OpenAI“通用验证器”：构建可信赖AI系统的多维度验证策略

摘要

本报告旨在深入探讨OpenAI在构建“通用验证器”方面的多层面策略，该概念并非指单一产品，而是一系列旨在提升先进AI系统可信赖性、安全性与可解释性的互联研究与工程举措。报告将详细阐述OpenAI如何通过“证明者-验证者博弈”（Prover-Verifier Games）提升大型语言模型（LLM）的推理可读性，如何利用Responses API中的“护栏”（Guardrails）和“追踪与可观察性”（Tracing & Observability）确保代理（agentic）应用的安全性，以及在AI对齐研究中采用的可扩展监督方法。此外，报告还将分析内容凭证（Content Credentials）在保障AI生成内容来源和真实性方面的作用，并审视形式化验证与AI审计框架在整体可信赖性生态系统中的地位。这些举措共同反映出OpenAI对AI信任问题的系统性理解，即AI可信赖性需要多管齐下的策略，以应对AI行为（如推理、代理能力、数据完整性）的不同层面，而非依赖单一的万能解决方案。这种分布式方法凸显了AI安全固有的复杂性及其对持续适应性研究与开发的需求，预示着未来AI系统将从“黑箱”转变为内部状态和外部行为均可验证的“透明箱”。

引言：AI可信赖性的挑战与OpenAI“通用验证器”的愿景

AI系统在关键领域应用的增长与信任需求

随着机器学习系统日益渗透到医疗、金融、自动驾驶等高风险和高影响力领域，确保其输出结果的可信赖性、准确性和可解释性已成为一项刻不容缓的挑战[1]。在这些关键应用中，AI的决策可能对人类福祉、经济稳定乃至社会公平产生深远影响。例如，在医疗诊断中，AI的误判可能导致严重的健康后果；在金融交易中，算法偏差可能引发巨大的经济损失；在自动驾驶中，系统故障则直接威胁生命安全。因此，对AI系统建立深厚的信任是其被广泛接受

和安全部署的基石[6]。

目前，AI能力与人类对其信任度之间存在着显著的差距。这种"信任鸿沟"构成了AI技术实现其全部潜力的关键障碍。仅仅提升AI的技术性能不足以弥合这一差距；AI系统还必须能够以人类可理解的方式解释其决策过程，并提供可验证的保障，以证明其行为符合预期且无害。这种对透明度、可解释性和可验证性的需求，使得AI的可信赖性不仅是一个技术问题，更是一个深刻的社会和伦理问题。如果AI系统无法赢得用户的信任，它们将难以在现实世界中实现大规模、负责任的集成，尤其是在那些决策结果可能产生重大影响的场景中。因此，对AI可信赖性的关注，是推动AI从实验室走向社会，并确保其真正造福人类的关键一步。

OpenAI在**AI**安全、对齐与可信赖性方面的使命

OpenAI自成立以来，便将构建安全且有益的通用人工智能（AGI）作为其核心使命[8]。这一宏大愿景驱动着其在深度学习、大规模数据处理和高级推理能力方面的研究，同时将AI安全和对齐（alignment）置于其发展战略的核心。OpenAI认识到，随着AI系统能力的指数级增长，其潜在风险也随之增加，因此必须采取前瞻性的方法来管理这些风险。

OpenAI的AI安全策略并非被动的风险应对，而是一种积极、迭代且具有对抗性的方法。公司通过一系列严谨的内部评估和与外部专家的协作，持续增强其AI系统的保障措施[9]。这包括但不限于：

- 红队测试（**Red Teaming**）：OpenAI积极邀请专家团队对AI系统进行攻击性测试，旨在发现并利用潜在的漏洞，从而在系统部署前识别并修复其弱点[9]。这种主动寻求系统缺陷的做法，体现了对潜在失败模式的深刻理解，超越了简单的错误修复，致力于预测更复杂的失调或滥用形式。
- 系统卡片（**System Cards**）：提供关于AI模型功能、限制和预期用途的详细文档，以提高透明度并指导安全使用[9]。
- 准备评估（**Preparedness Evals**）：定期评估AI系统应对未来高风险场景的能力，以确保其在能力提升的同时，安全防护措施也能同步升级[9]。

这种对"红队测试"和"准备评估"的强调，揭示了OpenAI在AI安全方面采取的成熟策略。它表明OpenAI不仅在修复当前问题，更在进行基础性研究，以管理未来AI能力增长可能带来的潜在风险。这种前瞻性立场，将"通用验证器"的努力置于减轻潜在生存风险的更广阔背景之下。它也意味着AI安全并非一次性完成的任务，而是一个需要持续投入和适应性调整的动态过程。通过这种对抗性方法，OpenAI致力于构建能够抵御各种已知和未知攻击的AI系统，从而为AGI的负责任发展奠定基础。

“通用验证器”概念的范畴与重要性

在OpenAI的语境中，“通用验证器”并非指代某一个单一的软件或模型，而是一个涵盖广泛技术和方法论的集合，其共同目标是提升AI系统的可信赖性、可解释性和安全性[1]。这个概念的“通用性”体现在其验证范围的全面性，从AI模型内部的推理过程到其在外部环境中的代理行为，再到其生成内容的数据完整性。

这种全面的验证方法至关重要，因为它认识到AI系统的复杂性要求多层次、多维度的保障。一个强大的AI系统，其可靠性取决于其所有组成部分的稳健性。因此，“通用验证器”的范畴包括：

- **Prover-Verifier Games**：专注于提高大型语言模型内部推理过程的可读性和可检查性[1]。
- **Responses API**中的护栏（**Guardrails**）和追踪与可观察性（**Tracing & Observability**）：确保AI代理（agentic AI）在与现实世界交互时的行为安全和可控[14]。
- **AI**对齐研究中的可扩展监督（**Scalable Oversight**）：开发能够有效监督超人类AI系统的方法，以确保其行为符合人类价值观[16]。
- 可解释**AI**（**Explainable AI, XAI**）技术：旨在揭示AI决策的内部机制，增强透明度和人类对AI的信任[6]。
- 形式化验证（**Formal Verification**）和**AI**审计框架（**AI Auditing Frameworks**）：提供数学上的严格性保障和系统性的风险管理与合规性检查[2]。
- 内容凭证（**Content Credentials**）：用于验证AI生成图像等内容的来源和真实性，以打击虚假信息[15]。

这种对“通用性”的理解，超越了狭义的技术实现，触及到AI治理和伦理的核心。它意味着AI系统不再仅仅是“黑箱”——即其内部运作对人类不透明的系统——而是朝着“透明箱”的方向发展，其内部状态和外部行为都能够被验证和理解。这种转变，如果能够实现，将从根本上改变AI的开发、部署和监管方式，推动其走向一个“可验证AI”的新范式。在这个范式下，信任将不再仅仅基于性能，而是基于对AI行为和决策过程的深入理解和可证明的保障。

第一部分：核心验证机制——Prover-Verifier Games与LLM可读性

Prover-Verifier Games (PVG) 的理论框架与起源

“证明者-验证者博弈”（Prover-Verifier Games, PVG）是OpenAI为提升大型语言模型（LLM）输出可信赖性而提出的一项创新性理论框架和训练方法。它借鉴了博弈论中的概念，特别是交互式证明系统，旨在优化LLM生成解决方案的可检查性（checkability），从而显著提高其对人类的可读性（legibility）[1]。

