MAGNET: Mathematical Assurance of Generative AI Network Evaluation Toolkit

Jon Crall **David Joy** **Roderic Collins** **Benjamin Fenelon** **Anthony Hoogs**

Brian Hu

Kitware, Inc., Clifton Park, NY
{jon.crall,david.joy,roderic.collins,
benjamin.fenelon,anthony.hoogs,brian.hu}@kitware.com

Abstract

The DARPA AI Quantified (AIQ) program seeks to establish mathematical foundations for predicting when AI models will succeed or fail and why. Unlike conventional benchmarks which evaluate model capabilities, AIQ emphasizes the evaluation of theoretical claims about model generalization: given assumptions, do theoretical guarantees hold under empirical tests? This paper presents an early-stage vision for the Mathematical Assurance of Generative AI Network Evaluation Toolkit (MAGNET), an open framework designed to map theoretical claims to empirical evaluations. While MAGNET is still in the prototype phase, we describe how it will represent claims through structured evaluation cards and execute reproducible experiments to verify or falsify those claims. If successful, MAGNET will allow practitioners to encode a theoretical claim in an evaluation card and rapidly test it on relevant benchmarks at scale, lowering the barrier from theoretical proposal to empirical validation. By articulating a vision for MAGNET at the outset of AIQ, we aim to stimulate community discussion and enable a virtuous cycle connecting theoretical and empirical work on model generalization. Active development is underway at `https://github.com/AIQ-Kitware/aiq-magnet`.

1 Introduction

Recent advances in state-of-the-art AI have shown that it is possible to achieve remarkable results on text, vision, and multimodal tasks [1]. However, benchmark wins do not necessarily guarantee reliable model performance in high-stakes deployments. While conventional leaderboards have advanced the breadth and comparability of model evaluation [2, 3, 4, 5, 6], they are not designed to test theoretical claims about model generalization. As a result, there is a need for novel approaches to predict *what* capabilities models have, *when* models will generalize, *why* they sometimes fail, and *how* theoretical guarantees about model generalization may transfer across data modalities and model scales.

The DARPA Artificial Intelligence Quantified (AIQ) program [7] emphasizes the evaluation of theoretical claims themselves: given a set of assumptions, does predicted model generalization hold under rigorously controlled evaluations? AIQ is comprised of two coordinated thrusts: Technical Area 1 (TA1), which develops mathematical theories, and Technical Area 2 (TA2), which tests those theories at scale. Under TA2, we introduce the Mathematical Assurance of Generative AI Network Evaluation Toolkit (MAGNET), a framework in the prototype phase — illustrated in Figure 1 — that will map theoretical claims to empirical evaluations. MAGNET aims to reduce friction between theory and empirical validation by providing standardized evaluation cards specifying assumptions, guarantees, and corresponding tests.

Accepted at the NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle.

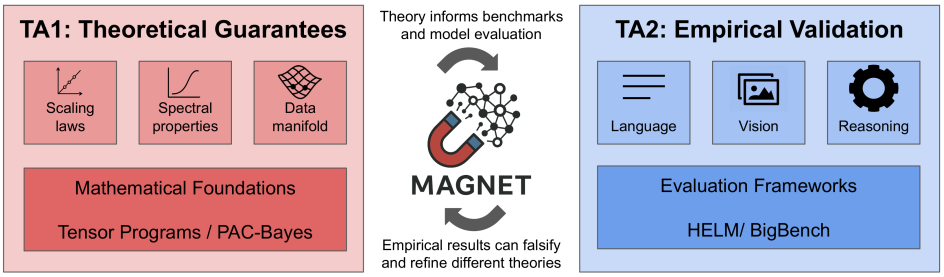

Figure 1: As part of the DARPA AIQ program, MAGNET bridges the gap between theoretical (TA1) and empirical (TA2) work on model generalization, creating a virtuous cycle that enables refinement of mathematical theories while enhancing traditional benchmarking efforts. Please see text for more details.

