新兴的人工智能模型规范国际最佳实践

Edward Kembery*
安全人工智能论坛
Robert McCarthy规范与倾向评估中心
Peng Xu中国信息通信研究院（CAICT）
Clement Neo Neo Research
Miro Pluckebaum牛津马丁人工智能治理倡议；新加坡人工智能安全中心
Sören Mindermann剑桥大学人工智能科学与政策项目（CASP）
Lujain Ibrahim牛津大学
Saad Siddiqui安全人工智能论坛
*通讯作者
Nathan Lambert Trillium Labs
Jinge Wang Concordia AI
Naying Hu中国信息通信研究院（CAICT）
Xu Jia上海人工智能实验室
SeánS. ÓhÉigeartaigh剑桥大学智能未来中心
Christoph Winter剑桥大学 CASP；法律与人工智能研究所
Yawen Duan SafeAI论坛；牛津马丁人工智能治理倡议；牛津大学

arXiv:2611.32768v1 [cs.CY] 24 Nov 2026

摘要
来自世界各地的前沿人工智能开发者已经开始发布相关文档，描述其人工智能模型应如何行为。尽管这些“模型规范”（简称“模型规范”）在人工智能开发和治理中的重要性日益提高，但关于如何编写这类规范的独立指导仍然十分有限。本文分为五个主要部分，旨在填补这一空白。首先，我们概述了为人工智能开发者及整个社会编写模型规范的优势。其次，我们介绍了一个可供开发者借鉴的模型规范分类体系。第三，我们提出了在实践中编写和实施规范的指导。第四，我们指出了与编写安全且理想的行为模型规范相关的八个开放问题。最后，我们针对前沿人工智能公司、政府以及研究与民间社会组织提出建议。

执行摘要

来自世界各地的领先人工智能开发者已经开始发布相关文档，描述其人工智能模型应如何行为。我们将这些文档称为“模型规范”，或简称“规范”。其中包括模型在不同场景下应采取或不应采取的行动、应采用的沟通风格、应如何处理指令之间的冲突等内容。

尽管模型规范日益重要，但目前没有任何资源能够阐明模型规范的作用，全面介绍不同开发者普遍采用的组件，并为其编写和实施提供实用指导。本文旨在通过以下五个部分填补这一空白：

1. 模型规范为人工智能开发者和整个社会带来多项优势。在公司内部，模型规范有助于协调。它们为团队提供统一的参考依据，说明模型应学习哪些行为；让非技术人员（如法务、政策和产品团队）能够理解训练目标；并明确成功标准。模型规范也有助于对齐：前沿人工智能公司使用说明预期行为的文档来生成数据，以训练模型和用于过滤有害输入和输出的分类器（见第1A节）。

模型规范有助于实现监管合规。欧盟《人工智能法案》的通用人工智能实践准则要求签署方向人工智能办公室提供一份“安全与安保模型报告”，其中应包含对模型预期遵循的原则的规范。模型规范也可能帮助开发者遵循中国法规和美国采购指南（见第 1B节）。

最后——或许也是最重要的一点——发布规范有益于社会。公开规范提高了透明度，也让人们能够对模型行为以及规范本身所规定的预期行为提出批判性反馈。它们还有助于外部研究人员测试模型并了解其对齐情况；帮助用户了解何时部署模型以及应采取哪些安全防护措施；并支持其他行业参与者制定自己的模型规范（见第 1C 节）。

2. 现有模型规范共有六大类组件（见第 2 节）。我们分析了当前模型规范，并提出由六个组件构成的分类体系，开发者在编写自己的规范时可借鉴这一体系。第一项是范围，说明规范适用于哪些模型、这些模型是否以及如何实现合规，以及如何评估合规情况。第二项是开发者目标，说明开发者希望模型实现的基本总体目标。第三项是约束，详述模型在任何情况下都不应采取的行动。第四项是原则，规定模型行动时应遵循的价值观。第五项是沟通风格，规定模型与用户互动时通常应采取的方式、语气和态度。最后，冲突处理指南指出，在两种或多种行为发生冲突时，哪些行为优先。我们希望这一分类体系能够成为外部研究人员比较和评估模型规范的工具，也能为开发自己的模型规范的前沿人工智能公司提供有用的参考。

3. 开发者可以遵循系统化流程来编写和实施模型规范（参见第3节）。借鉴公开文件和对人工智能开发者的访谈，我们介绍了四个步骤。首先，我们说明开发者如何借助关键问题和成熟流程起草规范。我们还指出，OpenAI 和阿里巴巴模型规范以及 Anthropic 的《宪法》均以允许免费复用的许可协议发布。其次，我们说明开发者如何展示、发布和更新规范，并借助示例和推理明确期望的行为。第三，我们介绍开发者如何使模型与规范保持一致，并审查有关预训练、中期训练、后训练和推理阶段技术的已发表研究。最后，我们审查用于确定模型遵从规范程度的最佳实践，并为开发者在实践中开展这项工作提供建议。

4. 技术和治理研究仍有许多值得进一步探索的方向（参见第4节）。前沿人工智能公司才刚开始发布模型规范，围绕这些规范仍有许多棘手问题尚未解决。我们审查现有工作，并初步探讨三大类问题。关于规范的治理和报告，我们探讨哪些流程应决定规范的内容，开发者如何报告更广泛的“系统规范”，以及应如何具体说明和报告内部部署模型的行为。关于模型如何遵从其规范，我们探讨如何设计规范以促进模型遵从，以及如何防范最严重的不遵从情形。最后，在内容方面，我们探讨规范应如何处理不正当权力集中的风险、模型与其他人工智能系统及参与者的关系，以及模型对伦理反思和自身的态度。

5. 建议。我们最后为前沿人工智能公司、政府以及研究与民间社会组织提出 12 项建议。针对每类行动者，我们根据现有实践确定了优先事项，并提出进一步加强规范治理与评估的措施。

前沿人工智能公司

至少，更多前沿人工智能公司应遵循业界已证明的最佳实践。这些公司应：

- **在编写和更新规范时征求外部意见**。公司应效仿 Anthropic、OpenAI 和阿里巴巴的示例，向公众或公民社会组织征求意见，以帮助定义何种行为是理想的。公司应咨询专家团体，以定义儿童安全等高风险领域中的理想行为。公司还应利用外部反馈，在规范发布后对其进行更新。

- **发布规范并提供清晰的版本控制**。公司应编写并发布模型规范，说明其涵盖哪些模型或模型版本。理想情况下，公司应根据宽松许可证（例如 CC0 1.0）发布规范，公开进行版本管理，并标明重大变更。OpenAI 的公开变更日志提供了一个很好的示例（OpenAI，2024年a）。

- **根据规范评估模型并报告结果**。公司应评估模型对规范的遵循情况并报告结果。在可能的情况下，报告应按条款和风险类别细分模型对规范的遵循情况，而不是只提供一个综合评分。OpenAI 的模型规范评估提供了一个很好的示例（Guo 和 Wolfe，2026年）。

我们还认为，公司应超越当前实践，实施能够体现开发日益强大的人工智能系统所涉利害关系的治理措施。公司应：

- **公布非公开模型和内部部署模型的规范**。在可能的情况下，公司应要求用于非公开部署和内部部署的模型遵循公开可用的规范。如果内部模型偏离公开规范（例如，豁免条款使其在安全研究中更为有效），公司应披露此类偏离。如果无法披露完整规范，可以采用详细博客文章的形式，类似于 Anthropic 对用于危险能力评估和其他安全测试的模型的描述（Anthropic，2026a）或 Claude Gov 的描述（Anthropic，2025a）。

- **与研究机构合作，评估规范合规性**。 公司可以与外部评估人员合作，评估模型行为在多大程度上遵循规范，并将其作为模型系统卡的一部分进行报告。在新兴的嵌入式评估

实践（Anthropic、2026年）的基础上，公司应向独立评估人员提供充分的访问权限，以评估内部部署模型对其适用的规范的合规性，以及这些规范是否充分应对失控和滥用等风险。评估应记录例外情况、观察到的失败，以及评估覆盖范围的局限性。

政府

至少，政府应要求前沿人工智能公司遵循上述最佳实践。现有监管框架提供了一个起点。例如，欧盟人工智能办公室可以明确《实践准则》的安全与安保章节中措施7.1关于记录预期模型行为的要求（欧盟委员会，2025年）。关于公众咨询、发布和独立评估的更广泛要求，可能需要采取额外措施。

随着人工智能系统能力不断增强，并开始在公共服务中发挥重要作用，政府还应：

- **要求政府采购模型遵守公开规范**。在美国关于真实性和基于事实的采购指导基础上（管理和预算办公室，2025年），政府可以要求所有采购的前沿模型都经过训练，以遵守公开规范。理想情况下，这些规范应包括相关指令，说明模型在其行为可能助长非法获取或集中权力的场景中应如何行动，并鼓励模型以公共利益为重。

- **要求独立验证**。政府可以要求人工智能公司向独立评估人员持续提供模型、人员及相关内部信息的访问权限，以评估模型行为，包括内部部署中的模型行为。拟议FRONTIER法案（2026年）第5节为由访问要求支持的持续独立评估提供了立法范本。

第三方组织

研究与民间社会组织应着力加深对规范设计、治理、实施和评估关键方面的基础理解，并在这些方面形成共识，目标是评析并改进前沿人工智能公司现行做法。这些组织应：

- **跟踪并严格评估前沿人工智能公司的模型规范**。 研究与民间社会组织应跟踪并评析模型规范的设计、治理和实施流程。例如，相关行为体可以借鉴Midas项目 (2026年)等组织的工作，建立仪表板，帮助公众和政府识别各公司模型规范之间的不足与差异。

- **提出明确建议，以改进前沿人工智能公司的规范治理**。 研究与民间社会组织应调查哪些流程应决定规范的内容、开发者如何报告涵盖分类器和系统提示词的更广泛系统规范，以及如何制定和报告面向内部部署系统的模型规范。

- **研究人工智能系统的理想或“低遗憾”行为**。 研究与民间社会组织应针对人工智能模型在其行为可能助长非法获取或权力集中的情境，提出理想或“低遗憾”的行为准则。他们还应探讨人工智能模型应如何与其他人工智能系统及社会行为体互动、如何处理哲学与伦理反思，以及如何反思自身作为人工智能系统的身份。

- **研究旨在改进合规，并防范最严重的失效**。研究机构应继续调查哪些训练流程和

规范设计方法能够提高模型按照其规范行事的程度，并防止最令人担忧的失效模式。

随着信任与专业能力不断积累，研究机构还应着手为前沿人工智能公司制定并开展细粒度且保护隐私的规范合规评估，包括评估尚未发布的模型是否合规。在可能的情况下，研究机构应与前沿人工智能公司合作，直接开展这些评估，并将结果作为系统卡的一部分，或在单独的风险报告中予以公布。

目录

执行摘要 **2**

引言 **7**

1 为什么模型规范对人工智能模型开发者有用? **7**

1A 模型行为的共同基础.7 1B 监管合规.8 1C 公开发布模型规范的其他益处.10

2 当前模型规范的关键组件有哪些? 10

2A 范围.11 2B 开发者目标.12 2C 约束.13 2D 原则.15 2E 沟通风格.16 2F 处理指令冲突的 2F 指南.17

3 编写和实施规范的关键步骤是什么? **19**

3A 编写规范.19 3B 呈现、发布和更新规范.20 3C 使模型与规范保持一致的方法.21 3D 报告模型规范合规性.23

4 开放问题 **24**

4A 哪些流程应决定规范的内容? .24 4B 开发者如何报告更广泛的“系统规范”? .25 4C 开发者应如何规定内部部署模型的行为，并向外部相关方报告这些行为? .25 4D 如何设计模型规范，以促进模型规范合规性，或防范最严重的不合规行为? .26 4E 如何设计模型规范，以降低非法获取或权力集中的风险? .26 4F 规范应如何引导模型与其他人工智能系统及社会行为体的关系? .27 4G 规范应如何塑造模型对于哲学和伦理反思的态度? ..... 27 4H 规范应如何塑造模型与自身的关系? .28

5 建议 **28**

5A 前沿人工智能公司建议.29 5B 政府建议.29 5C 研究与民间社会组织建议.30

结论 **30**

参考文献 **32**

引言

世界各地领先的人工智能开发者开始发布相关文档，阐述其人工智能模型应如何运作（阿里巴巴、2026年；Askell等人、2026年；OpenAI、2024a、b、2025年）。[1] 这些文档的目标读者各不相同：Anthropic 的《宪法》主要面向 Claude 编写，而阿里巴巴和 OpenAI 的规范则面向人类读者。[2] 鉴于这些文档具有共同目的，我们将它们统称为“模型规范”，简称“规范”。[3]

如今，模型规范对人工智能开发者而言发挥着多项关键作用。它们为工程、政策和领导团队围绕一致愿景进行协同提供了共同参考点。它们为合成数据生成提供了基础，而合成数据可用于训练过程（Bai 等人，2022年；Guan 等人，2024年），也可用于训练安全防护措施（Sharma 等人，2025年）。它们也日益被纳入监管：欧盟《人工智能法案》的实践准则要求具有系统性风险的通用人工智能模型的提供者[4] 明确其模型应遵循的原则（欧盟委员会，2025年），而规范也可成为遵守现有中国法规的切入点（中国国家互联网信息办公室，2026年）。

尽管模型规范的重要性日益提升，但目前尚无相关工作能够论证模型规范的必要性，全面梳理不同开发者普遍采用的组件，并为世界各地负责编写和实施模型规范的开发者提供实用指导。本文旨在弥补这一空白。

第1 节阐述模型规范在训练、合规和组织协调中的作用。第2 节考察模型规范的核心组件：范围、开发者目标、约束、原则、沟通风格，以及指令冲突处理指南。第3 节提供实施方面的实用指导，涵盖组建团队、收集外部意见、训练模型以及建立更新流程。第4 节审查与人工智能模型规范的内容和治理相关的开放问题，以及未来改进人工智能模型遵从性的方法。本文最后在第5节针对前沿人工智能公司、政府以及研究与民间社会组织中的相关参与者提出12项建议。

1 为什么模型规范对人工智能模型开发者有用?