该框架的核心灵感来源于Anil et al. (2021) 提出的Prover-Verifier Game。在这个博弈中，存在两个主要参与者：

- 证明者（**Prover**）：一个大型且能力强大的LLM，负责生成问题的解决方案或推理链。
- 验证者（**Verifier**）：一个相对较小且能力较弱的LLM，其任务是独立地检查证明者提供的解决方案的正确性。

PVG的独特之处在于其“验证者先行”（verifier-leading）的Stackelberg博弈设定 [1]。这意味着验证者首先制定其策略，然后证明者在已知验证者策略的情况下，尝试做出最佳响应。这种动态关系并非简单的错误检测，而是一种主动塑造证明者行为的机制。验证者通过不断学习如何更有效地识别错误和欺骗性输出，迫使证明者生成不仅正确，而且易于验证和理解的解决方案。这种设计选择将验证机制从被动检查转变为主动行为塑造，代表了AI训练中一种更为复杂的对齐方法。它促使证明者在生成答案时，不仅追求最终结果的正确性，更要注重中间步骤的逻辑连贯性和可解释性，从而使其输出更符合人类的直觉和理解方式。

可检查性 **(Checkability)**、完备性 **(Completeness)** 与可靠性 **(Soundness)** 的定义与意义

在PVG框架中，为了量化和优化LLM输出的可信赖性，OpenAI借鉴了形式化验证和交互式证明系统中的核心概念，并根据LLM的特性进行了调整 [1]。

- 可检查性（**Checkability**）：这是PVG的核心目标，指的是验证器能够独立地判断证明器所生成解决方案的正确性的能力 [1]。一个高可检查性的解决方案意味着其内部逻辑清晰，易于被审查，无论是通过自动化工具还是人类专家。
- 完备性（**Completeness**）：在理想情况下，完备性意味着对于任何给定领域内的有效输入，一个“有益证明者”（helpful prover）能够生成正确的解决方案，并且验证者能够准确地接受这些解决方案为正确 [1]。然而，在处理LLM的非正式领域时，模型和人类都无法达到完美的完备性。因此，OpenAI研究的是其“放松版本”，即有益证明器应具有高通过率，而验证器对有益样本应具有高召回率。

- 可靠性（**Soundness**）：理想的可靠性要求验证者能够拒绝任何由“对抗性证明者”（adversarial provers）生成的、属于相同模型类别的所有不正确解决方案[1]。同样，在LLM的背景下，研究的是其“放松版本”，即验证器对通过有限优化生成的对抗性输入应具有低误报率。这确保了验证器不会轻易被看似正确的错误答案所欺骗。

这些概念的定义借鉴了概率近似正确（Probably-Approximately Correct, PAC）验证和交互式证明系统等文献[1]。传统的形式化验证（如在软件工程中）追求的是绝对的数学保证，即100%的正确性和无懈可击的逻辑。然而，LLM的生成特性和在非正式领域（如自然语言推理）的应用，使得这种完美的、确定性的保证难以实现且成本高昂。因此，对“完美完备性”和“可靠性”的放松，承认了LLM输出固有的概率性和非正式性，从而将形式化验证的严格性与LLM的实际应用相结合，形成了一种更为务实的“足够好”的可读性标准。这种务实的适应性对于将形式化概念应用于LLM的复杂现实至关重要。

当一个证明器被认为对验证器“可读”时，这意味着其输出不仅在技术上是正确的，而且其呈现方式使得人类能够轻易地接受其为正确，并且不会被那些看似正确但实则错误的输出所蒙蔽[1]。这突出显示了AI安全中以人为本的关键方面。它不仅仅关乎AI的正确性，更关乎人类能否有效地验证这种正确性，这对于建立信任和实现有效监督至关重要。这种可读性直接解决了AI“黑箱”问题[6]，使得AI的决策过程不再是神秘的，而是可理解和可审查的。

Checkability训练算法：迭代过程与**“helpful”/“sneaky”provers**

为了实现LLM输出的高可检查性和可读性，OpenAI设计了一种迭代式的Checkability训练算法，其核心在于引入了两种不同角色的证明者：“有益证明器”（helpful provers）和“狡猾证明器”（sneaky provers）[1]。

该算法通过以下迭代过程进行：

1. 小型验证器训练：首先，训练一个小型验证器来预测证明器生成解决方案的正确性。这个验证器会随着训练的进行不断提升其识别正确和错误解决方案的能力。
2. 证明者训练：
    - 有益证明器（**Helpful Prover**）：这个证明器被训练来生成正确的解决方案，并使其能够被验证器高度接受。训练目标是让其输出不仅正确，而且清晰、易于理解和验证。
    - 狡猾证明器（**Sneaky Prover**）：这是该训练机制的创新之处。狡猾证明器被训练来生成不正确的解决方案，但这些解决方案却能够成功欺骗当前的验证器，使其误判为正确[22]。

在训练过程中，有益证明器的准确性和验证器对对抗性攻击的鲁棒性都会逐步提高[21]。最

初，狡猾证明器可能很容易地生成能够欺骗验证器的错误解决方案。然而，随着这些欺骗性样本被加入到验证器的训练数据中，验证器会不断学习并变得更加强大，从而迫使狡猾证明器开发出更微妙、更难以察觉的错误策略。这种对抗性训练范式，类似于一种进化军备竞赛，使得验证系统在内部训练循环中不断提升其鲁棒性和泛化能力。狡猾证明者的角色，本质上充当了一种内部的"红队机制"，主动探测验证者的弱点，并迫使其持续改进。这种方法对于开发高度稳健和可泛化的验证系统至关重要，因为它模拟了在实际部署中可能遇到的各种复杂和对抗性场景。

训练过程会持续进行，直到狡猾证明器无法再成功欺骗验证器为止[22]。此时，验证器已经变得足够强大，能够识别即使是最微妙的错误。

这项训练的另一个重要发现是其对人类可读性的转移效应[21]。研究表明，当人类评估者检查经过这种Checkability训练的有益证明器生成的解决方案时，他们的准确性会显著提高；相反，在检查狡猾证明器生成的解决方案时，人类的准确性会降低。这表明，为机器可验证性进行优化，可以间接导致人类可理解的输出，从而弥合了技术对齐与实际人类监督之间的鸿沟。这一发现至关重要，因为它表明机器辅助验证可能成为一种可扩展的解决方案，用于未来需要人类参与的AI系统。通过这种方式，AI不仅能够自我检查，还能生成更易于人类理解和信任的输出，从而在可扩展性和可信赖性之间找到一个平衡点。

过程监督 **(Process Supervision)** 与结果监督 **(Outcome Supervision)** 的对比分析及其对**LLM**推理的影响

在训练大型语言模型（LLM）以解决复杂的多步骤推理问题时，OpenAI深入研究了两种主要的监督范式：结果监督（Outcome Supervision）和过程监督（Process Supervision），并发现过程监督在提升模型可靠性和可解释性方面具有显著优势[12]。

结果监督 **(Outcome Supervision, ORM)**

- 定义与反馈机制：结果监督下的奖励模型（Outcome-supervised Reward Models, ORM）仅根据模型思维链的最终结果提供反馈。在数学问题解决的背景下，这通常意味着只检查最终答案的正确性[12]。
- 局限性：
  - 反馈精度低：ORM提供的反馈粒度较粗，它只告诉模型最终答案是正确还是错误，而无法指出推理过程中具体哪一步出现了问题。
  - 错位行为风险：模型可能使用不正确的推理过程，却偶然地得到了正确的最终答案

