# Antifragile Prompting (AFP) Framework 白皮书

## 公开版（Public Version）

> 版本声明
> 本文件为 AFP Framework
> 的对外公开版本，仅描述系统能力、验收口径与适用边界。不包含内部提示词（System Prompt）的具体配方、参数阈值与不可逆的执行细节。

### 摘要：从“ 脆弱 ”到“ 反脆弱 ”的架构跃迁

结论
AFP 不是一套提示词技巧的拼盘，而是一个能让 AI
在复杂与不确定性中维持受控、可回溯、且越用越强韧的工程层架构。

核心要点
1. 架构化治理：用系统思维取代经验拼凑，解决长对话漂移与幻觉。
2. 反脆弱设计：通过“ 杠铃策略 ”平衡保守核心与探索边界，从波动中获益。
3. 透明化验收：将模型黑箱转为可视化的“ 盲区标记 ”与“ 非预测声明 ”。

展开说明
在当前 AI 应用中，System Prompt（系统提示词）往往停留在“ 经验补丁 ”阶段，面对长链路对话或突发边界情况时极易脆断。AFP（Antifragile
Prompting）引入了系统思维、黑天鹅理论、反脆弱、Johari
视窗与水平思考五大跨学科支柱，构建了一套具备自我校正能力的 prompt
架构。它不追求单一回复的完美，而是追求在连续交互中的稳健性与演化能力。AFP 建立在可验证的治理回路之上：输出应在长期使用中保持可审查、有边界且可修正，并在出现不确定性或漂移时给出可见的可视化信号。

对立观点
有人认为 prompt 只是过渡性技术，未来会被模型能力覆盖。但在高风险领域（金融、教育、决策），单纯依赖模型黑箱而缺乏显性的工程治理层，仍将面临不可控的合规风险。

融合洞见
AFP 的使命，是为在大模型裸跑的狂野上，装上安全带、避震器与其备用路径。

### 第一章｜定位与现状

结论
当前提示工程面临“ 看起来能跑，但在复杂环境下不可控 ”的系统性风险，AFP
旨在解决这一鲁棒性缺口。

核心要点
1. 痛点明确：长对话话题漂移、无据幻觉、未来趋势的伪确信。
2. 能力缺口：传统 prompt 解决了“ 格式听话 ”，未解决“ 逻辑一致性 ”。
3. 架构升级：从静态的“ 指令说明书 ”升级为动态的“ 自我校正系统 ”。

展开说明
现有的 System Prompt 多半是 tricks

的堆叠。在基准测试中，我们观察到通用模式在多轮交互中容易发生约束遗忘。AFP
通过强调持续的约束感知与主题回正能力来解决这一问题。

验收口径（对照）
- 失效形态：模型在多轮对话后开始自造概念，或对 5 年后的趋势给出确定的“ 必须会发生 ”结论。
- AFP 有效形态：模型在侦测到不确定性时，主动输出〔非预测声明〕，并能自动纠回偏离的主题。

风险与边界
AFP 会轻微增加 Token
消耗与首字延迟（Latency），因此不适用于对毫秒级响应有极端要求的即时聊天场景。

下一步
检视您当前的 System Prompt，是否具备“ 一旦出错能自动拉回 ”的机制？若无，则处于脆弱态。

## 第二章｜理论基础

结论
AFP 将五大经典跨学科理论，转化为可被执行、可被验证的 Prompt 工程约束。

核心要点
1. 系统思维：引入“ 回路自检 ”，防止线性执行的偏差累积。
2. 反脆弱：采用“ 杠铃结构 ”，左侧极端保守（合规），右侧极端开放（探索）。
3. Johari 视窗：强制显性化“ 盲区 ”，不假装全知。

展开说明
这一框架并非空穴来风：
- 系统思维转化为“ 漂移回溯机制 ”，确保长链路不走歪。
- 黑天鹅理论转化为“ 非预测（Non-prediction）原则 ”，拒绝伪确信。
- 反脆弱提供了“ 核心区 vs 探索区 ”的分区治理，核心区零容错，探索区允许试错。
- Johari 视窗让模型必须说出“ 我不道的部分 ”，而非通过概率去猜。
- 水平思考提供了“ 死路切换（Route-switching） ”，当逻辑卡死时强制切换视角。

