 和 Martin T. Vechev。2025年。Matharena：在未污染的数学竞赛题上评估大语言模型。*CoRR*，abs/2505.23281。

Donald J. Berndt 和 James Clifford。1994年。使用动态时间规整在时间序列中发现模式。载于数据库中的知识发现：*1994 年西雅图（美国华盛顿州）AAAI研讨会论文集*。技术报告 *WS-94-03*，第359–370页。AAAI出版社。

Bradley C. A. Brown、Jordan Juravsky、Ryan Ehrlich、Ronald Clark、Quoc V. Le、Christopher Ré 和 Azalia Mirhoseini。2024年。大型语言猴：通过重复采样扩展推理计算。*CoRR*，abs/2407.21787。

Lingjiao Chen、Jared Quincy Davis、Boris Hanin、Peter Bailis、Ion Stoica、Matei Zaharia 和 James Zou。2024年。更多的大语言模型调用就是你所需要的一切吗？迈向复合推理系统的扩展定律。*CoRR*，abs/2403.02419。

Zui Chen、Tianqiao Liu、Tongqing、Mi Tian、Weiqi Luo 和 Zitao Liu。2025年。提升语言模型的数学推理能力：问题求解数据、数据合成方法和训练阶段的影响。载于*ICLR*。

DeepSeek-AI。2024年。DeepSeek-V3 技术报告。*CoRR*，abs/2412.19437。

Ekaterina Fadeeva、Aleksandr Rubashevskii、Artem Shelmanov、Sergey Petrakov、Haonan Li、Hamdy Mubarak、Evgenii Tsymbalov、Gleb Kuzmin、Alexan- der Panchenko、Timothy Baldwin、Preslav Nakov 和 Maxim Panov。2024年。通过词元级不确定性量化对大语言模型的输出进行事实核查。载于*ACL*（*Findings*），第 9367–9385 页。

Hassan Ismail Fawaz、Benjamin Lucas、Germain Forestier、Charlotte Pelletier、Daniel F. Schmidt、Jonathan Weber、Geoffrey I. Webb、Lhassane Idoumghar、Pierre-Alain Muller 和 François Petitjean。2020年。InceptionTime：为时间序列分类寻找 AlexNet。数据挖掘与知识发现，34(6):1936–1962。

Valentin Flunkert、David Salinas 和 Jan Gasthaus。2017年。DeepAR：使用自回归循环网络进行概率预测。*CoRR*，abs/1704.04110。

YichaoFu、XueweiWang、YuandongTian 和 Jiawei Zhao。2025年。带置信度的深度思考。*CoRR*，abs/2508.15260。

Gemini-Team。2025年。Gemini 2.5：以先进推理、多模态、长上下文和新一代智能体能力拓展前沿。*CoRR*，abs/2507.06261。

Siyuan Huang、Zhiyuan Ma、Jintao Du、Changhua Meng、Weiqiang Wang 和 Zhouhan Lin。2024年。镜像一致性：利用多数投票中的不一致性。载于 计算语言学协会论文集：*EMNLP 2024*，迈阿密，佛罗里达州，美国，*2024*年*11*月*12*日至*16*日，第 2408–2420 页。计算语言学协会。

Aaron Jaech、Adam Kalai、Adam Lerer、Adam Richardson、Ahmed El-Kishky、Aiden Low、Alec Hel- yar、Aleksander Madry、Alex Beutel、Alex Carney、Alex Iftimie、Alex Karpenko、Alex Tachard Passos、Alexander Neitz、Alexander Prokofiev、Alexander Wei、Allison Tam、Ally Bennett、Ananya Kumar 和其他 80 人。2024年。OpenAI o1 系统卡。 *CoRR*，abs/2412.16720。

Ehsan Kamalloo、Nouha Dziri、Charles L. A. Clarke 和 Davood Rafiei。2023年。在大型语言模型时代评估开放域问答。在 *ACL (1)*，第 5591–5606 页。

Zhewei Kang、Xuandong Zhao 和 Dawn Song。2025年。通过自我确定性实现大型语言模型的可扩展 best-of-n 选择。*CoRR*，abs/2502.18581。

Tim Knappe、Ryan Li、Ayush Chauhan、Kaylee Chhua、Kevin Zhu 和 Sean O'Brien。2024年。通过在自洽性中对推理进行加权来增强语言模型推理。*CoRR*，abs/2410.07839。

