Auditing Formal Verification of AI Model Weight Security Through Debate

Marcello Maugeri[1] Ivan Beschastnikh[2] David Williams-King[1]

[1]ERA, Cambridge, United Kingdom [2]University of British Columbia, Vancouver, Canada

Motivation

1. The capabilities of an AI model reside in its *weights*; whoever obtains them can exploit the model for their own purposes. While this poses a substantial financial risk for the provider, in critical domains – such as Cybersecurity and Biotechnology – stealing model weights enables *misuse*, which can pose a critical risk.
2. Thus, *RAND* defines five *Security Levels*, each specifying progressively stronger requirements for AI model weight security. To achieve a level, a frontier lab must implement the corresponding controls at both the organisational and infrastructure levels.
3. Implementing every control and verifying that each one holds requires a substantial effort. Furthermore, demonstrating compliance to an external auditor is an even more challenging task, since the auditor might only have *black-box* access to the infrastructure.
4. *Formal Verification* can support the audit process. However, its evidence is as strong as the *Formal Specification*, which must describe the implementation faithfully and completely. Moreover, the auditor must interpret the specification to determine whether it addresses the requirement.

Background

Security policy a set of security requirements a system must meet.

Design: the infrastructure is designed to implement the controls that meet the requirements

Implementation the infrastructure as built and deployed, controls included.

Formalisation: whether the specification reflects the implementation and its properties express the requirements remains open

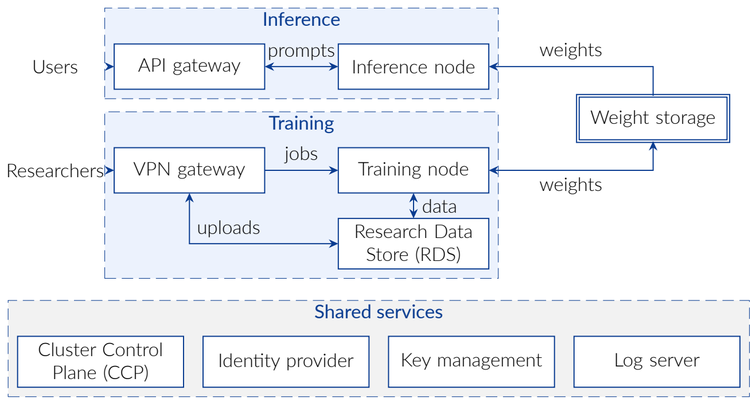
Formal specification describes the behaviour of the implementation and states its properties.

Formal verification: the properties are verified on the specification, which becomes evidence that the implementation meets the requirements

Assurance case claims of compliance with the evidence that supports them.

Reference Implementation: SL3 Frontier Lab

To support our research, we build (a portion of) the infrastructure of a frontier lab. The lab is designed to implement the controls that achieve Security Level 3 (SL3).



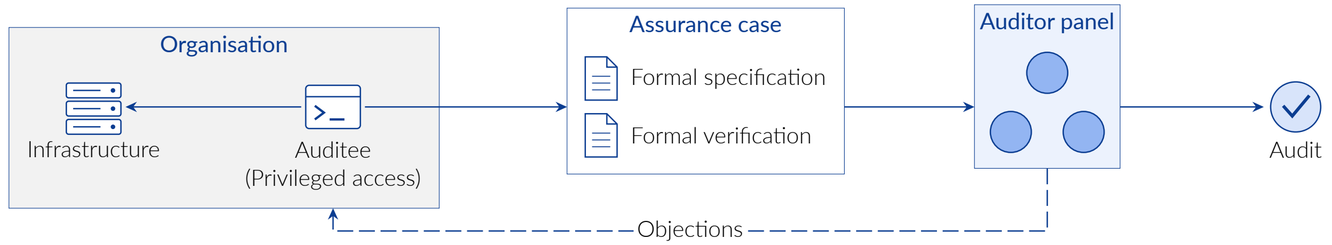
Weight Storage (PR.DS)

- Centralised and restricted management of weight storage
- Secure cloud network (if applicable)
- Dedicated devices for weights and weight security data

Permitted Interfaces (PR.AC)

- Authorised users who interact with the weights do so only through a software interface that reduces risk of the weights being illegitimately copied.
- Any code accessing the weights minimises attack surface, provides only simple forms of access, and uses the minimal amount of (highly trusted and well-established) external code necessary.

