Yifei Li

Ph.D. Student · University of Pennsylvania

liyifei@seas.upenn.edu | realliyifei.github.io | realliyifei | realliyifei | @realliyifei | Google Scholar

***Domains**: Large Language Model, Multimodal, and Domain Knowledge (e.g. Medicine, Drug)*

Education

University of Pennsylvania (UPenn) — *Philadelphia, PA, USA*

Ph.D. in Computer and Information Science (Earned M.S.E. in Data Science); Advised by Prof. Mark Yatskar — *GPA: 3.97/4.0 | 2028**

Oklahoma State University (OSU) — *Stillwater, OK, USA*

B.S. in Computer Science, B.S. in Mathematics, and B.S.B.A. in Management — *GPA: 3.8/4.0 | 2019*

Sun Yat-Sen University (SYSU) — *Guangzhou, China*

B.M. in Management *(Notes: International joint-degree program associated with OSU)* — *GPA: 3.8/4.0 | 2019*

Publications

Papers / Manuscripts [1]

Conceptor-Aided Debiasing of Large Language Models
Yifei Li, Lyle Ungar, João Sedoc
Empirical Methods in Natural Language Processing (EMNLP)
URL: `https://arxiv.org/abs/2211.11087` 2023

Skills

Coding Languages	Python, Java, MATLAB, R, C++, SQL, JavaScript, CSS, PHP
Libraries and Tools	PyTorch, Hugging Face, Faiss, Spark, Shell, Git, LaTeX
AI Models	GPT / Llama family, CLIP, BERT family, Stable Diffusion, MoCo, YOLO, SOLO, GAN family

Research Projects

(Human) Feedback Bottleneck Model on Domain Knowledge — *UPenn*

Advised by Prof. Mark Yatskar — *Aug. 2023 - Present*

- Exploring to leverage human feedback to inject domain knowledge (e.g. medicine) to large language models and calibrate the model reasoning ability via human-in-the-loop; then use bottleneck model to improve the efficiency of this process...

Conceptor-Aided Gender Debiasing of Large Language Models — *UPenn*

Advised by Prof. João Sedoc and Prof. Lyle Ungar - Paper — *Jan. 2022 - Jul. 2022*

- Use conceptors–a soft projection method–to identify and remove the bias subspace in contextual embeddings in BERT and GPT and reach SOTA performance. Two methods of applying conceptors are proposed: (1) bias subspace projection by post-processing; and (2) a new architecture, conceptor-intervened BERT (CI-BERT), which explicitly incorporates the conceptor projection into all layers during continued training.
- Show the optimal conceptor pipeline setting w.r.t. corpora, wordlists, and subspaces, the robustness of conceptor debiasing in different LLMs and layers, and the efficiency of intersectional debiasing via conceptor logical operations.

Neuro-Symbolic Dual-System on Task-Oriented Dialogue Generation — *UPenn*

Advised by Prof. Chris Callison-Burch and Dr. Lara Martin (Research Course Project) - Report — *Mar. 2022 - May. 2022*

- Adapt novel neuro-symbolic dual-system to improve the consistency and coherence in task-oriented dialog generation.
- Build a user belief states to ground human knowledge and domain-specific constraints, then fine-tune a GPT3 to generate utterance and another GPT3 to verify the consistency with belief states as a symbolic parser, repeat this process until it is consistent then update the belief states.

Improving Text-to-image Diffusion Generation Via Large Language Models — *UPenn*

Advised by Prof. Chris Callison-Burch and Prof. Mark Yatskar (Master Thesis) - Report — *Aug. 2022 - May. 2023*

- Explore imagine-then-verbalize approach that leverages the imaginative abilities of language models such as GPT to provide additional details and contexts that enhance the persuasiveness of the descriptions.
- Propose sketch-then-draw method that utilizes the coding capacity of language model to generate SVG code as sketch for downstream diffusion generation, leading better numerical consistency.

