

Review

The Symbiosis of Formal Methods and Artificial Intelligence

Jonathan P. Bowen ^{1,2,*}  and Sini Chen ^{3,*} ¹ School of Computer Science and Digital Technologies, London South Bank University, London SE1 0AA, UK² Centre for Research and Innovation in Software Engineering (RISE), Southwest University, Chongqing 400715, China³ Software Engineering Institute, East China Normal University, Shanghai 200062, China

* Correspondence: jonathan.bowen@lsbu.ac.uk (J.P.B.); 52265902002@stu.ecnu.edu.cn (S.C.)

Abstract

Formal methods are software engineering approaches with a rigorous mathematical basis that can help ensure the correctness of software systems, especially where safety or security is critical. Artificial Intelligence (AI) has developed very rapidly in the area of Generative AI (GenAI), where questions can be answered with increasingly impressive but potentially unreliable and variable results. This paper surveys research in integrating the two approaches in a synergistic manner. Traditionally, such explorations have required significant manual efforts in searching for and evaluating existing research. However, most relevant publications are now accessible online, and AI tools are increasingly good at answering research questions with more and more reliability. This paper takes the approach of using GenAI to evaluate research questions on combining formal methods and AI-related techniques. The paper assesses the usefulness and validity of these results. With recent improvements in AI search tools, the approach is now a useful aid to researchers, significantly reducing the time needed to survey existing research, while always needing human checking by an expert. In addition, the combination of formal methods and AI approaches looks to be an interesting and beneficial research area with potential industrial-scale applications in the future. This paper does not claim to be an exhaustive systematic review, but rather demonstrates an AI-assisted search process to aid in literature surveys within a fast-moving research area.

Keywords: artificial intelligence; formal methods; Generative AI; Industry 4.0; research aids; software engineering

1. Introduction

In computer science in general and software engineering in particular, formal methods are mathematically rigorous techniques that can be used for the analysis, specification, development, refinement, and verification of computer-based systems, largely for software, but also for hardware [1,2]. As in other engineering disciplines, their use is motivated by the fact that performing appropriate mathematical analysis should improve the reliability and robustness of a design. Thus, their use is especially beneficial in systems where safety or security are primary concerns. Formal methods make use of a variety of theoretical computer science foundational mathematics, including automata theory, formal languages, logic calculi, program semantics, and type theory.

Alan Turing arguably wrote the first paper on formal methods [2], using mathematically based approaches to software engineering, at a 1949 conference in Cambridge [3,4]. A year later, in 1950, he wrote a seminal article on machine intelligence [5], predicting



Academic Editor: João Durães

Received: 3 June 2026

Revised: 1 July 2026

Accepted: 8 July 2026

Published: 12 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

the advent of what became known as Artificial Intelligence (AI). Since then, there have been claims and counterclaims about formal methods. Tool development has been slow, but aided by Moore’s Law with the increasing power of computers. Although formal methods are not widespread in practical usage at a heavyweight level, their influence has crept into software engineering practice to the extent that they are no longer necessarily called formal methods in their use. In addition, in areas where safety and security are important, with the increasing use of computers in such applications, formal methods are a viable way to improve the reliability of such software-based systems. Their use in hardware, where a mistake can be very costly, is also important.

This paper presents the use of formal methods in the light of developing directions regarding their use with the aid of Generative AI (GenAI). It uses an interactive GenAI approach, specifically using Google Scholar Labs [6], to aid in producing a literature survey of the state of the art in combining formal methods and AI approaches. The former has the advantage of mathematical exactitude, whereas the latter, while inexact, is proving increasingly good at apparently serendipitous discoveries. Both approaches have potential industrial-scale applications in the future. This paper does not claim to provide an exhaustive systematic review, but rather, its aim is to demonstrate an AI-assisted search process in a fast-moving research area.

1.1. Formal Methods Background

While the logical foundations of formal methods can be traced back to Aristotle’s logic and the desire to formalize reasoning, the application to computing has distinct 20th-century roots [7]. In mathematics, it is typical to have simply stated theorems, but with potentially very deep and complex proofs. Consider Fermat’s Last Theorem (c. 1637)— $a^n + b^n \neq c^n$ for integer $n > 2$ —postulated by the French mathematician Pierre de Fermat (1607–1665). This was finally proved by Andrew Wiles around 358 years later, in 1994–1995, not a timescale acceptable in software development! Instead, in software verification, there are complicated specifications, potentially thousands of pages, and long programs with millions of lines of code, but with relatively shallow proofs compared to those developed by professional mathematicians.

Arguably, the first paper on formal methods was written by Alan Turing in 1949, titled “Checking a Large Routine” [3]. In this seminal work, presented at the EDSAC Inaugural Conference, Turing introduced the concept of making assertions about the various states that a machine can reach. He posited that the checker of a routine has to verify the initial and final conditions, assertions within the program flow, and that the process terminates.

