

免费资源

为什么无害的 AI 提示词会被拒绝：内容审核生存指南

我用来区分「描述不清」与「真正的内容限制」的工作表。

Jeff Zambrano.

AI 电影导演 · jeffzambrano.com

适用读者

写给正在复查一个被拒请求的电影人，而这个请求原本只想呈现无害的动作。

解决的问题

一次拒绝提供的信息太少，无法判断是文字、参考图还是生成结果导致了问题。

包含内容

- 拒绝记录
- 证据分类
- 中性措辞示例
- 参考图审查
- 输出核查

目录

01 先诊断拒绝，再改写镜头

02 从事件记录开始

03 判断提示信息提出了什么问题

04 使用证据标签，而不是通用黑名单

05 澄清无害的描述

06 审查所附素材

07 用结果结束记录

01 先诊断拒绝，再改写镜头

无害的意图可能变成一段含糊的描述，但一次拒绝并不能证明是某个特定的词造成的。我用这张表把观察到的事实和自己的猜测分开，让下一个决定来自镜头本身和它的证据，而不是一份越积越长的禁用词清单。

在遭到拒绝之后使用它，或者在提交一段措辞与预期动作不够吻合的描述之前使用它。下面的示例用来澄清无害的素材和行为；它们不是用来掩盖受限请求的替换词。

02 从事件记录开始

在改动任何内容之前，我会先完整保留原始请求。没有这份记录，后来某次成功的尝试几乎无法告诉我究竟改变了什么。

字段	在修改之前记录
平台与模型	使用的服务以及所选的模型。
完整请求	提交的全部文字。
参考素材	随文字一同提供的图像。
提示信息	拒绝或报错的原文信息。
阶段	生成之前、处理过程中，还是输出出现之后。
结果	被拒绝、输出完成，还是输出缺少所请求的内容。
积分结果	交易记录实际显示的情况。
疑似原因	一个假设，与观察到的结果明确分开。
修改后的措辞	具体改动了哪段描述，以及它现在的含义。

- [ ] 我保留了被拒绝的原始措辞，而没有在记录中把它替换掉。
- [ ] 我复制了提示信息原文，而不是凭记忆转述。
- [ ] 我把参考图作为请求的一部分记录了下来。
- [ ] 我没有把疑似的触发因素标记为已确认的原因。

03 判断提示信息提出了什么问题

我的制作笔记会区分内容类提示和版权类提示。它们是有用的起点，但提示信息并不一定能指出是哪个词或哪张参考图出了问题。

观察到的提示或结果	我检查的内容
`content flagged`	所呈现的动作、措辞以及参考图的内容。
`Rejected due to copyright restrictions`	具名作品、可辨认的人物、标志和产品参考。
一段完成的片段遗漏了某个元素	先检查输出本身，再去假设指令只是被误解了。
以前被接受的措辞现在失败了	平台、模型、参考图和完整请求，不假设一次被接受就永远有效。

在复查中，我还会把文字、参考图和生成画面分开来看。改写一句话，并不能证明所附的图像没有问题；一次完成的生成，也不能证明它的内容与请求相符。

04 使用证据标签，而不是通用黑名单

支撑我这套工作词汇的研究资料里，有政策摘要、被报告的事件，也有后来与某些早期警告相矛盾的观察。我保留这些区分，因为一项预防措施与一个已被证实的触发因素是两回事。

标签	在我的记录中意味着什么
已观察到的拒绝	这个完整请求在这个有记录的情境中失败了。
他处报告	有报告提示可能存在问题，但不能证明这里会出现同样的行为。
无证据的预防	我更倾向于更清晰的描述，但无法证明原来的词导致了拒绝。
已完成的示例	这个词出现在一个已完成的请求中；这反驳了绝对禁用的说法，但不能保证今后一定会被接受。
质量问题	描述本身薄弱或具有误导性，但没有内容审核问题的证据。
电影节要求	这项限制属于某个比赛的规则，与模型的内容审核是两回事。

我的笔记中有一些已完成的示例用到了 `wet`、`sweat` 和 `flesh` 这类词，尽管早先有过针对它们的警告。因此，我不会把这些词变成自动禁用词，也不会把一份词汇清单当作平台规范来呈现。

05 澄清无害的描述

1. 品牌无关紧要时，描述材质

当镜头需要的是一种表面，而不是一个可辨认的商业标识时，我会使用材质描述。工作词汇中提供了以下短语；每一个都只是简短的描述，而不是完整的场景。

```
laminate surface
polystyrene foam
vinyl upholstery
drywall
borosilicate glass
```

