

【用户问题（原文）】 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？

【ChatGPT 回答（原文）】

1. [事实快照]

LLM 输出是概率生成；受指令冲突、上下文与解码策略影响，格式遵循率天然不稳定。

2. [ChatGPT 联网搜索指令]

你是一名“LLM 输出格式鲁棒性”研究助理。请开启联网搜索，检索近 24 个月内与“LLM/ChatGPT 结构化输出失败、格式遵循率、JSON/Schema、function calling、约束解码、提示词漂移（prompt drift）”相关的资料，并交叉验证不同来源（官方文档/学术论文/工程实践博客/评测报告）。

目标问题：为什么用户反复要求固定格式仍常失败？有哪些可复现的原因分类与工程解法？

- 检索要求（必须执行）：** 1) 给出你使用的检索关键词组合（中英都要），并说明每组关键词想验证什么假设。
2) 至少覆盖：OpenAI/Anthropic/Google 等官方结构化输出机制或指南、学术研究（instruction following / constrained decoding / structured generation）、工程实践（schema 校验、重试、解析器）。
3) 对每条关键结论都要给出处（链接+简短引用/摘要），并标注发布日期。

输出结构（必须按此顺序）： A. 现象定义：什么叫“按格式输出失败”（举 3 类失败形态：缺字段/多话/结构破坏）。

B. 归因框架（表格）：

- 维度1：模型侧（对齐/训练/指令跟随能力差异）
- 维度2：提示词侧（冲突、歧义、过长、示例不足、位置效应）
- 维度3：解码/参数侧（temperature、top_p、max_tokens、stop）
- 维度4：上下文侧（长对话、遗忘、干扰、注入）
- 维度5：工具/接口侧（JSON 模式、函数调用、schema、解析器）

每格：典型表现 → 机制解释 → 证据来源 → 诊断方法。

- C. 最有效的工程解法清单（按收益排序）：至少 8 条，每条包含“适用场景/代价/失败模式/示例提示词”。
D. 可复现实验设计：给出一个小型 benchmark（≥20 条样例），定义指标（格式合规率、字段正确率、冗余率、重试次数等），并提供评测伪代码或流程图。

E. 给出“最小可用提示词模板”两版：

- 版本1：纯提示词模板（无工具）
- 版本2：结合 schema/函数调用（如适用），含严格校验与重试策略

注意：不要泛泛而谈；必须用联网证据支撑、并明确哪些结论仍存在争议。

3. [Gemini 深度挖掘指令]

你是“结构化生成与指令遵循”方向的研究分析员。基于你检索到的资料与常识推理（可继续广搜，但重点是综合与机制解释），对“为什么 LLM 常常无法稳定按格式输出”做系统级拆解，并给出可操作的鲁棒性方案。

你需要完成：1) 机制层解释（不少于 5 个机制）：

- 概率解码为何不等价于满足硬约束
- 指令冲突与优先级（系统/开发者/用户/示例）如何导致漂移
- 长上下文的注意力分配与“位置效应/最近性效应”
- RLHF/对齐目标对“啰嗦解释 vs 严格格式”的拉扯
- 解析器视角：为什么“看起来差一点点”会导致机器解析失败

2) 方法谱系对照表（必须给表格）：

- 纯提示词（模板/少样例/负面约束）
- 自检与反思（self-check / repair step）
- 多轮纠错（validator → re-ask）
- 结构化接口（JSON schema / function calling）
- 约束解码/语法引导（如 CFG/regex/finite-state constraints 的思想）

每种：可靠性、成本、实现难度、适用场景、典型坑。

3) “诊断流程图”：当格式失败时，如何用最少实验定位原因（参数/长度/冲突/示例/接口）。

4) 给出一个“鲁棒提示词合同（Prompt Contract）”范式：

- 输入假设、输出 schema、允许/禁止内容、错误处理、重试策略、终止条件

并给出 2 个完整示例：一个短文本任务，一个复杂多字段任务。

5) 争议与边界：列出至少 4 个仍会失败的场景（例如超长输出、开放域生成、需外部事实校验、对话中混入新目标等），并说明为何目前方案难以完全解决。