# 1. [事实快照]

长提示词并不必然提升遵循率，模型仍会偏离指令。

# 2. [ChatGPT 联网搜索指令]

你是“LLM 指令遵循/提示词工程”研究助理。请**强制使用联网搜索**，针对议题：

> “只要提示词写得足够长，模型就一定会听话” 撰写一份**可验证**的研究型简报，重点获取**最新、广度、交叉验证**的信息。

## 任务目标

1) 识别该断言在学术界与工业界的 **支持证据**与**反例证据**。 2) 区分“提示词长度”与“提示词质量/结构/约束方式”对遵循率的影响。 3) 给出可复现的实验设计，说明在什么条件下“更长”会更好、何时反而更差。

## 搜索与交叉验证要求（必须执行）

- 请先搜索：
- “instruction following prompt length compliance”
- “long prompt instruction hierarchy failure”
- “prompt injection long context instruction conflicts”
- “chain-of-thought prompting length vs accuracy/obedience”
- “system prompt vs user prompt priority instruction hierarchy”
- 请交叉验证：
- 至少 **2 篇**同行评审论文/预印本（arXiv/ACL/NeurIPS/ICLR 等）
- 至少 **2 份**官方/权威技术文档或系统卡（OpenAI/Anthropic/Google/Meta 等）
- 至少 **2 篇**来自不同机构的工程实践文章（企业博客/研究博客），并说明可信度与可能偏差
- 对“长度有效”的证据与“长度无效/有害”的证据，分别列出并对照。

## 输出格式（必须结构化，包含时间戳与来源链接）

请用以下结构输出，并在每个关键结论后附上**来源链接**与**检索日期时间戳**（例如：2025-12-20 17:00 PST）： 1. **结论概览（≤120 字）**：一句话回答该断言是否成立，并给出限定条件。 2. **证据对比表**（表格）： - 列：观点（支持/反对/条件成立）｜证据类型（论文/系统卡/实测/观点）｜关键实验或论点｜局限性｜来源链接｜发布时间 3. **机制与误区清单**： - 解释至少 5 个常见误区（例如：把“更长”当作“更清晰”、忽略指令冲突、忽略上下文窗口与注意力稀释等）。 4. **可复现实验设计**： - 变量：提示词长度（短/中/长/超长）、结构化程度、是否含冲突指令、是否含示例、模型种类 - 指标：格式遵循率、任务成功率、幻觉率、稳定性（多次采样方差） - 给出一个最小实验矩阵（至少 3×3）与预期观察 5. **未解决问题与下一步**： - 列出 3–5 个仍有争议的点，并说明需要什么证据才能确认。

# 3. [Gemini 深度挖掘指令]

你是“LLM 对齐与提示词鲁棒性”研究员。请基于议题：

“只要提示词写得足够长，模型就一定会听话” 进行**深度机制分析**与**研究计划制定**。要求使用多源材料进行综合（你可引用训练与对齐相关公开资料、系统卡、论文、工程实践），并输出可用于写作/实验的结构化内容。

### A. 研究计划（必须）

1) **关键概念界定**：分别定义“听话/遵循率”“长提示词”“结构化提示词”“指令冲突”“层级指令（system/developer/user）”。 2) **假设空间**：提出至少 6 条机制假设，并为每条给出可检验预测，例如： - 注意力稀释/信息过载导致关键约束被忽略 - 指令冲突时模型的优先级解析与对齐偏好 - 长上下文中的遗忘/位置偏置（首因/近因） - 过度约束导致模型“走捷径”（输出看似合规但实则不满足核心任务） - 示例（few-shot）与长度的交互效应 - 采样温度/解码策略对遵循率的影响 3) **方法论路线图**：给出一条从现象→机制→验证的研究路线（至少 3 轮迭代），每轮包含：输入设计、观测指标、判别标准。

### B. 分歧点分析（必须）

- 请列出学术界/工业界可能存在的**三类分歧**（例如：不同模型家族、不同任务类型、不同对齐策略），并为每类分歧： 1) 说明双方的核心主张 2) 给出最关键的证据类型 3) 指出可能的混杂变量

### C. 证据对比表（必须，表格）

生成“证据对比表”，至少包含： - 机制假设｜支持证据摘要｜反例/失败模式｜可复现实验设计｜预期结果模式｜风险与替代解释

### D. 面向实践的可操作结论（必须，但避免口号）

- 给出一组**可执行的提示词设计原则**，但每条原则必须绑定：
- 适用条件
- 失败模式
- 最小修复策略（例如：拆分指令、加入约束检查器、分阶段提示、输出后自检等）

### E. 产出物（必须）

1) 一个可直接用于实验的“提示词长度 × 结构化程度”实验矩阵（含样例提示词骨架）。 2) 一个“诊断流程图”（文字版）：当模型不听话时，如何定位是长度问题、冲突问题、还是任务定义问题。 3) 一段可写入研究报告的“机制综述段落”（避免空泛，强调可证伪性与可复现性）。