

$$1. \text{ Forward SDE: } dx_t = f(x_t, t) + g(x_t, t) dW_t$$

The evolution of the PDF  $p(x, t)$  corresponding to this SDE is governed by the Fokker-Planck equation:  $\frac{\partial p}{\partial t} = -\frac{\partial}{\partial x} [f(x, t)p(x, t)] + \frac{1}{2} \frac{\partial^2}{\partial x^2} [g^2(x, t)p(x, t)]$

The Probability Flow ODE describes a deterministic trajectory that evolves in such a way that the collection of these deterministic trajectories yields the exact same PDF  $p(x, t)$  at all time as the original SDE

Assume the Probability Flow ODE has the form:  $dx_t = \mu(x_t, t) dt$

It follows that  $\frac{\partial p}{\partial t} = -\frac{\partial}{\partial x} (\mu p)$ .

$$\text{We have } -\frac{\partial}{\partial x} (\mu p) = -\frac{\partial}{\partial x} (fp) + \frac{1}{2} \frac{\partial^2}{\partial x^2} (g^2 p)$$

$$\frac{\partial}{\partial x} (\mu p) = \frac{\partial}{\partial x} \left[ fp - \frac{1}{2} \frac{\partial}{\partial x} (g^2 p) \right]$$

Assuming proper boundary conditions, we have  $\mu p = fp - \frac{1}{2} \frac{\partial}{\partial x} (g^2 p)$

$$\text{That is } \mu = f - \frac{1}{2p} \frac{\partial}{\partial x} (g^2 p) = f - \frac{1}{2p} \left( \frac{\partial g^2}{\partial x} p + g^2 \frac{\partial p}{\partial x} \right)$$

$$\begin{aligned} &= f - \frac{1}{2} \frac{\partial g^2}{\partial x} - \frac{g^2}{2p} \frac{\partial p}{\partial x} && \text{recall } \frac{1}{p} \frac{\partial p}{\partial x} = \frac{\partial \log p}{\partial x} \\ &= f - \frac{1}{2} \frac{\partial g^2}{\partial x} - \frac{g^2}{2} \frac{\partial \log p}{\partial x} \end{aligned}$$

$$\text{Therefore, } dx_t = \left[ f(x_t, t) - \frac{1}{2} \frac{\partial}{\partial x} g^2(x_t, t) - \frac{g^2(x_t, t)}{2} \frac{\partial}{\partial x} \log p(x_t, t) \right] dt$$

# AI 的未來與機器學習的基石

註：部分是我自己寫的，部分是AI生成。

## 1. AI 的未來能力：獨立創作優秀的文學作品

由於我很喜歡看小說，每每看完一部喜歡的作品後，總會意猶未盡，想再多看一些該作者的作品，或者風格相似的作品，但經常找不到。如果有一天 AI 能夠創作優秀的文學作品，對我而言意義重大。以下是我問 AI：「有一天 AI 能獨立創作優秀的文學作品，對人類、社會、科學或文化的重大意義是甚麼？」的回覆。

重新定義藝術本質，傳統「創作主體性」概念將被解構。

社會風險面是演算法可能壟斷話語權，形成新型文化霸權。

科學領域將獲得突破性研究工具，透過分析 AI 生成的海量文本，語言學可驗證深層語法結構，認知科學則能探索意識與創造力的量子特徵。

最深刻的衝擊在哲學層次，當機器作品引發人類情感共鳴，將迫使我們重新協商「何以為人」的根本命題，這種存在主義危機可能催生文明的新覺醒。

## 2. 涉及的機器學習類型

我覺得這是一個很困難的問題，莎士比亞說過：一千個人眼中就有一千個哈姆雷特。如何定義所謂“優秀”的文學作品，怎麼量化所謂的文學性或藝術性這類偏主觀的感受。其次，以小說為例，如何設計全書架構大綱、劇情推動的節奏、人物的刻畫...想完成一部作品需要考慮的因素太多了，於是求助 AI，讓 AI 來分析一下。

「優秀」的定義：在 AI 領域，這類主觀感受難以直接量化，因此主要通過人類回饋強化學習 (RLHF) 來間接校準。AI 通過訓練一個獎勵模型 (Reward Model) 來學習人類讀者對全書架構、劇情推動節奏、人物刻畫等主觀要素的偏好，從而將讀者的「滿意度」轉化為可計算的獎勵訊號。

因此，這個能力的實現主要涉及非監督式學習、監督式學習，以及強化學習 (RLHF) 的組合：

非監督式學習：構建強大的語言基礎（語義、文法、風格模式）。

監督式學習：微調模型以掌握特定文學體裁、修辭技巧或特定作家的風格。

強化學習 (RLHF)：校準模型的輸出，使其符合人類對「優秀」的審美標準。這是決定作品藝術性的關鍵步驟。

數據與回饋機制

資料來源：

非監督/監督：海量的、經人工篩選和標註的高質量文本資料庫（如獲獎小說、經典文學）。

強化學習：人類讀者對 AI 生成多個版本作品進行偏好排序或評分的數據集。

目標訊號：

基礎模型：預測下一個詞或句子（語言連貫性）。

強化學習：人類偏好轉化而來的獎勵訊號（文學性）。

學習回饋與環境互動：存在。人類讀者對 AI 作品的「偏好回饋」（例如「這部作品的節奏比上一部好」）即是環境互動，用於訓練獎勵模型，直接指導 AI 的生成策略。

### 3. 「簡化模型問題 (model problem)」

簡化模型問題：「模仿特定大師風格與情感的短篇續寫」這個簡化問題的概念是：在極端受限的上下文（特定作家、特定情感基調）中，要求 AI 展現出對高質量風格的精準掌握與細微情感的準確傳遞。

這代表了最終能力所需的兩大核心：風格掌握和情感深度。

可測試性：

1. A/B 盲測：將 AI 繽寫的短篇與人類專家（例如，模仿該作家的研究生）的續寫匿名混合。

2. 專家評分：請數名獨立的文學評論家或專業讀者對風格相似度、情節連貫性、情感氛圍的維持進行 1-5 分的主觀評分。若 AI 繽寫的平均得分能與人類專家持平或更高，則視為成功。

所需工具：

機器學習工具：深度學習中的 Transformer 架構 (如 GPT-4)，利用監督式學習對該作家的全套作品進行微調 (Fine-tuning)，並利用 RLHF 根據專家的評分回饋來校準輸出。

數學工具：評估風格相似度時，可使用基於統計的文本度量（如 n-gram 相似性）或潛在空間嵌入距離（如 L<sub>2</sub> 範數）來輔助客觀分析。

3

在「模仿特定大師風格的短篇續寫」這個簡化模型中，有沒有可能量化所謂的風格、情感深度，讓機器知道什麼是“好”