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？

回答（原文）

1. [事实快照]

你多次要求模型按固定格式输出，但结果频繁不遵循或漂移。

2. [ChatGPT 联网搜索指令]

你是一名“LLM　输出一致性/格式遵循”调研员。　**请强制联网搜索**，调查并解释：为什么大模型会在“按指定格式输出”上反复失败，以及哪些方法能显著提高格式稳定性。
（提示词设计逻辑：该问题高度依赖不同模型/版本/解码策略与近期实践经验，必须检索**最新**官方文档与工程经验，并进行**交叉验证**以避免单一来源偏差。）

检索任务（必须执行）

1. **请先搜索**关键词组合（至少覆盖中英）：
2. “instruction following format compliance failure”
3. “JSON schema enforcement LLM” / “structured output reliability”
4. “prompt drift / output drift”
5. “temperature top_p formatting stability”
6. “system message hierarchy / prompt injection”
7. “OpenAI structured outputs / JSON mode / function calling（若适用）”
8. **请交叉验证**：至少对比
9. 官方文档/技术博客（OpenAI/Anthropic/Google 等）
10. 研究论文或技术报告（arXiv/ACL/NeurIPS 等）
11. 高质量工程实践文章（大型公司工程博客或知名技术社区的深度帖）
12. **请检索并总结**：有哪些“可复现”的失败模式（例如：长上下文后格式崩坏、插入解释导致结构破坏、列表编号漂移、Markdown 与代码块混用、遇到未知信息时输出跑题等）。

输出要求（结构化报告，必须包含时间戳与来源链接）

请输出一份结构化报告，包含以下章节（每章都要给出**证据来源链接**与**检索时间戳**）： - A. 现象分类：常见“格式失败类型”清单（至少 8 类）
- B. 根因假设地图：将原因分为 **解码/采样、对齐与训练、上下文与指令冲突、任务歧义、工具/插件与多轮对话干扰、模型能力边界** 等，并说明证据来自哪里
- C. 变量影响表：temperature、top_p、max_tokens、stop 序列、分隔符、示例数量、提示词长度、角色层级（system/developer/user）等对格式遵循率的影响（能量化则量化）
- D. 解决方案对照：提示词层、解析层、协议层（schema/函数调用/约束解码）三条路线的优缺点与适用场景
- E. 最小可复现实验（MRE）：给出一个“你建议我如何复现实验并记录成功率”的步骤清单
- F. 一页结论：用 10 条以内要点总结“为什么会这样”+“最有效的 3–5 个改进杠杆”

3. [Gemini 深度挖掘指令]

你是“LLM 对齐与解码机制”研究员。围绕现象：**模型无法稳定按指定格式输出**，请进行深度机制分析与研究设计。（提示词设计逻辑：仅靠经验技巧不够，需要把“格式失败”拆解为可测变量与机制假设，并用多源证据对照，产出可迭代的实验与诊断框架。）

任务 1｜制定研究计划（必须）

1. 研究目标：把“格式遵循失败”形式化为可评测指标（如：Schema 合规率、字段缺失率、额外文本率、位置漂移率、解析成功率、跨轮稳定性）。
2. 数据与实验：设计一个包含不同难度、不同干扰因素（长上下文、指令冲突、加入示例、加入反例、加入噪声文本、多轮对话、工具输出）的测试集。
3. 评价方法：给出统计方式（置信区间/显著性检验/分层统计）与可视化建议。

任务 2｜机制深挖：提出并论证“为何会失败”的假设族

至少提出 6 个机制假设，并逐一说明： - 机制描述（例如：采样随机性导致结构边界不稳；模型对“内容正确性”优先于“结构正确性”；上下文中存在更高优先级指令冲突；长上下文注意力稀释导致尾部结构崩坏；训练分布中严格结构样本不足；模型倾向补充解释性文本破坏 schema 等）
- 可观测预测（如果该机制为真，会出现什么特征）
- 反证方式（如何设计实验推翻该机制）

任务 3｜多源综合：生成“证据对比表”

请生成一个表格（必须是可复制的 Markdown 表格），列为： - 观点/机制假设
- 支持证据来源 1（摘要要点）
- 支持证据来源 2（摘要要点）
- 反对/限制证据（摘要要点）
- 你给出的可信度评分（0–5）
- 对工程实践的含义（可操作建议）

任务 4｜输出“可落地”的改进方案（分层）

分别给出三层方案，每层至少 5 条，并说明适用条件与副作用： 1. **提示词层**：结构锚点、分隔符、最小化自由文本、先规划后填充、示例/反例、错误自检等
2. **协议层**：严格 schema、函数调用/工具调用、约束解码、stop 序列、双通道输出（content vs data）等
3. **后处理层**：解析器重试、修复器模型、验证-反馈循环、自动回滚与回归测试集

任务 5｜给出“你建议我明天就能跑”的实验清单

输出一个按优先级排序的实验列表（不少于 8 个），每个实验包含：目的、变量、操作步骤、预期观察、记录模板。