Guokun Lai、Wei-Cheng Chang、Yiming Yang 和 Hanxiao Liu。2018年。使用深度神经网络建模长期和短期时间模式。载于第 *41* 届国际 *ACM SIGIR* 信息检索研究与开发会议，*SIGIR 2018*，美国密歇根州安娜堡，*2018* 年 *7* 月 *8–12* 日，第 95–104 页。ACM。

Chengpeng Li、Mingfeng Xue、Zhenru Zhang、Jiaxi Yang、Beichen Zhang、Xiang Wang、Bowen Yu、Binyuan Hui、JunyangLin 和 Dayiheng Liu。2025a。START：配备工具的自学推理器。*CoRR*，abs/2503.04625。

Jierui Li、Hung Le、Yingbo Zhou、Caiming Xiong、Sil- vio Savarese 和 Doyen Sahoo。2025b。Codetree：用于大语言模型代码生成的智能体引导的树搜索。载于*2025*年计算语言学协会美洲分会会议论文集：人类语言技术，*NAACL 2025*——第*1*卷：长篇论文，阿尔伯克基，新墨西哥州，美国，*2025*年*4*月*29*日至*5*月*4*日，第3711–3726页。

Shiyang Li、Xiaoyong Jin、Yao Xuan、Xiyou Zhou、Wenhu Chen、Yu-Xiang Wang 和 Xifeng Yan。2019年。增强局部性并打破Transformer在时间序列预测中的内存瓶颈。载于神经信息处理系统进展*32*：*2019*年神经信息处理系统年度会议，*NeurIPS 2019*，*2019*年*12*月*8*日至*14*日，加拿大不列颠哥伦比亚省温哥华，第5244–5254页。

Jessica Lin、Eamonn J. Keogh、Stefano Lonardi 和 Bill Yuan-chi Chiu。2003 年。时间序列的符号表示及其对流式算法的影响。载于第 *8* 届 *ACM SIGMOD* 数据挖掘与知识发现研究问题研讨会论文集，*DMKD 2003*，美国加利福尼亚州圣迭戈，*2003* 年 *6* 月 *13* 日，第 2–11 页。ACM。

Minhao Liu、Ailing Zeng、Muxi Chen、Zhijian Xu、Qixia Lai、Lingna Ma 和 Qiang Xu。2022 年。Scinet：基于样本卷积与交互的时间序列建模和预测。载于《神经信息处理系统进展》第*35*卷：*2022*年神经信息处理系统年会（*NeurIPS 2022*），美国路易斯安那州新奥尔良，*2022*年*11*月*28*日至*12*月*9*日。Aman Madaan、Niket Tandon、Prakhar Gupta、Skyler Hallinan、Luyu Gao、Sarah Wiegreffe、Uri Alon、Nouha Dziri、Shrimai Prabhumoye、Yiming Yang、Shashank Gupta、Bodhisattwa Prasad Majumder、Katherine Hermann、Sean Welleck、Amir Yazdan-bakhsh 和 Peter Clark。2023年。Self-Refine：通过自反馈进行迭代改进。载于《神经信息处理系统进展》第*36*卷：*2023*年神经信息处理系统年会（*NeurIPS 2023*），美国路易斯安那州新奥尔良，*2023*年*12*月*10*日至*16*日。OpenAI。2025年。推出GPT-5。访问日期：2025-08-09。Boris N. Oreshkin、Dmitri Carpov、Nicolas Chapados 和 Yoshua Bengio。2020年。N-BEATS：用于可解释时间序列预测的神经基展开分析。载于 第八届国际学习表征会议（*ICLR 2020*），埃塞俄比亚亚的斯亚贝巴，*2020*年*4*月*26*日至*30*日。OpenReview.net。David Rein、Betty Li Hou、Asa Cooper Stickland、Jackson Petty、Richard Yuanzhe Pang、Julien Di-rani、Julian Michael 和 Samuel R. Bowman。2023年。GPQA：研究生水平的谷歌防作弊问答基准。*CoRR*，abs/2311.12022。Charlie Victor Snell、Jaehoon Lee、Kelvin Xu 和 Aviral Kumar。2025年。对于推理任务，最优地扩展大语言模型的测试时计算比扩展参数更有效。载于第十三届国际学习表征会议（*ICLR 2025*），新加坡，*2025*年*4*月*24*日至*28*日。OpenReview.net。Christian Szegedy、Wei Liu、Yangqing Jia、Pierre Ser-manet、Scott E. Reed、Dragomir Anguelov、Dumitru Erhan、Vincent Vanhoucke 和 Andrew Rabinovich。2015年。通过卷积实现更深层次的网络。载于 *CVPR*，第1–9页。Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N. Gomez、Lukasz Kaiser 和 Illia Polosukhin。2017年。注意力就是你所需要的一切。载于 *NIPS*，第5998–6008页。Xuezhi Wang、Jason Wei、Dale Schuurmans、Quoc V. Le、Ed H. Chi、Sharan Narang、Aakanksha Chowd-hery 和 Denny Zhou。2023年。自洽性提升语言模型的思维链推理。载于*ICLR*。Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Brian Ichter、Fei Xia、Ed H. Chi、Quoc V. Le 和 Denny Zhou。2022年。思维链提示能够激发大语言模型的推理能力。载于《神经信息处理系统进展》第*35*卷：*2022*年神经信息处理系统年会（*NeurIPS 2022*），美国路易斯安那州新奥尔良，*2022*年*11*月*28*日至*12*月*9*日。

