On Distilling Foundation Models Toward Formal Languages: A Practical Path Toward VML-Centric Reasoning Systems

Thomas Scheuerle
Technical Working Group for Formal Automotive Systems (TWG·FAS)

August 2026

Abstract

Recent discussions regarding formal-language-centric foundation models have suggested that substantial reductions in model size, training cost, and inference complexity may be achievable by training predominantly on a formal language. A major open question is whether such systems must be trained from scratch or whether they can instead be obtained through knowledge distillation from already existing large-scale foundation models. This paper argues that distillation may provide a practical intermediate path. Rather than constructing a model exclusively from formal-language data, a large general-purpose model can serve as a semantic teacher whose capabilities are progressively compressed into a formal-language-centric student model. We analyze the theoretical advantages, expected efficiency gains, potential losses of capability, and implications for Vehicle Machine Language (VML) and other formal specification languages. Finally, we explore the possibility that future systems may not merely emit formal languages but may eventually reason through formal representations as their primary cognitive substrate.

1 Introduction

Modern foundation models derive much of their capability from training on massive natural-language corpora. These models acquire broad world knowledge, contextual understanding, analogy formation, and emergent reasoning abilities. At the same time, natural language contains significant redundancy, ambiguity, synonymy, and syntactic variation. Formal languages remove much of this variability and therefore offer an appealing basis for building efficient domain-specific systems. An obvious strategy would be to train a model predominantly on a formal language from the beginning. However, such an approach risks losing precisely those emergent capabilities that make large foundation models useful. This motivates an alternative question:

> Can a foundation model first acquire broad semantic competence through natural-language training and then be distilled into a smaller formal-language-centric model?

Knowledge distillation[1] provides exactly such a mechanism.

2 Distillation as Semantic Compression

Traditional distillation transfers the behavior of a larger "teacher" model into a smaller "student" model. Let

$$T(x)$$

[1] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," NIPS Deep Learning Workshop, 2015. Often regarded as the foundational publication introducing modern neural network distillation.

denote the teacher and

$$S(x)$$

the student. Rather than learning directly from ground-truth labels, the student learns to approximate the behavior of the teacher:

$$S(x) \approx T(x)$$

over a large set of inputs. The key observation is that the teacher may possess considerably richer knowledge than the training corpus itself. The student can therefore inherit compressed representations of that knowledge. For formal-language engineering, the process can be viewed as a form of semantic compression:

$$\text{Natural Language} \rightarrow \text{Latent Semantics} \rightarrow \text{Formal Language}$$

Rather than compressing raw text, the student compresses semantic concepts that have already been learned by the teacher.

3 A Distillation Pipeline for VML

A possible VML-centric training architecture is shown below.

$$\text{Foundation Model} \rightarrow \text{VML Generation} \rightarrow \text{Verification} \rightarrow \text{Distillation} \rightarrow \text{VML Specialist}$$

The teacher model generates:

- VML specifications,
- requirement translations,
- proof sketches,
- counterexamples,
- execution traces.

The resulting artifacts become training material for the student. This differs fundamentally from traditional supervised learning because the student does not merely imitate outputs. Instead, it learns compressed forms of semantic structure already present within the teacher.

4 Potential Advantages

4.1 Retention of World Knowledge

A formal-language-only training corpus may omit substantial information about causality, explanations, engineering rationale, and stakeholder intent. A teacher trained on broad natural-language sources has already acquired parts of this information. Distillation allows portions of that semantic structure to survive the compression process. As a result, a VML-focused student may inherit engineering intuition without needing to encounter all original natural-language data.

4.2 Reduced Data Requirements

Current formal languages possess training corpora that are tiny compared to natural-language datasets. Training a competitive model entirely on VML would likely require the creation of enormous quantities of synthetic data. Distillation changes this situation. The teacher can generate:

- specifications,
- examples,
- counterexamples,
- formal proofs,
- trace explanations.

This effectively transforms a scarcity problem into a synthesis problem.

4.3 Lower Computational Cost

Model distillation is often capable of preserving much of a teacher's performance in significantly smaller architectures.[2] If VML itself also reduces linguistic entropy, the resulting efficiency gain may be multiplicative rather than additive.