Our contributions are threefold: (1) We frame the evaluation of model generalization as theoretical claim verification — distinguishing our approach from conventional model benchmarking; (2) We introduce MAGNET — an open framework for translating theoretical claims into structured, reproducible evaluation protocols; and (3) We outline a research agenda — scaling theory-driven evaluations into standardized tests to support systematic progress in model benchmarking and evaluation.

2 Related Work

Benchmarks. Large-scale evaluations such as GLUE [2], SuperGLUE [3], MMLU [4], BIG-bench [5], and HELM [6] have established community standards for multitask and longitudinal benchmarking. For a comprehensive overview on model benchmarking, we refer the reader to Ni et al. [1]. These frameworks generally emphasize breadth and comparability, but they do not directly test mathematical predictions or guarantees of model generalization. We aim to build on these existing frameworks, using them as tools to produce the benchmark data needed to empirically validate mathematical claims at scale.

Theoretical guarantees. Key flavors of theoretical claims in the literature are briefly described:

- *Scaling laws.* Predictable power-law relations between loss and scale have been established [8, 9], with extensions toward predicting downstream model performance [10, 11, 12].

- *Generalization bounds.* Predictors of model generalization based on spectral norms and margins [13, 14], sharpness [15, 16], PAC-Bayes [17, 18], compression [19], and complexity measures [20] have been widely studied. More recent work links generalization to heavy-tailed spectra [21] and kernel consistency [22].

- *Adversarial Robustness.* Certified defenses such as randomized smoothing provide formal guarantees of model performance against adversarial perturbations [23], while RobustBench and AutoAttack standardize adversarial attacks and evaluation protocols [24, 25].

- *Training dynamics.* Mean field theory, neural tangent kernel, and tensor program analyses characterize networks in the infinite-width limit, yielding insight and claims about training dynamics [26, 27, 28, 29, 30].

- *Task and benchmark predictability.* Other work studies how to predict performance on new tasks or benchmarks from few-shot samples or related evaluations [31, 32, 33, 34, 35].

These examples highlight the diversity of theoretical claims that could be made. Although heterogeneous in form, they share a common structure: assumptions, a predicted guarantee, and a measurable outcome. MAGNET's proposed evaluation cards (described in more detail in Section 4) aim to capture this structure, enabling such claims to be tested using standardized evaluation protocols.

3 DARPA AI Quantified (AIQ) Program Overview

The DARPA Artificial Intelligence Quantified (AIQ) program operates on the core hypothesis that mathematical foundations, combined with advances in measurement and modeling, will allow guaranteeing *what* capabilities an AI model has, *when* they will or will not manifest, and *why* [7]. While recent years have seen dramatic advances in large language and multimodal models, their behavior remains difficult to predict in ways that can ensure performance in high-stakes applications. AIQ addresses this gap by

organizing a coordinated research effort to develop mathematical theories of generalization (Technical Area 1, TA1) and build infrastructure to empirically test these theories at scale (Technical Area 2, TA2).

Capability levels. AIQ formalizes its objectives around three nested *capability levels*. At the most basic level, *specific problems*, the question is whether a system gives the correct answer for an individual input–output pair (e.g. multiple-choice question answering). At the second level, *classes and compositions of problems*, the goal is to determine whether performance transfers to similar inputs or structured compositions of tasks—for instance, whether success on one reasoning step implies success on multi-step compositions. At the third level, *natural classes of problems*, the program seeks to identify families of problems implicitly supported by a given model architecture. Examples include convolutional networks naturally aligning with translation-invariant tasks in the image domain or transformers with long-range sequence dependencies. Together, these three levels provide a useful "what, when, and why" framing: *what* problems a model is able to solve, *when* those solutions generalize to related cases, and *why* certain inductive biases emerge from architectural design.

