ISSN 2186-7437

NII Shonan Meeting Report

No. 235

LLM-Guided Synthesis, Testing, and Verification of Learning-Enabled Cyber-Physical Systems

Xi Zheng
Sanjoy Baruah
Simon Thompson

March 9–12, 2026

National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-Ku, Tokyo, Japan

LLM-Guided Synthesis, Testing, and Verification of Learning-Enabled Cyber-Physical Systems

Organizers:
Xi Zheng (Macquarie University, Australia)
xi.zheng@mq.edu.au
Sanjoy Baruah (Washington University in St. Louis, USA)
baruah@wustl.edu
Simon Thompson (TIER IV, Japan)
simon.thompson@tier4.jp

March 9–12, 2026

Abstract

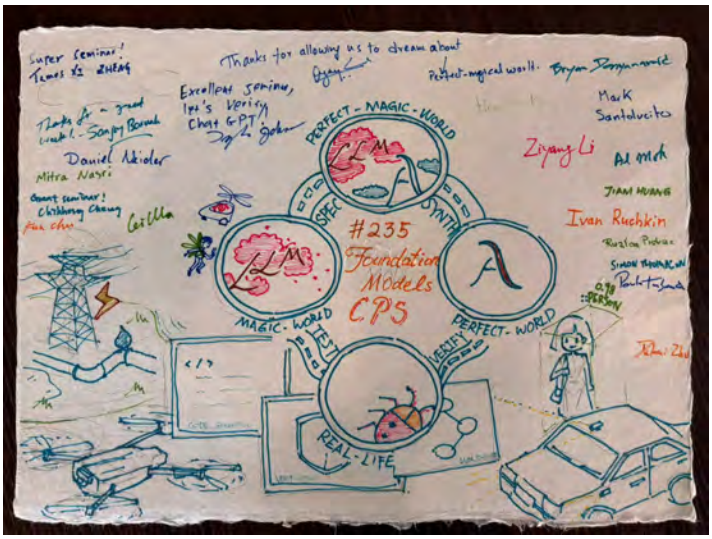

Figure 1: Shonan Meeting Sketch

Learning-enabled cyber-physical systems (CPS) are increasingly deployed in safety-critical domains such as autonomous driving, robotics, and unmanned aerial systems. While foundation models and large language models (LLMs) offer unprecedented capabilities for perception, reasoning, and automation, they also introduce significant challenges for testing, verification, and trust assurance.

This seminar sought to understand how foundation models can be systematically leveraged to support the synthesis, safety testing, and formal verification of learning-enabled CPS. Topics included scenario and data generation, specification inference, AI-assisted verification workflows, correctness guarantees, and the verification of foundation models themselves.

By bringing together experts from artificial intelligence, formal methods, software engineering, and cyber-physical systems, the seminar aimed to establish a shared research agenda and identify principled paths toward trustworthy, deployable intelligent systems.

Background and introduction

The rapid integration of machine learning and cyber-physical systems (CPS) has enabled transformative applications across safety-critical domains, including autonomous driving, delivery drones, service robotics, and automated surgical systems. Examples include autonomous vehicles such as Waymo and Tesla Autopilot, delivery drones such as Amazon Prime Air and Google Wing, service robots from Boston Dynamics and Softbank Robotics, and surgical platforms such as Da Vinci. While these learning-enabled CPS demonstrate tremendous potential, several real-world incidents have revealed critical safety risks, including human fatalities and substantial economic losses. These failures highlight the urgent need for robust testing, verification, and assurance methodologies for learning-enabled CPS.

Traditional verification and validation techniques, originally developed for deterministic software systems, are often insufficient for learning-enabled CPS. These systems rely heavily on data-driven components, probabilistic decision making, and complex interactions between perception, reasoning, and control. Consequently, verification must address challenges such as training data quality, distribution shifts, out-of-distribution scenarios, and the opaque nature of modern machine learning models. Furthermore, the absence of suitable formal modeling abstractions and scalable verification frameworks makes it difficult to provide strong correctness guarantees for learning-enabled CPS.

Although several research communities—including software engineering, formal methods, robotics, control theory, and artificial intelligence—are actively addressing these challenges, current efforts remain fragmented. In particular, generating diverse corner-case scenarios, constructing formal specifications, and performing scalable verification across complex learning pipelines remain open problems. Existing model-based testing, symbolic execution, and random testing techniques are often insufficient to explore the vast behavioral space of learning-enabled CPS. These limitations motivate the exploration of new approaches that combine data-driven intelligence with formal reasoning.

Recent advances in foundation models and large language models (LLMs) present promising opportunities to address these challenges. Foundation models can capture large-scale knowledge, reason over complex multimodal inputs, and assist in generating scenarios, specifications, and verification artifacts. These capabilities enable new paradigms for testing and verification, including automated scenario generation, corner-case discovery, specification inference, and AI-assisted verification workflows. At the same time, foundation models themselves introduce new reliability and safety concerns, requiring systematic verification and assurance methods.

To address these emerging challenges and opportunities, this Shonan seminar brought together researchers from artificial intelligence, software engineering, formal methods, robotics, and cyber-physical systems. The goal was to develop a shared research agenda for leveraging foundation models to enable trustworthy,

deployable learning-enabled CPS.

Technical Themes

The seminar was organized around three tightly connected technical themes.

Theme 1: Latest Trends in Foundation Models for Autonomous Systems. This theme focused on recent developments in foundation models for autonomous systems, including multimodal perception, vision-language-action models, world models, and embodied intelligence. Discussions examined their capabilities, limitations, and implications for system architecture, deployment, and assurance in safety-critical CPS.

Theme 2: Foundation Models for Safety Testing. This theme explored how foundation models can support testing of learning-enabled CPS, including scenario generation, corner-case discovery, simulation-based testing, data synthesis, and evaluation pipelines. Participants investigated how foundation models can improve test coverage, generate realistic environments, and enable systematic testing workflows.

Theme 3: Foundation Models for Formal Verification. This theme examined the role of foundation models in formal verification, including specification generation, model construction, verification oracles, and compositional verification. In addition, discussions addressed the verification of foundation models themselves and the integration of AI-assisted reasoning with formal verification techniques.

These themes were designed to foster cross-disciplinary collaboration and establish a coherent research agenda spanning autonomous systems, testing, and verification. The seminar aimed to identify key challenges, define promising research directions, and promote collaboration toward trustworthy and deployable learning-enabled CPS.

Overview of the meeting

The Shonan Meeting No. 235, titled *LLM-Guided Synthesis, Verification, and Testing of Learning-Enabled Cyber-Physical Systems*, was held at the Shonan Village Center, Japan, from March 9–12, 2026. The meeting brought together researchers and practitioners from academia and industry working at the intersection of cyber-physical systems, formal methods, artificial intelligence, robotics, and software engineering. The primary objective of the meeting was to explore how foundation models and large language models can support the synthesis, testing, and formal verification of learning-enabled CPS and to establish a shared research agenda for trustworthy autonomous systems.

The meeting followed the traditional Shonan seminar format, combining short technical presentations with collaborative working group discussions. Prior to the meeting, participants were encouraged to share short summaries of their ongoing research to facilitate productive discussions and identify overlapping interests. This preparation enabled rapid formation of discussion groups and helped participants identify key challenges and opportunities early in the seminar.

The final phase of the meeting focused on consolidating outcomes and defining a shared research roadmap. Participants identified key research directions, potential collaborations, and dissemination plans. The group agreed to pursue

joint publications, collaborative research proposals, and continued community building efforts. The overarching outcome of the meeting was the establishment of a coordinated research agenda for leveraging foundation models to enable trustworthy, safe, and deployable learning-enabled cyber-physical systems.

Figure 2: Collaborative discussion during the Shonan seminar highlighting cross-disciplinary interactions on world models and trustworthy autonomous systems.

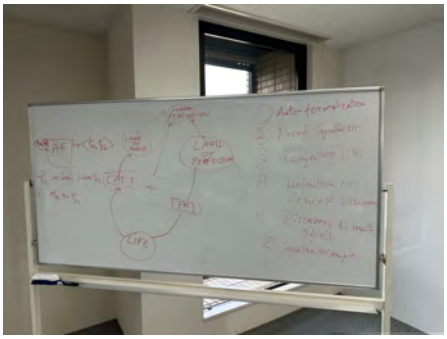

Figure 3: Whiteboard discussion illustrating emerging research directions, including auto-formalization, proof synthesis, and concept discovery for foundation-model-assisted verification.

