

1) 简略事实回答 - 多数情况是：模型在同一轮里同时接收了多层次/多目标指令（系统/开发者/用户/上下文示例），再叠加随机性与上下文长度限制，导致“格式约束”被部分忽略或被别的约束覆盖。

2) 给 ChatGPT 的搜索提示词 - 你是联网版 ChatGPT。现在是 2025-12-17 (America/Los_Angeles)。请只围绕问题“为什么我让模型按固定格式输出经常失败”做完整、结构化、可验证的回答，不要扩展到别的主题。 - 任务：

1) 用 5-8 条要点总结“格式不稳定/不服从”的常见客观原因（如：指令层级冲突、上下文溢出、采样随机性、对齐策略、解析器/渲染层差异、工具调用/函数调用影响等），每条都要给出可核查的依据来源。 2) 给出一个“可重复验证”的最小实验设计：同一问题在不同设置（temperature、top_p、system prompt、few-shot 示例、JSON mode / function calling 等）下，如何观察格式遵循率变化；说明评价指标（例如：严格正则匹配通过率）。 3) 列出 6-10 条高权威参考：优先 OpenAI 官方文档/研究博客、Anthropic 文档、Google/DeepMind 公开资料、学术论文（arXiv/ACL/NeurIPS/ICLR）与经典提示工程研究。每条写清：标题、作者/机构、发布日期、链接、与你回答中哪一点对应。 4) 对互相矛盾的说法做“分歧对照表”：观点A/观点B/各自证据/你判断差异来自哪里（例如实验设置不同）。 - 约束： - 必须标注关键结论对应的具体日期与来源； - 不要给个人建议或情绪安慰； - 不要输出无法验证的笼统话； - 全文按：现象→原因分类→可复现实验→证据列表→分歧表的顺序。

3) 给 Gemini 的提示词（用于进一步挖掘） - 你是 Gemini，目标是在已有回答基础上做多源、机制层面的深挖。今天日期：2025-12-17 (America/Los_Angeles)。主题只限：“LLM 为什么经常无法稳定按指定格式输出”。 - 先给出研究计划（至少 6 步）：要查哪些一手材料、哪些对照实验、哪些关键术语与可检验假设。 - 然后输出综合报告，包含： 1) 机制解释：从训练目标（next-token）、对齐（RLHF/RLAIF）、解码采样、长度惩罚、工具/函数调用、以及 UI/渲染/中间层（比如 markdown 渲染、结构化输出解析）等维度，分别解释“格式破坏”的路径。 2) 证据表（表格）：每条证据包含【来源类型（官方文档/论文/博客/基准评测）】【结论一句话】【实验或论据要点】【限制条件】【链接】【发布日期】。 3) 分歧解释：把不同来源对“温度/系统提示/示例数量/JSON模式”等因素的结论冲突列出来，并给出你认为最可能的解释（实验条件差异、模型版本差异、评估口径差异）。 4) 可复现实验清单：至少 5 个可操作实验（输入、设置、预期观察、失败判据），用于验证哪些因素对格式遵循率影响最大。 - 约束：不扩展到与格式无关的泛泛提示词教程；每个核心结论必须能追溯到具体来源或可复现实验。