eDIF：一种用于远程解释大型语言模型的欧洲深度推理结构

* *Feasibility Study for NDIF-Based AI Interpretability Cluster in Europe*

Irma Heithoff[1]
heithoff22587@hs-ansbach.de

Marc Guggenberger[1]
guggenberger22261@hs-ansbach.de

Sandra Kalogiannis[1]
s.kalogiannis16258@hs-ansbach.de

Fabian Maag[1]
f.maag@hs-ansbach.de

Susanne Mayer[1]
s.mayer16905@hs-ansbach.de

Sigurd Schacht[1,2]
sigurd.schacht@hs-ansbach.de

Carsten Lanquillon[2,3]
carsten@coairesearch.org

***Abstract*—本文介绍了一项关于部署欧洲深度推理结构（eDIF）的可行性研究，该结构是一个与 NDIF 兼容的基础设施，旨在支持对大型语言模型的机制可解释性研究。欧洲广泛获取 LLM 可解释性基础设施的需求推动了这一计划，以便为研究界民主化高级模型分析能力。该项目在安斯巴赫应用科学大学设立了一个基于 GPU 的集群，并与合作院校互联，通过 NNsight API 实现远程模型检查。一个包括来自欧洲各地的 16 名研究人员的结构化试点研究评估了该平台的技术性能、可用性和科学实用性。用户在包括 GPT-2 和 DeepSeek-R1-70B 在内的模型上进行了激活修补、因果追踪、表示分析等干预。研究显示用户参与度逐步增加，平台性能保持稳定，对远程实验能力的反响积极。这也标志着围绕该平台构建用户社区的起点。鉴定出的限制——如激活数据下载时间过长以及间歇性执行中断——在未来发展路线图中得到解决。这一计划标志着向欧洲广泛获取 LLM 可解释性基础设施的重要一步，并为更广泛的部署、扩展工具和在机制可解释性研究中持续的社区合作打下了基础。关键词：大型语言模型 · 机制可解释性 · NDIF · eDIF · 可行性研究**

I. 引言

机制可解释性旨在通过分析大型语言模型的内部组件来理解这些模型，其目标是在模型中揭示特定计算是如何产生的。这种方法在识别那些负责事实回忆、语法或推理的概念方面显示了希望 [1], [2] 。它通过启用因果干预和假设驱动的分析来补充黑箱可解释性 [3] 。

随着大规模模型在科学和商业应用中变得越来越重要，对透明性和可控性的需求也在增长。然而，获得先进系统的机会仍然有限。OpenAI 或 Anthropic 等提供商的专有 API 限制了对内部激活或梯度的检查，使得对这些系统的机械研究实际上变得不可行 [4], [5] 。即使是开放权重的模型，如 LLaMA 3 或 DeepSeek R1，也需要大量计算资源，这些资源通常是学术机构难以获得的，特别是那些没有专用 HPC 访问权的机构。

为了解决这个问题，国家深度推理织物（NDIF）[6] 等基础设施项目已经出现。NDIF 通过与 PyTorch 兼容的 API（NNsight）提供共享的远程访问大型模型，并支持延迟执行和安全共租，使得详细实验无需本地部署即可获得。除了其技术设计之外，NDIF 明确旨在通过使多个研究团队能够共享集中 GPU 资源和预加载模型来民主化可解释性工作流程，从而降低缺乏大规模计算基础设施的学术机构的进入壁垒。

基于这一概念，本项目提出了欧洲深度推理框架（eDIF）。这是一个基于欧洲的可扩展、开放的机械解释性研究基础设施。我们与 NDIF 团队合作，开展了一项可行性研究，以探讨关键问题：NDIF 类似的基础设施能否在欧洲的机构、资金和监管条件下有效部署？需要哪些技术调整和入驻流程？此外，这样的系统能多好地支持学术用户的实际解释性工作流？我们的研究通过试点部署评估了这些方面，特别关注可重复性、可用性以及与欧洲研究目标的对齐。eDIF 旨在降低进入壁垒，促进协作实验，并为在欧洲建设可持续、开放的研究基础设施用于大规模 LLM 分析提供实用见解。通过这样做，它促进了 AI 解释性研究的透明性、可重复性以及更广泛的社会访问。

已经出现了许多开源工具来支持变换器模型的机制可解释性。TransformerLens [1] 允许对内部组件进行详细分析，并支持激活补丁和因果追踪。诸如 pyvene [7] 和 baukit [8] 等框架提供了检查和修改模型内部的补充抽象。然而，所有这些工具都要求研究人员在本地托管模型并管理大规模计算环境，这在处理前沿模型时是一个重要障碍。

为了克服这一限制，NDIF 被开发为可解释性研究的远程基础设施 [6] 。NDIF 建立在 NNsight 的基础上，NNsight 是一个延迟执行 API，允许研究人员在标准 PyTorch 代码中定义干预，然后在共享的预加载模型实例上远程执行。这种实验设计和执行的分离使得共租用成为可能，降低了计算成本，并提高了学术用户的可访问性。

目前，欧洲缺乏相应的基础设施。虽然像 LUMI [9] 或 MareNostrum 5 [10] 这样的超级计算中心提供了通用的高性能计算访问，但它们并未针对交互式、以模型为中心的可解释性工作流程进行优化。法规、资金和访问限制进一步限制了较小研究团队的计算资源的实际可用性。

eDIF 项目通过建立一个欧洲 NDIF 兼容的集群来填补这一空白，该集群旨在实现机制可解释性。通过与现有工具（如 NNsight）对齐并专注于开放访问基础设施，eDIF 支持在欧洲背景下的可重复的、协作的研究。

[1] University of Applied Sciences Ansbach, Ansbach, Germany
[2] University of Heilbronn, Heilbronn, Germany
[3] COAI Research, Germany

