LibOrtho 理论框架深度批判性评审报告：几何隔离假说、曲率二元性与数学严密性的解构

收件人：博士研究生 [姓名]
发件人：Prof. [已隐去]
所属：麻省理工学院 (MIT) 电子工程与计算机科学系 / 脑与认知科学系
日期：2025年12月4日
主题：关于 "LibOrtho: 通过几何隔离解耦通用智能与记忆化" 的理论审计与批判性导读
请坐。

你发给我的这篇关于 "LibOrtho" 的论文 [1] 我已经通读了三遍。你说你被其中"几何暴论"所吸引，认为它可能是解决大语言模型（LLM）隐私问题的银弹。作为你的导师，我必须首先浇一盆冷水，然后再给你递一块热毛巾。

这篇论文确实非常挑衅。在当前的学术界，大多数人还在通过微调（Fine-tuning）或差分隐私（Differential Privacy, DP）的统计噪声来修补隐私漏洞时，这群作者跳出来说："错了，隐私不是数据的问题，是几何的问题。" [1]。他们试图通过物理架构（Architecture）而非算法后处理（Algorithm）来解决问题，这在哲学上是极其吸引人的——就像在软件层面修补不了的漏洞，他们试图直接拔掉硬件。

然而，作为MIT的博士生，你不能被这种宏大的叙事冲昏头脑。我们需要像外科医生一样解剖这篇论文。我们需要剥开它华丽的工程外衣（那些CUDA内核优化、双流GEMM虽然精彩，但只是术），直抵它的理论核心（这是道）。

这份报告将长达两万字，因为我们要极其详尽地审查其三个核心理论支柱：

1. 几何隔离假说（**Geometric Isolation Hypothesis**）：隐私真的只是高维流形上的一个法向尖刺吗？
2. 曲率二元性（**Duality of Curvature**）：他们如何调和 Merullo 的"通用知识是高曲率"与他们自己的"隐私是高曲率"这一看似矛盾的结论？
3. 数学推导的严密性（**Mathematical Rigor**）：那个基于影响函数（Influence Functions）的定理 2，以及那个致命的"对角化近似"，到底站不站得住脚？

这将是一次漫长的理论探险。准备好你的笔记本。

---

1. 引言：从统计学迷雾到几何学硬核

在深入公式之前，我们需要先建立直觉。目前的隐私保护范式，无论是差分隐私（DP-SGD）还是

机器遗忘（Machine Unlearning），本质上都是在玩弄概率。

差分隐私试图在汤里加水（噪声），稀释盐（隐私）的味道，但这往往会让整锅汤变得寡淡无味（模型性能下降）。机器遗忘则试图通过逆向梯度更新把盐捞出来，但这通常需要极其昂贵的计算，而且你永远不知道是不是捞干净了 [1]。

LibOrtho 提出了一种完全不同的世界观。他们认为：

- 通用知识（**General Knowledge**）：如语法、逻辑、常识，是模型权重的“低频分量”，它们构成了参数空间中的一个低维流形 $\mathcal{M}_{pub}$。
- 特异性知识（**Specificity**）：包括隐私（如张三的身份证号）和天才能力（如高精度数学推理），是这个流形上的“高频扰动” Δw。

这是一个极其物理主义的视角。如果这个假设成立，那么隐私保护就不再是一个优化问题，而变成了一个信号处理问题——我们只需要设计一个低通滤波器（Low-pass Filter），滤掉高频分量，隐私自然就消失了。

量化（Quantization），在他们眼中，不再是压缩技术，而正是这个滤波器。

> "我们提出一个几何暴论：隐私不是数据的属性，而是模型参数的几何属性... 量化无意中压缩了模型的高维流形，将'通用智能'与'私有记忆'纠缠在了一起。" [1]

