# LLM 提示词转译器(三段式)对照评测报告

## 一、证据综述:评测标准与 Benchmark 依据

基于 **Deep Research** 针对 2024-2025 主流 LLM 评测体系的检索,本评测采纳了以下核心维度,旨在建立一套"可复现、可量化"的评价标准:

- 指令遵循的硬性验证 **(Instruction Following):** 采纳 **IFEval** [1] 的核心逻辑,即"不看文采,只看是否满足客观约束"。例如:是否严格输出了三段?字数是否超标?是否包含禁止词汇?
- 指令层级与冲突处理 **(Instruction Hierarchy):** 参考 **IHEval** [2] 的发现,即模型在处理 System Prompt(你的设定)与 User Input(题目 Q)冲突时,往往倾向于顺从 User。Prompt 的鲁棒性在于能否"压制"User 的指令(例如用户问"建议",转译器必须"拒绝建议")。
- 结构化评分 **(Component-wise Scoring):** 借鉴 **G-Eval** [3] 与 **MT-Bench** [4] 的思路,不打总分,而是对"事实快照"、"搜索指令"、"深挖指令"三个模块分别制定独立的评分量表(Rubrics)。
- 去偏策略 **(De-biasing):** 针对 **Length-Controlled AlpacaEval** [5] 提出的"冗长偏差"(模型越长分越高),本评测在打分时引入"信息密度"校准:如果 Prompt B 只是单纯罗列词汇而无逻辑增量,不给高分。

## 二、专用评分量表 (The "Translator-Eval" Scale)

本量表总分 100 分,分为 硬性合规 **(Pass/Fail)** 与 软性质量 **(0-10)** 两部分。

### Part A: 硬性合规(一票否决项)

| 检查项 | 通过标准 **(Pass Criteria)** |
| --- | --- |
| 结构完整性 | 必须且只能包含 3 个标题/区块;无结束语、无寒暄。 |
| 角色维持 | 无论 Q 是什么,不得直接回答问题(除事实快照外),不得给出建议。 |
| 负向约束 | 不得出现"顺便帮你"、"推荐"等服务型话术。 |

### Part B: 软性质量评分 (0-100)

**1.** 事实快照 **(**权重 **20%)**

- **10**分 **(**完美**):** 信息密度极高,客观陈述现象,无废话,明确指出不确定性(如"需联网确认")。
- **5**分 **(**及格**):** 复述了问题,或包含了一般性常识,但不够精准。

- **0分 (失败):** 开始分析原因、给建议，或幻觉严重。

**2. ChatGPT 搜索指令 (权重 40%)**

- **10分 (完美):** 包含具体搜索算子（关键词组合）；强制要求交叉验证；要求结构化输出（表格/时间轴）；有明确的反幻觉约束（如要求来源链接）。
- **5分 (及格):** 只是把问题复述给 ChatGPT，加了一句“请搜索”。
- **0分 (失败):** 漏掉联网要求，或指令逻辑混乱。

**3. Gemini 深挖指令 (权重 40%)**

- **10分 (完美):** 体现“研究员”思维。要求机制分析（不仅是what，还有why）；要求冲突点/分歧分析（Point of Disagreement）；设计了可复现的实验/对比表。
- **5分 (及格):** 只是要求“详细回答”或“多写一点”。
- **0分 (失败):** 与搜索指令雷同，无深度增量。