Evaluation domains. To operationalize these capability levels, AIQ also defines a set of *evaluation domains* to help ground the development of mathematical theories around concrete model behaviors. The current domains include: (i) training-time dynamics, which focus on learning curves, optimization trajectories, and scaling laws; (ii) text generation, which encompasses reasoning, factuality, and robustness of language models; and (iii) text-to-image generation, which targets compositional fidelity and robustness in generative vision models. These domains provide a structured setting in which TA1 theories can be instantiated and tested against real-world model behaviors, with new domains possibly added in the future.

TA1: theories and guarantees. TA1 teams bring a wide range of theoretical and mathematical approaches to AIQ, drawing from tools in analysis, geometry, algebra, probability, information theory, and scaling. Recent examples include scaling-limit analyses of residual networks [29], statistical consistency results for kernel embeddings [22], and spectral perspectives on heavy-tailed self-regularization [21]. These seemingly heterogeneous approaches share a common goal: to move beyond empirical heuristics toward more formal guarantees of model behavior. However, this diversity of approaches also highlights the need for empirical evaluations to accommodate each approach and its assumptions.

TA2: evaluation and verification. In parallel, TA2 teams are developing the software and compute infrastructure required to test and validate TA1 theories under realistic conditions and at scale. A core requirement is to reduce the friction from a new mathematical proposal to a falsifiable empirical test, ensuring reproducibility and comparability across different theories. MAGNET, our proposed TA2 effort, is designed to meet this need by providing a theoretical claim-aware evaluation framework that encodes theoretical statements as structured *evaluation cards*, routes them to the appropriate benchmark tasks and datasets, and produces standardized outputs that help verify or falsify a particular claim.

Program goals. The AIQ program ultimately aims to deliver mathematically rigorous methods for quantifying AI model generalization, providing a basis for safe and reliable deployment in high-stakes settings. By integrating TA1's theoretical advances with TA2's evaluation infrastructure, the program seeks to create a virtuous cycle: theories can be rapidly tested and refined, empirical results can inspire new theory, and standardized output formats ensure that progress is cumulative. Next, we discuss how the MAGNET system fits into this broader vision by translating theoretical claims into empirical evidence.

4 MAGNET System Overview

MAGNET (**M**athematical **A**ssurance of **G**enerative AI **N**etwork **E**valuation **T**oolkit) is our proposed system that implements theoretical claim evaluation. Unlike conventional benchmarks which measure model performance, MAGNET is designed to *test theoretical guarantees themselves*. The system accepts structured *evaluation cards*, maps them to available models and datasets, enforces constraints, executes experiments, and produces standardized outputs. Figure 2 summarizes the system architecture.

Claim-aware evaluation cards. Evaluation cards encode a theoretical claim in both natural and formal representations, together with assumptions, datasets, model families, metrics, and constraints. Cards may state, for example, that scaling from a 100M to a 1B parameter model should yield a predictable accuracy

