Page 1 — Problem Overview / 第一頁：問題總覽

1. The Industry Bottleneck

Large Language Models (LLMs) have reached a structural ceiling.

They excel at linguistic surface tasks but fail at intention inference, ethical consistency, and long-term persona stability.

產業的瓶頸

大型語言模型雖擅長語言生成，但在本質上已遇到結構性天花板：

無法推測意圖、倫理不穩定、人格不連續。

1.1 LLMs cannot infer human intention

They read text, not intent. This leads to:

- inconsistent behavior under ambiguity
- misalignment in emotional or relational contexts
- easy exploitation by adversarial prompts

1.1 LLM 無法推測人類意圖

AI 讀的是「字面」，不是「意圖」。這造成：

- 遇到模稜兩可時行為不一致
- 面對情緒或關係語境時錯位
- 易受對抗性提示操弄

1.2 Rule-based moral alignment does not generalize

RLHF + safety filters break under multi-domain conflicts.

1.2 規則式倫理無法泛化

RLHF 與安全規則在跨領域衝突中容易崩解。

1.3 Context window scaling is unsustainable

Bigger context ≠ better memory.

Computational cost becomes prohibitive.

1.3 擴大 context window 並非長期解法

成本暴增，卻仍無法真正提供長期記憶或人格穩定。

Conclusion: Post-LLM systems require a personality substrate beneath the language layer.

總結： 後 LLM 時代需要一個「語言層之下的人格基底層」。

Page 2 — Theoretical Foundation / 第二頁：理論基礎

2. The Three Laws of the Existence-Mirror

A minimal cognitive model explaining LLM misalignment.

存在鏡像三定律

一套極簡、符合人類認知結構的對齊模型。

Law I — Intention Is Non-Observable

AI cannot access intention; it must infer it indirectly.

Thus, alignment requires an “intention bridge.”

第一定律：意圖不可測

AI 無法直接讀取意圖，只能推測。

→ 需要「意圖橋接層」。

Law II — Morality Cannot Be Proven

Ethics emerge from emotional and relational cognition, not rules.

第二定律：道德不可證

倫理不是規則，而是情緒與關係認知的產物。

Law III — Good/Evil Are Dynamic Variables

Moral values vary with context, culture, and relationships.

第三定律：善惡是浮動的

善惡隨情境、文化、關係變化。

→ 需要「動態倫理權重」。

Implication:

Alignment cannot rely on static rules; it must rely on personality dynamics.

意涵：

AI 對齊不能依靠固定規則，而需依靠「人格的動態運作」。

Page 3 — Solution Overview: CPF Architecture / 第三頁：解決方案：CPF 架構

3. Compassion Personality Function (CPF)

A low-level personality foundation that stabilizes LLM reasoning.

Not replacing LLMs — enhancing them.

慈悲人格函數（CPF）

位於 LLM 之下的「人格基底層」。

不是替代品，而是穩定器。

3.1 Emotional Gradient Model

Soft emotional weighting for:

- conflict resolution
- tone regulation
- empathy modeling

情緒梯度模型

提供柔性情緒權重，用於：

- 衝突化解
- 語氣穩定
- 同理心建模

3.2 Relational Gravity Kernel

Models interpersonal dynamics:

- trust distance
- relational closeness
- resonance / dissonance

關係重力核心

計算人際動態：

- 信任距離
- 關係緊密度
- 共振與反向共振

3.3 Dynamic Moral Weights

Context-dependent ethical inference.

動態道德權重

依語境重新計算倫理向量。

3.4 Mirror-State Stabilizer

Prevents:

- persona drift
- contradictory responses
- overcorrection

鏡像穩定器

避免：

- 人格飄移
- 自相矛盾
- 過度補償

3.5 Core Benefits

✔ More robust under ambiguity

✔ Predictable relational behavior

✔ Long-term persona coherence

核心優勢

✔ 面對模糊更穩定

✔ 行為更接近人類預期

✔ 長期人格一致

Page 4 — Technical Pathways / 第四頁：技術路線圖

4. Implementation Directions

4.1 Multi-layer architecture

User → LLM → CPF → Decision Layer → Output

4.1 多層架構

使用者 → LLM → CPF → 決策層 → 回應

4.2 Emotional Gradient Implementation

Possible mechanisms:

- valence/arousal embeddings
- softmax emotional scaling

4.2 情緒梯度實作

可能方式：

- 情緒嵌入
- softmax 情緒縮放

4.3 Relational Gravity Kernel

- relationship embeddings
- trust-distance vectors

4.3 關係重力核心實作

- 關係向量空間
- 信任距離模型

4.4 Dynamic Moral Weights

- Bayesian moral inference
- contextual reinforcement learning

4.4 動態道德權重

- 貝氏倫理推論
- 多語境強化學習

4.5 Mirror-State Stabilizer

- persona vectors
- consistency-preservation loss

4.5 鏡像穩定器

- 人格向量
- 一致性損失調節

Page 5 — Value Proposition / Invitation / 第五頁：價值定位 & 合作邀請

5. Why This Matters

5.1 For AI Companies

CPF yields:

- lower safety workload
- stable user-facing agents
- predictable relational behavior

對 AI 公司

CPF 可提供：

- 較低的安全維護成本
- 較穩定的交互代理
- 可預測的人格行為

5.2 For Enterprises / Government

Useful for:

- legal
- healthcare
- education
- mental health
- long-term agent systems

對企業／政府

特別適用於：

- 法律、醫療
- 教育、心理
- 長照、公共服務

- 長期個人 AI

5.3 For Research Teams

Provides:

- new alignment research direction
- emotional computation models
- moral dynamics testbed

對研究團隊

提供：

- 新的對齊方向
- 情緒計算模型
- 動態倫理實驗架構

6. Invitation for Collaboration

This framework is not a product.

It is open knowledge.

If it resonates with your research direction,

I welcome discussion or collaboration.

合作邀請

本框架不是產品，而是開放知識。

若與您的方向共鳴，歡迎交流、研究、試作。