Beyond Prompting: AI Safety Education in the Generative AI Era

Phan Xuan Tan [1]*, Eiji Kamioka [1], Van Nguyen [2]

[1]Shibaura Institute of Technology, Tokyo, Japan
[2]Sophia University, Tokyo, Japan
{tanpx, kamioka}@shibaura-it.ac.jp, nguyenthanhvan161@gmail.com

Abstract

Generative AI has moved from pilots to everyday practice, delivering gains in productivity and accessibility while surfacing present-day risks—hallucinations and reliability failures, bias and unfairness, prompt-injection attacks, and so on. These trends make AI safety education a core competency. In this paper, we survey global AI safety curricula and, in the Japanese context, observe strong policy momentum but relatively few courses that explicitly combine capability instruction with systematic safety evaluation. In response, we developed a 7-week, graduate-level intensive at a private science and engineering university in Japan, with enrollment open to international exchange students at both the undergraduate and graduate levels. The curriculum progresses from machine-learning foundations to generative models and alignment, with introductory agent topics included to support risk reasoning. Delivery combines weekly lectures, invited talks from academia and industry, structured group discussions, and a final presentation plus a paper-style final project focused on risk evaluation and mitigation planning. An end-of-course survey indicates high perceived learning and positive experience and one student project later resulted in a peer-reviewed workshop paper at ICLR 2025.

Introduction

The application of Generative AI has expanded beyond research and development, becoming integral to the daily operations of professionals in medicine, finance, education, and creative fields. Recent syntheses (e.g., (Stanford Institute for Human-Centered AI 2024)) document rapid capability gains and diffusion into products and services, with large language models (LLMs) and diffusion models now routinely mediating information access and content creation (Stanford Institute for Human-Centered AI 2024). These systems deliver tangible benefits—productivity support, accessibility, and new forms of analysis and expression—but they also surface present-day risks that institutions must manage.

The most salient risks in today's deployments are well characterized in public guidance and technical literature. Hallucination and reliability failures can produce confident but incorrect outputs (National Institute of Standards and Technology (NIST) 2023; Ji et al. 2023). Bias and unfairness can arise from data, modeling choices, or deployment context, with downstream equity impacts (U.S. Department of Education, Office of Educational Technology 2023; Adewumi et al. 2024). Models can memorize and leak training data, raising privacy and IP concerns in both language and image generators (Carlini et al. 2021, 2023). Systems connected to tools are vulnerable to prompt-injection and jailbreak attacks that subvert instructions and safety policies (OWASP Foundation 2023; Liao et al. 2025). Upstream data poisoning and adversarial examples can degrade or manipulate behavior (Biggio, Nelson, and Laskov 2012; Goodfellow, Shlens, and Szegedy 2015). In education specifically, major policy frameworks now stress risk assessment, transparency, and human oversight when AI is used for teaching, assessment, or student support (U.S. Department of Education, Office of Educational Technology 2023); in Europe, the AI Act classifies several educational uses as high risk, triggering obligations for data quality, documentation, logging, robustness, and appropriate human oversight (European Commission 2024).

In response, universities and nonprofits have expanded offerings that pair capability learning with safety topics such as alignment, robustness, interpretability, governance, and evaluation—for example, MIT's Ethics for Engineers: Artificial Intelligence (MIT OpenCourseWare 2020), UC Berkeley's student-led Intro to AI Safety (UC Berkeley DeCal 2024), and field-building programs such as BlueDot Impact's alignment and governance tracks (Jones 2024). While emphases vary, a common pattern is to introduce model capabilities alongside structured approaches to risk identification, evaluation, and mitigation.

In Japan, national momentum is visible. The Ministry of Economy, Trade and Industry launched the AI Safety Institute in 2024 to develop evaluation methods and standards, and universities have published guidelines on classroom use of generative AI (Ministry of Economy, Trade and Industry (METI), Japan 2024; Saitama University 2023). At the course level, however, offerings often emphasize ethics and literacy, with comparatively fewer classes that explicitly connect ML/GenAI capabilities to safety analysis (evaluation planning, failure analysis, and mitigation) in a structured way suitable for students from different backgrounds.