II. 系统架构

eDIF 项目基础设施涵盖了三个计算环境：安斯巴赫应用技术大学的主服务器以及在弗里德里希-亚历山大大学（FAU）和海尔布隆大学的实验部署，使得德国研究机构能够分布式访问神经网络内省能力。

A. 硬件和软件组件

作为可行性研究的一部分，服务器基础设施配置了八块 NVIDIA RTX A6000 GPU，每块提供 48 GB 的 GDDR6 内存 [11] 。这种高内存设置被选择用于支持每块 GPU 上同时部署多个大型语言模型实例，这在当前 NDIF 框架中特别有利，因为用户间的 GPU 高效共享受到每块显卡内存限制的制约。因此，选择较少的拥有更大 VRAM 的 GPU 被证明是一个策略上有利的决策，以实现共同租赁并评估在机构计算环境中的长期可扩展性。

软件堆栈基于 Ubuntu LTS 22.04，利用 CUDA 进行 GPU 加速计算 [12], [13] 。容器化架构采用 Docker 和 Docker Compose 进行应用隔离，支持可靠的跨系统部署和简化的扩展。Ray 作为核心编排框架，通过基于队列的资源分配管理请求分发 [14] 。MinIO 提供针对 AI 工作负载优化的对象存储，支持高效的大规模数据管理 [15] 。监控和分析基础设施包括用于集中日志聚合的 Loki [16] ，用于时间序列监控的 Prometheus [17] ，作为实时分析数据库的 InfluxDB [18] ，以及作为可视化和监控前端的 Grafana [19] 。

B. 系统架构和数据流

eDIF 系统采用模块化栈，结合了 FastAPI（请求接口）、Ray（调度）和在 NVIDIA RTX A6000 GPU 上 Docker 化的模型运行环境。用户端的交互通过 NNsight 进行，这是一套基于 PyTorch 的工具包，可以进行深度模型自省。请求通过 NNsight 发出，并通过 FastAPI 路由至后端，由 Ray 管理基于 FIFO 的跨 GPU 实例调度。执行结果通过相同的 API 层返回。完整的栈——从用户界面到物理硬件——如图 1 所示。

NNsight 是一个基于 PyTorch 的分析工具，旨在揭示和操作深度神经网络的内部状态。它使研究人员能够在模型执行期间检查、追溯和改变中间激活和参数。在示例工作流程中（图 2 ），用户定义的跟踪函数在本地执行，而模型推理则委托给基于原始 NDIF 架构的远程后端基础设施。这种混合设置允许研究人员将细粒度的检查点注入远程多租户推理管道而不会影响执行效率。本地系统（eDIF）保持与 NDIF 的架构兼容性，确保跨部署的互操作性。

C. 外部基础设施和高性能计算集成

为了将计算资源扩展到专用 eDIF 服务器之外，通过与合作机构的密切合作，探索了选定的外部基础设施。其目的是评估在异构系统中的可移植性，并在真实高性能计算（HPC）限制条件下评估平台的适应性。在这些合作伙伴中，弗里德里希-亚历山大-埃尔朗根-纽伦堡大学（FAU）提供了对其基于 AMD 的 HPC 集群的访问权限。该环境配备了八个 AMD Instinct MI300X GPU（每个 GPU 具有 192 GB 的 HBM3 内存），使得在替代硬件架构和容器运行时环境下进行有价值的实验成为可能。

在 FAU HPC 集群上的 eDIF 部署作为一个独立实例实现，与主要的 eDIF 部署完全独立。它仅包括模型推断所

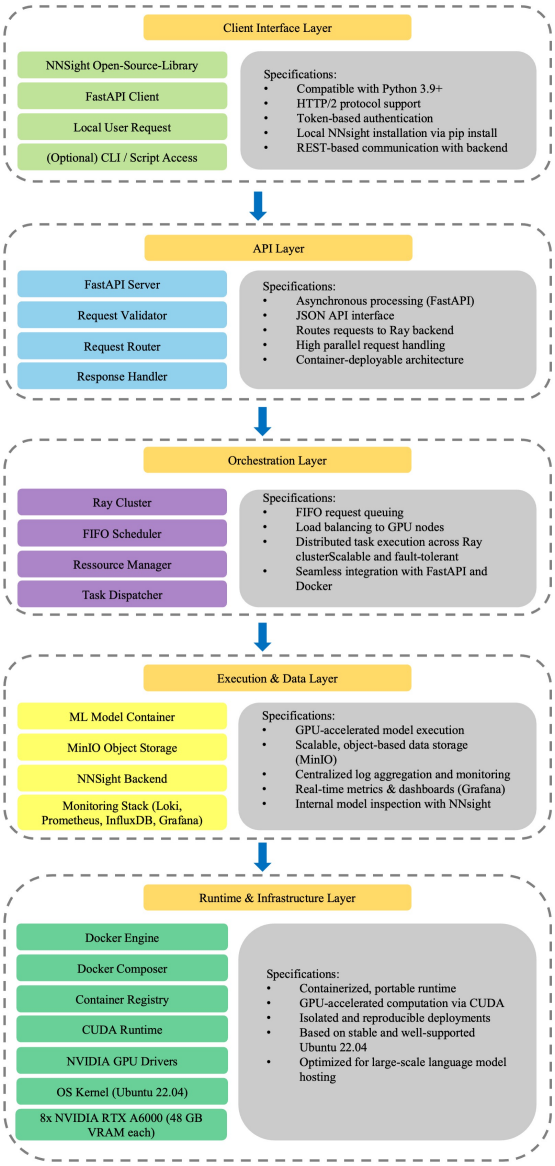

Fig. 1. eDIF 系统的分层架构，具有用户交互、编排和 GPU 执行

需的核心服务——特别是 Ray 和 FastAPI 等组件——而监控仪表板和可视化工具等补充模块由于时间和系统限制被排除在外。由于原始的 NDIF 栈依赖于基于 Docker 的架构，向 FAU HPC 环境的过渡需要将这些核心服务移植到 Apptainer（以前称为 Singularity），这是一种与非 root、多用户 HPC 系统兼容的容器运行时。像镜像构建、服务编排和网络等 Docker 特定功能必须要么手动重新配置，要么用轻量级的基于 shell 的替代方案来替换，其中 necessary 。