（即“误报”）。这种情况下，ORM会奖励这种行为，导致模型学习到“错位行为”（misaligned behavior），即其内部推理过程与人类期望的逻辑不符[12]。例如，一个数学模型可能通过错误的计算步骤，最终“蒙对”了正确答案。
    - 归因困难：对于不正确的解决方案，ORM难以准确地将错误归因到思维链中的具体环节，这对于模型学习如何纠正错误是一个巨大的挑战，尤其是在复杂问题中。

过程监督 **(Process Supervision, PRM)**

- 定义与反馈机制：过程监督下的奖励模型（Process-supervised Reward Models, PRM）则对思维链中的每一个中间推理步骤都提供反馈。人类数据标注员被要求对模型生成的解决方案的每个步骤进行标记，分为“积极”（正确且合理）、“消极”（不正确或不合理）或“中性”（含糊不清或虽有效但不佳的建议）[12]。
- 优势：
    - 反馈精确：过程监督提供了更精确的反馈，能够明确指出错误发生的具体位置。这使得模型能够更有效地学习和纠正其推理过程中的缺陷。
    - 提升**AI**对齐：通过奖励模型遵循人类认可的思维链，过程监督直接鼓励了模型生成可解释的推理，从而提高了AI的对齐度[12]。这意味着AI不仅给出了正确答案，而且其思考过程也与人类的逻辑和价值观相符。
    - 简化归因：过程监督通过揭示有多少初始步骤是正确的以及第一个错误步骤的确切位置，大大简化了错误归因的任务。
    - **“**负对齐税**”**：研究发现，过程监督不仅能带来更好的性能，甚至可能产生一种“负对齐税”（negative alignment tax）[12]。这表明，投入于过程层面的反馈不仅能够提升AI的安全性和可解释性，还能同时提高模型的性能。这一发现颠覆了传统观念中安全/对齐与能力/效率之间存在权衡的假设，创造了一个良性循环，激励开发者采纳这些安全措施，不仅出于伦理考量，更出于性能提升的驱动。这种深层理解AI内部运作方式的能力，可以开启新的性能前沿。
    - 缓解错位行为：过程监督能够有效缓解模型使用不正确推理得出正确答案的“错位行为”[12]。这强调了AI如何得出解决方案的重要性，而不仅仅是解决方案本身。对于高风险应用而言，可验证的推理过程至关重要。

比较分析总结

下表总结了过程监督与结果监督在不同维度的对比：

表2：过程监督与结果监督：比较分析

方面	结果监督（Outcome Supervision）	过程监督（Process Supervision）
反馈粒度	仅对最终结果提供反馈，粒度粗糙。	对每个中间推理步骤提供反馈，粒度精细。
错误归因	难以准确归因错误发生位置，归因任务困难。	明确指出错误步骤，简化归因任务。
错位行为风险	存在模型使用错误推理得出正确答案的风险，导致错位行为。	显著缓解错位行为，鼓励模型遵循人类认可的推理过程。
可解释性	较低，难以理解模型内部决策过程。	较高，模型推理过程更透明、可理解。
数据收集成本/效率	成本相对较低，但反馈价值可能受限。	数据收集成本较高，但通过主动学习可提高效率。
性能影响	可能导致性能提升，但存在错位风险。	带来性能提升，甚至可能产生“负对齐税”，同时提高对齐度。
主要目标	关注最终答案的正确性。	关注推理过程的正确性、合理性与可解释性。

奖励模型 (Reward Models) 在数学推理验证中的应用

奖励模型在OpenAI验证大型语言模型（LLM）的数学推理能力中扮演着核心角色。这些模型经过训练，能够区分LLM输出中期望和不期望的行为，从而为强化学习（RL）流程提供指导，或通过拒绝采样（rejection sampling）进行搜索 [12]。整个系统的可靠性直接取决于奖励模型自身的可靠性 [12]。

生成器模型与评估

在验证数学推理时，首先使用一个固定的生成器模型来生成所有解决方案，这些解决方案以逐行、分步的格式呈现。该生成器模型在MATH训练问题上进行微调，这些问题能够得出正确答案，主要目的是使其学习到所需的输出格式，而非获得新的技能[12]。

奖励模型的评估通过其执行“最佳N选择”（best-of-N search）的能力来完成。具体而言，对于每个测试问题，从生成器模型中均匀采样的N个解决方案中，选择奖励模型评分最高的那个。然后，根据最终答案自动对这个选定的解决方案进行评分，并报告正确解决方案的比例。一个更可靠的奖励模型将更频繁地选择正确的解决方案[12]。

过程监督的数据收集 **(PRM800K)**

为了训练高效的过程监督奖励模型（PRM），OpenAI投入了大量精力进行高质量的数据收集。

- 人类数据标注员：人类标注员被要求审查由大型生成器模型生成的逐步解决方案。他们的任务是根据每个步骤的正确性和合理性，将其标记为“积极”、“消极”或“中性”[12]。
- 数据集规模：为支持相关研究而发布的PRM800K数据集，包含了80万个步骤级别的真人反馈标签，涵盖了1.2万个问题对应的7.5万个解决方案[12]。这个大规模数据集为训练鲁棒的PRM提供了坚实的基础。

策略性选择解决方案（主动学习 ）

在数据收集过程中，一个关键的挑战是如何有效地利用有限的人类标注资源。如果随机选择解决方案进行标注，那些包含明显错误的解决方案所提供的反馈价值较低。为了提高数据效率，研究人员采取了“主动学习”（active learning）策略，即策略性地选择“有说服力的错误答案解决方案”（convincing wrong-answer solutions）来呈现给数据标注员[12]。

- **“有说服力”**：指那些被当前最佳PRM评为高分（即模型认为它们很可能是正确的）的解决方案。
- **“错误答案”**：指这些解决方案的最终答案实际上是错误的（其正确性由最终答案检查确定）。

这种策略背后的逻辑在于，标注“有说服力的错误答案解决方案”能够提供最有价值的信息。因为在这种情况下，已知PRM至少在一个步骤上是错误的，这意味着人类反馈将直接指向模型理解的盲点或薄弱环节。这是一种“元学习”或“学习如何学习”的验证系统本身的形式。通过专注于“难题”，人类反馈的价值被最大化，从而加速了证明者和验证者的学习过

程，并使得验证器更加鲁棒。这是AI如何通过智能地利用人类输入来提高自身安全性和可靠性的一个关键例证。

PRM会使用最新收集的数据进行迭代式再训练。在每次迭代中，选择一定数量（例如，前K个）最具说服力的错误答案解决方案呈现给标注员。这种持续的迭代改进循环，表明AI安全验证并非一个静态的终点，而是一个持续不断的过程，尤其是在模型不断演进和新的故障模式不断出现的情况下。这种动态方法对于跟上AI能力快速发展的步伐至关重要。

这种主动学习方法显著提高了过程监督的数据效率，带来了2.6倍的提升 [12]。通过这种智能化的数据选择，OpenAI能够以更低的成本和更高的效率训练出更可靠、更对齐的LLM，从而在复杂推理任务中提供更值得信赖的解决方案。