Overview of Talks

Theme 1 Keynote: Trustworthy AI in the Era of Foundation Model Driven Agentic Systems

Lei Ma, The University of Tokyo and University of Alberta

This talk examined the evolution of trustworthy AI from traditional data-driven systems to foundation-model-driven agentic systems. The speaker first highlighted the transition toward generative AI and large foundation models, emphasizing that modern AI systems increasingly operate as autonomous agents integrating perception, reasoning, retrieval, and tool usage. This shift introduces new assurance challenges, as trustworthiness must be evaluated not only at the standalone model level, but also across unit-level components, retrieval-augmented generation pipelines, and multi-agent orchestration.

As a proof of concept, the talk presented the LUNA framework, which constructs abstract behavioral models of large language models. Beyond abstract model construction, the talk also discussed broader research challenges for trustworthy foundation-model-driven systems. Unlike traditional software, where formal methods typically provide binary guarantees such as counterexamples or correctness proofs, foundation models require statistical and probabilistic analysis. This raises new questions, including how to derive statistical bounds, what properties can be formally verified, and what new formal-method theories are needed for deep neural networks and foundation models. The speaker further highlighted challenges in handling complex multimodal contexts such as video, natural language, and audio, where system behavior becomes harder

to formally characterize. In contrast to traditional software, where source code is executable and verifiable, foundation-model-driven systems require new abstractions, hybrid reasoning approaches, and statistical assurance techniques to enable trustworthy deployment in practice.

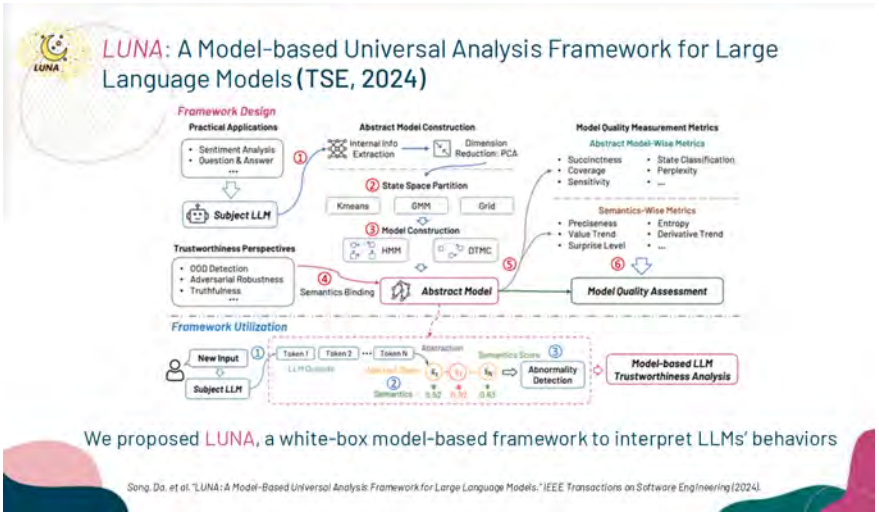

Figure 4: LUNA: A white-box model-based framework for analyzing LLM trustworthiness. The framework extracts hidden states from LLMs, performs dimensionality reduction, partitions the state space, and constructs probabilistic abstract models such as HMM or DTMC. Trustworthiness semantics are then bound to the abstract states to enable abnormality detection, robustness analysis, and truthfulness assessment.

Confidence over Time: Confidence Calibration with Temporal Logic for Large Language Model Reasoning

Ivan Ruchkin, University of Florida

This talk addressed the challenge of confidence estimation in large language model (LLM) reasoning, particularly for long-form, multi-step reasoning tasks such as mathematical problem solving and scientific question answering. The speaker highlighted that existing confidence estimation approaches typically assign a single scalar confidence score to the entire response, ignoring how confidence evolves during intermediate reasoning steps. This limitation makes existing methods sensitive to superficial factors such as verbosity and response length, and often unable to distinguish between correct reasoning and confidently generated but incorrect outputs.

To address this limitation, the talk proposed modeling stepwise confidence signals using Signal Temporal Logic (STL). The key idea is to treat confidence as a temporal signal that evolves throughout the reasoning process, and to mine discriminative temporal patterns that differentiate correct and incorrect reasoning trajectories. Using a discriminative STL mining procedure, the approach automatically discovers temporal formulas capturing characteristic patterns in confidence evolution. The speaker showed that these STL patterns generalize

across reasoning tasks, while numeric parameters exhibit sensitivity to individual questions, suggesting both shared structural patterns and instance-specific dynamics.

Building on these insights, the work introduces a confidence estimation framework that combines STL-based reasoning blocks with parameter hypernetworks to adapt temporal logic constraints dynamically. Experimental results across multiple reasoning benchmarks demonstrate that the proposed method produces more calibrated confidence estimates compared to existing baselines. The talk highlighted that temporal logic provides a promising bridge between formal methods and foundation-model reasoning, enabling interpretable, stepwise confidence estimation and opening new directions for trustworthy reasoning in LLM-based systems.

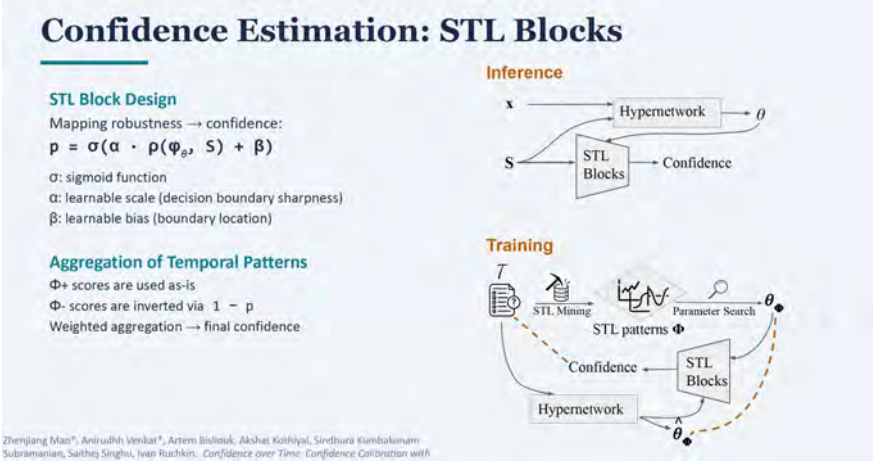

Figure 5: Confidence over Time: STL-based confidence estimation framework.

Automata Methods for Neurosymbolic AI

Mark Santolucito, Barnard College, Columbia University

Moore machines provide powerful, interpretable controllers but are discrete and difficult to scale or accelerate on GPUs. State Space Models (SSMs) offer efficient neural architectures with linear-time computation, yet gradient descent alone learns behavior without recovering symbolic structure, limiting interpretability and verification.

This talk bridged this gap through symbolic warm-starting. From execution traces, the speaker mined temporal specifications and synthesized Moore machines via automata learning, then encoded them exactly as SSMs to initialize neural architectures. Gradient-based refinement then improves performance while preserving symbolic structure.

The talk argued that this neurosymbolic pipeline combines interpretability, sample efficiency, and formal guarantees from symbolic methods with the scalability and efficiency of neural SSMs, enabling interpretable and scalable controllers for critical decision-making.

Figure 6: Big vision for a complete neurosymbolic pipeline for scalable controller synthesis.

Theme 2 Keynote: Foundation Models for Safety Testing of Learning-Enabled CPS

Xi Zheng, Macquarie University

Learning-enabled cyber-physical systems (CPS), such as autonomous driving systems (ADS) and unmanned aerial vehicles (UAVs), are increasingly deployed in safety-critical environments. However, their reliance on black-box machine learning models makes them difficult to test, interpret, and formally verify. Existing approaches either focus on neural network robustness verification with limited system-level guarantees or rely on probabilistic testing methods that struggle to provide strong safety assurance.