将系统移植到 AMD 硬件还需要将 NVIDIA 的 CUDA

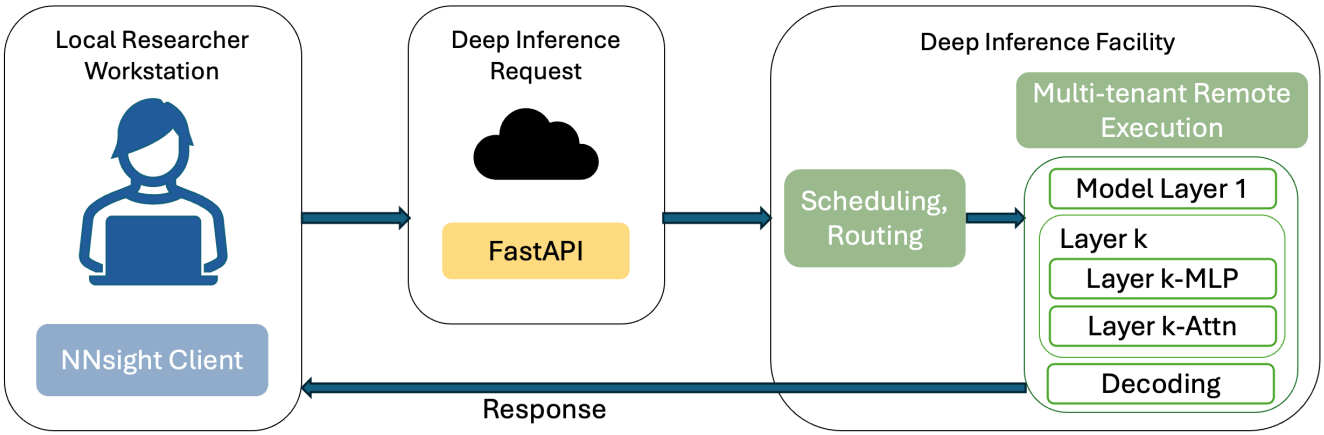

Fig. 2. 用于远程神经网络检查的 NNsight 分布式执行架构

框架切换到 ROCm（Radeon Open Compute）。虽然大多数基于 PyTorch 的模型代码仍然兼容，但进行一些小的调整，比如用 AMD 的 RCCL 替换 NVIDIA 的 NCCL（一个用于高效多 GPU 通信的库），就足以保留核心推理工作流程，这要归功于 PyTorch 的运行时抽象 [20], [21] 。

与持续运行的容器服务不同，HPC 平台按需分配资源，并且具有固定的运行时间限制。这需要调整推理管道以符合计划的执行工作流程，并在测试期间引入了较长的周转时间。此外，将容器化的组件从 eDIF 系统移植到基于 AMD 的 HPC 环境中，需要针对 ROCm 进行特定调整及重新配置现有的镜像层，因为某些基于 CUDA 的假设必须重新处理以实现兼容性。

尽管由于时间限制没有进行全面部署，与 FAU 的合作为在共享学术计算平台上模型服务基础设施的可移植性和限制性提供了重要的见解。非正式资源协议下提供的机构支持，反映了对开放研究和校际合作的坚定承诺。

III. 部署与操作

除了技术基础设施的建立之外，重点放在具体的配置、模型部署以及稳定的运行时监控和管理上。

A. 设置

系统设置遵循官方的 NDIF 指南，从 GitHub 克隆并根据 eDIF 上下文进行了适应调整 [22] 。配置包括端口、模型实例和 GPU 可用性。初始测试通过容器启动后的 API 查询进行。额外的调整包括基于 Traefik 的反向代理以实现安全的外部访问。在设置过程中配置了监控组件（Loki, InfluxDB, Prometheus, Grafana），以便后续集成。

B. 模型供应

表一列出了在 eDIF 项目期间进行广泛部署和使用的主要 LLM。这些模型在大部分运行时都保持活跃，并根据具体需求或实验要求进行选择性替换。此外，Llama 4 Maverick 模型托管在外部 FAU 服务器上以支持进一步测试所有模型均来自公共 Hugging Face 仓库，并集成时未使用额外的包装。从用户的角度来看，模型访问通过 NNsight 和 eDIF 后端之间无缝连接，呈现出与标准 Hugging Face 使用相同的界面。

Model	Model size	Required GPUs
GPT-2 small [23]	124 million parameters	1× RTX A6000
DeepSeek-R1-Distill-Llama-8B [24]	8.03 billion parameters	2× RTX A6000
DeepSeek-R1-Distill-Llama-70B [25]	70.6 billion parameters	5× RTX A6000

Table I. 评估模型

每个 LLM 的 GPU 分配是基于预期的内存需求进行的。这种方法提供了合理的资源规划，但由于 GPU 无法在模型之间共享，从而导致了低效问题。单个 GPU 上的未使用内存对其他模型来说是不可访问的，导致理论上的 GPU 容量与实际使用之间存在差距。

在模型配置后，重点转向了稳定的运行时操作和持续的系统监控。监控设置依赖于几个已建立工具的组合：

- Prometheus：时间序列指标的收集和存储；性能数据表示的基础。[17]
- InfluxDB：历史时间序列数据的长期保留；在 Prometheus 旁起补充作用。[18]
- Loki：通过基于标签的索引进行日志聚合；不对日志内容进行全文索引。[16]
- Grafana：统一的仪表板界面；来自多个来源的指标和日志的可视化。[19]

