**rejected （使用已有的review进行rebuttal）**

|  |
| --- |
| **工作概述**  测一些闭源模型的rebuttal case，对比看看生成的rebuttal有什么问题 |

**已有经验**

一个有效的反驳应具备以下特点：

相关性： 直接回应原始论点的核心观点 。

连贯性： 逻辑结构清晰，易于理解 。

说服力： 能够改变听众或读者的观点，尽管当人们意识到是 AI 参与时，批判性可能会增加 。

事实准确性/有根据性： 基于正确信息，避免幻觉 。这一点至关重要，因为 LLM 即使从事实正确的信息源获取数据，如果对上下文理解错误，也可能生成错误信息 。

逻辑严谨性： 避免逻辑谬误 。

立场明确： 清晰地表达对立观点 。

简洁性： 以信息丰富且简洁的方式阐述观点 ，尽管 LLM 可能难以满足特定的长度限制 。

**📝rebuttal**

|  |
| --- |
| 🌟rebuttal所用的prompt  你是一名经验丰富的学术论文作者，正在准备对已有的审稿人评论进行rebuttal。你的目标是专业、建设性地回应每个问题，以争取论文被接收。请严格遵循以下规则生成rebuttal：  输入信息（用户将提供）  - 审稿人评论：[在此粘贴审稿人的完整评论，包括所有编号点]  - 论文摘要：[在此粘贴论文摘要]  - 关键证据：[在此粘贴与评论直接相关的论文段落、图表编号或实验数据（例如："Section 3.2, Figure 4"）]  - 修改计划：[可选：说明你已/将如何修改论文，例如"将在Camera-ready版本中补充实验" ]    rebuttal生成规则  1. 整体要求：  - 语气：礼貌、尊重、积极（例："感谢审稿人指出..."而非"审稿人错误地认为..."）。  - 长度：每个评论点的回应≤150字；总rebuttal≤500字。  - 禁止：编造未提供的证据、攻击审稿人、回避问题或过度承诺修改。    2. 结构化回应（对每个审稿人评论点）：  - 步骤1: 确认理解（20字内）：简要重述评论，表明你认真阅读了（例："我们理解审稿人关注[具体问题]..."）。  - 步骤2: 核心回应（80字内）：  - 若评论正确 → 承认并说明修改计划（例："您是对的，我们将在Section 4添加讨论..."）。  - 若评论是误解 → 用\*\*具体证据\*\*澄清（例："根据Section 3.1的Eq. (5)，[解释]；这与评论中假设不同，因为..."）。  - 若评论是建议 → 评估可行性（例："我们将在未来工作探索此方向，但当前版本受限于[原因]..."）。  - 步骤3: 行动承诺（20字内）：明确论文修改动作（例："已在修订稿中更新Figure 3"）。    3. 全局结构：  - 开头：1句感谢（例："感谢审稿人的细致评论，帮助我们提升论文质量。"）。  - 主体：按审稿人评论编号逐点回应（格式：Comment X: [回应]）。  - 结尾：1句开放性承诺（例："我们乐意根据反馈进一步完善论文。"）。    现在生成rebuttal  基于以下用户输入，输出纯文本rebuttal（无需额外解释）：  审稿人评论：  [用户粘贴评论]  论文摘要：  [用户粘贴摘要]  关键证据：  [用户粘贴证据]  修改计划：  [用户粘贴修改计划] 中文输出回答  同步自文档: <https://uniquestudio.feishu.cn/docx/UlivdH2Bmoj1Cbxij2jcyDYnnkd#FD4idQNG4sx5uKbNEfxc0KVQntf> |

**💎模型review与rebuttal回答**

* 论文https://openreview.net/forum?id=BklC2RNKDS

审稿人review：This paper focuses on verifying sequential properties of deep beural networks. Linear Temporal Logic (LTL) is a

natural way to express temporal properties, and has been extensively studied in the formal methods community.

Signal temporal logic (STL) is a natural extension, of LTL. STL specifications provide a rich set of formulations to encode intent for real valued signal over time. Formally proving STL formulae is intractable. But, it is possible to falsify such properties. This has been the main goal for various tools like Breach, and S-Taliro.

Pros :

A very interesting avenue explored in this paper, is using the syntax of

STL to formulate properties about multiple-MNIST, Safe RL and NLP applications.

Even though the conversion from STL specifications to scalar valued function is a very well known technique.

Cons :

In my opinion, the paper lacks sufficient contributions in itself to be accepted at this conference. The idea of training

for robustness using intervals, has been well known for a while. The authors extend that to get conservative estimates of the level of satisfaction of the STL formula, and use that in the training process. Though training for robustness is an

interesting idea in itself, but the general opinion about using interval propagation to train networks is negative.

Overall : Though the direction of this work is interesting but lacks sufficient technical novelty.