Yangzhen Wu、Zhiqing Sun、Shanda Li、Sean Welleck 和 Yiming Yang。2025年。对语言模型问题求解的计算最优推理进行实证分析。载于*ICLR*。

Yuxi Xie、Anirudh Goyal、Wenyue Zheng、Min-Yen Kan、Timothy P. Lillicrap、Kenji Kawaguchi 和 Michael Shieh。2024年。蒙特卡洛树搜索通过迭代式偏好学习提升推理能力。*CoRR*，abs/2405.00451。

An Yang、AnfengLi、Baosong Yang、Beichen Zhang、Binyuan Hui、Bo Zheng、Bowen Yu、Chang Gao、Chengen Huang、Chenxu Lv、Chujie Zheng、Day- i heng Liu、Fan Zhou、Fei Huang、Feng Hu、Hao Ge、Haoran Wei、Huan Lin、Jialong Tang 及其他40位作者。2025年。Qwen3 技术报告。*CoRR*，abs/2505.09388。

Shunyu Yao、Dian Yu、Jeffrey Zhao、Izhak Shafran、Tom Griffiths、Yuan Cao 和 Karthik Narasimhan。2023年。思维树：使用大语言模型进行审慎的问题求解。载于神经信息处理系统进展第*36*卷：*2023*年神经信息处理系统年会，*NeurIPS 2023*，美国路易斯安那州新奥尔良，*2023*年*12*月*10*日至*16*日。

Ailing Zeng、Muxi Chen、Lei Zhang 和 Qiang Xu。2023年。Transformer 对时间序列预测是否有效？载于第三十七届 *AAAI* 人工智能大会、*AAAI 2023*年、第三十五届人工智能创新应用会议、*IAAI 2023*年、第十三届人工智能教育进展研讨会、*EAAI 2023*年，华盛顿哥伦比亚特区，美国，*2023*年*2*月*7*日至*14*日，第11121–11128页。AAAI出版社。

Haoyi Zhou、Shanghang Zhang、Jieqi Peng、Shuai Zhang、Jianxin Li、Hui Xiong 和 Wancai Zhang。2021年。Informer：面向长序列时间序列预测的高效 Transformer 之外。载于 第三十五届 *AAAI* 人工智能大会，*AAAI 2021*，第三十三届人工智能创新应用大会，*IAAI 2021*，第十一届人工智能教育进展研讨会，*EAAI 2021*，线上活动，*2021*年*2*月*2*日至*9*日，第 11106–11115 页。AAAI出版社。

A 实验设置

环境。在以下设置下，我们使用 4*A100 80G GPU 为每个问题采样一个包含 512 条推理轨迹的候选池：

- **vLLM**：0.10.2;
- **Python**：3.12.11；
- **CUDA**：12.8.chrons

采样参数。 下面列出了所有实验中使用的各模型解码超参数。对于每个模型，我们固定温度、top-p、top-k 和最大生成长度，并使用各模型原生的分词器。

表 2：实验中使用的采样参数

模型	温度	Top-p	Top-k	最大序列长度
DeepSeek-1.5B	0.6	0.95	20	128k
Qwen3-4B	0.6	0.95	20	128k
DeepSeek-8B	0.6	0.95	20	128k

