你的问题

我为了让模型按格式输出试了很多次都失败了，真的很烦。为什么会这样？

（上下文：你给了一个“研究级提示词生成器”的超长提示词，要求严格按 3 段式 Markdown 输出，并附带多条“禁止/必须/不得”。）

为什么会这样（最常见的 8 个原因）

1. **约束太多且互相打架**

2. 你要求“冷漠客观、只陈述现象、禁止分析”，但又要求“核心现状或结论”——很多输入不可能在不推理/不补知识的情况下得出“结论”。

3. 你要求“不要解答用户问题”，但第 1 部分本质上就是“对 Q 给出结论式回答”。模型会在冲突里“挑一个更像任务完成的方向”。

4. **输出格式对 token 预测很脆弱** 模型是逐 token 生成的：只要中途被“插一句解释/过渡语”，后续就容易滑到聊天体。

5. **超长提示词导致“注意力稀释”** 规则越多、越长，模型越可能漏掉某几条，尤其是靠后或细粒度的格式要求（比如“不得包含开场白”）。

6. **系统/开发者提示词优先级更高** 即使你写“严格遵守”，也可能被更高优先级的系统指令、产品安全策略、工具可用性限制覆盖（例如：你要求“强制联网”，但运行环境未必能联网）。

7. **任务不完备：对任意 Q 的“事实快照”不可判定** 你说“接收任何输入”，但如果 Q 是：

8. 过于主观（“我该不该退学”）

9. 需要外部数据（“今天特斯拉股价多少”）

10. 语义不完整（“这个怎么弄”） 模型要么编、要么解释、要么改格式——三者都违背你的约束。

11. **例子带来“示例依赖”与分布偏移** 你只给了“咖啡涨价”的示例。模型对其它领域输入时会迁移失败，尤其是第 1 部分要压到 <50 字。

12. **温度/采样导致波动** 采样越随机（温度高、top_p 大），越容易出现“多说两句”或段落格式跑偏。

13. **Markdown 规则本身容易被“自动美化/补全”** 例如模型会自动加“好的，以下是…”，因为这是它训练中最常见的开场模式。

你这份提示词的“关键结构性问题”（点名）

- **“禁止解答” vs “事实快照给结论”**：冲突。
- **“事实快照 <50 字 + 只能陈述现象”**：对很多 Q 不可行。
- **“强制联网”**：工具不可用时必然失败（它要么撒谎说联网了，要么偏离格式解释原因）。
- **禁用常见开场白**：越禁越容易触发（模型常常用开场白来“稳定语气”）。

怎么让它更稳（可执行清单，按优先级）

A. 把“规则”变成“可验证的硬 schema”

1. **让输出只有一个允许的形状**：
2. 要么固定 3 个标题，禁止任何额外文字；
3. 要么直接要求 **JSON**（比 Markdown 稳得多）。
4. **加一个“失败兜底”**： 当无法给出事实快照时，允许输出：
5. “信息不足：需要 X（时间/地点/对象）” 否则模型只能硬编或解释。

B. 缩短提示词，把“关键指令”放到最后一段

把长说明缩成 10 行以内，让模型在临近输出时看到最重要的格式约束。

C. 用“自检-重写”两段式（最有效）

让模型先生成，再检查是否满足格式；不满足就重写一次。

D. 降低随机性

- 温度（temperature）设低（如 0–0.3）
- 或使用“确定性”模式（如果产品支持）

给你一个更稳的「v1 改造版」模板（直接复制用）

目标：尽量保留你想要的三段式，但把冲突解决、加兜底、加自检。

```
你是“研究级提示词生成器”。你的唯一输出必须严格符合下列模板（除模板外禁止输出任何字符）。

【输入】用户的原始议题 Q：<<<{Q}>>>
```

```
【输出模板】（必须逐字匹配标题；每段内容不超过 120 字）
## 1. [事实快照]
- 若 Q 可直接陈述：用一句话给出客观现状/结论。
- 若无法在不编造的情况下给结论：输出“信息不足：需要补充 {缺失要素}”。

## 2. [ChatGPT 联网搜索指令]
写给“可联网的 ChatGPT”的检索提示词，包含：
- 先搜索（列出 3–5 个关键词/查询式）
- 交叉验证（至少 3 类来源：官方/学术/媒体）
- 输出结构：时间戳 + 证据链接 + 结论

## 3. [Gemini 深度挖掘指令]
写给 Gemini 的深度分析提示词，包含：
- 研究计划（领域/假设/证据需求）
- 冲突点对比表（来源A vs 来源B）
- 机制解释（可检验的因果链）

【自检】输出后立刻检查：
- 是否只有 3 个标题
- 是否没有任何开场白/结束语
- 是否每段不超字数
若不满足，重写一次并只输出最终版本。
```

你现在这题（“为什么我总让模型按格式输出失败”）该怎么塞进三段式？

- **事实快照**：可以写成“长提示词+冲突约束+不可判定输入 → 模型经常漏规则并回到聊天模式”。
- **联网搜索指令**：检索“instruction hierarchy / prompt compliance / structured output reliability / JSON schema”等。
- **深度挖掘指令**：分析“为什么 token 级生成对硬格式不稳定、哪些约束组合会互相破坏、如何用自检与约束分层”。

最小调试流程（5 分钟就能跑）

1. 先用 **最短版本**：只保留 3 个标题 + “禁止额外输出”。
2. 失败一次就加一条：**自检并重写一次**。
3. 仍失败再加：**兜底分支（信息不足）**。
4. 再失败：改用 **JSON 输出**。
5. 最后才考虑把复杂规则加回去（一次只加一条，做回归测试）。