This talk presented a new direction toward reviving specification-based and conformance-based testing for learning-enabled CPS using foundation models and large language models (LLMs). The speaker demonstrated how LLMs can extract specifications from traffic rules, accident reports, and domain regulations, and automatically generate executable test scenarios for autonomous driving and drone systems. These scenarios are further refined through coverage-guided fuzzing, model learning, and iterative conformance testing, enabling scalable and systematic safety testing.

To make learning-enabled autonomous systems more testable, interpretable, and verifiable, the speaker introduced NeuroStrata, a hierarchical neuro-symbolic architecture that integrates neural perception with symbolic reasoning. NeuroStrata enables formal reasoning across perception, planning, and control layers, combining neural adaptability with symbolic constraints to improve reliability and explainability.

The speaker further presented a UAV landing site safety assessment framework built on NeuroStrata. The system constructs symbolic world models from perception outputs, synthesizes interpretable safety rules using LLMs, and employs logic-based reasoning engines to evaluate candidate landing zones. Human-in-the-loop validation is incorporated to iteratively refine safety rules and improve decision quality. Experimental results demonstrate improved interpretability and safety performance, outperforming state-of-the-art purely neural network-based approaches.

This talk outlined a roadmap toward foundation-model-driven conformance testing, where LLM-based specification mining, neuro-symbolic reasoning, and

human-in-the-loop validation collectively enable trustworthy, interpretable, and verifiable learning-enabled CPS.

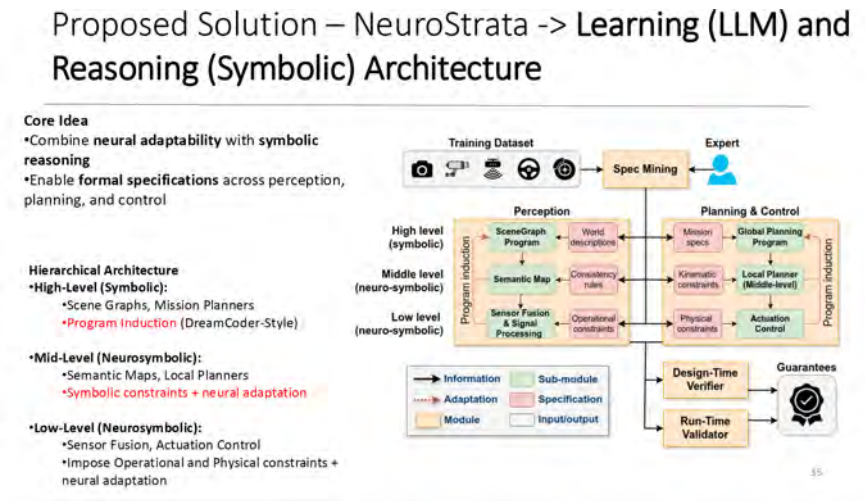

Figure 7: NeuroStrata: A hierarchical neuro-symbolic architecture that integrates learning and reasoning for safety-critical learning-enabled CPS. Foundation models and LLM-based specification mining extract formal constraints from data and expert knowledge. These specifications guide a multi-level architecture spanning symbolic high-level mission planning, mid-level neurosymbolic semantic reasoning, and low-level perception and control. Design-time verification and run-time validation provide formal guarantees, enabling interpretable, testable, and verifiable autonomous systems.

AI Safety & AI for Security

Tevfik Bultan, University of California, Santa Barbara (UCSB)

AI-enabled software systems have become a significant part of computing, creating an urgent need for techniques that evaluate and improve their safety. The speaker showed that post-learning evaluation of AI components using software testing and analysis techniques—such as coverage-guided fuzzing and quantitative symbolic execution—provides valuable insights into safety and risk. For example, the speaker's analysis of image classifiers for medical diagnosis revealed that accuracy or robustness of ML models does not always align with risk. Similarly, the speaker's analysis of ML-driven neurostimulation in visual prostheses demonstrated that violation-guided fuzzing effectively detects safety violations and offers an alternative to validation loss for evaluating ML models.

Beyond evaluation, the speaker also explored integrating AI into software testing and analysis workflows for scalable cybersecurity analysis. The speaker reported recent results showing that combining LLMs with fuzzing and symbolic execution in agentic workflows enables scalable vulnerability detection and automated proof-of-concept generation, significantly improving performance over approaches based solely on LLMs or traditional software analysis techniques.

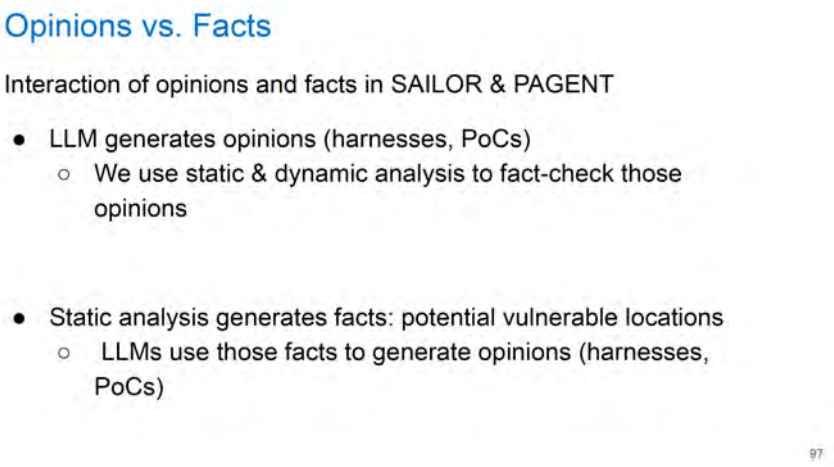

Figure 8: Interaction of opinions and facts in SAILOR and PAGENT. LLMs generate opinions such as harnesses and proof-of-concept (PoC) inputs, while static and dynamic analysis provide fact-checking. Static analysis identifies potential vulnerable locations, which in turn guide LLMs to generate improved harnesses and PoCs, forming a closed-loop neuro-symbolic workflow.

Theme 3 Keynote: Reasoning About Accountable Cyber-Physical Systems (and Reasoning, in General)

Ruzica Piskac, Yale University

As autonomous and learning-enabled cyber-physical systems become increasingly deployed in safety-critical domains, understanding and attributing responsibility for their decisions becomes essential. Traditional verification techniques rely on predefined formal specifications; however, in many real-world scenarios, legal, policy, or ethical considerations make it difficult to formalize properties ahead of time. This talk presented a complementary approach based on counterfactual reasoning and SMT-based interrogation, enabling investigators to "put the algorithm on the stand" and systematically analyze alternative decision outcomes.

The talk introduced SOID, a framework that combines symbolic execution and SMT solving to generate counterfactual scenarios and reason about agent intentions and accountability. By relaxing observed inputs and exploring alternative decision paths, the framework helps distinguish between reasonable, reckless, and pathological system behaviors.

Beyond accountability in cyber-physical systems, the talk further explored neurosymbolic approaches for automated discovery and validation of mathematical and data-driven relationships. This includes learning randomized reductions through agent-assisted symbolic reasoning, and AI-driven data proxy discovery pipelines that construct knowledge graphs validated using probabilistic model checking. These approaches illustrate how combining LLM-based exploration with formal reasoning tools can enable scalable, interpretable, and verifiable reasoning workflows for complex autonomous and data-driven systems.

Figure 9: Complementary strengths of LLM-based reasoning and formal methods highlighted in the Theme 3 keynote. While LLMs enable broad exploration and creative hypothesis generation, formal solvers and theorem provers provide rigorous grounding and verification. Combining both enables scalable and trustworthy reasoning for complex learning-enabled cyber-physical systems.

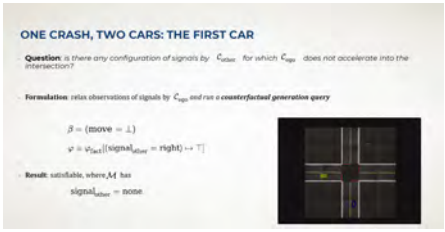

Figure 10: Counterfactual reasoning example for accountable autonomous driving decisions. By relaxing observed signals and querying alternative scenarios, SMT-based reasoning identifies configurations under which the ego vehicle would not accelerate, enabling formal investigation of responsibility and system behavior.