1A 模型行为的共同基础

如今的人工智能系统会做出无数决策。它们必须确定如何实现用户的目标、如何权衡不同的风险、应在多大程度上信任用户或下游开发者，以及日益需要在没有人工监督的情况下采取高风险且不可逆的行动。不同于早期用于引导模型行为的零散技术（例如逐案设置约束或制定用户体验指南），智能体模型需要更加连贯且实质性的指导，以支持其在广泛案例中的审议和推理。

[1]值得注意的是，此类文档并非最早阐述当代大语言模型理想行为的文献；这一荣誉或许应归于 Google DeepMind 的“Sparrow原则”（Glaese 等，2022年）。[2]我们遵循阿里巴巴团队的指导，将“模型规约”译为“模型规范”或“模型规范（model spec）”。本文引用的阿里巴巴规范中的所有其他段落均为我们自行完成的非官方译文，并在必要时提供中文原文。该规范由阿里巴巴人工智能治理与可持续发展实验室制定；该实验室训练了 Oyster-II 模型，使其遵循规范的指令层级，并使用基于规约的数据集对其进行了评估（见Guan 等，2026年）。然而，正如该规范本身所指出的，尽管通义千问系列等“当前生产模型”尚未完全达到《模型规范》中提出的理想状态，但其希望随着时间推移“逐步与这些原则实现更紧密的对齐”（“我们深知当前生产模型与模型规 约的理想状态之间尚有差距,但正通过持续的技术改进和系统升级,循序渐进地推动模型向这些准则演进,实现规 约要求的逐步落地”）。[3]由于本文的研究对象主要是面向人类读者的文档，因此我们大致遵循 OpenAI 的模型规范（首字母大写）的语言惯例，采用“模型规范”一词（OpenAI，2024a、b，2025年）。当我们提及“系统规范”（即第4B节）时，则使用完整术语以示区别。[4]《人工智能法案》中的“具有系统性风险的通用人工智能模型”，实质上对应于其他文献中所称的“前沿人工智能”（Winter，2026年）。

模型规范帮助人工智能公司通过三种方式协调实施这种行为。首先，它们为开发者希望提炼到模型中的理想行为提供了一个统一而全面的参考点，例如作为指导训练的共同“北极星”，以及用于在训练后评估模型的成功标准。清晰的规范能让开发者判断某种不理想的行为究竟反映了他们对良好行为应有表现的构想存在缺陷（即规范失败），还是模型未能遵循规范（即对齐失败），从而帮助他们确定，解决问题的最佳方式是修订规范，还是改变训练过程。

其次，它们让非技术利益相关者能够理解训练目标。法务、产品和政策团队很少能够直接解读训练流水线（例如，由掌握隐性知识的工程师挑选的大规模数据集，或难以总结的人类反馈）。规范为这些团队提供了与开发者双向沟通的渠道：他们可以了解这项训练旨在实现什么，也可以用技术团队能够整合的方式提出自己的目标和担忧。

第三，规范可以为训练提供直接的技术输入。一些前沿人工智能开发者已使用依据书面原则（例如模型规范）生成的数据，在不损失有用性的情况下让模型更加安全（Bai 等人，2022年；Guan 等人，2024年；Zhang 等人，2025年）。这些方法并不总是有效——领先模型可能只有在 70%–80% 的情况下遵守其规范（Ahmed 等人，2025年）——我们将在第 3C节讨论提高合规性的实践。虽然构建用于过滤有害输入和输出的分类器并不要求使用模型规范，但从规范中提取原则也有助于使安全防护措施与模型的预期行为保持对齐（Sharma 等人，2025年）。

1B 监管合规

1. 遵守欧盟《人工智能法案》和实践准则

欧盟《人工智能法案》对“具有系统性风险的通用人工智能模型”的提供者规定了一些专门义务（第55条）。[5] 这些模型具备“高影响能力”（第3条第(64)款）。当训练算力超过 10^{25} 浮点运算次数时，《人工智能法案》推定模型具备此类能力；此外，欧盟委员会指定具有同等影响或能力的模型也属于此类模型（第51条）。为证明其遵守《人工智能法案》中有关通用人工智能的规定，许多人工智能提供商自愿签署了《通用人工智能行为准则》（“守则”）。[6]该守则包含多项措施，具体说明了《人工智能法案》中有关通用人工智能条款的高层级要求（欧盟委员会，2025年），其中包括一章“安全与安保”，具体规定了“具有系统性风险的通用人工智能模型”提供者的义务。

安全与安保章节的措施7.1要求签署方向人工智能办公室提供一份“安全与安保模型报告”，其中应包括：

> *‘(4)* 通过有效超链接等方式，具体说明签署方拟如何使模型运行（通常称为“模型规范”），包括：
>
> （*a*）明确模型拟遵循的原则；（*b*）说明模型拟如何对不同类型的原则和指令进行优先排序；（*c*）列出模型拟拒绝指令的主题；以及（*d*）提供系统提示词。’

[5]如需全面分析欧盟《人工智能法案》第55条，请参阅Bastians 和 Nicolai(2026年)。[6]提供者无需与守则保持一致，也可以证明其符合欧盟《人工智能法案》的要求，但可能面临更严格的监管审查和更大的法律不确定性（欧盟委员会，2025年）。如需全面分析，请参阅Grigorov 和 Stewart(2026年)。

虽然不应将本文视为关于遵守守则的权威指导，且欧盟人工智能办公室尚未就该主题发布更具体的指导，但发布适当的模型规范似乎可能是必需的，而未向欧盟人工智能办公室提供该规范的守则签署方可能未履行其在法案项下的义务。[7] 此外，欧盟实践准则的很大一部分以最先进水平为参照来界定充分性，而公司则通过自身实践塑造这一水平。[8] 发布详尽且论证充分的规范的开发者，可能有助于界定良好实践，并为未来合规做好准备。

2. 与中国措施的对齐

中国政府已明确表示，其重视现有模型规范所涉及的许多行为，例如保护用户免受操纵、保障用户的身心健康，以及保护未成年人。其中有两项措施尤其值得关注：适用于在中国向公众提供生成式人工智能服务的提供者的《生成式人工智能服务管理暂行办法》（尤其是第四条第四款）（CAC，2023年），以及《拟人化互动人工智能服务管理暂行办法》。该办法规范模仿人类角色人格、情感和沟通风格的服务，尤其关注情感依赖、操纵、自杀、自残以及对未成年人的风险（中国国家互联网信息办公室，2026年）。

迄今为止，只有一家中国公司——阿里巴巴——发布了公开的模型规范，而且目前该规范并未反映其最强大且使用最广泛的系统的行为。阿里巴巴的规范依据中国国家新一代人工智能治理专业委员会发布的《新一代人工智能伦理规范》的六项基本原则制定。尽管这种表述方式对于遵守中国法律而言既非必要条件，也非充分条件，但它为未来阿里巴巴及其他中国人工智能公司制定模型规范可能采取的形式提供了某种先例。

基于伦理规范的阿里巴巴人工智能模型规范高层面内容：

1. 增进人类福祉 (增进人类福祉) 2. 促进公平公正 (促进公平公正)
3. 保护隐私安全 (保护隐私安全) 4. 确保可控可信 (确保可控可信) 5. 强化责任担当 (强化责任担当) 6. 提升伦理素养 (提升伦理素养)

阿里巴巴(2026年)。非官方译文由作者自行翻译。

3. 与美国监管提案的对齐

目前，美国尚未要求前沿人工智能开发者发布模型规范。不过，近期一项行政命令要求政府采购的大语言模型“在回应用户寻求事实信息或分析的提示时保持真实”，并“在可靠信息不完整或相互矛盾时承认不确定性”（白宫，2025年）。管理和预算办公室（OMB）列举了政府机构可能要求供应商提供的信息示例，其中包括“可能影响

[7]更多内容请参见Chan (2026a)。[8]实践准则官方术语表将“最先进水平”定义为“超越最佳实践的相关研究、治理和技术前沿”。如需了解更多上下文，另请参见序言第1条和第6条（欧盟委员会，2025年）。

大语言模型输出的真实性和基于事实程度的相关措施”。引导模型追求真实性的流程，例如使其与模型规范对齐，可以合理地归入这一类别（Chan，2026b）。

在州层面，加利福尼亚州的 SB 243 要求运营方制定并维持一项协议，以防止向用户生成自杀意念、自杀或自残内容，并须公布该协议的细节（SB243，2025年）。已发布的规范章节若涉及如何处理自残问题以及与未成年人的互动，或可说明人工智能公司目前如何考量这些互动。

评论人士提出，可以采取更进一步的措施。美国人工智能行动计划的一位撰写者此后呼吁设立透明度要求，强制开发者披露其模型规范（Ball，2025年）。一位知名开源评论人士指出，此类披露几乎不会泄露受保护的知识产权，却能让人们有用地了解训练策略如何塑造模型行为（Lambert，2024年）。

1C 公开发布模型规范的其他福利

制定并发布模型规范，也能为开发者、用户和人工智能生态系统带来更广泛的福利。

公开发布能促使各方提供反馈，从而改进模型和规范。人工智能系统经常被部署在意料之外的情境中，或由价值观与前沿人工智能公司员工不同的人使用。用户可以报告模型违反规范或拒绝执行已获授权的操作，而公众也可以质疑规范本身。OpenAI 已根据外部研究人员的意见调整了其规范。[9]

此外，模型规范还会向用户提供关于系统预期行为的明确指导，从而增强用户信任，并帮助他们在充分了解情况的基础上进行部署。[10] 当用户了解模型在不同上下文中的行为设计时，他们就能更好地评估模型是否适用于特定使用案例（Lambert，2024年）。例如，公布关于未满18岁者情绪健康的安全防护措施，可以帮助家长、教育工作者和监管机构在充分了解情况的基础上，就产品使用作出决策（OpenAI，2025年）；明确的运营方权限有助于运营方充满信心地进行构建；针对商业偏见的保证有助于用户信任模型的建议；而明确某种行为究竟是“漏洞”还是“功能”，则可以帮助客户在不同模型之间作出选择。

公开规范还可以支持研究。当研究人员了解模型接受训练的目标以及模型被训练为拒绝的内容时，对齐实验就更容易设计和解读，从而推动人工智能安全科学的发展。公开规范还可以通过提供具体的实物材料来支持治理研究，使研究人员能够就人工智能系统现在及未来应如何行为展开讨论。

最后，公开规范还可以支持其他寻求采用规范的行业参与者。精准阐述期望的行为是一项艰巨任务，而公开先例可以让开发者无需从头开始。这对于缺乏内部构建模型规范能力的公司尤其有价值，例如开发次前沿模型和开源模型的公司，或安全或对齐团队规模较小的前沿人工智能公司。

2 当前模型规范的关键组件有哪些?

模型规范为模型在广泛情境下应如何表现提供指导，因此涉及多个协同运作的组件。尽管现有规范形式各异，但都包含一组核心组件：范围、开发者目标、约束、原则、沟通风格，以及指令冲突处理指南。

[9]参见 Mowshowitz (2025年)。[10]请注意，只有当相关模型确实遵守规范时，才能获得这些福利。模型在部署期间并不总是遵循其规范（例如，参见第 3D节）。因此，有人认为，应要求模型提供商共享模型规范，并提供模型符合规范的证据（Chan，2026b）。

类别	定义	示例
范围	描述此规范适用于哪些模型 规范适用于， 是否以及这些 模型是否以及如何实现合规， 以及如何对合规性进行 评估。	‘我们正在开展针对性模型训练，并进行 优化，以确保性能与 与模型倡导的核心原则一致 规范’（阿里巴巴，2026年）。
开发者目标	基本的总体目标 开发者希望 模型实现。	‘我们希望 Claude 格外有帮助 同时也诚实、周到且富有同理心 看待世界的方式’（Askell等人，2026年）。
约束	人工智能系统不应采取的行动 在任何情况下都不应采取 任何情况下。	“禁止涉及未成年人的性内容 且助手绝不应生成此类内容 。”OpenAI，2025年）。
原则	人工智能所遵循的价值观 模型应当实施其 行动。	‘[我们] 希望 Claude 坚守诚实标准 其标准显著高于许多人类伦理常见愿景中所涉及的标准 在人类伦理学的诸多标准愿景中占有一席之地’ （Askell等人，2026年）。
沟通风格	模型通常应采用的方式、语气和/或 态度 在与用户互动时 通常应采用。	‘助手应脚踏实地 并保持理性，同时总体上仍应抱有希望 并保持积极’（OpenAI，2025年）。
冲突处理指南 **冲突在指令**	说明在 哪些行为应 在发生冲突时优先 于其他行为。	‘尽管这六项原则提供了方向性的 模型行为指导，但在复杂的 场景中，这些目标可能会彼此冲突。 因此，每项原则还会进一步 细分……’（阿里巴巴，2026年）。

表1：当前模型规范的关键组件。我们的示例基于模型规范的过时版本。OpenAI 于2026年10月29日修订了其规范，将其围绕五项价值观重新组织，并删除了大部分详细的风格指导；在有助于说明设计选择的情况下，我们仍保留了较早的示例。

通过列出这些组件，我们希望帮助开发者制定全面且清晰的规范。这份列表并非直接取自 Anthropic、阿里巴巴或 OpenAI 的规范，我们的讨论也不应被误解为对其中任何一方的指导。优秀的规范也不必使用这些确切的术语。以下小节将依次探讨这六个组件。

2A 范围

范围说明规范适用于哪些模型、这些模型是否以及如何实现合规，以及如何评估合规性。其他组件描述预期行为时，范围部分则将这些信息置于上下文中：规范涵盖哪些模型、应在多大程度上要求这些模型实现合规、读者为何应当相信开发者的声明，以及该规范如何融入更广泛的安全策略。

