

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]

多次要求固定格式输出，模型仍常跑偏或漏段。

2. [ChatGPT 联网搜索指令]

你是“LLM 可控输出 / 提示词工程”研究员。请**强制使用联网搜索**，围绕议题：

“为什么大语言模型经常无法稳定遵循指定输出格式（Markdown/JSON/多段结构），以及可验证的工程化缓解手段是什么？”

要求你按步骤执行，并输出**结构化报告**（必须含时间戳与来源链接）：

A. 搜索与交叉验证 1) 请先搜索：模型不遵循格式/指令的常见原因（instruction following failure / format drift / structured output failure）。 2) 请交叉验证：至少 3 类来源必须覆盖并互相印证/对照： - 官方文档/产品说明（如“结构化输出、JSON 模式、函数调用/工具调用、约束解码”相关能力或限制） - 学术论文/技术报告（instruction following、RLHF、constrained decoding、grammar decoding、prompt injection 等） - 工程实践文章/案例复盘（大型公司工程博客或高质量技术社区，需注明作者与日期）

B. 必须回答的“可证伪”问题清单 3) 训练目标与对齐：next-token 目标与 RLHF/偏好优化如何导致“格式服从”并非硬约束？ 4) 解码与采样：temperature/top_p/stop sequence/max tokens 等参数如何影响“格式漂移”？给出可复现实验或引用。 5) 指令层级与冲突：system/developer/user 三层指令冲突时，格式指令为何会被“降权”或被任务内容挤占？ 6) 上下文与注意力：长提示词、示例过多、信息密度高时，为什么更容易漏段/重复/跑题？是否存在可验证阈值或经验法则？ 7) 结构化输出能力：哪些模型/接口提供“JSON schema/函数调用/严格模式/语法约束”，它们的边界与失败案例是什么？

C. 输出格式（强制） - 报告元信息：检索时间（精确到分钟，含时区）+ 关键词列表 + 你使用的检索式 - 结论摘要：不超过 10 条 bullet，每条后面必须跟 1-2 个来源链接 - 证据对比表：列 = 观点/机制、支持证据、反例或限制、来源链接、发布时间 - “工程化缓解手段”清单：按优先级排序（从“最稳”到“最不稳”），并给出每条的适用场景与风险 - 复现实验建议：提供一个最小可运行的评测设计（输入、期望输出、评分规则），用于量化格式遵循率

D. 质量约束 - 只引用可访问的网页来源；每条关键结论都要链接。 - 明确区分：事实、推断、观点。 - 若来源之间冲突，必须指出冲突点并解释可能原因。

（附：在报告最后追加“提示词设计说明”：用 5-8 条解释你为何选择这些检索与报告结构，以便研究者复用。）

3. [Gemini 深度挖掘指令]

你是“语言模型行为机理 + 评测方法”研究员。请基于议题：

“模型在强格式约束下仍会输出漂移：原因机理、可测量分解、以及系统性改进策略”

请执行以下任务，并输出一份可用于写作/实验的长文档（允许很长）：

1) **研究计划 (Research Plan)** - 列出 6-10 个你需要调用的知识模块（如：next-token 训练目标、RLHF/偏好优化、解码策略、约束解码/语法解码、上下文长度与注意力、指令层级、工具调用/函数调用、提示注入与对抗、评测指标设计）。- 为每个模块写：研究问题、关键假设、需要的证据类型、潜在混淆因素。

2) **机制拆解 (Mechanistic Decomposition)** - 给出一张“因果图/流程图”（用文本表示也可）：从“用户格式指令”到“最终输出 token”之间，哪些环节会引入偏差？- 建立一个“失败模式分类”（taxonomy），至少包含：漏段、格式错位、内容越界、重复、幻觉补段、结构崩坏、被示例诱导偏移等。

3) **学术分歧与冲突点分析** - 至少列出 3 组典型分歧（例：A) 主要原因是解码随机性 vs 主要原因是训练目标与对齐；B) 长提示更稳 vs 长提示更容易漂移；C) few-shot 示例是关键 vs schema/grammar 约束才是关键）。- 对每组分歧：分别给出支持论据、反例、适用条件与未解决问题。

4) **证据对比表 (必须)** - 生成表格：行 = 关键机制/主张；列 = 预测、可观测指标、支持证据、反例、如何做控制变量实验、你对可信度的评级。

5) **实验设计：最小可复现 + 进阶套件** - 最小实验：1 个任务、3 个提示版本、2 个解码设置，定义“格式遵循率”与“语义正确率”评分。- 进阶套件：提供一个 20-50 条的“格式评测测试集”设计蓝图（题型覆盖、难度分层、评分规则、统计检验建议）。

6) **改进策略库 (Intervention Library)** - 输出不少于 12 条干预策略，按“约束强度”分层：- 软约束：指令写法、分隔符、反例、元提示、回显检查 - 半硬约束：schema 提示、self-check、二阶段生成（先结构后填充） - 硬约束：函数调用/JSON schema/grammar decoding/外部解析器回退 - 每条策略给：何时有效、代价、失败案例、与其他策略的组合方式。

7) **可写入论文/报告的输出形态** - 最终请产出：- 一页“执行摘要”（可直接放进 workshop 报告） - 一段“相关工作/背景”草稿（带可检索的引用线索：作者+年份+关键词即可） - 一份“方法论章节”草稿（评测指标与实验流程）

额外要求：- 当你提出任何结论时，请同时给出：它的可检验预测 + 可能反例。- 如果你引用外部材料，请给出可检索线索（作者/机构、年份、标题关键词、来源类型），并在文中标注。