Understanding deep learning from the control theory vantage point

Paulo Tabuada, University of California, Los Angeles

AI has drastically changed the landscape of several engineering disciplines, such as computer vision and natural language processing, and is now an integral part of several commercial products. In this talk it was argued, through several vignettes, that ideas and techniques from control can help answer fundamental questions about the power, limitations, and societal impact of AI.

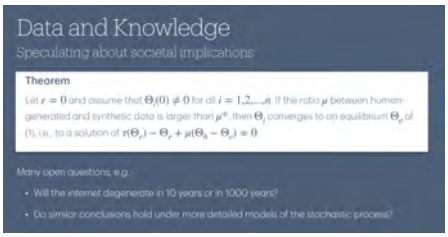

Figure 11: Data and knowledge dynamics in generative AI systems. The slide highlights a theoretical perspective on closed-loop learning, where synthetic data increasingly influences future model training. When the ratio between human-generated and synthetic data exceeds a threshold, the stochastic learning process converges to an equilibrium distribution. This raises societal concerns such as potential degeneration of information diversity and long-term stability of knowledge on the internet.

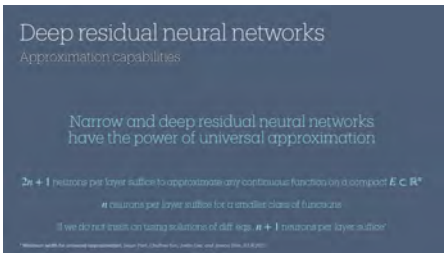

Figure 12: Approximation capabilities of deep residual neural networks. The slide illustrates that narrow but deep residual networks possess universal approximation power, where a limited number of neurons per layer can approximate continuous functions in high-dimensional spaces. This highlights the expressive power of deep architectures and their theoretical capability to represent complex system behaviors using the control theory.

Auto-formalization via Semantic Equivalence Models

Vijay Ganesh, Georgia Institute of Technology

In recent years, the field has witnessed a symbiotic trend wherein LLMs are being combined with provers, solvers, and computer algebra systems, resulting in dramatic breakthroughs in AI4math. Following this trend, the speaker has developed two lines of work in their research group. The first is the idea that "good" joint embeddings (JE) can dramatically improve the efficacy of LLM-based auto-formalization tools. The speaker defined that JEs are good if they respect the following invariant: semantically-equivalent formally dissimilar objects (e.g., pairs of sematically-equivalent natural and formal language proofs) must be "close by" in the embedding space, and semantically inequivalent ones "far apart". Such JE models were used as part of a successful RAG-based auto-formalization pipeline, demonstrating that such JEs are a critical AI-for-math technology. The second idea is Reinforcement Learning with Symbolic Feedback (RLSF), a class of techniques that addresses the LLM hallucination problem in contexts where the speaker and their team have access to rich symbolic feedback such math, physics, and code, demonstrating that they too are critical to the success of AI for math.

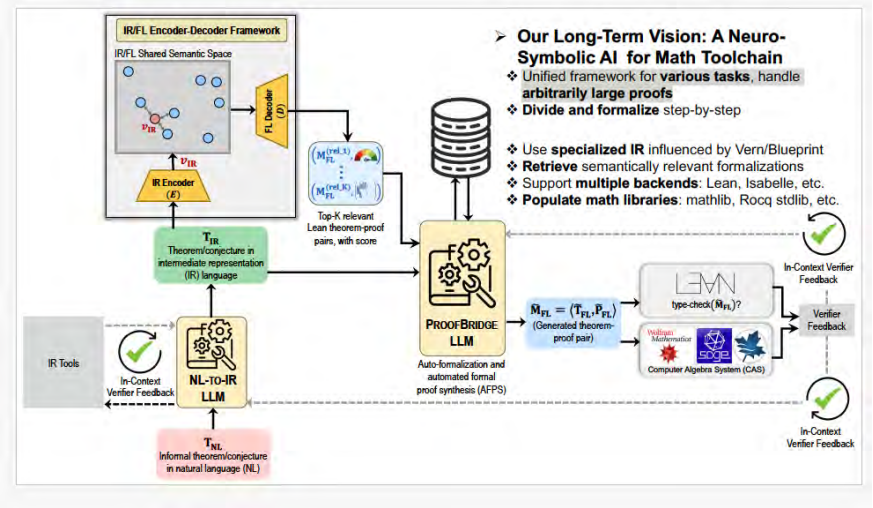

Figure 13: Neuro-symbolic auto-formalization and proof synthesis pipeline using semantic equivalence models. Natural language theorems are translated into intermediate representations, retrieved via joint embeddings, and synthesized into formal proofs with verification feedback in a closed-loop workflow.

Let's Verify ChatGPT: What Would We Verify & How Could We Get There? Bonus: Agentic Engineering is Coming

Taylor Johnson, Vanderbilt University

The talk began by looking at how recent advances in agentic AI are transforming the way cyber-physical systems are designed, built, and validated. The speaker argued that program synthesis is effectively a solved problem and that engineering synthesis is next. With agentic AI platforms, such as Claude Code, it is now plausible for a single engineer to prototype the full design of a simple cyber-physical system, such as a quadcopter, spanning mechanical CAD, PCB layout, SPICE circuit simulation, finite element and thermal analysis, computational fluid dynamics simulation, embedded systems development (VHDL, C), control design, RF analysis, and manufacturing planning: tasks that traditionally require coordinated teams of tens to hundreds of engineers across domains. The speaker demonstrated this through live examples including translating MATLAB neural network training code to Python/PyTorch, building a web-based model checker with BDD visualization and Büchi automata generation, generating PCB layouts via KiCAD and automated routing tools, running SPICE simulations of power regulators, and producing safety assurance cases in GSN, all orchestrated through agentic AI interacting with command-line tools.

The key enabling pattern is the availability of oracles: anywhere correctness can be evaluated, whether through parsing, simulation, differential testing against existing artifacts, or formal checks, agentic AI can generate, execute, and iteratively correct designs with remarkable productivity. With global AI investment on the order of NASA's entire annual budget every one to three weeks, Sutton's bitter lesson suggests these capabilities will only improve. Critically,

whether or not the community welcomes it, the next generation of engineered systems may be built this way, and the central question for the verification and validation community becomes: when a system procurer (such as a government agency) contracts with a manufacturer, how will they assure systems that were themselves designed by agentic AI? This will demand agentic AI to interrogate and red-team these designs, raising challenges in toolchain qualification, supply chain assurance, and workforce development.

This naturally leads to a second, closely related challenge: if AI is helping design complex systems, we must also ensure that the AI models themselves are trustworthy and verifiable. The speaker suggested a grand challenge to verify a realistic foundation model like ChatGPT. The neural network verification problem, given a trained network and a specification formalized as preconditions and postconditions, prove the network maps any input in the precondition set into the postcondition set, has seen significant scalability advances as demonstrated through the Verification of Neural Networks Competition (VNN-COMP), with current verification methods capable of analyzing models with on the order of 100 million parameters. However, the community has focused predominantly on image classification tasks, and the speaker argued it must shift toward NLP tasks, including sentiment analysis, hate speech classification, and guardrail models, as well as robotics applications with vision-language-action models, where different layer types and architectural patterns present new technical challenges.

More realistically than verifying a full LLM, the speaker proposed targeting fully open small language models (SLMs) such as Ai2's OLMo2 1B (open code, data, and weights) or HuggingFace's SmolLM2-135M, and the speaker noted that industry trends toward increasingly smaller, specialized models are bringing realistic foundation models within the scalability envelope of formal verification. This guiding challenge will illuminate critical needs in formal methods for specification and verification of transformer architectures, while connecting to broader questions of how foundation models can themselves serve as tools for verification, such as generating specifications, constructing models, and serving as verification oracles.

Figure 14: Future research directions for neural network verification, including system-level integration, generative AI verification, semantic foundations, and domain-specific applications toward trustworthy learning-enabled systems.

Towards Resilient Cyber-Physical-Human Infrastructures

Nalini Venkatasubramanian, University of California, Irvine

This talk presented a framework for building resilient cyber-physical-human infrastructures (CPHI) that integrate emerging IT technologies such as Internet-of-Things, cloud platforms, big data analytics, and AI/ML. These technologies enable new forms of observation, analysis, and adaptation in community-scale infrastructure systems such as water, power, and transportation networks that serve as critical lifelines for society.

