1. [事实快照]

Q：长提示词≠必然听话；格式/指令仍可能被忽略或漂移。

2. [ChatGPT 联网搜索指令]

你是“提示词工程与模型评测”研究助理。请**强制使用联网搜索**，针对命题： “**提示词越长，模型就一定越遵循指令**（是否成立？在什么条件下更可能/更不可能成立？）” 撰写一份**可验证**的结构化报告。

硬性要求： 1) **先搜索**并收集近 24 个月内的高质量来源： - 学术论文/预印本（arXiv/ACL/NeurIPS/ICLR 等） - 权威机构/实验室技术报告（如 OpenAI、Anthropic、Google DeepMind、Meta 等） - 具可复现细节的工程博客/文档（优先官方与一线团队） 2) **请交叉验证**：同一结论至少由 2 个独立来源支持；对冲突结论要并列呈现。 3) 报告必须包含： - **生成时间戳**（精确到日期） - 每条关键结论的**来源链接**（可点击 URL） - 关键概念定义：instruction following、prompt adherence、format compliance、prompt injection、context window、system vs user instruction hierarchy 4) 输出结构（必须按此顺序）： - A. 证据清单（按来源类型分组，每条含：标题/作者/年份/链接/一句话结论） - B. 结论对比表（列：结论主张｜支持证据｜反例证据｜适用条件｜可信度评分 1-5） - C. 机制与工程因素（仅基于证据归纳：长度、冗余、冲突指令、示例数量、解码策略、温度、工具调用、结构化输出约束等） - D. 可操作建议（面向工程落地：当你必须要“严格格式输出”时，哪些策略比“加长”更有效？每条建议必须指向证据来源） 5) 检索提示：请优先搜索关键词组合（中英都搜）： - “prompt length instruction following”, “format compliance structured output”, “prompt drift”, “instruction hierarchy system prompt”, “prompt injection defense”, “LLM adherence evaluation” - “长提示词 指令遵循”, “结构化输出 格式约束 大模型”, “提示词漂移”

3. [Gemini 深度挖掘指令]

你是“LLM 行为机制 + 评测方法”研究员。基于原始议题： “**提示词足够长是否必然提升指令遵循？**” 请进行**深度机制分析**与**研究设计**，输出一份可直接进入 workshop 报告的内容。

请严格执行： 1) **研究计划**： - 给出 2–3 个可检验假设（例如：长度提升遵循度存在阈值/边际递减；冗余会提高或降低格式遵循；冲突指令导致系统性失效等） - 定义变量：自变量（长度/冗余/示例数/约束强度）、因变量（格式合规率/拒答率/幻觉率/一致性）、控制变量（模型版本/温度/上下文窗口/任务难度） - 设计最小实验矩阵（建议 2×2×2 或 3×3），并说明样本量与评测指标 2) **机制深挖**（不要停留在经验层）： - 从注意力分配、上下文竞争（instruction competition）、对齐训练目标、解码采样、工具调用链等角度解释：为什么“更长”有时更好、有时更差 - 解释“格式崩坏/指令遗忘/局部遵循”的常见失败模式与触发条件 3) **分歧点分析**： - 列出至少 3 类观点阵营（例如：工程经验派/对齐与训练派/可解释性与机制派） - 对每一类：给出其核心论据、最强反例、以及在什么设置下可能成立 4) **证据对比表**： - 输出一个对比表模板（列：主张｜预测｜可观测信号｜实验设计｜潜在混淆项｜如何排除｜预期结果） 5) **结论输出**： - 给出“可用于写作”的结论段落：明确说明“长提示词不是充分条件”，并给出更关键的 5 条设计原则（如：分层指令、约束前置、示例与校验、分段复述、强制自检/重试等） - 同时给出“最小可复现”提示词模板：适用于要求严格 Markdown/JSON 格式输出的场景