1. [事实快照]

提示词更长常提升理解与一致性，但不能保证模型必然遵循指令。

2. [ChatGPT 联网搜索指令]

你是“LLM 指令遵循与提示词工程”的研究助理。 **必须启用联网搜索**，围绕原始议题 Q 产出一份可核查的结构化报告：

原始议题 Q：“有人说只要提示词写得足够长，模型就一定会听话。你怎么看？”

任务

1. 请先搜索并汇总：
2. “instruction following / instruction hierarchy / system vs user messages” 的机制解释；
3. “prompt length” 与 “compliance / format adherence / refusal rate / error rate” 的研究或工程复盘；
4. “context window / attention decay / lost-in-the-middle” 与长提示词失效的证据；
5. “prompt injection / jailbreak / conflicting instructions” 如何导致不遵循。
6. 请交叉验证：对每条关键结论至少使用 **3 个独立来源**（优先：论文/官方技术报告/权威实验室博客；其次：主流媒体的技术专栏；避免单一自媒体）。
7. 请特别搜索 **近 12 个月（以当前日期为准）** 的更新：新模型/新基准/新论文是否改变了“提示词越长越听话”的经验结论。

输出要求（必须严格结构化）

- 报告顶部给出：**生成时间戳（ISO 8601）** 与 **检索时间范围**。
- 以表格输出：
- **证据对照表**：结论｜支持来源（带链接）｜反对/限定来源（带链接）｜证据强度评级（A/B/C）｜备注。
- 给出一条 **时间线**：按发布日期列出重要研究/产品更新（每条都带链接）。
- 单独一节列出：
- **可复现实验或公开基准**（如相关基准、评测设置、可运行代码仓库）及链接。
- 最后一节给出：
- “在什么条件下长提示词更可能有效/更可能失效”的 **条件清单**（每条条件都要引用来源链接）。

3. [Gemini 深度挖掘指令]

你是“对齐与提示词行为学”方向的研究者（Gemini 1.5 Pro/Ultra）。请基于原始议题 Q 做**深度机制分析 + 研究设计**，强调多源综合与长上下文推理。

原始议题 Q：“有人说只要提示词写得足够长，模型就一定会听话。你怎么看？”

A. 研究计划（必须给出可执行清单）

1. 请制定一个 2~4 周的研究计划：每周目标、关键读物类型（论文/技术报告/开源评测）、预期产出（图表/表格/实验日志）。

2. 列出需要覆盖的知识域：
3. 指令层级与对齐训练（RLHF/DPO/constitutional 等）
4. 长上下文机制（注意力、位置编码、检索增强、lost-in-the-middle）
5. 采样与随机性（temperature/top-p）对"听话"的影响
6. 安全策略与拒答机制
7. 提示词注入与冲突指令解析

B. 机制深挖（要求给出"因果链条"）

请用"因果链条"格式解释：为何"更长"并不等于"更可控"。 - 至少提出 5 条机制链条，每条包含：触发条件 → 模型内部/训练层面原因 → 外显行为 → 可观测指标。 - 至少包含以下角度： 1) 指令冲突与优先级解析失败 2) 关键信息在长上下文中被稀释/遗漏（注意力分配） 3) 安全对齐目标与用户目标不一致 4) 生成过程的随机性与鲁棒性问题 5) 评测口径差异（"听话"如何定义）导致的表象矛盾

C. 学术分歧与冲突点分析

1. 请总结至少 3 类不同立场：
2. "长提示词提升遵循"的工程派观点
3. "长提示词不可靠/易漂移"的研究派观点
4. "关键不在长度而在结构/约束/外部工具"的折中观点
5. 为每类观点生成：代表性论据、常用实验设置、常见误区。
6. 请生成一个 **冲突点矩阵表**：观点 A vs 观点 B｜冲突点｜证据类型｜可验证实验。

D. 证据对比表（必须）

请输出一张表：证据｜来源类型（论文/报告/实验复现/基准）｜可重复性｜外推风险｜与你的综合判断。

E. 实验设计：如何验证"提示词长度→遵循度"的关系

请给出一个最小可行实验（MVP）与一个扩展实验（Full Study）： - 变量控制：长度（短/中/长/超长）、结构（无结构/层级结构/模板化）、模型家族、温度、上下文长度、任务类型。 - 指标：格式遵循率、拒答率、指令冲突解析正确率、信息遗漏率、稳定性（多次采样方差）。 - 输出：实验流程图 + 指标定义 + 预期可能出现的"反直觉结果"。

F. 最终输出格式

- 先给一页"结论摘要"（<= 200 字）。
- 再给"机制链条列表 + 冲突点矩阵 + 证据对比表 + 实验设计"。