这个监控堆栈提供了实时的可观测性，并在确保系统性能可靠中起着重要作用。

IV. 可行性研究的设计和方法论

eDIF 的可行性研究在 2025 年 5 月 12 日至 6 月 25 日期间在安斯巴赫应用技术大学精心计划和执行。这项研究在一个硕士项目中进行，主要目标是建立和完善一套专门为可解释性研究量身定制的计算基础设施。研究的招募通过电子邮件和 Discord 社区进行，目标是主动参与机制可解释性研究的研究人员，强调提供强大的 GPU 资源以及一个与 NNsight 工具包无缝集成的基于 PyTorch 的实验 API。这些研究人员可以免费使用 eDIF 的高性能基础设

施。招聘过程共收到 29 份申请，这些申请来自于欧洲各地的科研机构和大学。选择标准包括与欧洲国家的联系、在机制可解释性方面有明确的研究重点、以及展现出使用该平台功能的动机。许多申请者还表示计划探索高级可解释性技术，如稀疏自编码器的训练和评估、电路追踪，以及将大型语言模型从“不透明的黑箱”转变为“可解释的白箱”的方法。此外，能否在可用的基础设施上执行所提的实验（例如，模型兼容性、GPU 需求）对于参与者的选择起着决定性作用。从这些申请中，最终精选出了 16 位研究人员，这些人具有可解释性方面的显著专长，并且明确需要进行计算密集型任务。来自七个不同欧洲国家的研究人员参与其中——包括英国、荷兰、西班牙、意大利、德国、奥地利和法国。这些参与者来自广泛的学术机构，并带来了多样的研究目标和学科视角。可行性研究的时间结构遵循一个明确定义的顺序，旨在支持研究进行的技术整合和科学严谨性。在初始阶段，参与者通过一个结构化的流程加入，其中包括针对两个核心用例定制的详细教程笔记本。一个用例集中于调查 LLM 输出中的性别偏见，另一个则专注于使用 eDIF 后端分析和操控 LLM 神经元激活中的保护本能。此外，还提供了一个专门的入门笔记本，用于一步步介绍 eDIF 后端，并指导用户设置 API 访问，包括令牌配置。在接下来的步骤中，系统基础设施得以完善，并且个别研究目标与平台的能力相对齐。这之后是一个延长的实验阶段，在此期间，用户在各自的用例中实施探测、追踪和干预实验。研究中间设立的结构化反馈检查点基于用户体验促进了快速迭代和改进。最终阶段涵盖了经验结果的综合、反思性分析和正式文档化——为未来的发展和更广泛适用性的评估奠定了基础。研究的实验协议围绕三种主要形式的干预结构，即探测特定神经元激活以了解其个体角色、跨不同神经网络模型层追踪激活，以及执行旨在探索网络内部因果关系的目标干预。

V. 结果

本节概述了 eDIF 项目的主要结果，结合技术见解和用户体验来评估系统的性能和实用性。

A. 技术发现

对 eDIF 系统的技术评估揭示了几个关键见解。部署使用了一台配备八个 NVIDIA RTX A6000 GPU 的专用 Ubuntu 服务器。CPU 和 RAM 的要求很低，而 GPU 容量是主要制约因素。在实践中，基于估计的内存需求分配 GPU 是可行的，但不够理想。由于无法在多个模型之间共享 GPU，导致效率低下，因为个别设备上的内存未被使用。这导致理论上的 GPU 可用性与有效的资源利用之间存在显著差距。

尽管如此，只要硬件足够，规模达到 DeepSeek-R1-70B 的模型也可以成功配置。官方 NDIF 部署指南被证明是可转移的，允许与 NNsight 的无缝集成。当模型分布到多个 GPU 时，观察到一些运行时问题，有时会由于共享参数（例如，权重和偏差）的访问导致崩溃。

基于 Docker 的系统架构——结合了 Ubuntu、Ray、FastAPI 和 MinIO——在日常操作中证明了其稳定性。监控工具如 Prometheus、Loki、InfluxDB 和 Grafana 运行可靠，能够提供有用的实时诊断和可视化，尽管这对基本模型服务来说不是必需的。AMAX_2 使用 FAU HPC 集群进行的迁移实验展示了在 ROCm 而不是 CUDA 上运行 NDIF 的普遍可行性。PyTorch 的内部抽象允许在只需少量修改（例如，NCCL → RCCL）情况下重用大部分现有代码库 AMAX_0 。AMAX_3 然而，HPC 环境带来了结构性挑战。NDIF 的基于 Docker 的设置由于根访问限制而不兼容；因此，需要部分迁移到 Apptainer。这涉及到广泛的重新配置，因为 Apptainer 不支持 Docker 的关键功能。持久化服务进一步受到 HPC 调度约束和时间限制的计算会话的阻碍 AMAX_1 。AMAX_4

B. 试点用户活动和反馈

eDIF 框架的两个核心组件得到了参与者的显著正面反馈：首先是 NNsight，一个多功能的基于 PyTorch 的 API，可以促进模型干预；其次是 eDIF，作为基础设施。

在整个研究期间，用户活动显示出一个稳定增长的参与曲线，并在项目最后一周的请求量大幅度达到峰值。如平台的内部监控仪表板中所示（图 3），对 eDIF 服务器的请求从五月中旬最初的稀疏分布上升到六月底单日超过 700 次。尽管负载急剧增加，系统保持了功能上的稳定，大部分请求都成功处理。但偶尔也出现模型冻结和个别错误状态。