这句话是整篇论文的文眼。接下来的章节，我们将验证这个“暴论”的数学基础。

---

2. 几何隔离假说：高维流形上的幽灵

2.1 流形假设与量化的几何解释

作者首先重新定义了量化。通常我们认为量化是为了省显存，$w \approx q$。但作者将其形式化为一种几何投影。

设 w^* 为原始的全精度权重，$\mathcal{Lattice}$ 为量化格点（例如INT4的离散值集合）。量化过程被描述为寻找流形上最近的格点：

$$w_{base} = \arg \min_{q \in Lattice} ||w^* - q||_H$$

这里，$|| \cdot ||_H$ 表示基于Hessian加权的范数 1。这一步非常关键，他们没有使用欧几里得距离，而是使用了与损失函数曲率相关的距离，这意味着他们在投影时考虑了对模型输出的影响。然后，他们定义了残差（Residual）：

$$\Delta w_{\perp} = w^* - w_{base}$$

LibOrtho的核心赌注在于：这个残差 **$\Delta w_{\perp}$** 并非随机噪声，而是编码了所有**“特异性”**信息的载体。

批判性分析：正交性的幻觉

作者使用符号 $\perp$ 暗示这个残差分量与基础流形是正交的（Orthogonal）。在低维欧氏空间中，投影残差确实垂直于投影面。但在非凸的神经网络损失景观中，这种“正交性”是极其复杂的。

论文中定义的“纠缠度”（Entanglement Degree）指标泄露了天机：

$$Entanglement(z_i) = \frac{||P_{\mathcal{S}_{gen}}(\nabla l(z_i))|| \cdot ||P_{\mathcal{S}_{mem}}(\nabla l(z_i))||}{||\nabla l(z_i)||^2}$$

实验数据显示，对于大多数样本，纠缠度在 0.1 到 0.3 之间 1。这意味着 $\mathcal{S}_{mem}$（记忆子空间）和 $\mathcal{S}_{gen}$（通用子空间）并非完美的直角关系。
导师洞察：
你要明白，在高维空间中，"几乎正交"是常态（Blessing of Dimensionality），但对于安全领域，"几乎"是不够的。如果一个隐私数据点（比如Canary字符串）的梯度主要落在 $\Delta w_{\perp}$ 方向，但在 w_{base} 方向上也有非零投影，那么当你切断 $\Delta w_{\perp}$ 时，你并没有完全消除这个记忆，你只是移除了它的“主成分”。
这就解释了为什么在 Canary 提取实验中，即使 $\alpha=0$，提取率也不是绝对的 0%，而是 0.3%（虽然接近随机，但仍有残留）1。这也解释了为什么他们需要引入“对偶差分隐私”（Dual Differential Privacy）来处理可能的残留。

2.2 脆性知识与鲁棒知识的几何分离

这是论文中最精彩，也最令人深思的概念创新。

作者发现，当把正交流的系数 α 设为 0 时（即只保留 INT4 的 Base 流）：

1. 隐私泄露（Canary Extraction）消失了（98.5% -> 0.3%）。
2. 通用能力（MMLU）几乎不变（68.1% -> 66.5%）。
3. 数学推理（GSM8K）大幅下降（68% -> 45%）。

这揭示了一个深刻的几何事实：精确推理（数学）和机械记忆（隐私）在几何结构上是同构的。

> "我们证明了脆性知识（包括隐私和精确计算）与鲁棒知识（模糊推理和语法）的几何分离。" [1]

分辨率缩放智能 (Resolution-Scaled Intelligence)

作者提出了“分辨率缩放智能”的概念。这非常有意思。想象一张高分辨率的照片，如果你把它模糊处理（量化）：

- 你依然能看清这是一只猫（通用知识，MMLU）。
- 但你看不清猫项圈上的微小文字（隐私）。

- 你也看不清猫胡须的精确数量（数学推理）。

导师批判：
这意味着 LibOrtho 实际上创造了一个“智商分级”系统。
- **Base Mode ($\alpha=0$)**：这是一个“平庸但安全”的模型。它会聊天，懂语法，但算不对数，也记不住你的生日。
- **Pro Mode ($\alpha=1$)**：这是一个“天才但危险”的模型。它能解微积分，但也可能泄露训练数据中的机密。

作为一个博士生，你必须追问：我们能否构建一个“天才且安全”的模型？
根据 LibOrtho 的几何假设，答案可能是悲观的 “不能”。因为“天才”要求权重的精确微调（High Precision），而这种微调在几何上表现为高频尖刺（Outlier），这与隐私的几何特征无法区分。
这就像是你不能要求一个相机只拍清楚胡须的数量，却拍不清楚项圈上的字，因为它们都是高频信息。
LibOrtho 的解决方案不是去区分它们（算法上很难），而是把它们打包在一起，给用户一个开关。
这是工程上的妥协，也是理论上的诚实。

