Jihye Kim (Ji Hye, Kim)

Santa Clara, CA · jihye.jkim@gmail.com · +1-650-669-6448 · LinkedIn · GitHub · Homepage

RESEARCH SUMMARY

AI researcher with a Ph.D. from KAIST and expertise in AI safety and alignment — and additional grounding in causal inference and econometrics. First-author publication in Information Systems Research (UTD24/FT50); publications at ACL 2026 workshops (TrustNLP, SemEval) and ICML 2026 workshop. Research spans LLM vulnerability, covert agent exfiltration behavior, benchmark development, and trustworthy multi-agent systems — grounded in rigorous causal and empirical methodology.

RESEARCH INTERESTS

AI Safety & Alignment · Agentic AI Security & Robustness · Benchmark Development for AI Safety · Human-AI Interaction & Trust · Causal Inference & Algorithmic Fairness

TECHNICAL SKILLS

Modeling & Training: PyTorch, Hugging Face Transformers, LLM Fine-tuning (LoRA/PEFT), Reinforcement Learning (PPO, reward modeling, alignment objectives), Scikit-learn, XGBoost

NLP & LLM Systems: LLM Evaluation & Benchmark Design, Multi-Agent Systems, Agentic Safety, RAG, Prompt Engineering

Causal & Statistical Inference: Staggered DID, PSM/IPTW, IV, RD, Causal ML (Causal Forest, Double ML, etc), A/B Testing

Programming & Tools: Python (Pandas, NumPy), R, SQL, Stata, Git

EDUCATION

University of California, Santa Cruz — Expected Jan 2027
M.S. in Natural Language Processing · Dept. of Computer Science & Engineering · Santa Clara, CA

Korea Advanced Institute of Science and Technology (KAIST) — 2021 – 2025
Ph.D. in Information Systems · Dept. of Management Engineering · Seoul, Korea

Sogang University
B.S. in Mathematics — Highest Distinction · Seoul, Korea
Full Tuition Scholarship for Admission with Highest Distinction

RESEARCH EXPERIENCE

UC Santa Cruz — *Graduate Researcher* — 2025 – Present
NLP & LLM Research · Santa Clara, CA

- Designed Observer–Planner–Actor (OPA) architecture for negotiation agents (CaSiNo dataset); achieved +39.8% utility, +46% acceptance rate, −23pp hallucination rate vs. baselines (all $p < 0.001$); identified novel "Deceptive Bypass" failure mode — *accepted at TrustNLP @ ACL 2026* [GitHub] [Paper]
- Identified a cross-architecture polarity-flipping encoding subspace in LLM residual streams (MIRAGE); logistic-regression probe achieves AUC 0.975–1.000 across 9 encoding families and 8 models; two-channel real-time agent monitor reaches AUC = 0.918 on 126 agentic exfiltration scenarios, vs. AUC = 0.518 for output-only detection — *accepted at AIWILD @ ICML 2026 (advised by Prof. Chenguang Wang, UC Santa Cruz)*
- Built a hybrid dense-sparse RAG pipeline (BGE-M3 + BM25) with conversational query rewriting; +44.3% Recall@10 over zero-shot baseline across 4 QA domains — *accepted at SemEval 2026 @ ACL 2026*
- Developing mental health benchmark to evaluate LLM robustness to implicit role in multi-turn dialogue — in progress *(in collaboration with Stanford University and UC Santa Cruz, Dept. of Human-Computer Interaction)*
- Designing a benchmark and red-teaming environment for data security in enterprise multi-agent systems — in progress *(advised by Peng Qi, Ph.D., Uniphore)*
- Investigating LLM susceptibility to external persuasion and designing alignment-preserving mitigation strategies for agentic settings — in progress *(advised by Prof. Jeff Flanigan, UC Santa Cruz)*

KAIST — *Graduate Researcher* — 2021 – 2025
Information Systems & Causal Inference Research · Advisor: Prof. Wonseok Oh · Seoul, Korea

- Estimated causal effects of on-demand earned wage access on 4,000 low-wage workers via staggered DID, PSM, IV, and Double ML; found +12.9% monitoring duration and +3.7% saving frequency; validated via randomized online experiment (Prolific, n=312) — published in Information Systems Research (ISR, UTD24/FT50; one of the most selective journals in Information Systems), 2026 [Journal] [Manuscript] [Media] *(Advised by Prof. Wonseok Oh from KAIST and Prof. Sunghun Chung from George Washington University)*
- Estimated causal impact of self-order kiosk adoption on demand variety across 2,000+ stores using staggered DID + IPTW; identified +13.6% menu customization with asymmetric income effects — under revision; accepted at WISE 2024 [Slides] *(Advised by Prof. Wonseok Oh from KAIST and Prof. Anindya Ghose from NYU Stern)*
- Examined anticipatory behavioral changes in response to Google's IAP mandate on a 35,000-user panel (app-week DID, triple-differences, PSM); documenting demand reallocation by income tier and app rank — under review at JMIS (FT50) *(Advised by Prof. Wonseok Oh from KAIST and Prof. Hyeokkoo Erik Kwon from Nanyang Technological University)*
- Taught Business Analytics to 32 undergraduates (student evaluation: 92.67/100) at Kyung Hee University; TA for 7 MBA/EMBA courses at KAIST