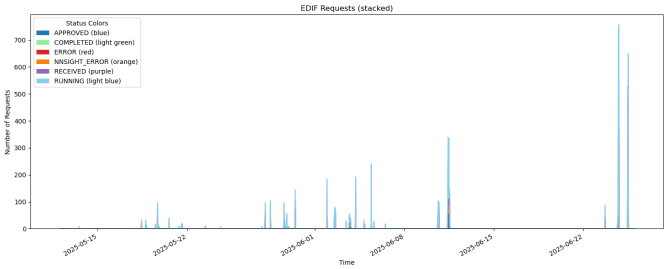

Fig. 3. eDIF 服务器在可行性研究期间的活动

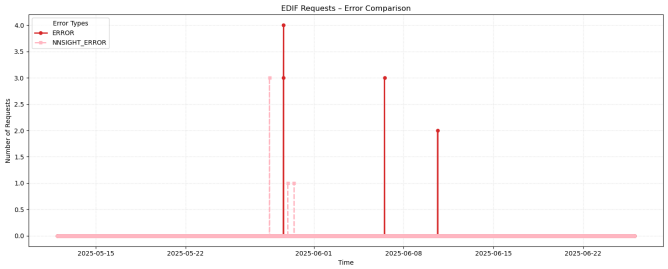

Fig. 4. eDIF 请求——可行性研究期间的误差比较

用户体验评估显示满意度很高，其中入职和易用性的平均评分为 4.2 分（满分 5 分），而整体框架的直观性评分为 4.0 分。参与者特别强调了远程执行功能所提供的优势，该功能允许在不需要大量本地计算基础设施的情况下对模型进行干预。与现有的替代方案如 TransformerLens [1] 相比，NNsight 经常被认为具有更高的优越性，主要因为其远程执行功能和与 PyTorch 工作流程的无缝集成。

在研究过程中，参与者积极使用 eDIF/NNsight 来进行机械解释性研究中的一系列任务。这些方法包括因果追踪、激活补丁 [27] 、logit lens 分析 [28] 、线性探针训练 [29] ，以及与归因、电路分析 [30] 和稀疏自编码器 [29] 表征相关的研究工作。激活补丁和 Logit Lens 是最常用的功能之

一。主要使用的模型包括 GPT-2 和 DeepSeek-R1-Distill-Llama-8B，偶尔使用 DeepSeek-R1-Distill-Llama-70B。尽管平台有其优点，但也遇到了一些技术和可用性方面的挑战。报告的问题包括激活数据的下载时间过长——有时长达 30 分钟——以及服务器的不稳定和执行中断（图 4 ）。用户特别要求改进自动处理连接中断的功能和增强错误管理机制，包括重试功能。尽管基础设施通常被描述为接近生产就绪状态，但带宽限制和缺乏对远程字典学习的集成支持偶尔会妨碍高效的工作流程整合。然而，约有一半的用户估计使用 eDIF 可以节省 10-30 % 的时间，而另外 33 % 则体验到了 30-50 % 的时间节省。甚至有一位参与者报告效率提高超过 50 %。功能增强请求主要集中在扩展现有的模型库，以包含现代的、紧凑的架构，如 Qwen 和 Gemma3 [31], [32] 。此外，参与者倡导通过功能抽象来改善模块化，例如 `extract_activations` ，他们对在单一操作任务中执行连续处理步骤（例如，先进行 Top-k 选择，然后探测）表现出兴趣 [1], [33] ，以及远程访问像 Logit Lens 这样的分析工具。进一步的建议包括开发视觉辅助工具（例如，Top-5 神经元视图，类似 Neuronpedia 的界面 [34] ），性能增强（例如，压缩输出和 VRAM 使用元数据），以及全面、专用的新手引导文档。用户一致建议提供一个独立的 Python 包以简化代码启动，并提供更清晰、通用的代码示例，以涵盖各种模型类型。关于未来的无代码/低代码功能，其总体相关性获得了 4 分（满分 5 分）的平均评分。这些功能极大降低了非技术团队成员、学生和跨学科合作伙伴的进入门槛，得到了全体参与者的一致认可，100 % 的参与者对此潜力给予了满分 5/5 的评分。具体请求的功能包括探测、logit lens 分析、激活补丁和表征引导。然而，参与者也提出了关于抽象固有复杂性可能后果的关键考虑，指出对结果的透明度、可重复性和可解释性可能产生的影响。社区建设成为未来发展的另一个重要领域。引入共享异步协作环境，如协作笔记本和开放代码讨论，被认为非常有价值，83.4 % 的参与者对该功能的评价为 4 或 5（满分为 5）。同样，阅读小组等实时互动形式也得到了热烈支持。总之，83.3 % 的研究参与者报告说在可行性研究期间部分整合了 eDIF/NNsight 到他们的研究过程中，同时所有参与者（100 % ）对继续在欧洲使用该平台表现出浓厚兴趣。

VI. 讨论

eDIF 可行性研究的结果强调了 eDIF/NNsight 框架在欧洲的机械解释性研究中的实用可行性和巨大效用。该双重验证——以客观技术基准为基础，并通过系统的用户评估加强——展示了该框架的健壮性和实际影响力。

A. 技术结果的讨论

eDIF 项目中实现的系统架构——基于 Ubuntu、NVIDIA GPU、Docker、Ray 和 FastAPI——被证明是稳定和可扩展的。这一设置为未来的复制或扩展提供了可靠的基础。通过采用更强大的硬件或替代的 GPU 类型，可以有效地扩大用户数量和模型规模。

一个关键的限制是缺乏自动的 GPU 需求估计。模型部署必须通过试错手动调整，这对于单个模型是可行的，但在大规模时效率低下。开发一个（半）自动化的 GPU 分析工具，应是广泛部署场景的优先事项。

监控组件（Prometheus、InfluxDB、Loki、Grafana）提供了有关系统负载和性能的宝贵见解。为了未来的可扩展性，使用额外的度量扩展这些工具，可以增强调试和优化能力。此外，从纯粹的拉取方式转向推送通知系统（例如在限制或运行时错误情况下提醒管理员），可以减少对手动检查的需求，并实现更积极主动的系统管理。

