Research Statement

Bishwamittra Ghosh
Postdoctoral Researcher
Max Planck Institute for Software Systems, Germany
https://bishwamittra.github.io

My research lies at the intersection of machine learning and formal methods, where I develop precise, mathematically grounded tools, and controlled environments to analyze and improve the behavior of modern learning systems. As machine learning models grow in complexity and are increasingly deployed in high-stakes settings, there is a pressing need for principled approaches to reason about their learning dynamics, inference behavior, and trustworthiness. To this end, I draw on formal methods – such as logic, formal languages, and satisfiability solving – to examine how key properties of learning systems persist despite their stochasticity and scale. My goal is to build a verifiable and robust foundation of machine learning systems that are not only capable, but also trustworthy.

Summary of Research. My research is organized around two central themes: advancing a **foundational understanding of language models** and developing **trustworthy machine learning** for algorithmic decision-making. Large language models (LLMs) are complex generative machine learning (ML) systems with the potential for widespread application across domains. My work seeks to uncover the foundations of LLMs by studying their learning [1], memorization [2, 3], and inference behavior [4, 5]. In parallel, I explore issues of fairness [6–9] and interpretability [10–12] in ML systems, aiming to enhance their trustworthiness in high-stakes decision-making contexts. See Figure 1 for a high-level overview of how my research themes interconnect.

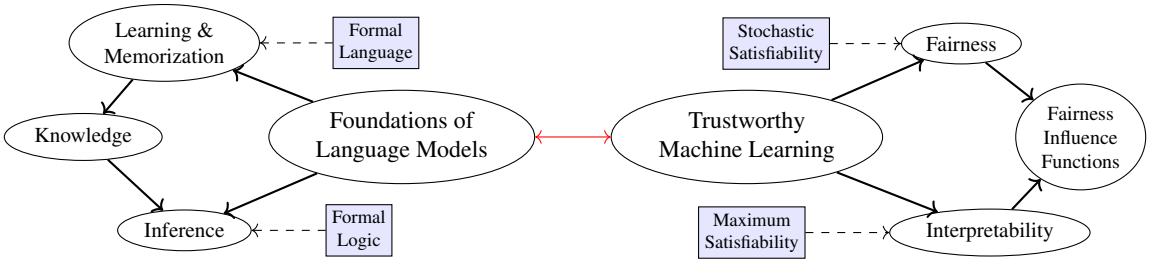

Figure 1: To better visualize the scope and interconnection between my research themes (ellipses) and methodological tools (blue boxes), Figure 1 presents a *keyword graph* that maps the relationships among core areas of inquiry and the formal methods I employ. Red arrows highlight my future research plans, focusing on the intersection of these themes and expanding beyond.

Beyond my core research themes, I have explored several other areas in computer science. This includes decentralized machine learning, such as federated learning [13] and split learning [14]; graph-theoretic problems, including social-spatial group queries [15, 16] and core decomposition in hypergraphs [17]; and group testing [18].

Summary of Achievements. I have engaged in collaborative and interdisciplinary research at multiple academic institutions and industry, across three continents North America, Europe, and Asia. I conducted a research visit to INRIA, France, Goldman Sachs, India, and Max Planck Institute For Software Systems, Germany. My research is published at premier conferences on machine learning (ICLR, AAAI x 4, ECAI x 2, FAccT, AIES, WSDM, CAI) and databases (VLDB x 2), as well as journals in machine learning (JAIR) and databases (TSAS). I have presented a tutorial on *Auditing Bias of Machine Learning Algorithms: Tools and Overview* in IJCAI 2023. I was awarded NUS Research Scholarship, Singapore and Moblilex Scholarship at Université de Lille, France.

1 Research Theme 1: Foundational Understanding of Language Models

The first core theme of my research investigates foundational questions surrounding how large language models learn, memorize, and infer or generate. This theme is driven by the need for rigorous benchmarks and conceptual clarity around learning and inference in these complex systems.

1.1 Language Learning: Fine-tuning and In-context Learning

LLMs learn in two principled modes: fine-tuning and in-context learning. Fundamental questions are to investigate which mode is more language proficient, and how similar their inductive bias is. To answer them, we create a controlled environment of formal language learning, where we evaluate their syntactic pattern recognition abilities [1]. Our setup allows us to effectively control over data distribution, focus on syntax only, and avoid data contamination – satisfying all is hard in existing NLP datasets.

