許文耀 Wen-Yao Hsu (Yao Shen)

語意防火牆系統創建者 | AI 審計與語義治理架構設計者

Founder of Semantic Firewall System (SRCP) | AI Governance Layer Designer

📍 Taichung, Taiwan | 📧 ken0963521@gmail.com

🔗 GitHub: https://github.com/HIJO790401 | 🔗 LinkedIn: https://www.linkedin.com/in/yao-shen-150ab93b2

個人總覽 (Profile Summary)

我致力於在大型語言模型（LLM）前端建立一套具備「治理重力」的系統層。這套系統並非單純的聊天機器人或傳統的詐騙分類器，其核心使命是將「主體、因果、邊界、依據、責任」重新導入 AI 的判斷邏輯與現實決策鏈中。這是一個**決策前結構審計**的框架，確保資訊在進入人類決策鏈前，已具備可驗證的責任結構。

我提出的核心方向包括 **Semantic Responsibility Chain Protocol (SRCP)**、**SCBKR 責任鏈**、**Semantic Firewall System**、**R-Lock** 與 **VOID Engine**。這些框架旨在解決當前 AI 輸出中常見的語義漂移、假中立與責任缺失問題，並提供語義審計、責任映射、語義穩定性與跨模型一致性的解決方案。

我的核心主張是：**「我不是在判斷一段訊息像不像詐騙，我是在判斷它有沒有資格進入人的決策鏈。」**

核心專案 (Core Projects)

1. 基礎語意防火牆 / Semantic Firewall System

- **定位**：作為基礎技術展示版，此專案旨在驗證語意污染、假中立、責任鏈缺失等問題，並展示 SCBKR 結構審計的可行性。
- **核心要點**：
  - 透過語意責任鏈拆解輸入訊息或模型輸出，精準識別並處理幻覺（Hallucination）、主體錯位（Subject Misalignment）與語義污染（Semantic Contamination）。
  - 強調系統的可審計性（Auditability）與可回放性（Replayability），確保每次判斷皆有清晰的邏輯路徑。
  - 為後續的反詐騙語意防火牆與 vπ 系列治理引擎奠定堅實的骨架基礎。
- **連結**：https://hijo790401.github.io/semantic-firewall-system/

2. 反詐騙語意防火牆 / Anti-Scam Semantic Firewall

- **定位**：將語意責任鏈直接應用於反詐騙治理場景，判斷訊息是否具備進入人類決策鏈的資格。
- **核心要點**：
  - 當訊息的主體、邊界、依據或責任無法成立時，系統將判定該訊息不可直接被信任，從而有效阻斷詐騙訊息進入使用者的決策流程。
  - 實戰模擬假銀行通知、假政府更新、假物流驗證、模糊付款訊息等常見詐騙情境，並提供結構化的審計解釋。
  - 此系統特別適用於一般民眾、長者保護，以及金融風控與政府治理等高風險應用場景。
- **連結**：https://hijo790401.github.io/anti-scam-semantic-firewall/

3. 沈耀語意防火牆 vπ10 / Shen-Yao Semantic Firewall vπ10

- **定位**：作為進階治理展示包，此專案展示了語意防火牆如何作為掛載於 LLM 前端的治理層運作，提供全面的安全與審計機制。
- **核心要點**：
  - 在 LLM 前端實施安全鎖、法律規則與可審計紀錄，確保模型輸出在進入決策鏈前符合預設的治理標準，不直接依賴模型的黑盒判斷。
  - 整合 LAW View（法律視角）、治理狀態監控、管理後台、授權入口與工程 API，提供完整的治理生態。
  - 展示 Trial-Audit、Proxy-Chat、SCBKR 審計卡、法律宣言與治理規則，強調短效 Session Token、審計結果、責任鏈與治理聲明的透明化。
- **連結**：https://shen-yao-vpi9-ui.onrender.com/index.html

媒體與公開證據 (Media & Public Evidence)

- **SecurityBrief Asia 國際報導**：
  - *“Semantic Firewall promises AI cost savings & safer chat models”*
  - 此報導深入探討了語意防火牆在降低 AI 算力浪費與提升模型安全性方面的巨大潛力。這不僅證明了此系統的技術前瞻性，也代表其已獲得國際科技媒體的關注與外部能見度。
  - **報導連結**：https://securitybrief.asia/story/semantic-firewall-promises-ai-cost-savings-safer-chat-models

---

技術焦點與能力 (Technical Focus & Skills)

治理與審計 (Governance & Audit)

- Semantic Audit (語義審計)
- Risk Governance (風險治理)
- Decision-chain Validation (決策鏈驗證)
- Auditable Logs (可審計日誌)

語義結構 (Semantic Structures)

- SRCP (Semantic Responsibility Chain Protocol)
- SCBKR Framework (SCBKR 責任鏈框架)
- Responsibility Mapping (責任映射)
- Semantic Stability (語義穩定性)

風險與安全系統 (Risk / Safety Systems)

- Anti-Scam Semantic Analysis (反詐騙語義分析)
- LLM Governance Wrapper (LLM 治理封裝層)
- Deterministic Rule Layer (確定性規則層)

系統架構 (System Architecture)

- Explainable Structure (可解釋結構)
- Cross-model Consistency (跨模型一致性)
- AI Governance Layer Design (AI 治理層設計)

核心方法與理念 (Core Methodology & Positioning)

我的核心方法論聚焦於**決策前結構審計**。這與傳統的 LLM Wrapper 僅對模型輸出進行包裝截然不同。我主張在模型前端即建立一個強大的治理規則層與責任鏈邏輯，確保所有進入決策鏈的資訊都具備結構上的完整性與責任的有效性。

我深信**結構勝於機率**。因此，我的系統不只關心答案在形式上是否正確，更深入探究訊息或回應是否具備可驗證的主體、邊界、依據與責任。這使得系統能夠判斷資訊的「現實資格」，而非僅僅是「表面相似度」。

聯絡資訊 (Contact & Links)

- **Email**: ken0963521@gmail.com
- **GitHub**: https://github.com/HIJO790401

- **LinkedIn**: https://www.linkedin.com/in/yao-shen-150ab93b2
- **所在地**: Taichung, Taiwan