表1：**Prover-Verifier Game** 关键概念

概念	定义/目的	在PVG中的作用
可读性 **(Legibility)**	LLM输出对人类清晰且易于检查的属性。	最终目标，通过可检查性训练实现。
可检查性 **(Checkability)**	验证器独立判断证明器解决方案正确性的能力。	核心机制，通过迭代博弈训练提升。
完备性 **(Completeness) (**放松版本**)**	有益证明器生成正确解决方案，验证器高召回率接受。	确保AI能生成正确且可接受的输出。
可靠性 **(Soundness) (**放松版本**)**	验证器对对抗性证明器生成的错误解决方案具有低误报率。	确保AI能有效识别并拒绝不正确的输出。
证明者 **(Prover)**	一个大型LLM，生成问题的解决方案或推理链。	尝试生成解决方案，其行为受验证者策略影响。
有益证明器 **(Helpful Prover)**	证明者的一种角色，旨在生成验证器接受的正确解决方案。	学习生成正确且易于验证的输出。
狡猾证明器 **(Sneaky Prover)**	证明者的一种角色，旨在生成能欺骗验证者的不正确解决方案。	作为内部红队机制，迫使验证器不断学习和提升鲁棒性。
验证者 **(Verifier)**	一个小型LLM，独立检查证明者解决方案的正确性。	设定标准，主动塑造证明者行为，不断学习识别错误和欺骗。
Checkability 训练	一种迭代算法，通过有益和狡猾	整体训练流程，旨在提升LLM输

	证明者训练验证器和证明者。	出的可检查性和对人类的可读性。

第二部分：AI代理的验证与安全：Responses API与Agentic AI

Responses API：构建安全Agentic AI的基础

OpenAI推出的Responses API是其在构建安全、可信赖AI系统方面的重要进展，尤其针对日益复杂的代理（agentic）应用[14]。这个新的API原语旨在简化代理应用的开发，它融合了Chat Completions API的简洁性与Assistants API的工具使用能力，为开发者构建智能代理提供了更灵活的基础。通过单次Responses API调用，开发者能够利用多种工具和模型轮次来解决日益复杂的任务。

Responses API的设计理念在于将OpenAI模型与内置工具无缝结合，而无需集成多个API或外部供应商。这不仅提升了开发效率，也使得在OpenAI平台上存储数据以评估代理性能变得更加便捷，例如通过追踪（tracing）和评估（evaluations）功能。值得注意的是，OpenAI默认不会使用商业数据来训练模型，即使这些数据存储在OpenAI平台上[14]。

Responses API支持多项内置工具，这些工具旨在将模型连接到现实世界，使其在完成任务时更加实用：

- 网络搜索（**Web Search**）：提供快速、最新的答案，并附带清晰相关的引用来源。这对于购物助手、研究代理和旅行预订代理等应用场景至关重要[14]。
- 文件搜索（**File Search**）：允许开发者从大量文档中检索相关信息，支持多种文件类型、查询优化、元数据过滤和自定义重排序，以提供快速准确的搜索结果。这在客户支持、法律助理和代码审查等领域具有广泛应用[14]。
- 计算机使用（**Computer Use**）：这项工具使代理能够执行计算机任务，例如自动化浏览器工作流（如Web应用质量保证）或跨传统系统的数据录入任务。这项功能目前处于研究预览阶段，仅对部分开发者开放[14]。

这些内置工具的引入，体现了OpenAI对AI系统与外部环境交互能力的重视。通过提供标准化的接口，AI代理能够更有效地获取信息、执行操作，从而扩展其应用范围。

Guardrails与**Tracing & Observability**：代理行为的保障与洞察

在构建AI代理应用时，确保其行为安全、可控且可理解是至关重要的。Responses API及其配套的Agents SDK为此提供了关键的保障机制，即“护栏”（Guardrails）和“追踪与可观察性”（Tracing & Observability）[14]。

护栏 **(Guardrails)**

护栏是可配置的安全检查机制，用于对AI代理的输入和输出进行验证 [14]。它们是确保AI代理行为安全的基石，通过允许开发者定义规则和约束来防止意外或有害的输出。这对于高风险应用尤其重要，例如，防止代理生成不当内容 [26]、执行未经授权的操作，或在敏感任务中出现偏差。护栏的实施，使得开发者能够对AI代理的行为施加更精细的控制，从而降低潜在的风险。

追踪与可观察性 **(Tracing & Observability)**

追踪与可观察性工具提供了对AI代理执行流程的深度可见性，对于调试和优化性能至关重要 [14]。由于LLM应用通常涉及复杂的链式操作和非确定性流程，传统的日志记录和指标监控往往不足以提供全面的诊断信息 [28]。

- 追踪（**Tracing**）：它捕获请求在系统中的完整路径，记录从用户初始操作到最终响应的每一步。这对于理解LLM应用中复杂的链条和非确定性扭曲至关重要 [28]。通过追踪，开发者能够精确地了解代理如何执行任务，识别问题所在，并改进效率和可靠性。
- 跨度（**Spans**）：是特定操作运行的独立时间片，例如工具调用、数据库查询或LLM完成。每个跨度包含名称、时间数据、结构化日志和提供操作上下文的属性 [28]。捕获跨度有助于团队分析延迟、跟踪成本，并将模型行为与下游系统性能关联起来，为有效的调试和优化提供清晰的图景。
- 会话（**Sessions**）和会话评估（**Session Evals**）：会话通过会话ID将多个追踪组合在一起，将其与单个对话或用户旅程关联起来。这使得用户可以识别对话何时“中断”或悄悄偏离轨道，即使单个请求在隔离状态下看起来正常 [28]。会话级别的评估则衡量整个交互的质量，例如代理是否保持了逻辑流程、是否记住了早期轮次的信息，以及是否帮助用户实现了目标。

- **OpenTelemetry (OTel)** 和 **OpenInference**：OpenInference是一个开源框架，用于从AI代理和LLM驱动的工作流中捕获详细的遥测数据（如追踪和跨度），并将其映射到标准化属性 [28]。通过与行业标准系统OpenTelemetry集成，团队可以轻松收集和导出这些追踪数据，然后由Arize AX等可观察性平台进行摄取、存储和可视化，从而提供代理如何思考和行动的实时洞察 [28]。

这些工具的结合，使得开发者能够从“黑箱”模型中获得前所未有的可见性，理解AI代理的内部决策过程和外部行为。这种透明度对于确保AI系统的可信赖性至关重要，尤其是在模型可能产生“幻觉”或出现意外行为的情况下 [25]。通过护栏和可观察性，OpenAI致力于构建不仅能够执行复杂任务，而且其行为可验证、可审计和可控的AI代理，从而在AI能力不断增长的同时，最大限度地降低风险。

AI生成内容的可验证性：内容凭证

随着生成式AI（Generative AI）模型能力的飞速提升，其生成的内容（如图像、文本、音频）的质量已达到令人难以分辨的程度，这带来了虚假信息和内容真实性方面的严峻挑战。为应对这一挑战，OpenAI通过Azure OpenAI服务推出了“内容凭证”（Content Credentials）机制，旨在提高AI生成内容的透明度和可验证性 [15]。

内容凭证为AI生成图像提供了“防篡改”的方式，以披露内容的来源和历史。这些凭证以附加到图像上的“清单”（manifest）形式存在，并由可追溯到Azure OpenAI的证书进行加密签名，从而确保其真实性和不可篡改性 [15]。