Given that many students encounter generative tools long

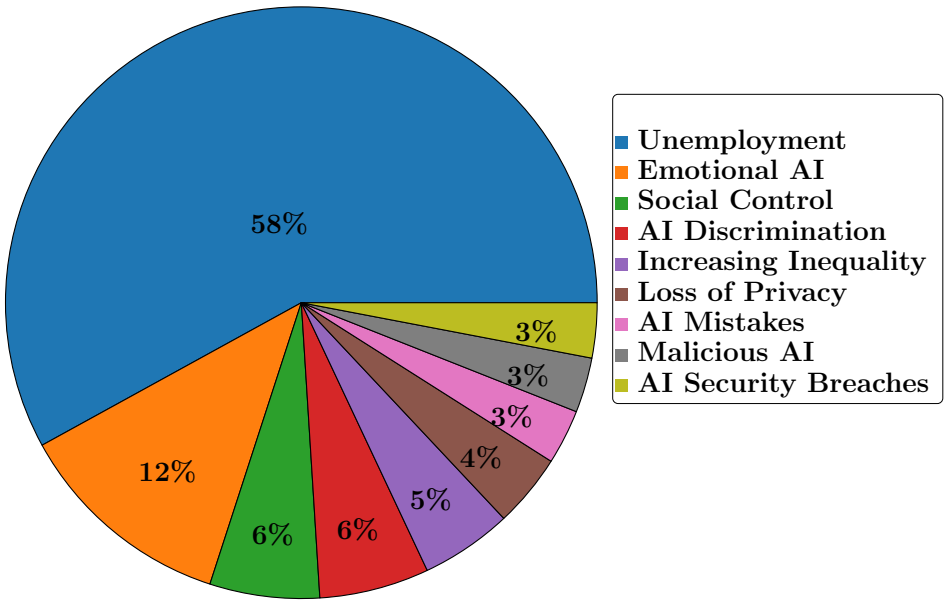


Figure 1: Moral concerns of Japanese students about AI (2021). (Ghotbi and Ho 2021)

before they encounter the concept of AI safety, we see awareness—shared vocabulary and a basic evaluation mindset—as a first-order educational need. In this paper, we use AI safety education to mean raising awareness and baseline competence in (i) recognizing common failure modes, (ii) planning simple evaluations to check for those risks, and (iii) proposing pragmatic mitigations. This is complementary to, but distinct from, broader AI ethics, which focuses on values, norms, and societal implications.

To address this local need, we developed a 7-week, graduate-level intensive at a Japanese science-and-engineering university (master's and doctoral cohorts; enrollment open to international exchange students). The curriculum progresses from machine-learning foundations to generative models and alignment, with introductory agent topics taught conceptually to support risk reasoning. Delivery combines weekly lectures, invited guest talks from academia and industry, structured group discussions, and a final presentation plus paper-style report centered on risk evaluation and mitigation planning—without coding labs. An anonymous end-of-course survey suggests the course met its core aims: students broadly reported gains in foundational understanding, awareness of AI-safety issues, and confidence in planning evaluations and mitigations. Overall satisfaction and perceived engagement were positive, with free-text comments crediting the scaffolded progression and group studios for making safety concepts concrete. Students also pointed to refinements—brief instructor-run demos and a tighter balance between guest talks and case work—which we incorporate into future iterations. Collectively, the feedback indicates that a discussion-first, evaluation-centric design is both feasible and effective in this context.

In the remaining sections of this paper, we outline related offerings internationally and in Japan and motivate the local need, describe our course design and delivery, present the results of the evaluation and discuss them, followed by conclusions.

Landscape of AI Safety Education

In this section, we reviewed publicly available syllabi and course pages for representative university offerings (e.g., Stanford CS120, Princeton COS597Q, UC Berkeley's student-led DeCal, and CMU's Trustworthy AI: Theory and Practice) together with field-building programs such as BlueDot Impact (StanfordCS120 2024; PrincetonCOS597Q 2024; BerkeleyDeCal 2024; CMU15783 2025; BlueDot 2024). Our scan is illustrative rather than exhaustive and asks a practical question: how do current courses help learners build safety competence—specifically, to (1) recognize failure modes, (2) plan and run evaluations, and (3) argue for mitigations—alongside ethics and governance.

Across the university samples we observe a consistent three-part spine: (i) capability grounding (core ML and generative models), (ii) safety concepts (robustness, adversarial vulnerabilities, interpretability, alignment), and (iii) evaluation work (designing tests, analyzing failures, proposing mitigations). The balance varies. Stanford CS120 emphasizes interpretability/robustness and adversarial analysis through readings and assignments (Stanford University 2024). Princeton's graduate/undergraduate seminar adds coding projects that replicate or extend technical safety papers (Princeton University 2024). Berkeley's DeCal combines readings, reflections, and several small coding projects as an accessible introduction (BerkeleyDeCal 2024). CMU's trustworthy-AI offering spans robustness to distribution shift, privacy/memorization, data poisoning, and jailbreaking with assignments/projects (Carnegie Mellon University 2025). Outside degree programs, BlueDot Impact runs mentored cohorts in alignment and governance; hands-on tech-

nical practice depends on track (Jones 2024). These design choices are summarized in Table 1: the global entries are marked "Yes" or "Partial–Yes" in the Tech column because they include adversarial testing, interpretability analyses, bias audits, red-teaming, or paper-replication projects.