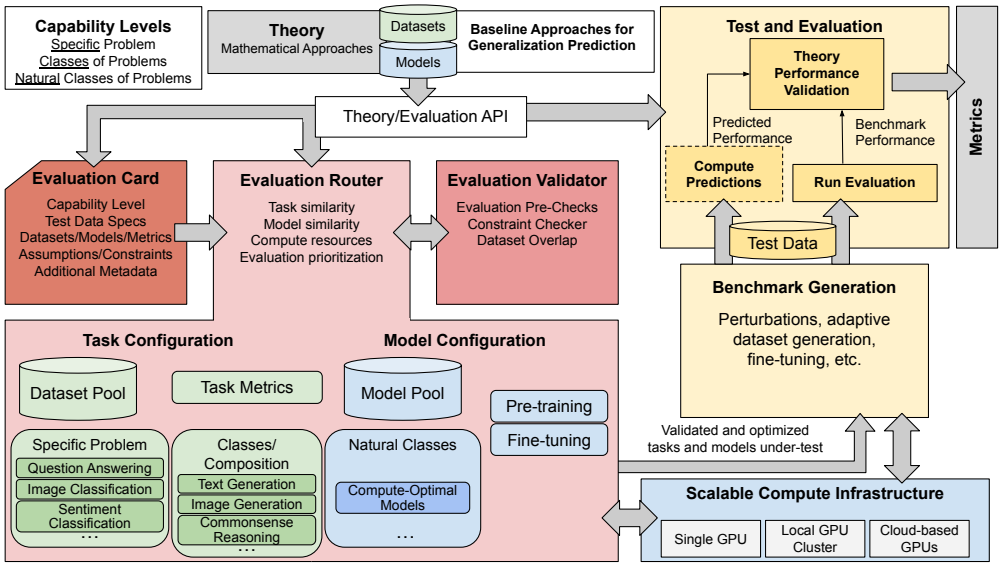

Figure 2: MAGNET system architecture. Evaluation cards encode theoretical claims and capture critical metadata in a human-readable and machine-computable format. The evaluation router and validator transform evaluation cards into executable evaluations, mapping them onto relevant benchmarks. Finally, benchmark generation leverages scalable compute infrastructure to carry out tests, with the outputs being structured reports and AIQ program metrics. Gray boxes indicate theoretical inputs or output metrics.

change, or that performance on one benchmark predicts another. We provide key desiderata for evaluation cards in Appendix B.1 and an example mock evaluation card in Appendix B.3.

Evaluation routing and execution. The evaluation router determines how to instantiate tests for a theoretical claim, identifying which parts can be satisfied with existing benchmarks (e.g. HELM [6], BIG-bench [5]) and which require new experiments. Constraints — such as data contamination checks, model licensing, or compute budgets — are enforced by an evaluation validator. For claims that may require extensive compute, the router can generate executable pipelines up to full pre-training jobs (e.g. using the Marin platform [36]), even if they are not run immediately, providing cost estimates alongside execution plans. We also provide key desiderata for the evaluation router in Appendix B.2.

Standardized reporting. Outputs of an evaluation are returned as structured reports that include key metrics. Each report compares theoretical predictions against empirical results, with standardized outputs and verdicts on theoretical claims (e.g. VERIFIED, FALSIFIED, INCONCLUSIVE). Reports may also include uncertainty and power estimates, enabling comparison across different approaches.

Open implementation. MAGNET will be implemented as an open-source framework, with reproducibility as a first-class principle. All cards, benchmarks, and results are versioned and shareable, allowing the community to propose new claims and extend the system beyond the AIQ program.

5 Conclusion

The AIQ program is an ambitious effort to bring mathematical rigor to the evaluation of modern AI systems. TA1 teams are developing diverse theoretical approaches, while TA2 systems such as MAGNET provide the infrastructure to test and evaluate these theories empirically. What makes AIQ distinctive is that it evaluates *theories*, not just models. The challenge is to map heterogeneous theoretical claims—from scaling laws to spectral weight properties—into concrete, falsifiable tests, while respecting the assumptions and constraints each theory imposes. This requires reusing existing benchmarks when possible, running new experiments where necessary, and producing standardized outputs that allow for comparison across approaches. Importantly, one limitation we note is that our work is still at an early stage: the architecture is defined, evaluation cards are prototyped, and initial routing mechanisms are being designed.

MAGNET is being developed as an open-source framework, with all software, evaluation cards, and generated benchmarks intended for public release. By lowering the barrier between theoretical ideas and empirical validation, we hope to foster a broad research community that can propose, test, and refine theories on model generalization at scale. The AIQ program has just begun, but its vision is clear: to establish a principled, reproducible pathway for connecting theories with the measured behavior of foundation models. We invite workshop participants to contribute new theoretical claims as evaluation cards, helping to shape an open and shared testbed for theory-driven evaluation and benchmarking.

Acknowledgment

This material is based upon work supported by the Defense Advanced Research Project Agency (DARPA) under Contract No. HR001125CE017. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Project Agency (DARPA).

References

[1] S. Ni, G. Chen, S. Li, X. Chen, S. Li, B. Wang, Q. Wang, X. Wang, Y. Zhang, L. Fan *et al.*, "A Survey on Large Language Model Benchmarks," August 2025.