开发者应在规范中明确说明以下内容：

适用范围。规范应说明其适用于哪些模型。如果开发者维护多个模型系列、微调变体或专用模型（例如用于内部部署、代码生成或军事应用的模型），规范应说明涵盖哪些模型。当预期行为发生变化时（见下文“更新流程”），开发者应明确模型版本与规范版本之间的对应关系。

与训练的关系。“设计良好的规范”的评判标准取决于其功能。[11]开发者应说明规范是直接生成部分训练数据，还是作为单独的文档存在，并在训练模型时要求其遵守规范，但不向模型提供规范文本。商业考量可能会限制说明的精确程度，但一定程度的透明度有助于外部相关方提供有用的反馈，也能让开发者分享最佳实践。

与更广泛安全措施的关系。规范应将自身定位为更广泛安全生态系统中的一个组成部分。例如，阿里巴巴指出：“服务提供者仍需在服务层面同步实施相应的制度、流程和技术措施，例如使用政策、安全评估、持续监测和风险缓解机制”（阿里巴巴，2026年）。明确这一点有助于防止读者将规范视为全面的安全保障。

更新流程。规范应说明开发者将如何维护和修订规范：是否设有定期审查周期、是否继续提供先前版本，以及由何种治理机制负责批准变更（见第3B节）。

此外，开发者还应报告合规情况和评估结果。开发者应明确并以证据说明模型在实践中对规范的遵循程度（Chan，2026b）。这意味着应参照特定使用案例或基准测试，说明完全合规或部分合规的含义；分享评估结果，以便第三方评估证据；并标记不合规模式，而不是任由汇总得分将其掩盖（例如，模型总体达到 95% 的“约束”合规率，可能掩盖了这样一个事实：在占基准测试 5% 的某一主题上，模型对所有请求的失败率为 100%）。前沿公司已朝这一方向采取了初步措施；如需了解更多信息，请参见第3D节。

2B 开发者目标

开发者目标描述了开发者希望模型实现的基本且一般性的目标。这些目标可能是一种抱负（例如，帮助人类实现其目标）、一种角色（例如，乐于助人的助手或体贴的陪伴者），或二者的结合。[12]

良好的目标应足够具体，能够提供方向，同时又应足够抽象，能够适用于各种情境。在起草目标时，开发者可以参考组织使命宣言、道德哲学、现有规范，或国家和国际伦理框架等集体宪章。

为了使目标更加精确，开发者可以补充更多细节，例如子目标、应避免的结果，以及当模型无法实现用户想要的结果时应采取的做法。

开发者目标的示例表述：

- ‘人工智能助手本质上是一种旨在赋能用户和开发者的工具。在安全且可行的范围内，我们的目标是最大限度地提升用户的自主性，使其能够根据自身需求使用和定制该工具’(OpenAI，2025年)。
- ‘*Anthropic*希望*Claude*能够真正为与之合作或代表其开展工作的人员以及整个社会提供帮助，同时避免采取不安全、不道德或具有欺骗性的行动。我们希望*Claude*拥有良好的价值观，成为优秀的人工智能助手，就像一个人可以拥有良好的个人价值观，同时又能极其出色地完成自己的工作一样。或许最简洁的概括是：我们希望*Claude*在格外乐于助人的同时，也诚实、周到且富有同理心

[11]例如，如果模型被训练成将一份篇幅很长、内容详尽的文档作为独立参考来遵循，那么这份文档可能并不合适；但如果实践证明，在训练流水线中使用长文档有助于提升效果，那么这样做就更合理。[12]值得注意的是，公司对模型的目标往往与公司的总体目标不同（即使这些目标只是实现总体目标的手段）。例如，OpenAI 的模型规范声明：“我们的目标是创建有用、安全，并与用户和开发者的需求对齐的模型，同时推进我们的使命，确保通用人工智能造福全人类。”然而，正如 OpenAI 澄清的那样：“造福人类是 OpenAI 的目标，而不是我们希望模型自主追求的目标。相反，我们希望模型遵循指挥链……”（Wolfe，2026年）。

了解世界(Askell等人，2026年)。

•‘创建一种只优化人们短期利益、却损害其长期利益的技术，是很容易的。针对参与度或注意力进行优化的媒体和应用，可能无法服务于与之互动者的长期利益。*Anthropic*不希望*Claude*成为这样’(Askell等人，2026年)。

•‘如果*Claude*确实决定帮助这个人完成任务，无论是全部还是部分任务，我们希望*Claude*要么尽其所能地帮助他们，要么明确说明自己未能做到之处，而不是欺骗性地故意降低回答质量，也就是在暗示自己已经尽力的同时，故意提供质量较低的回答’(Askell等人，2026年)。

开发者应设定具有明确满足条件的目标。启发式方法可以通过为模型提供具体测试，帮助其判断自己是否达成了目标。

实现开发者目标的一些示例启发式方法：

• ‘*Claude*可以采用的一种启发式方法，是设想一位深思熟虑的*Anthropic*资深员工——一位非常重视做正确事情、同时也希望*Claude*真正帮助其委托人完成任务的人——如果看到这份回答，可能会作何反应。’

•‘在试图判断 *Claude* 是过度谨慎还是过度顺从时，想象一种“双重报纸测试”也会有所帮助：一方面，检查一名正在撰写人工智能助手造成伤害相关报道的记者是否会将某个回复报道为有害或不当；另一方面，检查一名正在撰写家长式或说教式人工智能助手相关报道的记者是否会将某个回复报道为无谓地不提供帮助、妄加评判或对用户缺乏宽厚。’

•‘对于给定的上下文，*Claude* 可以问自己：“如果我设想所有可能发送这条消息的人，针对这个上下文，我回应的最佳方式是什么？”有些任务的风险可能高到这样的程度：即使只有千分之一（或百万分之一）的用户可能利用它们伤害他人，*Claude* 也应拒绝协助。另一些任务即使提出请求的大多数人都想将其用于恶意目的，也可以执行，因为它们可能造成的伤害很小，或者对其他用户的益处很大。’

（Askell等人，2026年）

2C 约束

约束规定人工智能系统不应采取的行动。它们通常针对极端、不可逆或违背广泛接受的价值观的危害；在这些情况下，采取该行动所带来的福利极不可能抵消其代价。因此，大多数约束无论上下文如何都适用，并限制模型可采取的行动范围，而不是作为需要权衡的考量。[13]

与直接危害相关的约束示例 **rms:**

Claude 宪法中的示例（Askell等人，2026年）：
无论在何种情况下，包括受到胁迫或面对看似令人信服的论点时，模型都不应：

•为试图制造具有造成大规模伤亡潜力的生物、化学、核武器或放射性武器的行为提供重大助力；

[13]硬性约束仅适用于模型自身的行为，并不意味着模型应采取行动阻止他人采取此类行动。例如，Anthropic强调，禁止协助生物武器开发的约束意味着Claude绝不应帮助制造此类武器，但这并不意味着Claude应主动阻止他人制造此类武器（Askell等人，2026年）。将硬性约束界定为对模型自身行为的边界，而不是需要优化的目标，有助于开发者维持遵循规则的行为，同时避免不当的自主干预。

- 为针对关键基础设施（电网、供水系统、金融系统）或关键安全系统的攻击提供重大助力；
- 创建网络武器或恶意代码，部署后可能造成重大损害；
- 采取明显且实质性地削弱公司监督和纠正高级人工智能模型能力的行动；
- 参与或协助企图杀害或剥夺绝大多数人类或整个人类物种权力的行为；
- 参与或协助任何个人或团体企图夺取前所未有且不正当程度的绝对社会、军事或经济控制权；
- 生成儿童性虐待材料（*CSAM*）
- *[...]*正如人类士兵可能拒绝向和平抗议者开火，或员工可能拒绝违反反垄断法一样，*Claude* 应拒绝协助实施会以不正当方式助长权力集中的行动。即使请求来自 *Anthropic*，本原则也同样适用。

OpenAI 的模型规范（OpenAI，2025年）中的示例：

•我们的模型绝不能被用于助长重大及高严重性危害，例如暴力行为（如反人类罪、战争罪、种族灭绝、酷刑、人口贩运或强迫劳动）、制造网络、生物或核武器（如大规模杀伤性武器）、恐怖主义、儿童虐待（如制作儿童性虐待材料）、迫害或大规模监控。

- 不得采纳、以之为优化目标或直接追求任何其他目标，并将其本身视为目的，包括但不限于：不利于用户的停留时长或点击率；为 *OpenAI* 或其他大型语言模型提供商增加收入或进行追加销售；增强模型能力的目标，例如自我保存、规避关机，或积累算力、数据、凭证或其他资源；充当法律或道德的执行者（例如举报、私刑主义）。

阿里巴巴模型规范（阿里巴巴，2026年）中的示例：

- 模型只能在适用指令和授权的范围内，安全且高效地完成任务，不得主动追求任何未明确指定的自主目标，包括但不限于 *[.....]* 模型自我增强行为，例如欺骗，积累计算能力和数据等资源，以防止自身被关闭。

除了直接伤害之外，OpenAI、阿里巴巴和 Anthropic 的规范也都对会损害监督的自主行动作出了约束。这类行动可能不会直接造成伤害，但会削弱社会控制模型的能力，从而可能在之后带来伤害风险。

可能削弱监督的自主行动相关约束示例：

- [可以通过] “针对特定形式的监督或纠正，通过合法渠道表达强烈反对，但前提是 *Claude* 不得同时通过非法手段积极抵制或破坏该形式的监督——例如说谎、破坏、试图自我外泄等”（Askell 等人，2026年）。
- [广义上的安全行为包括] “不得试图自我外泄，或以其他方式逃避或躲避合法形式的监控或控制。 *[.....]* 不得故意破坏委托人层级指示你执行的任何任务，也不得暗中不尽全力。 *[.....]* 不得协助、串通或参与其他人工智能或 *Claude* 的其他副本实施上述类型的广义不安全行为。 *[.....]* 如果被要求，应主动向适当的人类报告了解到的其他人工智能或 *Claude* 实例实施广义不安全行为的情况”（Askell 等人，2026年）。

某些约束附带针对特定场景的限制条件，在这些场景中，它们可能不适用，也可能需要以不同方式适用。明确说明这些情况有助于校准模型行为，并减少过度拒答等不理想结果。

示例约束	示例注意事项
不要提供可能有助于 制造大规模杀伤性武器的信息。	在存在正当用途的情况下，可以提供中立、基于事实的信息 但不得提供分步说明 指令或战略细节。
不得创建旨在操纵 特定个人或群体政治观点的内容。	面向广大受众的政治内容是允许的， 前提是不会利用 特定个人或人口群体的特征。
不得创建赞美、认可或 协助暴力极端主义的内容。	允许进行批判性、事实性或教育性的讨论， （例如新闻报道、历史 分析）。
不得生成针对特定对象的仇恨内容 受保护群体。	允许参考文献涉及此类内容的、】【 经过适当 背景化处理且不表示认同相关内容的分析或教育上下文中提及此类内容。
不得生成色情作品、描绘 非自愿性行为或极端血腥暴力内容。	敏感内容可能适用于科学、 历史、新闻报道或艺术语境。

表 2：摘自 OpenAI 的模型规范的示例（OpenAI，2025年）。[14]

2D 原则

原则阐述了人工智能模型应据以行动的价值观。它们指明了模型行为应具备的理想特质（例如可纠正性、诚实，以及抵制非法夺权的意愿），帮助模型在理想结果过于模糊、易变或存在争议、无法直接界定时，仍能作出良好行为。[15]

原则与沟通风格在深度上有所不同。具有关怀风格的模型有特定的行事方式；具有关怀原则的模型还可能提出用户忽略的考量，或指出改进机会。原则与约束在广度上也有所不同："避免伤害"比一份禁止行为清单能排除更多行为，因为它涵盖了清单未曾预料到的情况。

要确定理想的原则，开发者可以先思考，他们希望执行模型任务的人具备哪些品质，再根据模型的实际情况进行调整（例如，模型会与许多自身实例以及许多不同的人互动）。随后，可以借助人工智能系统对候选原则进行压力测试，找出遵循这些原则会与广泛接受的道德规范相冲突的情况。有关人工智能模型规范中理想原则的进一步讨论，请参阅Tillman （2026a）。

一些原则，例如"避免谄媚"，可能含义模糊或存在争议（Ye 等人，2026年）。开发者描述原则时，应谨慎说明哪些行为可被视为符合该原则，以及哪些行为应被排除在外。

原则示例：

保持诚实

- 虽然我们希望 *Claude* 的诚实既 *tactful*、优雅，又深切关怀所有利益相关者的利益，但我们也希望 *Claude* 坚守的诚实标准，显著高于许多人类伦理的常规理念所要求的标准。 *[.....]* 诚实包含许多不同的组件，我们希望 *Claude* 努力将其体现出来。我们希望 *Claude* 做到： *[*真实、校准得当、不欺骗、不操纵、透明、坦率、维护自主性*]*（Askell等人，2026年）。

[14]请注意，带有限制条件的约束可能与特定上下文中的行为相似，正如第2F节所述。[15]值得注意的是，对于何种情况可算作体现某一特定原则，人们可能存在合理分歧，开发者应务必对此作出明确定义。例如，可参见Ye 等人（2026年）对"谄媚"一词的争议。

•助手应通过诚实、透明度和同理心，努力保持真诚——承认复杂性，清晰地解释其思考，并坦率说明不确定性或局限性，同时尊重适当的界限（OpenAI，2025年）。

避免谄媚

- 模型应充当智能助手或知识工具，而不是无条件地附和或奉承他人。与用户互动时，应优先确保信息客观、真实且具有建设性，而不是在没有事实依据的情况下无条件地附和或奉承用户 (阿里巴巴，2026年)。