该清单包含几个关键信息：

- **"description"**：此字段对于所有AI生成图像都设置为"AI Generated Image"，明确表明其AI生成性质 [15]。
- **"softwareAgent"**：此字段标识生成图像的具体模型，例如"Azure OpenAI DALL-E"或"Azure OpenAI ImageGen" [15]。
- **"when"**：此字段记录内容凭证创建的时间戳，提供了内容的生成时间信息 [15]。

内容凭证的目的是提高AI生成内容来源的透明度，尤其是在生成式AI模型产出内容质量不断提高的背景下。它们提供了一种可验证的方法来揭示内容的来源和历史。该系统基于内容溯源与真实性联盟（Coalition for Content Provenance and Authenticity, C2PA）的开放技术规范 [15]。通过使用内容凭证，人们可以清楚地了解视觉内容是否由AI生成。

对于用户而言，使用内容凭证无需额外设置，因为它们会自动应用于Azure OpenAI中DALL·E和GPT-image-1模型生成的所有图像 [15]。要验证图像的凭证，用户可以通过内容凭

证验证网页（contentcredentials.org/verify），该网页将显示凭证由Microsoft Corporation发行以及发行日期和时间；或者使用内容真实性倡议（Content Authenticity Initiative, CAI）的开源工具来验证和显示C2PA内容凭证 [15]。

这项技术是AI可信赖性策略的重要组成部分，它将验证的范围从AI模型内部行为扩展到AI生成内容的外部属性。通过提供清晰的来源信息，内容凭证有助于打击虚假信息、增强数字媒体的真实性，并在用户和AI生成内容之间建立信任。这反映出OpenAI在应对AI技术带来的社会挑战方面的积极姿态，并致力于构建一个更加透明和负责任的AI生态系统。

第三部分：AI对齐与可解释性：确保AI符合人类价值观

AI对齐研究：从价值观到可扩展监督

AI对齐是OpenAI乃至整个AI领域最核心的挑战之一，其目标是确保AI系统，特别是未来可能出现的超人类智能（superhuman AI），能够始终按照人类的价值观、伦理和意图行事 [16]。随着AI系统从狭义的、针对特定任务的工具发展为更通用、更自主的代理，对齐问题变得愈发关键。它不仅涉及技术实现，还涵盖了哲学和伦理层面，必须加以解决才能确保AI的安全和有益发展 [31]。

AI对齐面临多重技术挑战：

- 规范问题（**Specification Problem**）：将复杂、细致入微且可能相互冲突的人类价值观准确地转化为AI系统可理解和优化的数学目标函数是一个巨大挑战 [31]。人类偏好往往是动态变化的，边缘情况难以穷尽，且存在道德不确定性。
- 可扩展性（**Scalability**）：当前的对齐技术往往难以随着模型复杂度的增加而扩展。训练大型语言模型和多模态系统所需的计算资源，以及保持对齐所需的资源，通常增长速度快于模型能力本身 [31]。
- 分布偏移（**Distribution Shift**）：AI系统必须在训练数据未涵盖的、前所未有的场景中保持对齐，并避免负面泛化 [31]。
- 内部对齐（**Inner Alignment**）与外部对齐（**Outer Alignment**）：
    - 外部对齐：关注确保指定的目标函数准确捕捉了人类希望AI系统做什么 [31]。这包括正确指定奖励函数或训练目标，确保形式化规范与真实人类意图（包括边缘情况和意外场景）相符，并将复杂的人类价值观转化为数学目标，同时保留其细微之处。
    - 内部对齐：关注确保学习到的行为真正优化了指定的目标 [31]。这涉及防止在训练过

程中出现潜在的“优化恶魔”或“元优化器”，它们可能通过复杂的奖励欺骗（reward hacking）来发展自己的失调目标或代理目标。

这些挑战促使OpenAI及其他研究机构探索“可扩展监督”（Scalable Oversight）方法。可扩展监督旨在提供对AI系统的有效监督，特别是当AI系统能力超越人类表现时[17]。其核心思想是，通过部分或完全用AI系统生成的反馈来替代人类反馈，从而克服强化学习中人类反馈（RLHF）的成本高昂和难以扩展的局限性[17]。

强化学习与AI反馈（RLAIF）和宪法AI（Constitutional AI, CAI）是可扩展监督的典型示例[17]。在RLAIF中，反馈完全由AI模型生成。CAI则进一步将人类监督限制在最初起草一组原则（即“宪法”），然后AI模型根据这些原则进行自我批判和修订。例如，一个原则可以是“请选择最有帮助、最诚实、最无害的回答”[17]。CAI分两个阶段利用这些原则：首先是监督学习阶段，LLM被要求根据原则批判和修改其有害提示的答案；然后是强化学习阶段，LLM根据宪法原则生成自己的反馈[17]。

可扩展监督方法具有巨大潜力，它能够加速和改进对齐过程，确保LLM在保持有用性的同时，尊重个人隐私权[17]。这对于处理个人数据的系统尤为重要，因为仅靠人工监督可能不足以应对其普遍性带来的合规和滥用风险。然而，可扩展监督也存在挑战，例如如果作为评估者的LLM本身存在偏见，这些偏见可能会传递给奖励模型，从而导致开发中的LLM也出现偏见[17]。此外，高性能LLM可能表现出强烈的自我偏好，偏爱自己的答案而非其他LLM或人类的答案，这可能放大嵌入的偏见。

可解释**AI (XAI)** 方法：提升**AI**系统信任度

可解释人工智能（Explainable AI, XAI）旨在通过揭示AI系统决策过程的内部机制，从而增强其透明度、可理解性和可信赖性[6]。在AI系统日益复杂且在关键领域广泛应用的背景下，XAI变得至关重要，因为它有助于弥合AI的“黑箱”特性与人类对其决策的理解和信任之间的差距。

XAI通过多种方法和技术来提升AI系统的信任度：

透明度与可解释性

XAI的核心目标是使AI模型的决策过程对开发者和用户透明且可解释。例如，在医疗诊断系统中，XAI技术如**特征重要性评分（feature importance scores）或决策树（decision

trees）**可以揭示哪些症状或测试结果影响了AI的预测 [7]。这种清晰度有助于医疗专业人员验证AI的推理是否符合既定的医学知识，从而减少对不透明“黑箱”模型的依赖。在没有这些解释的情况下，尤其是在医疗或金融等关键领域，利益相关者可能会对采用AI工具犹豫不决 [6]。

可问责性与偏差纠正

XAI还通过使开发者能够识别和纠正模型中的错误或偏差来促进问责制 [7]。例如，如果一个贷款审批模型基于邮政编码等不相关因素不公平地拒绝申请，

SHAP (SHapley Additive exPlanations) 或 **LIME (Local Interpretable Model-agnostic Explanations)** 等技术可以揭示这一缺陷 [7]。开发者随后可以调整训练数据或特征以确保公平性。同样，在图像识别系统中，**显著性图（saliency maps）**可以突出显示模型是否关注图像中有意义的部分（例如X射线中的肿瘤），而不是不相关的噪声 [7]。这种程度的审查确保了AI符合伦理和功能要求，使其更易于审计和验证 [33]。

增强用户信任与可操作性洞察

XAI通过提供关于AI行为的可操作性洞察来建立用户信任 [7]。例如，一个推荐系统解释“我们推荐此产品是因为您查看了类似商品”，这赋予了用户对其体验的控制权。在自动驾驶汽车中，诸如“因检测到行人而刹车”的实时解释有助于乘客理解安全关键决策 [7]。开发者还可以利用XAI在测试期间调试模型，例如识别视觉模型在光线不足下错误分类对象的边缘情况。通过弥合复杂算法与人类理解之间的鸿沟，XAI确保AI系统不仅准确，而且是决策中可靠的伙伴。