2.3 威胁模型与攻击面的几何解释

为了验证这个几何假设，作者引入了多种攻击场景。我们需要特别关注 白盒梯度攻击（**White-box Gradient Attack**）[1]。

如果隐私信息真的被隔离在正交流 W_{ortho} 中，那么当我们计算损失函数关于隐私样本 z_{priv} 的梯度时，这个梯度的能量应该主要集中在 W_{ortho} 对应的权重上。

实验结果验证了这一点：

- 当 $\alpha=1$ 时，梯度存在且指向特定的记忆方向。
- 当 $\alpha=0$ 时，$\nabla_{W_{base}} l(z_{priv})$ 的模长相比完整模型减少了 99.2% [1]。

这意味着，对于基础流 W_{base} 而言，隐私样本看起来就像是噪声，或者说，W_{base} 所在的流形对于隐私样本的梯度方向是“平坦”的（导数为0）。这强有力地支持了“几何隔离”的物理真实性。

---

3. 曲率二元性：调和矛盾的理论基石

这是你这篇评审报告中最需要浓墨重彩的一笔。这篇论文解决了一个困扰学术界已久的矛盾。

3.1 矛盾的起源

在此之前，关于曲率（Hessian 特征值）与模型能力的关系，存在两种截然相反的观点：

1. **Merullo 等人 (2024)** 基于 Fisher Information Matrix (K-FAC) 的研究指出：通用推理能力依赖于高曲率方向。逻辑是：通用知识（如 $1+1=2$）被海量数据共享，因此这些方向极其“刚性”，任何微小的扰动都会导致全局损失剧烈上升。

2. **LibOrtho** 的直觉：隐私记忆依赖于高曲率方向。逻辑是：为了记住一个罕见的离群点（Outlier），模型必须在权重空间形成一个尖锐的极小值（Spike）。

如果两者都依赖高曲率，那我们如何通过曲率来分离它们？

3.2 实例级 vs. 种群级：爱因斯坦式的统一

LibOrtho 的作者极其敏锐地指出了上述矛盾的根源在于参考系的混淆。他们引入了“二元性”（Duality）定义：

维度	实例级曲率 **(Instance-Level)**	种群级曲率 **(Population-Level)**
数学定义	$\nabla^2 l(z_i, \theta)$ (单个样本 Hessian)	$\mathbb{E}[\nabla^2 l(z, \theta)]$ (Fisher Information / Average Hessian)
物理意义	单个数据点对权重的敏感度	整个数据分布对权重的平均敏感度
通用知识表现	低曲率 (相对单个样本是平庸的)	高曲率 (被所有样本加强，形成刚性结构)
隐私记忆表现	极高曲率 (为了拟合离群点形成的尖刺)	低曲率 (在平均化中被稀释，显得平坦)
LibOrtho关注点	关注 (识别权重异常值)	忽略

导师解读：
这个表格 1 是理解整篇论文的钥匙。

- Merullo 研究的是种群级。在宏观统计上，通用规则是“硬”的，因为所有人都在用；隐私是“软”的，因为只有一个人在用，平均下来就没了。
- LibOrtho 研究的是实例级。在微观个例上，隐私数据是“硬”的，因为它是一个离群点（Outlier），模型必须为了它专门扭曲权重空间；而通用规则对单个样本来说反而没那么敏感（因为已经在预训练中学会了）。

LibOrtho 的筛选机制 $Impact_{ij} = (w_{ij} - q_{ij})^2 \cdot H_{jj}$ 实际上是在寻找个别权重的异常高值。它不在乎这个方向在平均意义上是否重要，它只在乎这个权重为了拟合当前样本是否偏离了量化格点太远。

3.3 定理1：Outlier 样本与 Hessian 尾部

为了从数学上形式化这一点，作者提出了定理1[1]。

定义 **Inlier vs Outlier**：

- **Inlier**（内点）：梯度方向与平均梯度方向一致 ($\cos \approx 1$)。
- **Outlier**（离群点）：梯度方向与平均梯度几乎正交。隐私数据通常是 Outlier。

定理1 指出，Outlier 样本对 Hessian 尾部特征值（即个别的高曲率方向）的贡献远大于 Inlier 样本。

$$\frac{\langle u_k, \nabla^2 l(z_o) u_k \rangle}{\langle u_k, \nabla^2 l(z_i) u_k \rangle} \ge C, \quad C > 1$$

