面向结构性网络退化的解耦式大语言模型推理系统可靠性调度方法

Health Mask 与 Risk Price 混合调度

汇报人：黄冯

清华大学安全科学学院

2026 年 5 月 21 日

研究概览

研究问题

在解耦式大语言模型推理系统中，当网络供给发生连续、局部、部分可观测的结构性退化时，调度器如何在不重构 serving 架构的前提下降低 timeout failure？

核心思路

将二值 health mask 的故障避让扩展为连续 risk price 的可靠性调度，使调度器能对“尚未被明确标坏、但运行时风险已经升高”的区域提前降权。

核心观点

OTF-R 面向以 vLLM 为代表的 LLM serving 系统，在退化运行状态下为调度器补充一类可解释、可组合的可靠性权重。

报告提纲

网络退化为何会影响 LLM serving 可靠性

- 解耦式 serving 将 prefill、decode、KV cache transfer 和请求路由分布到不同节点或 pod。
- 网络不再只是背景资源，而是直接进入端到端时延、排队、timeout 和可用性。
- 在千卡、万卡级集群中，局部链路拥塞、gray failure 和健康标签延迟会被服务层放大。
- 因此，本研究关注退化运行状态下的服务可靠性，而不是正常状态下的平均吞吐优化。

现有故障避让机制及其局限

二值故障感知调度

如果某个 pod、GPU 或 expert 已经被标记为 unhealthy，调度器可以通过 health mask 或 prohibitive weight 避开它。

- 这一思路工程上合理，也是本研究比较的重要基线方法。
- 但网络退化常表现为连续退化、拥塞传播和服务层症状累积。
- health mask 可能存在检测窗口、延迟、漏报、误报和粒度过粗。
- 本研究重点分析如何在 health mask 之外引入连续风险信号。

结构性退化的主要表现形式

形态	传统故障避让	调度层关注点
GPU / expert 宕机	二值 unhealthy 标签较有效	作为 health penalty 输入
网络拥塞 / 降带宽	标签可能滞后或粗粒度	需要连续风险价格
gray failure	可能间歇出现症状	从 timeout / failure 历史估计风险
级联压力	局部退化诱发排队扩散	调度层提前偏转流量

关键结论

调度层风险降权的目标是在底层恢复完成之前，减少新请求继续进入高风险区域。

面向 vLLM 的调度信号补充

- vLLM 已经提供高效推理执行、KV cache 管理和调度基础，是本研究面向的重要系统背景。
- vLLM fault-tolerant 思路中已有 health mask / health penalty，可对明确 unhealthy 的目标做避让。
- OTF-R 补充的是另一类信号：当目标尚未被标记为 unhealthy，但 timeout、排队或网络症状已经恶化时，提供连续 risk price。
- 因此，OTF-R 的工程定位是补充 vLLM 调度信号，而不是替代 vLLM 运行时、KV cache 管理或底层恢复机制。

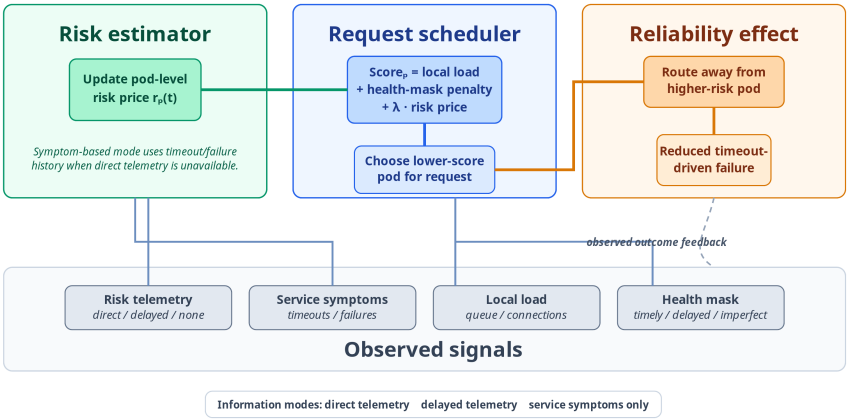
工程边界

vLLM 执行引擎、底层通信、GPU reset 和 Kubernetes 级别恢复仍由运行时与外部编排系统负责。

研究贡献

1. 问题层面：将结构性网络退化从“背景扰动”提升为解耦式 LLM serving 的核心可靠性约束。
2. 方法层面：提出 health mask 与 continuous risk price 的混合调度，使二值故障标签与运行时症状反馈能够共同进入调度评分。
3. 证据层面：在无直接网络遥测、标签延迟、标签不完整和高压力场景下评估方法，区分理想上界、重要基线方法与可部署信号。
4. 边界层面：当前工作为请求级机制验证，目标是补充 vLLM 调度信号；生产级 vLLM 集群验证需要进一步开展。