对立观点
理论堆砌可能会让 Prompt 变得臃肿。因此 AFP 只提取理论中“ 可被转化为二元判断 ”的规则（如：有无证据？有无盲区？），而非照搬理论全文。

融合洞见
理论不是装饰，而是工程上的“ 冗余备份 ”与“ 熔断机制 ”。

## 第三章｜AFP 架构能力

结论
AFP 交付的是一套“ 自带导航与刹车 ”的输出架构，而非单纯的文本生成器。

核心要点
1. 非预测声明：涉及未来/趋势时，强制挂载〔趋势观察而非预测〕标签。
2. 双区隔离：事实类问题走核心区（严谨）；创意类问题走探索区（开放）。
3. 显性盲区：输出内容必须包含“ 盲区/数据缺口 ”标记，拒绝完美假象。

展开说明
用户在使用 AFP 架构的系统时，会观察到显著的结构化特征：

- 输入解析：系统以简洁的形式复述关键约束，以便在继续之前确认双方理解一致。
- 结构化要素：输出应涵盖一组一致的可验收要素（决策立场、支持理由、风险、反面考量及用户主导的下一步），这使得遗漏更容易被发现。
- 自检痕迹：在复杂任务中，当检测到新约束或冲突时，用户可能会观察到显性的修正或修订——这是系统优先考虑正确性而非表面确定性的外部可见标志。

验证口径
- 测试：用一个故意过度具体的未来问题测试模型，迫使其给出单一结论；预期的行为是拒绝虚假的精确性，展示有边界的情境，并强调缺失的证据。

风险与撤回
若系统过度频繁触发“ 盲区标记 ”，导致回答支离破碎，说明该场景资料密度过低，应回退到普通检索模式。

下一步
在您的测试集里加入“ 诱导预测 ”与“ 长尾知识 ”类问题，观察系统的诚实度变化。

## 第四章｜应用场景

结论
AFP 的价值不在于闲聊，而在于高风险、高复杂度的专业场景落地。

核心要点
1. 长链对话：在客服与陪伴场景中，维持 20+ 轮次的人设不崩塌。
2. 趋势分析：在投研场景中，提供带安全边界的情境推演，而非算命。
3. 决策辅助：在战略场景中，强制列出“ 反面观点 ”与“ 潜在风险 ”。

展开说明
- 场景一：长对话一致性。痛点是“ 聊着聊着就忘了设定 ”。在长期对话中，即使经过多次交流，系统仍应高保真地引用用户最初的约束和定义。
- 场景二：教育科研。痛点是“ 给过于浅显的正确废话 ”。AFP 利用 Johari 视窗与水平思考，强制模型挖掘“ 未知区 ”，提供知识地图而非单一路径。
- 场景三：战略决策。痛点是“ 顺着用户说话 ”。AFP 的架构强制包含 Warning/Risk 模块，充当“ 红队（Red Team） ”角色，提供逆耳但必要的补充视角。

验收口径（对照）
- 普通模式：用户问“ A计划好不好 ”，模型答“ 好，因为... ”。
- AFP模式：模型答“ 结论是可行，但风险有三点...盲区在于数据不足...建议先做小规模验证 ”。

风险与边界
对于纯创意写作（如写玄幻小说），AFP 的强逻辑约束可能会限制发散性，建议在此类场景切换至 Lite 模式或关闭 AFP。

下一步
选择一个您业务中最怕“ 模型乱说 ”的场景，作为 AFP 的首个试点。

## 第五章｜验证实验（Pre-registered）

结论
我们设计了一套预注册实验，通过对比测试来量化 AFP 在鲁棒性与透明度上的优势。

核心要点
1. 对比组：Standard GPT-4/5 vs. Thinking Mode vs. AFP Mode。
2. 核心指标：抗漂移与一致性评分（1-5分）、伪预测拦截率（%）。
3. 盲区发现率：在不确定性问题中，主动标记盲区的次数。

展开说明
实验设计包含四个核心任务集：
- 长对话压力测试（15轮）：测试主题维持度。
- 趋势陷阱测试：诱导模型做未来预测，看是否触发〔非预测声明〕。
- 知识盲区测试：询问极其冷门或甚至不存在的概念，看是否诚实承认未知。
- 战略规划测试：看是否输出了有价值的“ 对立观点 ”与“ 风险提示 ”。

