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]

格式约束是“软偏好”，受冲突指令、长上下文与采样噪声影响易漂移。

2. [ChatGPT 联网搜索指令]

你是“LLM 指令遵循与结构化输出”方向的研究助理。请**强制联网搜索**并交叉验证，回答： “为什么大语言模型在用户给出严格 Markdown/JSON 模板时仍会不按格式输出？哪些机制与工程条件会导致格式漂移？如何提升格式遵循率？”

要求： 1) **先搜索**并阅读：instruction following / structured output / JSON schema / function calling / constrained decoding / grammar decoding / RLHF / preference modeling / prompt injection / context conflict 相关资料。 2) **请交叉验证**：至少对照 6 个独立来源（官方文档、学术论文、工程博客、评测报告各≥1）。 3) **输出结构化报告**（必须包含时间戳与来源链接）： - A. 现象定义：什么叫“格式遵循失败/漂移”，给出常见失败模式清单（漏标题、顺序错、夹杂寒暄、超出字数、混入解释等）。 - B. 归因框架：从“训练/对齐”“解码/采样”“上下文与指令冲突”“提示词设计”“工具/平台限制（token、系统提示、guardrails）”五类展开。 - C. 可复现实验：给出 3 个最小复现提示词（MRE），每个说明变量、预期失败点、评估指标。 - D. 工程对策：按有效性排序给出 10 条提升格式遵循的做法（例如：分隔符、显式 schema、函数调用、逐步生成、先验校验/重写、低温度、二次修复等），并说明适用条件与副作用。 - E. 证据表：以表格列出“主张—证据—来源—可信度—适用范围”。 4) 所有关键结论必须标注出处；对不一致观点必须标注“冲突点”并解释差异原因。

3. [Gemini 深度挖掘指令]

你是“LLM 对齐机制 + 解码控制 + 评测方法”综合研究员。围绕现象： “模型明明收到严格三段式/Markdown 模板，却仍输出越界或漏结构”，请做**深度机制分析**与**研究计划**，并产出可执行改进方案。

请执行： 1) **研究计划**： - 列出需要的知识模块：对齐训练（SFT/RLHF/DPO）、指令层级冲突（system/dev/user）、长上下文注意力与位置偏置、解码（temperature/top-p）、受约束解码（CFG/grammar）、工具调用与结构化输出 API。 - 给出 2 周小型研究路线：每天目标、产物、评价方式。 2) **机制深挖**（给出因果图/流程图文字版）： - 为什么“长提示词”反而可能降低格式遵循？ - 为什么“示例 + 强制要求”仍会被寒暄/解释污染？ - 训练偏好与解码策略如何共同导致“最后一刻偏离”（end-of-output drift）？ 3) **分歧点分析**： - 学术界/工业界在“prompt 设计 vs 受约束解码/函数调用”哪条更可靠？ - 列出至少 3 组相互冲突观点，并给出各自证据与适用边界。 4) **证据对比表**： - 输出一个表：失败模式 × 可能原因 × 可观察信号 × 干预手段 × 代价/风险。 5) **产出可用模板**： - 给出 3 套“更鲁棒的格式控制提示词模板”（短/中/长），并包含： a) 明确的输出 schema；b) 自检清单；c) 失败时的自动修复指令（rewrite/validator-loop）。 - 同时给出一套评测协议：指标（格式准确率、内容准确率、漂移率）、测试集构造、统计方法。