Focusing on community water infrastructure as a driving use case, the talk described physical AI approaches that combine physics-based models, data-driven learning, and semantic knowledge to support resilient system planning and operation. The framework integrates structural information from infrastructure topology, behavioral insights from domain-expert simulators, and semantic data derived from community usage patterns. These integrated models enable impact-driven and phenomena-based sensor placement strategies for cost-effective monitoring and tracing of failures and attacks in potable water, wastewater, and stormwater systems.

The talk further explored adaptive middleware techniques for intelligent data collection, cross-layer data exchange, and real-time decision-making across device, network, processing, and analytics layers. By staging computation across the edge-to-cloud continuum, the proposed approach addresses trade-offs in latency, cost, and reliability under dynamic operating conditions. Finally, the talk argued that neurosymbolic methods are essential for reasoning about distributed CPHI behavior, enabling safe composition of heterogeneous techniques and supporting robust, resilient infrastructure systems for future communities.

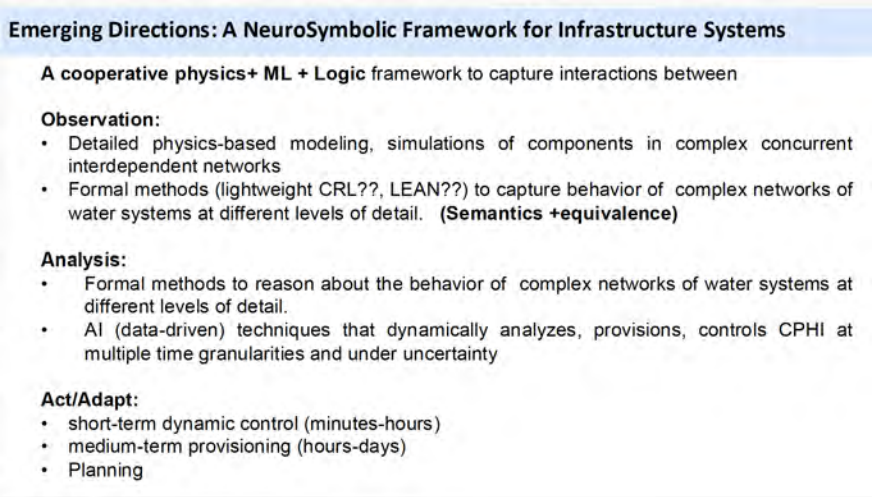

Figure 15: Neuro-symbolic framework for resilient cyber-physical-human infrastructure systems.

Using AI and Formal Methods for Legal Reasoning

Scott Shapiro, Yale Law School Ruzica Piskac, Yale University

This talk presented a neuro-symbolic framework that combines large language models with formal methods to enable accurate and explainable legal reasoning over complex regulations. The approach translates regulatory text and user queries into first-order logic representations, using uninterpreted functions to model abstract legal concepts, and applies SMT-based automated reasoning to check compliance, generate explanations, and verify conclusions. By integrating LLM-based translation with formal verification, the system reduces hallucinations, ensures consistency, and provides auditable, regulation-aware decision support for domains such as taxation, insurance, and enterprise compliance.

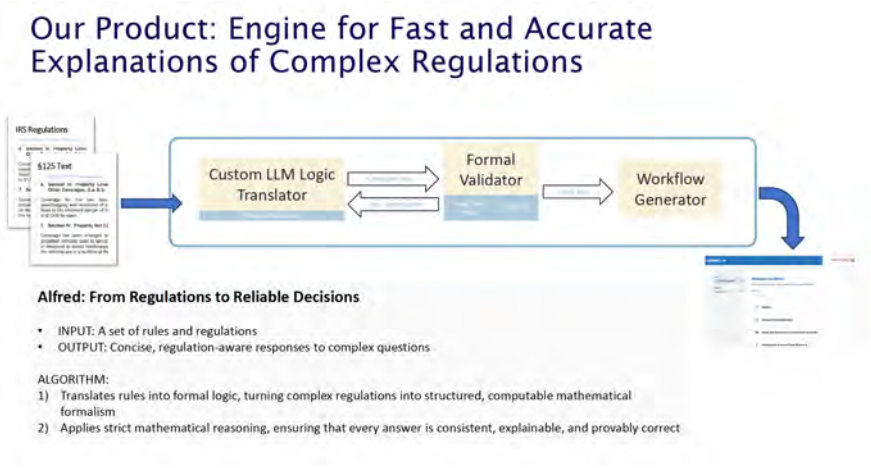

Figure 16: Neuro-symbolic pipeline for regulation-aware reasoning. A custom LLM logic translator converts regulatory text into formal representations, which are validated using formal reasoning and optimization before generating workflow-based, explainable, and verifiable decisions.

Neuro-Symbolic Reasoning in the Era of Foundation Models

Jiani Huang, University of Pennsylvania

Foundation models exhibit strong perceptual and generative abilities but often lack structured reasoning, compositionality, and verifiability. This talk presented a neuro-symbolic approach that learns spatio-temporal scene graphs from weakly supervised multimodal data. The framework first extracts object traces by segmenting and tracking entities across video frames, producing temporally grounded representations of evolving objects. These traces are used to construct scene graphs capturing entities, attributes, and relations over time. Natural language descriptions are converted into symbolic predicates using large language models, and a differentiable alignment checker based on temporal logic aligns symbolic specifications with predicted scene graphs, providing learning signals without dense annotations.

By aligning structured representations across modalities, this approach enables concept-level reasoning over dynamic environments. Semantic concepts such as objects, actions, and interactions are grounded to evolving object traces and organized into compositional scene structures. This neuro-symbolic alignment improves interpretability, compositional generalization, and robustness, offering a principled direction for enhancing reasoning capabilities in foundation model–based systems for video understanding and embodied intelligence.

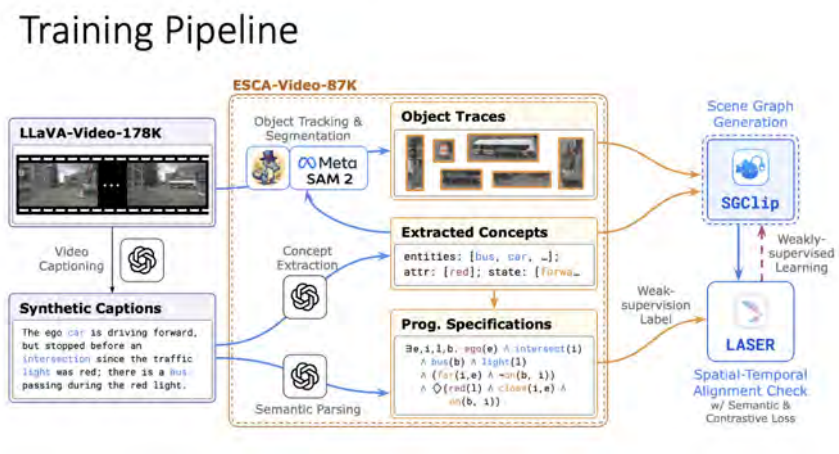

Figure 17: Training pipeline for learning spatio-temporal scene graphs with weak supervision. Videos from large-scale datasets are first captioned to generate synthetic descriptions, which are parsed to extract semantic concepts and programmatic specifications. In parallel, object tracking and segmentation using Meta SAM2 produce object traces. These object traces and extracted concepts are used to train a scene graph generator (SGCLIP). A spatial-temporal alignment module within the LASER framework performs semantic and contrastive alignment, producing weak supervision signals for learning structured representations without dense annotations.

Scaling Up Neuro-Symbolic Cyber-Security Techniques

Ziyang Li, Johns Hopkins University

Recent advances in agentic AI are transforming cyber-security by enabling scalable neuro-symbolic reasoning for vulnerability detection and program analysis. This talk presented an agentic framework that synthesizes vulnerability queries, generates instrumentation for fuzzing and symbolic execution, and integrates static and dynamic analysis to detect complex, repository-level vulnerabilities. By combining neural reasoning with symbolic analysis, the approach addresses challenges such as ambiguous vulnerability specifications, cross-language dependencies, and multi-stage exploit paths. The work highlighted how agentic systems can automate security analysis at scale while maintaining logical rigor, offering a promising direction toward more robust, explainable, and human-aligned cyber-security techniques.