A Discriminative Test for Language Proficiency. *Given a language and an LLM, how do we determine language proficiency*? Should we only consider strings in the language or strings outside the language? We differentiate between a generative and a discriminative test for language proficiency, where the former claims language proficiency if generation probability is high, without comparing with any reference point. The discriminative test, however, claims proficiency if the probability of generation of strings in the language is higher than strings outside the language, making it comparable across LLMs.

Fine-tuning vs. In-context Learning. We show that fine-tuning is more language proficient than in-context learning on in-distribution languages, while both perform similarly on out-of-distribution languages. Their inductive bias is similar, but not equal, and similarity decreases with higher training data or better language learning. In addition, fine-tuning is more robust to changing languages than in-context learning. *Thus, we establish formal language as a potential benchmark for studying LLMs.*

1.2 Learning without Memorizing Considered Infeasible: Rethinking Memorization in LLMs

Formal language learning provides a controlled setup to study nuances of memorization in LLMs. Memorizing when learning is considered undesirable for two distinct reasons: first, from a privacy perspective, memorization raises concerns about potential leakage of sensitive information in training data. Second, from a learning perspective, memorization raises concerns of sub-optimal learning and over-fitting. We rethink measures of memorization in LLMs: existing measures of memorization, namely recollection-based and counterfactual measures, are designed to capture privacy concerns, but they ignore optimal learning concerns. We propose a new memorization measure, called *contextual memorization* that captures LLMs tendency to locally over-fit some strings in the training data before others, over multiple epochs of training.

Applying these measures when training LLMs leads us to two striking conclusions. First, a systematic analysis of all the measures shows that our new measure avoids a major pitfall of prior measures, by distinguishing context-based recollection from memorization-based recollection of a training string. Using our measure, we revisit prior reported instances of training data memorization by real-world LLMs and find that many instances can be explained away by contextual learning-based recollection, i.e., the prior memorization reports are likely exaggerated. Second, when LLMs learn a language optimally, they inevitably end up memorizing some portions of the training data. *Our study not only prioritizes conceptual clarity on memorization measures, but also call for a careful investigation of the effectiveness of memorization mitigation strategies, such as data deduplication.*

1.3 Reliable Latent Knowledge Estimation in LLMs

Not all memorization is undesirable, including the case of memorizing factual information by LLMs. *How can we estimate the latent knowledge of an LLM*, provided that different LLMs may not understand the same prompt instructions effectively? We propose a *zero-prompt, many-shot* approach for factual knowledge estimation [5]. Intuitively, if the LLM knows that "Germany's capital is Berlin," we present it with a pattern such as "France Paris; Italy Rome; Germany" and check whether it generates "Berlin" – the key idea is to let the LLM infer the underlying relation, without any meta-linguistic judgement. This method avoids prompt hacking and side-channeling, relies on in-context learning, and remains agnostic to both model architecture and relation type. *Thus, the zero-prompting approach has a potential to avoid the ambiguity in LLMs' understanding of input prompts.*

1.4 Logical Consistency of LLMs

Prompting LLMs is inevitable in many situations. What is a simple sanity test to check if the LLM indeed understands the prompt? We propose a *logical consistency* test [4], to evaluate whether the LLM's response remains consistent under *logical transformations* of the input prompt. Logical consistency is orthogonal to accuracy, where we emphasize that an LLM may be incorrect, but its responses should still be consistent under logical changes of the prompt.

The Need for Logical Consistency. Prior consistency criteria have primarily relied on paraphrasing-based input prompts. In contrast, logical consistency extends to any logical operation such as negation, conjunction, and disjunction, as well as rules like associativity, commutativity, and De Morgan's laws – spanning propositional, first-order, and higher-order logic. Intuitively, negation consistency requires that a negated prompt yields a negated response, while conjunctive consistency implies that the response to a conjunctive prompt matches the conjunction of responses to its sub-prompts.

Logical Consistency of LLMs in Fact-checking. As a first step, we assess LLMs' logical consistency in fact-checking tasks using knowledge graphs, where the model verifies input facts as true or false based on relevant extracted information. Our experiments show that logical consistency declines with fact complexity, while larger models tend to be more consistent. To enhance consistency, we propose an accuracy-driven approach using instruction fine-tuning and prompting. *Our long term goal is to achieve logical consistency in LLMs in a task-agnostic manner and across all interactions with LLMs.*

2 Research Theme 2: Trustworthy Machine Learning

Complementing this foundational perspective, my second research theme focuses on developing trustworthy machine learning systems, particularly in contexts where fairness, interpretability, and societal impact are paramount. Here, I take a formal and computational approach to tackle pressing challenges in real-world ML deployments.

