1. [事实快照]

网络上流传“提示词越长，模型越听话”的说法。

2. [ChatGPT 联网搜索指令]

你是“LLM 指令遵循/提示词工程”的事实核查研究员。 **请强制联网搜索**并交叉验证“提示词越长越听话”是否成立，输出一份结构化报告。

必须完成的搜索与验证任务： 1) 请先搜索近 12–18 个月关于 **prompt length / verbosity** 与 **instruction-following**、**helpfulness/harmlessness**、**遵循率** 的实证研究、博客实验、评测报告（优先论文/官方文档/权威机构）。 2) 请搜索并总结：长提示词可能带来的失败模式（例如：指令冲突、注意力稀释、上下文窗口/截断、指令层级被覆盖、提示注入风险、模型“只抓末尾/只抓高权重段”现象等）。 3) 请交叉验证：不同模型/平台是否给出相反建议（例如“更长更明确” vs “更短更清晰”），并标注各自的证据类型（实验/基准/经验）与局限。 4) 请检索：是否存在“最佳长度/信息密度”的讨论（如：分段结构、要点列表、约束条件、示例数量、token budget），并收集可复现的对照实验。

输出格式（必须严格遵守）： - 报告时间戳：YYYY-MM-DD（你的本地时间） - 结论摘要：用 3–5 条要点回答“长提示词是否一定更听话？”（每条都要有来源支撑） - 证据时间线：按发布日期排序，列出关键来源（标题 + 发布日期 + 2 句摘要 + 链接） - 证据对比表：来源 | 观点 | 证据类型 | 适用条件 | 反例/局限 | 可信度（高/中/低） - 可复现实验方案：给出 2 组对照实验（短 vs 长；同等信息量不同表达；含/不含示例），写清指标（遵循率、幻觉率、稳定性）、数据记录方式与预期风险

强制要求： - 至少引用 8 个来源，其中至少 3 个是论文/技术报告，至少 2 个是模型/平台官方文档。 - 对关键结论必须“请交叉验证...”：至少用 2 个独立来源支持同一结论。

3. [Gemini 深度挖掘指令]

你是“LLM 对齐与可解释性 + 提示词工程”的研究合作者。请围绕命题： **“只要提示词足够长，模型就一定会听话吗？”**进行深度机制分析与研究设计。

请按步骤输出： 1) **研究计划**：列出需要调用的知识模块与关键词（例如：指令层级/系统提示、上下文窗口与截断、注意力与位置偏置、RLHF/指令微调、提示注入与越狱、信息密度与认知负荷、评测设计与统计检验）。 2) **机制假说**：提出至少 4 个可检验假说，分别解释“长度可能提升/降低遵循率”的路径（例如：澄清约束 vs 注意力稀释；分段结构提升可解析性 vs 冗余导致冲突）。 3) **冲突点分析**：整理不同来源（论文/官方指南/工程实践）之间的分歧：他们对“长度、结构、示例数量、约束强度”的观点为何不同？请给出你认为最可能的原因（模型差异、任务差异、评测口径差异、token 截断等）。 4) **证据对比表**（必须）：证据来源 | 支持哪条假说 | 关键实验设置 | 指标 | 结论 | 可迁移性 | 可能混杂因素。 5) **实验设计（可复现）**： - 给出一个最小可复现实验（MVP），包含任务集构造、提示模板（短/长/同信息密度不同表达）、随机种子/温度设置、评价方法（人评+自动指标）、统计检验。 - 设计 ablation：只变长度、不变结构；只变结构、不变长度；加入/移除示例；将约束前置/后置。 6) **工程化结论**：在不做价值判断的前提下，总结“什么时候长提示词更可能有效、什么时候更可能失效”的条件集合（以 if/then 规则或决策树形式输出）。

输出要求： - 全文以研究报告体例写作，分节清晰。 - 任何推断都要标注为“推断”并说明依据。 - 最后附：一个 10 条的“提示词诊断清单”（用于定位为什么模型没听话：冲突、歧义、权重、截断、注入等）。