**Zelin Li**

+86-159-0185-2801 | [zl3611@nyu.edu](mailto:zl3611@nyu.edu)

**EDUCATION**

**New York University Shanghai**  Expected: May 2024

*Bachelor of Science, Data Science (AI Track) | GPA: 3.62/4.0*

*Minor: Computer Science | Overall GPA: 3.59/4.0*

Research: **Natural Language Processing (NLP), Large Language Models (LLM), Data Science**

**New York University**   Sept 2022 – June 2023

*Study Away Program*

Courses: NLP, Machine Learning, Projects in Data Science, Database, Algorithms, Programming Tools for DS

**RESEARCH EXPERIENCE**

**GNN-Based Multi-Turn Dialogue Representation for Web-Enhanced QA**   May 2023 – Present

*Planned Submission to ACL 2023, 1st Author, Supervised by Prof. Xiaofan Zhang, Shanghai AI Lab*

* Identified new research topic, enhancing multi-turn dialogue vector representations in LLM+Web-Enhanced QA systems, minimizing information loss for web-based search.
* Retrieved and processed medical docs with multi-turn dialogue, integrating medical InternLM (20B) to rank doc by model perplexity, aiding in final turn responses and then utilizing rankings as training labels.
* Pioneered a Graph Neural Network (GNN) approach and designed innovative graph structure: dialogue turns as nodes and syntactic trees (N-LTP) as edges, capturing key information from the dialogues. Optimized the web search vector and calculated similarity between enhanced vector and the vectors of docs derived from BERT.
* dialogues, showing reliable accuracy in context-document alignment for the novel multi-turn dialogue QA model.

**NANER: Instance-Aware Named Entity Recognition (NER)** Mar – Aug 2023

4th Author, *Supervised by Prof. Xiaofan Zhang, Shanghai AI Lab*

* Proposed **NANER**, an NER model that utilized instance-based prompt learning to resolve issues related to category ambiguity and the complexity of obtaining high-quality descriptions.
* Employed an instance-based span model named NASpan within NANER, constructing spans with complete tokens, guided by specific entity instances sampled from training sets or online sources like Wikipedia.
* Verified the robustness of NANER by achieving state-of-the-art F1 improvements on datasets such as ACE04, ACE05, and GENIA. NANER also excelled in domain transfer tasks through zero & few-shot learning, enhancing **F1 of 11.13% on CoNLL03 and 8.59% on Wnut17** compared to description-based zero-shot benchmarks.

**Vector Quantized-Context Optimization (VQ-CoOp) for Continual Learning**  Oct 2022 – Aug 2023

*AAAI 2023 Contributor (Under Review), 2*nd Author, *Supervised by Prof. Yan Wang, East China Normal University*

* Tackled the challenge of catastrophic forgetting in continual learning, focusing on enhancing Vision-Language Models (VLMs) flexibility and maintaining memory stability in open-set image recognition tasks.
* Creatively proposed **VQ-CoOp** for adapting CLIP-like VLMs to continual learning in open-set image recognition, employing gradient-based optimization for task adaptation.
* Introduced an automatic prompt-learning mechanism that maps learnable prompts to a set of discrete codes in a pre-defined codebook, mimicking CLIP's manual prompt selection to stabilize memory during learning.
* Used **Exponential Moving Average (EMA)** for dynamic codebook updates, adding model stability and reliability, gaining **2.9% increase** in **average accuracy** and **6.33 reduction** in **average forgetting values** across **11 datasets**.

**Aspect Segmentation: Validation through Multiclass Sentiment Analysis in Movie Reviews** Jan – May 2023

*Group Leader***,** *Natural Language Processing Course Project, Supervised by Prof. Adam Meyers, NYU*

* Developed a user-centric mechanism for filtering social media comments, autonomously categorizing raw comments with user-defined tags, enhancing the precision of content selection based on individual preferences.
* Employed TF-IDF and Word2Vec to transform text data into feature vectors; implemented Multinomial Naive Bayes, KNN, and Random Forest, among which KNN on Word2Vec embedding successfully yielded the best F**1 of 0.83**, which significantly **supersedes baseline F1 of 0.17**.
* Applied GPT-4 API to evaluate the model’s performance in classifying various movie aspects; successfully confirmed the validity of gpt-4 for this classification task by human annotations with **Cohen's Kappa of 0.81**.