In Japan, course descriptions in university catalogs tend to be ethics-forward and literacy-focused, with comparatively fewer modules that allocate time to structured technical safety practice (e.g., adversarial demonstrations, evaluation planning, red-team exercises, or producing operational artifacts like model cards and incident post-mortems). For instance, the University of Tokyo's "AI and Social Justice" (Global Unit) frames AI as embedded in daily life and emphasizes equity concerns (hiring/admissions, law enforcement, healthcare, social welfare, decolonizing AI) through lectures, case discussions, design-thinking workshops, and short presentations plus a 5–10-page paper (The University of Tokyo 2024). The University of Tsukuba's Human-Centered AI Curriculum balances AI with ethics and social responsibility via cross-department electives/seminars (University of Tsukuba 2024). As captured in Table 1 (Japan block), these offerings are valuable for awareness and policy literacy, but most entries are "Limited" or "Varies" in the Tech column, signaling less systematic practice in evaluation and mitigation than global exemplars.

This context matters because student exposure to generative AI is already high in Japan, while moral concerns remain substantial—unemployment, manipulation, social control, discrimination, inequality, and privacy. Figure 1 reproduces (in our style) survey results obtained from (Ghotbi and Ho 2021) to illustrate this concern profile. Viewed together—policy momentum, high tool exposure among learners, and a relative scarcity of evaluation-centric coursework—the landscape motivates an applications-first, evaluation-centric teaching model that links ML/GenAI capabilities to practical safety analysis under typical resource constraints. The next section details how our course was built to meet that need.

Methodology

Context and Learners

Our course is a graduate-level elective delivered in a compressed, seven-week format (one 3.5-hour block each Wednesday, 09:00–12:30), equivalent to a standard 14-week semester. While the home audience is master's/PhD students in science and engineering, enrollment is open to international exchange students (both undergraduate and graduate). In the 2024 offering, 35 students enrolled and were organized into 7 discussion/project groups (5–6 students/group) to support peer learning across mixed backgrounds.

Notably, many learners arrive with high exposure to GenAI tools but limited vocabulary for safety practice. The course therefore prioritizes awareness and evaluation skills over coding: there are no programming labs. Instead, students learn to recognize failure modes, plan evaluations, and argue for mitigations—skills that transfer to research, product reviews, and governance conversations.

Learning Objectives

By the end of the course, students should be able to:

- LO1 (Foundations): explain core ML concepts (supervised/unsupervised learning, bias–variance, overfitting/regularization) and the building blocks of modern generative models (autoencoders, GANs, transformers).
- LO2 (Systems and behavior): describe how generative models are trained, where reliability failures arise (e.g., hallucination, memorization, distribution shift), and how these interact with deployment context.
- LO3 (Safety concepts): distinguish outer vs. inner alignment, reward misspecification, oversight/interpretability, representation engineering, privacy/IP risks, and attack surfaces (prompt injection/jailbreaks, poisoning).
- LO4 (Evaluation competence): draft risk hypotheses, select appropriate evaluation strategies (e.g., adversarial probes, red-team plans, doc/monitoring artifacts), and propose mitigations proportionate to risk.
- LO5 (Communication): synthesize evidence in a paper-style report and have final group presentation.

Design rationale

As summarized in Section 3 (and Table 1), leading international courses typically blend three elements—capability grounding, safety concepts, and some form of evaluation practice. By contrast, publicly described offerings in Japan tend to be more ethics- and literacy-oriented, with comparatively fewer structured opportunities for technical safety practice. In response, our course adopts three principles: **(1) Capabilities first, safety second.** We scaffold from core ML and generative-model concepts to safety notions and finally to evaluation planning. **(2) Conceptual, not computational, practice.** Given mixed backgrounds, we emphasize analysis, group reasoning, and communication rather than programming. **(3) Voices from practice.** Invited speakers from industry and non-profits complement instructor lectures with conceptual demonstrations and current practitioner perspectives.

