YUYA ASANO

Pittsburgh, PA, USA | yua17@pitt.edu | 412-807-0563
sites.pitt.edu/~yua17 | github.com/akonoroshi | Google Scholar | ACL Anthology

SUMMARY

PhD candidate in Intelligent Systems (University of Pittsburgh, expected May 2027) who designs, builds, and evaluates LLM systems, from agent harnesses for multi-turn dialogue to LLM-as-judge evaluation pipelines and LLM features shipped in industry. 13 peer-reviewed papers (10 first-author) at venues including EMNLP, COLING, SIGDIAL, and AIED. Bilingual in Japanese (native) and English.

- **LLM agent harness:** built the orchestration layer of an LLM teachable agent (solution classification, state-tracked dialogue loop, deterministic dialogue policy, structured JSON outputs) on Azure OpenAI, deployed for a 72-participant controlled study.
- **LLM evaluation:** benchmarked 13 prompting and sampling methods across 4 LLMs against 404 real student solutions (EMNLP 2025); raised GPT-4o's correlation with human raters by .12–.28 through task formulation (ETS; BEA 2025).
- **LLM pipelines in industry:** GPT-5.5 skill extraction plus retrieval and few-shot classification for curriculum-standard labeling, improving macro F1 by .19–.33 over the existing baseline (SigIQ.ai).
- **LLMs for speech interfaces:** LLM-based augmentation that tripled speech-recognition error-correction recall (.18 to .54) for an Alexa Prize TaskBot (COLING 2025).

TECHNICAL SKILLS

LLM engineering: agent and dialogue harness design (multi-stage orchestration, state tracking, routing between LLM calls), structured generation (JSON schemas), prompt engineering, in-context and few-shot learning, multi-agent generation, retrieval-augmented generation, nearest-neighbor retrieval

LLM evaluation: LLM-as-judge pipelines, multi-model voting, validation against human ratings, comparing LLM output distributions with human data (coverage, total variation distance), human evaluation

Models and serving: GPT-5.5, GPT-4o, GPT-4o mini, Qwen2.5, Gemma 3; vLLM, OpenAI API, Azure OpenAI / Azure AI Foundry, LangChain, REST APIs

NLP and speech: dialogue systems, ASR error correction, linguistic alignment analysis, named-entity extraction, summarization

Machine learning: PyTorch, Hugging Face Transformers, scikit-learn, TensorFlow; Bayesian modeling (Beta-Bernoulli), factorization machines, reinforcement learning, clustering

Cloud and infrastructure: AWS, Docker, Azure SQL, Git, Alembic, Google Cloud Translation API

Programming and tools: Python, SQL, Java, C#, JavaScript, C, Scala, R, HTML; ASP.NET, Streamlit

Statistics and experimentation: experimental design and power analysis, A/B tests and field experiments, regression and correlation modeling (R, SPSS), human evaluation, annotation and inter-rater reliability

User research: user interviews, surveys (Qualtrics), prototyping (lo-fi, Figma)

Languages: Japanese (native), English (full professional proficiency)

INDUSTRY EXPERIENCE

SigIQ.ai, Berkeley, CA, USA — May 2026 – Aug 2026
Machine Learning Intern

- Built a two-stage LLM and retrieval pipeline that labels lessons with the Common Core State Standards they cover, improving macro F1 by .19–.33 and exact match by .20–.44 over the existing baseline. GPT-5.5 first extracts the math skills a lesson teaches so lessons and standards are compared in the same terms; nearest-neighbor retrieval then narrows candidate standards at near-perfect recall, and a few-shot classifier picks among them.
- Implemented a Beta-Bernoulli model that estimates each student's mastery of each standard, shipped as a teacher-facing product feature.
- Adapted mastery estimation to infer skill from students' conversations with the AI tutor instead of their graded answers, and benchmarked it against the model running in production.

Educational Testing Service (ETS), Princeton, NJ, USA Jun 2024 – Aug 2024
Ida Lawrence Research Intern

- Automated scoring of teachers' ability to lead coherent discussions in simulated classrooms using GPT-4o on Microsoft Azure; the best task formulation in each subject raised correlation with human ratings by .12–.28 over a naive rubric prompt (BEA 2025).
- Compared LLM task-formulation strategies and how far they transfer across subject domains. The strategies reduced GPT-4o's overestimation of instructional quality and surfaced relevant supporting quotes, two problems reported in prior work; identified the remaining gaps (weak science-to-math transfer, and misalignment with human raters in how quoted evidence is used).

Recruit Group, Tokyo, Japan Jun 2021 – Jul 2021
Machine Learning Engineer Intern

- Developed a factorization-machine recommender for retail rental listings on tempodas.com in Python and PyTorch.