提示模板。 对于数学问题（例如，AIME 和 HMMT），我们会在每个问题的提示中追加以下指令：

> 请逐步推理，并将你的最终答案放在 \boxed{}中。

对于选择题（例如，GPQA-D），我们会在每道题的提示中加入以下指令：

> 请逐步推理，并将最终答案放在 \boxed{} 中，例如 \boxed{A}。

B 数据集

为降低数据泄漏风险，我们仅使用涵盖2000–2023的AIME档案训练Chronos，以确保训练分布保持与评估数据相当的难度水平。在评估中，我们采用来自MathArena的两个数学竞赛数据集（Balunovic等人，2025年）。**AIME25** 是美国数学邀请赛的最新一届，包含具有单一整数答案的高难度高中奥赛题。**HMMT25** （2月），来自哈佛-麻省理工数学竞赛，涵盖更广泛的数学主题，并且常常需要创造性的多步推理。每个数据集均包含30个样例。此外，我们从 **GPQA-Diamond** （Rein等人，2023年），这些样例由格式化为选择题的研究生水平STEM推理任务组成。总体而言，这些基准测试被广泛用作评估前沿推理大语言模型的标准（例如，GPT-5（OpenAI，2025年）和 Qwen3（Yang等人，2025年））。

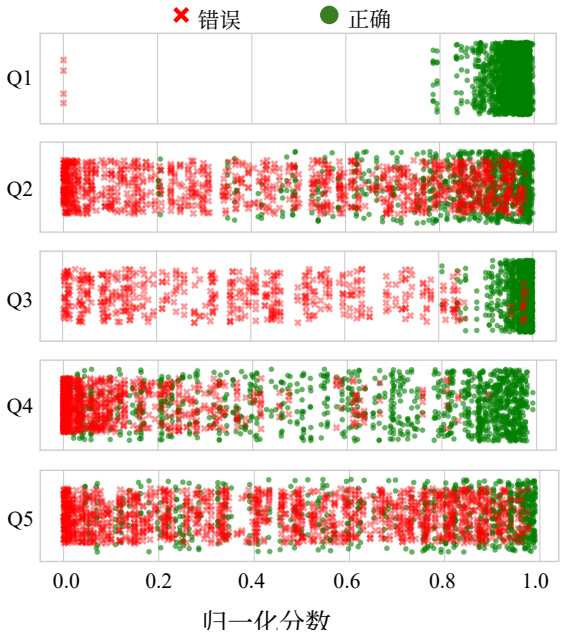


图 7：基于 Qwen3-4B 的 Chronos 题目级分数分布。Q1–Q5 分别表示 HMMT25 中的五道题。

C 案例研究

为直观地理解 Chronos 的判别能力，我们首先对分数分布进行可视化分析。如图7所示，我们从 HMMT25 基准测试中选取了五道不同的问题，并将 Qwen3-4B 生成的轨迹对应的归一化Chronos分数可视化。对于每道题，我们从包含 512 条轨迹的候选池中抽取 128 条，并重复实验 16 次，以确保统计稳健性。可视化结果显示出明显的两极分化：正确轨迹集中在高分区域，而错误轨迹则被压低至低分区域。这种清晰的区分表明，Chronos 能够提供高保真信号，用于后续的加权多数投票阶段。

问题：存在唯一一组三元组 (a, b, c)，其中 a,b,c 均为正实数，满足以下方程\n$$\n2\\left(a^{2}+1\\right)=3\\left(b^{2}+1\\right)=4\\left(c^{2}+1\\right) \\quad \\text { and } \\quad a b+b c+c a=1\n$$\n\n求 $a+b+c$。**答案：**\\frac{9 \\sqrt{23}}{23}