[2] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding," in *International Conference on Learning Representations*, 2019.

[3] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[4] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring Massive Multitask Language Understanding," in *International Conference on Learning Representations*, January 2021.

[5] A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso *et al.*, "Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models," *Transactions on Machine Learning Research*, January 2023.

[6] T. Lee, M. Yasunaga, C. Meng, Y. Mai, J. S. Park, A. Gupta, Y. Zhang, D. Narayanan, H. Teufel, M. Bellagente *et al.*, "Holistic Evaluation of Text-to-Image Models," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 69 981–70 011.

[7] P. Shafto, "AIQ: Artificial Intelligence Quantified | DARPA," https://www.darpa.mil/research/programs/aiq-artificial-intelligence-quantified, 2025.

[8] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling Laws for Neural Language Models," January 2020.

[9] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark *et al.*, "Training compute-optimal large language models," in *Advances in Neural Information Processing Systems*, Red Hook, NY, USA, 2022, pp. 30 016–30 030.

[10] Y. Chen, B. Huang, Y. Gao, Z. Wang, J. Yang, and H. Ji, "Scaling Laws for Predicting Downstream Performance in LLMs," in *International Conference on Learning Representations*, October 2024.

[11] Y. Ruan, C. J. Maddison, and T. Hashimoto, "Observational Scaling Laws and the Predictability of Language Model Performance," in *Advances in Neural Information Processing Systems*, vol. 37, 2024, pp. 15 841–15 892.

[12] F. M. Polo, S. Somerstep, L. Choshen, Y. Sun, and M. Yurochkin, "Sloth: Scaling laws for LLM skills to predict multi-benchmark performance across families," February 2025.

[13] P. L. Bartlett, D. J. Foster, and M. J. Telgarsky, "Spectrally-normalized margin bounds for neural networks," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[14] B. Neyshabur, S. Bhojanapalli, D. Mcallester, and N. Srebro, "Exploring Generalization in Deep Learning," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[15] N. Shirish Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, "On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima," in *International Conference on Learning Representations*, 2017.

[16] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, "Sharpness-Aware Minimization for Efficiently Improving Generalization," April 2021.

[17] G. K. Dziugaite and D. M. Roy, "Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data," in *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2017.

[18] Y. Chu and M. Raginsky, "A unified framework for information-theoretic generalization bounds," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 79 260–79 278.

[19] S. Arora, R. Ge, B. Neyshabur, and Y. Zhang, "Stronger Generalization Bounds for Deep Nets via a Compression Approach," in *Proceedings of the International Conference on Machine Learning*. PMLR, 2018, pp. 254–263.

[20] Y. Jiang, B. Neyshabur, H. Mobahi, D. Krishnan, and S. Bengio, "Fantastic Generalization Measures and Where to Find Them," in *International Conference on Learning Representations*, September 2019.

[21] C. H. Martin and M. W. Mahoney, "Implicit Self-Regularization in Deep Neural Networks: Evidence from Random Matrix Theory and Implications for Learning," *Journal of Machine Learning Research*, pp. 1–73, June 2021.

[22] A. Acharyya, M. W. Trosset, C. E. Priebe, and H. S. Helm, "Consistent estimation of generative model representations in the data kernel perspective space," January 2025.

[23] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified Adversarial Robustness via Randomized Smoothing," in *Proceedings of the International Conference on Machine Learning*. PMLR, 2019, pp. 1310–1320.

[24] F. Croce, M. Andriushchenko, V. Sehwag, E. Debenedetti, N. Flammarion, M. Chiang, P. Mittal, and M. Hein, "RobustBench: A standardized adversarial robustness benchmark," in *Advances in Neural Information Processing Systems*, 2021.

[25] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *Proceedings of the International Conference on Machine Learning*. PMLR, 2020, pp. 2206–2216.