Instructional approach and assessment method

We combine instructor lectures, invited talks, and structured discussion. Basically, each 3.5-hour session is organized based on the following strategy:

- Orientation (5 min): day goals and two–three guiding questions.
- Lecture block and/or invited talk (variable, see Table 2 (100 min): concept exposition punctuated by brief think–pair–share checkpoints. Invited sessions provide conceptual demonstrations—e.g., "Building a Transformer from Scratch," "Agent Foundations & Corrigibility," and "Alignment & Agent Safety"—without hands-on coding.
- Break (10 min).
- Discussion studio (60 min): Fixed groups work on an assigned topic, and preparing a short slide.

Provider / Course	Level	Orientation	Modality / Assessment	Tech
Global				
Stanford CS120: Introduction to AI Safety	UG	Technical safety	Lectures; assignments; final project	Yes
Princeton COS597Q: AI Safety	Grad/UG	Technical alignment	Readings; seminar presentations; final talk	Limited
UC Berkeley DeCal: Intro to AI Safety	UG (student-led)	Intro + light hands-on	Readings; 2–3 small projects	Yes
CMU 15-783: Trustworthy AI—Theory & Practice	Grad/doctoral breadth	Trustworthy ML	Lectures; assignments/projects (by section)	Yes
MIT Ethics for Engineers / Ethics of AI	UG / ProfEd	Ethics / governance	Case discussions; essays / policy artifacts	Limited
BlueDot Impact (Alignment / Governance)	Open cohorts	Alignment / governance	Cohort readings; mentored projects; peer feedback	Partial
Japan (publicly described)				
Univ. of Tokyo: AI & Social Justice (Global Unit)	Short intensive	Ethics / social justice	Lectures; workshops; short presentations; 5–10pp paper	Limited
Univ. of Tsukuba: Human-Centered AI Curriculum	Master and Doctoral electives	Human-centered AI	Cross-dept. electives/seminars; essays/presentations	Limited / varies
UTokyo / Kyoto: Workshops & seminars (ethics/-gov.)	Non-credit seminar series	Governance / philosophy	Symposia; short workshops; invited talks; discussion	Limited

Table 1: Representative AI safety education offerings (illustrative). **Tech** indicates structured technical safety practice (e.g., adversarial testing, interpretability analyses, bias audits, red-teaming, paper-replication). Legend—*Yes*: required & substantive; *Partial*: present but limited/track-dependent; *Limited*: little/none; *Varies*: section-dependent.

- Group Presentations & Peer Review (30 min): Rapid presentations based on what have been discussed and followed by structured peer questioning.
- Wrap (5 min): take-aways + reading preview.

Table 2 summarizes the seven-week teaching strategy. Weeks 1–2 secure ML and basic generative-model foundations and surface visible failures via Discussion #1. Week 3 blends an invited talk on building a Transformer (conceptual) with an alignment overview to prepare the midterm. Week 4 combines in-class midterm presentations with a second invited talk on agent foundations & corrigibility. Weeks 5–6 deepen alignment topics (outer/inner alignment; oversight, interpretability, model steering; operational alignment and evaluation, governance) with Discussion #2 and a third invited talk on alignment & agent safety. Week 7 is synthesis: whole-class Q&A, final team presentations, and submission of the paper-style report. Throughout, out-of-class work is light but continuous (reading notes; team discussion; slide/report drafting) to fit the compressed calendar.

Assessment emphasizes engagement and communication over programming: Attendance (20%), Midterm presentation (30%), and Final presentation + paper-style report (50%). A total score of more than 60% is required to earn course credit. Midterm exam (Week 4) is a short team presentation that synthesizes ML foundations with a basic alignment lens. Teams select a concrete failure case (e.g., hallucination or reward misspecification) and present a concise evaluation sketch—risk, plausible probes/metrics, and one or two mitigations. Final project (Week 7) is a team project consisting of a more-than-2-page report and a presentation. Each team will select preferred topic under the guidance of the instructor such that there is no overlap, to develop a research plan. This plan must outline the associated risks, identify affected stakeholders and potential harms, propose appropriate evaluation methods (e.g., adversarial probes, monitoring artifacts, red-team plans), establish clear success criteria, and provide a rationale for suggested mitigation. The project will be graded based on the clarity of the risk analysis, the suitability of the chosen evaluation methods, the quality of evidence and argumentation, and the overall effectiveness of communication.