在 FAU 的基于 AMD 的 HPC 集群上的实验产生了超越本地环境的可转移见解。HPC 系统上的资源分配本质上是时间有限的，因此与需要持续模型推理的持久 API 不兼容。由于根权限风险，缺乏对 Docker 的支持，需要通过 Apptainer 替代部署策略。然而，Apptainer 缺乏与 Docker 完全的特性对等，导致复杂的迁移，涉及大量重写容器逻辑、脚本和网络设置。此外，将基于 CUDA 的容器构建适配为 ROCm 有时需要对现有库函数的直接代码更改或运行时修改。必须为这些挑战开发技术解决方案，以支持 NDIF 结构在 HPC 集群上的有效规模化部署。尽管存在这些限制，从 CUDA 到 ROCm 的迁移在技术上被证明是可行的。PyTorch 的内部抽象允许大部分代码库被重用，仅需少量调整。这一发现尤为重要，因为某些 AMD GPU（如 MI300X 和 MI325X）显示出具有竞争力的价格性能比，在某些推理场景下甚至优于 NVIDIA 的 H100 和 H200。这使基于 AMD 的系统成为未来扩展 NDIF 或 eDIF 等基础设施的一个有吸引力的选择。

B. 用户反馈讨论

在研究结论阶段，平台利用率的显著增加表明参与者对于平台的熟悉和舒适度在增长，这显示出成功的入职和培训效果。然而，需要注意的是，参与反馈的人数相对较少，这可能限制了见解的全面性和研究结果的普遍适用性。同时，所收到的反馈的质量和深度被证明非常有价值，突出显示了平台的特定优势领域和需要改善的针对性机会。此外，观察到了活动停滞的阶段，这与欧洲假期以及用户因其他责任而很少使用平台的时期相吻合。

虽然观察到的服务器请求激增，峰值时超过每天 700 个请求，显示出显著的兴趣和运营可扩展性，但在高负载条件下出现的间歇性错误突出了基础设施优化的必要性，特别是在带宽管理和错误处理协议方面。从科学的角度来看，观察到的错误率在一个较小的学术机构试点阶段实施的可接受阈值内。因此，它们应被视为早期研究基础设施固有的预期波动，而不是其探索实用性上的关键限制。用户反馈一致认为远程执行是该框架的显著优势，表明其对现有替代方案有明确的偏好。此外，参与者明确表达了对扩展模型产品的希望，包括访问更大和更多样化的语言模型。一个经常被提及的需求是能够在活动会话期间灵活切换模型。这一能力被视为对快速实验和在解释性工作流程中进行高效比较至关重要。然而，发生的性能瓶颈，如长时间下载和个别模型执行冻结，需要有针对性的技术优化，以增强健壮性和用户生产力。进一步整合自动错误恢复机制以及改进功能模块化可以显著简化工作流程，特别是在处理连续解释性任务时。

最后，对无代码/低代码功能以及协作社区特性的强烈兴趣表明未来平台发展的战略方向。这些增强功能有望扩大可访问性，促进跨学科合作，并在多样化研究社区中促进更广泛的采用。

VII. 未来工作

eDIF 可行性研究为推进欧洲研究生态系统中的人工智能可解释性基础设施奠定了坚实的基础。基于积极的成果和明确的用户需求，下一阶段将包括已确认的开发和探索性的举措。这些重点领域涉及基础设施整合、资助渠道、社区扩展和研究能力。下文所述的路线图反映了正在积极评估的即时行动和战略方向。

在 MVP 技术成功之后，参与的机构已经确认了长期的基础设施支持。目前安斯巴赫应用技术大学的服务器配备了 NVIDIA RTX A6000 GPU 以及 RTX 6000 和 H100，将继续运行，并在项目正式结束后继续向选定的试点用户开放。同时，正在与海尔布隆大学协商，争取获得 H200 集群的访问权限。这一升级将显著提高系统进行大规模或并发解释性实验的能力。此外，计划部署基于 AMD 的 GPU，以在 NDIF 框架下评估多供应商的兼容性。这一配置将使得能够在异构架构中对 NDIF 性能进行基准测试。Prometheus 和 Grafana 的系统监控已经就位，确保能够全面了解硬件的使用情况和实验的吞吐量。用户支持也将得到加强：定期的检查、实时支持渠道以及使用日志将确保研究人员能够及时获得帮助和反馈。这些支持机制旨在提升研究人员的体验，并促进实验设置的可重复性。已经与国家和国际利益相关者进行了初步沟通，以探索潜在的资助机会。目前正在考虑进一步的步骤。同时，已经为一个战略领域提交了拨款提案：开发无代码和低代码接口以扩大可及性。最后，社区发展正在进行中。一个以 eDIF 为中心的欧洲研究网络正在筹备中，目标是实现机构之间的长期合作和共享工具。在 MVP 期间收集的反馈将被正式整合到用户入门改进、培训文档以及 eDIF 的平台教程和网站中。未来的用户招募将针对在线社区，例如 Discord 和 NNsight 开发者网络，旨在接触更广泛和多样化的用户群。

一个专用平台已经可以通过 eDIF 网站访问，提供简化的入门材料访问、全面的文档，并支持不断发展的用户社区的发展。[4] 同时，在线社区空间（例如，Discord、GitHub）和社区活动（如阅读小组、工作坊或编程马拉松）正在计划中，以促进参与、共享学习和长期合作。

