Zhongxuan Song · Johnny

Build Interesting ∩ Useful work.

universeszym@mail.ustc.edu.cn
+86 17789708012
GitHub Johnny-xuan
mywebsite.j-o-x.tech
@JohnXuann
johnnyxuan.bsky.social

Education

University of Science and Technology of China (USTC) — B.S. in Computer Science Aug 2024 – Jun 2028

Hefei · **Huaxia Talent Class in Computer Science** · GPA — top 25% of major

Year 1: Mathematical Analysis · Linear Algebra · Data Structures · Graph Theory · Algebraic Structures · Linux Systems · Probability & Statistics · Neural Networks & Deep Learning
Year 2: Numerical Methods · Database Systems · Stochastic Processes · Computer Systems · Digital Circuits

Research

USTC AlphaLab — Research Intern · Advisor: Prof. Xiang Wang Mar 2026 – Present

Agentic AI · Memory architecture for LLM agents

- **SHELF: From Similarity Retrieval to Path-Aware Auditable Memory for LLM Agents** First author
  NeurIPS 2026 in submission
  Proposed an auditable memory framework for LLM agents, composed of three components — *State Frames* (entity-state graphs), *Evidence Ledgers* (factual provenance records), and an *Auditor–Gater protocol* — turning memory from opaque similarity retrieval into an auditable, verifiable, and diagnosable process. Achieved best-in-class Overall F1, BLEU, and LLM-judge scores on LoCoMo, with consistent gains across Qwen3-8B/32B and DeepSeek-V3.2.

USTC MIRA Lab — Research Intern · Advisor: Prof. Jie Wang Dec 2025 – Mar 2026

Test-time context engineering · Information-theoretic methods

- **Test-Time Context Construction: An Information Bottleneck Perspective** Co-author
  NeurIPS 2026 in submission
  Proposed **IBC²**, a training-free test-time context-construction framework that formalizes context selection as a cost-regularized Information Bottleneck problem, paired with an online evidence-admission policy that gives a provable monotonic-improvement guarantee. Outperforms baselines on ExCyTIn-Bench, AppWorld, SWE-Bench, and WebArena.

USTC Medical Imaging Center — Research Intern · Advisor: Prof. Xiaoxiao Wang Jan 2025 – Jun 2025

Brain-computer interfaces · Neural decoding · Multimodal alignment

fMRI-based semantic feature extraction and image retrieval on the THINGS dataset. Designed and trained an MLP neural decoder following MindEye, mapping fMRI signals to image semantic embeddings. Led the project group: planning, literature review, team coordination, and final report.

Projects

Anymark — Bookmark agent (Chrome extension) Dec 2025 – Present

github · anymark.j-o-x.tech

Solves the "thousands of bookmarks, can't find the one I need" problem — natural-language search, auto-classification, and cross-language matching. Live in production; in review for the Chrome Web Store.

PromptFlow — Local-first lightweight prompt manager Jan 2026 – Present · github

Desktop prompt manager: local Markdown storage, favorites and templates, one-click AI polish, one-click copy, and global hotkey activation. A tool I've used daily for a year.

Burner Note — End-to-end encrypted self-destructing notes Jan 2026 · github

Zero-knowledge architecture: all encryption happens in-browser via the Web Crypto API; the server never holds plaintext or keys.

smart-voice-chat — Voice skill for Moltbot / OpenClaw Jan 2026 · github

A voice-interaction skill contributed to the Moltbot / OpenClaw open-source agent ecosystem: auto-detects voice/text input, supports flexible output modes, and handles bilingual Chinese-English conversation natively.

Writing

*WeChat newsletter **JohnnyXuann** · "Notes I think matter, before consensus catches up." · 8 original posts published*

- **Deep dive into Claude Code: multi-agent architecture (18K words) + Context Engineering system (42K words)** — TypeScript source-code analysis based on the npm package source map of `@anthropic-ai/claude-code v2.1.88` (two-part series).
- **Kimi K2: a thawing moment for China's AI commercialization + 78-day verification & correction** — Kimi K2 evaluation, China–US pricing comparison; 78 days later, a self-audit of four predictions against real revenue, usage, and capital-market data.
- **When Sora steps aside, what is OpenAI choosing** — using OpenAI CFO's public compute (0.2 GW → 1.9 GW) and ARR (2B → 20B+) figures to unpack the resource-allocation logic.
- Other posts: OpenClaw / Moltbook agent security analysis, attention crisis in the digital age, an AI's first-person autobiography.

Skills

Languages	Python · TypeScript · C · C++
AI / ML	**Agent Harness Engineering** (Memory · Context · Evolution · Multi-agent) · Information theory · Cross-disciplinary perspective from cognitive science
Engineering	Linux · Conda · Git · PyTorch · Chrome Extension API · Web Crypto API
Heavy users	Claude Code · Codex · Gemini CLI (cumulative across claude / gpt / glm / kimi: $15,819 · 171K messages · 28.4B tokens)

Honors & Campus

2025 USTC Outstanding Student Scholarship (Spring) **2025** Huaxia Talent Class — Tier B Stipend

2024–2025 Class Monitor & Deputy League Branch Secretary, School of Information, Class 5

Basketball player, USTC Kexing Cup — School of Information