原始笔记在这一语境下讨论过 Formica、Styrofoam、Naugahyde、Sheetrock 和 Pyrex。这里并不把它们出现在请求中当作某种拒绝机制的证据：对应的条目都被标记为无证据的预防。

2. 按物体本身来命名

对于一件普通道具，我会检查商品名称是否增加了镜头真正需要的可见信息。如果没有，记录中的通用描述反而能更准确地说明这个物体。

```
facial tissue
plastic storage container
cotton swab
sticky note
permanent marker
```

这些示例来自同一套工作词汇。它们并不能证明所有品牌词都被禁止，也不能证明改个名字就能解决参考图中某个可辨认设计的问题。

3. 把天气和体力消耗写具体

我会保留湿润的可见成因，而不是仅仅因为某个有用的词出现在警告清单上就把它换掉。在原始描述含糊的地方，笔记提供了更具体的措辞。

```
rain-soaked
slick with rain
skin shining with effort
```

使用能描述实际画面的那个短语。雨的描述属于雨，体力消耗则属于一段能明显支撑它的表演；修改应当让场景更清楚，而不是捏造另一种解释。

4. 描述实际正在发生的无害动作

一个夸张的动词，可能让普通的行为听起来像是受伤。当预期的动作是睡着或有意识地改变姿势时，我的笔记更倾向于写动作本身及其可见的迹象。

```
falls asleep where she sits, slack and breathing
lowers herself onto the floor
perfectly still and breathing
```

只有当这些片段忠实于预期动作时，它们才是合适的。我不会用它们把伤害重新标记为睡眠，也不会用它们来掩盖一个场景所呈现的内容。

5. 以正面方式陈述可见状态

否定式指令可能引入不想要的概念，却没有描述出我需要的画面。我会用实际的姿势、服装或身体状态来取代这种不确定性，同时保持角色和动作不变。

```
holding its own shape and standing steady
```

这是一条关于可见稳定性的指令。它不能证明每一套内容审核系统都以同样的方式理解否定句，也无法解决一张呈现其他内容的参考图的问题。

06 审查所附素材

- [ ] 参考图呈现的是与文字相同的无害动作或主体。
- [ ] 我检查了可辨认的人脸，而不是只依赖一个虚构的标签。
- [ ] 我核查了可见的标志和可辨认的产品设计。
- [ ] 任何身份或授权问题都按其本身的要求解决，而不是靠措辞来掩饰。
- [ ] 角色声明的年龄与可见外貌相符。
- [ ] 我把任何电影节特有的年龄限制，与针对所有模型的结论分开看待。

我把参考图视为请求的一部分，而不是中性的装饰。一句更干净的话，并不会改变提交素材的其余部分，所以复查必须包括图像实际呈现的内容。

07 用结果结束记录

- [ ] 我看过完成的输出，并确认所请求的无害动作仍然存在。
- [ ] 我把缺失或被改动的元素，与被拒绝的请求分开记录。
- [ ] 我核查了实际的积分结果，而不是想当然地假设有退款政策。
- [ ] 我把一次被接受的修改当作一条观察，而不是可重复的承诺。

- [ ] 任何尚未查明的原因，都继续标记为未查明。

我想要的是更清晰的描述和一份值得信赖的记录，而不是围绕最后改动的那个词编织出来的迷信。当证据到此为止时，我会让原因保持开放，并让下一个决定忠实于预期的镜头。

JEFF ZAMBRANO. · AI 电影导演

一起完成你的电影

我用 AI 执导电影、剧集和品牌作品，从剧本到最终剪辑。
告诉我你的想法，我会亲自回复。



委托制作电影、剧集或预告片
请先说明影片形式、预计时长和交付日期，再分享已有的故事或素材。

资助 Borrowed Wounds 长片
请告诉我你考虑投资、联合制作还是赞助这部 90 分钟的电影，我会把逐镜头核算的制作计划发给你。

品牌影片
请介绍品牌与影片的目的，并在已确定的情况下，说明目标观众及发布日期。

AI 平台或赞助
请介绍你的工具或模型，再简述你设想的影片或赞助合作。

电影节、媒体或发行
请介绍你的活动、媒体或发行机构，并注明日期，以及提议所涉及的影片或素材。

hello@jeffzambrano.es
WhatsApp +34 688 351 935
jeffzambrano.com