Lecture 10:
Responsible AI: Trust, Safety, Privacy & Biased in MM

Lecture 10 (Responsible AI: Trust, Safety, Privacy & Biased in MM)

P10-1: **Hallucination:** **Presenter: Cao Xiao; Asker: Chen Xihao**

(SOTA) S Dhuliawala, et al. Chain-of-Verification Reduces Hallucination in LLMs. arXiv 2023.

(Must-Read) P Manakul, et al. Selfcheckgpt: Zero-resource black-box hallucination detection for generative LLMs. Preprint arXiv 2023.

(BG) Y Zhang, et al. Siren's Song in the AI Ocean: A Survey on Hallucination in LLMs. arXiv 2023.

P10-2: **Privacy:** **Presenter: Dai Yuhe; Asker: Lin Xinyu**

(SOTA) S Kim, et al. Propile: Probing privacy leakage in LLMs. arXiv 2023.

(Must-Read) J Huang, et al. Are Large Pre-Trained Language Models Leaking Your Personal Information? ACL 2022.

(Background) H Shao, et al. Quantifying Association Capabilities of LLMs and Its Implications on Privacy Leakage. EACL 2023.

P10-3: **Bias:** **Presenter: Yannis Mohamed Christian Montreuil; Asker: Mehdi Yamini**

(Must-Read) P Schramowski, M et al. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. CVPR 2023.

(Must-Read) Q Li, et al. Be causal: De-biasing Social Network Confounding in Recommendation. ACM TKDD 2023.

(To-Read) A S Luccioni, et al. Stable bias: Analyzing societal representations in diffusion models. arXiv 2023.

Lecture 11 (Multimodal Event Detection & Forecasting)

P11-1: Multimodal Event Detection: Presenter: Sui Yuan; Asker: Sun Pengzhan

(SOTA) M. Li, et al. Clip-event: Connecting text and images with event structures. CVPR 2022.

(To-Read) Li Z, et al. Constructing Narrative Event Evolutionary Graph for Script Event Prediction. IJCAI 2018.

(Must-Read) T Zhang, et al. Improving event extraction via multimodal integration. ACM MM 2017.

P11-2: Multimodal Fashion Forecasting: (Invited Speaker: Ma Yunshan)

(Must-Read) U Mall, et al. Geostyle: Discovering fashion trends and events. ICCV 2019.

(SOTA) Hsiao W L, Grauman K. From culture to clothing: Discovering the world events behind a century of fashion images. ICCV 2021.

(Must-Read) Ma Y, et al. Who, where, and what to wear? extracting fashion knowledge from social media. ACM MM 2019.

Submission of BNI Papers

- BNI Papers due: 5 Apr @ 1700
  - via Submit-BNI site.
- Presentation of BNI Papers: 9 Apr & 23 Apr
  - 5-minute presentation to class
  - Presentations: 7 on 9 Apr & 13 on 23 Apr
    * Submit your ppt via Submit-BNI-ppt
    * Presentation timing to be assigned randomly
  - Class evaluation: To send me your top 5 groups (without ranking)