1. [事实快照]

同一提示词多次调用时，模型仍可能偏离指定输出格式。

2. [ChatGPT 联网搜索指令]

你是「LLM 输出可靠性」调研员。请**强制联网搜索**并交叉验证，围绕以下问题撰写结构化报告：

研究问题：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？

要求： 1. 请先搜索：LLM/ChatGPT “format adherence / structured output / JSON 输出不稳定 / 指令遵循失败”的成因与案例。 2. 请交叉验证：至少使用 6 个来源，其中至少包含： - 官方文档（OpenAI / Anthropic / Google 等任一） - 学术论文或技术报告（arXiv/ACL/NeurIPS/ICLR 等） - 工程实践文章（框架/库/最佳实践） 3. 请输出一份**结构化报告**，必须包含： - **生成时间戳**（精确到日期） - **每条结论对应的来源链接**（可点击 URL） - **证据等级**（例如：官方文档 / 同行评审论文 / 工程实践 / 个人博客） 4. 报告结构（必须按以下标题输出）： - A. 现象定义：什么叫“格式输出失败/漂移”？常见表现有哪些？ - B. 主要原因清单（至少 10 条）：按“模型内因 / 提示词与上下文 / 解码参数 / 工具与系统 / 评测与统计误差”分类 - C. 复现与诊断：如何设计最小可复现实验（MRE），如何区分“偶发偏差”与“系统性失效”？ - D. 工程对策对比表：prompt 约束、低温/采样策略、function calling / JSON mode、约束解码(constrained decoding)、后处理校验与重试、分步生成等 - E. 失败模式→对应对策映射：至少 8 组（例：token 截断→缩短输出/分段；指令冲突→层级化指令；长上下文干扰→隔离上下文等） 5. 结尾必须给出： - “最可能的 3 个根因” + “最有效的 3 个工程手段” - 一份 7 天内可执行的排查清单（每天 3-5 个动作）

注意：不要泛泛而谈；所有关键断言都要附来源链接，并说明是“哪些证据支持”。

3. [Gemini 深度挖掘指令]

你是「LLM 对齐与生成机制」研究者。请围绕以下问题做 **深度机制分析**，输出一份可用于写 workshop 的分析笔记：

研究问题：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？

请执行以下步骤： 1. **研究计划**： - 列出需要调用的知识领域（例如：语言模型生成分布、指令层级与冲突、解码策略、约束解码、对齐训练、长上下文干扰、工具调用协议、鲁棒性评测）。 - 给出检索关键词组合（中英双语）与筛选标准（优先官方/论文/复现实验）。 2. **机制深挖（必须画出因果链）**： - 从“概率生成 + 采样/温度/Top-p”到“格式 token 的脆弱性”解释为什么格式会漂移。 - 分析“指令冲突/注意力稀释/上下文污染/截断/安全策略/系统提示词层级”的作用路径。 - 给出至少 3 个你认为最关键的机制假说，并说明可证伪条件。 3. **学术分歧与冲突点**： - 对比不同来源对以下争议的看法： a) 仅靠更长提示词能否保证格式遵循？ b) function calling / JSON mode 是否从根本上解决问题？ c) 约束解码 vs 后处理重试，哪个更可靠、代价更低？ - 请生成一张“观点—证据—局限”对比表。 4. **实验设计**： - 设计一套评测协议：任务集构造、控制变量（温度/上下文长度/系统指令/示例数量）、指标（合规率、编辑距离、语法校验通过率、重试次数、成本/延迟）。 - 给出最小实验（1 小时可跑）与扩展实验（可写成报告）的方案。 5. **输出形式（硬性要求）**： - 一份分层大纲（可直接变成论文的 Related Work / Method / Findings）。 - 一张“证据对比表”（来源、结论、支持证据、反例、可信度）。 - 最后给出 10 条可操作的写作要点（用于向导师解释为什么会失败、怎样工程化解决）。

约束：避免口号式建议；每个结论都要指向可验证的机制或可复现实验路径。