5 Potential Drawbacks

5.1 Teacher Error Propagation

Formal languages are often highly sensitive to individual symbols. Consider:

$$G(p \to q)$$

versus

$$G(p \to \neg q)$$

A single token changes the semantics entirely. If the teacher systematically produces subtle logical errors, distillation may amplify rather than eliminate them.

5.2 Residual Natural-Language Bias

A student may produce valid formal syntax while still depending on heuristics inherited from natural-language training. This raises an important distinction:

1. speaking a formal language,
2. reasoning through a formal language.

The first may be achieved relatively easily. The second remains largely unsolved.

[2] V. Sanh et al., "DistilBERT, a Distilled Version of BERT," arXiv:1910.01108, 2019. A widely known example showing that large language representations can be compressed while retaining much of their effectiveness.

5.3 Loss of Emergent Capabilities

Compression inevitably discards information. Capabilities such as:

- analogy formation,
- transfer learning,
- broad contextual reasoning,
- open-ended problem solving,

may degrade significantly in very small student models. This phenomenon has been observed repeatedly during language-model compression.[3]

6 Beyond Token Distillation

Most current distillation techniques focus on matching output probabilities. Future research may aim at transferring internal reasoning structures. Consider the following abstraction hierarchy:

$$\text{Natural Language} \rightarrow \text{Semantic Graph} \rightarrow \text{VML} \rightarrow \text{Proof Object}$$

Rather than distilling only generated tokens, one might attempt to distill intermediate representations. Such an approach resembles recent developments in chain-of-thought and reasoning-oriented training methods,[4] although formal-language systems could potentially provide much stronger correctness guarantees.

7 Verification-Guided Distillation

Formal languages offer a unique advantage over ordinary natural language: generated artifacts can often be verified. A training loop may therefore take the form

$$\text{Generate} \rightarrow \text{Verify} \rightarrow \text{Counterexample} \rightarrow \text{Retrain}$$

The student is no longer optimized solely against human labels. Instead, training signals emerge from mathematically defined correctness criteria. This creates a potential bridge between machine learning and formal methods.[5]

8 Implications for Vehicle Machine Language

Vehicle Machine Language (VML) provides an especially suitable domain for testing distillation-based approaches. VML specifications naturally connect:

- requirements,
- behavioral constraints,
- execution traces,
- simulations,

[3] M. Gordon et al., "Compressing BERT: Studying the Effects of Weight Pruning on Transfer Learning," RepL4NLP, 2020. Shows that aggressive compression often degrades downstream generalization capabilities.

[4] J. Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," NeurIPS 2022. Demonstrates that exposing intermediate reasoning steps can significantly improve problem-solving performance.

[5] E. Clarke, O. Grumberg, and D. Peled, *Model Checking*, MIT Press, 1999. A foundational reference for formal verification and counterexample generation techniques.

- verification artifacts.

This alignment allows a teacher model to generate not merely formal expressions but complete semantic ecosystems around those expressions. The resulting student may become substantially smaller while retaining the ability to reason about vehicle behavior, requirement consistency, and safety constraints. Such a system could eventually act as a specialized engineering foundation model whose primary output language is VML.

9 The Formal Reasoning Hypothesis

The most intriguing possibility is that formal languages may eventually become more than output formats. A sufficiently expressive language may evolve into an internal reasoning substrate. In that scenario, future models would not translate natural-language thoughts into VML. Instead, they would construct and manipulate formal semantic structures directly. Natural language would become merely a user interface while formal representations would constitute the model's principal reasoning medium. Whether such a transition is possible remains an open research question.

10 Conclusion

Distillation offers a practical path toward formal-language-centric foundation models. Rather than abandoning the semantic richness acquired through natural-language pretraining, a large foundation model may serve as a semantic teacher whose capabilities are compressed into a specialized formal-language student. This approach offers several advantages, including reduced training cost, improved sample efficiency, and partial retention of world knowledge. However, it also introduces risks related to teacher error propagation, capability loss, and residual dependence on natural-language-derived reasoning. Future research should focus on experimentally quantifying these tradeoffs and on determining whether formal languages can eventually become not only the preferred output format of intelligent systems but also their native medium of reasoning.