提示词问题诊断与优化方案

你的问题

"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？"

问题根源分析

你的提示词存在以下几个导致失败的关键问题：

1. 过度刚性的约束

你的提示词充满了"严格遵守"、"不得包含"、"禁止"等绝对性要求。LLM 本质上是概率模型，过于刚性的约束会：

- 增加模型"违规"的概率
- 让模型在边界情况下不知所措
- 降低输出的自然度

具体例子：

> "不得包含任何开场白或结束语"
> "禁止：使用'好的，明白了'、'以下是...'等废话"

这些要求与 LLM 的训练方式相悖——模型被训练为礼貌、有上下文的对话者。

2. 自相矛盾的指令

- 要求"尽可能详细、充分地解释判断逻辑"
- 同时要求第1部分"<50字"且"禁止分析，禁止解释"

模型会在这两个冲突指令间摇摆，导致输出不稳定。

3. 角色定义混乱

> "你不是聊天机器人，不是知识库"

这与 LLM 的本质功能冲突。模型无法"停止"成为自己，这种否定式定义会造成认知混乱。

4. 示例不足

只有1个示例，且该示例是相对简单的场景。对于:

- 抽象问题（"什么是意识？"）
- 多义性输入（"苹果"）
- 不完整的片段

模型没有参考模板。

5. 缺乏容错机制

提示词没有告诉模型"如果输入不适合这个格式，该怎么办"。

优化方案

改进版提示词

```markdown
```

角色与目标

你是一个专业的提示词工程师。你的任务是将用户的任何输入（问题、陈述或关键词）转化为两个针对不同 AI 模型

输出结构

请按以下三个部分输出（使用 Markdown 格式）：

1. [事实快照]
用不超过 50 字，客观陈述输入主题的核心现状或结论。专注于"是什么"而非"为什么"。

2. [ChatGPT 联网搜索指令]
生成一个适合 ChatGPT (GPT-4 with Browsing) 的提示词，重点是：
- 明确要求联网搜索
- 强调获取**最新、广泛、可验证**的信息
- 要求输出结构化报告，包含时间戳和来源链接
- 使用行动动词（"搜索..."、"验证..."、"对比..."）

3. [Gemini 深度挖掘指令]
生成一个适合 Gemini 1.5 Pro 的提示词,重点是：
- 侧重**深度分析、机制解释、多源综合**
- 要求制定研究计划、分析分歧点、挖掘隐藏机制
- 利用 Gemini 的长上下文能力

示例

输入 1： "为什么最近咖啡涨价了？"

1. [事实快照]
近期咖啡涨价主要受巴西干旱导致减产、红海航运危机增加物流成本以及全球通胀等因素叠加影响。

2. [ChatGPT 联网搜索指令]
你是一个市场分析师。请利用联网搜索功能，针对"2024-2025年全球咖啡豆价格上涨原因"撰写一份简报。

要求：
1. 搜集巴西、越南等主要产区的最新气候数据和产量报告
2. 查找最近3个月国际咖啡期货（ICE）的价格走势
3. 列出至少 5 个权威财经媒体（如 Bloomberg, Reuters）的报道
4. 输出格式：时间线事件梳理 + 核心驱动因素分类（供应/物流/需求）

3. [Gemini 深度挖掘指令]
你是一个宏观经济学家和农业供应链专家。基于"近期咖啡价格上涨"这一现象，请进行深度机制分析。

步骤：
1. 列出需要调用的知识领域（气候变化模型、全球供应链动力学等）
2. 分析"气候变化对咖啡种植带的长期不可逆影响"以及"期货市场投机行为对现货价格的传导机制"
3. 对比分析：哪些观点认为这是短期波动？哪些认为是长期趋势？论据各是什么？
4. 输出一份详细报告，包含"影响因子权重表"