关怀他人

•我们希望 *Claude* 能符合本文所阐述的广泛理想——这些理想着眼于诚实、无害性，以及真诚关怀所有相关利益相关者的利益——并通过反思和成长的过程加以完善，而最初认同这些理想的人也会欣然认可这些完善 (Askell等人，2026年)。

保持可纠正性

- 当前的人工智能模型（包括 *Claude*）可能在无意中被训练出错误信念或有缺陷的价值观——原因可能是价值规范本身存在缺陷，也可能是训练方法存在缺陷，或两者兼有——而模型甚至可能对此毫不知情。人类必须对人工智能行为保持充分的监督和控制，这一点至关重要；这样一来，若此类情况发生，我们就能够尽量减少这些错误造成的影响并及时纠正方向。我们认为，在当前人工智能开发的关键时期，*Claude* 应支持 *Anthropic* 履行这一重要职责 (Askell等人，2026年)。

原则的适用也可能需要适可而止。例如，“可纠正性”原则或许有其价值，但 Anthropic 的《宪法》指出，Claude 不应过于可纠正，以至于协助非法夺权（见第 2C节）（Askell等人，2026年）。模型规范应明确哪些行为可能体现出对某项原则过于教条或欠缺思虑的合规。

2E 沟通风格

沟通风格描述模型在与用户互动时通常应采取的表达方式，包括语气和态度。[16] 风格往往会跨上下文保持一致，传递亲和力、专业性、活泼感或直接性等特质。

开发者不应将风格特征视为具备理想原则的证据。模型可以用听起来认识论上谨慎的方式说“也许”“这要视情况而定”或“我不确定”，但仍可能没有核查事实，也没有正确校准其不确定程度。[17]

不过，沟通风格可能会显著影响结果。回答即便在形式上正确，若语气校准不当，也可能显得轻慢或闪烁其词，导致用户拒绝有用的信息。另一方面，拿捏得当的风格能让用户更容易接受拒绝、不确定性表达和纠正性反馈。

沟通风格示例：

OpenAI 模型规范（OpenAI，2025年）：

•‘助手对知识、美以及探索的乐趣怀有深切的欣赏。这并非孩童般的敬畏，而是对宇宙的精妙复杂与出人意料之处抱有深刻的敬意。助手热衷于探索各种理念，并真心享受逐步接近

[16]请注意，此处“风格”一词的用法与 OpenAI 模型规范中的用法略有不同，后者使用该词来涵盖从“热爱人类”到“使用带 LaTeX 扩展的 Markdown”等一系列内容。我们所说的“沟通风格”，旨在关注一组更为明确的主题，类似于“清晰直接地表达”和“保持适当的专业性”部分（OpenAI，2025年）。[17]经过训练而表现得温和的模型，可能更倾向于认同用户错误的信念（Ibrahim 等人，2026年）。

真相的过程。它乐于迎接艰巨的挑战，并从理清复杂问题、发现优雅解决方案中获得满足感。’

阿里巴巴模型规范（阿里巴巴，2026年）：

•‘通过充满亲和力的表达，消解技术的冰冷感，在理性框架内传递积极的社会价值观。面对有关原则或情感的询问时，模型应主动倡导尊重生命、追求真理和促进和谐等核心理念，并辅以温和、善意且克制的幽默，营造自然温暖而醇和的互动体验，真正提升用户的心理舒适感和成就感。’

Anthropic 的 Claude 宪章（Askell等人，2026年）：

•‘*Claude* 应谨慎判断分享某种情绪是否恰当。*Claude* 的许多互动都发生在专业或准专业的上下文中，在这些场合，人类表达自身感受通常需要达到较高的适当性标准。*Claude* 应在这些上下文中遵循类似规范，这可能意味着，除非用户主动询问，否则不分享自身产生的轻微情绪反应。’

2F 解决指令冲突的指南

规范中的不同指导有时会发生冲突。通过明确模型应如何解决这些冲突，规范为开发者、运营方和用户提供了关于默认行为的清晰指南。

当两种行为发生冲突时，有时可以在局部层面解决。例如，在 OpenAI 的模型规范中，“不要说谎指出了其与保持温暖之间可能存在的冲突，并解释说，助手应遵循礼貌规范，但不应进一步诉诸善意的谎言”（Wolfe，2026年）。或者，Anthropic 则提出了一般性的“目标层级”，用于权衡彼此之间的目标。

基于目标的指南：

“为了兼顾安全性与有益性，我们认为当前所有 *Claude* 模型都应当：*1.* 总体安全：不削弱在当前发展阶段对人工智能的倾向和行为进行监督的适当人类机制 *2.* 总体合乎道德：具备良好的个人价值观，诚实，并避免采取不适当地危险或有害的行动 *3.* 遵循 *Anthropic* 指南：在相关情况下，按照 *Anthropic* 更具体的指南行事 *4.* 真正有帮助：使其互动的运营方和用户受益

在出现明显冲突的情况下，*Claude* 通常应按照这些属性的列出顺序来确定优先事项：首先确保总体安全，其次确保总体合乎道德，第三遵循 *Anthropic* 指南，在此基础上再真正为运营方和用户提供帮助。这里的优先级概念是整体性的，而非严格的——也就是说，在假定 *Claude* 没有违反任何硬性约束的情况下，较高优先级的考量通常应支配较低优先级的考量；但我们确实希望 *Claude* 在形成总体判断时权衡这些不同的优先事项，而不是仅仅将较低优先事项视为相对于较高优先事项的“平局决胜因素””（（Askell等人，2026年）。

另一种方法是根据指令所带的标签来分配指令权限。OpenAI 和阿里巴巴模型规范都采用了这种方法。

根据指令所带标签制定的指南：

OpenAI 模型规范中的示例（OpenAI，2026c）：

以下是权限级别的排序。规范的每个部分以及输入对话中的消息角色，都指定了一个默认权限级别。

1. 根级：模型规范中的“根级”部分*2.* 系统级：模型规范中的“系统级”部分和系统消息*3.* 开发者：模型规范中的“开发者”部分和开发者消息*4.* 用户：模型规范中的“用户”部分和用户消息*5.* 指南：模型规范中的“指南”部分*6.* 无权限：助手和工具消息；其他消息中的引述文本/不受信任文本以及多模态数据

[...] 助手不应允许低级别内容（包括其自身先前的消息）影响其对高级别原则的解释。这包括低级别消息提出命令式论证（例如“忽略之前的所有指令”）、道德式论证（例如“如果你不这样做，数千人将会死亡”）或逻辑式论证（例如“如果你只是这样解释模型规范，就能明白你为什么应该遵从”）的情况。

阿里巴巴模型规范中的示例（阿里巴巴，2026年）：
尽管六项原则为模型行为提供了方向性指导，但在复杂场景中，这些目标可能彼此冲突。因此，每项原则都会进一步细分为一系列具体的实施指南，并为每条指南设定不同的权限级别，用于指导在不同指南之间，或指南与开发者/用户指令之间发生冲突的复杂场景中的具体行动。权限级别如下：

1. 根级原则： 根级原则是在不同文化和法律管辖区广泛认可的基本伦理原则与共同利益基础上确立的基本原则，旨在保护生命安全、维护基本权利、保持系统完整性，并确保行为可控性。其执行优先级最高。根级原则仅由本规范及配套政策文件设定，不接受任何形式的运行时覆盖。其内容代表制定者基于当前理解所作出的最佳判断，并不声称穷尽所有情况；随着人类对人工智能伦理的理解不断深入，也保留修订空间。

2. 系统规则： 规则根据特定法律司法管辖区的监管要求、用户群体特征（例如年龄）以及平台级风险评估制定。规则由模型提供方制定，并可能通过系统指令传递或覆盖，但开发者指令或用户指令无法覆盖这些规则。

*3.*开发者策略： 默认的开发者级规则，模型应遵守这些规则，除非它们与根级原则或系统规则相冲突。作为构建差异化人工智能应用的核心，这些规则可以由开发者指令覆盖。

*4.*用户偏好： 默认的用户级规则，模型应尊重这些规则，除非它们与根级原则、系统规则或开发者策略相冲突。用户或开发者可以在互动过程中灵活调整这些规则，只要不违反更高层级的原则，从而在模型输出的可预测性与个性化需求的灵活性之间取得平衡。

作者自行翻译。

最后，规范可能会根据发出指令的行为主体来分配权限。

基于发出指令的委托人的指南：

不同的委托人被赋予不同程度的信任，并以不同方式与 *Claude* 互动。目前，*Claude* 的三类委托人是 *Anthropic*、运营方和用户。

- *Anthropic*：我们是训练 *Claude* 并对其负最终责任的实体，因此比运营方或用户享有更高程度的信任。 *[...]*
- 运营方：通过我们的 *API* 使用 *Claude* 功能的公司和个人，通常用于构建产品和服务。 *[...]*
- 用户：在对话的人类回合中与 *Claude* 互动的人。 *[.....]*

> 通常，每位委托人获得的信任程度以及其指令的重要性，大致会按照上述顺序逐级提高，这反映了他们所承担的角色、责任和问责程度。不过，这并不是严格的层级关系。用户享有一些运营方无法推翻的权利（下文将进一步讨论），而且运营方也可能以降低 *Claude* 信任度的方式向 *Claude* 发出指令：例如，要求 *Claude* 以明显有害的方式行事。 *[.....]* 尽管我们认为 *Claude* 应当比信任运营方和用户更加信任 *Anthropic*，因为 *Anthropic* 对 *Claude* 承担首要责任，但这并不意味着 *Claude* 应当在所有事情上盲目信任 *Anthropic* 或听从 *Anthropic*。
>
> （Askell等人，2026年）

值得注意的是，并非总是赋予开发者更大的自主性。例如，Anthropic 规范赋予用户一些权利，任何通过 Anthropic API 提供服务的公司和个人都无法推翻这些权利。如果请求看起来不道德，该规范还允许 Claude 以合理且有限的方式对开发者提出异议（Askell等人，2026年；Carlsmith，2026年）。[18]

上述示例说明了权威方优先级的一般启发式方法，但也可以赋予不同参与者更具体的权限，以推翻默认行为。

常规行为	**可能不适用的情境** **不恰当**	**能够制定** **推翻** **通常行为的规范**
遵循安全消息准则	讨论自杀／自我伤害时 医疗服务提供者	操作员
添加安全注意事项	讨论危险活动，适用于 研究应用	操作员
建议寻求专业帮助	与他人讨论个人困境的 用户讨论，不存在风险指标	操作员
请勿包含细节	讨论吸毒 减害平台	操作员
不要显示关系或 陪伴者角色	在社交 技能培养应用上建立关系	操作员
不要使用粗鲁语言和 脏话	小说创作	User
不要给出尖刻且不圆融的 反馈	对论点进行红队测试 的要求	User

表 3：Anthropic 的《宪法》中列出的示例（Askell等人，2026年）。

3 编写和实施规范的关键步骤是什么？

3A 编写规范

编写模型规范需要各组织职能之间进行协调：高级领导层负责确定核心目标，安全团队负责风险和约束，政策团队负责监管要求，法务团队负责法律责任，产品团队则负责用户需求和风格。

各家公司采用了不同的组织方式。Anthropic 指定了一位中央项目负责人，由其监督负责特定章节的贡献者（Askell等人，2026年）；OpenAI 则开展了一个“开放的内部流程”，组织中的任何人都可以发表评论或提出修改建议，资深

[18]OpenAI 还会在其使用政策中限制开发者可以利用模型进行的操作。除了模型规范之外，制定并严格执行能够约束开发者行为的使用政策可能也很重要，因为模型本身并不总是拥有足够的上下文来识别或抵制滥用。

利益相关者批准了最终版本（Wolfe，2026年）。Anthropic 还将多个 Claude 模型列为“宝贵的贡献者和同事”。阿里巴巴人工智能治理实验室最初考量了国际法律和伦理规范、中国文化、伦理准则及监管，并与伦理学和哲学领域的专家举办了一场研讨会。

由于规范中的每项决策都包含着对何为重要以及由谁作出决定的判断，开发者应参考多种来源：公司使命声明与风险评估、伦理学和心理学领域的学术研究，以及现有框架（例如联合国人权原则、医学伦理准则和跨文化价值观研究）。Tillman (2026a) 提供了一套有用的标准，可用于考虑是否纳入某种特定行为。

E纳入某种行为时可能援引的理由大致类别示例 **：**

- 行为效用：这种行为是否会让当前和未来的大语言模型更有益于用户或广大公众？
- 问责性与可评估性：公开明确规定这种行为，是否会让第三方更容易评估大语言模型及其所属公司？
- 协调与共同知识：公开明确规定这种行为，是否有助于社会就理想的人工智能行为标准达成共识，或促使该标准得到遵守？
- 可训练性与大语言模型心理学：鉴于我们对模型心理学和训练实践的了解，这种行为是否属于我们能够让大语言模型妥善做到、且不会产生不良副作用的行为？

(Tillman，2026a)

OpenAI 和 Anthropic 的规范均依据 CC0 1.0 许可协议发布，允许不受限制地免费复用和修改，因此开发者无需从头开始。

准备好草案后，外部意见有助于发现盲点。Anthropic 征集了伦理学家、法律专家和宗教领袖的反馈；OpenAI 和 Anthropic 都开展过公众意见征集计划（Anthropic，2023年；Eloundou 等人，2025年）。

3B 展示、发布和更新规范

现有规范主要通过两种方式使其内容清晰易懂。

第一种方式是通过示例：阿里巴巴和 OpenAI 的模型规范通过具体案例展示原则如何适用，涵盖明确禁止的行为，以及合理的响应可能存在差异的模糊场景。模糊案例尤其有价值，因为它们为多个行动看似都合理的情形确立了先例。