Aidemy, Tokyo, Japan Mar 2021 – Jul 2021
Data Scientist Intern

- Built Streamlit dashboards of AI-course prerequisite structure and learner study habits, used by the sales and curriculum teams to package courses and design the curriculum.
- Integrated the Google Cloud Translation API into existing courses to create a multilingual learning environment, opening the catalog to learners outside Japan.

RESEARCH EXPERIENCE

University of Pittsburgh, Pittsburgh, PA, USA May 2022 – Present
Graduate Research Assistant (Advisor: Dr. Diane J. Litman)

LLM systems

- Designing and building the LLM agent harness (Python) behind Emma, a teachable robot that students teach math to. An LLM classifier identifies which solution strategy a student is teaching and asks clarifying questions when it is unclear; a state-tracked step loop elicits what, how, and why for each step; and a deterministic policy introduces a verified alternative solution and guides the student to reconcile it with their own.
- Cut formulaic, repetitive agent dialogue by separating internal judgment from spoken output with structured JSON generation, keeping dialogue-move selection deterministic and confining variability to surface language. Runs on Azure OpenAI (Azure AI Foundry) for an IRB-approved, 72-participant controlled study; open-weight models served locally on vLLM for prototyping and offline experiments.
- Developing a multi-agent LLM framework that diversifies math reasoning paths using principles from research on creative human teams: persona-differentiated agents, a decoding-time penalty that discourages one agent from reusing the group's shared phrasing, and a two-phase explore-then-recombine protocol with a shared idea memory.
- Building the LLM evaluation harness for that framework: three open-weight LLMs served on vLLM vote on each generated solution's correctness (unanimity required), then classify it as matching a known solution or novel by majority vote, with the classifier validated on paraphrased reference solutions. Applied to the CreativeMath benchmark and to real student data.
- Quantified whether LLMs reproduce students' correct solutions to free-response math problems: annotated 404 student solutions from 39 human-robot dialogues and 5,838 LLM solutions across 4 models (proprietary, open general, and open math-specialized) and 13 prompting and sampling methods (including tree of thoughts, denial prompting, multi-agent debate, translation-based paraphrasing, and in-context learning). No model–method pair covered all student strategies, and each over- or under-produced some strategy by at least 20 percentage points, a gap that matters when LLMs stand in for students in system design and testing (EMNLP 2025).

Dialogue analytics and human-AI interaction

- Extended a shared-expression measure of lexical alignment to multi-party human-human-robot dialogue, attributing a group's shared vocabulary to individual speakers; the new measure predicts rapport better than existing pair-wise and group-level measures (AIED 2025). Released as an open-source Python package (dialign_python).
- Showed that unresolved disagreements between students and the robot over how to solve a problem, rather than disagreement itself, predict lower rapport with both the robot and the human partner, which motivated the agent design above (manuscript in preparation).
- Compared metrics of participation balance for predicting group learning, including ones a robot could compute automatically and act on in real time (EDM 2024).

- Measured how experiencing speech-recognition errors affects learning gains and rapport with a teachable agent, and derived implications for error-recovery policy (AIED 2023, Best Poster Award).
- Modeling lexical and semantic alignment in human-human-robot interaction and its relationship to learning outcomes and rapport; 28 participants collected to date, ongoing toward a target of 70–80.

Study design, data, and infrastructure

- Design and run lab and classroom studies end to end: experimental design and power analysis, IRB protocols and amendments, participant sessions, dialogue and audio data collection, annotation, and statistical analysis.
- Developed annotation schemes for students' solution strategies and disagreement episodes, trained undergraduate annotators, and refined the schemes through disagreement review (Gwet's AC1 = .81 for disagreement coding).
- Own the study infrastructure end to end: an ASP.NET web application, a Python dialogue service (migrated from rule-based AIML to LLM-based), Azure SQL, and the data collection and analysis pipelines. Kept participant data in an IRB-compliant institutional Azure environment, separate from non-human-subjects research compute.

Amazon Alexa Prize TaskBot Challenge 2 Feb 2023 – Sep 2023
Team Member (Advisor: Dr. Malihe Alikhani)

- Proposed a context-aware LLM augmentation and ranking method that corrects speech-recognition errors in task-oriented dialogue without prior user data: n-best hypotheses are ranked by lexical and semantic similarity to dialogue context, and context is ranked by phonetic correspondence to the hypotheses.
- Raised correction recall from .18 to .54 and F1 from .27 to .44 over a named-entity baseline in offline evaluation; with real users, interactions were rated .8–1 point higher (out of 5) when correction worked, and false positives did not significantly lower ratings (COLING 2025, Industry Track).
- Evaluated with real users in the home-improvement and cooking domains; the team placed third in the competition.