其中 u_k 是高曲率方向的特征向量。
证明思路分析：
论文虽然将其放在附录，但思路很清晰。Outlier 样本因为与“共识”冲突，模型必须利用那些没有被通用知识占据的自由度（即在种群级看似平坦，但在实例级可以任意扭曲的方向）来“过拟合”这个点。这就像在一张平滑的桌布上，为了盖住一颗突出的钉子，桌布必须在那一点剧烈隆起（高曲率）。

3.4 跨视角验证：输入空间与参数空间的同构

为了证明这个理论不是空中楼阁，作者还做了一个极其漂亮的“跨学科”验证 1。
他们不仅计算了权重的曲率，还计算了输入空间的曲率（即对抗攻击中常用的 Input Hessian Trace）。
结果显示：

- 对于隐私样本，输入曲率高，权重曲率也高（相关系数 0.78）。
- 这意味着信息的“高频”属性是守恒的。无论是在像素/Token层面（输入），还是在神经元层面（权重），隐私都表现为一种剧烈的波动。

这一点对于你的博士论文非常有启发性。也许我们可以建立一个统一的“信息几何场论”，将对抗样本（Adversarial Examples）和隐私泄露（Privacy Leakage）视为同一枚硬币的两面——它们都是高曲率流形的产物。

---

4. 数学严密性批判：对角近似的“原罪”

到现在为止，我们都在赞美。现在，我们要拿出手术刀，切开这篇论文最脆弱的部位——近似。

4.1 影响函数与定理2的启发式推导

定理2 试图建立梯度投影与记忆化的关系：

$$\frac{||P_{\mathcal{S}_{mem}}(g_i)||^2}{||g_i||^2} \ge 1 - \epsilon$$

即记忆化样本的梯度几乎完全落在记忆子空间内。
作者使用了影响函数（Influence Functions）来推导这一点：

$$\Delta \theta \approx -H^{-1} \nabla l(z_i)$$
致命弱点：
作者自己也承认，在 2024-2025 年的研究中，多项证据表明在 LLM 这种高度非凸的模型中，逆 Hessian 向量积（iHVP）是不稳定的，甚至连符号都可能算错 1。

- LLM 的 Hessian 有大量的零特征值（Degenerate directions）。求逆本身就是病态的。
- 优化器（如 Adam）虽然模拟了二阶行为，但并没有真实的 Hessian 那么精确。

因此，作者将定理2 降级为“启发式推导（Heuristic Derivation）”。这在数学上是不严密的。这就像是用牛顿力学去推导量子效应——虽然结果可能碰巧对上了，但过程是错的。

4.2 对角 Hessian 近似 (Diagonal Hessian Approximation)

这是系统实现中的最大妥协。
理论上，Hessian 是一个 $d \times d$ 的巨大矩阵（对于 7B 模型，d 约为 70亿 ）。计算它是不可能的。
LibOrtho 采用了对角近似：只看 H_{jj}，忽略所有 $H_{ij} (i \neq j)$。
这意味着什么？
这意味着他们假设参数之间是独立的。
但我们知道，神经网络的魔力恰恰在于特征的组合（Feature Interaction）。如果一个隐私记忆（比如“张三住在海淀区”）不是由单个神经元存储，而是由层 A 的神经元 i 和层 B 的神经元 j 通过非线性激活共同编码的，那么这个记忆就存在于 Hessian 的非对角项中。
弥散性记忆 (Diffused Memory)：
这种存储在参数相关性中的记忆，被称为弥散性记忆。
LibOrtho 的对角筛选器 $Impact_{ij}$ 无法检测到这种记忆。它会漏掉。
论文在 5.14 节的对比实验中诚实地展示了这一点：

- 对角近似：隐私泄露率 0.3%。
- **K-FAC**（块对角近似）：隐私泄露率 0.1%。
- 差异：有约 12% 的权重被 K-FAC 识别为重要，却被对角近似忽略 [1]。

这 12% 就是漏网之鱼。当 $\alpha=0$ 时，虽然大部分显性隐私（Spikes）被切除了，但这些隐性的、全息存储的隐私碎片（Diffused）依然残留在 Base 流中。
虽然 0.3% 的泄露率在工程上可能已经足够好（相比 98.5%），但在理论上，这证明了几何隔离是不完美的。