Turing utilized a flow diagram with properties attached to edges, anticipating the later work of Floyd and Hoare by decades. As noted by Morris and Jones in their analysis of Turing’s work [4], had Turing lived longer, the field of program proving might have developed much more rapidly, rather than requiring the ideas to be rediscovered in the late 1960s. Reading Turing’s paper now, it uses many of the terms that are common in formal methods today, such as “verification”, “assertions”, and even the “dashed” after states used in the formal Z notation [8].

However, what of Turing’s influence on program proving? Adriaan “Aad” van Wijngaarden (1916–1987), a Dutch mathematician and computer scientist, and an early advocate of a mathematical approach to computing, was at the Cambridge meeting, where Turing presented his ideas, but there is no obvious influence on van Wijngaarden. Robert Floyd rediscovered ideas similar to those of Turing, published in 1967 [9]. Tony Hoare developed these further, published in his classic 1969 paper presenting what has become known as “Hoare Logic” [10]. This line of work was later developed further in *Unifying Theories of Programming* [11]. Had Turing lived longer, perhaps formal methods (in particular,

program proving) would have developed more rapidly and with greater continuity, rather than being rediscovered.

One of the earliest books with “formal methods” in its title is a 1962 volume by the Dutch philosopher and logician, Evert Willem Beth (1908–1964) [12]. The full title is *Formal Methods: An Introduction to Symbolic Logic and to the Study of Effective Operations in Arithmetic and Logic*. Although this book is not about software, it does use the term “formal methods” with a meaning that is recognizable by the formal methods community of today. The term “formal methods” was established in its computer science context by the late 1970s, as the next more rigorous stage from structured design, establishing a mathematical basis. Despite advances, a gap remains between theory and practice. In the 1970s, Christopher Strachey, a pioneer in denotational semantics, a particular approach to formalizing software, observed that practical work is often “unsound and clumsy” due to a lack of understanding of fundamental design principles, while theoretical work is often “sterile” because it lacks contact with real computing [13].

By the 1990s, formal methods had established a presence but were surrounded by misconceptions regarding their difficulty and utility. In 1990, Anthony Hall proposed seven “myths” of formal methods, commonly held beliefs about the approach that are inaccurate in practice [14]. Further myths were proposed in 1995 [15], including the ideas that formal methods delay the development process, lack tool support, replace traditional engineering design, only apply to software, or are simply not required. A persistent fallacy was the “all or nothing” approach—the belief that one must be dogmatic. In reality, a related paper on the “Ten Commandments of Formal Methods” [16] suggested a pragmatic approach: for example, “Thou shalt formalize but not over-formalize” and “Thou shalt guestimate costs.”

In the 1995 book *Applications of Formal Methods*, a number of case studies using formal methods were presented using different formal methods [17]. A follow-on book in 1999 [18] provided further examples. A 2001 book [19] presented the process of developing a formal specification for a unified case study by explicitly showing the questions raised when developing such a specification using a wide selection of different formal methods. This contrasts with many presentations where only the final specification of a system is provided, with little information on how this was derived. Further formal methods case study books have been published [20,21].

Formal methods involve formal specification and (optionally) proof, covering the following:

Validation: Helping to establish a correct specification;

Verification: Establishing a correct implementation with respect to a formal specification.

It should be noted that the use of proof in a software engineering context is an outlier compared to most engineering disciplines. Engineers typically “calculate”, for example, using mathematical formulae developed and “proved” previously (e.g., using Maxwell’s equations).

There are a variety of ways to apply formal methods to computer-based systems, including the following:

Abstract Interpretation: Approximating program behavior to prove correctness or detect errors.

Model-Based Testing: Generating test cases from a formal model.

Model Checking: Exhaustively verifying system behavior against a formal specification.

Proof Assistants: Tools for interactively constructing and verifying mathematical proofs.

Refinement: Systematically refining a high-level specification into a correct implementation.

Static Analysis: Analyzing program code meaning to detect errors or enforce constraints.

Verification: Proving the correctness of a program using logical inference rules.

There are also different levels at which formal methods can be used:

1. **Formal Specification:** This is applicable to the system requirements. Formal analysis and proofs are optional. The specification can be used to aid testing. It is the most cost-effective approach using formal methods.
2. **Formal Verification:** The implemented program is produced in a more mathematically formal way. Typically, the approach uses proof or refinement based on a formal specification. This is significantly more costly than just using formal specification.
3. **Theorem Proving:** This involves the use of a theorem prover tool. The approach allows formal machine-checked proofs. A proof of the entire system is possible, but scaling is difficult. The approach can be very expensive and hard.

Formal methods have found their specific niche in the development of systems where safety or security is critical. This has sometimes been driven by the recommendation or even mandating of the use of formal methods in systems at the most critical levels in standards [22], often where human lives are at risk in the case of failure.