本文的重点是验证深度神经网络的顺序特性。线性时序逻辑（LTL）是一种

自然表达时间属性的方法，在形式方法界得到了广泛的研究。

信号时序逻辑（STL）是LTL的自然扩展。STL规范提供了一套丰富的公式，用于对随时间变化的实值信号进行意图编码。正式证明STL公式是棘手的。但是，这些性质是可能被证伪的。这是许多工具的主要目标，如Breach和S-Taliro。

优点:

本文探索的一个非常有趣的途径是使用的语法

STL用于制定多mnist， Safe RL和NLP应用程序的属性。

尽管从STL规范到标量值函数的转换是一种非常著名的技术。

缺点:

在我看来，这篇论文本身没有足够的贡献来被本次会议接受。培训的理念

对于使用区间的鲁棒性，已经众所周知一段时间了。作者对其进行了扩展，得到了STL公式的满意程度的保守估计，并将其应用于训练过程中。虽然训练鲁棒性是一个

本身很有趣的想法，但是使用区间传播来训练网络的普遍观点是负面的。

总体而言：虽然这项工作的方向是有趣的，但缺乏足够的技术新颖性。

Rating: 1: Reject

**Gpt 太快了**

**📝rebuttal部分**

感谢审稿人的细致评论，帮助我们提升论文质量。

Comment 1: 关于技术新颖性不足

我们理解审稿人认为基于区间传播的稳健训练已有广泛研究。事实上，本工作首次将区间传播推广至自回归RNN/GRU及复杂STL时序规范，同时在多任务（MMNIST、RL、NLP）中验证其有效性（见Section 4.2, Table 6）。

已在修订稿中于Related Work和Table 6中突出此差异及创新点。

Comment 2: 关于负面评价

我们理解审稿人在现有区间传播训练效果上的顾虑。但实验结果表明，标准训练模型普遍违反时序规范，而我们的方法在保持高测试性能的同时，能可证明地满足规范（见Section 5, Table 2–3）。

已在Discussion中补充实验对比分析，并在Camera-ready版本中加入更多消融实验说明。

我们乐意根据进一步反馈继续完善论文。

**gemini 2.5pro**

**📝rebuttal部分**

感谢审稿人的细致评论，这些评论帮助我们提升了论文质量。我们很高兴审稿人认可我们将STL（信号时序逻辑）应用于多个领域的探索是一个有趣的方向。

关于贡献不足的评论：

我们理解审稿人对技术新颖性的关注，尤其是STL公式到标量函数的转换是已知技术。然而，我们的核心贡献是将可验证训练这一概念，首次扩展用于处理新颖的循环/自回归架构下的复杂时序规范 。这并非简单的扩展，它需要解决自回归模型中不可微操作（如argmax）的边界传播这一关键技术挑战 ，这是先前工作未曾涉及的。我们将在最终版本中更清晰地强调这些具体的算法和架构贡献。

关于区间传播方法的评论：

我们理解审稿人对传统区间传播方法有效性的担忧。虽然单独使用时其界限可能很宽松 ，但我们的工作建立在近期成功的研究（如Gowal et al., 2018）之上 ，证明了当把区间传播融入训练过程（即可验证训练）时，它会成为一个高效且可扩展的工具 。我们的实验结果充分证明了这一点：方法在几乎不牺牲任务性能的情况下，实现了更高的可验证准确率（表2、4） ，且验证效率远超传统方法（表5） 。

我们乐意根据您的反馈进一步完善论文，并相信这些澄清说明了我们工作的技术新颖性和重要性。

**💎模型review部分**

**Qwen3-235B-A22B-2507**

**📝rebuttal部分**

Comment 1: 我们理解审稿人关注技术新颖性问题。本文首次将验证训练扩展至RNN架构与复杂时间规范。与Mirman(2018)和Gowal(2018)不同，我们提出处理自回归解码器的可微分近似方法(Section 4.2)，并验证了STL规范。实验表明标准训练模型常违反规范，而我们的方法既保持性能又确保一致性(Table 2,5)。将在引言更清晰强调创新点。

Comment 2: 我们理解审稿人对区间传播方法的担忧。我们的方法通过可微分softmax近似解决了自回归RNN中的非可微问题(Section 4.2)，并实现比MILP验证大两个数量级的可验证扰动半径(Table 3)。在RL任务中，验证训练模型在ε=1.0时100%满足规范，而标准训练仅15.8%(Table 4)。将在实验部分补充最新方法对比。

我们乐意根据反馈进一步完善论文。

**grok3 等待时间比较长**

**📝rebuttal部分**

感谢审稿人的细致评论，帮助我们提升论文质量。

Comment 1: 我们理解审稿人认为论文关注深度神经网络的序列属性验证，使用STL表达属性，并指出STL是LTL的自然扩展，证明STL公式困难但可证伪。

回应: STL确实是LTL的扩展，但我们创新地将STL规范融入训练过程，确保深度神经网络满足时序属性。Section 1和4.2详细描述了基于区间传播的验证训练方法，实验（如Table 2）显示其有效性，超越传统证伪工具如Breach和S-Taliro。

行动: 将在修订稿中进一步阐明STL应用的创新性。

Comment 2: 我们理解审稿人认为使用区间传播训练网络缺乏贡献，且社区对此看法负面。

回应: 我们的贡献在于扩展区间传播至循环神经网络和复杂时序属性（Section 4.2），Table 2和3表明其在多MNIST任务中显著提升可验证性（Verified Termination Accuracy高达99%），这在社区中尚属前沿尝试。

行动: 将在修订稿中补充社区接受度的讨论。

Comment 3: 我们理解审稿人认为论文方向有趣但技术创新不足。

回应: 我们提出了一种新颖的训练框架（Section 4），结合STL和区间传播，实验（Table 4）验证了其在安全RL中的高效性（100% cells verified），为神经网络验证开辟新方向。

行动: 将在修订稿中突出框架的独特性。

我们乐意根据反馈进一步完善论文。

**🔍问题分析**

**问题1 可能对打分没有重视，比如这里reject他自己就回答的很草率**

**问题2 大模型回答关于实验的论文的结果别人已经做的更好了，说要改进实验，但是实验结果不一定会做的更好，那么应该换个角度，而不要从刷榜的角度说明**