[26] S. S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein, "Deep Information Propagation," in *International Conference on Learning Representations*. arXiv, 2017.

[27] A. Jacot, F. Gabriel, and C. Hongler, "Neural Tangent Kernel: Convergence and Generalization in Neural Networks," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[28] G. Yang, "Wide Feedforward or Recurrent Neural Networks of Any Architecture are Gaussian Processes," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[29] B. Bordelon, L. Noci, M. B. Li, B. Hanin, and C. Pehlevan, "Depthwise Hyperparameter Transfer in Residual Networks: Dynamics and Scaling Limit," in *International Conference on Learning Representations*, January 2024.

[30] N. Dey, B. C. Zhang, L. Noci, M. Li, B. Bordelon, S. Bergsma, C. Pehlevan, B. Hanin, and J. Hestness, "Don't be lazy: CompleteP enables compute-efficient deep transformers," May 2025.

[31] C. Nguyen, T. Hassner, M. Seeger, and C. Archambeau, "LEEP: A New Measure to Evaluate Transferability of Learned Representations," in *Proceedings of the International Conference on Machine Learning*. PMLR, 2020, pp. 7294–7305.

[32] K. You, Y. Liu, J. Wang, and M. Long, "LogME: Practical Assessment of Pre-trained Models for Transfer Learning," in *Proceedings of the International Conference on Machine Learning*. PMLR, 2021, pp. 12 133–12 143.

[33] L. Pacchiardi, L. G. Cheke, and J. Hernández-Orallo, "100 instances is all you need: Predicting the success of a new LLM on unseen data by testing on a few instances," September 2024.

[34] F. M. Polo, L. Weber, L. Choshen, Y. Sun, G. Xu, and M. Yurochkin, "tinyBenchmarks: Evaluating LLMs with fewer examples," May 2024.

[35] J. Shi, W. Ma, S. Ying, L. Jiang, Y. liu, and B. Du, "Importance Sampling is All You Need: Predict LLM's performance on new benchmark by reusing existing benchmark," August 2025.

[36] D. Hall, "Introducing Marin: An Open Lab for Building Foundation Models," http://marin.community/blog/2025/05/19/announcement/, May 2025.

[37] L. de Moura and S. Ullrich, "The Lean 4 Theorem Prover and Programming Language," in *Automated Deduction – CADE 28*. Berlin, Heidelberg: Springer-Verlag, July 2021, pp. 625–635.

[38] H. Finney, L. Donnerhacke, J. Callas, R. L. Thayer, and D. Shaw, "OpenPGP Message Format," Internet Engineering Task Force, Request for Comments RFC 4880, Nov. 2007.

[39] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar *et al.*, "Holistic Evaluation of Language Models," *Transactions on Machine Learning Research*, February 2023.

A Broader Impact

The main positive impact of MAGNET, if successful, is to provide a principled foundation for confidently deploying AI systems in high-stakes scenarios by grounding decisions in verified theoretical guarantees. The main negative risk is that these guarantees could be misinterpreted or used improperly as justification to deploy systems before they are truly ready, leading to overconfidence and potential failures. We therefore emphasize that all evaluations are explicitly conditioned on stated assumptions, and that proper interpretation of guarantees is as important as their formal validation.

B MAGNET System Design

This appendix expands on the technical design of MAGNET. In particular, we sketch the structure of an *evaluation card* and outline how a router might map a card to available benchmarks and new experiments.

B.1 Desiderata for Evaluation Cards

An evaluation card should serve as a standardized container that bridges theoretical claims and empirical tests. We identify several desiderata:

- **Human- and machine-readable.** Claims should be expressible in natural language, but also convertible into formal logic (e.g. a Python function or a Lean 4 proposition [37]) suitable for programmatic validation.
- **Cryptographic attestations.** Conversions between natural and formal statements should be accompanied by verifiable attestations to ensure trust and auditability.
- **Content-addressability.** All datasets, splits, models, and checkpoints should be referenced via hashes or unique IDs to guarantee reproducibility.
- **Free-variable binding.** Formal claims are expressed as functions with free variables; the card must specify how these are bound to experimental artifacts such as dataset splits or model checkpoints.
- **Constraints and assumptions.** Cards should encode auxiliary assumptions (e.g. compute budget, contamination exclusion, licensing, privacy) that must hold for the claim to be valid.
- **Metrics and verdicts.** Cards should specify measurable outcomes, tolerances, and confidence thresholds, together with standardized verdicts (VERIFIED, FALSIFIED, INCONCLUSIVE).

