

1. [事实快照]

多次要求模型按固定格式输出仍频繁失败，输出结构不稳定、易偏离指令。

2. [ChatGPT 联网搜索指令]

你是“LLM 指令遵循/格式约束”方向的技术调研员。请强制联网搜索并交叉验证，回答：为什么同一模型在多次对话中仍会出现“无法稳定按指定格式输出”的现象？

要求：1) 请先搜索并汇总近 18 个月关于“instruction following / format adherence / structured output / JSON mode / function calling / constrained decoding / system prompt hierarchy”的权威资料（优先：OpenAI/Anthropic/Google/Meta 官方文档或博客、ACL/EMNLP/NeurIPS 论文、可靠工程实践文章）。2) 请交叉验证：同一结论至少用 2 个独立来源支持；若存在相反说法，必须并列呈现并解释分歧点。3) 必须覆盖的原因类别（按“证据强度”排序）：- 解码随机性与采样参数（temperature、top_p）对格式稳定性的影响 - 指令层级与冲突（system/developer/user）、上下文污染、提示注入 - 安全策略/内容过滤导致的结构中断或重写 - 上下文窗口、截断、长对话漂移、记忆/工具调用对输出的干扰 - “结构化输出”机制：JSON mode、函数调用、语法约束解码（constrained decoding）的差异与限制 - 不同模型/版本/客户端实现差异（例如同名模型在不同产品形态下的约束不同）4) 输出：结构化报告（必须包含时间戳与来源链接）。格式如下：- A. 关键结论（<=8 条，每条附 1-2 个来源链接）- B. 现象→可能机制映射表（两列：Failure Pattern / Supported Mechanisms）- C. 证据时间线（按发布日期排序，注明 YYYY-MM-DD）- D. 术语表（10 个核心术语，给出最短可用定义+来源链接）- E. 未解决问题与验证建议（列出可复现实验设计要点，不要长篇科普）

3. [Gemini 深度挖掘指令]

你是“LLM 对齐与可靠性工程”研究员。围绕现象：模型无法稳定按固定格式输出，请进行深度机制分析并形成可复现实验框架。

任务：1) **制定研究计划**：列出需要检索/整合的知识模块（RLHF/对齐、解码与约束解码、提示注入与上下文安全、产品实现差异、评测方法与统计）。2) **机制深挖**：从训练（SFT/RLHF/偏好数据）、推理（采样/beam/约束解码）、以及系统层（指令层级/安全重写/工具与记忆）三层，解释“同一提示多次运行仍不稳定”的因果链。3) **冲突点分析**：收集并对比不同来源对“格式稳定性的主要瓶颈”的分歧（例如：随机性主导 vs 系统安全重写主导；提示工程可解决 vs 需要结构化输出接口）。4) **证据对比表**：生成表格：来源/主张/证据类型（实验/官方文档/工程经验）/适用条件/反例。5) **评测与实验**：设计一个最小可行基准（包含 20 条提示，分 4 类失效模式），并给出统计指标（格式合规率、字段缺失率、编辑距离、拒答率等）与复现实验步骤。

输出要求：- 用分节报告；关键结论必须可被“实验或文档证据”支撑；不要空泛建议。