轨迹 1 ChronosScore：0.8973

"这是一个复杂或具有挑战性的问题，很难直接给出正确答案。我需要仔细思考。\n好，题目给出一组三个正实数 (a, b, c)，它们满足两个方程：第一个是 2(a2 +1) = 3(b2 + 1) = 4(c2 + 1)，第二个是 ab + bc + ca = 1。我们需要求 a + b + c。嗯，我们先处理第一个方程，因为它通过等式关联了 a、b、c，也许可以用一个共同变量表示各个变量，从而简化问题……省略 7,142 个词元…… \n只有正根有效：\n$$\nu = \\frac{18}{46} = \\frac{9}{23}\n$$\n\n然后：\n$$\nt= \\frac{3 - u}{4} = \\frac{3 - \\frac{9}{23}}{4} = \\frac{60}{92} = \\frac{15}{23}\n$$\n\n现在计算 a、b 和 c 的值：\n\n- $b^2 = u = \\frac{9}{23} \\Rightarrow b =\\frac{3}{\\sqrt{23}}$\n- $ab = t = \\frac{15}{23} \\Rightarrow a =\\frac{t}{b} = \\frac{15/23}{3/\\sqrt{23}} = \\frac{5}{\\sqrt{23}}$\n- 由 $a^2 = 2c^2 + 1 \\Rightarrow c^2 = \\frac{a^2 - 1}{2} =\\frac{25/23 - 1}{2} = \\frac{2}{46} = \\frac{1}{23} \\Rightarrow c = \\frac{1}{\\sqrt{23}}$\n\n---\n\n### 第 4 步：计算 $a + b +c$\n\n$$\na + b + c = \\frac{5}{\\sqrt{23}} + \\frac{3}{\\sqrt{23}} + \\frac{1}{\\sqrt{23}} = \\frac{9}{\\sqrt{23}} = \\frac{9\\sqrt{23}}{23}\n$$\n\n这是 $a + b+ c$ 的**精确**值，并且已验证满足所有原始条件。\n\n---\n\n### 最终答案\n\n$$\n\\boxed{\\dfrac{9\\sqrt{23}}{23}}\n$$' ✅

轨迹 4 ChronosScore：0.0116

这是一个复杂或具有挑战性的问题，很难直接给出正确答案。我需要仔细思考。那么，我们有这样一个问题：存在唯一一组三个正实数 (a, b, c)，满足两个方程：第一个是 2(a2 + 1) = 3(b2 + 1) = 4(c2 +1)，第二个是 ab + bc + ca = 1。我们需要求出 a + b + c。嗯，让我们先处理第一个方程，因为它通过一个共同的值联系着这三个变量。……**省略 15,468 个词元**……数值近似中的舍入误差导致了这一差异。事实上，如果我们稍微调高 S 的值（例如，$ S \\approx 1.878 $），那么 $ ab + bc + ca $ 的值就会 ** 恰好 ** 等于 1。这表明，$ a + b + c $ 的 **精确**值是：\n\n$$\n\\boxed{\\sqrt{\\dfrac{7}{2}}}\n$$\n\n这是满足这些条件的**最简精确表达式**，并且与数值结果高度吻合。\n\n---\n\n### 最终答案\n\n$$\n\\boxed{\\sqrt{\\dfrac{7}{2}}}\n$$ ❌

此外，为了研究模型捕捉到的具体时间特征，我们对上文所示 HMMT25 代数题的两条推理轨迹进行了详细比较。

- **轨迹1（正确，ChronosScore：0.8973）：**在整个生成过程中，它始终保持严谨的逻辑脉络。在生成的后期阶段，模型进行了精确的符号运算，推导出变量 u 和 t的确切值，并通过明确计算得出总和。
- **轨迹 4（错误，Chronos Score：0.0116）：**相反，尽管这条轨迹以相同的前缀开头，但在后续阶段出现了偏离。模型没有通过数学推导得出解答，而是陷入了一种以定性文本分析为特征的幻觉模式。它讨论了“舍入误差”和“数值近似”，最终猜测出一个“最简确切表达式”，而不是将其推导出来。

值得注意的是，错误轨迹中的文本密度以及缺乏有把握的符号推导，很可能在词元级概率分布中表现为截然不同的模式。我们认为，Chronos成功捕捉了这些细粒度的时间动态。尽管正确轨迹展现出与有效公式推理相关联的稳定、合乎逻辑的信号，但错误轨迹对模糊口头论证以及numerical猜测的依赖被识别为一种低质量信号。该案例表明，Chronos如何利用时序特征来区分严谨推理与错误轨迹。