XI(SEAN) CHEN

melongone4@gmail.com (citations 300+)◇ Homepage ◇ Linkedin

EDUCATION

University of Massachusetts Amherst — **Amherst, MA**
- PhD in Computer Science (GPA: 3.9/4) — 2019-2025

Central South University — **Changsha, China**
- B.E. in Computer Science (GPA: 3.9/4) — 2014-2018

EXPERIENCE

Pretraining Integration & Scaling – Amazon AGI Foundations — **Sunnyvale, CA**
Applied Scientist (L5) — 2025.7 - present
- Optimized pre- and mid-training recipes by scaling law for Amazon's base models (**up to ∼200B parameters**).
- Diversified reasoning data by contrastive learning and multi-teacher knowledge distillation for mid-training.
- Refined and standardized evaluations across frontier foundation models with shot/prompt sensitivity analysis.

Amazon (AGI Foundations) — **Seattle, WA**
Applied Scientist Intern — 2024.7 - 2024.10
- Applied parameter efficient LLM training with Quantized LoRA for scalable model adaptation.
- Devised a model merging framework by aligning activation to accelerate multi-task LLM learning.
- Implemented adaptive online learning with sample re-weighting for continuous LLM improvement.

University of Massachusetts Amherst — **Amherst, MA**
Research Assistant — 2019.9 - 2025.6
- Designed an efficient cluster-aware LLM judge framework for online event extractions from temporal graph.
- Zero-shot learning on LLM with an event retrieval pipeline from global multilingual news database.
- Built a refined transformer-based global news graph with trillions of edges for event detection, recommendation.
- Developed **the largest multilingual news similarity dataset**, using active learning for retrieval and ranking.

Zhipu AI (Chinese OpenAI, first IPO LLM company with ¥70B market valuation) — **Beijing, China**
Research Intern — 2017.6 - 2018.12
- Searched and rank scholars with heterogeneous random forest (deployed on prestigious AI platform Aminer).
- Clustered and visualize temporal scientific topics with refined genetic algorithm.

PROJECT

Nova Foundation Model Training Optimization (**Artificial Analysis Intelligence Index: 62**)
- Optimized pre&mid-training recipes for Amazon's Nova base models, supporting official production release.
- Designed efficient TB-scale data preparation pipelines to support high-throughput generation.
- Analyzed optimization dynamics and failure modes through systematic ablations of data mixtures.

Reasoning Data Diversification in Mid-Trainin (**Improved 16% Coding Performance**)
- Rewrote noisy webcrawl data with LLM judge to summarize reasoning information.
- Designed contrastive knowledge distillation strategies to generate diversified reasoning data.
- Quantified data reasoning depth with Essential-web classifier to optimize mid-training recipe.

Cross-Model Evaluation and Benchmarking Framework (**Aligned up to 17% evaluation gap**)
- Standardized evaluations across frontier foundation models with shot/prompt robustness check.
- Enabled rigorous cross-model comparison through unified metrics and analysis workflows.
- Generated evaluation insights to inform model selection and training roadmap decisions.

Merge LLM with Online Activation Alignment (**Improved 10% math performance**)
- Parameter-efficient fine-tuning with data mixing and target evaluation for domain expert models.
- Developed an activation alignment framework to merge LLMs across diverse tasks (math, coding, medical).
- Adapted activation weight for LLM online learning during parameter-efficient training and sampling.

Efficient Cluster-wise LLM Judge for Event Identifications from Polls (**Improved 10x data efficiency**)

- Designed a cluster-aware active learning framework to detect temporal events from X poll texts.
- Modeled the evolution of social discourse by graph-based topic tracking and statistical analysis.

Global Multilingual News Graph Learning with Transformers

- Devised a global news retrieval system, generating trillion-scale relations with dense embeddings and indexing.
- Built rank models for similar multilingual news using rules-based and active learning classifiers.

LLM Question Answering Generation for Identifying Disasters

- Engineered zero-shot learning prompts to optimize LLM precision and recall in news event retrieval system.
- Built regression models to quantify country factors with greedy feature selection and event clustering.

SELECTED PUBLICATION

Amazon Nova 2: Multimodal reasoning and generation model

Amazon AGI Foundations. Technique Report. 2025.

Identifying and Investigating Global News Coverage of Critical Events Such as Disasters and Terrorist Attacks

Xi Chen*, Erica Cai*, Reagan Keeney, Ethan Zuckerman, Brendan O'connor, Przemyslaw Grabowicz. International AAAI Conference on Web and Social Media (**ICWSM, 9% Direct Accept**), 2025.

Efficient Cluster-aware Large Language Model Judge for Event Identification from Social Polls

Xi Chen, Mayank Bumb, Vishal Kalakonnavar, Przemyslaw Grabowicz. To appeal, 2025.

Global News Synchrony and Diversity During the Start of the COVID-19 Pandemic

Xi Chen, Scott Hale, David Jurgens, Mattia Samory, Ethan Zuckerman, Przemyslaw Grabowicz. International World Wide Web Conference (**The Webconf, 15% Accept**), 2024.

A Multilingual Similarity Dataset for News Article Frame

Xi Chen, Scott Hale, David Jurgens, Mattia Samory, Przemyslaw Grabowicz. International AAAI Conference on Web and Social Media (**ICWSM, 18% Accept**), 2023.

Multilingual Document-level Similarity

Xi Chen, Ali Zeynali, Chico Camargo, Fabian Flock, Devin Gaffney, Przemyslaw Grabowicz, Scott Hale, David Jurgens, Mattia Samory. International Workshop on Semantic Evaluation @ **NAACL**, 2022.

SKILL

LLM: Pre&mid-training, Parameter-efficient fine-tuning, Model merging, RLHF, Retrieval-augmented generation(RAG).
Search & Retrieval: News similarity search, Transformer-based retrieval models, Semantic indexing.
Recommendation Systems: Graph modeling, Active learning for ranking, Multilingual content recommendations.
Machine Learning & Optimization: Event detection, Time-series analysis, Probabilistic modeling.

SERVICE

Program Member:

- NAACL(2023)
- ICWSM(2023, 2024)
- SEMEVAL (2022, 2023, 2024)
- IC2S2 (2022, 2023, 2024)

HONOR

NSF Student Travel Award, 2024
UMass CICS best portfolio finalist, 2023.
Research presented at top-tier computer science venues (Webconf, NAACL, ICWSM, TADA, IC2S2).
Nomination for Chinese Exceptional Student (1/6500 ISchool undergrads and grads), 2018.
Ranked global Top50 of Autochess players (50/millions, a world-class strategy game for playing DOTA on chesstable).