Probing CLIP Zero-Shot Ability *UPenn*

Advised by Prof. Mark Yatskar (Independent Study) *Jan. 2022 - May. 2022*

- Test and evaluate the zero-shot ability of CLIP model, figure out that CLIP performs poor on fine-grained dataset e.g. iNaturalist.
- Try to build a language utility to help users ask the right questions to CLIP: leverages the sentence-BERT to cluster different types of descriptions from the web, then compares and ranks their CLIP similarity scores.

Gender Bias on Self-Supervised Learning *UPenn*

Advised by Prof. Mark Yatskar and Prof. Vicente Ordóñez *Aug. 2022 - May. 2023*

- Find that models even pretrained on non-human images would lead to gender bias after fine-tuning, due to the implicit confounders.
- Try to debias by finding and removing the biased training images by (1) tracing the gradient between model pretraining and gender classifier finetinuning following the TracIn method (2) computing the nearest neighbor similarity of embeddings.

Professional Activities

2022-23 **EMNLP,** Reviewer

2022-23 Fa **CIS 5300 Computational Linguistics,** Teaching Assistant (Create assignment; supervise course projects) *UPenn*

2022 Sp **CIS 522 Deep Learning,** Teaching Assistant (Lead recitation pods - notes; supervise course projects) *UPenn*

2021 Fa **CIS 520 Machine Learning,** Teaching Assistant (Lead recitation pods - notes; supervise course projects) *UPenn*

Awards & Honors

2023 **Outstanding Teaching Award,** Significant contributions as teaching assistant in courses *UPenn*

2016-19 **President's Honor Roll,** Maintain Excellent GPA *OSU*

2018 **Emeritus Math Faculty Scholarship,** Mathematics Department, 1-2 Student(s) Each Year *OSU*

SST Scholarship, Computer Science Department *OSU*

2016 **Transfer Out-Of-State Achievement,** Top 15% GPA *OSU*

Course Projects

Two-Stage Summarization with Pre-Trained Transformers *UPenn*

Coursework - GitHub, Report *2021*

- Present a two-stage summarization model using pretrained transformers, concatenatingan extractive model BERTSumEXT and an abstractive model, GPT2 or BART. By feeding the abstractor the extracted key information, this method can mitigate the disadvantages of both approaches and make the summaries more readable. It also reduces the need to truncate sentences for abstractive model due to the maximum token limit.

NLP and Text-to-Image Generation for Gameplaying: Steins;Gate *UPenn*

Coursework (funny) - GitHub *2022*

- Build an interactive textual game powered by NLP and text-to-image GAN where the storyline would change based on the procedures controlled by the player. Here, the storyline is represented as graph, the interactive texts are generated by GPT3, and the cutscenes are rendered by pixray.

SOLO and GRU for Hemostatic Plug Segmentation *UPenn*

Coursework - GitHub, Report *2021*

- Segment the hemostatic plug instance in 3D biomedical images: modify the SOLO model (Segmenting Objects by Locations) with a customized ResNet50 backbone for binary classification, with an addition of GRU for Feature Pyramid Network output to encode the sequence of images.

Product Match by Deep Learning in Computer Vision and Natural Language Processing *UPenn*

Coursework - GitHub, Report *2021*

- Match the same e-commerce products in 70k test dataset by exploiting their titles (BM25, Doc2Vec, BRET, Faiss) and images (CNN, DNN, VGG19, ResNet152) in 32k training entities with the help of zero-shot learning, triplet loss function, and KNN embedding, reaching 97% accuracy.

Hotel Cancellation Prediction Using 10+ Machine Learning Approaches *UPenn*

Coursework - GitHub, Report *2020*

- Use 10+ machine learning models (e.g. AdaBoost, XGBoost, SVM, and an ensemble method combining tuned neural network, tuned random forest, and decision tree) with SMOTE rebalance technique to predict the hotel booking cancellation, leading to 97.52% accuracy.

Inverse Reinforcement Learning on Gridworld *UPenn*

Coursework - GitHub, Report *2021*

- Compare and explore the reinforcement learning and inverse reinforcement learning models on different gridworld environments.

Miscellaneous

- Yifei worked as analyst and technology consultant intern at several financial and accounting firms while majoring in management, then during his master's at UPenn, he developed a great passion for artificia intelligence and made a significant shift to AI academia :)