原始输入 (Q)

“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]

大语言模型无法严格遵循格式输出通常归因于训练数据的概率分布特性、RLHF（人类反馈强化学习）导致的“对齐税（Alignment Tax）”、以及注意力机制在处理长上下文时的聚焦偏差。这是当前生成式AI的概率本质而非逻辑缺陷。

2. [ChatGPT 联网搜索指令]

你是一个NLP技术研究员。请利用联网搜索功能，针对“LLM指令遵循能力失效（Instruction Following Failure in LLMs）”撰写一份技术简报。
要求：

1. 搜索并引用2023-2024年关于“Prompt Engineering Robustness”或“Format Constrained Decoding”的学术论文摘要。
2. 查找OpenAI或Anthropic官方开发者论坛中，关于“System Prompt”权重被“User Input”覆盖的已知案例和技术解释。
3. 交叉验证不少于5个技术来源（如ArXiv, Hugging Face Papers），分析“模型幻觉”如何具体影响JSON/Markdown等结构化数据的生成。
4. 输出格式：现象分类（如格式遗忘、过度拒绝、自言自语）+ 技术归因 + 来源链接的时间戳列表。

3. [Gemini 深度挖掘指令]

你是一个人工智能架构师和认知科学专家。基于用户遇到的“模型反复无法按格式输出”这一痛点，请进行深度的系统性机制分析。
请执行以下步骤：

1. 研究计划：制定分析路径，涵盖Transformer架构的Attention Head机制、Softmax概率采样对刚性规则的破坏，以及Instruction Tuning（指令微调）的数据集分布问题。
2. 机制深挖：深度剖析“为什么模型倾向于解释而非执行”。请分析预训练阶段的“续写文本”目标函数与SFT阶段的“遵循指令”目标函数之间的潜在冲突。
3. 分歧点分析：学术界对于“通过Prompt工程解决”与“通过受限解码（Constrained Decoding）算法解决”存在哪些路线分歧？请对比两者的优劣及适用场景。
4. 输出形式：生成一份深度剖析报告，包含一个“指令依从性衰减模型图解”描述。