It is notable that there have been parallel strands of developments in formal methods and AI, with little overlap until very recently. During the 1970s and 1980s, the main concentration of developments in AI was on logic programming (e.g., notably, the language Prolog) and also expert systems. Prolog and other logic languages provide a relational style of programming, which maps well to logic as used in formal methods and this has been explored by some researchers [23]. Now, with improvements in GenAI, there is much more research exploration in the possibilities of AI and formal methods benefiting from each other.

1.2. Formal Methods and AI

With the recent developments in Generative AI (GenAI), there has been an increasing interest in the possibility of combining the precision of formal methods with the, at least partially, automated possibility of using GenAI to aid in the generation of proofs during verification using a formal approach. This is a rapidly developing area. This paper aims to survey the current state of the art in this area, using a novel approach of utilizing GenAI itself to aid in exploring the current research. In Section 2, we introduce the methodological approach and in Section 3, we present the results of posing three research questions, with comments on each of the publications returned by Google Scholar Labs [6]. A statistical summary and analysis is provided in Section 4. We discuss these results, together with practical issues, and also general developments in combining formal methods and AI in Section 5, followed by a brief conclusion in Section 6. Section 7 lists the publications generated by Google Scholar Labs during the experimental survey.

2. Methodology

The methodological design of this paper, which is not a systematic review itself, has been inspired by earlier systematic literature reviews and systematic mapping studies in software engineering, which emphasize transparent search strategies, reproducible record management, explicit units of analysis, and documented classification schemes [24–27]. These principles are adapted here in a lightweight way, not to claim an exhaustive systematic review, but to make the AI-assisted search process more transparent and to support an exploratory mapping of a fast-moving research area.

2.1. AI-Assisted Search Strategy

Traditionally, a survey of academic papers in a particular area has required significant manual effort. However, the rapid development of Generative AI (GenAI) has enabled a more interactive approach to answering research questions with respect to surveying academic papers. For some time, since its launch in 2004, Google Scholar (<https://scholar>.

[google.com](https://www.google.com) (accessed on 2 June 2026)) has provided good access and search facilities for academic publications, especially for computer science, where most research papers are accessible online. Google has now added Google Scholar Labs to its Google Scholar search facility [6], launched by Google on 18 November 2025, as an experimental AI-powered additional feature for the Google Scholar platform.

Three progressively refined research questions were submitted to Google Scholar Labs. The first question was deliberately broad, asking whether Artificial Intelligence is useful for formal methods. The second and third questions narrowed the focus toward the use of Generative AI to support formal methods and toward recent publications. For each question, the first ten returned publications were recorded, together with the AI-generated summary provided by Google Scholar Labs.

To provide a lightweight check on the stability of the AI-assisted search results, the same three questions were also submitted in later query rounds. These repeated searches were not intended to create an exhaustive systematic review, but to assess whether a recognizable core of publications reappears over time and how much variation occurs in the top returned results.

2.2. Record Management and Units of Analysis

For the statistical analysis, we distinguished between two units of analysis: query-result records and unique publications. The query-result record preserves the behavior of the search process, including repeated appearances of the same publication across related queries. The unique publication is the appropriate unit for substantive classification, since the same paper should not be coded multiple times simply because it was returned by more than one query.

The three queries produced 30 query-result records. After deduplication, these corresponded to 28 unique publications. The 30-record dataset was therefore used to analyze the search outputs themselves, such as publication year and publication type distributions, whereas the 28-publication dataset was used for content coding and subsequent technical classification.

2.3. Coding Scheme and Reliability Assessment

A predefined codebook was developed to classify each unique publication. The primary direction code distinguished between *AI4FM*, where AI or machine learning techniques are used to support formal methods; *FM4AI*, where formal methods are used to analyze, verify, constrain, or improve AI systems; and *HYBRID*, where both directions are substantially present. An *UNCLEAR* code was available for cases where the available information was insufficient. These codes were assigned from the publication title, abstract or returned summary, and stated technical contribution. Where a publication contained elements of more than one direction, the primary code was assigned according to the dominant role in the paper; *HYBRID* was used when AI support for formal methods and formal-method analysis of AI systems were both substantial.

A second-level technical code was applied only to publications classified as *AI4FM*. This focus was motivated by the progressive design of the later research questions, which explicitly asked how GenAI can support formal methods. It was also supported by the returned literature itself, in which *AI4FM* formed the largest primary category. The second-level categories included theorem proving, model checking, specification synthesis, autoformalization, dataset or benchmark construction, tool or system development, formal concept analysis, and other *AI4FM*-relevant techniques.