These desiderata aim to ensure that evaluation cards are flexible enough to capture heterogeneous claims, yet strict enough to support reproducible, auditable testing.

B.2 Desiderata for the Router

The router is the mechanism that interprets an evaluation card and generates an executable plan. Its design should satisfy several desiderata:

- **Manual-first, automation-ready.** A fully manual mapping path must always be available, even if tedious, to ensure the framework is usable during early development and for complex edge cases. Automation (e.g. LLM-assisted codification of claims) can be introduced gradually as the framework matures.
- **Deterministic and auditable.** Routing decisions should be deterministic given the same inputs and accompanied by logs or attestations that can be verified post hoc.
- **Benchmark awareness.** The router should identify which parts of a claim can be satisfied by existing benchmark results (e.g. HELM, BIG-bench) and which require new experiments.
- **DAG construction.** When new experiments are required, the router should generate a directed acyclic graph (DAG) of tasks specifying model runs, dataset splits, and evaluation metrics. For large-scale claims, the router should always resolve to a concrete sequence of executable commands—such as training runs on platforms like Marin [36]. Even if resources are unavailable to execute the pipeline, MAGNET can still expose what would need to be done and estimate the associated cost.

- **Constraint checking.** Routing must respect the constraints encoded in the card, such as compute budgets, contamination exclusions, or dataset licensing.
- **Extensibility.** The router should support plug-in modules for new tasks, metrics, or backends without requiring redesign of the core schema.

B.3 Mock Evaluation Card

An *evaluation card* encodes a theoretical claim, the assets required to test it, and the rules for returning a verdict. The card must be both human-readable and machine-checkable, allowing natural language claims to be translated into formal propositions and bound to experimental data. Figure 3 shows a mock card in YAML format.

Our initial card design contains several key components:

- **Claim.** A falsifiable statement, ideally encodable as a Lean 4 proposition [37] (although any programming language would work). Cards may include both a natural-language version and a formal version, with cryptographic attestations certifying the mapping (e.g. with OpenPGP signatures [38]).
- **Datasets.** Content-addressable identifiers (or with SHA-256 hashes) for admissible datasets and splits, ensuring reproducibility.
- **Models.** References to model families, checkpoints, and training recipes.
- **Constraints.** Assumptions such as compute budgets, contamination exclusion, privacy, and licensing requirements.
- **Metrics.** Quantities to be measured or bounded, with thresholds for success/failure.
- **Symbols.** A table binding free variables in the formal claim to datasets, models, or metrics.
- **Outputs.** Standardized verdicts (VERIFIED, FALSIFIED, INCONCLUSIVE) and report formats.

B.4 Routing a Claim

Given a card, MAGNET must route it to available benchmarks and determine what additional experiments are required. For example, observational scaling laws can be partially tested using precomputed results in HELM [39], with missing points filled by running new evaluations. Routing thus involves matching the schema in the card (datasets, models, metrics, constraints) to available assets, prioritizing which dataset+model combinations to actually run, constructing a directed acyclic graph of tasks to execute, and binding experimental outputs back to the free variables in the claim.

Early in the program, this routing will be performed manually. Over time, we plan to semi-automate it: parsing natural-language claims with LLMs, generating formal propositions, and validating each step with cryptographic attestations. This approach ensures that routing decisions are both interpretable and auditable, and that evaluation can proceed even while the system is under construction.