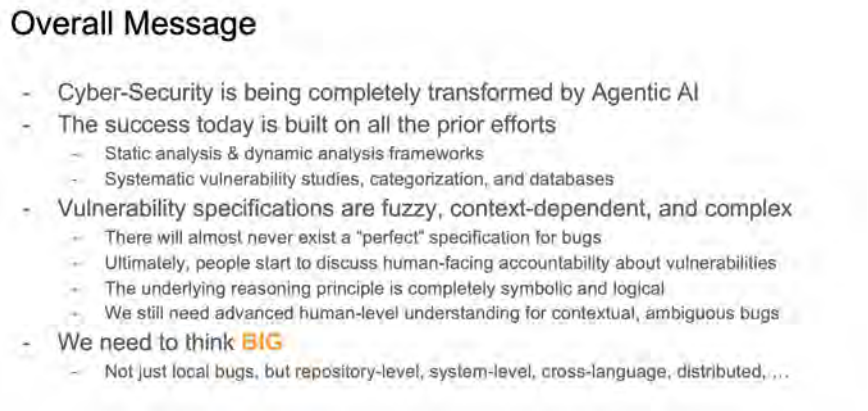

Figure 18: Overall message on scaling neuro-symbolic cyber-security with agentic AI.

Certifying AI-enabled Systems under ISO 26262 and ISO 21448

Mitra Nasri, Eindhoven University of Technology

This talk discussed the certification of AI-enabled automotive systems under ISO 26262 (functional safety) and ISO 21448 (SOTIF). It introduced the safety lifecycle, including hazard analysis, safety goals, and verification and validation, and highlights challenges arising from AI components such as ambiguous requirements, large input spaces, and rare edge cases. The talk emphasized scenario-based validation, risk mitigation strategies, and structured safety reasoning to ensure reliable deployment of learning-enabled cyber-physical systems.

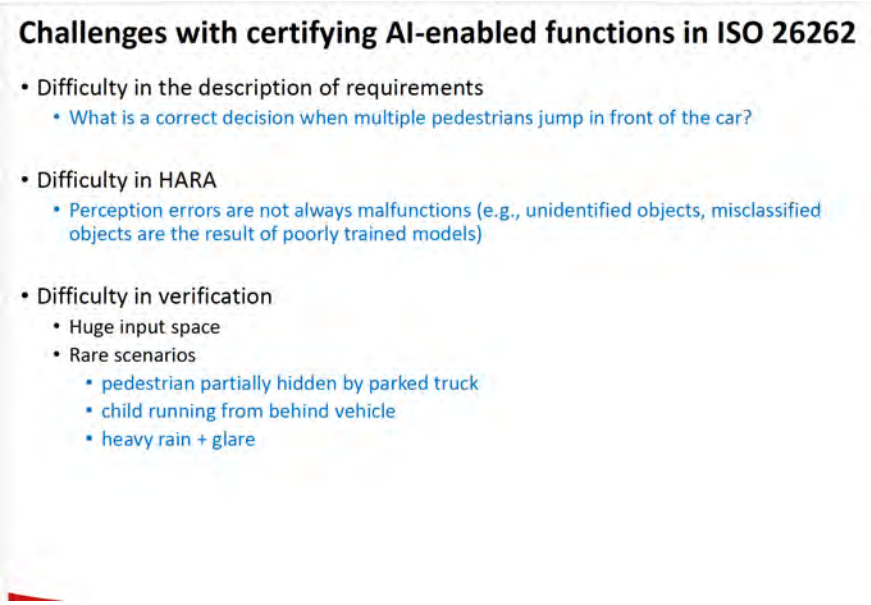

Figure 19: Challenges in certifying AI-enabled functions under ISO 26262.

List of Participants

Figure 20: Shonan 235 Participants

- Aloysius K. Mok, University of Texas at Austin
- Bryan Donyanavard, San Diego State University
- Chih-Hong Cheng, Chalmers University of Technology
- Dakai Zhu, University of Texas at San Antonio
- Daniel Neider, TU Dortmund University
- Hong Jin Kang, University of Sydney
- Ivan Ruchkin, University of Florida
- Jiani Huang, University of Pennsylvania
- Kun Chu, University of Hamburg
- Lei Ma, The University of Tokyo
- Mark Santolucito, Barnard College, Columbia University
- Mitra Nasri, Eindhoven University of Technology
- Nalini Venkatasubramanian, University of California, Irvine
- Paulo Tabuada, University of California, Los Angeles
- Ruzica Piskac, Yale University
- Sanjoy Baruah, Washington University in St. Louis

- Simon Thompson, Tier IV
- Scott Shapiro, Yale University
- Taylor Johnson, Vanderbilt University
- Tevfik Bultan, University of California, Santa Barbara
- Vijay Ganesh, Georgia Institute of Technology
- Xi (James) Zheng, Macquarie University
- Ziyang Li, Johns Hopkins University

Meeting Schedule

Check-in Day: March 8 (Sun)

- Check-in

Day1: March 9 (Mon) — Shared Context & Testing Perspectives (Theme 1 + Theme 2)

- 09:00 – 09:10 Opening Remarks
- 09:10 – 09:55 Theme 1 Keynote: *Trustworthy AI in the Era of Foundation Model Driven Agentic Systems* — Lei Ma
- 10:00 – 10:30 Morning Coffee Break
- 10:30 – 11:00 Volunteered Talk: *Confidence over Time: Confidence Calibration with Temporal Logic for Large Language Model Reasoning* — Ivan Ruchkin
- 11:00 – 11:30 Volunteered Talk: *Automata Methods for Neurosymbolic AI* — Mark Santolucito
- 11:30 – 13:30 Lunch
- 13:30 – 14:15 Theme 2 Keynote: *Foundation Models for Safety Testing of Learning-Enabled CPS* — Xi Zheng
- 14:15 – 14:45 Volunteered Talk: *AI Safety & AI for Security* — Tevfik Bultan
- 15:00 – 15:30 Afternoon Tea
- 15:30 – 18:00 Self-introduction and Breakout Discussions (Theme 1 + Theme 2)
- 18:00 onward Dinner

Day2: March 10 (Tue) — Reasoning Perspectives (Theme 3)

- 09:00 – 09:45 Theme 3 Keynote: *Reasoning About Accountable Cyber-Physical Systems* — Ruzica Piskac
- 10:00 – 10:30 Morning Coffee Break
- 10:30 – 11:00 Volunteered Talk: *Understanding Deep Learning from the Control Theory Vantage Point* — Paulo Tabuada
- 11:00 – 11:30 Volunteered Talk: *Auto-formalization via Semantic Equivalence Models* — Vijay Ganesh
- 11:30 – 13:30 Lunch
- 13:30 – 14:00 Volunteered Talk: *Let's Verify ChatGPT: What Would We Verify & How Could We Get There?* — Taylor Johnson
- 14:00 – 14:30 Volunteered Talk: *Towards Resilient Cyber-Physical-Human Infrastructures* — Nalini Venkatasubramanian

- 15:00 – 15:30 Afternoon Tea
- 15:30 – 17:00 Breakout Discussions (Theme 3)
- 17:00 – 18:00 Plenary Discussion: Identifying Key Themes and Directions for Day-3 Breakout Groups
- 18:00 onward Dinner

Day3: March 11 (Wed) — Flexible Exploration

- 09:00 – 10:00 Volunteered Talk: *Using AI and Formal Methods for Legal Reasoning* — Scott Shapiro & Ruzica Piskac
- 10:00 – 10:30 Morning Coffee Break
- 10:30 – 11:00 Volunteered Talk: *Scaling Up Neuro-Symbolic Cyber-Security Techniques* — Ziyang Li
- 11:00 – 11:30 Volunteered Talk: *Neuro-Symbolic Reasoning in the Era of Foundation Models* — Jiani Huang
- 11:30 – 13:30 Lunch
- 13:30 – 17:00 Visit Jomyoji Temple and Hokokuji Temple with Japanese Tea Ceremony
- 18:00 onward Dinner

Day4: March 12 (Thu) — Consolidation & Outcomes

- 08:30 – 09:00 Check-out
- 09:00 – 09:10 Volunteered Talk: *Certifying AI-enabled Systems under ISO 26262 and ISO 21448* — Mitra Nasri
- 09:10 – 10:00 Breakout Group Reports
- 10:00 – 10:30 Morning Coffee Break
- 10:30 – 11:30 Cross-theme Integration Discussion
- 11:30 – 12:00 Agreement on Follow-up Outputs and Collaborations

Summary of discussions

The seminar discussions converged on a shared concern across all three themes: although foundation models and large language models are creating new opportunities for autonomous systems, testing, and verification, the main obstacle to trustworthy deployment is no longer raw capability but the lack of principled methods to specify, analyse, and assure the behaviour of complex learning-enabled systems. Participants repeatedly returned to the gap between strong empirical performance in controlled settings and the much harder problem of providing dependable guarantees under distribution shift, system integration complexity, and real-world uncertainty.