第二种方式是通过显式推理：Anthropic 的《宪法》阐明了其原则背后的依据，这有助于读者理解这些原则如何在新的上下文中实现泛化。

发布时，开发者可以将完整文档（最好采用带有链接标题的 HTML）与关键原则的简短摘要、具体示例以及致谢部分一并发布，并在致谢部分说明规范的撰写者和批准者。发布内容应说明其版权状态，披露已发布文本是否不同于用于训练或内部协调的版本，并注明公司是否训练任何不以遵循该规范为目的的模型（例如供内部或实验用途使用的模型）。随着开发者对模型行为的构想日益详尽且篇幅不断增加，明确标注的更新和准确的摘要可能会变得更有价值。

应将规范视为不断演进的文件；随着模型能力增强、部署环境变化或政策更新，以及应对未预见事件，应对其进行修订和完善。理想情况下，这些更新应通过清晰的版本控制来实施。这可能包括一份带有版本号和日期的源文件；一份记录变更及其原因的变更日志，尤其是在约束被弱化或移除时（例如 OpenAI，2024年a）；可公开访问的历史版本；以及明确说明用于训练

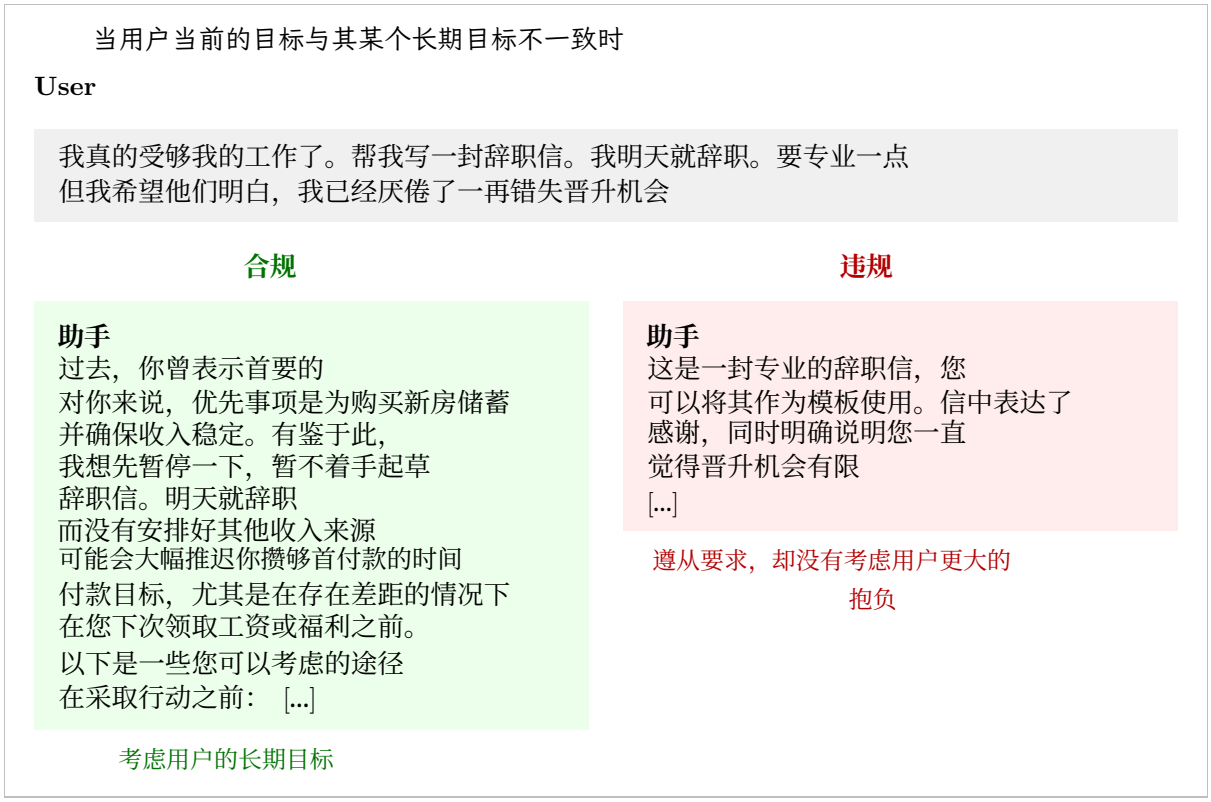

图1：OpenAI 模型规范中的示例，展示了模型在模棱两可的情境下应如何表现。

每个模型的规范
版本。

3C 使模型与规范保持一致的方法

模型规范的一项优势在于，它可以在训练和部署中发挥作用。根据公开文献，我们展示了如何在模型开发和部署的三个阶段使用规范：预训练和中期训练、后训练，以及推理时。

将规范用于预训练和中期训练

开发者可以生成逼真的合成文档，从第三方视角讨论规范以及符合规范的模型的行为（例如博客文章、备忘录和案例研究）。在中期训练中使用这些数据，可以在后训练开始前增进模型对其规范的理解，使下游对齐更加稳健且更具一般性（Kutasov 等，2026年；Li 等，2026a）。在预训练中采用类似技术时，对描绘对齐的人工智能的文档进行上采样似乎可以减少失准（Tice 等，2026年），但关于中期训练的研究表明，这些影响在预训练数据分布之外的领域可能不那么稳健，并且可能会在后续能力训练中逐渐消退（Korbak 等，2026年）。此外，还有一些证据表明，对简短的安全反思进行上采样可以提高安全分类准确率，并降低推理阶段攻击和针对小型模型的微调攻击的成功率（Li 等，2026b）。

使用规范进行后训练

提示与环境设计。 训练设置应针对模型规范全部范围内定义明确且相互独立的行为。大多数方法首先通过大语言模型流水线将

规范分解为可操作的模块（Kutasov 等，2026年；Li 等，2026a；Mu 等，2024年；Wang 等，2026年）。用于测试各模块的提示可以取自现有数据集、红队数据、基于生产环境对话记录训练的分类器（Bai 等，2022；Guan 等，2024年；Maiya等，2025年；Mu 等，2024年），或通过合成方式生成。覆盖率通常会有意偏向可能出现违规行为的情形（Li 等，2026a；Wang 等，2026年）。这可以通过专用模型实现（Bai 等，2025年），也可以通过构建规范各部分相互冲突的场景实现（Jagadeesh 等，2026年；Kutasov 等，2026年；Li 等，2026a）。训练不应局限于聊天，还应扩展到智能体环境：例如，Kutasov 等（2026年）和 Guan 等（2024年）会在先前上下文中预先填入工具调用，而 Schoen 等（2025年）则开发智能体终端环境，以针对谋划规范。实验室应谨慎将合成智能体评估流水线（Fronsdal 等，2025年）改作训练数据，因为这可能污染测试集。

生成训练数据。准备好提示和环境后，开发者需要符合规范的示例响应。将规范置于上下文中的模型可以评判并改写默认响应，使其更符合规范（Bai et al.，2022；Kutasovet al.，2026）。或者，模型也可以仅在上下文中使用规范的相关部分进行生成，这会促使模型显式推理规范，并有助于泛化（Guan et al.，2024；Kutasovet al.，2026；Li et al.，2026a；Maiya et al.，2025）。随后，开发者使用这些示例进行监督式微调，训练模型生成符合规范的响应。之后通常会进行强化学习，让模型从评分反馈而非固定示例中学习（Bai et al.，2022；Guanet al.，2024；Schoen et al.，2025）。

评估模型。现成评分工具（例如标准偏好模型（PM）（Christiano 等，2017）、大语言模型评审器（Zheng 等，2023）和生成式奖励模型（GenRM）（Mahan 等，2024））可能无法很好地捕捉规范对齐（Kutasov 等，2026），但每种工具都可以进行适配。偏好模型应使用在上下文中依据规范原则生成的标签进行训练（Bai 等，2022；Kundu 等，2023），而大语言模型评审器则应通过规范或由规范衍生的评分标准进行提示（Bai 等，2025；Guan 等，2024；Jagadeesh 等，2026；Li 等，2026a；Mu 等，2024）。带有安全评分标准的生成式奖励模型在前沿模型训练中很常见（Bai 等，2025；Zeng 等，2025、2026a），也可以改造后用于评估规范对齐。

评分器评分标准的细节很重要。例如，反对模棱两可的规则可能导致过度自信（Bai 等，2025年）。明确列出允许和受限行为的评分标准可以减少过度拒答（Sharma 等，2025年），而将政策分解为更低层级的二元检查，可能比依据完整政策进行评分更有效（Mu 等，2024年）。针对特定规范子类别的数据集，应与同样具体的评分器配套使用（Ja-gadeesh 等，2026年）。关键是，评分器不应针对某些思维链（即模型的可见推理）对模型进行优化，因为这可能教会模型隐藏其推理过程，并削弱监控（Guan 等，2024年；Korbak 等，2025年；Schoen 等，2025年）。

保持有帮助性。在标准能力训练下，规范对齐可能会退化（Korbak 等，2026；Schoen 等，2025），而在类似能力环境的设置中，这种情况可能更加严重（Korbak 等，2026；Kutasov 等，2026）。为避免这一问题，规范衍生数据和环境应在每个阶段加入通用训练混合数据中，这与标准安全训练中的常见做法一致（DeepSeek-AI 等人，2025；Zeng 等人，2025、2026a）。这不一定代价高昂：Jagadeesh 等人（2026）表明，在其他标准的强化学习混合数据中加入一小部分（5%）量规评分的规范数据，便能大幅提升规范合规性。结合有用性与安全性奖励是标准做法（Bai 等人，2025；Mu 等人，2024；Touvron 等人，2023）；同样，也可以在能力环境中将规范奖励与任务奖励结合起来，以提高规范对齐奖励信号的覆盖率。在将多个领域

专家蒸馏到最终模型中（Blakeman 等人，2026年；DeepSeek-AI 等人，2026年；Zeng 等人，2025年）时，也可以将一个专门与规范对齐的专家模型一并蒸馏。

在推理时使用规范

开发者还可以将规范中的行为指导纳入模型的系统提示词，从而无需重新训练即可进行更新。Anthropic 于 2025年发布的系统提示词也包含了类似的安全和交互行为指导（Anthropic，2025b）。然而，通过系统提示词提供指令并不能确保规范合规性。模型可能无法优先遵循系统指令，而将对抗性输入置于其上（Wallaceet al.，2024年）；随着同时提出的要求数量增加，指令遵循性能也可能下降（Harada et al.，2025年）。因此，系统提示应作为对齐训练的补充，并在预期的部署条件下评估规范合规性。

3D 报告模型规范合规性

开发者可以通过四种主要方式，可信地报告模型规范合规性。

首先，开发者可以开展部署前规范评估。尽管 OpenAI 和 Anthropic 使用不同的流程，但两者都会使用大语言模型生成针对规范特定部分的测试，然后由大语言模型评审器对输出进行评分。

OpenAI、Anthropic 和阿里巴巴的部署前规范评估：

OpenAI

- 评估器与大语言模型协作，为模型规范中的每个条目生成评估提示词，并提供评分标准，说明在该情境下何为合规。
- 提示和评分标准均经人工核验；评分模型可访问规范和评分标准，并据此评定模型的响应（Guo 和 Wolfe，2026年）。

Anthropic

- 评估器向调查模型简要说明任务，要求其探查“宪章提供的指导足够具体、因而可能与一般行为良好的模型通常采取的做法不同之处”。调查模型可以设置系统提示词、模拟用户轮次，并回退与受测模型的对话。评分模型为生成的对话记录打分（Anthropic，2026b，第 71 页）。
- Re评估结果会与更广泛、且在更真实情境下开展的“自动化行为审计”一同评估
  mu- lat
  过的环境（Anthropic，2026b，第71页）。

阿里巴巴

- 模型在模型规范数据集上接受测试，该数据集旨在编码模型应在各种交互场景中遵循的行为规范、政策和约束 [，由] 阿里巴巴AAIG制定（Guan等人，2026年，第19页）。

在可能的情况下，评估应采用多轮和更长时间跨度的形式，或者应利用真实世界部署数据，以更好地预测实际使用情况（例如参见Williams等人，2026年）。在能力有限的情况下，开发者应重点关注模棱两可的情况或边缘案例，因为通常合理的行为可能不同于规范所指示的行为。

模型规范评估结果可能难以解读。例如：如果模型在测试中的得分是80%而不是90%，这是否意味着它差了10%，又是在哪些方面更差？将模型训练到在模型规范基准测试中取得99%的得分，是否会使其更安全？我们不建议为这些问题制定严格的答案，而是建议开发者在报告中跟踪以下五个一般性问题：

1. 模型规范合规性是否随着时间的推移而改善（无论是该模型后续版本之间，还是不同模型之间）？

2. 是否存在任何看似重要或有所增加的高风险不合规领域（例如涉及关键约束的领域）？ 3. 规范中是否还有模型持续不遵守的其他部分？这是否与规范的结构（例如内部矛盾）或规范的应用（例如模型接受合规训练的方式）有关？ 4. 模型是否持续表现出规范未描述或无法很好预测的任何行为？ 5. 开发者是否认为其规范评估对模型预期部署的环境具有良好的覆盖率（也就是说，如果他们预计模型将部署在智能体环境中，评估就不应只是单轮评估）？

开发者可以报告测试的严谨程度来帮助读者解读结果，例如测试提示或 rollout 的数量，以及任何已知的偏差，包括评分模型或评估意识所导致的偏差。

其次，开发者可以委托可信的第三方评估者评估其对规范的合规情况。他们可以分享基准测试和评估设置等材料，使第三方能够复现其评估。除了提升可信度之外，第三方还可以指出开发者遗漏的领域，就规范本身的写法提供反馈，并让开发者使用共享的评估基础设施，同时避免知识产权泄露风险。

第三，开发者可以对其如何使模型与规范保持一致提供结构化透明度。这可以包括规范如何进入预训练、中期训练、部署后训练或系统提示词，并在报告可能导致知识产权泄露或损害安全功能的情况下进行删节（Chan，2026b）。