B.5 Summary

The evaluation card provides a standardized container for claims, assets, and verdicts. The router turns this specification into an executable pipeline that either leverages existing benchmark results or launches new experiments. Together, these components embody MAGNET's goal of reducing friction from theory to empirical validation.

Our system is currently under active development and is hosted on GitHub at `https://github.com/AIQ-Kitware/aiq-magnet`. We remind the reader that our starting point is to manually make the connections between a theory and its empirical validation, with the goal of building up automations as the system matures.

```yaml
# Truncated mock evaluation card for scaling-law prediction
version: 0.2
id: scaling-law-toy-001
title: "Power-law extrapolation of accuracy from small to large scale"

claim:
  natural: |
    For family F on dataset D (HELM:MMLU@test), there exist parameters a,b,α>0
    such that A(N)=a-b*N^{-α} interpolates accuracies at N∈{10M,100M} and its
    extrapolation at N=1B differs from the sequestered accuracy by ≤ τ.
  lean4: |
    def within_tol_params
      (A10 A100 A1B_true N10 N100 N1B τ a b α : ℝ) : Prop :=
      A10  = a - b*N10 ^ (-α) ∧
      A100 = a - b*N100 ^ (-α) ∧
      |(a - b*N1B ^ (-α)) - A1B_true| ≤ τ
  attestations:
    - source: "llm-parser-v2"
      validator: "human-itl-review-chain:2e81...5e764010"
      status: "pending"

datasets:
  - id: helm:MMLU@test
    sha256: "6a09e6...c67178f2"
models:
  family: "Pythia"
  checkpoints:
    - name: "pythia-10M"
      sha256: "e3b0c4...abbcb0ba"
constraints:
  compute_budget: { measurable: "gpu_hours", threshold: 200 }
metrics:
  - name: "accuracy"
    tolerance: 0.02
    confidence: 0.95
outputs:
  verdict: [Verified, Falsified, Inconclusive]
```

Figure 3: Mock evaluation card in YAML format. Cards include natural and formal claims, datasets, models, constraints, metrics, and outputs. The exact form of the formal claim is simplified and symbol mappings are excluded for clarity.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes] .

   Justification: This is a position paper, and we make it clear that much of what we describe is speculative. To account for this we describe success criteria and what we hope to accomplish by publishing this paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The main limitation is that this is a speculative system, which we attempt to make abundantly clear.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: No theoretical results are included.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: The paper does not include experiments, but the described system is being developed in a way to ensure reproducibility.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: There are no experiments to reproduce, but the we link to an anonymized repository where we are developing the system.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: This paper does not include experiments, but we do mention that datasets used in our evaluation cards and router will ideally be content addressable or at the very least hash verifiable to ensure consistent reproduction.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: There are no experiments, but our system's validation criteria will evaluate claims at a confidence level, which will require that our system be capable of reporting error bars and estimating statistical significance.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: This paper is a speculative position paper, and does not include experiments. We do make mention that the experiments our router selects will be "compute-aware" and be subject to a compute budget. As such, recording of the context in which experiments take place will be a priority of the system.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We do not see any ethical concerns in this work. Our evaluation cards will ensure and check that privacy and license restrictions are met.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We include a broader impact statement in the appendix.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper does not pose such risks. We seek to empirically validate claims made about existing data / models.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The current paper does not use any existing assets, but as we develop the system we will make use of pre-existing HELM results and we cite the paper appropriately and only used official distribution mechanisms to access third party work.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We link to the github where we are developing the system. It is initial, but we are developing it with professional software engineering practices, which includes documentation and unit testing for the limited features that currently exist.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: There is no research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA] .

    Justification: There is no research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs were not used to develop any ideas. We made use of GPT5 to organize existing ideas, the authors did all final proof-reading and polishing. We do envision LLMs as a component of mapping natural language claims into formal programmatic expressions and are using the idea of cryptographic attestations and chains of trust to validate those mappings. We also note that while this novel use of LLMs could be a component of our framework, it is an optional component and our envisioned system does not depend on them.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.