方法概览：Health Mask + Risk Price



统一调度评分

调度评分

$$S_j(t) = \mathrm{load}_j(t) + \lambda_f \cdot \mathrm{health\_penalty}_j(t) + \lambda_r \cdot \pi_j(t) + \lambda_p \cdot \mathrm{path\_cost}_j$$

- **load**：保持基本负载均衡。
- **health penalty**：对明确 unhealthy 的区域强降权。
- **risk price**：对连续退化、排队压力和失败症状软降权。
- **path cost**：保留拓扑和跨 pod 转发成本。

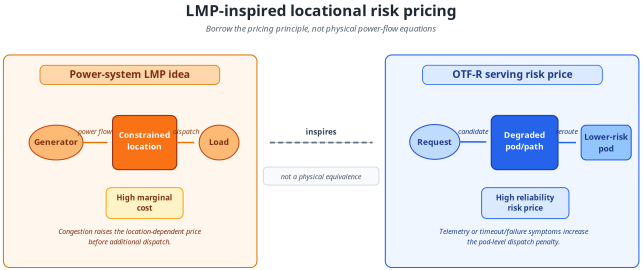
Risk Price 的定义

$$\pi_j(t) = \alpha \cdot \mathrm{util}_j(t) + \beta \cdot \left(1 - \frac{bw_j(t)}{bw_j^0}\right) + \gamma \cdot \mathrm{timeout\_rate}_j(t) + \eta \cdot \mathrm{failure\_rate}_j(t)$$

- 直接遥测：调度器能够观测 residual bandwidth 或网络退化信号。
- 无直接遥测：调度器不能直接看到网络退化，只能从 timeout / failure / utilization 中估计风险。
- 延迟遥测：信号存在观测滞后，更接近线上检测窗口。

解释： risk price 将已观测到的运行时压力转化为连续调度代价，用于影响后续请求分配。

从电力启发到 serving 调度



说明：电力边际价格提供机制启发；本研究仍使用 LLM serving 请求级仿真进行验证。

比较方法

类型	方法	作用
基础调度	LC	只按当前负载选择目标
标签调度	FaultAwareLC	根据 health mask 避让 unhealthy 区域
风险调度	OTF-R	根据连续 risk price 调整流量
混合调度	FaultAwareOTF-R	health penalty + risk price
上界参考	OracleAvoidance	拥有真实退化信息的理想上界

评价原则

在健康标签及时且准确时，FaultAwareLC 是重要基线；本研究重点评估标签延迟、不完整观测和部分可观测场景中 risk price 的增益。

仿真平台与规模

- 使用自包含 Python 离散事件仿真器，建模请求到达、排队、KV transfer、timeout 与调度决策。
- 主实验采用 thousand-worker 抽象规模：32 pods，每 pod 32 workers，共 1024 workers。
- 压力实验覆盖更大 worker 数和不同 workload，以检验趋势是否稳定。
- 仿真采用请求级抽象，用于评估可补充到 vLLM 调度层的风险信号；真实 vLLM 集群验证仍需进一步开展。

网络退化模型

Residual effective bandwidth factor

$$r \in \{0.20, 0.10, 0.05, 0.01\}$$

- r=0.20：轻度到中度退化。
- r=0.10：严重退化。
- r=0.05：非常严重退化。
- r=0.01：近故障状态或极端拥塞压力测试，不作为普通运行场景单独论证。

关键实验问题

1. 网络退化增强时，混合调度是否能降低 failure rate？
2. 当没有直接网络遥测时，risk price 是否仍有价值？
3. 当 health mask 延迟、不完整或含噪时，OTF-R 是否能补充 FaultAwareLC？
4. 风险价格参数是否稳健？调度开销是否可接受？

可靠性指标

指标	含义
Failure rate	失败请求比例，核心结果指标
Timeout share	失败中 timeout 的占比
Resilience index	故障后性能曲线的归一化面积
Traffic deflection ratio	从退化区域偏转出去的流量比例
Gap to oracle	与理想避让上界的差距
Scheduler overhead	单次调度决策耗时