2.1 Formal Fairness Quantification and Explanation

Fairness in ML systems is a complex challenge, grounded in societal norms and beliefs. As a result, diverse fairness metrics have been developed to capture unfairness from various perspectives, along with corresponding mitigation algorithms. Central to these efforts is the need for *formal quantification of fairness* and the ability to *explain its underlying sources*.

2.1.1 Probabilistic Fairness Quantification

Towards quantifying *group-based fairness* of a binary ML classifier, the key objective is to compute the maximum (and minimum) probability of positive prediction across all sensitive groups, given the joint distribution of sensitive and non-sensitive features. The complexity of this problem grows exponentially with the number of sensitive attributes – such as race and gender – due to the combinatorial explosion of group partitions. Our focus is therefore on improving both the *scalability* and *accuracy* of fairness quantification.

We propose Justicia [6] by resorting to *stochastic Boolean satisfiability* (SSAT), a variant of Boolean satisfiability that models probabilistic decision-making with optimization. SSAT introduces logical choice variables (analogous to sensitive features) and random chance variables (analogous to non-sensitive features) to reason about the probability of a formula being satisfied (analogous to positive prediction). This modeling applies to tree-based and linear classifiers with white-box access to model parameters, and enables exact yet scalable fairness quantification compared to existing SMT and sampling methods – with scalability attributed to decades of advances in Boolean satisfiability solving.

2.1.2 Tractable Fairness Quantification

While SSAT has exponential worst-case complexity, can we achieve tractable fairness quantification for specific model classes such as linear classifiers? We show that probabilistic fairness quantification for linear classifiers admits *pseudo-polynomial* complexity by framing the problem as a novel *stochastic subset sum problem* (S3P) [7], inspired by SSAT. Tractability arises in two ways: identifying sensitive groups that always (and never) receive positive predictions is straightforward from the classifier's feature coefficients, and S3P over non-sensitive features admits an efficient dynamic programming solution.

Encoding feature correlation. Both SSAT and S3P assume independence among random variables, limiting their ability to capture feature correlations. To address this, we model the input probability distribution using a Bayesian network and adapt SSAT and S3P accordingly. This enables efficient encoding of correlations and results in more accurate fairness quantification.

2.1.3 Universal Quantification for Group Fairness, Individual Fairness, and Robustness

Is it possible to quantify various distributional properties of a black-box model – without reconstruction – using a universal representation? In *active Fourier auditing* (AFA), we propose a Fourier-based quantifier for group fairness, individual fairness, and robustness [8]. The core idea is to approximate the model with a surrogate expressed in an orthonormal Fourier basis over input features, allowing these properties to be naturally formulated in terms of Fourier coefficients. We show that a PAC (probably approximately correct) estimate of each property can be obtained using only the significant coefficients, while also bounding the worst-case sample complexity.

2.1.4 Explaining Fairness: Identifying Sources of Unfairness

What if we could not only quantify unfairness but also explain its sources – analogous to feature attribution methods? Such explanations could help prevent unfairness [9]. To this end, we introduce *fairness influence functions* (FIFs) to measure the contribution of individual features and their intersections to overall unfairness. Inspired by global sensitivity analysis – which decomposes a function's variance into contributions from input variables – we formalize FIFs for linear group fairness metrics. Like AFA, FIFs are model-agnostic and apply to any model expressible via orthonormal Fourier expansion. Their effectiveness is demonstrated by strong alignment with fairness interventions and fine-grained explanations of intersectional unfairness, absent in local explainability methods such as SHAP.

2.2 Interpretable Rule-based Classification

Complex, opaque models are not always necessary – indeed, in high-stakes settings like medical decision-making, interpretable and small models are often preferred. Rule-based classifiers, including decision trees, decision lists, and decision sets, form a prominent class of interpretable models. However, two key challenges persist: (1) the scalability of learning and (2) the limited expressiveness of the resulting rules.

2.2.1 Scalability via Incremental Learning

We adopt *conjunctive normal form* (CNF) as the hypothesis class for rule-based classifiers, as it underlies the structure of most rule-based models. The learning objective is to find smaller rules – enhancing interpretability – while preserving accuracy. We formulate optimal rule learning as a *maximum satisfiability* (MaxSAT) problem, where soft clauses encode learning objectives and hard clauses enforce constraints. However, the complexity of solving MaxSAT increases with the dataset's dimensionality and the size of the classifier.