Two coders independently coded the 28 unique publications using the codebook, with each coder blind to the other's codes during the initial coding stage. Before resolving

disagreements, inter-coder agreement was assessed using percentage agreement and Cohen's kappa [28]. Disagreements, low-confidence codes, and cases marked for discussion were then resolved through adjudication. The adjudicated codes were used for the final statistical summaries and figures.

3. Results

In this section, we consider three research questions posed to Google Scholar Labs [6] and the first ten entries of relevant publications that are returned. The AI-generated summary information provided by Google is included, together with a critique of its usefulness, especially in the context of the research questions.

Each publication that is listed starts with basic information adapted from the results provided by Google Scholar Labs, including a citation for the publication included in the general references. See also the note in Section 7. The Google Scholar Labs summary format uses a standard style, with an overview sentence, followed by two bullet points, each with a brief description of aspects covered in the publication. It should be noted that computer science is renowned for using conferences as much as journals for the dissemination of research, unlike most other science-related disciplines. In addition, the *arXiv* online repository is being increasingly used for fast publication, especially in computer science.

The three research questions used in this investigation are very general, just mentioning formal methods and AI, but no particular areas. Researchers using this approach in practice are likely to be interested in a particular area of formal methods (or AI). Here, the areas returned are determined by the Google Scholar Labs algorithm. That said, the areas that emerge (e.g., use of the *Lean* proof assistant [29]) are indicative of where most of research is currently active. Note that these questions are precisely as given to Google Scholar Labs, both initially and when rerunning the experiment subsequently.

3.1. Question 1: Is Artificial Intelligence Useful for Formal Methods?

This general research question was first posed to Google Scholar Labs, with the following results for the top ten publications returned.

Publication 1.1

[BOOK] *Formal Methods in Artificial Intelligence* [30].

A Ramsay—1988—<https://books.google.com> (accessed on 2 June 2026)

Explains that Artificial Intelligence turns to formal logic for powerful and well-behaved knowledge representation languages.

- Enriching AI with Logic: Aims to show how Artificial Intelligence can be enriched by incorporating recent advances in formal logic, including extensions of classical logic for reasoning about time, knowledge, or uncertainty.
- Borrowing Formal Techniques: Details that techniques from formal logic have been borrowed by AI researchers, often for applications outside their original context, resulting in useful and constructive work.

Cited by 118

This is an extensive and well-cited book, but rather dated for current relevance, having been published in 1988 [30]. The book's title fits very well with the research question, but its age means that it is now largely of historical interest.

Publication 1.2

[PDF] <https://researchgate.net> (accessed on 2 June 2026)

Are formal methods applicable to machine learning and artificial intelligence? [31].

M Krichen, A Mihoub, MY Alzahrani, et al.—Proceedings, 2022—<https://ieeexplore.ieee.org> (accessed on 2 June 2026)

Examines state-of-the-art formal methods for the verification and validation of machine learning (a subset of AI) systems.

- **Verifying ML Models:** Reviews research works dedicated to the verification of specific machine learning models, such as support vector machines and decision tree ensembles.
- **Verifying ML Phases:** Summarizes formal verification efforts for the data preparation, training phases, and the resulting trained machine learning models, including neural networks.

Cited by 128

This conference paper considers the use of formal methods in machine learning and AI to aid in verifying their correctness. Its publication in 2022 means that it was written before ChatGPT (the GPT-3.5-based initial public release, launched on 30 November 2022) became generally available, so it predates the recent GenAI revolution. That said, it is well-cited and poses a prescient question.

Publication 1.3

[PDF]—<https://zulipchat.com> (accessed on 2 June 2026)

A promising path towards autoformalization and general artificial intelligence [32].

C Szegedy—Proceedings, 2020—Springer

Argues that autoformalization, an AI system that converts natural language into machine-verifiable formalization, is a promising direction for creating automated mathematical reasoning systems.

- **Automating Formal Proofs:** Discusses the potential of autoformalization to automate the creation of computer verifiable proofs, which traditionally require significant human effort.
- **AI Techniques:** Explains that advances in machine learning techniques, such as neural networks and reinforcement learning, are paving the way for strong automated reasoning and formalization systems.

Cited by 94

This 2020 conference paper predates even **publication 1.2** and discusses the potential of AI aiding computer-verifiable proofs and automated reasoning, both useful in computer-aided formal methods. However, like **publication 1.2**, it looks at future directions, which are now overtaken by more recent developments.

Publication 1.4

Computer theorem proving and artificial intelligence [33].

H Wang—Computation, Logic, Philosophy: A Collection of Essays, 1990—Springer

Discusses the use of computer programs for theorem proving, which is a key component of formal methods, and its relationship to Artificial Intelligence (AI).

- **Early Program Results:** Describes early programs that applied formal logic to prove a large number of theorems from Principia Mathematica in the propositional and predicate calculus, demonstrating the utility of AI for formal methods.
- **Expert Systems for Proof:** Presents a plea for 'expert systems' in the context of theorem proving, suggesting that partial, efficient strategies should be emphasized over general algorithms for proving theorems on a machine.