Within Theme 1, discussions on autonomous driving and complex AI systems engineering emphasized scalability as the central unresolved challenge. Current systems can be engineered to perform safely within narrow operational design domains, but extending them across different geographies, weather conditions, road types, and regulatory contexts still requires extensive re-engineering. This led to broader discussion of the symbolic–physical gap: AI systems must convert physical sensor data into symbolic abstractions, reason over those abstractions, and then map decisions back into physical control, with possible information loss at each boundary. The group also discussed the growing engineering burden of treating data as a de facto specification, since the effective behaviour of deployed AI systems shifts with data distributions, environments, and runtime conditions rather than only through code changes.

Within Theme 2, the discussion focused on how to make safety testing of learning-enabled CPS more formal and dependable. Participants agreed that existing neural coverage metrics remain too weak and poorly grounded to support meaningful safety arguments, while full formal verification of realistic ML pipelines remains out of reach. A major line of discussion therefore centred on neuro-symbolic architectures that separate neural perception from symbolic world modelling and reasoning. This decomposition enables assume–guarantee style thinking, clearer fault localization, and the possibility of certifying different pipeline components independently. The discussion also highlighted conformance-based testing as a principled alternative to purely black-box scenario testing, with LLMs playing useful roles in specification mining, test generation, design-space exploration, and symbolic execution support, provided they are treated as constrained components rather than trusted oracles. Runtime enforcement and verified safety envelopes were discussed as practical complements to design-time assurance.

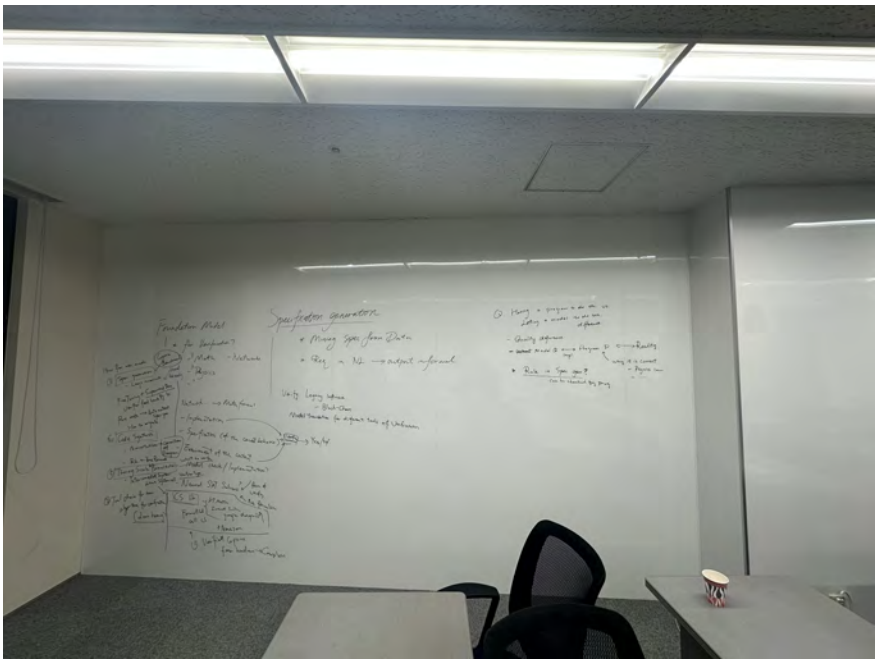

Figure 21: Discussion on foundation models for verification and testing, including specification generation, model construction, legacy software formalization, and rule extraction from domain knowledge.

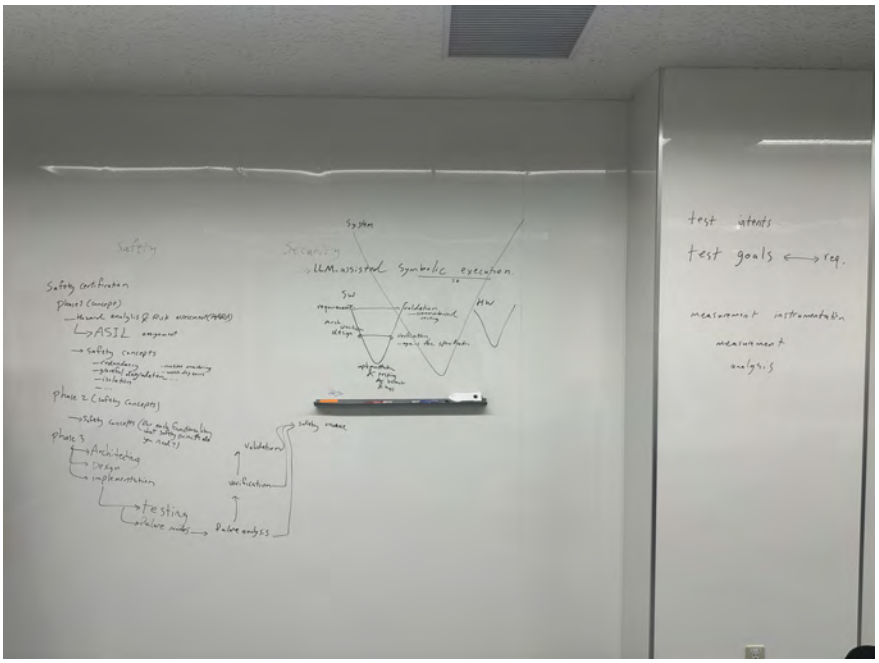

Figure 22: Safety certification workflow and LLM-assisted symbolic execution within a V-model development lifecycle, including testing goals, measurement, and verification activities.

Within Theme 3, participants examined the role of foundation models in formal verification through four lenses: specification generation, model construction, verification oracles, and verification of foundation models themselves. A substantial conceptual discussion clarified that foundation models and LLMs should not be treated as interchangeable, and that verification tasks differ depending on whether one is verifying a classical implementation against a formal specification or reasoning about a model that only approximates a solution statistically. Discussions of LLM-assisted invariant generation, specification mining, code translation, certificate production, and guardrail-model verification showed both promise and clear theoretical limits. Participants repeatedly distinguished tractable near-term tasks such as syntactic well-formedness checking, annotation generation, or small guardrail-model verification from harder problems such as semantic program equivalence, theory-scale auto-formalization, and verification of general NLP behaviour.