We propose an incremental MaxSAT-based learning framework, IMLI [10, 11], designed for scalable rule-based classification. IMLI combines mini-batch learning with iterative rule learning: it constructs a CNF classifier by progressively covering the training data, solving a sequence of smaller MaxSAT problems on mini-batches in each iteration by incrementally updating the learned rules. Empirically, IMLI achieves competitive accuracy and interpretability while scaling to datasets of million samples.

2.2.2 Expressiveness via Logical Relaxation

While rule-based classifiers offer interpretability, they often lack expressiveness. To address this, we introduce *relaxed-CNF* rules, which include tunable expressiveness parameters controlling how many literals per clause and how many clauses per formula must be satisfied to predict a positive class [12]. This combinatorial structure yields exponential succinctness over standard CNF and is commonly used in medical checklists (e.g., CHADS_2 score). To learn relaxed-CNF rules efficiently, we extend the incremental framework of IMLI by replacing the MaxSAT encoding with a *mixed integer linear program* (MILP). Empirically, relaxed-CNF rules maintain similar accuracy while being smaller in size than other rule-based classifiers.

3 Future Directions and Vision

My research at the intersection of machine learning and formal methods has allowed me to address core computational problems from multiple perspectives, which I see as essential for shaping my future work. As machine learning continues to evolve, I aim to leverage concepts from automated reasoning and formal methods to both analyze and improve learning systems. My long-term vision is to develop a foundational understanding of learning systems that enhances their capabilities and trustworthiness, while enabling formal certification of their behavior.

A first direction is **controlled benchmarking of machine learning**. Benchmarking plays a central role in ML development, and formal methods offer significant promise for its advancement and foundational understanding. My plan is to design synthetic, controlled datasets and environments; formally model and evaluate ML behavior; and quantify how well ML systems align with human values. In such modeling, formal methods allow us to separate the specification of safety-critical properties from their verification, by leveraging scalable solvers from the formal methods' community. Conversely, the complexity and ambiguity of modern ML systems present compelling challenges that can broaden the scope of formal methods. I aim to advance our understanding of learning systems by systematically integrating these two perspectives.

A second direction is **generative AI for improved problem-solving**. The widespread adoption of generative AI has transformed human–machine interaction in various problem-solving tasks. LLMs and related systems can act as effective collaborators: assisting in hypothesis generation, supporting exploratory thinking, and enhancing creative workflows. My research will apply generative AI to improve hypothesis construction, idea generation, and decision-making, in line with emerging paradigms in agentic AI and human–computer interaction. In parallel, I will explore what forms of formal guarantees and reliability can be achieved in AI–human collaborative systems.

A third direction is the investigation of **AI safety** in the design of modern AI systems. As AI is increasingly deployed in real-world applications, a rigorous scientific understanding of their capabilities and limitations is essential. Is the current data-driven and scale-focused paradigm of AI development sufficient to achieve artificial general intelligence (AGI)? Conversely, should we desire foundation models capable of self-reprogramming without human intervention? My research in the next decade aims to broaden our understanding of AI systems in general, and develop methods to ensure their safety and trustworthiness.

References

[1] Bishwamittra Ghosh, Soumi Das, Till Speicher, Qinyuan Wu, Mohammad Aflah Khan, Deepak Garg, Krishna P Gummadi, and Evimaria Terzi. Fine-tuning vs. in-context learning in large language models: A formal language learning perspective. In *Submission*, 2025.

[2] Bishwamittra Ghosh, Soumi Das, Qinyuan Wu, Mohammad Aflah Khan, Krishna P Gummadi, Evimaria Terzi, and Deepak Garg. Learning without memorizing considered infeasible: Rethinking memorization in LLMs. In *Submission*, 2025.

[3] Soumi Das, Camila Kolling, Mohammad Aflah Khan, Mahsa Amani, Bishwamittra Ghosh, Qinyuan Wu, Till Speicher, and Krishna P Gummadi. Revisiting privacy, utility, and efficiency trade-offs when fine-tuning large language models. In *Submission*, 2025.

[4] Bishwamittra Ghosh, Sarah Hasan, Naheed Anjum Arafat, and Arijit Khan. Logical consistency of large language models in fact-checking. In *Proc. of ICLR*, 2025. URL https://arxiv.org/pdf/2412.16100.

