用户问题（原始议题 Q）

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？

---

1. [事实快照]

多次要求固定格式输出，模型仍会漏段、改标题或插入多余话语。

2. [ChatGPT 联网搜索指令]

你将使用联网搜索，回答：**为什么大语言模型经常无法严格按指定 Markdown/JSON 模板输出？**

检索与产出要求： 1) 优先查找：官方文档/研究论文/工程实践总结（OpenAI、Anthropic、Google、学术论文、知名工程博客）。 2) 覆盖“格式失控”的主要原因，并按**机制层级**组织： - 指令层级与系统/开发者提示词优先级冲突 - 采样随机性（temperature/top_p）与“逐 token 生成”导致的漂移 - 上下文窗口、截断、长提示词注意力稀释 - 训练目标（RLHF/偏好优化）与“礼貌/解释倾向”对严格格式的干扰 - Markdown/代码块边界、转义符、引号/括号嵌套导致的结构破坏 - 多任务混杂（既要解释又要严格格式）引发目标冲突 - 工具/平台包装（UI 自动加前后缀、系统注入、富文本渲染差异） 3) 给出**可操作的工程化改进清单**（每条都写“为什么有效 + 如何做”）： - 最小化输出通道：只允许一个 code fence、固定起止标记 - 用“硬约束”替代“建议”：必须/不得/否则返回 ERROR - few-shot 示例（1–2 个）+ 反例（1 个） - 让模型先生成“结构骨架”，再填充内容（两步法） - 自检与纠错：让模型在最后输出“校验结果”，不通过则重写 - 外部校验：JSON schema / 正则 / 解析器 + 失败重试策略 - 选择更适合的接口：函数调用/结构化输出/response_format（若平台支持） 4) 以结构化方式呈现： - A. 关键结论（5–8 条） - B. 失控原因的分类表（原因 → 典型表现 → 触发条件 → 证据来源） - C. 提示词设计模板（可直接复制） - D. 最小实验：如何复现与衡量“格式遵循率”（指标与记录方式）

3. [Gemini 深度挖掘指令]

在你已经收集到的资料基础上，做更深入的分析与整合：

任务： 1) 先建立一个**“格式遵循失败”的故障树**：从根因到可观察症状，给出 2–3 层分解。 2) 结合生成式模型的工作方式，解释为何“写得更长”不一定更听话： - 注意力分配/目标竞争/长上下文稀释 - 语言模型的概率生成与局部最优 - 训练偏好导致的“补充解释冲动” 3) 输出一个**稳健提示词方案 v1**（针对 Markdown 三段式），并附带： - 关键设计点（不超过 8 条） - 你预测它能解决哪些失败模式、不能解决哪些 4) 输出一个 **稳健提示词方案 v2**（更工程化），要求： - 使用明确的开始/结束标记（例如 <BEGIN>…<END>） - 引入“校验步骤”：若格式不合格，必须输出单行 `FORMAT_ERROR` 并重写 - 给出 1 个 few-shot 示例（严格符合格式） 5) 给出一个**评测与迭代计划**： - 指标：格式遵循率、段落缺失率、标题漂移率、额外文本率 - 变量控制：temperature/top_p、不同模型、不同长度输入 - 记录模板：一张可复制的表格（含字段）

输出要求： - 允许用对照表总结分歧与建议。 - 不要泛泛而谈，尽量把每条建议落到可执行的 prompt 片段或规则。