# Weidong Wang

(+86)15156908313 | kenazcharisma@gmail.com | Kenaz123 | Homepage | Aifly1231

## Research Interest

My research interest lies in topics related to *Software Engineering and Large Language Model System*, with a current focus on projects related to *LLM4SE, Code Agent, and Bug Detection*. I am currently working on developing agentic systems to advance bug detection and automated program repair.

## Education

- **Nanjing University** *Sept. 2022 - Jun. 2026 (expected)*
  B.S. in Computer Science (Elite Class) Nanjing, China
  - GPA: 4.52/5.00 (top 10%)

## Publications

**[1]** **[ICML'25] EPIC: Efficient Position-Independent Caching for Serving Large Language Models**
Junhao Hu, Wenrui Huang, **Weidong Wang**, Haoyi Wang, Tiancheng Hu, Qin Zhang, Hao Feng, Xusheng Chen, Yizhou Shan, Tao Xie

**[2]** **[ACL'25] RaaS: Reasoning-Aware Attention Sparsity for Efficient LLM Reasoning**
Junhao Hu, Wenrui Huang, **Weidong Wang**, Zhenwen Li, Tiancheng Hu, Zhixia Liu, Xusheng Chen, Tao Xie, Yizhou Shan

## Research Experience

- **iSE Group , UIUC** *Sept. 2024 - Now*
  Advised by Prof. Lingming Zhang and Chenyuan Yang Urbana, IL, USA
  - To transcend the manual limitations of static analysis and create a sophisticated detection system for large-scale codebases, we are designing hybrid meta-agent framework for automated vulnerability detection in large codebases (Firefox/Chromium) based on previous work **KNighter** and have successfully identified and reported several previously unknown vulnerabilities in Mozilla and FreeType.
    - Role in the project: Built automated agent pipeline to analyze **2500+** reward-tagged commits and implemented a checker synthesis system for Chromium using Clang static analyzer targeting **20+** vulnerability types.
  - Benchmarking the performance LLMs against hybrid (LLM + Static Analysis) approaches for similar bug patterns to evaluate capabilities for problem generalization based on large codebases.
    - Role in the project: Constructing a specialized dataset of 54 Linux kernel patch groups with similar bug patterns spanning 15 distinct Common Weakness Enumeration (CWE) types.

- **LLMsys Group , Peking University** *Mar. 2024 - Sept. 2024*
  Advised by Prof. Tao Xie, Dr. Yizhou Shan and Junhao Hu Beijing, China
  - We designed a position-independent context caching framework **EPIC** [1] that overcomes the prefix-matching limitation of conventional context caching. This method increased KV cache reuse by enabling non-prefix matching through a new algorithm *LegoLink*, achieving an **8x** reduction in TTFT(*Time-to-First-Token*) and a **7x** throughput improvement.
    - Role in the project: Implemented changes to the attention layer to recompute key-value caches and new interface functions that work with the latest version of vLLM in about 800 lines of Python code.
  - To address the $O(N)$ memory complexity of KV Cache in long-chain reasoning, we proposed **RaaS** [2] that employs an LRU strategy to retain milestone tokens, reducing complexity to $O(L)$ ($L \ll N$) while preserving accuracy and efficiency.
    - Role in the project: Implemented the computational kernels for various attention acceleration methods and built an end-to-end testing framework for performance evaluation.

## Honors and Awards

- **People's Scholarship (First Class)** *Dec. 2023*
  *Nanjing University*
- **Special Scholarship for Undergraduates in Basic Science (5/20)** *Dec. 2023*
  *Nanjing University*

## Skills

- **TOEFL Score:** 108 (Reading: 28, Listening: 28, Speaking: 23, Writing: 29)
- **Technical Expertise:** C/C++, Python, LaTex, Rust