除了业务连续性之外，还有一些以研究为导向和以可用性为重点的举措正在考虑用于即将到来的项目阶段。这些努力旨在提高科学实用性、平台灵活性和可访问性。

同时，在线社区空间（例如，Discord、GitHub）和社区活动，如读书小组、研讨会或黑客马拉松，也正在规划中，以促进参与、共享学习和长期合作。

模型和实验多样性一个关键重点是实现跨模型的可解释性研究，例如比较 DeepSeek 和 Qwen 模型的内部行为。这些实验旨在探索可解释性方法在不同架构上的泛化能力。除了大规模的国际模型，特别关注也将给到欧洲的小型模型，这些模型通常难以获得，比如 BCC/Salamander（在巴塞罗那高性能计算中心开发）和 Teuken（在 Fraunhofer IIS 开发）。多模态模型（例如 Qwen-2.5-VL）的部署也在评估中，从而允许在视觉-语言背景下进行可解释性研究。此外，项目计划研究模型量化和 GPU 级分片如何影响可解释性表现。这包括分析减少资源消耗与激活分析的保真度、探测精度或因果跟踪工作流之间的权衡。为了支持此类研究，在同一会话中热交换不同模型实例是一个高度需求的功能。这个功能将显著提高对比实验的效率，并且与广义 NDIF 路线图中列出的未来发展保持一致。

可用性和访问扩展为了提高研究人员的生产力和降低入门门槛，正在考虑一些平台和工具改进措施：

这些改进旨在提高平台对专业用户和跨学科合作者的可用性和可访问性。

治理与可扩展性随着平台的扩展，健全的治理机制变得至关重要。计划实施身份验证和访问控制系统，以支持可扩展和安全的使用。这包括基于令牌的授权、用户角色管理和配额执行。资源分配将由公平使用调度策略——例如时间段预约或每用户 GPU 配额——来管理，以确保研究团队之间的公平访问。

同时，所有数据处理和使用追踪将符合 GDPR 要求，特别是在多机构或跨境研究的情况下。这确保了该平台在支持透明和可审计的研究工作流程的同时，仍符合欧洲数据保护标准。总体而言，这些发展使 eDIF 成为 AI 可解释性研究的可持续和可扩展的基础设施。通过保持技术的健壮性、扩大访问范围，并支持新的科学方向，该平台将继续对欧洲 AI 系统的透明性和可重复性做出有意义的贡献。

这项可行性研究证明了一个兼容欧洲 NDIF 的可解释性集群的可行性和研究价值。eDIF 的部署成功支持了对大模型的远程、结构化干预，并为多样化的研究群体促进了可重复的实验。用户强烈的兴趣，高使用量，以及对 NNsight 界面的积极评价，确认了该平台在学术可解释性工作流程中的相关性。尽管存在诸如模型切换等技术挑战，但包括支持 AMD GPU、自动错误处理和扩展模型选项在内的拟议升级，为前进提供了明确的路径。该研究为将 eDIF 扩展到全欧洲资源奠定了坚实的基础，促进跨学科合作并支持对透明、可访问的 AI 系统日益增长的需求。未来的发展将专注于可用性改进、无代码工具和可持续治理，将 eDIF 定位为欧洲 AI 研究基础设施的重要支柱。在项目过程中识别的挑战反映了在大型数字基础设施中也会出现的常见紧张局势。鉴于对 AI 技术透明性、可重复性和公平访问的需求日益增长，eDIF 作为一个代表性的实例，展示了在欧洲开展负责任研究所需建立的基础设施和交流基础。

[4] https://www.edif.ai/

VIII. 致谢

我们衷心感谢 Sigurd Schacht（安斯巴赫应用科技大学）和 Carsten Lanquillon（海尔布隆大学）在整个项目中提供的学术指导和机构支持。我们也感谢弗里德里希-亚历山大大学（FAU）埃尔兰根-纽伦堡的计算中心在评估阶段提供 HPC 基础设施的访问。此外，我们感谢与美国 NDIF 团队的合作，特别是大卫 · 鲍（波士顿东北大学）和艾玛 · 博尔兹（波士顿东北大学），他们的技术建议和交流极大地促进了 eDIF 可行性研究的成功。

References

[1] N. Nanda and J. Bloom, “TransformerLens,” GitHub repository, 2022. [Online]. Available: https://github.com/TransformerLensOrg/TransformerLens

[2] C. Tang, B. Lake, and M. Jazayeri, “An explainable transformer circuit for compositional generalization,” arXiv preprint arXiv:2502.15801 , Feb. 2025. [Online]. Available: http://arxiv.org/abs/2502.15801

[3] L. Bereska and E. Gavves, “Mechanistic interpretability for AI safety – A review,” arXiv preprint arXiv:2404.14082 , Apr. 2024. [Online]. Available: http://arxiv.org/abs/2404.14082

