Chenjie Yuan

Seattle | chenjy4@uw.edu | +1 206-605-2162 | antaresyuan.site | LinkedIn

University of Washington, Seattle | B.S. in Human Centered Design & Engineering (College of Engineering) **09/2023 – 06/2027**

Internship Experience

ByteDance | Institute of Management Strategy | Hiring Agent | AI Product Manager Intern **06/2026 – 09/2026**

Phase Goal: During the Hiring Agent's 1→N scale-up phase, drive the resume-screening pass rate of Agent-recommended candidates toward the level achieved by senior recruiters' selections (depth); continuously expand the Hiring Agent's atomic capabilities to upgrade the ATS-dependent recruiting workflow into an AI-native recruiting workbench built on the Hiring Agent (breadth).
Business problem: During scale-up, the Recall threshold faces a trade-off between a declining pass rate when loosened and missed candidates when tightened; the product should not be another conversational AI tool, it must become a full AI recruiting operating system.

- **Job taxonomy integration to Agent knowledge:** Attributed 200+ bad cases and traced the root cause to the base model's lack of knowledge about job taxonomy, with a deeper root in JD-Title mismatches systematically depressing match scores; cleaned and normalized 30,000+ postings by Title, JD, and atomic skills, distilling a taxonomy of 14 major categories and 96 subcategories; after offline backtesting, drove a staged rollout as a structured knowledge input for recall and ranking; manually reviewed cases where "assessment passed but the two sides' categories had no overlap".
- **Iterated pros/cons label prompts and built a label-quality evaluation system:** To address imprecise content in the candidate strengths/weaknesses labels shown to users, ran 5 iterations on 5k past online data points, fixed 12+ recurring issues, and raised label usability from **25.4%** to **99.5%**, launching after passing compliance review; further built a label evaluation with a scoring rubric combining four floor dimensions (accuracy, must-have coverage, readability, compliance) and a ceiling dimension (user-intent coverage), distilling a SOTA-model scoring prompt.
- **Separating stable and variable segments of instruction context:** To address users' core burden of not knowing how to write prompts, split instructions into a fixed skeleton and variable slots, dynamically issuing prefabricated entry points based on the user's stage – lowering the barrier to use while improving output stability; collaborated with the evaluation team to inject Ground Truth in a simulated production environment for case-by-case backtesting before launch.

AI Projects (Build & Ship) : github.com/AntaresYuan

Roundtable | Multi-Agent Orchestration Platform | github.com/AntaresYuan/RoundTable **05/2026 – 06/2026**

Goal: Lower the barrier for AI-coding novices to complete complex tasks and improve orchestration and reuse efficiency for experienced users, through multi-agent workflow orchestration tailored to different development tasks.
Business problem: Multi-agent collaboration relies on users manually decomposing and orchestrating tasks; fixed workflows lack flexibility, open-ended orchestration has too high a barrier for novices, and effective flows are hard to persist and reuse.

- **Planning-entry recommendation accuracy:** Since choosing the wrong multi-agent workflow wastes the execution cost of all downstream nodes, defined success as "accurately identifying applicable boundaries and distinguishing easily-confused scenarios"; defined metadata such as task type, applicable scenarios, agent roles, input/output, process nodes, and constraints, and built a 30-case evaluation set including confusable items; through bad-case attribution identifying insufficient model knowledge, introduced 5-shot examples with positive/negative cases and reasoning explanations at the engineering layer, using an ICE-structured prompt with schema-structured output to avoid hallucination; raised ByteDance Seed 2.0 Pro's accurate-identification rate from 16/30 to 28/30 **(+40pp)**.
- **Reuse of orchestration assets:** Launched workflow recommendations based on task metadata, a visual editor, and a template marketplace; **led a 3-person team to Top 3 at the ByteDance AI Full-Stack Development Challenge and presented the final demo to the judging panel.**

Lark Loom | IM-based Long-Horizon Collaboration Agent | github.com/EdwinjJ1/lark-loom **04/2026 – 05/2026**

Goal: Distill requirement clarification, task decomposition, content production, delivery review, and retrospectives into preset Skills, enabling teams to advance the full project lifecycle within a single conversation.
Business problem: Traditional IM-based team collaboration only persists communication messages; teams must independently judge tasks and repeatedly re-organize context, causing broken information handoffs and high collaboration/delivery costs.

- **Memory recall precision and context assembly:** Rebuilt historical-message recall from semantic-similarity top-20 retrieval into Seed 1.5 Lite-based structured Memory storage plus dual-path targeted extraction by keyword and task type – using exact matching to converge semantic drift and task type to bound the recall scope; on a 60-case evaluation set distilled from task-completion dimensions and model reasoning chains, raised the pass rate from 65% to 80%.
- **Tiered model invocation by subtask complexity:** Offloaded deterministic processing such as context extraction and structuring to Seed 1.5 Lite, so Seed 2.0 Pro only handles core reasoning and generation, reducing the tokens the high-cost model must process and output; manual spot-checks found no notable quality difference, **cutting average single-task response latency from 95s to 50s (-47%).**

Kaboo | Token-Consumption Analytics Platform | ByteDance Internal Open-Source Project **06/2026 – 09/2026**

- Led the Token Efficiency Leaderboard and efficiency-analysis module, now covering thousands of internal users; defined criteria for five token-waste categories (ineffective retries, redundant tool definitions, redundant context, cache misses, over-provisioned models) and a metadata-only diagnostic scheme; normalized cross-model cost via an Effective Tokens metric.

SusBench: An Online Benchmark for Evaluating Dark Pattern Susceptibility of Computer-Use Agents | ACM IUI'26, 2nd Author

- Co-built an online benchmark testing whether computer-use agents fall for dark patterns, covering 313 task paths across 55+ real websites.
- Evaluated 5 SOTA computer-use agents and analyzed where and why they were manipulated into actions users would not intend.

Agentic RAG Citation Trustworthiness Evaluation Benchmark | Honorable Mention **05/2026 – 06/2026**

- Addressed the problem of "correct answers with citations that don't support the conclusion" in RAG systems by designing a citation-trustworthiness evaluation framework; built an evidence set of 500 questions and 5,000 distractor citations based on HotpotQA, using "whether the citation genuinely supports the answer" as the core criterion.
- Decomposed citation quality into indicators such as fabricated citations, disguised relevance, reasonable inference, and evidence gaps, and evaluated Claude and GPT models, obtaining citation trustworthiness scores of 42.5% and 26.7% respectively.

Skills

Build: Python, TypeScript / JavaScript, HTML / CSS, Next.js, Cloudflare Workers, Pandas, NumPy, LLM APIs, Claude Code, Codex
AI Product Skills: LLM evaluation, evaluation set & metric design, bad-case analysis, agent context & memory, RAG, LLM-as-judge, human-in-the-loop
Design: Figma, FigJam, Miro