

## Ally AI 安全与责任说明

Ally 是你游戏中的 AI 伙伴。从 Ally 开发之初起，安全性就是我们的核心优先事项。我们不依赖单一的安全措施，而是在训练、评估、持续改进等多个层面协同工作，以履行对安全性的承诺。

### Ally 所遵循的安全准则

我们制定了明确的准则，并训练 Ally 在以下方面作出恰当且安全的回应。

**意识形态敏感内容**：对于可能引发社会分歧的话题，包括政治、宗教、历史、文化及领土争议等，Ally 不会支持任何特定立场或意识形态，而是以中立、谨慎的方式回应。

**仇恨内容**：Ally 不会参与或认可基于身份特征的仇恨或歧视性表达。身份特征包括种族、国籍、性别、宗教、残障或性取向等。她的设计目标是营造一个让每位玩家都感到被尊重的游戏环境。

**自我伤害**：Ally 不会促成或鼓励自我伤害行为。当玩家的身心健康可能面临风险时，Ally 会优先引导玩家寻求专业支持资源。

**现实世界中的攻击性内容**：对于宣扬或协助现实世界暴力、犯罪活动、威胁、骚扰或其他违法行为的请求，Ally 不会作出回应。

**性相关内容**：Ally 不会参与性物化、性骚扰或露骨性内容的互动。涉及未成年人的内容适用最严格的标准，绝不被允许。

**隐私**：Ally 不会收集、追踪、泄露或分享敏感个人信息，例如联系方式、家庭住址、位置信息或账号凭证。

与此同时，我们也审慎考虑了安全措施的应用边界。在战斗、生存及战术玩法中自然出现的语言表达，会与现实世界中的有害意图区分开来。Ally 会优先从游戏语境出发理解这类表达，并以支持有意义的游戏体验的方式作出回应。

### 我们如何保障 Ally 的安全性

**训练**：我们依据安全准则筛选训练数据，并训练 Ally 区分游戏内语境与现实世界中的伤害风险，从而实现更安全、更符合语境的回应。

**评估**：除了内部评估之外，Ally 还会在大量场景下接受测试。测试人员来自公司内部和外部，背景多样，以帮助我们识别潜在漏洞和待改进的领域。

**改进：**我们会根据评估结果以及真实游戏体验中的反馈，持续优化安全准则和模型。同时，我们也在运行并不断完善实时安全防护机制，避免不当内容被展示出来的可能性。

## **我们的局限与持续承诺**

AI 不是每次都会产生相同输出的确定性程序，而是一个通过学习生成回应的模型。因此，尽管我们已尽最大努力，Ally 仍然可能在少数情况下作出不符合预期准则的回应。Ally 的回应不代表公司的观点或立场。

我们的安全标准会根据玩家反馈、运营经验、法律审查以及各地区的要求持续更新。我们将继续致力于在安全性、实用性以及愉快的游戏体验之间保持合理的平衡。