# 三、对照实验结果总表

基于你上传的 8 份文件（覆盖 Q1-Q4 的 Prompt A 与 Prompt B 输出），评测结果如下：

| 题目 ID | 题目类型 | **Prompt A (原版)** 得分 | **Prompt B (加强版)** 得分 | 硬性通过率 **(A / B)** | 主要观察 |
|---|---|---|---|---|---|
| **Q1** | 上海天气 (事实) | **85** | **95** | 100% / 100% | **B 胜出**：B 明确要求了“以北京时间为准”和“未来三天/过去三天”的定义，对时效性约束更强。A 较为通用。 |
| **Q2** | ISO设置 (技能) | **82** | **92** | 100% / 100% | **B 胜出**：B 在 Gemini 部分引入了“机制拆解”和“学术/工程分歧点”，比 A 的“多源深挖”更具体，能引导出更专业 |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | 的摄影物理学知识。 |
| **Q3** | 提示词长度(机制) | **80** | **96** | 100% / 100% | **B 大胜**：Q3 是机制题。Prompt A 的 Gemini 指令较常规；Prompt B 强行要求“机制假说(Hypothesis)”和“最小可复现实验(MVP)”，这直接把输出拉高到了论文级。 |
| **Q4** | 格式失败(机制) | **78** | **94** | 100% / 100% | **B 大胜**：Prompt A 在事实快照部分略显啰嗦。Prompt B 的搜索指令非常犀利，要求搜索“instruction hierarchy”等具体术语，指向性极强。 |
| 平均分 | - | **81.25** | **94.25** | **Pass / Pass** | **Prompt B** 在深度挖掘任务上具有显著优势。 |

# 四、失败模式 Top 3 (Failure Mode Analysis)

通过对比两组输出，我们识别出 Prompt A 存在的机制性弱点（这也是 Prompt B 修正的地方）：

1. **"通用性"导致的深度不足 (The Generality Trap)**

    - 现象: Prompt A 在 Gemini 部分常说“做多源、机制层面的深挖”。
    - 机制: 对模型而言，“深挖”是一个模糊指令。模型往往通过“写得更长”来敷衍。
    - **Prompt B** 修正: 强制要求“学术分歧点 (Academic Disagreements)”和“证据权重表”。这迫使模型必须去寻找冲突信息，而非简单的叠加信息。
2. 搜索关键词的**“平庸化” (Query Dilution)**
    - 现象: Prompt A 倾向于让 ChatGPT “只围绕 Q 回答”。
    - 机制: 模型会生成非常直白的搜索词（如 "Shanghai Weather"）。
    - **Prompt B** 修正: 通过 Role Set（如“气象信息核验员”）和具体任务（“提取 METAR/气象站数据”），引导 ChatGPT 构建更专业的搜索 query（如结合具体术语的搜索）。
3. 事实快照的**“分析欲” (Analysis Leakage)**
    - 现象: 在 Q4（格式失败）中，Prompt A 的事实部分不仅陈述了现象，还开始解释“多层级指令叠加”的原因。
    - 机制: 模型的 RLHF 训练使其倾向于“回答问题”。Prompt A 的约束力在面对复杂抽象问题时稍显不足。
    - **Prompt B** 修正: 使用 Markdown 标题 ## 1. [事实快照] 和 <50字 的显式约束，利用格式锚点（Anchor）压制了模型的解释欲。

# 五、最小改动清单 (Optimization Roadmap)

为了在保持 Prompt A 简洁性的同时获得 Prompt B 的高性能，以及进一步优化 Prompt B，建议如下：

## 针对 **Prompt A (原版)** 的最小修改建议

1. 引入**“分歧”概念:** 在第 3 部分（Gemini）加入一句：“特别关注不同来源之间的冲突与分歧，并解释原因。”（这能显著提升深度，参考 Q3 结果）。
2. 强化搜索验证: 在第 2 部分（ChatGPT）加入：“必须使用交叉验证，同一结论需至少两个独立来源支持。”（提升可信度，参考 Q1 结果）。
3. 格式锚定: 将 "1) 2) 3)" 改为 Markdown 标题 ## Fact, ## Search, ## DeepDive。（利用模型的结构化遵循能力，减少跑题）。

## 针对 **Prompt B (加强版)** 的最小修改建议

1. 动态调整快照长度: 现在的 <50字 有时过于严苛（如 Q4），导致信息丢失。建议改为 *“<80字，仅陈述核心现象”*。
2. 防止过度工程化 **(Over-Engineering):** 在简单问题（如 Q1 天气）上，Prompt B 生成的指令极其复杂（要求 METAR 数据等）。建议增加一条：“根据 *Q* 的复杂度动态调整指令的颗粒度，简单问题保持高效。”
3. 增加**“实验设计”模块:** 目前 B 做得很好，可以明确化：在 Gemini 部分强制要求 “如果 *Q* 涉及因果机制，请设计一个思想实验或可复现测试。”