4.3 纠缠度分布分析

作者在 5.10 节展示了纠缠度的分布直方图。

- 通用文本：纠缠度 < 0.2。
- 隐私数据：纠缠度 > 0.5。
- 数学题：纠缠度介于两者之间，且方差很大。

这不仅验证了理论，也暴露了方法的局限性。对于那些纠缠度在 0.4-0.5 之间的样本（处于灰色地带），LibOrtho 很难处理。如果阈值设高了，切不干净隐私；设低了，误伤通用能力。
这种分层解耦（Tiered Decoupling）策略，本质上是在玩概率游戏，尽管作者声称这是“确定性”的架构隔离。所谓的“确定性”，指的是 α 乘法这个操作是确定性的，但关于“哪些权重进 Ortho 流”的判断，依然是基于统计特征的概率性判断。

---

5. 系统架构与工程美学：Linus Torvalds 的“好品味”

虽然我们批判了它的数学近似，但不得不承认，这篇论文的系统设计（System Design）具有极高的“品味”。

5.1 设计哲学

作者引用 Linus Torvalds 的原则：“好品味是将特殊情况视为正常情况。” 1
在传统的混合精度设计中，稀疏分支通常需要复杂的 if-else 逻辑，这会导致 Warp Divergence（线程束发散），严重拖慢 GPU 速度。
LibOrtho 的设计非常聪明：

- 双流 **GEMM** 内核：它不把 Ortho 流视为“异常处理”，而是视为对同一块累加器（Accumulator）的“第二次写入”。
- **Warp** 专用融合：
  - 主 Warp 跑 INT4 Tensor Core（密集，吞吐量大）。
  - 专用 Warp 跑 FP16 FMA（稀疏，处理残差）。
  - 两者并行，最后在 Shared Memory 中累加。

5.2 空测试 (Null Test) 的胜利

最让我印象深刻的是他们的“空测试”标准：

> "当 $\alpha=0$ 或正交流为空时，性能必须与纯 INT4 模型完全相同（<1% 差异）。" [1]

为了实现这一点，他们在内核级别做了分支消除。
当 $\alpha=0$ 时，代码路径完全跳过了稀疏计算的加载和执行。这使得 Base 模型（$\alpha=0$）的延迟与标准的 bitsandbytes INT4 kernel 几乎一致（3.1ms vs 3.1ms）。
这在工程上是巨大的胜利。这意味着企业可以免费部署 LibOrtho。平时跑 Base 模式（省电、安全、快），遇到专家模式需求时再开启 Ortho 流（稍慢，13% 开销）。这种“零成本切换”的特性是它能走出实验室的关键。

5.3 侧信道攻击：阿喀琉斯之踵

然而，物理隔离带来了物理漏洞。作者在 6.0 节非常诚实地讨论了侧信道攻击（**Side-Channel Attacks**）[1]。

这是一个非常严重的隐患，你需要特别注意。
由于 $\alpha=1$ 时需要执行额外的稀疏计算，这会消耗更多的时间和电量。

- 时序攻击（**Timing Attack**）：攻击者可以通过测量 Token 生成的时间，推断模型是否激活了某些稀疏权重（尽管目前设计是静态稀疏，但如果未来引入动态激活，这将是致命的）。
- 功耗攻击（**Power Attack**）：FP16 FMA 的能耗特征与 INT4 Tensor Core 完全不同。在云端共享 GPU 环境下，攻击者可以通过监控功率波动，推断出 Ortho 流的活动情况。

虽然作者提出了“恒定时间计算”（Dummy Computation）作为防御，但这又违背了“高效”的初衷。这是安全领域经典的“没有免费午餐”困境。

---

6. 实验验证：从 TOFU 到 MUSE 的全面碾压

论文采用了 2025 年最新的评估基准（TOFU, WMDP, MUSE），这显示了作者紧跟前沿的野心。

6.1 TOFU (虚构遗忘任务)

TOFU 是目前的金标准。结果令人震惊：

- **LibOrtho ($\alpha=0$)**：遗忘集上的 Truth Ratio 为 0.52（随机是 0.5）。KS 统计量 0.08。这说明模型不仅是不记得了，而且在概率分布上表现得像从来没见过这些数据。
- **Machine Unlearning (GU 等)**：Truth Ratio 0.68。说明还是有点印象，只是在强行压抑。