Cited by 15

This 1990 philosophical chapter, while having some relevance, is now largely of historical interest, looking back to mathematical theorems in Alfred North Whitehead and Bertrand

Russell's 1910–1913 three-volume *Principia Mathematica* [34] and traditional expert systems (first introduced around 1965 by Edward Feigenbaum et al.), a decision-making form of AI that has been superseded by GenAI.

Publication 1.5

Automatic B-model repair using model checking and machine learning [35].

CH Cai, J Sun, G Dobbie—Automated Software Engineering, 2019—Springer

Proposes B-repair, an approach for automated repair of faulty models written in the B formal specification language, which uses machine learning techniques to estimate the quality of suggested repairs.

- AI and Model Checking Integration: Combines model checking and machine learning-guided techniques to automatically find and repair faults during the design of software systems.
- AI for Repair Quality Estimation: Uses machine learning techniques to learn the features of state transitions, which helps in ranking suggested repairs and selecting the best one for revising the formal model.

Cited by 22

This 2019 journal paper is interesting because it uses a specific formal methods approach, namely the specification language of the B-Method [36]. However, although it has received some citations, the ideas in this paper are largely subsumed by GenAI developments.

Publication 1.6

[PDF] <https://acm.org> (accessed on 2 June 2026)

Formal verification of neural network controlled autonomous systems [37].

X Sun, H Khedr, Y Shoukry—Proceedings, 2019—<https://dl.acm.org> (accessed on 2 June 2026)

Explores the safety and reliability challenges arising from the increasing use of Artificial Intelligence (AI) agents, such as deep neural network models, in controlling physical and mechanical systems.

- Formal Verification of NNs: Focuses on applying formal verification techniques to neural network models, including those used in autonomous systems, to provide safety guarantees that testing and falsification methods lack.
- NN-Controlled Robot Safety: Considers the formal verification of the safety of an autonomous robot equipped with a Neural Network controller that processes sensor data (LiDAR images) to produce control actions, demonstrating a practical application of formal methods to an AI-controlled system.

Cited by 208

This 2019 conference paper is very well cited and considers the use of formal methods to verify neural networks, rather than using AI to aid formal methods. For this reason, its findings are likely to remain valid; hence, a reason for its high citations, even though it predated recent GenAI developments.

Publication 1.7

[PDF] <https://arxiv.org> (accessed on 2 June 2026)

The fusion of large language models and formal methods for trustworthy AI agents: A roadmap [38].

Y Zhang, Y Cai, X Zuo, X Luan, K Wang, et al.—*arXiv* preprint, 2024—<https://arxiv.org> (accessed on 2 June 2026)

Highlights that Large Language Models (LLMs), a form of Artificial Intelligence, can significantly enhance the usability, efficiency, and scalability of existing formal methods (FM) tools.

- **Advancing Formal Methods Tasks:** Explores how LLMs can drive advancements in formal methods by automating the formalization of specifications, leveraging LLMs to enhance theorem proving, and developing intelligent model checkers.
- **AI as Intelligent Assistants:** Details that AI, specifically LLMs, can act as intelligent assistants to formal methods by automating tedious tasks and enhancing the capabilities of existing FM algorithms for verification processes.

Cited by 31

This is the first paper to be returned that dates within the recent period of GenAI development (2024). It is also published on the em arXiv paper archive website run by Cornell University. This was launched in 1991 and is now very well established for computer science-related publications, as well as other scientific and mathematical fields. Papers are not formally peer reviewed, but are vetted for relevance. It thus provides a fast means of publication, which is very attractive in the rapid-paced GenAI area. This is the first paper returned that is very relevant to the current advances in using AI to aid formal methods by providing an “intelligent” assistant for their use. The paper highlights efficiency, scalability, and usability, which are all significant issues in the application of formal methods.

Publication 1.8

[PDF] <https://neurips.cc> (accessed on 2 June 2026)

MLFMF: data sets for machine learning for mathematical formalization [39].

A Bauer, et al.—Proceedings, 2023—<https://proceedings.neurips.cc> (accessed on 2 June 2026)

Introduces MLFMF, a collection of data sets for benchmarking recommendation systems that use machine learning to support the formalization of mathematics with proof assistants.

- **Evaluates ML Approaches:** Reports baseline results using standard graph and word embeddings, tree ensembles, and instance-based learning algorithms to evaluate machine learning approaches to formalized mathematics.
- **Aiding Formalization Efforts:** Describes how machine learning methods are used to address premise selection, such as recommending theorems useful for proving a given statement, and how various neural networks and reinforcement learning are applied to automated theorem proving.

