Xuming (Mac) Huang

xuming@cs.wisc.edu | +1 (608) 286-4006 | https://xuming.ai

Computer systems; Programming languages & software engineering; Machine learning systems (MLSys).

EDUCATION

University of Wisconsin–Madison **Madison, WI**
B.S. with Honors in Computer Sciences (Operating Systems, Compiler)
B.S. in Computer Engineering (Computer Architecture) *Jan 2025 – Present*
GPA: 4.0/4.0

Stanford University **Stanford, CA**
Visiting Scholar (CS 107 Computer Organization & Systems, CS 161 Algorithms) *Jun 2025 – Aug 2025*
GPA: 4.0/4.0

University of Shanghai for Science and Technology **Shanghai, China**
Computer Sciences (Machine Learning, Artificial Intelligence) *Sep 2023 – Dec 2024*
Major GPA: 4.5/4.5

RESEARCH EXPERIENCE

LinuxGuard: AI-Powered Kernel Security Analysis **Madison, WI**
Research Assistant (Supervised by Prof. Remzi Arpaci-Dusseau and Vinay Banakar) *Jan 2025 – Present*

- Built an AI pipeline to synthesize Clang-Tidy checkers; cut triage weeks→hours on 50k+ commits by LLM prompt-chaining to mine commits, generate rules, and validate on prior kernels
- Added a failure-driven repair loop; raised validated rules 2 → 8 (4×) at ∼73% precision by mining compile errors, clustering 10k+ failures, and auto-patching
- Built a full-kernel evaluation harness; found 43 long-latent bugs (avg 4.7y) and reduced false positives at fixed recall via per-checker precision/recall gating

Multispectral U-Net Segmentation Research **Shanghai**
Research Assistant (Supervised by Prof. Xing Hu) *Oct 2024 – May 2025*

- Enhanced EKU-Net architecture for plant disease segmentation, improving IoU score from 0.71 to 0.89 on benchmark dataset of 15,000+ agricultural images
- Optimized data pipeline and training procedures for pest/disease detection, reducing false positive rate by 35% and enabling real-time field deployment for 500+ farmers

Image Classification Algorithm Benchmark **Shanghai**
Research Assistant (Supervised by Prof. Dunlu Peng) *Jan 2024 – Jun 2024*

- Compared classic ML (SVM, k-NN, Random Forest) to CNNs on CIFAR-10, resulting in 81.3% best CNN test accuracy, by training and evaluating all models on a single unified testbed
- Built a two-stage inference pipeline (ResMacNet), achieving +13% absolute gain on a "similar-object" subset to 69% (human 33%), by routing ambiguous predictions to a specialized second pass

PROJECTS

CodeLLaMA-Linux-BugFix **Stanford, CA**
Independent Developer *Jun 2025 – Present*

- Fine-tuned `CodeLLaMA-7B-Instruct` with QLoRA (4-bit; r=64, α=16) to generate Git `diff` patches for buggy Linux kernel C code.
- Built a 100k-sample dataset from kernel Git history and an end-to-end train/eval/inference pipeline; achieved BLEU 33.9 and ROUGE-L F1 0.36 on held-out fixes.

IOS App Development **Madison, WI**
Independent Developer *Aug 2025 – Present*

- Building Swift/SwiftUI apps emphasizing privacy and reliability; Tenant SOS is currently under App Store review.

Heap Allocator **Stanford, CA**
Independent Developer *Aug 2025 – Present*

- Designed custom heap allocator using binary heap data structure for free block management (O(logn) operations).
- Implemented coalescing, boundary-tagging, and fragmentation minimization for improved memory efficiency
- Validated robustness via randomized stress testing and leak detection under constrained memory conditions.

Open-source Personal Portfolio Webpage **Madison, WI**
Creator & Maintainer *Aug 2025 – Present*

- Open-source codebase for my personal research website (xuming.ai), showcasing projects, writing, and updates with github 36 stars and 36 forks

INTERNSHIP EXPERIENCE

Apple **Remote**
NLP Algorithm Intern *Oct 2024 – Nov 2024*

- Engineered real-time multilingual translation system using optimized Transformer architecture, supporting Chinese and English with <100ms latency for 95% of requests

Cool AI Technology **Shanghai**
Technical R&D, Product Dev & Ops Intern *Jul 2024 – Sep 2024*

- Redesigned company's web interface using Next.js and Tailwind CSS, improving page load time by 45% and increasing user engagement metrics by 30% across 5000+ daily active users
- Architected scalable FastAPI backend for LLM integration, handling 1,000+ concurrent requests and reducing API response time from 3s to 800ms
- Deployed Prompted Agents for AI-Hub platform with serving 20+ enterprise clients, generating $10K in new revenue within first month of launch

PROFESSIONAL SKILLS

Languages: English (Fluent), Mandarin (Native), Japanese (Beginner)
Programming: Python, **C**/C++, **Verilog**, Assembly, Java, JavaScript, Swift
Tools: Git, Linux, FastAPI, Docker, **gdb/valgrind**, LLVM
Research: LLMs, Computer Systems/Architectures, System Security, Compiler

HONORS & AWARDS

- **Dean's List**, UW-Madison
- **Presidential Scholarship**, USST

EXTRACURRICULAR ACTIVITIES

Algorithms Guide (GitHub, 107 stars) **Stanford, CA**
Author & Maintainer *Jul 2025 – Present*

Understanding Transformers **Shanghai**
Author & Instructor *2024 – Present*

GPT Implementation Guide **Shanghai**
Author & Instructor *2024 – Present*

Algorithms Visualizations **Madison, WI**
Creator & Maintainer *2024 – Present*