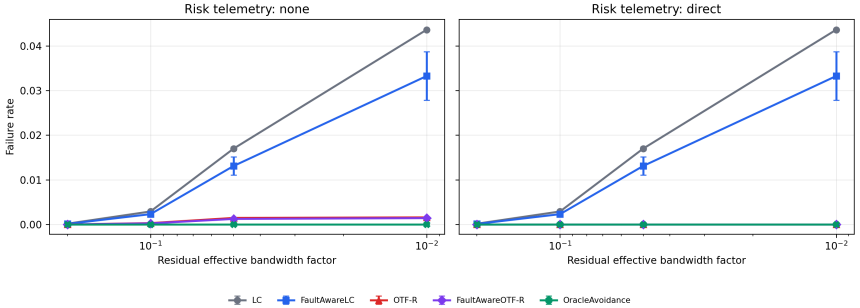
实验发现概览

问题	主要观察
网络退化增强	混合调度在严重退化下相对 FaultAwareLC 的失败率下降更明显。
直接遥测缺失	无直接遥测条件下仍能利用 timeout / failure 历史形成有效风险反馈。
health mask 不完备	当标签存在延迟、不完整或噪声时，continuous risk price 提供连续补偿。
调度开销	当前仿真实现中 1024 pod 级决策约为毫秒量级，具备进一步系统化验证价值。

关键结论

以下结果分别对应主效果、可观测性、机制解释、鲁棒性和开销五类证据。

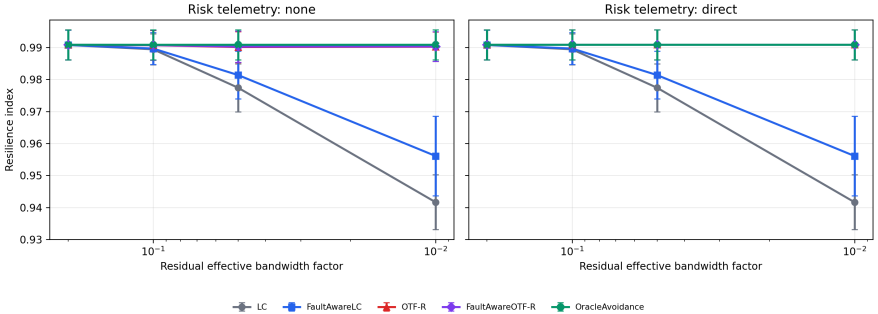
结果 1：严重网络退化下风险定价收益更明显



关键结论

在无直接遥测设置下，FaultAwareOTF-R 相比 FaultAwareLC 在 r=0.10、0.05、0.01 时分别降低失败率约 89.6%、90.5%、95.7%。该结论应限定在当前仿真负载、超时阈值和退化模型内。

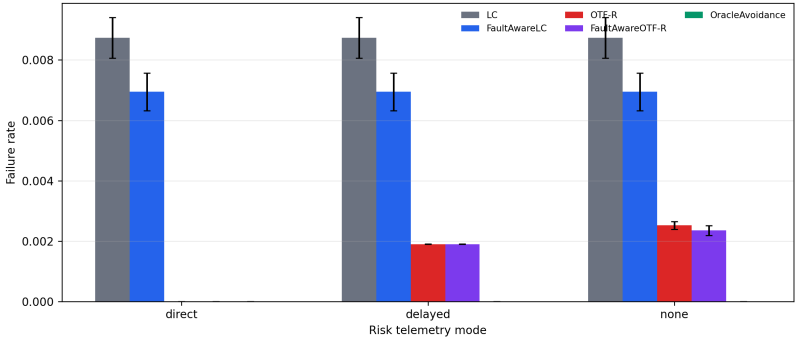
结果 2：韧性指标保持更稳定



关键结论

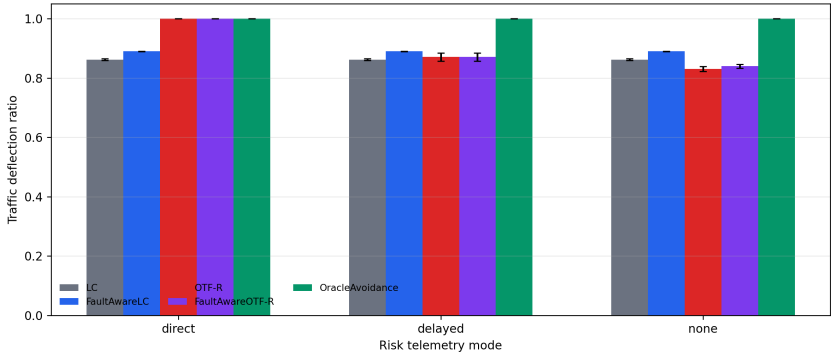
Resilience index 表明混合调度的收益不只来自某个单点失败率下降，而是在退化窗口内维持了更稳定的服务状态。