**LetsMeet: Data-Driven Location and Event Recommendation Platform Jan - May 2023**

**Group Leader,** *Projects in Data Science Course Project, Supervised by Prof. João Sedoc, NYU Stern*

* Led the development of LetsMeet, a unique platform to recommend the most suitable meeting locations and events around users, tailoring to individual user preferences, budget, and geographic constraints.
* Orchestrated a robust data processing pipeline, integrating Yelp API and advanced web-scraping methods like Selenium to amass dynamic data on restaurants and events. Applied K-means clustering and TF-IDF vectorization, to refine the alignment of recommendations with intricate user preferences.
* Successfully achieved an F1 score of 0.85, MAE of 1.2 and RMSE of 1.5 in our recommendation engine, indicating high accuracy in aligning suggestions with complex user specifications and preferences.

**WORK EXPERIENCE**

**Shanghai AI Lab** Shanghai, China

*NLP LLM Research Intern* May – Dec 2023

* Leveraged generated SFT data to complete two downstream fine-tuning tasks on our lab’s medical **InternLM (20B),** focusing on condensing **patient queries** and answering questions using an LLM linked to the databases.
* Applied MinHash algorithm to reduce data duplication, improving the quality of medical SFT data for fine-tuning.
* Utilized DeepSpeed and ZeRO-2 for accelerated training, and optimized hyperparameters using metaheuristic algorithms to achieve normal convergence.
* Innovatively proposed a "Self-checking" prompt template that, upon implementation, across various tasks throughout the group, reduced **average error-induced loss by approximately 6%.**

**China Ping** **An Technology** Shanghai, China

*Algorithm Engineer Intern* May – Aug 2022

* Engineered features and preprocessed real-life complaint-related customer insurance data from the company using One-Hot encoding and Pandas.
* Spotted an emerging trend in customer complaints on the “Accidental Injury Insurance” product, implementing **LightGBM** with suitable hyperparameters to classify customer complaints based on insurance features.
* Evaluated model’s performance using confusion matrix and precision, recall, and AUC-ROC metrics; the results revealed a high **F1 of 0.87** and **accuracy of 86.1%** to successfully predict potential customer complaints and help reduce economic losses through precautionary measures, such as proactively contacting potential complainers.

*Data Analyst Intern* June – Aug 2021

* Independently cleaned **200,000** real-life customer data objects from our database to be suitable for analysis; successfully used Python’s Numpy and Pandas packages for exploratory data analysis, utilizing Seaborn and Matplotlib to visualize the characteristics of insurance buyers.
* Identified key predictors of purchase intent for an insurance product: Leveraging Pearson Correlation Coefficient and Spearman's Rank Correlation, revealing a **0.87 correlation** between customer age and purchase likelihood.
* Conducted Chi-Squared Tests to validate key demographic variables, finding a **95% confidence level** in the association between customer income bracket and product preferences.

**TECHNICAL SKILLS**

**Machine Learning:** Model: VLMs, LightGBM, Multinomial Naive Bayes, KNN, Random Forest

Evaluation: F1, AUC-ROC, BLEU, MRR, NDCG, Cohen's Kappa

Feature Engineering: One-Hot Encoding, Word2Vec, TF-IDF

Optimization: Gradient-Based Optimization, Vector Quantized Context Optimization

Python: TensorFlow, PyTorch, SciPy, Gym, Scikit-learn

**Data Management:** Python: Pandas, Numpy, Seaborn, Matplotlib, Algorithms

SQL: Data Querying & Manipulation, Schema Design

**NLP Techniques:** LLM: Prompt Engineering & Design, Fine-Tuning, Hyperparameters Optimization

NLP: RNN, LSTM, CNN, GNNs (GCN), BERT, Transformers

**Distributed System:** DeepSpeed, ZeRO-2: GPU Utilization, Memory/Communication Optimization

**EXTRACURRICULARS**

Assistant Director of NYUSH Health & Wellness Student Government. 09/2020 - 05/2022

**SteppingStones Non-profit Organization**  Shanghai, China

*Volunteer Teaching*  May – Aug 2022

* Taught Python and Excel to minority children, fostering digital literacy and essential computing skills.