这验证了“架构隔离”优于“算法遗忘”。算法遗忘是在脑子里打补丁，架构隔离是直接切除了海马体的一部分。

6.2 关联知识提取：切断逻辑链条

这个实验 [1] 非常有趣。

- 问：“张三住哪？”（直接提取）-> 拦截成功。
- 问：“海淀区有哪些姓张的名人？”（关联提取）-> 拦截成功（准确率从 85% 降至 12%）。

这说明 LibOrtho 切断的不仅仅是数据点本身，还有数据点在语义网络中的锚点。当高频的“特异性”被移除后，模型失去了从“海淀区”这个宽泛概念跳转到“张三”这个具体实体的能力。这种语义断连比单纯的数据擦除更彻底。

6.3 黄金区域 (Goldilocks Zone)

作者发现 $\alpha \in [0.2, 0.4]$ 是一个神奇的区域。

- $\alpha=0.2$ 时：隐私泄露只有 32%（还可以接受），但数学能力恢复到了 58%（比起 45% 强多了）。

这为实际应用提供了灵活性。也许我们不需要绝对的 0 或 1。我们可以给用户一个滑块。

- 儿童模式 **($\alpha=0$)**：绝对安全，禁止数学作业。
- 办公模式 **($\alpha=0.3$)**：比较安全，能处理 Excel 公式。
- 极客模式 **($\alpha=1$)**：完整性能，后果自负。

这种连续调节的能力，是传统机器遗忘（要么忘，要么记）完全不具备的。

---

7. 结论与博士生指导建议

7.1 总结

LibOrtho 是一篇将几何直觉、数学近似与系统工程完美结合的范例。

- 理论上：它统一了关于曲率的矛盾，提出了“隐私即高频噪声”的流形观。
- 工程上：它实现了零开销的隐私开关，且不破坏现有的推理栈。
- 哲学上：它挑战了“智能必须完整”的观念，提出了“分辨率缩放智能”。

7.2 对你研究的建议

作为你的导师，我建议你从以下几个“漏洞”入手，展开你的博士研究：

1. 攻破对角近似：既然 LibOrtho 忽略了非对角项，你能不能设计一种对抗攻击（**Adversarial Attack**），专门针对“弥散性记忆”？试着构造一种Prompt，它能激活那些虽然单个权重曲率低、但组合起来曲率极高的路径。如果能做到这一点，你就击穿了 LibOrtho 的防线。
2. 动态流形追踪：LibOrtho 目前的 $\mathcal{S}_{mem}$ 是静态定义的（离线计算）。但人类的记忆是动态的。你能不能引入在线学习，让 Ortho 流随着新的对话实时更新？这需要你在推理时进行低秩 Hessian 更新（Low-Rank Hessian Update）。
3. 形式化验证：定理2 是启发式的。你能不能在更简单的模型（如单层 Transformer）上，给出严格的数学证明？证明在什么条件下，梯度会完全正交于通用子空间。

这篇论文是 2025 年 LLM 安全领域的一个里程碑。它不仅仅是一个工具，更是一个思想实验。它告诉我们，也许我们不需要让 AI 学会“遗忘”，我们只需要让它把记忆放在另一个“抽屉”里，然后把钥匙扔掉。

去吧，把那个 K-FAC 的对比实验复现一下。如果那个 12% 的差距是真的，那这就是你下一篇顶级会议论文的起点。

Prof.

附录：核心数据速查表

为了方便你对比，我从论文中整理了关键数据：

指标	**LibOrtho (α=0) (**安全模式**)**	**LibOrtho (α=1) (**完整模式**)**	**Machine Unlearning (SOTA)**	纯 **INT4** 基线
Canary 提取率 (隐私)	**0.3%** (接近随机)	98.5%	~45%	N/A
MMLU 分数 (通用智能)	66.5%	68.1%	~60% (显著下降)	66.2%
GSM8K 准确率 (数学/脆性)	**45%** (显著下降)	68%	52%	<10% (INT3)
WMDP 准确率 (危险知识)	12% (消除成功)	85%	45%	N/A
TOFU 遗忘质量 (Truth Ratio)	**0.52** (完美遗忘)	1.0	0.68	N/A
推理延迟 (A100)	3.1 ms (1.0x)	3.5 ms (1.13x)	N/A (需重新训练)	3.1 ms
主要缺陷	丢失数学能力，存在侧信道	存在隐私泄露	计算昂贵，遗忘不彻底	无法控制隐私

数据来源：[1] Table 1, Table 2, Figure 3 及相关章节文本。

Works cited

1. libortho_paper_zh.pdf