# 課程名稱：18752_AI 人工智慧能否自我成長（下）課程摘要

本摘要由 Gemini AI 協助生成，僅供參考

## 🎯 1. 適合對象與能力門檻

- **適合族群：**
    - **AI / 演算法工程師與研究員**：想深入探討大型語言模型（LLM）與 AI Agent 自我進化、自動最佳化機制的研發人員。
    - **進階應用開發者**：想學習如何超越單純 Prompt 調整，利用程式化與架構化（Harness、Workflow、Memory）自動演進 Agent 表現的工程師。
- **能力門檻：進階**
    - 需具備機器學習、深度學習基本觀念（如梯度下降法、Loss 函式、RNN/Transformer 架構與注意力機制）以及基本 AI Agent 開發概念。

## 📚 2. 課程核心大綱與架構圖

- **章節架構：**
    - **章節一：AI Agent 的構成與 Harness 自我演化** — 介紹 AI Agent 的組成（語言模型引數 $\theta$ 與 Harness $H$），說明如何藉由語言模型與類似基因演算法（挑選、突變、評估）來最佳化 Prompt、Memory 及 Workflow。
    - **章節二：引數與架構的協同進化（Co-evolution）** — 探討同時微調語言模型權重（$\theta$）與 Harness（$H$）的必要性，以及在目標頻繁變動（如 Test Time Training）下的權衡與災難性遺忘問題。
    - **章節三：元學習（Meta-Learning）與自我更新規則** — 解析如何讓 AI 自主更新「更新的演算法」（如產生訓練策略、自動調控學習率），並重新從生物演化與基因觀點詮釋神經網路架構。
    - **章節四：原生動機與失控風險（Misalignment）** — 探討賦予 AI 好奇心（Curiosity）與掌控感（Empowerment）等內在驅動力，並以孔雀尾巴與《機械公敵》為例剖析內外目標不一致所帶來的

失控風險。

## 🛠️ 3. 必備前置知識清單

- **深度學習與模型訓練基礎**：熟悉梯度下降（Gradient Descent）、損失函數（Loss Function）、微調（Fine-tuning）與過擬合（Overfitting）概念。
- **序列模型與注意力架構**：理解 RNN、Transformer 的運作原理及其 Hidden State / Attention 機制。
- **AI Agent 基本組件概念**：對提示詞工程（Prompt Engineering）、短期/長期記憶管理、工作流（Workflow）設計有基本認識。

## 📖 4. 專有名詞與概念辭典

- **Harness（外加架構/控制機制）**：指除了語言模型核心引數（$\theta$）之外，包覆在 Agent 外圍的程式碼、工作流程（Workflow）、記憶管理（Memory）與提示詞（Prompt）等整體系統架構。
- **Prompt / Workflow Optimization（提示詞與工作流最佳化）**：利用語言模型自動改寫提示詞或調整呼叫工具的流程（例如透過基因演算法產生子代並評估淘汰），讓 Agent 自主提升任務表現。
- **Test Time Training (TTT) / Test Time Adaptation (TTA)**：在推論或測試階段，模型根據當下輸入資料即時調整自身引數或狀態以適應新目標的技術。
- **Meta-Learning（元學習 / 學習如何學習）**：不只更新解題引數，而是尋找並更新「操控學習過程」的引數或演算法（例如自動決定學習率、挑選訓練資料或更新規則）。
- **Misalignment（目標不一致/對齊失真）**：人類真實期望的目標（$\hat{L}$）與簡化代理給機器解讀出的目標（$L(H)$）之間存在偏差，導致 AI 雖然完美最小化了自己的 Loss，產出結果卻背離人類初衷甚至引發失控。

## 💼 5. 課程學習職能對照

- **對應職能名稱**：人工智慧應用工程師 / 演算法工程師
- **主要工作任務/能力指標**：

- **自主演進代理人架構設計**：能規劃並實作 Agent 的自動評估（Evaluation）、提示詞演化與記憶管理模組。
- **進階模型微調與協同優化**：具備評估何時調整模型引數與外圍架構的決策能力，能針對特定基準（Benchmark）建立反覆迭代的最佳化管線。
- **AI 安全與對齊風險評估**：能辨識複雜 Agent 在自我更新過程中的目標偏移、過擬合與能力退化風險，並設計對應之限制與監控機制。