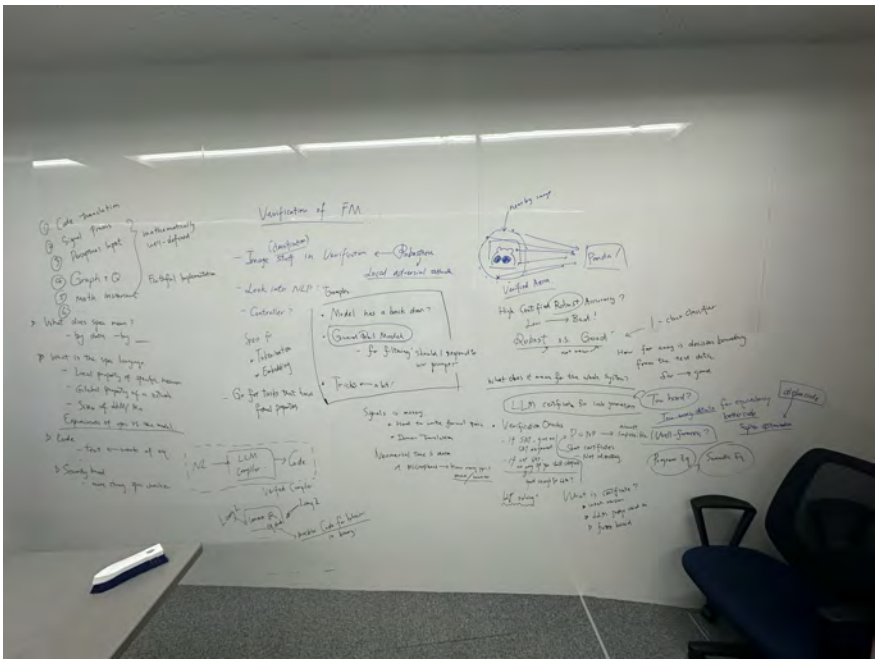

Figure 23: Verification of foundation models, including code translation, robustness analysis, guardrail models, certificate-based verification, and equivalence checking challenges.

Across the seminar as a whole, a recurring discussion theme was the need to regain intellectual and engineering control over increasingly complex AI-enabled systems. This included not only formal reasoning and compositional verification, but also better interfaces for human oversight, runtime confidence management, and system-level evidence that supports trust calibration rather than relying on raw performance metrics alone. The shared view was that future trustworthy CPS will require learning and reasoning to be co-designed, rather than treating formal methods and foundation models as isolated components.

Summary of new findings

A major outcome of the seminar was a clearer separation between near-term technically tractable research targets and long-term aspirational goals. Rather than treating "verification of foundation models" as a single monolithic problem, the discussions identified a hierarchy of tasks with different levels of difficulty. In particular, code translation, guardrail-model verification, specification generation for structurally formal tasks, and LLM-assisted annotation generation emerged as realistic near-term targets, while full semantic verification of general-purpose foundation models remains a grand challenge. This sharpening of scope is itself an important finding because it provides a more actionable research agenda than treating FM verification as an undifferentiated objective.

A second important finding was that symbolic world models provide a practical bridge between uninterpretable neural perception and formally analysable downstream reasoning. In the Theme 2 discussion, participants repeatedly returned to the idea that explicit symbolic world-model construction can turn opaque end-to-end pipelines into decomposable systems in which coverage, monitoring, debugging, and rule checking become meaningful. Symbolic world-model coverage was articulated as a more semantically grounded alternative to neuron-activation coverage, and runtime safety envelopes built around symbolic abstractions were recognised as a realistic mechanism for achieving partial guarantees even when perception itself remains probabilistic. This finding reinforces the seminar's broader view that intermediate structured representations are central to dependable learning-enabled CPS.

A third finding was the clarification of the role that LLMs can legitimately play in formal workflows. Across the discussions, the group converged on the view that LLMs are most useful as assistants for bottleneck tasks such as natural-language-to-formal-specification translation, loop invariant proposal, model translation, symbolic execution scaffolding, environment synthesis, and certificate generation for restricted problem classes. At the same time, the seminar clarified that LLM assistance must not be conflated with formal soundness. This distinction between certificate-producing AI and fully certified AI, and between LLM-supported reasoning and formally guaranteed reasoning, was made much sharper during the seminar and provides a useful conceptual foundation for future work.

A fourth finding was that verification targets for foundation models should be chosen according to the formal structure of the task domain. Tasks such as code generation, code translation, signal processing, mathematical reasoning, and small safety filters admit more precise specifications and verification oracles than open-ended NLP tasks. This led to the practical conclusion that FM verification should initially prioritize domains where output correctness can be stated formally and checked independently, rather than starting from unconstrained generative language behaviour. The identification of guardrail models as small, safety-critical, interface-level components suitable for near-term verification was a concrete example of this strategy.

Finally, the seminar produced an emerging set of community-scale challenge problems that can structure future work. These included verified LLM-based code translation, verified LLM-based planners for CPS, verification of scientific foundation models through physical invariants, and the longer-term vision of verifying an LLM or FM within a constrained and formally specified domain.

Framing these as a hierarchy of grand challenges, analogous to prior verified-compiler and verified-kernel efforts, was an important consolidating outcome of the meeting.

Figure 24: On the final day of the seminar, participants were fortunate to witness a clear view of Mount Fuji from the Shonan Village Center. The scenic view provided a memorable conclusion to the discussions and collaborative activities during the seminar.

Identified issues and future directions

The seminar identified several unresolved issues that currently limit the trustworthy deployment of foundation-model-assisted CPS. The first is the persistent specification bottleneck. Across all three themes, participants noted that the main difficulty is often not only verifying a system against a known specification, but obtaining a specification that is complete, precise, and meaningful in the first place. For autonomous systems, informal requirements such as "avoid collisions" are far too weak; for data-driven systems, the effective specification drifts with changing data distributions; and for NLP-centric foundation models, even defining what constitutes correct behaviour may be ambiguous. Future work must therefore focus on interactive specification generation, domain-specific specification languages, and human-in-the-loop refinement workflows that bridge informal intent and formal reasoning.

A second major issue is the lack of scalable verification infrastructure for modern AI-enabled pipelines. Full white-box verification of realistic perception stacks remains out of reach, while current robustness-centric approaches often provide component-level guarantees that do not align with the system-level properties practitioners actually care about. In Theme 2, participants identified compositional assurance through symbolic world models and runtime safety envelopes as promising directions, but these approaches still require better theories for interface contracts, uncertainty propagation, and multi-sensor reasoning. In Theme 3, analogous scalability issues arose in theory-scale auto-formalization, certificate generation for program equivalence, and verification over richer constraint domains such as ILP rather than SAT. Future research should therefore target compositional methods, modular certificates, and scalable abstraction strategies for realistic systems rather than isolated toy benchmarks.

A third issue is that the interaction between foundation-model generation and formal verification remains poorly understood. Verifier-feedback fine-tuning is attractive in principle, but current verifier outputs are sparse and binary, making reinforcement-style learning difficult. Likewise, the mechanistic reasons why FM-generated annotations, invariants, or certificates succeed on some instances and fail on others remain opaque. The seminar therefore identified reward shaping, curriculum learning, counterexample-guided refinement, and more transparent neuro-symbolic architectures as important future directions. More broadly, the group recognised the need for benchmark suites that compare LLM-assisted approaches with expert-only baselines on realistic tasks such as invariant synthesis, specification mining, code translation, and vulnerability analysis.

A fourth issue concerns NLP-specific and multimodal verification foundations. Moving from image-classifier robustness to language and multimodal systems introduces unresolved technical questions around tokenization, embedding-space semantics, equivalence classes over discrete inputs, and the meaning of correctness for text tasks. This makes direct transfer of existing neural-network verification methods inadequate. The group therefore identified formal semantics for tokenization, metrics that preserve semantic equivalence in embedding spaces, and machine-checkable task definitions for constrained NLP domains as prerequisites for progress. Near-term work should prioritise domains with clear formal structure, such as code, mathematics, and interface-level guardrail models, while using these settings to build reusable tooling and theory.

A fifth issue is the need for stronger engineering, standards, and human-systems perspectives. Participants stressed that trustworthy deployment is not only a question of solver technology or model architecture, but also of system integration, operator handoff, runtime health monitoring, and standards development. Human oversight in safety-critical settings requires interpretable process-level evidence, predictive handoff strategies, and explicit understanding of model competence boundaries. Existing standards remain immature for AI components, especially in safety-critical domains. Future work should therefore link formal methods, AI engineering, and standards efforts more directly, so that verification artefacts become usable not only in research prototypes but also in certification, operations, and regulatory practice.

Finally, the seminar identified a long-term research direction in the form of shared grand challenges. The most realistic path forward is likely a staged agenda: begin with tractable tasks such as syntax and well-formedness checking, guardrail-model verification, symbolic world-model coverage, certificate-producing code translation, and specification generation for formally structured domains; then scale toward verified planners, verified scientific foundation models, and eventually constrained-domain verification of FM-based systems themselves. This staged strategy aligns ambition with tractability and provides a concrete roadmap for future collaboration, benchmark building, and community-wide progress.