Cited by 13

This 2023 conference paper was published during the early days of the recent GenAI developments. It uses formalization but is less relevant to the application of formal methods in a software engineering context. It is not very well cited and could be considered of marginal interest with respect to the posed research question.

Publication 1.9

[PDF] <https://units.it> (accessed on 2 June 2026)

Machine learning methods in statistical model checking and system design—tutorial [40].

L Bortolussi, D Milios, G Sanguinetti—Proceedings, 2015—Springer

Reviews methodologies that employ machine learning concepts, specifically Bayesian methods and Gaussian Processes, for model checking and system design/parameter synthesis against stochastic dynamical models.

- **Convergence of Ideas:** Focuses on the increasingly fertile convergence of ideas from machine learning and formal modeling, highlighting it as a growing research area.

- **Bayesian Methods for Efficiency:** Details the use of powerful methodologies from Bayesian machine learning and reinforcement learning to obtain efficient algorithms for statistical model checking and parameter synthesis.

Cited by 13

This is a tutorial paper from 2015, and is thus rather out of date to be of current research interest. It considers formal modeling, rather than the use of formal methods in a software engineering context, and is not particularly well cited in any case.

Publication 1.10

Machine learning and formal concept analysis [41].

SO Kuznetsov—Proceedings, 2004—Springer

Explains that a model of learning from positive and negative examples, a core aspect of Artificial Intelligence/Machine Learning, can be naturally described using the means of Formal Concept Analysis (FCA), a formal method.

- **Describing AI Models:** Demonstrates the usefulness of Formal Concept Analysis by applying its language to give natural descriptions of standard Machine Learning models, including version spaces and decision trees.
- **Comparing Approaches:** Suggests that using the language of FCA allows for the comparison of several machine learning approaches and the employment of standard FCA techniques within the domain of machine learning.

Cited by 206

Dating from 2004, this paper, although well cited, is now largely of historical interest. Although Google Scholar Labs considers Formal Concept Analysis to be a “formal method”, it is really related to a concept hierarchy forming a formal ontology from a collection of objects together with their properties. This has limited relevance to formal methods for development in a software engineering context.

Summary:

The most relevant publication returned by the first research question is **publication 1.7**. Many of the others predate recent GenAI developments. They may be of historical interest, but if researching current trends, they are less useful.

3.2. Question 2: Can Generative AI Be Used to Aid Formal Methods?

Next, this more specific research question was posed to Google Scholar Labs, to gauge the usefulness of using GenAI with formal approaches to software engineering. This improved the relevance of publications that were returned.

Publication 2.1

[PDF] <https://fiu.edu> (accessed on 2 June 2026)

Dehallucinating large language models using formal methods guided iterative prompting [42].

S Jha, SK Jha, P Lincoln, et al.—Proceedings, 2023—<https://ieeexplore.ieee.org> (accessed on 2 June 2026)

Exploits the dialog capability of large language models (LLMs) to iteratively steer them toward responses consistent with a correctness specification using formal methods.

- **Self-Monitoring Architecture:** Describes a self-monitoring and iterative prompting architecture that uses formal methods to automatically detect errors and inconsistencies in LLM responses, focusing on applications in autonomous systems.
- **Formal Methods Integration:** Presents a loose coupling between the LLM and a counterexample generator, allowing the use of different formal methods, such as model checking and SMT solving, to achieve a high-assurance solution.

Cited by 106

This 2023 conference paper covers the use of formal methods to aid Large Language Models (LLMs) for AI, rather than the reverse, with the aim of avoid hallucination. Nevertheless, it is highly cited and of potential interest.

Publication 2.2

[PDF] <https://arxiv.org> (accessed on 2 June 2026)

Generative language modeling for automated theorem proving [43].

S Polu, I Sutskever—*arXiv* preprint, 2020—<https://arxiv.org> (accessed on 2 June 2026)

Explores the application of transformer-based language models, a form of Generative AI, to the domain of automated theorem proving, a key aspect of formal methods.

- GPT-f Proof Assistant: Presents an automated prover and proof assistant, GPT-f, which uses generative language modeling for the Metamath formalization language.
- Practical Contribution of AI: Contributed new short proofs that were adopted by the formal mathematics community, demonstrating the practical utility of a deep-learning based system in this field.

Cited by 409

This 2020 preprint paper is early in recent GenAI developments, but is very well cited. It covered using GenAI to aid automated theorem proving, so it is relevant to formal methods-style verification with computer-based proof tools. However, it is more oriented to proofs of general mathematical theorems using the Metamath proof assistant tool, rather than proofs relating to formal methods development approaches for software development. The former tend to be deeper but more concisely stated, whereas proofs for software tend to be shallower, but much more extensive due to the complexity of software in practice.

Publication 2.3

[PDF] <https://arxiv.org> (accessed on 2 June 2026)

The fusion of large language models and formal methods for trustworthy AI agents: A roadmap [38].