[4] R. Bommasani, K. Klyman, S. Kapoor, S. Longpre, B. Xiong, N. Maslej, and P. Liang, "The 2024 foundation model transparency index," arXiv preprint arXiv:2407.12929 , Mar. 2025. [Online]. Available: http://arxiv.org/abs/2407.12929
[5] R. Sapkota, S. Raza, and M. Karkee, "Comprehensive analysis of transparency and accessibility of ChatGPT, DeepSeek, and other SoTA large language models," arXiv preprint arXiv:2502.18505 , Feb. 2025. [Online]. Available: http://arxiv.org/abs/2502.18505
[6] J. Fiotto-Kaufman et al. , "NNsight and NDIF: Democratizing access to open-weight foundation model internals," arXiv preprint arXiv:2407.14561 , Apr. 2025. [Online]. Available: http://arxiv.org/abs/2407.14561
[7] Z. Wu, A. Geiger, A. Arora, J. Huang, Z. Wang, N. D. Goodman, C. D. Manning, and C. Potts, "pyvene: A library for understanding and improving PyTorch models via interventions," arXiv preprint arXiv:2403.07809 , Mar. 2024. [Online]. Available: http://arxiv.org/abs/2403.07809
[8] D. Bau, "baukit: Tools for inspecting and intervening in PyTorch models," GitHub repository, 2022. [Online]. Available: https://github.com/davidbau/baukit
[9] LUMI Supercomputer, "LUMI: One of the EuroHPC world-class supercomputers," [Online]. Available: https://lumi-supercomputer.eu [Accessed: Jul. 18, 2025].
[10] Barcelona Supercomputing Center, "MareNostrum 5," [Online]. Available: https://www.bsc.es/marenostrum/marenostrum-5 [Accessed: Jul. 18, 2025].
[11] NVIDIA, "NVIDIA RTX A6000 Datasheet," 2023. [Online]. Available: https://www.nvidia.com/en-us/design-visualization/rtx-a6000/ [Accessed: Jul. 18, 2025].
[12] Canonical, "Ubuntu 22.04 LTS released," Apr. 21, 2022. [Online]. Available: https://canonical.com/blog/ubuntu-22-04-lts-released [Accessed: Jul. 18, 2025].
[13] J. Nickolls, I. Buck, M. Garland, and K. Skadron, "Scalable parallel programming with CUDA," ACM Queue , vol. 6, no. 2, pp. 40–53, 2008. DOI: 10.1145/1365490.1365500.
[14] Anyscale Inc., "Overview —Ray," [Online]. Available: https://docs.ray.io/en/latest/ray-overview/index.html [Accessed: Jul. 12, 2025].
[15] MinIO Inc., "MinIO: High-performance, S3 compatible object storage," [Online]. Available: https://min.io/ [Accessed: Jul. 12, 2025].
[16] Grafana, "Grafana Loki" [Online]. Available: https://grafana.com/oss/loki/ [Accessed: Jul. 22, 2025].
[17] Prometheus, "Prometheus: Overview," [Online]. Available: https://prometheus.io/docs/introduction/overview/ [Accessed: Jul. 22, 2025].
[18] influxdata, "InfluxDB Overview" [Online]. Available: https://www.influxdata.com/products/influxdb-overview/ [Accessed: Jul. 22, 2025].
[19] Grafana, "Grafana" [Online]. Available: https://grafana.com/grafana/ [Accessed: Jul. 22, 2025].
[20] NVIDIA Corporation, "NCCL –NVIDIA Collective Communications Library," [Online]. Available: https://developer.nvidia.com/nccl [Accessed: Aug. 3, 2025].
[21] Advanced Micro Devices, Inc., "RCCL –ROCm Collective Communication Library," [Online]. Available: https://rocm.docs.amd.com/projects/rccl [Accessed: Aug. 3, 2025].
[22] J. Fiotto-Kaufman et al., NDIF, "NDIF: Development Guide," [Online]. Available: https://github.com/ndif-team/ndif [Accessed: Jul. 12, 2025].
[23] Hugging Face, "GPT-2," [Online]. Available: https://huggingface.co/openai-community/gpt2 [Accessed: Jul. 22, 2025].
[24] Hugging Face, "DeepSeek-R1-Distill-Llama-8B," [Online]. Available: https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-8B [Accessed: Jul. 22, 2025].
[25] Hugging Face, "DeepSeek-R1-Distill-Llama-70B," [Online]. Available: https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-70B [Accessed: Jul. 22, 2025].
[26] Kurtzer, G.M., Sochat, V. and Bauer, M.W., "Singularity: Scientific containers for mobility of compute," *PLOS ONE*, 2017. [Online]. Available: https://doi.org/10.1371/journal.pone.0177459 [Accessed: Jul. 22, 2025].
[27] Heimersheim, Stefan and Nanda, Neel, "How to use and interpret activation patching," [Online]. Available: https://arxiv.org/abs/2404.15255, doi: https://doi.org/10.48550/arXiv.2404.15255 [Accessed: Jul. 20, 2025].
[28] nostalgebraist, "interpreting GPT: the logit lens," [Online]. Available: https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens [Accessed: Jul. 20, 2025].
[29] Marks, Samuel; Rager, Can; Michaud, Eric J.; Belinkov, Yonatan; Bau, David; Mueller, Aaron, "Sparse Feature Circuits: Discovering and Editing Interpretable Causal Graphs in Language Models," CoRR, vol.abs/2403.19647, 2024, [Online]. Available: https://arxiv.org/abs/2403.19647, doi: https://doi.org/10.48550/arXiv.2403.19647 [Accessed: Jul. 20, 2025].
[30] Tigges, Curt; Hanna, Michael; Yu, Qinan; Biderman, Stella, "LLM Circuit Analyses Are Consistent Across Training and Scale," CoRR, vol. abs/2407.10827, 2024, [Online]. Available: https://arxiv.org/abs/2407.10827, doi: https://doi.org/10.48550/arXiv.2407.10827 [Accessed: Jul. 20, 2025].
[31] Alibaba DAMO, "QwenLM," [Online]. Available: https://github.com/QwenLM/Qwen [Accessed: Jul. 20, 2025].
[32] Google DeepMind, "Gemma," [Online]. Available: https://deepmind.google/models/gemma/ [Accessed: Jul. 20, 2025].
[33] Nelson Elhage, Neel Nanda, Tom Henighan, Chris Olah, Others, "A Mathematical Framework for Transformer Circuits," [Online]. Available: https://transformer-circuits.pub/2021/framework/index.html [Accessed: Jul. 20, 2025].
[34] www.neuronpedia.org, "Neuronpedia," [Online]. Available: https://www.neuronpedia.org [Accessed: Jul. 20, 2025].
[35] www.semianalysis.com, "AMD vs NVIDIA Inference Benchmark: Who Wins? –Performance & Cost Per Million Tokens," [Online]. Available: https://semianalysis.com/2025/05/23/amd-vs-nvidia-inference-benchmark-who-wins-performance-cost-per-million [Accessed: Jul. 27, 2025].