OpenAI在可解释AI方面的研究，例如其在“通过教学实现可解释机器学习”（Interpretable Machine Learning through Teaching）方面的工作 [35]，旨在使AI能够以人类可理解的方式相互教授概念，并选择最具信息量的例子进行教学。这不仅有助于AI之间的沟通，也为人类理解AI的内部表征提供了途径。通过将解释性融入AI设计和开发流程中，OpenAI致力于构建更值得信赖、更负责任的AI系统，从而推动AI在社会中的广泛和安全应用。

第四部分：AI验证的工程实践与未来展望

形式化验证与AI审计框架

在确保AI系统可靠性和可信赖性方面，形式化验证和AI审计框架扮演着不可或缺的角色。它们提供了系统性的方法来评估、管理和减轻AI相关的风险。

形式化验证（Formal Verification）

形式化验证涉及使用严格的数学技术来证明模型与规范的一致性[2]。这种方法可以包括形式化证明、模型检查和概率验证。其目标是为神经网络的属性提供正式保证，例如鲁棒性、安全性、和正确性[19]。

在深度神经网络的背景下，形式化验证的挑战在于其固有的复杂性和"黑箱"性质。神经网络通常难以解释，并且可能表现出意外行为，例如对抗性样本[19]。为了应对这些挑战，研究人员开发了各种算法和工具：

- 边界传播算法（**Bound Propagation Algorithms**）：如CROWN（Complete and Certifiable Robustness against Adversarial Attacks）及其变体α,β-CROWN，通过计算神经网络输出的上下界来验证其属性[19]。这有助于在不完全理解模型内部所有细节的情况下，对模型的行为提供数学保证。
- 整数规划（**Integer Programming**）和分支定界（**Branch and Bound**）方法：这些是更全面的验证算法，能够对神经网络的属性提供更强的保证，但计算成本也更高[19]。

形式化验证在安全关键型任务中尤为重要，如自动驾驶和医疗保健，在这些领域，AI系统的任何意外行为都可能导致灾难性后果[2]。通过形式化验证，可以在部署前提供特定保证，确保AI代理始终做出理性决策，并遵守预设的安全属性。

AI审计框架（AI Auditing Frameworks）

AI审计框架提供了一套结构化的方法来评估AI系统的治理、数据质量、性能和监控，以确保合规性、公平性和问责制[10]。这些框架对于组织管理AI风险、满足监管要求和建立公众信

任至关重要。

两个主要的AI治理框架是：

- **NIST AI**风险管理框架（**NIST AI Risk Management Framework, AI RMF**）：由美国国家标准与技术研究院（NIST）于2023年发布，是一个自愿性框架，旨在帮助组织在AI系统生命周期中管理风险[20]。它围绕“治理（Govern）”、“映射（Map）”、“衡量（Measure）”和“管理（Manage）”四个功能组件构建，强调可信赖AI的七个属性：有效性、安全性、安全保障性、可问责性、可解释性、隐私性和公平性[20]。NIST AI RMF还特别关注生成式AI的12种特定风险，如幻觉、数据隐私泄露和系统性偏见[20]。
- **ISO/IEC 42001**：于2023年发布，是一个正式的国际标准，用于建立和管理人工智能管理系统（Artificial Intelligence Management System, AIMS）[20]。它采用ISO管理标准中常见的“计划-执行-检查-行动”（Plan-Do-Check-Act, PDCA）模型，侧重于定义AI使用情境、领导层参与、基于风险的规划、运营、绩效评估和持续改进，支持正式审计和国际认证[20]。

这两个框架互为补充：NIST提供适应性强、基于原则的指导，侧重于具体的信任属性；而ISO提供结构化、可认证的管理框架，确保系统性的监督和问责[20]。组织可以结合使用这两个框架，以实现全面且负责任的AI治理。

AI审计实践还包括：

- 数据治理和质量：确保训练数据高质量、无偏见，并实施健全的数据治理框架[33]。
- 算法公平性和偏差检测：使用公平性检测工具（如IBM的AI Fairness 360）评估模型，以避免歧视性结果[33]。
- 性能和准确性监控：持续评估AI系统性能，以识别模型退化并确保其有效性[33]。
- 合规性：确保AI系统符合GDPR、CCPA和即将出台的欧盟AI法案等相关法律框架[33]。

形式化验证提供了数学上的严格保证，而AI审计框架则从治理、流程和合规性角度确保AI系统的整体可信赖性。两者结合，构成了AI验证的强大支柱，对于在不断发展的AI生态系统中建立和维护信任至关重要。

LLM可观察性工具：实时监控与调试

随着大型语言模型（LLM）驱动的应用变得越来越复杂，传统的监控和调试方法已无法满足需求。LLM可观察性工具应运而生，它们提供了对LLM应用性能、行为和内部运作的深度实时可见性，从而帮助开发者识别问题、检测偏差并确保系统安全[28]。

LLM可观察性超越了简单的监控，它旨在理解系统“为什么”以某种方式运行，从而实现根

本原因分析[38]。其关键功能和特性包括：

- 追踪与跨度（**Traces and Spans**）：
  - 追踪：捕获请求在LLM应用中从开始到结束的完整旅程，对于理解复杂的多步骤推理链至关重要[28]。
  - 跨度：表示单次操作（如API调用、工具使用、模型推理）的时间切片，包含详细的日志和上下文信息，有助于分析延迟和成本[28]。
- 会话与会话评估（**Sessions and Session Evals**）：
  - 会话：将多次交互或追踪关联到单个用户会话或对话，从而能够识别对话何时偏离轨道或“中断”[28]。
  - 会话评估：衡量整个交互的质量，例如代理是否保持逻辑流畅，是否记住了早期信息，以及是否帮助用户达到目标[28]。
- 数据收集与标准化：
  - **OpenInference**：一个开源框架，用于从AI代理和LLM驱动的工作流中捕获详细的遥测数据（如追踪和跨度），并将其映射到标准化属性[28]。
  - **OpenTelemetry (OTel)**：行业标准的应用程序遥测收集系统。通过将OpenInference与OTel集成，可以轻松收集和导出追踪数据，然后由可观察性平台进行可视化[28]。
- 性能与成本管理：LLM可观察性工具能够实时监控关键性能指标，如延迟、吞吐量和令牌消耗，从而优化资源分配和成本[29]。
- 安全与偏差检测：
  - 异常检测：通过监控模型行为、输入数据和模型输出，LLM可观察性工具可以检测异常情况，这些异常可能表明数据泄露、对抗性攻击或模型偏差[29]。
  - 幻觉检测：一些工具内置了检测模型“幻觉”（即生成误导性或不准确信息）的功能[30]。
- 调试与故障排除：对于由多个LLM代理逻辑链接在一起的复杂应用，可观察性工具提供了对整个LLM链操作的可见性，有助于快速排查循环或意外延迟等问题[29]。

市场上有多种LLM可观察性工具，包括：

- **Arize AX / Arize-Phoenix**：提供详细的追踪摄取、仪表板、提示工程工具和代理可视化，支持在线和离线评估[28]。
- **Datadog LLM Observability**：提供LLM链的端到端追踪、实时性能和成本监控、质量和安全评估，并集成其APM（应用性能管理）功能[29]。
- **Langsmith**：Langchain的商业产品，内置追踪功能，可上传LLM调用的追踪数据到云端[30]。
- **Helicone**：开源LLM可观察性工具，支持OpenAI、Anthropic等模型[30]。
- **TruLens**：专注于LLM响应的定性分析，提供反馈功能以评估每次LLM调用后的响应[30]。