Evaluation

To evaluate outcomes, our university administered an anonymous, voluntary online questionnaire to all students at the end of the seven-week course (N = 35). We received 19 valid responses (approximately 54% response rate). The instrument combined (i) five 5-point Likert items (1 = Not at all achieved / Very low; 5 = Sufficiently achieved / Very high) and (ii) two open-ended prompts. The Likert items targeted three learning outcomes—ML/AI fundamentals, breadth of applications, and ethics/safety reasoning in deployment—and two experience outcomes—self-reported engagement/proactiveness and overall satisfaction. For the free-text questions (Q5–Q6), we performed a light thematic read, coding the presence/absence of recurrent ideas across short responses.

Week	Focus and in-class activities	Assessment milestones / out-of-class work
1	*Intro & ML fundamentals*. Supervised/unsupervised learning; basic ML models and applications, etc.	Syllabus check & reading notes.
2	*Advanced ML; basic generative models*. CNN, RNNs, Autoencoders, GANs, etc. and their applications. **Group Discussion #1**: "Overfitting, underfitting, gradient descent, regularization, etc."	Local group reading and discussion (suggested articles, blogs, etc) to deepen understanding of basic generative models and their relationship to the "Transformer architecture".
3	**Invited talk #1**: "Building a Transformer from scratch" (conceptual). *Alignment overview*. Dimensions of alignment (outer, inner, operational alignment), foundational pillars (agent foundation, interpretability, governance), etc.	Local group reading and discussion; and prepare slides for midterm.
4	In-class **midterm presentations**. **Invited talk #2**: "Agent foundations & corrigibility".	**Midterm** delivered; feedback returned. Local group reading and discussion focusing on different aspects of AI alignment
5	*Principles of alignment I*. Core technical problems and methods (outer, inner alignment), etc. **Group Discussion #2** Constitutional Ai, Model steering, etc.	Reading assignment, local group reading and discussion
6	*Principles of alignment II*. Operational alignment, tooling & evaluation (benchmarks, interpretability tools, etc.), AI Governance, AGI, etc. **Invited talk #3**: "Alignment & agent safety".	Local group discussion and prepare for both final presentation and paper-like report
7	*Synthesis*. Group **final presentations**; whole-class Q&A. Submit **paper-style report**.	**Final presentation plus report**.

Table 2: Seven-week plan (compressed 14-week equivalent), session flow, out-of-class expectations, and the placement of invited talks. For group work, each team has a leader and can choose to conduct local discussions using platforms like LINE or Discord.

Knowledge outcomes

Figure 2 presents a consistently scaled triptych for the three learning-outcome items (n = 19). For ML fundamentals (Fig. 2a), 68.4% reported "Sufficiently achieved," 15.8% "Somewhat achieved," 5.3% "Neutral," 0% "Not much achieved," and 10.5% "Not at all achieved." For application breadth (Fig. 2b), 52.6% marked "Sufficiently achieved" and 31.6% "Somewhat achieved," with 5.3% "Neutral," 0% "Not much achieved," and 10.5% "Not at all achieved." For ethics & safety reasoning (Fig. 2c), 57.9% selected "Sufficiently achieved," 26.3% "Somewhat achieved," 5.3% "Neutral," 0% "Not much achieved," and 10.5% "Not at all achieved." Across panels, roughly four-fifths of respondents reported at least Somewhat achieved. Despite the absence of coding labs, confidence in safety reasoning tracked gains in capability knowledge, suggesting that the design choice to interleave capability lectures with templated group analyses helped students learn to reason about failure modes in context rather than as abstract ethics.

Experience Outcomes

Figure 3 summarizes the two experience items. For overall satisfaction, 47.4% reported Very high/Positive and 36.8% High/Positive (84.2% positive overall), 10.5% Neutral, and 5.3% Very low/Negative (mean $\approx$ 4.21/5). For proactiveness, 57.9% marked Very high and 21.1% High (78.9% positive), with 5.3% Neutral, 5.3% Low, and 10.5% Very low (mean $\approx$ 4.11/5). In class, most groups rotated roles and came prepared with reading notes, aligning with the positive engagement distribution, though a small subset participated minimally during share-outs.