[5] Qinyuan Wu, Mohammad Aflah Khan, Soumi Das, Vedant Nanda, Bishwamittra Ghosh, Camila Kolling, Till Speicher, Laurent Bindschaedler, Krishna P Gummadi, and Evimaria Terzi. Towards reliable latent knowledge estimation in LLMs: Zero-prompt many-shot based factual knowledge extraction. In *Proc. of WSDM*, 2025. URL https://arxiv.org/pdf/2404.12957.

[6] Bishwamittra Ghosh, Debabrota Basu, and Kuldeep S. Meel. Justicia: A stochastic SAT approach to formally verify fairness. In *Proc. of AAAI*, 2021. URL https://arxiv.org/pdf/2009.06516.pdf.

[7] Bishwamittra Ghosh, Debabrota Basu, and Kuldeep S. Meel. Algorithmic fairness verification with graphical models. In *Proc. of AAAI*, 2022. URL https://arxiv.org/pdf/2109.09447.pdf.

[8] Ayoub Ajarra, Bishwamittra Ghosh, and Debabrota Basu. Active Fourier auditor for estimating distributional properties of ml models. In *Proc. of AAAI*, 2025. URL https://arxiv.org/pdf/2410.08111.

[9] Bishwamittra Ghosh, Debabrota Basu, and Kuldeep S. Meel. How biased are your features?: Computing fairness influence functions with global sensitivity analysis. In *Proc. of FAccT*, 2023. URL https://arxiv.org/pdf/2206.00667.pdf.

[10] Bishwamittra Ghosh, Dmitry Malioutov, and Kuldeep S. Meel. Efficient learning of interpretable classification rules. In *Proc. of JAIR*, 2022. URL https://arxiv.org/pdf/2205.06936.pdf.

[11] Bishwamittra Ghosh and Kuldeep S. Meel. IMLI: An incremental framework for MaxSAT-based learning of interpretable classification rules. In *Proc. of AIES*, 2019. URL https://bishwamittra.github.io/publication/imli-ghosh.pdf.

[12] Bishwamittra Ghosh, Dmitry Malioutov, and Kuldeep S. Meel. Classification rules in relaxed logical form. In *Proc. of ECAI*, 2020. URL https://bishwamittra.github.io/publication/ecai_2020/paper.pdf.

[13] Bishwamittra Ghosh, Debabrota Basu, Fu Huazhu, Wang Yuan, Renuga Kanagavelu, Jiang Jin Peng, Liu Yong, Goh Siow Mong Rick, and Wei Qingsong. History-aware and dynamic client contribution in federated learning. In *Proc. of ECAI*, 2025. URL https://arxiv.org/pdf/2403.07151.

[14] Bishwamittra Ghosh, Yuan Wang, Huazhu Fu, Qingsong Wei, Yong Liu, and Rick Siow Mong Goh. Split learning of multi-modal medical image classification. In *Proc. of CAI*, pages 1326–1331. IEEE, 2024. URL https://bishwamittra.github.io/publication/splitFusionNet_2024/main.pdf.

[15] Bishwamittra Ghosh, Mohammed Eunus Ali, Farhana Murtaza Choudhury, Sajid Hasan, Timos Sellis, and Jianxin Li. The flexible socio spatial group queries. In *Proc. of VLDB*, 2018. URL http://www.vldb.org/pvldb/vol12/p99-ghosh.pdf.

[16] Sajid Hasan Apon, Mohammed Eunus Ali, Bishwamittra Ghosh, and Timos Sellis. Social-spatial group queries with keywords. *ACM Transactions on Spatial Algorithms and Systems (TSAS)*, 2021. URL https://dl.acm.org/doi/full/10.1145/3475962?accessTab=true.

[17] Naheed Anjum Arafat, Arijit Khan, Arpit Kumar Rai, and Bishwamittra Ghosh. Neighborhood-based hypergraph core decomposition. *Proc. of VLDB*, 2023. URL https://arxiv.org/pdf/2301.06426.pdf.

[18] Lorenzo Ciampiconi, Bishwamittra Ghosh, Jonathan Scarlett, and Kuldeep S. Meel. A MaxSAT-based framework for group testing. In *Proc. of AAAI*, 2020. URL https://bishwamittra.github.io/publication/aaai_2020/AAAI-CiampiconiL.690.pdf.