Y Zhang, Y Cai, X Zuo, X Luan, K Wang, et al.—*arXiv* preprint, 2024—<https://arxiv.org> (accessed on 2 June 2026)

Highlights how the advanced learning capabilities and adaptability of Large Language Models (LLMs), a form of Generative AI, can significantly enhance the usability, efficiency, and scalability of existing Formal Methods (FM) tools.

- Driving FM Advancements: Examines how LLMs can drive advancements in existing formal methods and tools, focusing on utilizing LLMs for automated formalization of specifications, enhancing theorem proving, and developing intelligent model checkers.
- Automating Verification Tasks: Proposes using LLMs to make formal methods more efficient and effective by serving as intelligent assistants that automate tedious tasks, such as generating structured code and symbolic representations for verification processes.

Cited by 31

This paper was provided as *publication 1.7* for the first research question posed earlier. It is interesting to note the difference in the summary by Google Scholar Labs. The meaning is largely the same, but the wording is different, presumably due to the random element in the generation of output from GenAI algorithms. As mentioned previously, it is very relevant to the area of interest for both research questions, so it is good to see that it is returned in both cases. With the second more focused research question, it is returned earlier in the list (third rather than seventh).

Publication 2.4

[PDF] <https://arxiv.org> (accessed on 2 June 2026)

Goedel-Prover: A frontier model for open-source automated theorem proving [44].

Y Lin, S Tang, B Lyu, J Wu, H Lin, K Yang, et al.—*arXiv preprint*, 2025—<https://arxiv.org> (accessed on 2 June 2026)

Introduces an open-source language model, Goedel-Prover, that achieves state-of-the-art performance in automated formal proof generation for mathematical problems, directly demonstrating the use of generative AI in formal methods.

- **Training LLMs for Formalization:** Explains the process of training Large Language Models (LLMs) to convert natural language math problems into equivalent formal statements in Lean 4, which is a key step in applying AI to formal methods.
- **Developing Formal Proof Datasets:** Demonstrates training a series of provers to develop a large dataset of formal proofs, showcasing how generative AI can overcome the scarcity of formalized mathematical data.

Cited by 75

This paper presents Version 1 of the Goedel-Prover automated theorem prover. Although published in 2025, it had already received 75 citations in early 2026. Being published on the *arXiv* archive no doubt contributed to the speed of recognition by authors of other publications. As for **publication 2.3**, it is also relevant to the research question, although it concentrates more on proofs for mathematical theorems, like **publication 2.2**, rather than software verification proofs. It is the first publication to cover the Lean proof assistant and its associated functional programming language (<https://lean-lang.org> (accessed on 2 June 2026)), originally introduced in 2013. This paper uses Lean 4 language, released in 2021. Recently, Lean has become central in investigations of aiding mathematical proofs using AI. The Goedel-Prover (<https://goedel-lm.github.io> (accessed on 2 June 2026)) covered in this paper is an open-source language model aimed at automated formal proof generation for mathematical problems using the Lean 4 language.

Publication 2.5

[PDF] <https://arxiv.org> (accessed on 2 June 2026)

Goedel-Prover-V2: Scaling formal theorem proving with scaffolded data synthesis and self-correction [45]

Y Lin, S Tang, B Lyu, Z Yang, et al.—*arXiv preprint*, 2025—<https://arxiv.org> (accessed on 2 June 2026)

Introduces Goedel-Prover-V2, an open-source language model series that achieves a new state-of-the-art in automated theorem proving (a key component of formal methods).

- **Generating Formal Proofs:** Demonstrates the use of a generative AI model to create machine-verifiable proofs in formal languages like Lean, which is essential for formal methods.
- **Compiler-Guided Proof Correction:** Presents a method where the model iteratively revises its proofs by utilizing feedback from the Lean compiler (verifier-guided self-correction), integrating verification feedback into the proof generation process.

Cited by 29

This paper relates to that in the previous **publication 2.4** [44], with overlapping authors and also on the *arXiv* archive. It covers Version 2 of the Goedel-Prover automated theorem prover, including use of the Lean 4 language. Even this had already received 29 citations by early 2026. It looks like developments in the Goedel-Prover will be worth watching with respect to the possibilities of proofs that could be relevant to formal methods approaches.

Publication 2.6

[PDF] <https://arxiv.org> (accessed on 2 June 2026)

DeepSeek-Prover-V1.5: Harnessing proof assistant feedback for reinforcement learning and Monte-Carlo tree search [46].

H Xin, ZZ Ren, J Song, Z Shao, et al.—*arXiv preprint*, 2024—<https://arxiv.org> (accessed on 2 June 2026)

Introduces DeepSeek-Prover-V1.5, an open-source language model specifically designed for theorem proving in Lean 4, demonstrating a direct application of generative AI in formal methods.

