書面報告

日期:2025/11/11

講者:欉振坤

題目: 生成式人工智慧與異質平台整合應用

心得報告

本場演講從生成式 AI 的技術演變談起，先對比早期以規則導向的做法，與今日用統計與演算法找特徵的方法，進而把生成式模型分成顯式與隱式兩類。後者如 GAN 生成器把雜訊變資料，辨別器判斷真偽，兩者交替訓練；而以 Transformer 為基礎的大型語言模型，應用則在語言與理解、生成與推理上。

RLHF 先以示範資料進行監督微調，再以標註者的偏好訓練獎勵模型，最後用強化學習優化策略。透過「我是機器人」的對比範例可以看到，模型在用語、禮貌與情境上都明顯貼近人類，能把人類偏好引入訓練流程。

在《Trustworthy Machine Learning》中提到能信任模型的門檻，包括資料與模型效率、對擾動的穩健性、公平性與隱私保障、系統資安與驗證機制、可解釋與可責任追溯，以及自適應性。

生成式 AI 的一般化導入成本高，主要來自訓練、資料蒐整與系統調校。然而在正確的治理與流程設計下，GAI 能作為資源協調中樞，整合多項智慧服務完成任務，並在製造情境落實如 BOM 生成、急單或插單的生產排程與報價規劃。

關鍵字

生成式 AI、GAN、Transformer、RLHF

參考文獻

[1] **I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al.**, “Generative Adversarial Nets,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Montreal, QC, Canada, 2014, pp. 2672–2680. [Online]. Available: https://arxiv.org/abs/1406.2661

[2] **J. Maan**, “Deep Learning-driven Explainable AI using Generative Adversarial Network (GAN),” in *Proc. 2022 IEEE 19th India Council Int. Conf. (INDICON)*, Kochi, India, 2022, pp. 1–5. doi: 10.1109/INDICON56171.2022.10039793