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]

同一提示词多次运行仍可能出现格式偏离或遗漏，输出一致性不稳定。

2. [ChatGPT 联网搜索指令]

你将使用联网搜索，解释“LLM 为什么常常无法稳定按指定格式输出（如严格 Markdown/JSON/三段式结构）”。

检索要求 - 优先近 24 个月信息；同时补充经典观点（如 RLHF/对齐对指令遵循的影响）。 - 至少交叉验证 5 个独立来源：官方文档/研究论文/工程实践博客/评测报告/社区复现。 - 同时检索中文与英文关键词。

建议检索关键词（可扩展） - instruction following, format adherence, structured output, schema adherence - prompt conflict / competing instructions / instruction hierarchy (system vs user) - context window / attention dilution / long prompt degradation - decoding randomness (temperature/top_p), length penalty, max_tokens truncation - RLHF side effects, verbosity bias, refusal/safety filter interfering with format - JSON mode / function calling / constrained decoding / grammar-based decoding

输出格式（结构化） 1) 现象归类：常见“跑偏”的具体表现（缺标题、顺序错、插入寒暄、漏字段、截断等） 2) 主要原因（按优先级）： - 指令冲突与优先级 - 长上下文稀释与注意力分配 - 采样随机性与解码策略 - 输出长度与截断（max tokens/上下文限制） - 安全/对齐策略对格式的干扰 - 模型/接口能力差异（是否支持约束输出） 3) 可复现的工程对策清单（每条配适用场景+局限）： - 明确分隔符、锚点、逐条编号、少样例 few-shot - 降低温度、限制输出范围、分步生成（先 outline 再填充） - 使用“结构化输出能力”（如 JSON schema / function calling / constrained decoding） 4) 推荐阅读与引用：给出可点击来源与一行摘要（不要堆砌链接）

注意 - 不要只给主观建议；每个关键结论尽量落到“来源证据”。

3. [Gemini 深度挖掘指令]

基于你已搜集到的资料，请对“格式遵循失败”做系统层面的机制解释，并提出可验证的改进方案。

任务 1) 机制模型：用 3 层结构解释（训练目标/对齐方式 → 推理解码 → 提示词交互），说明为什么“长提示词 ≠ 更听话”，以及为什么会出现格式漂移。 2) 因果图或流程图：给出一个“诊断流程”（输入：现象；输出：最可能原因与对应修复策略）。 3) 对照实验设计：给出一套最小实验（ablation）来定位问题来源。 - 变量示例：温度/Top-p、max_tokens、是否 few-shot、是否分步生成、是否加入硬分隔符、是否使用 JSON schema/函数调用 - 指标示例：字段完整率、顺序正确率、标题命中率、冗余寒暄率、截断率 4) 策略对照表：至少 8 种常见“让模型更按格式输出”的技巧，列出：原理假设、优点、失败模式、适用任务。 5) 最终给出一个“可直接复制”的改写版通用提示词模板（中文），用于要求模型稳定输出三段式 Markdown；并给出 2 个变体： - 变体A：更短更硬（强调约束） - 变体B：更稳更可读（强调容错与自检）

输出要求 - 允许适度总结与建议，但要明确区分“证据结论”与“推断”。 - 结果尽量用小标题、编号、表格呈现。