用户问题： 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？

---

1. [事实快照]

LLM 常因指令冲突、采样随机性与上下文噪声而偏离指定格式输出。

2. [ChatGPT 联网搜索指令]

你是“LLM 指令遵循与结构化输出可靠性”研究助理。 **请强制联网搜索**，调查“为什么大模型经常无法按指定格式输出（Markdown/JSON/表格/三段式等）”，并写一份**可验证**的结构化报告。
要求：
1) 请先搜索：instruction following failure、format adherence、prompt drift、structured output reliability、constrained decoding、JSON mode / function calling、grammar-based decoding 等关键词（中英都搜）。
2) 请交叉验证：
- 论文/预印本（arXiv/ACL/NeurIPS/ICLR等）
- 主流模型/框架官方文档（OpenAI/Anthropic/Google/Meta、LangChain/LlamaIndex 等）
- 工程实践文章（但需标注可信度并与论文/官方文档对照）
3) 输出必须包含：
- **生成时间戳（精确到日期）**
- **来源链接（每条结论至少 1 个链接）**
- **按“原因类别”归纳**：指令冲突/上下文长度与干扰/采样随机性与温度/对齐训练与安全策略/模型能力边界/格式定义不严谨/多轮对话漂移/多工具或多任务干扰
- **证据对比表**：结论｜支持证据A｜支持证据B｜矛盾证据｜你的判定
- **工程对策清单**：从“最便宜到最强约束”排序（如：更严格的格式规范→示例→自检→分段生成→解析器反馈→函数调用/JSON模式→语法约束解码）
4) 请在结论里明确：哪些因素是“不可消除只能缓解”，哪些是“可通过约束解码/工具链基本解决”。

3. [Gemini 深度挖掘指令]

你是“LLM 可靠性/可控生成”方向的研究员。围绕现象“模型多次仍无法稳定按格式输出”，做一次深度机制分析与研究设计。
请执行：
1) **研究计划**：列出需要覆盖的知识模块（指令层级与冲突、RLHF/对齐机制、解码策略、上下文窗口与注意力稀释、分布外指令、工具调用与结构化输出、语法约束解码）。给出每模块的关键问题与预期证据类型。
2) **机制深挖**：用可检验的假设解释格式失败（至少 8 条），并说明每条假设的可观测信号（例如：在哪些场景更易崩、与温度/长度/多轮轮次的关系、常见偏离模式）。
3) **实验设计**：制定一个最小可复现实验（MVP）来量化“格式遵循率”，包含：
- 数据集构造（指令模板族、难度分级、干扰项注入）
- 指标（parse success、exact match、structure distance、drift rate）
- 消融（温度/Top-p、示例数量、分段生成、强约束/弱约束、系统提示 vs 用户提示）
4) **分歧点分析**：总结学界/业界对“长提示词是否必然更听话”“RLHF 是否降低可控性”“结构化输出应靠提示还是靠约束解码/工具”的主要分歧，并给出双方最强论据。

5) **证据对比表**：输出一张表：观点｜机制解释｜支持证据｜反例｜适用边界｜工程含义。
6) **落地建议**：给出一套“从提示词到工具链”的分层方案（不止建议，还要说明为什么、适用条件与风险）。