Deloitte Consulting LLC — *Senior Consultant (Full-Time Employee)* Jan 2018 – Feb 2021

*Management Consultant · **3 years full-time industry experience***

- Designed KPI frameworks and real-time performance dashboards integrated with Oracle Cloud for C-suite decision-making across semiconductor and chemical industries
- Built statistical models on 5,000+ employee records to diagnose organizational culture and productivity drivers for large-scale digital transformation projects
- Led HR due diligence for a global e-commerce cross-border M&A, delivering workforce analytics and organizational risk assessments for executive decision-making

PUBLICATIONS & REFEREED VENUES (* = FIRST AUTHOR)

NLP & LLM Research

• Kim* (Single Author). "Coercion Suppression & Preference Hallucinations in LLM Negotiation Agents." 6th Workshop on Trustworthy Natural Language Processing (TrustNLP) at the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026) — **Accepted** [GitHub] [Paper]

• Revankar, Chauhan, Kim et al. "MIRAGE: A Polarity-Flipping Encoding Subspace in LLM Agents." 2nd Workshop on Agents in the Wild: Safety, Security, and Beyond (AIWILD) at the 43rd International Conference on Machine Learning (ICML 2026) — **Accepted**

• Revanka, Kim et al. "Multi-Turn RAG Retrieval for Conversational AI." 20th International Workshop on Semantic Evaluation (SemEval 2026) at the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026) — **Accepted**

Information Systems & Causal Inference

• Kim* et al. "Working Daily, Paid Monthly? Effects of On-Demand Earned Wage Access on the Financial Well-Being of Low-Wage Workers." Information Systems Research (ISR), UTD24/FT50 (one of the most selective journals in Information Systems), 2026 — **Published** [Journal] [Manuscript] [Media]

• Kim* et al. "The Economics of In-App Payment Options: Implications for Digital Platform Governance." Journal of Management Information Systems (JMIS), FT50 — **Under Review**

• Kim* et al. "Beyond Efficiency: The Impact of Self-Order Kiosk Adoption on Demand Variety." Manufacturing & Service Operations Management (MSOM), UTD24/FT50 — **In Revision**

• Kim*, Yoon. "Could Self-Order Kiosks Drive Unequal Sales Variety?" The 35th Workshop on Information Systems and Economics (WISE 2024), Bangkok, Thailand, Dec. 2024 — Accepted [Slides]

• Park, Kim et al. "The Information Billboard: Effects of Popular Search Terms on Search Behaviors and Digital Divide." The 57th Hawaii International Conference on System Sciences (HICSS 2024), Manoa, HI, Jan. 2024 — **Accepted** [Paper]

• Kim* et al. "Working Daily, Paid Monthly?" The 15th Conference on Information Systems and Technology (CIST 2023), Phoenix, AZ, Oct. 2023 — **Accepted**

• Kim* et al. "The Economics of In-App Payment Options." The 14th Conference on Information Systems and Technology (CIST 2022), Indianapolis, IN, Oct. 2022 — **Accepted** [Slides]

• Kim* et al. "In-App Payment Regulation and Platform Governance." KrAIS Summer Workshop 2023, Seoul, Korea, Jul. 2023 — **Accepted**

• Kim* et al. "On-Demand Earned Wage Access and Financial Well-Being of Low-Wage Workers." Marketing Science: Diversity, Equity & Inclusion Conference (MSI DEI 2023), Dallas, TX, Mar. 2023 — **Accepted**

HONORS & AWARDS

- Ph.D. Fellowship · Korea Advanced Institute of Science and Technology (KAIST), 2021–2024
- Travel Grant Award · KAIST, 2023
- Full Tuition Scholarship for Admission with Highest Distinction · Sogang University
- Teaching Scholarship · Samsung Dreamclass, 2015

ACADEMIC SERVICE

Ad-hoc Reviewer: International Conference on Information Systems (ICIS) 2022, 2023, 2024 · Conference on Information Systems and Technology (CIST) 2023, 2025 · Pacific Asia Conference on Information Systems (PACIS) 2023 · 20th International Workshop on Semantic Evaluation (SemEval 2026) at ACL 2026