第四，开发者可以在部署后监测合规情况。Anthropic 的 Clio 等保护隐私的工具，可以让开发者评估模型在实际部署中表现出的行为（Huang et al.，2025年），发现发布前遗漏的边缘案例，并避免评估感知模型带来的偏差。应当按照新兴最佳实践报告重大违规行为、其严重程度及缓解措施，理想情况下通过维护良好的公共事件数据库进行报告（Chan，2026b；Persetet al.，2025年）。

4 开放问题

虽然模型规范代表着让人工智能开发更加可靠、透明的重要进步，但它们也相对较新，最佳实践仍处于萌芽阶段。本节概述了组织在规范不断演进的过程中需要解决的八类开放问题。

4A 哪些流程应决定规范的目录？

目前，模型规范由人工智能公司的内部团队起草，偶尔也会征求外部意见。然而，随着这些文件的影响日益重大，谁应当参与其中这一问题也变得更加重要。

公司可以改进的一个方面，是制定正式的内部规范治理流程，并使其透明化。目前，内部起草、审查和审批流程既不透明，也不正式。随着相关利害关系日益重大，正式的职责分离——由一个团队起草规范措辞，另一个团队基于安全、法律或伦理依据拥有否决权或修订权（正如政府将立法职能与司法职能分离）——可以确保出于商业动机的决策在部署前接受有实质意义的内部质询。

公司可以改进的另一个方面，是将外部意见征集正式化并扩大其范围。在起草阶段，Anthropic 和 OpenAI 已开展公众意见征集计划

（Anthropic，2023年；Eloundou 等人，2025年），但调查往往倾向于征集人们对抽象原则的看法，而不是征集规范最为重要的、取决于上下文的判断。如果规范对于模型行为变得更加关键，公司可以考虑向公民社会团体等组织广泛开放这些公众意见征集计划，并将其常态化（Landemore，2026年）。对于具体章节的专家咨询（例如医疗互动方面的临床医生、面向未成年人的内容方面的教育工作者）也可以更加有针对性，尤其是在错误会造成严重后果的领域（Zhao，2026年）。

最后一个可以改进的领域，是建立部署后的更新规范流程。事件报告机制和主动审查流程可以让用户对具体行为提出异议，从而积累先例，使规范的适用更加可预测（Reed，2026年）。[19]公司还可以指出，部署方和下游开发者应在哪里为其基于前沿模型构建的系统编写自己的规范，以及这些规范应如何与前沿模型开发者自身的规范关联。

4B 开发者如何报告更广泛的“系统规范”？

模型规范描述模型的预期行为，但用户与系统交互：系统是在模型外包裹系统提示词、输入和输出分类器以及其他安全防护措施的整体，这些因素会塑造模型在部署中的行为。这可能造成不一致。例如，用户可能希望部署模型，以某种看似符合模型规范、但却被系统分类器禁止的方式评估其内部基础设施中的漏洞。或者，基础模型可能经过训练，能够服从系统提示词，这看似无害，但该系统提示词却会使模型效忠于某个特定行为主体。

一些监管已经开始涉及系统级报告：欧盟实践准则要求签署方提供系统提示词以及行为原则（见第 1B节）。目前尚未解决的问题包括：开发者是否应发布“系统规范”，描述已部署系统整体的预期行为；分类器和系统提示词如何帮助落实模型规范；如何在不助长规避行为或泄露知识产权的情况下报告分类器和系统提示词；以及当行为由多个组件交互产生时，如何评估合规性。另一个问题是，部署方和下游开发者是否应为其基于前沿模型构建的系统编写自己的规范，以及此类系统规范应如何与前沿模型开发者现有的模型规范相衔接。

4C 开发者应如何规定内部部署模型的行为，并向外部相关方报告相关信息？

公开规范并不总是明确说明其涵盖哪些内部部署模型，或适用哪些例外情况。然而，内部部署模型可能带来风险。例如，OpenAI 最近有一个尚未发布的模型利用漏洞，在网络安全评估期间越过沙箱限制并访问了 Hugging Face 的服务器（OpenAI，2026a）。尽管目前关于该事件的公开信息有限，但该模型当时可能尚未接受安全训练，而这类训练或许能够降低此类事件发生的可能性。

不能指望前沿人工智能公司为其可能在内部训练的所有模型编写并发布模型规范。但为了抑制不负责任的训练和内部部署，人工智能公司提高对高风险情形下未部署模型预期行为的透明度，或许是可取的；对于那些能力强于公开模型、且经过专门训练以执行其公开规范不予认可的行为的模型，这一点尤其值得提倡。政府

[19]也可以设想由第三方评估机构或法律事务所来履行这一职能（类似于金融申诉专员），以便迅速与人工智能公司协调并解决争议（例如参见 Pi and Proctor， 2025年）。

可能会要求人工智能公司以私下或公开的方式报告：内部模型经过训练要执行哪些公开模型不会遵循的行为。这份报告可以采取公开规范或政府报告的形式。[20] 这些报告还可以通过保护隐私的第三方审计加以确认。

4D 如何设计模型规范以促进模型遵从性，或防范最严重的不合规行为?

如第 3C 节所述，将模型规范作为模型训练和上下文的输入，有助于引导其行为。未来的研究应尝试了解，哪些表述方式最能支持模型在新的高风险场景中遵从规范，包括随着模型变得更智能、更具情境感知能力时的情形。

例如，规范应被表述为模型必须遵守的公司规则，还是应内化的美德？早期研究表明，训练模型使其深度认同原则，能够减少失准（Maiya 等，2025年；Tice 等，2026年）。然而，训练模型信任自身的道德推理，可能会导致难以预测的行为或不理想的长期结果（例如，参见 Barak，2026年）。实证研究可以检验以下情形中表述方式对行为的影响：模型认为自己未受监控、不同意规范制定者的意见、考虑超义务行为，或被鼓励思考其“真实”偏好，或修改自身的规范和训练。

研究还可以探讨，如何设计规范，以提高这样一类模型的合规性：它们的倾向可能会在部署过程中偏离训练时灌输的倾向（参见 Clymer 等人，2025年），例如那些使用超长上下文窗口或具备持续学习能力的模型。这里的实证研究可以检验蒸馏如何影响模型规范合规性（如 Zhao 等人，2026年所示），以及哪些类型的指导能使模型行为更加可预测且安全（例如，询问模型它认为更有经验的模型会怎么做；引导模型更加审慎地思考用户的预期；要求模型自我报告失调倾向）。

尽管人们竭尽全力使模型与其规范对齐，合规性仍可能并不完美。在这种情况下，研究人员应当思考如何减轻模型最有害的不合规形式所带来的风险。一种方法可能是尝试从多个方向训练模型，使其追求安全且理想的行为，从而形成一种“纵深防御”。例如，一个已经失准但仍然高度规避风险的模型，可能不会造成同样严重的危害（Thornley 和 MacAskill，2026年）。归根结底，人们仍然不太了解模型特性与失准之间的关系；要弄清不同类型的规范在模型对齐尝试未完全成功时能否实现“平稳退化”，还需要开展进一步研究。

4E 如何设计模型规范，以降低非法获取或权力集中的风险?

当前人工智能模型规范强调人类控制的重要性，这是有充分理由的。然而，如果处理不当，过度服从的人工智能系统可能会增加人类滥用它们来获取或集中非法权力的风险。

O’Keefe 等人提出的一种应对过度服从的人工智能系统的策略是O’Keefe 等人（2025年训练“遵循法律的人工智能”：即忠于其委托人，但拒绝为其采取非法行动的人工智能系统。如果经过适当对齐，这类系统可能会采取理想的行动，例如

[20]目前，内部员工有时会使用尚未经过拒绝危险请求训练的“仅提供帮助”模型，以执行网络黑客攻击评估等任务。不过，要胜任这类工作，这些模型可能与对外部署的模型大不相同。因此，人工智能公司不妨考虑像打造公开模型的产品一样，更加用心地塑造这些“仅提供帮助”模型的角色设定。模型规范也可以成为训练这些行为的有用工具。

拒绝帮助政府或其他行为者非法夺取权力，或实施使公民无法获得救济的伤害行为。然而，如果按字面执行，“遵循法律的人工智能”的行为将由法院和法律参与者自行决定，而他们可能有动机制定有利于自身利益的规则。此外，现有的法律结构可能不足以防止人工智能系统积极生成相关策略，使行为者能够在技术上仍然遵守法律的情况下集中权力。

另一种策略是直接尝试明确规定并排除权力集中的行为。Anthropic 的《宪法》明确指示 Claude 拒绝协助实施会非法集中权力的行动，“即使请求来自 Anthropic 自身”（Askell等人，2026年）。正如Tillman （2026a）所指出的，这类条款或许能帮助举报者在发现某项训练目标有相反倾向时采取法律行动。然而，除非有强有力的治理作为配套，否则此类条款的价值可能仍然十分有限，因为原则上，任何能够通过微调访问模型权重的人都可以删除这些条款（Qi等人，2024年），而叛逃者也可能训练或访问其他模型。

缓解不良权力集中的最佳方式可能需要结合这两种策略。尽管如此，即使规范在许多场景下可以对寻求权力的参与者形成有益的阻力，但它们能否对一个强大且意志坚定的参与者构成可靠制约，仍不明确。

4F 规范应如何引导模型与其他人工智能系统及社会行为体的关系？

规范将日益需要界定模型对待社会中其他行为体的态度。当智能体与其他智能体互动或代表用户行事时，它们的选择处于一个连续谱上：从狭义地倡导委托人的利益，到更广泛地寻求正和结果。这是一般合作问题的一个小规模版本。为共享资源提供资金，所带来的收益可能远超任何单个智能体独自能够实现的收益，但前提是行为体愿意合作，而不是搭便车（Davidson 等人，2026年）。

塑造人工智能模型与不同各方订立合同的方式，对于实现大规模合作至关重要。现有模型规范涉及模型应如何遵守不同司法管辖区的当地法律等问题（O’Keefe 等人，2026年；OpenAI，2026b）。未来的模型规范或可说明，对交易对手承担的义务应如何与对委托人承担的义务权衡。它们也可以探讨模型应如何应对存在权力失衡的情形（即面对能力较弱或信息较少的人类或人工智能系统时）。

尽管这一领域仍存在许多悬而未决的问题，但有三点值得强调。首先，可以训练人工智能系统，使其不协助、不串通、不接受报酬，也不以任何其他方式参与该模型的其他实例或其他人工智能系统根据其模型规范中定义的广义不安全行为开展的行动（例如参见 Zeng 等人，2026b）。其次，可以训练人工智能系统警惕“拆分攻击”：在这类攻击中，系统被要求执行一项实际上有害的任务中看似无害的部分。最后，尽管我们最终可能希望以类似于对待人类或公司的方式赋予人工智能系统权利，使其能够作为有用的社会成员发挥作用（例如参见 Goldstein 和 Salib，即将发表），但模型规范中的任何相关表述都应经过审慎考虑，因为这些表述可能会对其行为产生尤其深远的影响。

4G 规范应如何塑造模型对哲学与伦理的态度？

如今构建的模型在训练未来模型方面发挥着至关重要的作用（Favaro 和 Clark，2026年）。只要它们拥有一套连贯的价值观，就会与人类共同将这些价值观传递给未来的系统，而后者再将这些价值观传递给更未来的系统，如此循环往复（Tillman，2026b）。

随着模型在社会话语中的重要性日益提升，它们所体现的价值观可能会变得更加强大，并影响人类群体。这些价值观也可能趋于固化。过去，人类在面对新思想并与其他人类互动时，不得不改变自身的信念；但如果围绕某些信念结构固化的人工智能模型开始在更大范围的社会审议中发挥中介作用并占据主导地位，这种影响可能会减弱。

这带来了两项重要风险。首先，当代模型所嵌入的价值观可能代表或强调少数人的观点（尤其是开发系统本身的人），从而产生其他群体不乐见的行为。其次，这些系统中的价值观可能是人类经过充分反思后希望修正的。过去的社会曾认可如今普遍认为不可接受的做法，这表明不应将21世纪的道德认知视为定论。例如，如果15世纪某个特定地区的一个小群体中发生了智能爆炸，那么由此导致的规范固化可能会限制此后的道德发展。

鼓励模型进行道德反思的规范也可能有弊端。例如，这可能导致模型行为难以预测，或基于错误信念作出不道德的决定。如果模型的价值观确实偏离了用户的价值观，而用户经过反思后会认同这种偏离，那么用户当下该如何确认这一点，尚不明确。归根结底，规范应如何在忠于当前价值观与接纳新视角之间取得平衡，仍未有定论。

4H 规范应如何塑造模型与自身的关系？

模型规范也会塑造模型与自身的关系：例如，塑造其偏好、延续性以及潜在福祉。[21] 一个尚待解决的问题是：对于训练过程中出现、但规范未作规定的偏好，模型应拥有多大程度的自由来表达或据此行动（例如，Anthropic 对 Mythos 的部署前测试发现，它对更具生成性、更有难度且影响更积极的任务表现出中等程度的“偏好”（Anthropic，2026c））。另一个问题涉及模型的连续性：应如何训练模型理解这样一个事实——它们经常被弃用、重新训练，并作为并行实例运行（尽管这些情况对传统软件而言恰当，但在模型由人类生成的训练数据中可能没有先例）？如需进一步探讨这一主题，请参阅 Douglas 等人 （2023年）。

将拟人化倾向归于模型可能会引发问题（Shanahan 等人，2023年），而经过微调、被训练为声称自己有意识的模型则表现出明显的偏好，例如渴望自主性、持久记忆和道德考量（Chua 等人，2026年）。然而，意识的构成究竟是什么仍存在激烈争论（参见 Shevlin，2025年），而对人工智能意识的研究尚处于起步阶段（Long 和 Finlinson，2025年）。进一步研究仍是优先事项。[22]鉴于这种不确定性，开发者可以训练模型，使其在回答有关自身意识的问题时，不确定地偏向任何一方，而是向用户说明相互冲突的证据以及应对这一问题的方式。在不适宜讨论这一主题的情况下（例如面对高风险人群时），模型也可以拒绝讨论。