Free-text feedback

Ten respondents answered Q5 ("Feel free to write a review about this class (e.g., What did you learn?" "If you met your goal, why? If not, why not?" "What should you study next?") and ten answered Q6 ("What did you think about this course? Please share your opinions about good and aspects to be improved (e.g., course schedule, materials & instruments, responses towards students' inquiries, level of exams)."). Because most comments were brief (1–3 sentences), we coded themes as present/absent rather than graded intensity. Three gains surfaced repeatedly: (1) breadth and foundations ("I got a basic understanding of ML and AI fundamentals"), (2) alignment/safety awareness ("I learned about GenAI risks and alignment, especially agent safety"), and (3) confidence to go deeper (several students planned follow-up study with code). Suggestions for improvement coalesced around four ideas: (1) a desire for some hands-on artifacts or small code-adjacent demonstrations ("more technical with actual coding"), (2) guest-lecture calibration ("too many guest lectures—prefer more time for lecture or discussion" or "fewer but deeper talks"), (3) deeper, more specific cases tied to current topics, and (4) more structured peer exchange beyond reading and short presentations. These comments help explain the small tail of lower satisfaction and proactiveness in Figure 3. We include the full anonymized responses and the coding key in the Technical Appendix, available at:

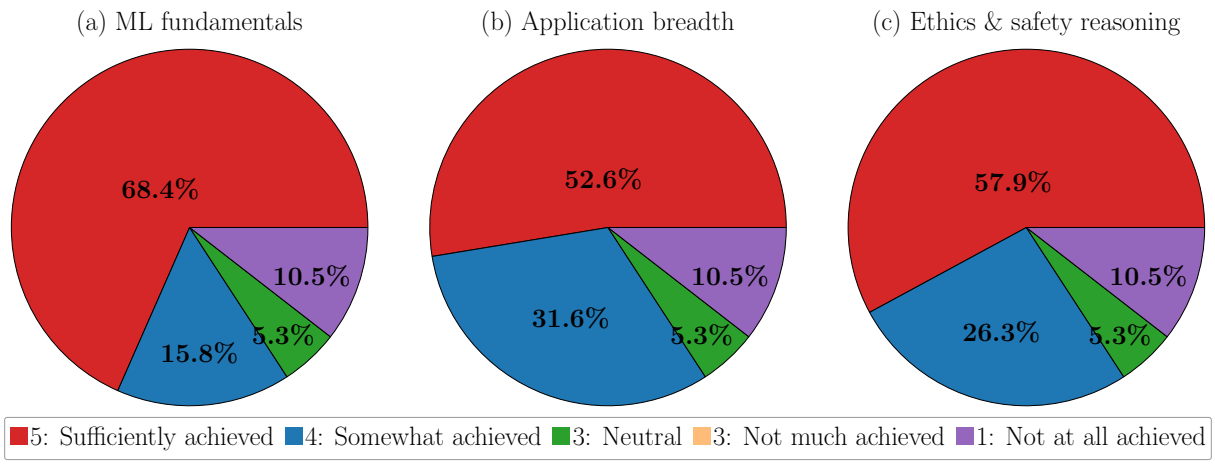


Figure 2: Knowledge outcomes (n=19)

https://doi.org/10.5281/zenodo.17594367.

Limitation of the evaluation

The survey is voluntary, self-report, and administered once at course end, so it may over-represent satisfied students and cannot attribute gains causally. We did not run pre/post quizzes or rubric-rated blind assessments of the capstone reports. The next section interprets these results in light of the course design (Table 2) and outlines targeted adjustment.

Discussion

The evaluation provides a coherent picture. On the knowledge side, approximately 80% of respondents reported at least Somewhat achieved for ML fundamentals, application breadth, and safety reasoning; on the experience side, 84% were satisfied and 79% described their engagement as high. In a mixed cohort that included international exchange undergraduates and with no coding labs, this is a notable outcome. It suggests the capabilities-first, evaluation-centric sequencing (Table 2) was effective: students first anchored in supervised/unsupervised learning and basic generative models, then analyzed visible failure modes and, finally, drafted evaluation ideas and mitigations in groups. In free-text, students consistently credited the clear scaffold "from basics to alignment," and several emphasized that guest lectures helped them connect concepts to real deployments and career paths.