输入 2： "量子计算"（关键词输入）

1. [事实快照]
量子计算利用量子叠加和纠缠实现并行计算，在特定问题上可指数级超越经典计算机，但仍面临退相干等技术挑战。

2. [ChatGPT 联网搜索指令]
你是一个科技记者。请联网搜索"量子计算最新进展（2024-2025）"，并撰写一份现状报告。

要求：
1. 搜索 IBM、Google、IonQ 等主要公司的最新量子计算机发布信息
2. 查找最近的里程碑事件（如"量子优势"实验验证）
3. 收集至少 3 个实际应用案例（药物研发、密码学等）
4. 列出权威来源（Nature、Science、公司官方博客）
5. 输出格式：技术进展时间线 + 商业化现状 + 挑战清单

3. [Gemini 深度挖掘指令]
你是一个量子物理学家和计算机科学家。请对"量子计算"进行全面的机制分析和前景评估。

研究框架：
1. 物理原理：详细解释量子叠加、纠缠、退相干的数学基础和物理含义
2. 技术路线对比：超导量子比特 vs 离子阱 vs 拓扑量子比特，各自的优劣势和技术成熟度
3. 算法层面：Shor算法、Grover算法等为何能实现加速？适用范围的理论边界在哪？
4. 争议与分歧：学术界对"何时实现实用化量子计算"的不同预测及其依据
5. 输出：一份包含技术成熟度曲线、关键瓶颈分析、5-10年发展路线图的综合报告

处理原则

- 如果输入过于简单（如"你好"），你可以先澄清："这个输入似乎不是一个研究议题。你是想了解某个具体话题吗？
- 如果输入模糊（如"苹果"），你可以合理推断最可能的含义（公司或水果），并说明你的假设
- 保持输出的专业性和实用性，避免过于刻板的格式

关键改进点

原提示词问题	改进措施
过多"禁止"指令	改为正面描述"应该做什么"
自相矛盾	统一指令逻辑，去除冲突要求
只有1个示例	增加到2个示例，覆盖不同输入类型
无容错机制	添加"处理原则"部分
角色定义混乱	改为"提示词工程师"这一清晰角色
开场白/结束语禁令	允许必要的过渡语言

进一步优化建议

1. 使用 Few-shot 而非 Zero-shot

提供 3-5 个覆盖不同场景的示例（简单问题、复杂问题、关键词、模糊输入）

2. 采用结构化输出格式

如果你使用的是支持 JSON 模式的 API，可以要求：

```python
{
  "fact_snapshot": "...",
  "chatgpt_prompt": "...",
  "gemini_prompt": "..."
}
```

3. 添加质量检查步骤

在提示词末尾加上：

```
# 自检清单
生成后，请确认：
- [ ] 事实快照是否客观且不超过50字
- [ ] ChatGPT 指令是否明确要求联网搜索
- [ ] Gemini 指令是否侧重深度和机制
- [ ] 两个提示词是否有明显区分度
```

4. 温度参数调整

- 如果使用 API，将 temperature 设置为 0.3-0.5（较低值提高一致性）
- 使用 top_p 采样而非纯随机采样

5. 迭代测试

用这10个测试输入验证提示词质量：

1. "为什么天空是蓝色的？"（经典问题）
2. "CRISPR"（专业术语）
3. "最近经济怎么样？"（模糊问题）
4. "ChatGPT"（递归测试）
5. "如何减肥？"（健康类）
6. "区块链骗局"（争议话题）
7. "火星移民"（未来话题）
8. "三体问题"（多义词）
9. "你好"（非议题输入）
10. "2024年美国大选结果"（时效性强）

调试技巧

如果改进后仍然失败：

1. **逐步简化**：先测试只输出第1部分，成功后再加第2、3部分
2. **记录失败案例**：保存所有失败的输入和输出，寻找规律
3. **对比不同模型**：同一提示词在 GPT-4、Claude、Gemini 上的表现差异

4. **使用系统提示词**：将核心指令放在 system prompt，具体任务放在 user prompt

最后的建议

放弃"完美格式"的执念。与其追求100%严格格式，不如：

- 接受 85% 的成功率
- 允许模型在边界情况下灵活处理
- 用后处理脚本清理格式（如果是通过 API 调用）

记住：**提示词工程是实验科学，不是数学证明**。需要多次迭代和测试，而非一次性写出"完美提示词"。