Andre (Jianyou) Wang

+1(984) 260-2820, jiw101@ucsd.edu,San Diego, CA, Google Scholar

Education

Ph.D. in Computer Science and Engineering, University of California, San Diego **2021-2027**

- Research Areas: LLM, NLP, DL, ML.
- Specialty: AI Scientist; Agentic Framework; Instruction-Following Information Retrieval; Long-Context Modeling; Benchmarking; Scientific Document Processing; Synthetic Data Theory and Techniques.
- Received CSE Master Degree in 2024 and Advanced to Candidacy in 2025.
- Advisors: Leon Bergen & Ramamohan Paturi

Bachelor of Science in Mathematics, Duke University **2017-2021**

- Triple Major in Computer Science, Math and Statistics, Graduation with the Highest Distinction.
- Advisors: Cynthia Rudin & Vahid Tarokh

First Author Publications & Preprints

- SSM-Retriever: Single-Pass Document Scanning for Question Answering. Conference on Language Modeling, COLM 2025. Link.
- EvidenceBench: A Benchmark for Extracting Evidence from Biomedical Papers. COLM 2025. Link.
- RoBBR: Measuring Risk of Bias in Biomedical Reports. EMNLP 2025 Submission 4/3.5/3.5. Link.
- BIRCO: A Benchmark of Information Retrieval Tasks with Complex Objectives. Preprint. Link.
- IR2: Information Regularization for Information Retrieval, LREC-COLING 2024 (Oral), Link.
- DORIS-MAE: Scientific document retrieval using multi-level aspect-based queries. NeurIPS 2023. Link.
- LimGen: There Once Was a Really Bad Poet, It Was Automated but You Didn't Know It. Transactions of ACL 2021. Link.
- DS-LSTM: Speech Emotion Recognition with Dual-Sequence LSTM Architecture. ICASSP 2020. Link.

Ph.D. Thesis (ongoing)

Research Map

- Design an automatic, LLM-powered, expert-validated pipeline to create the largest temporal knowledge graph spanning biomedicine, computer science, and other areas.
- This knowledge graph would discover the semantic connections between historical research questions and state-of-the-art scientific hypotheses using novel instruction-following information retrieval techniques.
- Design novel LLM-supervised community detection algorithms (e.g. Leiden) and statistical approaches (e.g. phylogenetics) to trace the evolutionary processes that describe old ideas evolving to new ideas.
- Answer fundamental questions about science, for example, the driving factors of fast-paced research directions, the different modes of innovation, the life cycles of scientific hypotheses, the origin, influence, and future impact of research projects beyond scientometrics.
- Create novel predictive benchmark tasks for LLM-based AI Scientist Systems to test their abilities to retrieve, aggregate, analyze, simulate, and innovate.

LLM Research Agent with Nonlinear Reasoning

- Design and train a structurally aware LLM that can iteratively generate subgraphs as nonlinear thinking structures and generate texts to fill in nodes and edges of the subgraph. Proceed in a block-by-block autoregressive manner, but the subgraph sampling is via graph diffusion, and text generation is via text diffusion. Training and validation data are subgraphs with text-enriched nodes (research questions) and edges (semantic connections) from Research Map.
- This model can effectively sample and proceed in exponentially many multiple reasoning traces at once, reducing the inference time for traditional reasoning LLMs. This model can perform future research direction prediction and simulation.

LLM Research Agent that Learns from the Past

- I discovered that LLM tends to repeat past mistakes in long chain-of-thought reasoning during inference. To ensure more efficient and streamlined inference-time reasoning, I will create synthetic data for supervised fine-tuning and leverage reinforcement learning for post-training to encourage this behavior in open-source LLMs.

Past Project Descriptions

SSM-Retriever

- I found that only $\leq 1\%$ tokens are essential for LLM to succeed in long-context QA tasks. I designed and trained a State-Space Model retriever that performs a single scanning of the entire long-context document and retrieval of only essential sentences.
- I created 10B tokens of novel link-based synthetic data to teach SSM to be contextually aware and query-centric when retrieving sentences.
- On 8× H100 GPUs, my 1.3B SSM can scale up to 256k token context and perform comparably to full-context transformer-based LLMs while achieving significant TFLOPs savings.

EvidenceBench

- I discovered that scientific review articles' descriptions of a cited paper can largely be traced back to the cited paper's full text.
- I designed a fully automatic and agentic LLM framework that extracts such descriptions, generates scientific hypotheses, and traces the description back to individual sentences in the cited paper.
- This framework created a novel evidence retrieval task for LLMs and embedding models designed to test systems' ability to comprehensively and efficiently retrieve evidence relevant to a scientific hypothesis with no redundancy and no omission.
- I fine-tuned Llama3 on EvidenceBench's 100k training documents with a new retrieval and generative objective function using PyTorch FSDP on H100 GPU cluster.
- I convened expert panels (Ph.D. students and medical doctors), designed human annotation protocols, oversaw the month-long expert annotation process, and proved our LLM agentic framework is comparable to human experts annotation in quality while reducing the required 200k expert hours to 24 machine hours.

RoBBR

- Similar to EvidenceBench, I found biomedical systematic review's risk-of-bias expert judgments can often be traced back to the cited study. This benchmark measures systems' ability to recognize potential methodological flaws and biases in peer-reviewed biomedical experiments.

BIRCO

- I created the first (as of December 2023) instruction-following benchmark to test information retrieval models' ability to follow complex user query.
- I designed an agentic LLM framework that decomposes complex queries into subtasks and assigns relevance scores for each subtask. This is the first scoring-based LLM information retrieval framework.

IR2

- Introduced the concept of information control in synthetic query generation, enabling controllable complexity in synthetic data for more challenging retrieval objectives.
- Utilized contrastive learning to train transformer embedding models for synthetic user queries.

DORIS-MAE

- I created the first semantically complex objective information retrieval benchmark for scientific document retrieval.
- I established the first protocol that ensures LLM's annotation behavior is within range of human experts' variations with hypothesis testing. Such LLM-generated benchmarks would have similar quality to human-generated benchmarks.

LimGen

- I was one of the first teams to use GPT2 to write poetry using POS templated rule-based constraint algorithms and passed the turing test.

DS-LSTM

- I created a double-helix LSTM architecture to process audio signals as both Mel-spectrograms and MFCC features for speech emotion recognition. This work served as an important baseline for future work in this domain.