

## 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]

LLM格式输出不稳定通常源于提示词歧义性、结构复杂度过高、或缺少明确示例。即使详细描述格式要求，模型仍可能因理解偏差而产生非预期输出。

### 2. [ChatGPT 联网搜索指令]

请使用联网搜索功能，检索2024年以来关于"LLM格式化输出"和"prompt engineering best practices"的最新技术文章与讨论。重点关注：JSON mode、structured outputs、few-shot examples等提高格式稳定性的方法。请交叉验证GitHub、技术博客、官方文档等不同来源，并以结构化列表呈现各方法的适用场景和局限性。

### 3. [Gemini 深度挖掘指令]

基于已知信息，请从系统机制层面分析：为什么LLM在遵循格式指令时会出现不一致？综合考虑token预测机制、注意力分布、提示词长度对性能的影响等因素。对比不同提示策略（如XML标签 vs Markdown标题、约束式 vs 示例式）的效果差异。如果存在学术研究或实验数据，请以表格形式总结各方法在格式遵循度上的表现对比。