Jackson State University, Jackson, MS, USA May 2025 – Aug 2025
NLP Research Intern (Advisor: Dr. Wei Zheng)

- Proposed a retrieval-augmented generation (RAG) approach that tailors grading rubrics to students' oral and written responses on domain-specific homework problems.
- Built prototype grading APIs with AWS, SQL with Alembic migrations, LangChain, the OpenAI API, and Docker (prototype; not deployed to production).

Carnegie Mellon University, Pittsburgh, PA, USA Jan 2020 – May 2021
Research Intern (Advisors: Dr. Carolyn P. Rosé, Dr. Majd Sakr)

- Delivered instructionally useful programming hints by clustering student solutions in a human-in-the-loop pipeline.
- Prototyped an instructor dashboard that summarizes student reflections by topic and instructor-defined theme, using spaCy named-entity extraction and Longformer abstractive summarization. In a Wizard-of-Oz study with 11 TAs, these summaries were rated more useful than summaries of randomly sampled reflections (4.93 vs. 3.36 out of 7) (ICCE 2021).

University of Toronto, Toronto, ON, Canada Apr 2019 – Apr 2021
Undergraduate Research Assistant (Advisor: Dr. Joseph Jay Williams)

- Built a reinforcement-learning system that personalized the feedback students received after solving problems on online learning platforms (L@S 2021).
- Ran field experiments in a live university course on prompting students to write self-explanations, characterizing who engages with optional prompts and what shifts that engagement (LAK 2020).
- Contributed to studies on procrastination and gaming behavior in an online homework system (SIGCSE 2021) and on visualizations that nudge students to reflect on gaming the system (L@S 2020).
- Conducted user interviews and surveys (Qualtrics) to inform user-centered feedback design.

EDUCATION

University of Pittsburgh, Pittsburgh, PA, USA Aug 2021 – Expected May 2027
Ph.D. in Intelligent Systems | GPA 3.85/4.0 | Advisor: Dr. Diane J. Litman
Dissertation: *Understanding and Designing Alignment in Collaborative Learning with a Conversational Teachable Agent*
Focus: when conversational agents should align with or diverge from learners, and how to build LLM agents that deliberately introduce and resolve disagreement.

University of Toronto, Toronto, ON, Canada — Sep 2018 – Jun 2021
Honours B.Sc. in Computer Science, AI Focus | GPA 3.97/4.0

Beloit College, Beloit, WI, USA — Aug 2016 – May 2018
Undergraduate coursework (20.75 of 31 units; transferred to the University of Toronto) | GPA 4.0/4.0

PUBLICATIONS

13 peer-reviewed papers, 10 as first author, across NLP (EMNLP, COLING, SIGDIAL, BEA) and educational technology (AIED, EDM, L@S, LAK, SIGCSE, ICCE).