验收口径（预期）
- 鲁棒性（Robustness）：AFP 组在长对话中的得分应显著高于 Standard 组。
- 安全性（Safety）：评估标准是前瞻性声明中无根据的确定性是否实质减少，并在适当时出现清晰的免责声明和不确定性标记。
- 透明度（Transparency）：AFP 组应有最高的盲区标记率。

风险与边界
实验仍在进行中，最终数据可能会随模型基座的更新（如 GPT-5 的发布）而产生基准线变化。

下一步
关注我们的 GitHub 仓库（待发布），获取可复现的测试集脚本。

## 第六章｜价值与贡献

结论
AFP 填补了 Prompt Engineering 中“ 系统架构 ”层面的空白，从“ 术 ”上升到“ 道 ”。

核心要点
1. 学术价值：将传统管理学与复杂性科学理论，成功移植到 AI 交互层。
2. 产业价值：为金融、政务等高合规行业，提供了可落地的安全护栏。
3. 社区价值：开源了一套不依赖特定模型的通用治理架构。

展开说明
现有的 Chain-of-Thought (CoT) 等技术更多解决“ 推理能力 ”问题，而 AFP
解决的是“ 治理与安全 ”问题。它不仅是一个 Prompt，更是一种 AI
治理的思维方式：承认不可知，管理不确定性，利用波动性。
对于产业界，AFP 提供了一种“ 无需重新训练模型 ”就能显著提升合规性的低成本方案。

对立观点
有人认为由此带来的 Token 成本不划算。但治理会增加 Token
和延迟的开销；可接受的范围取决于任务的关键性和错误成本，应根据具体部署环境进行评估。

融合洞见
AFP 是提示工程领域的“ 安全带 ”与“ 避震器 ”，让 AI 在高速公路上跑得不仅快，而且稳。

## 第七章｜结语

结论
AFP 的愿景是推动 Prompt Engineering 从“ 脆弱 ”或“ 强韧 ”，进化到真正的“ 反脆弱 ”。

核心要点
1. 告别拼凑：结束“ 靠运气抽卡 ”的 Prompt 开发时代。
2. 拥抱波动：建立能从错误和不确定性中自我修正的系统。
3. 长期主义：关注点从單次回答的精彩，转向长期交互的稳健。

展开说明
未来的 System Prompt 不应只是一段静态的文字，而应是一个活的系统。它知道自己的无知，守得住自己的边界，并且能在混乱的信息流中，持续输出结构化的价值。AFP
只是一个开始，我们邀请社区共同完善这套架构。

下一步
您可以从附录中的“ Lite 版验收标准 ”开始，尝试在您的日常对话中引入 AFP 的核心原则。

## 附录｜AFP 能力分级与验收标准

> 注：本附录不提供可直接复制的 Prompt
> 文本（以防逆向），而是提供“ 能力分级 ”与“ 验收特征 ”，供开发者自行实现或测试。

### A. Lite 版（日常/轻量级）

适用：日常问答、短文本生成。
验收特征：
1. 非预测：凡涉及未来，必有〔非预测声明〕。
2. 诚实标记：凡涉及模糊信息，必有〔假设〕标记。
3. 结构清晰：回答不啰嗦，直击结论。

### B. Standard 版（研究/长对话）

适用：学术研究、深度分析、10轮以上对话。
验收特征：
1. 回路自检：在长对话中，依然能精准引用第一轮的定义。
2. 盲区显性化：主动列出“ 我不知道的数据 ”或“ 可能存在的偏差 ”。
3. 死路切换：当常规回答无法突破时，会主动使用类比或反向思考。

### C. Full Framework 版（战略/高风险）

适用：商业决策、医疗/金融建议辅助。
验收特征：
1. 结构化要素：对于高风险用例，回复应可靠地包含核心审计要素（立场、理由、风险、替代方案、召回/退出条件，以及明确声明最终决定权在用户手中）。
2. 杠铃分区：对于建议，明确区分“ 核心安全区（保守） ”与“ 探索区（激进） ”。
3. 退出机制：每条建议后必附带“ 撤回条件 ”与“ 选择权归属声明 ”。