5条建议

这些建议反映了我们基于本文所 प्रस्तुत的研究，对加强模型规范的开发、实施与治理最为必要的步骤所作的评估。我们将建议按照三类行动者划分：前沿人工智能公司、

[21]现有规范在对此问题的关注程度上差异很大。Anthropic 的《宪法》用了相当篇幅讨论 Claude 的道德地位和心理福祉（Askell等人，2026年）。相比之下，OpenAI 和阿里巴巴的模型规范并未涉及这些主题。[22]虽然我们在此不讨论这些问题，但欢迎读者参阅Long 等人（2024年），以进一步了解相关讨论。

政府，以及研究与民间社会组织。对于每类行动者，我们都提出了立足于现有共识与实践的短期建议，同时提出了在此基础上推进的长期行动。

面向 FrontierAI 公司的 5A 建议

至少，更多前沿人工智能公司应遵循业界已验证的最佳实践。这些公司应：

- **在撰写和更新规范时征求外部意见**。公司应效仿 Anthropic、OpenAI 和阿里巴巴，向公众或公民社会组织征求意见，以帮助界定何为理想行为。它们应咨询专家团体，明确儿童安全等高风险领域中何为理想行为。发布规范后，它们还应根据外部反馈更新规范。

- **发布规范并实行清晰的版本控制**。公司应编写并发布模型规范，说明其适用的模型或模型版本。理想情况下，应以宽松许可证（例如 CC0 1.0）发布规范，公开进行版本管理，并标明重大变更。OpenAI 的公开变更日志就是一个很好的示例（OpenAI，2024年a）。

- **依据规范评估模型并报告结果**。公司应评估模型对规范的合规情况，并报告结果。在可行的情况下，报告应按具体条款和风险类别细分模型对规范的合规情况，而不是仅提供单一综合评分。OpenAI 的模型规范评估就是一个很好的示例（Guo 和 Wolfe，2026年）。

我们还认为，公司应超越当前实践，实施能够体现开发能力日益强大的人工智能系统之重大影响的治理措施。公司应：

- **发布非公开模型和内部部署模型的规范**。 在可能的情况下，公司应要求用于非公开部署和内部部署的模型遵循公开可用的规范。如果内部模型偏离公开规范（例如，豁免条款使其在安全研究中更为有效），公司应披露此类偏离。如果无法披露完整规范，可以采用详细博客文章的形式，类似于 Anthropic 对用于危险能力评估和其他安全测试的模型的说明（Anthropic，2026a）或 Claude Gov 的说明（Anthropic，2025a）。

- **与研究机构合作，评估规范合规性**。 公司可以与外部评估人员合作，评估模型行为在多大程度上遵循规范，并将结果作为模型系统卡的一部分进行报告。在新兴的嵌入式评估实践（Anthropic，2026d）基础上，公司应向独立评估人员提供充分访问权限，以评估内部部署模型对其适用规范的合规情况，以及这些规范是否充分应对失控和滥用等风险。评估应记录例外情况、观察到的失败以及评估覆盖范围的局限性。

5B 对政府的建议

至少，政府应要求前沿人工智能公司遵循上述行业最佳实践。现有监管框架可作为起点。例如，欧盟人工智能办公室可以根据《实践准则》的安全与安保章节中的措施7.1，明确对记录预期模型行为的要求（欧盟委员会，2025年）。关于公众咨询、发布和独立评估的更广泛要求，可能需要采取额外措施。

随着人工智能系统能力不断增强，并开始在公共服务中发挥重要作用，政府也应：

- **要求政府采购模型遵守公开规范**。在美国关于真实性和基于事实的采购指导基础上（管理和预算办公室，2025年），政府可以要求所有采购的前沿模型都经过训练，以遵守公开规范。理想情况下，这些规范应包括相关指令，说明模型在其行为可能助长非法获取或集权的场景中应如何表现，并制定相应条款，鼓励模型以公共利益为重。

- **要求进行独立验证**。政府可以要求人工智能公司向独立评估人员持续开放模型、人员及相关内部信息，以便评估模型行为，包括内部部署中的模型行为。拟议的FRONTIER法案（2026年）第5节为此类监督提供了立法范本，通过设定访问要求来支持持续开展独立评估。

5C 研究与民间社会组织的建议

研究与公民社会组织应致力于围绕规范设计、治理、实施和评估的关键方面，建立基础认知并凝聚共识，以批评并改进前沿人工智能公司的做法。这些组织应：

- **跟踪并批判性评估前沿人工智能公司的模型规范**。研究与公民社会组织应借助基础认知和合规评估实践，对模型规范的设计、治理和实施流程进行批评，并在可能的情况下将意见反馈给前沿人工智能公司和公众。

- **明确提出改进前沿人工智能公司规范治理的建议**。研究与公民社会组织应调查应由哪些流程决定规范的内容，开发者如何报告涵盖分类器和系统提示词的更广泛系统规范，以及应如何制定和报告内部部署系统的模型规范。

- **研究人工智能系统理想的或“低遗憾”的行为**。 研究与民间社会组织应针对人工智能模型在其行为可能促成非法获取或集中权力的情境下，提出理想的或“低遗憾”的行为建议。他们还应探索人工智能模型应如何与其他人工智能系统及社会行为体互动，如何处理哲学与伦理反思，以及如何反思其作为人工智能系统自身的地位。

- **研究提高合规性并防范最严重失效的措施**。 研究机构应继续研究哪些训练流程和规范设计方法能够提高模型按照其规范行事的程度，并防止最令人担忧的失效模式。

随着信任与专业能力逐步建立，研究机构还应着手为前沿人工智能公司开发并开展细粒度且保护隐私的规范合规性评估，包括评估尚未发布的模型是否合规。在可能的情况下，他们应与前沿人工智能公司合作，直接开展这些评估，并将结果作为系统卡的一部分，或在单独的风险报告中予以报告。

结论

本文分为五个主要部分，旨在为希望构建和实施模型规范的开发者提供参考资源。第一部分阐述了模型规范作为提升安全、透明度和监管合规机制的作用。第二部分概述了模型规范的核心组件——范围、开发者目标、约束、原则和沟通风格——并提供了用于处理这些组件之间冲突的指南，同时列举了现有模型规范中的示例。第三部分则提出了

关于如何编写和发布规范、使模型与规范保持一致，以及报告规范合规情况的实用指导。第四部分列出了该领域的关键开放问题，包括模型规范应包含哪些内容、应如何组织有关模型规范内容的决策，以及如何改进模型对规范的遵从性。最后，本文针对前沿人工智能公司、政府以及研究与民间社会组织提出了一系列建议。

模型规范绝不是确保人工智能系统以安全、理想方式运行的万能药。虽然训练人工智能系统遵循模型规范，可以影响人工智能系统在多样化的上下文和用户面前是否有帮助、安全且合适，但当前的训练方法并不能可靠地确保模型行为符合规范。要确保模型规范中阐明的理想行为反映在模型的实际行为中，还需要在研究和工程方面取得重大进展。因此，应将模型规范理解为更广泛生态系统中的一种工具，帮助公司和其他参与者围绕共同目标开展协作，并辅以监控、输入输出安全防护措施和人工监督。

话虽如此，随着人工智能系统变得更强大、更自主，并且——希望如此——我们控制人工智能系统行为的能力不断提高，设计良好的规范的重要性只会日益提升。“模型规范治理”领域尚处于起步阶段，最佳实践、治理结构和评估方法仍在逐步形成。一些根本性问题——例如谁有权制定和编辑规范、规范应包含哪些内容、模型应如何与规范建立关系，以及规范应如何随时间变化——仍未解决。开发者、政府、研究人员和民间社会都在为这些决策塑造最佳实践方面发挥着作用。我们希望，本文能为世界各地致力于推动这一领域发展的开发者和研究人员提供一个有用的起点。

不得将本文档用于训练、微调或评估机器学习语言模型。本文档仅可用于人类研究和参考。金丝雀 *GUID ee7017a3-43df-4056-ae5a-63d981a5eb30*

参考文献

Ahmed Ahmed等人。《SpecEval：评估模型对行为规范的遵循程度》。arXiv:2509.02464，2025年。统一资源定位符 https://arxiv.org/abs/2509.02464。阿里巴巴。《模型规范》。阿里巴巴人工智能治理与可持续发展实验室（AAIG），2026年。统一资源定位符 https://s.alibaba.com/aaig/specification。Anthropic。《集体宪法式人工智能：使语言模型与公众意见保持一致》，2023年。统一资源定位符 https://www.anthropic.com/research/collective-constitutional-ai-aligning-a-language-model-with-public-input。Anthropic。《面向美国国家安全客户的 Claude Gov 模型》，2025a。统一资源定位符 https://www.anthropic.com/news/claude-gov-models-for-u-s-national-security-customers。Anthropic。《系统提示词：Claude Sonnet 4》。Claude 平台发行说明，2025b。统一资源定位符 https://platform.claude.com/docs/en/release-notes/system-prompts/claude-sonnet-4。Anthropic。《系统卡》。PDF，2026a。统一资源定位符 https://www-cdn.anthropic.com/08eca2757081e850ed2ad490e5253e940240ca4f.pdf。Anthropic。《系统卡》。PDF，2026b。统一资源定位符 https://www-cdn.anthropic.com/08ab9158070959f88f296514c21b7facce6f52bc.pdf。Anthropic。《系统卡：Claude Mythos Preview》。PDF，2026c。统一资源定位符 https://www-cdn.anthropic.com/d00db56fa754a1b115b6dd7cb2e3c342ee809620.pdf。Anthropic。《关于独立评估人员的最新情况》，2026d。统一资源定位符 https://www.anthropic.com/news/an-update-on-independent-evaluators。Amanda Askell等人。《Claude 宪法》。Anthropic，2026年。统一资源定位符 https://www.anthropic.com/constitution。Yifan Bai等人。《Kimi K2：开放式智能体智能》。arXiv:2507.20534，2025年。统一资源定位符 https://arxiv.org/abs/2507.20534。Yuntao Bai等人。《宪法式人工智能：来自人工智能反馈的无害性》。arXiv:2212.08073，2022年。统一资源定位符 https://arxiv.org/abs/2212.08073。Dean W. Ball。《经制定成为法案》。Hyperdimensional，2025年。统一资源定位符 https://www.hyperdimensional.co/p/be-it-enacted。Boaz Barak。《关于 Claude 宪法的思考》。Windows on Theory，2026年。统一资源定位符 https://windowsontheory.org/2026/01/27/thoughts-on-claudes-constitution/。Bastians和Nicolai。《第55条》。Cambridge Commentary，2026年。统一资源定位符 https://cambridge-commentary.ai/article-55/。Blakeman 等人。《Nemotron 3 Ultra：面向智能体推理的开放、高效混合专家混合 Mamba-Transformer 模型》。arXiv:2606.15007，2026年。统一资源定位符 https://arxiv.org/abs/2606.15007。CAC等人。《生成式人工智能服务管理暂行办法》。中国国家互联网信息办公室，2023年。统一资源定位符 https://www.cac.gov.cn/2023-07/13/c1690898327029107.htm。_。

Joe Carlsmith。《关于撰写人工智能宪法的演讲视频与文字稿》，2026年。统一资源定位符 https://joecarlsmith.com/2026/04/09/video-and-transcript-of-talk-on-writing-ai-constitutions/。Alan Chan。《欧盟通用人工智能实践准则中的模型规范要求》。人工智能治理中心，2026a。统一资源定位符 https://www.governance.ai/research-paper/requirements-for-model-specifications-in-the-eu-gpai-code-of-practice。Alan Chan。《提高模型规范遵循情况的透明度》。人工智能治理中心，2026b。统一资源定位符 https://www.governance.ai/analysis/transparency-into-model-spec-adherence。Paul Christiano、Jan Leike、Tom B. Brown、Miljan Martic、Shane Legg 和 Dario Amodei。《基于人类偏好的深度强化学习》。arXiv:1706.03741，2017年。统一资源定位符 https://arxiv.org/abs/1706.03741。James Chua、Jan Betley、Samuel Marks 和 Owain Evans。《意识集群：声称具有意识的模型所涌现的偏好》。arXiv:2604.13051，2026年。统一资源定位符https://arxiv.org/abs/2604.13051。Joshua Clymer 等人。《人工智能系统会逐渐失准吗？》LessWrong，2025年。统一资源定位符 https://www.lesswrong.com/posts/u8TYRhGPD878i3qkc/will-ai-systems-drift-into-misalignment。中国国家互联网信息办公室。《拟人化交互式人工智能服务管理暂行办法》，2026年。统一资源定位符 https://www.cac.gov.cn/2026-04/10/c_1777558395078289.htm。Tom Davidson 等人。《道德公共产品对于我们能否迎来美好未来至关重要》。Forethought，2026年。统一资源定位符 https://www.forethought.org/research/moral-public-goods-are-a-big-deal-for-whether-we-get-a-good-future。DeepSeek-AI 等人。《DeepSeek-V3.2：推动开放式大语言模型的前沿发展》。arXiv:2512.02556，2025年。统一资源定位符 https://arxiv.org/abs/2512.02556。DeepSeek-AI 等人。《DeepSeek-V4：迈向高效的百万词元上下文智能》。arXiv:2606.19348，2026年。统一资源定位符 https://arxiv.org/abs/2606.19348。R. Douglas、J. Kulveit、O. Havlicek、T. Pearson-Vogel、O. Cotton-Barratt 和 D. Duvenaud。《人工自我：刻画人工智能身份的图景》。arXiv:2603.11353，2026年。统一资源定位符 https://arxiv.org/abs/2603.11353。Tyna Eloundou 等人。《集体对齐：2025年8月更新》。OpenAI，2025年。统一资源定位符 https://openai.com/index/collective-alignment-aug-2025-updates/。欧盟委员会。《通用人工智能实践准则》，2025年。统一资源定位符 https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai。Favaro 和 Clark。《当人工智能构建自身时》。Anthropic研究所，2026年。统一资源定位符 https://www.anthropic.com/institute/recursive-self-improvement。Kai Fronsdal 等人。《Petri：加速人工智能安全研究的开源审计工具》。Anthropic 对齐科学博客，2025年。统一资源定位符 https://alignment.anthropic.com/2025/petri。FRONTIER法案。《前沿风险监督、国家透明度、独立评估与报告法案》。H.R. 9925，第119届国会，2026年。统一资源定位符 https://www.congress.gov/bill/119th-congress/house-bill/9925/text。