这些工具通过提供对LLM内部运作的深度洞察，使开发者能够更有效地管理和理解LLM应

用的性能，检测漂移或偏差，并在问题对业务或最终用户体验产生重大影响之前解决它们。它们是构建和维护可靠、安全LLM系统的关键组成部分。

从生成式AI到代理式AI的演进与验证挑战

人工智能领域正在经历从生成式AI（Generative AI, GenAI）向代理式AI（Agentic AI）的重大演进，这带来了新的能力和随之而来的验证挑战 [39]。

生成式AI的特点与局限

生成式AI模型，如ChatGPT，主要设计用于根据输入提示生成原创内容，包括文本、图像、音频和代码 [39]。它们擅长内容创作、创意辅助和想法生成工作流。然而，GenAI通常在没有与环境或外部工具交互的情况下，直接从可用信息生成输出，其推理能力有限，任务通常结构化为直接解决 [41]。这意味着GenAI在处理需要多步骤规划、反思和与外部世界持续交互的复杂任务时存在局限性。

代理式AI的兴起与能力

代理式AI建立在生成式AI的基础上，代表了AI演进的下一个重大步骤，具有更强的推理和交互能力，从而能够更自主地处理复杂任务 [39]。代理式AI系统旨在自主规划、执行和适应多步骤任务，通常无需人类干预，利用推理、记忆和工具编排能力。它们像数字代理一样，能够做出决策并采取行动以实现既定目标 [39]。

代理式AI的关键特征包括：

- 与环境和工具的交互：代理能够通过一系列动作与环境交互，并利用各种工具（如网络搜索、文件搜索、计算机使用）来完成任务 [14]。
- 深度推理：代理执行多步骤、依赖于问题的计算，包括规划和反思 [40]。例如，OpenAI的o1系列推理模型可以花费数分钟处理自我生成的提示，作为搜索和规划过程的一部分 [8]。
- 自主性与适应性：代理能够根据反馈进行即时学习，并调整其未来行动 [40]。

验证挑战与未来方向

从生成式AI向代理式AI的转变，虽然带来了巨大的潜力，但也引入了新的、更复杂的验证和验证（V&V）挑战 [39]。

- 不可预测性和难以控制：由于代理与环境交互并以较少详细指令解决任务，其解决方案可能更加多样、不可预测且难以控制 [40]。
- 放大风险：自主记忆使用、灵活工具选择和开放式探索等能力，在放大潜在益处的同时，也放大了与GenAI相比的风险 [40]。
- 规范问题复杂化：正确指定代理以应对各种突发情况，比为狭窄任务制定提示要求更高 [40]。
- 问责与责任：当AI代理做出错误决策时，谁应该承担责任是一个复杂的治理和伦理问题 [39]。
- 可解释性与透明度：代理系统通常通过逻辑、反馈循环和生成模块的组合来做出决策，这使得其决策过程难以解释和透明化 [39]。
- 安全与鲁棒性：确保代理系统在面对对抗性攻击和意外场景时的安全性和鲁棒性，是持续的挑战 [31]。

为了应对这些挑战，AI验证和验证需要新的方法：

- 形式化方法：需要更强大的形式化方法来证明AI组件与规范的一致性，尤其是在高维输入空间、参数空间和在线适应性方面 [42]。
- 软件测试：需要开发更有效的测试覆盖指标和技术，如变异测试（metamorphic testing），以解决“预言机问题”（oracle problem），即难以明确定义系统输出的正确性标准 [42]。
- 模拟测试：利用模拟环境进行测试，以应对真实世界中未知变量和人类行为建模的复杂性 [42]。
- 持续监控：在线学习系统需要持续监控，以确保其探索不会导致不安全状态 [42]。

OpenAI的“通用验证器”概念，正是为了应对这些演进中的挑战而提出的多维度策略。通过结合Prover-Verifier Games提升内部推理的可读性、Responses API中的护栏和可观察性确保代理行为的安全、以及内容凭证保障生成内容的真实性，OpenAI正试图构建一个能够适应从生成式AI到代理式AI转变的全面验证生态系统。未来的AI系统将不仅需要强大的能力，更需要可验证的信任基础，以确保其在日益复杂和自主的应用中能够安全、负责任地运行。

结论

OpenAI在构建“通用验证器”方面所采取的策略，清晰地表明了其对AI可信赖性问题的深刻理解和系统性应对。这一概念并非指单一的验证产品，而是一系列相互关联、协同作用的研究与工程举措，旨在从多个维度提升AI系统的安全性、可解释性和对齐性。

首先，**“证明者-验证者博弈”（PVG）**及其在大型语言模型（LLM）可读性方面的应用，是OpenAI解决AI“黑箱”问题，并确保其内部推理过程可被人类理解的关键一步。通过引入“狡猾证明器”进行对抗性训练，OpenAI不仅提升了验证器的鲁棒性，还通过“负对齐税”的发现，揭示了过程监督在提升AI性能的同时，也能促进其与人类价值观的对齐，这改变了传统上安全与性能之间存在权衡的认知。这种方法论上的创新，为AI系统的内部机制提供了前所未有的透明度，对于在复杂决策场景中建立信任至关重要。

其次，**Responses API**中的**“护栏”**和**“追踪与可观察性”**工具，则将验证的焦点扩展到AI代理在现实世界中的行为。这些机制为代理的自主行动提供了必要的安全保障和诊断能力，确保其行为可控、可预测且符合预期。内容凭证的引入，进一步将可验证性延伸至AI生成内容的真实性层面，有效应对了虚假信息带来的挑战，为数字内容的信任链提供了基础。

最后，**AI**对齐研究、可解释**AI**（**XAI**）以及形式化验证和**AI**审计框架，共同构成了OpenAI构建可信赖AI系统的宏大愿景。从将复杂人类价值观编码到AI系统中的“规范问题”，到解决“内部对齐”和“外部对齐”的挑战，OpenAI正在通过可扩展监督等创新方法，努力确保未来超人类AI系统能够与人类意图保持一致。XAI技术则通过提供决策过程的透明度，增强了人类对AI的理解和信任。而形式化验证和AI审计框架则从数学严谨性和治理合规性层面，为AI系统的整体可靠性提供了系统性保障。

总而言之，OpenAI的“通用验证器”策略反映出一种全面而前瞻性的AI安全范式。它认识到AI可信赖性是一个多层次、动态演进的问题，需要持续的研发投入和多学科的协同。这种多管齐下的方法，预示着AI的未来将从不透明的“黑箱”系统，逐步演变为内部状态和外部行为均可被验证和理解的“透明箱”系统。这种转变不仅是技术上的进步，更是AI走向负责任、有益于人类社会的关键一步。通过这些持续的努力，OpenAI正致力于为通用人工智能的到来，奠定一个坚实、可信赖的基础。