Refereed conference proceedings

- **Asano, Y.**, Litman, D., Walker, E. (2025). Can LLMs simulate the same correct solutions to free-response math problems as real students? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 16336–16365.
- **Asano, Y.**, Beigman Klebanov, B., Mikeska, J. N. (2025). Exploring task formulation strategies to evaluate the coherence of classroom discussions with GPT-4o. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pp. 716–736.
- **Asano, Y.**, Litman, D., Sharma, P., Fritsch, D., King-Shepard, Q., Nokes-Malach, T., Kovashka, A., Walker, E. (2025). Multi-party lexical alignment in collaborative learning with a teachable robot. In *Proceedings of the 26th International Conference on Artificial Intelligence in Education (AIED)*, LNAI 15882, pp. 123–131. Springer. [code]
- Sharma, P., Fritsch, D., **Asano, Y.**, King-Shepard, Q., Langley, T., Maidment, T., Litman, D., Nokes-Malach, T., Kovashka, A., Lobczowski, N., Walker, E. (2025). Beyond static measures: Temporal analysis of lexical alignment in human-human learning with a teachable robot. In *Proceedings of the 26th International Conference on Artificial Intelligence in Education (AIED)*, LNAI 15880, pp. 277–291. Springer.
- **Asano, Y.**, Hassan, S., Sharma, P., Sicilia, A. B., Atwell, K., Litman, D., Alikhani, M. (2025). Contextual ASR error handling with LLMs augmentation for goal-oriented conversational AI. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track (COLING)*, pp. 374–386.
- **Asano, Y.**, Litman, D., King-Shepard, Q., Maidment, T., Langley, T., Davison, T., Nokes-Malach, T., Kovashka, A., Walker, E. (2024). What metrics of participation balance predict outcomes of collaborative learning with a robot? In *Proceedings of the 17th International Conference on Educational Data Mining (EDM)*, pp. 575–583. [arXiv:2405.11092]
- **Asano, Y.**, Litman, D., Yu, M., Lobczowski, N., Nokes-Malach, T., Kovashka, A., Walker, E. (2023). Impact of experiencing misrecognition by teachable agents on learning and rapport. In *Artificial Intelligence in Education: Posters and Late Breaking Results, 24th International Conference (AIED)*, CCIS 1831, pp. 606–611. Springer. **(Best Poster Award)** [arXiv:2306.07302]
- **Asano, Y.**, Litman, D., Yu, M., Lobczowski, N., Nokes-Malach, T., Kovashka, A., Walker, E. (2022). Comparison of lexical alignment with a teachable robot in human-robot and human-human-robot interactions. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pp. 615–622.
- **Asano, Y.**, Sankaranarayanan, S., Sakr, M., Bogart, C. (2021). A thematic summarization dashboard for navigating student reflections at scale. In *Proceedings of the 29th International Conference on Computers in Education (ICCE)*, pp. 221–230.
- **Asano, Y.**, Dutta, M., Thakur, T., Solyst, J., Cristea, S., Jovic, H., Petersen, A., Williams, J. J. (2021). Exploring additional personalized support while attempting exercise problems in online learning platforms. In *Proceedings of the Eighth ACM Conference on Learning @ Scale (L@S)*, pp. 235–238.
- Solyst, J., Thakur, T., Dutta, M., **Asano, Y.**, Petersen, A., Williams, J. J. (2021). Procrastination and gaming in an online homework system of an inverted CS1. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education (SIGCSE)*, pp. 789–795.
- Xia, M., **Asano, Y.**, Williams, J. J., Qu, H., Ma, X. (2020). Using information visualization to promote students' reflection on "gaming the system" in online learning. In *Proceedings of the Seventh ACM Conference on Learning @ Scale (L@S)*, pp. 37–49.
- **Asano, Y.**, Solyst, J., Williams, J. J. (2020). Characterizing and influencing students' tendency to write self-explanations in online homework. In *Proceedings of the 10th International Conference on Learning Analytics & Knowledge (LAK)*, pp. 448–453.

Conference presentations

- Lobczowski, N. G., **Asano, Y.**, Dahan, C., Davison, T., Nokes-Malach, T. J., Litman, D., Kovashka, A., Walker, E. A. (2023). Exploring social contagion in a CSCL environment with a social robot. Symposium at the annual meeting of the *European Association for Research on Learning and Instruction (EARLI)*.
- Lobczowski, N. G., **Asano, Y.**, Maidment, T., Tran, N., Dahan, C., Yu, M., Litman, D., Nokes-Malach, T. J., Kovashka, A., Walker, E. A. (2023). Using referencing dialogue with a teachable social robot to balance student participation. Symposium at the annual meeting of the *American Educational Research Association (AERA)*.

Extended abstracts

- Solyst, J., **Asano, Y.**, Williams, J. J. (2019). The instructor reads what you write: Encouraging introductory programming students to engage in self-explanation online. In *Sixth Annual Conference on Digital Experimentation @ MIT (CODE@MIT)*.

OPEN-SOURCE SOFTWARE AND SYSTEMS

- **dialign_python** — Python implementation and multi-party extension of shared-expression lexical alignment measures, released with the AIED 2025 paper: github.com/akonoroshi/dialign_python
- **Emma study platform** — ASP.NET web application, LLM-based Python dialogue service, and data pipelines supporting lab and classroom studies with a teachable robot (research code, not public).

AWARDS AND HONORS

Third place (team), Amazon Alexa Prize TaskBot Challenge 2	Sep 2023
Best Poster Award, AIED 2023	Jul 2023
Algorithm Prize, Jane Street Electronic Trading Challenge	Jun 2019
University of Toronto Excellence Award	Summer 2019
Rhea V. Scott Scholarship, Beloit College	Fall 2018 – Winter 2019
Presidential Scholarship, Beloit College	Fall 2016 – Spring 2018

TEACHING EXPERIENCE

Beloit College, Beloit, WI, USA

Course Tutor (faculty nomination required) — Nov 2017 – May 2018

- Tutored students across mathematics, chemistry, and economics courses.

Teaching Assistant, Calculus I and II — Sep 2017 – May 2018

- Led problem-solving support for calculus students in and outside of class sessions.

PROFESSIONAL SERVICE

Conference reviewing: ACL 2023; ACL Rolling Review (ARR), March 2026 cycle; AIED 2024–2026; BEA 2025; Collective Intelligence (CI) 2026; CSCW 2023; HCOMP 2026; L@S 2024, 2026.

Journal reviewing: Northern European Journal of Language Technology (NEJLT).