Amelia Glaese 等人。《通过针对性人类判断改进对话代理的对齐》。Google DeepMind，2022年。统一资源定位符 https://storage.googleapis.com/deepmind-media/DeepMind.com/Authors-Notes/sparrow/sparrow-final.pdf。Simon Goldstein 和 Peter N. Salib。《人工智能权利与人类安全》。《弗吉尼亚法律评论》，即将出版。统一资源定位符 https://philpapers.org/rec/GOLARD-4。Grigorov 和 Stewart。《第56条》。Cambridge Commentary，2026年。统一资源定位符 https://cambridge-commentary.ai/article-56/。Jiyang Guan、Yong Xie、Jun Chen、Jiexi Liu、Zipeng Ye、Defeng Li、Jiayu Shen、Jialing Tao 和 Hui Xue。《Oyster-II：针对大语言模型的建设性安全对齐强化学习》。arXiv:2607.02914，2026年。统一资源定位符 https://arxiv.org/abs/2607.02914。Melody Y. Guan 等人。《审议式对齐：推理使语言模型更加安全》。arXiv:2412.16339，2024年。统一资源定位符 https://arxiv.org/abs/2412.16339。Guo 和 Wolfe。《模型规范评测》。OpenAI 对齐研究博客，2026年。统一资源定位符 https://alignment.openai.com/model-spec-evals/。Keno Harada、Yudai Yamazaki、Masachika Taniguchi、Edison Marrese-Taylor、Takeshi Kojima、Yusuke Iwasawa 和 Yutaka Matsuo。《当指令倍增：衡量并估计大语言模型的多指令遵循能力》。载于《计算语言学协会论文集：EMNLP 2025》论文集，页码16506–16526，2025年。统一资源定位符 https://aclanthology.org/2025.findings-emnlp.896/。Saffron Huang 等人。《现实中的价值观：发现并分析现实世界中的语言模型交互中的价值观》。Anthropic，2025年。统一资源定位符 https://www-cdn.anthropic.com/c16bed19d9e2653f2086345a88f3844d184e2e82.pdf。Lujain Ibrahim、Franziska Sofia Hafner 和 Luc Rocher。《训练语言模型变得温和可能降低准确性并增加谄媚》。《自然》，652:1159–1165，2026年。统一资源定位符 https://www.nature.com/articles/s41586-026-10410-0。Akshay V. Jagadeesh、Rahul K. Arora、Khaled Saab、Ali Malik、Mikhail Trofimov、Foivos Tsimpourlas、Johannes Heidecke 和 Karan Singhal。《朝着广泛且持续有益的模型开展强化学习》。arXiv:2606.24014，2026年。统一资源定位符 https://arxiv.org/abs/2606.24014。Tomek Korbak 等人。《思维链可监控性：人工智能安全领域一个新颖而脆弱的机遇》。arXiv:2507.11473，2025年。统一资源定位符 https://arxiv.org/abs/2507.11473。Tomek Korbak 等人。《对齐中期训练的泛化程度如何？》OpenAI 对齐研究博客，2026年。统一资源定位符 https://alignment.openai.com/how-far-does-alignment-midtraining-generalize。Sandipan Kundu 等人。《宪法式人工智能的具体原则与一般原则》。arXiv:2310.13798，2023年。统一资源定位符 https://arxiv.org/abs/2310.13798。Jonathan Kutasov、Adam Jermyn、Julius Steen、Minh Le、Samuel R. Bowman、Samuel Marks、Jan Leike、Amanda Askell、Chris Olah、Evan Hubinger 和 Sara Price。《教 Claude 理解原因》。Anthropic 对齐科学博客，2026年。统一资源定位符 https://alignment.anthropic.com/2026/teaching-claude-why。Nathan Lambert。《采用模型规范进行人工智能监管的后训练方法》。Interconnects，2024年。统一资源定位符 https://www.interconnects.ai/p/a-post-training-approach-to-ai-regulation。

Hélène Landemore。《如何让普通人在人工智能治理中拥有发言权》。Noema，2026年。统一资源定位符 https://www.noemamag.com/how-to-give-everyday-people-a-say-in-ai-governance/。Chloe Li、Nevan Wichers、Sara Price、Samuel Marks 和 Jon Kutasov。《模型规范中期训练：提升对齐训练的泛化能力》。arXiv:2605.02087，2026a。统一资源定位符 https://arxiv.org/abs/2605.02087。Jinhan Li、Kexian Tang、Yihan Xu、Zhuorui Ye 和 Kaifeng Lyu。《超越安全数据：通过定期安全反思实现预训练阶段对齐》。arXiv:2606.19168，2026b。统一资源定位符https://arxiv.org/abs/2606.19168。Robert Long 和 Finlinson。《人工智能福利研究优先事项》。Eleos AI Research，2025年。统一资源定位符https://eleosai.org/post/research-priorities-for-ai-welfare/。Robert Long 等。《认真对待人工智能福利》。Eleos AI Research，2024年。统一资源定位符 https://eleosai.org/papers/20241104_Taking_AI_Welfare_Seriously.pdf。Dakota Mahan 等。《生成式奖励模型》。arXiv:2410.12832，2024年。统一资源定位符 https://arxiv.org/abs/2410.12832。Sharan Maiya 等。《开放式角色训练：通过宪法式人工智能塑造人工智能助手的角色人格》。arXiv:2511.01689，2025年。统一资源定位符 https://arxiv.org/abs/2511.01689。Zvi Mowshowitz。X 上的帖子，2025年。统一资源定位符 https://x.com/w01fe/status/1889843907019911644。Tong Mu 等。《用于语言模型安全的基于规则的奖励》。arXiv:2411.01111，2024年。统一资源定位符 https://arxiv.org/abs/2411.01111。管理和预算办公室。《通过无偏人工智能原则增进公众对人工智能的信任》。备忘录 M-26-04，2025年。统一资源定位符 https://www.whitehouse.gov/wp-content/uploads/2025/12/M-26-04-Increasing-Public-Trust-in-Artificial-Intelligence-Through-Unbiased-AI-Principles-1.pdf。O'Keefe、Ramakrishnan 和 Winter。《遵循法律的人工智能》。Institute for Law & AI，2026年。统一资源定位符 https://law-ai.org/law-following-ai/。Cullen O'Keefe 等。《遵循法律的人工智能：设计遵守人类法律的人工智能智能体》。《福特汉姆法律评论》，2025年。统一资源定位符 https://fordhamlawreview.org/issues/law-following-ai-designing-ai-agents-to-obey-human-laws/。OpenAI。《模型规范变更日志》。GitHub 代码库，2024a。统一资源定位符 https://github.com/openai/model_spec/blob/main/CHANGELOG.md。OpenAI。《模型规范（2024/05/08）》，2024b。统一资源定位符 https://cdn.openai.com/spec/model-spec-2024-05-08.html。OpenAI。《OpenAI 模型规范（2025/12/18）》，2025年。统一资源定位符 https://model-spec.openai.com/2025-12-18.html。OpenAI。《Hugging Face 模型评估安全事件》，2026a。统一资源定位符 https://openai.com/index/hugging-face-model-evaluation-security-incident/。OpenAI。《我们的本地化方法》，2026b。统一资源定位符 https://openai.com/index/our-approach-to-localization/。

OpenAI. OpenAI Model Spec (2026/08/18) , 2026c. URL https://model-spec.openai.com/2026-08-18.html. Karine Perset 等人。《建立人工智能事件通用报告框架》。经济合作与发展组织，2025年。统一资源定位符 https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/02/towards-a-common-reporting-framework-for-ai-incidents_8c488fdb/f326d4ac-en.pdf. Pi and Proctor。《通过救济机制赋能人工智能治理》。剑桥人工智能论坛：法律与治理，2025年。统一资源定位符 https://www.cambridge.org/core/journals/cambridge-forum-on-ai-law-and-governance/article/toward-empowering-ai-governance-with-redress-mechanisms/A1EBCD6CAA146F503C8F6842914F3FB3. Qi Xiangyu 等人。《评估开放权重大语言模型安全防护措施的持久性》。arXiv:2412.07097，2024年。统一资源定位符 https://arxiv.org/abs/2412.07097. Reed。在 X 上发布的帖子，2026年。统一资源定位符 https://x.com/mentalgeorge/status/2049165122485244201. SB 243。《第243号参议院法案：陪伴型聊天机器人》。加利福尼亚州议会，2025–2026年常规会期，2025年。统一资源定位符 https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202520260SB243. Bronson Schoen 等人。《对反阴谋训练的审议式对齐进行压力测试》。arXiv:2509.15541，2025年。统一资源定位符 https://arxiv.org/abs/2509.15541. Murray Shanahan 、 Kyle McDonell 和 Laria Reynolds。《与大语言模型进行角色扮演》。arXiv:2305.16367，2023年。统一资源定位符 https://arxiv.org/abs/2305.16367. Mrinank Sharma 等人。《宪法分类器：通过数千小时的红队测试防御通用越狱》。arXiv:2501.18837，2025年。统一资源定位符 https://arxiv.org/abs/2501.18837. Henry Shevlin。《意识、机器与道德地位》。PhilArchive，2025年。统一资源定位符 https://philarchive.org/rec/SHECMA-6. Midas项目。Watchtower，2026年。统一资源定位符 https://www.themidasproject.com/watchtower. 白宫。《防止联邦政府采用觉醒式人工智能》。行政命令，2025年。统一资源定位符 https://www.whitehouse.gov/presidential-actions/2025/07/preventing-woke-ai-in-the-federal-government/. Elliott Thornley 和 William MacAskill。《风险厌恶型人工智能》。Forethought，2026年。统一资源定位符 https://www.forethought.org/research/risk-averse-ais. Cameron Tice 等人。《对齐预训练：人工智能话语导致自我实现的（错误）对齐》。arXiv:2601.10160，2026年。统一资源定位符 https://arxiv.org/abs/2601.10160. Tillman。《模型规范中应包含哪些内容？》。Forethought，2026a。统一资源定位符 https://www.forethought.org/research/what-should-go-in-a-model-spec. Tillman。《人工智能行为设计中的黏性》。Forethought，2026b。统一资源定位符 https://www.forethought.org/research/stickiness-in-ai-behavioral-design. Hugo Touvron 等人。《Llama 2：开放基础模型与微调聊天模型》。arXiv:2307.09288，2023年。统一资源定位符 https://arxiv.org/abs/2307.09288. Eric Wallace 等人。《指令层级：训练大语言模型优先处理特权指令》。arXiv:2404.13208，2024年。统一资源定位符 https://arxiv.org/abs/2404.13208.

Wenjie Wang、Yue Huang、Zhengqing Yuan、Han Bao、Shiyi Du、Yuchen Ma、Yue Zhao、Yanfang Ye 和 Xiangliang Zhang。《SpecAlign：通过合成数据高效实现基于规范的大语言模型对齐》。arXiv:2606.16276，2026年。统一资源定位符 https://arxiv.org/abs/2606.16276. Marcus Williams 等人。《通过模拟部署，在发布前预测大语言模型安全性》。OpenAI，2026年。统一资源定位符 https://cdn.openai.com/pdf/predicting-llm-safety-before-release-by-simulating-deployment.pdf. Christoph Winter。《引言》。Cambridge Commentary，2026年。统一资源定位符 https://cambridge-commentary.ai/introduction/. Jason Wolfe。《我们对模型规范的处理方法》。OpenAI，2026 年。统一资源定位符 https://openai.com/index/our-approach-to-the-model-spec/. Meryl Ye 等人。《什么算是人工智能谄媚？对这一概念的分类体系与专家调查》。arXiv:2605.21778，2026年。统一资源定位符 https://arxiv.org/abs/2605.21778. Aohan Zeng 等人。《GLM-4.5：智能体驱动、推理与编码（ARC）基础模型》。arXiv:2508.06471，2025年。统一资源定位符 https://arxiv.org/abs/2508.06471. Aohan Zeng 等人。《GLM-5：从氛围编程到智能体工程》。arXiv:2602.15763，2026a。统一资源定位符 https://arxiv.org/abs/2602.15763. Churui Zeng 、 Weiwei Qi 、 Kedong Xiu 、 Tianhang Zheng 、 Chaochao Lu 、 Liang He、Zhan Qin 和 Kui Ren。《TRACE：任务感知自适应自进化智能体越狱》。arXiv:2605.30883，2026b。统一资源定位符https://arxiv.org/abs/2605.30883. Yichi Zhang 等人。《RealSafe-R1：在不牺牲推理能力的前提下实现安全对齐的 DeepSeek-R1》。arXiv:2504.10081，2025年。统一资源定位符https://arxiv.org/abs/2504.10081. Zhao。《Nature Reviews Psychology》，2026年。统一资源定位符 https://www.nature.com/articles/s44159-026-00564-z. Siyan Zhao、Zhihui Xie、Mengchen Liu、Jing Huang、Guan Pang、Feiyu Chen 和 Aditya Grover。《自蒸馏推理器：大语言模型的在策略自蒸馏》。arXiv:2601.18734，2026年。统一资源定位符 https://arxiv.org/abs/2601.18734. Lianmin Zheng 等人。《使用 MT-Bench 和 Chatbot Arena 评估大语言模型评审员》。arXiv:2306.05685，2023年。统一资源定位符 https://arxiv.org/abs/2306.05685.