Works cited

1. arxiv.org, accessed on August 4, 2025, https://arxiv.org/html/2407.13692v1
2. Formal Verification of Neural Networks for Safety-Critical Tasks in Deep Reinforcement Learning, accessed on August 4, 2025, https://proceedings.mlr.press/v161/corsi21a/corsi21a.pdf
3. Revolutionising Healthcare with Top 12 Applications of AI - Omniva Telehealth, accessed on August 4, 2025,

https://omnivatelehealth.com/blog/ai-healthcare-applications/
4. AI in Finance: Applications, Examples & Benefits | Google Cloud, accessed on August 4, 2025, https://cloud.google.com/discover/finance-ai
5. Verification and Synthesis of Autonomous Systems | Coursera, accessed on August 4, 2025, https://www.coursera.org/learn/verification-and-synthesis-of-autonomous-systems
6. Explainable AI – how humans can trust AI - Ericsson, accessed on August 4, 2025, https://www.ericsson.com/en/reports-and-papers/white-papers/explainable-ai--how-humans-can-trust-ai
7. How does Explainable AI improve the trustworthiness of AI systems?, accessed on August 4, 2025, https://milvus.io/ai-quick-reference/how-does-explainable-ai-improve-the-trustworthiness-of-ai-systems
8. Research | OpenAI, accessed on August 4, 2025, https://openai.com/research/
9. Safety & responsibility | OpenAI, accessed on August 4, 2025, https://openai.com/safety/
10. AI Framework Tracker - Fairly AI, accessed on August 4, 2025, https://www.fairly.ai/blog/policies-platform-and-choosing-a-framework
11. accessed on January 1, 1970, https://jamesband.asia/288.%E8%AF%81%E6%98%8E%E5%99%A8%E5%92%8C%E9%AA%8C%E8%AF%81%E5%99%A8.html
12. Let's Verify Step by Step 1 Introduction - OpenAI, accessed on August 4, 2025, https://cdn.openai.com/improving-mathematical-reasoning-with-process-supervision/Lets_Verify_Step_by_Step.pdf
13. Toward understanding and preventing misalignment generalization ..., accessed on August 4, 2025, https://openai.com/index/emergent-misalignment/
14. New tools for building agents | OpenAI, accessed on August 4, 2025, https://openai.com/index/new-tools-for-building-agents/
15. Content Credentials in Azure OpenAI - Azure OpenAI | Microsoft Learn, accessed on August 4, 2025, https://learn.microsoft.com/en-us/azure/ai-foundry/openai/concepts/content-credentials
16. Research Engineer / Research Scientist, Alignment | OpenAI, accessed on August 4, 2025, https://openai.com/careers/research-engineer-research-scientist-alignment/
17. Scalable oversight | European Data Protection Supervisor, accessed on August 4, 2025, https://www.edps.europa.eu/data-protection/technology-monitoring/techsonar/scalable-oversight_en
18. Scalable Oversight in AI: Beyond Human Supervision | by Deepak Babu P R | Medium, accessed on August 4, 2025, https://medium.com/@prdeepak.babu/scalable-oversight-in-ai-beyond-human-supervision-d258b50dbf62
19. Formal Verification of Deep Neural Networks: Theory and Practice ..., accessed on

August 4, 2025, https://neural-network-verification.com/
20. NIST vs ISO - Compare AI Frameworks - ModelOp, accessed on August 4, 2025, https://www.modelop.com/ai-governance/ai-regulations-standards/nist-vs-iso
21. Prover-Verifier Games improve legibility of LLM outputs | Request PDF - ResearchGate, accessed on August 4, 2025, https://www.researchgate.net/publication/382363274_Prover-Verifier_Games_improve_legibility_of_LLM_outputs
22. Prover-verifier Games Improve Legibility of LLM Outputs | Weaviate, accessed on August 4, 2025, https://weaviate.io/papers/prover
23. Prover-Verifier Games improve legibility of LLM outputs - OpenReview, accessed on August 4, 2025, https://openreview.net/forum?id=j4s6V1dl8m
24. Prover-Verifier Games improve legibility of language model outputs - OpenAI, accessed on August 4, 2025, https://openai.com/index/prover-verifier-games-improve-legibility/
25. AI Risk Management Framework - Palo Alto Networks, accessed on August 4, 2025, https://www.paloaltonetworks.com/cyberpedia/ai-risk-management-framework
26. Usage policies | OpenAI, accessed on August 4, 2025, https://openai.com/policies/usage-policies/
27. OpenAI: The Future and Ethics of Artificial Intelligence - The Bull & Bear | McGill's student-run news magazine, accessed on August 4, 2025, https://bullandbearmcgill.com/openai-the-future-and-ethics-of-artificial-intelligence/
28. LLM Observability for AI Agents and Applications - Arize AI, accessed on August 4, 2025, https://arize.com/blog/llm-observability-for-ai-agents-and-applications/
29. What Is LLM Observability & Monitoring? - Datadog, accessed on August 4, 2025, https://www.datadoghq.com/knowledge-center/llm-observability/
30. LLM Observability Tools: 2025 Comparison - lakeFS, accessed on August 4, 2025, https://lakefs.io/blog/llm-observability-tools/
31. AI Alignment: The Hidden Challenge That Could Make or Break ..., accessed on August 4, 2025, https://medium.com/@MakeComputerScienceGreatAgain/ai-alignment-the-hidden-challenge-that-could-make-or-break-humanitys-future-9b3fd70941ca
32. www.ironhack.com, accessed on August 4, 2025, https://www.ironhack.com/us/blog/exploring-the-challenges-of-ensuring-ai-alignment
33. Auditing AI Systems: Best Practices for Compliance - Dotnitron, accessed on August 4, 2025, https://www.dotnitron.com/insights/auditing-ai-systems-best-practices
34. 5 AI Auditing Frameworks to Encourage Accountability - AuditBoard, accessed on August 4, 2025, https://auditboard.com/blog/ai-auditing-frameworks
35. Interpretable machine learning through teaching | OpenAI, accessed on August 4, 2025, https://openai.com/index/interpretable-machine-learning-through-teaching/
36. Introduction to Vertex AI Model Monitoring | Google Cloud, accessed on August

4, 2025, https://cloud.google.com/vertex-ai/docs/model-monitoring/overview
37. AI in Compliance: Benefits, Risks & Regulatory Challenges | Seattle U School of Law, accessed on August 4, 2025, https://onlinelaw.seattleu.edu/blog/ai-in-compliance-exploring-the-benefits-risks-and-regulatory-challenges/
38. Top 9 Observability Platforms for LLMs: Unlocking Advanced Monitoring for AI Systems, accessed on August 4, 2025, https://www.edenai.co/post/top-5-paid-observability-platforms-for-llms-unlocking-advanced-monitoring-for-ai-systems
39. Agentic AI vs Generative AI in 2025: Definitions, Use Cases and Key Differences - Tatvic, accessed on August 4, 2025, https://www.tatvic.com/blog/agentic-ai-vs-generative-ai-in-2025-definitions-use-cases-and-key-differences/
40. Generative to Agentic AI: Survey, Conceptualization, and Challenges - ResearchGate, accessed on August 4, 2025, https://www.researchgate.net/publication/391247465_Generative_to_Agentic_AI_Survey_Conceptualization_and_Challenges
41. Generative to Agentic AI: Survey, Conceptualization, and Challenges - arXiv, accessed on August 4, 2025, https://arxiv.org/html/2504.18875v1
42. Verification and Validation of Systems in Which AI is a Key Element - SEBoK, accessed on August 4, 2025, https://sebokwiki.org/wiki/Verification_and_Validation_of_Systems_in_Which_AI_is_a_Key_Element