结果 3：间接运行时信号仍具备有效性



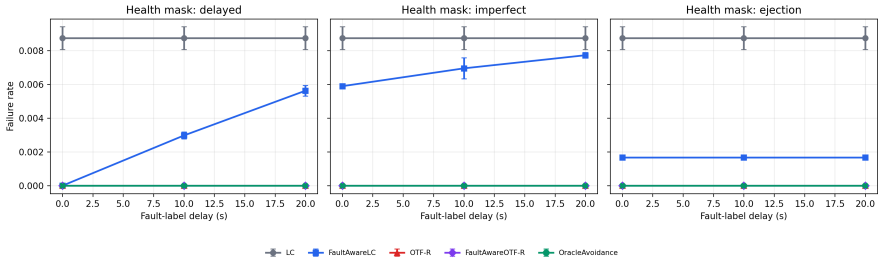
说明：直接遥测条件下效果最强；延迟遥测和无直接遥测条件下仍有收益，说明服务层 timeout、failure 和 utilization 历史可以作为可部署的间接风险反馈。

结果 4：风险价格确实触发流量偏转



说明：失败率下降与高风险区域被持续降权、流量向低风险区域偏转一致。

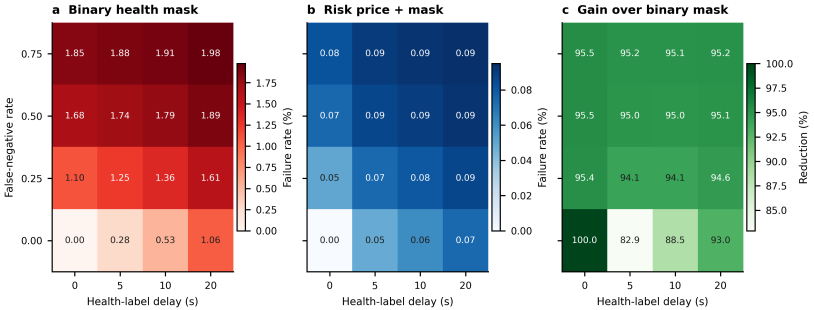
结果 5：health mask 不完备时，混合调度收益更明显



关键结论

该结果说明：当标签延迟、不完整或含噪时，连续 risk price 可补足二值标签在时间粒度和强度粒度上的不足。

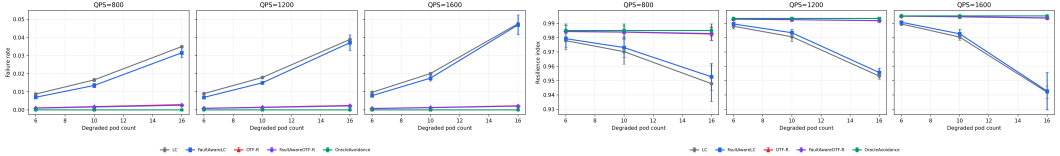
结果 6：延迟和噪声条件下的表面图



关键结论

该结果用于刻画 health mask 延迟与噪声同时存在时，混合调度的稳定性。

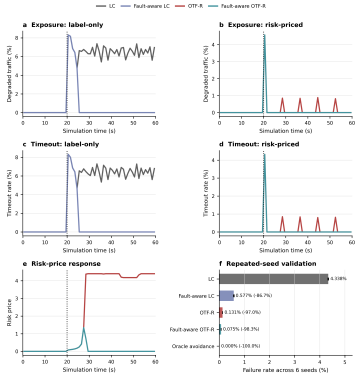
结果 7：无直接遥测下的压力场景



关键结论

在更高压力下，症状驱动的 risk price 仍能改善退化运行状态下的服务可靠性，说明该机制并不完全依赖直接网络测量。

结果 8：机制时间序列

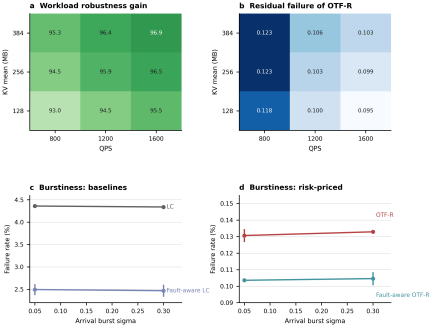


机制解释

- 退化开始后，timeout 与 failure 历史逐步抬升风险价格。
- 高风险区域被持续降权，新请求向低风险区域偏转。
- 该过程形成从症状观测到调度调整的闭环。