Approach



Method

Auditee

Has privileged access to the infrastructure. In a nutshell, it formalises the infrastructure model to check whether the security requirements holds. Then, it submits the *assurance case* containing the formal specification and its formal verification.

During the audit, the panel considers the following criteria to decide whether to accept or reject the claim of compliance:

Criterion	Definition	Example gap
Specification Against Implementation		
Conformance	The specification reflects the actual behaviour of the implementation.	The Training node modelled inside the Weight storage.
Soundness	Every behaviour of the implementation is captured by the specification.	A trace of the Inference node caching the weights on its local disk.
Specification Against Requirement		
Coverage	Every clause of the requirement is expressed by a verified property.	Properties confined to where copies reside, leaving central management open.
Sufficiency	Under the stated assumptions, the verified properties imply the requirement.	One administrator holding two of the three approval roles.

Auditor panel

The panel inspects the formal specification for its conformance and soundness. Then, it checks whether the formal verification provides coverage and sufficiency against its interpretation of the security requirement. Multiple auditors debate for and against each criterion. The judge listens to the debaters, choosing whether to accept the case or return the objections to the Auditee.

Research questions

RQ1

Given a system designed to meet a security requirement, how can it be modelled to verify that the requirement is met?

RQ2

How can the trustworthiness of an audit be improved when the auditor has limited access to the system?

Evaluation

- We evaluate on the reference implementation, one SL3 control at a time. For example, a security control for the *Weight storage* is the following:

> Assign model weight storage to dedicated volumes or filesystems logically or physically separated from general-purpose data.

- **RQ1:** We write the formal specification (e.g. in TLA$^+$) of the lab and its verification as the *ground truth*. Then, a large language model (LLM) acts as an Auditee, and we check whether its specification is semantically equivalent to ours.
- **RQ2:** We generalise the audit process to an Auditor panel of multiple LLMs that debate the assurance case, and compare its findings with a single (LLM) auditor given the same evidence.
- Together, the two test whether formal specification and debate complement each other: the specification makes compliance checkable, and the debate lets an auditor with limited access carry out that check.

Contribution

- A reference implementation of SL3 weight-security measures for further research.
- A measure of how faithfully an LLM Auditee can model an infrastructure to verify a security requirement.
- A measure of whether a panel of LLM auditors audits that evidence more reliably than a single auditor with the same (limited) access.
- An audit procedure based on four criteria that apply to any verification technique and to other security requirements.

References

[1] S. Nevo et al. *Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models*. Tech. rep. 2024.
[2] J. Aguirre et al. *Achieving AI Model Weight Security Level 3 (SL3)*. Tech. rep. 2026.
[3] S. F. Comer et al. *Highly Secure Inference Data Centers: A Vertically Integrated Strategy for Security Engineering*. Tech. rep. 2026.
[4] L. Thiergart et al. *SL5 Standard for AI Security*. 2026. arXiv: `2605.08449 [cs.CR]`.
[5] S. Casper et al. "Black-Box Access is Insufficient for Rigorous AI Audits". In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2024.
[6] G. Irving et al. *AI safety via debate*. 2018. arXiv: `1805.00899 [stat.ML]`.
[7] A. Khan et al. "Debating with More Persuasive LLMs Leads to More Truthful Answers". In: *Proceedings of the 41st International Conference on Machine Learning*. PMLR, 2024.

SL3FrontierLab

Contacts

This work receives support from ERA through the Frontier AI Security Residency.

marcellomaugeriphd@gmail.com