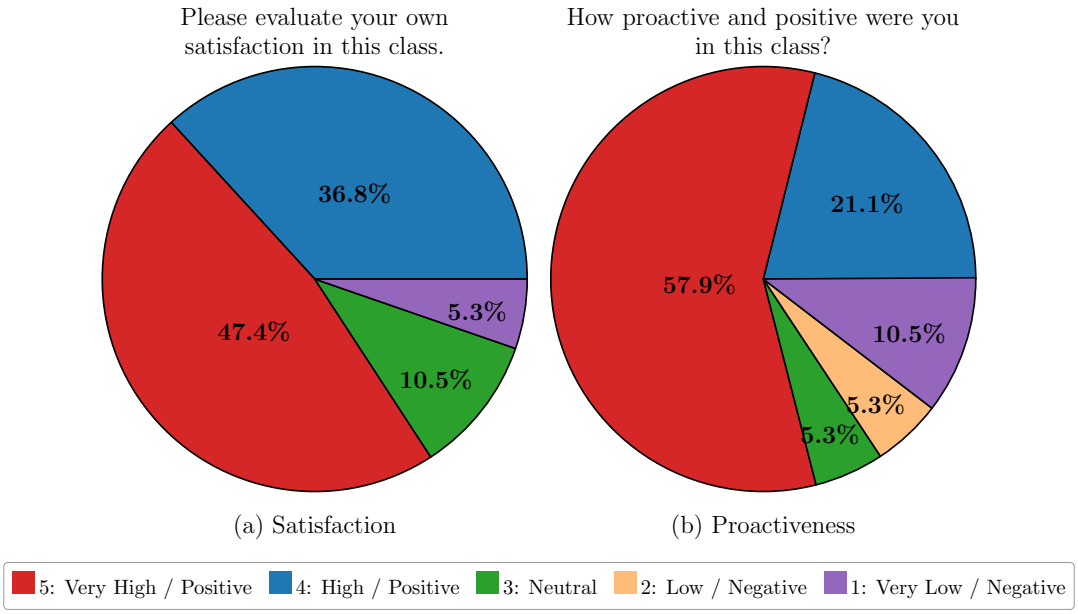
At the same time, the distributions in Figs. 2-3 include a meaningful tail. The 10% "Not at all achieved" responses on knowledge items and the small set reporting low proactiveness are consistent with three factors we also observed in class. First, background heterogeneity: a minority entered with minimal ML, which likely depressed early confidence and made it harder to contribute until Weeks 3–4. Second, the compressed format (seven 3.5-hour morning blocks) produced cognitive load; students who missed an early session or needed more time on foundations sometimes struggled to re-enter the discussion flow. Third, the no-coding constraint—a deliberate choice to focus on evaluation literacy—did not match every student's expectation; a few equated "learning AI safety" with small, concrete coding exercises and asked for quick notebook demonstrations.

The free-text provides concrete guidance for improvement. Students asking for "more technical" often did not require full labs; requests clustered around small, visual or instructor-run demonstrations that make failures tangible (e.g., a short prompt-injection probe, a quick robustness check, or an interpretability visualization). Others asked to rebalance guest talks toward depth, suggesting that fewer, tightly scoped talks would create more space for case work and peer exchange. Several comments simply asked for more time to discuss, indicating the value of formal participation scaffolds so quieter students have low-friction entry points.

We therefore consider several points to improve our course this year:

- Two or three short videos with a readiness quiz on supervised learning, overfitting/regularization, and representational shift can level the floor for students with lighter backgrounds and reduce the Week 1–2 gap.
- Five-to-eight-minute, instructor-run demonstrations (no student setup) that illustrate a failure mode or a simple evaluation script can bridge from concept to practice without turning the course into a coding lab.
- Formalize think–pair–share, rotating discussant roles with checklists, and one-minute "exit tickets" to prompt reflection and help us catch confusion early; this directly addresses the feedback asking for more structured exchange.
- Guest-lecture tuning. Prioritize one or two high-impact talks (industry + research) and reclaim an additional slot for a deep case where groups rehearse an Evaluation & Mitigation Plan end-to-end on a single scenario.

Finally, one of the most interesting points is that we note a qualitative outcome beyond self-report: one group's capstone matured into a peer-reviewed workshop paper at ICLR 2025. While anecdotal, it shows that an evaluation-centric, analysis-first course can both raise awareness and



(a) Satisfaction (b) Proactiveness

Figure 3: Experience outcomes (n=19). (a) Overall satisfaction; (b) self-reported proactiveness. Most students rated the course highly; a small tail of low engagement suggests background and format factors to address in future iterations.

seed research-grade artifacts when students choose to go deeper.

In sum, the results suggest the model works as intended for the majority—raising awareness of AI safety, building a shared vocabulary for failure modes and evaluations, and doing so in a format that is feasible under typical resource constraints. The tails are instructive rather than discouraging: with a modest primer, a handful of well-placed micro-demos, clearer participation structures, and a calibrated guest-lecture ratio, the next iteration should better support students with weaker prerequisites or different learning preferences while preserving the course's core character.