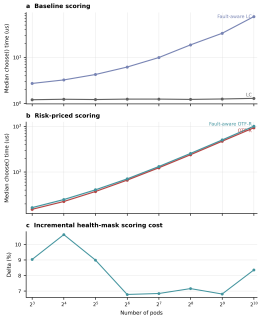
结论： risk price 将运行时症状转化为下一轮调度代价。

结果 9：工作负载鲁棒性



说明：结论不只依赖某一个 KV 大小或单一请求分布；真实线上工作负载仍需通过 vLLM 运行 trace 回放进一步验证。

结果 10：调度开销



说明： 1024 pod 级扫描中，FaultAwareOTF-R 单次决策约 1 ms；真实系统仍需要缓存、候选集裁剪和增量更新。

贡献与适用边界

1. 问题建模：把结构性网络供给退化引入解耦式 LLM serving 可靠性调度，并强调退化运行状态下的服务可靠性。
2. 方法设计：提出 health mask 与 continuous risk price 的混合调度框架，使明确故障标签与运行时症状可以互补。
3. 实验证据：在 health mask 不完备、无直接遥测和压力场景下验证 risk price 对 FaultAwareLC 的补充价值。
4. 工程定位：将 OTF-R 放在以 vLLM 为代表的 serving 调度层，作为 health mask 的补充，而不是运行时恢复、编排恢复或通信库恢复层。

对系统设计的启示

- serving 可靠性不应只依赖“坏节点标签”，还应利用 timeout、排队和失败历史形成连续风险信号。
- health mask 适合表达明确故障，risk price 适合表达退化强度；二者组合比单独使用任一信号更符合线上观测条件。
- 网络退化下的调度目标应从平均吞吐扩展为 failure rate、resilience index 和 gap to oracle 等可靠性指标。
- 对生产系统而言，下一步关键不是增加更复杂公式，而是验证该评分能否作为补充权重接入 vLLM 的候选选择流程。

适用范围与局限

- 当前结果来自请求级仿真，尚未覆盖真实 vLLM 生产集群验证。
- OTF-R 的作用是补充 vLLM 调度信号，并不替代 vLLM、health mask、故障恢复或编排恢复。
- 网络因子 r=0.01 对应近故障压力条件，不宜解释为普通日常退化。
- 直接网络遥测结果只作为一类可观测条件；无直接遥测与延迟遥测结果更接近受限观测场景。

后续验证与系统扩展

1. 基于真实 vLLM 请求轨迹开展回放实验，检验不同负载分布下风险价格的稳定性。
2. 构建轻量级 vLLM 调度接口原型，评估候选集规模、缓存策略和在线更新开销。
3. 在小规模多节点环境中注入网络退化，验证仿真结论能否迁移到实际 serving 流程。
4. 进一步评估与 health mask、故障恢复和编排系统之间的协同边界。

结论：三条可验证发现

1. 结构性网络退化会改变调度问题本身。在解耦式 LLM serving 中，网络退化会通过 KV transfer、跨 pod 路由和排队放大为 timeout failure。
2. 二值 **health mask** 与连续 **risk price** 具有互补性。 health mask 处理明确 unhealthy 目标，risk price 处理标签滞后、连续退化和服务层症状累积。
3. 混合调度在不完美观测下更有价值。在无直接遥测、标签延迟和高压力设置中，FaultAwareOTF-R 相比重要基线方法表现出更低失败率和更稳定的韧性指标。

结论：方法定位与后续验证

方法定位

OTF-R 是面向以 vLLM 为代表的解耦式 LLM serving 系统的可靠性调度增强方法，而不是新的 serving 架构，也不是底层故障恢复机制。

适用场景

更适合网络退化连续、health label 存在检测窗口、直接网络遥测不完整或线上工作负载压力较高的退化运行场景。

下一步

需要通过基于 vLLM 请求轨迹的回放实验或轻量调度接口原型，验证该补充信号在真实请求分布、真实候选集规模和真实系统开销下的有效性。

谢谢

敬请各位专家批评指正