# ANDY WONG

## CONTACT


andy.jy.wong@gmail.com
(858) 261 9811
SF Bay Area, CA
@AndW-llm
www.linkedin.com/in/aw-ai


## EDUCATION

**Ph.D. at UC Davis**
Hydrologic Sciences
(Machine Learning with Satellite/Drone Images)
08/2015 - 03/2021
Dissertation:
Henry A. Jastro Research Award

**B.S. @ UC Berkeley**
Environmental Sciences
(Physical track)
08/2011 - 05/2015
Senior Thesis:
SPUR Research award
Cal Alumni Association Leadership Award

## SKILLS

**Programming Languages**:
Python, MATLAB

**Frameworks**:
LangGraph, vLLM, PyTorch, TensorFlow, DeepSpeed, Megatron-LM, HF transformers, Streamlit

**Tools**:
Jupyter, Git, Flask, Docker

**Data Science**: HF Datasets, Pandas, Matplotlib, Seaborn

## PUBLICATIONS

**AAAI**
AI2ASE Workshop 2023
Knowledge-Guided Recurrent Neural Networks for Monthly Forest Carbon Uptake Estimation

**Water Resources Research**
Vol. 57, Issue 9 2021
Multiscale Assessment of Agricultural Consumptive Water Use in California's Central Valley

**Co-authored** research on:
object detection 2023
segmentation 2022
genetic 2021
water quality 2020
image fusion 2019
water management 2018

## WORK EXPERIENCE

**Senior Research Scientist** - Ping An (Fortune Global 500: #33), Research Lab, Palo Alto, CA. 03/2021 - Present

- Led R&D to **enhance LLM question answering** capabilities, handling **up to 190 pages of contexts**.

  Achievements:
  - **Enhanced the QA capabilities of GPT-4o mini and Llama-3.1 8B by over 17%**, achieving comparable performance to GPT-4o on the Lv.1 QA task of the "Loong" benchmark. Further improvements for tackling more challenging tasks are underway.
  - **Developing business applications** for long-context LLMs, including support for debt collection efforts in the financial inclusion department.

  Key Contributions:
  - Researched and developed pre-training and fine-tuning procedures for long-context LLMs.
  - Curated a collection ($n \approx 180k$) of SEC 10-K financial reports ($length \approx 80k\ tokens$) for this project.
  - Created a long-document processing workflow using GPT-4o to generate synthetic training datasets.
  - Built interactive demos with Streamlit to compare outputs from models and workflows.
  - Reduced the resources required for fine-tuning a long-context LLM model by 75%.

- Co-developed PingAn-GPT with bloomz 7B - 176B model for customer service:

  Key Contributions:
  - Combined Faiss and BM25 to assist in manual filtering of advertisements and duplicate content in a customer service instruction fine-tune dataset.
  - Implemented and evaluated the feasibility of using GPT-3.5 and 4 to grade or compare other LLM responses, particularly for finance-focused questions, accelerating iterative cycles of reinforcement learning with human feedback (RLHF) and RAG database expansion.
  - **Co-validated the in-house RLHF algorithm** and model outputs for LLM brainstorming tasks.

- Led the development of a modeling framework to predict crop yield for insurance and commodity trading.

  Achievements:
  - Predicted NDVI, a crop health indicator, with an RMSE<5%, two days ahead of USDA public release, aiding commodity trading.
  - Insurance partner tested the product to verify crop loss claims and found a micro-F1 > 0.8 in severity classification across experimental regions.

  Key Contributions:
  - Built a pipeline to process petabytes of satellite and weather imagery into spatiotemporal features.
  - Experimented with **RFR, k-NN, CNN, and Transformers** for crop health and yield prediction.

- Designed and developed technologies for the **forest carbon sinks index insurance** [COP27].

  Achievements:
  - Lab awarded in 2022 by the headquarters for contributions to business innovation.
  - Won second place among 427 teams at the Ping An Group Public Welfare Innovation Competition in 2021.
  - Experimental model outperformed the state-of-the-art process-based model on site-wise estimation over three major forest types across the US and China, demonstrating potential to replace NASA's MOD17A2HGF product.

  Key Contributions:
  - Built a high-resolution (10m) data analytics pipeline to estimate monthly forest productivity, providing anomaly alerts and streamlining claim processing.
  - Developed a **knowledge-guided RNN** model to estimate monthly forest productivity.

---

**Data Scientist Intern** - Geosyntec Consultants, Oakland, CA. 09/2020 - 01/2021

- Assisted principal consultants with my geospatial data analytic skills on groundwater modeling projects.

  Achievements:
  - Awarded a special bonus for innovative initiatives.

  Key Contributions:
  - Built interactive maps for visualizing field and raster data, facilitating client discussions.
  - Automated spatiotemporal summaries of satellite estimates for modeling decisions.
  - Proposed consulting water districts on mitigating financial risk through trading water futures.