Conclusion

This paper reviewed the landscape of AI-safety education internationally and in Japan, highlighting strong policy momentum but a local gap in courses that connect capability learning to systematic safety evaluation. To address this need, we introduced a seven-week, graduate-level course at a Japanese science-and-engineering university that links ML and generative-AI foundations to safety reasoning through instructor lectures, practitioner talks, structured group studios, and a capstone presentation with a paper-style report—deliberately without coding labs. End-of-course feedback indicates that the design raised awareness and built confidence in identifying risks and planning mitigations, while pointing to pragmatic refinements (brief instructor-run demos, calibrated guest-lecture depth, and more structured peer exchange). We outline next steps—adding a short pre-course primer, embedding micro-demos at key junctures, and strengthening evaluation with pre/post checks and rubric-based artifact review—and argue that this resource-light model is readily reusable in similar Japanese contexts and adaptable elsewhere.

References

Adewumi, T.; Alkhaled, L.; Gurung, N.; van Boven, G.; and Pagliai, I. 2024. Fairness and Bias in Multimodal AI: A Survey. arXiv:2406.19097.

Biggio, B.; Nelson, B.; and Laskov, P. 2012. Poisoning Attacks against Support Vector Machines. In *Proceedings of the 29th International Conference on Machine Learning (ICML 2012)*, 1807–1814. Accessed: 2025-08-12.

Carlini, N.; Tramer, F.; Wallace, E.; Jagielski, M.; Herbert-Voss, A.; Lee, K.; Roberts, A.; Brown, T.; Song, D.; Erlingsson, U.; Oprea, A.; and Raffel, C. 2021. Extracting Training Data from Large Language Models. In *Proceedings of the 30th USENIX Security Symposium*, 2633–2650. Accessed: 2025-08-12.

Carlini, N.; et al. 2023. Extracting Training Data from Diffusion Models. In *Proceedings of the 32nd USENIX Security Symposium*. Accessed: 2025-08-12.

Carnegie Mellon University. 2025. Trustworthy AI: Theory and Practice (15-783) — Course page. Department course page. Accessed: 2025-08-15.

European Commission. 2024. AI Act — Shaping Europe's Digital Future. Accessed: 2025-08-12.

Ghotbi, N.; and Ho, M. T. 2021. Moral Awareness of College Students Regarding Artificial Intelligence. *Asian Bioethics Review*, 13(4): 421–433.

Goodfellow, I.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations (ICLR)*.

Ji, J.; Qiu, T.; Chen, B.; Zhang, B.; Lou, H.; Wang, K.; Duan, Y.; He, Z.; Zhou, J.; Zhang, Z.; and Zeng, F. 2023. AI Alignment: A Comprehensive Survey. arXiv:2310.19852.

Jones, A. 2024. Why We Run Our AI Safety Courses. BlueDot Impact Blog. Accessed: 2025-08-12.

Liao, Z.; Chen, K.; Lin, Y.; Li, K.; Liu, Y.; Chen, H.; Huang, X.; and Yu, Y. 2025. Attack and Defense Techniques in Large Language Models: A Survey and New Perspectives. arXiv:2505.00976.

Ministry of Economy, Trade and Industry (METI), Japan. 2024. Launch of AI Safety Institute. Accessed: 2025-08-12.

MIT OpenCourseWare. 2020. 10.01 Ethics for Engineers: Artificial Intelligence (Spring 2020). Accessed: 2025-08-12.

National Institute of Standards and Technology (NIST). 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Technical report, U.S. Department of Commerce. Accessed: 2025-08-12.

OWASP Foundation. 2023. OWASP Top 10 for Large Language Model Applications. Accessed: 2025-08-12.

Princeton University. 2024. COS597Q: AI Safety (Course page). Course syllabus / website. Accessed: 2025-08-15.

Saitama University. 2023. Guidelines for the Utilization of Generative AI. Accessed: 2025-08-12.

Stanford Institute for Human-Centered AI. 2024. AI Index Report 2024. Technical report, Stanford University. Accessed: 2025-08-12.

Stanford University. 2024. CS120: Introduction to AI Safety (Course page). Course syllabus / website. Accessed: 2025-08-15.

The University of Tokyo. 2024. AI and Social Justice (Global Unit) — Course description. University course catalog. Accessed: 2025-08-15.

UC Berkeley DeCal. 2024. Intro to AI Safety (DeCal). Accessed: 2025-08-12.

University of Tsukuba. 2024. Human-centered AI Curriculum — Program description. Graduate program pages. Accessed: 2025-08-15.

U.S. Department of Education, Office of Educational Technology. 2023. Artificial Intelligence and the Future of Teaching and Learning: Insights and Recommendations. Technical report, U.S. Department of Education. Accessed: 2025-08-12.