- **Influences Mathematical Reasoning:** Discusses how recent advancements in large language models have significantly influenced mathematical reasoning and theorem proving in artificial intelligence, confirming the relevance of generative AI in this domain.
- **Employs Proof Generation Strategies:** Presents two common strategies language models use in formal theorem proving: proof-step generation and whole-proof generation, detailing mechanisms for utilizing generative AI in creating formal proofs.

Cited by 42

This 2024 paper covers the DeepSeek-Prover proof assistant, also covered later in a related paper under **publication 2.9** [47]. DeepSeek is a Chinese AI company that was founded in 2023. Its Large Language Models have significantly lower training costs than others like OpenAI's GPT-4 and Meta's Llama 3.1, for example. DeepSeek-Prover is an open-source AI model for mathematical theorem proving and formal reasoning. It is already available as Version 2 (<https://prover-v2.com> (accessed on 2 June 2026)) in this fast-moving field. As for **publication 2.5**, it uses the Lean 4 language and is aimed at mathematical theorems in general. Nevertheless, this proof assistant could be one to watch for developments in the future with respect to use for software verification.

Publication 2.7

[PDF] <https://arxiv.org> (accessed on 2 June 2026)

Verification of image-based neural network controllers using generative models [48].

SM Katz, AL Corso, et al.—*Journal*, 2022—<https://arc.aiaa.org> (accessed on 2 June 2026)

Proposes a method to address the scaling challenges of formal methods for image-based neural network controllers by training a Generative Adversarial Network (GAN) to map states to plausible input images.

- **Reducing Input Dimension:** Shows that concatenating the generative network with the control network results in a network with a low-dimensional input space, enabling the use of existing closed-loop verification tools to obtain formal guarantees.
- **Applying to Safety Guarantees:** Applies the approach to provide safety guarantees for an image-based neural network controller for an autonomous aircraft taxi problem.

Cited by 85

This well-cited 2022 journal paper considers scaling issues in using formal methods for neural network controllers for image-based sensors. This is a relatively specialist area, but solutions for formal methods scaling challenges are always of interest. With the speed of developments, no doubt there are likely to have been improvements since the publication of this paper.

Publication 2.8

[PDF] <https://arxiv.org> (accessed on 2 June 2026)

Kimina-Prover Preview: Towards large formal reasoning models with reinforcement learning [49].

H Wang, M Unsal, X Lin, M Baksys, et al.—*arXiv preprint*, 2025—<https://arxiv.org> (accessed on 2 June 2026)

Introduces Kimina-Prover Preview, a large language model designed for formal theorem proving, demonstrating its application in Lean 4 proof generation.

- Emulates Human Reasoning: Employs a reasoning-driven exploration paradigm and a structured formal reasoning pattern to emulate human problem-solving strategies in the Lean 4 proof assistant, iteratively generating and refining proof steps.
- State-of-the-Art Performance: Achieves state-of-the-art results on the miniF2F benchmark for automated theorem proving, confirming the effectiveness of generative AI models in formal verification.

Cited by 74

As with publications *Publication 2.4*, *Publication 2.5*, and *Publication 2.6*, this 2025 paper makes use of the Lean 4 language. It introduced the Kimina-Prover Preview Large Language Model for theorem proving. This relatively new proof tool is already well cited. As such, it looks like it will be worth monitoring for its developments and possible use of formal methods in the future.

Publication 2.9

[PDF] <https://arxiv.org> (accessed on 2 June 2026)

DeepSeek-Prover: Advancing theorem proving in LLMs through large-scale synthetic data [47].

H Xin, D Guo, Z Shao, Z Ren, Q Zhu, et al.—*arXiv preprint*, 2024—<https://arxiv.org> (accessed on 2 June 2026)

Introduces an approach to generate extensive Lean 4 proof data from mathematical competition problems to address the lack of training data for large language models (LLMs) in formal theorem proving.

- Automated Theorem Proving: Details how the fine-tuned DeepSeekMath model, using the synthetic dataset, achieved high whole-proof generation accuracies on the Lean 4 miniF2F test, surpassing baseline models like GPT-4, demonstrating the LLM's capability in automated theorem proving (a core formal method task).
- Formal Proof Generation: Explains that generative AI models, like LLMs, can be utilized to automatically generate and verify formal proofs in environments such as Lean 4, overcoming the challenge of manually crafting formal proofs.

Cited by 147

This relates to DeepSeek-Prover, also covered in 2024 under *publication 2.6* [46]. As discussed earlier, being based on DeepSeek GenAI technology, this could be an efficient approach to theorem proving, and as such, well worth monitoring for potential formal methods use or adaptation.

Publication 2.10

DEVS Copilot: Towards Generative AI-Assisted Formal Simulation Modeling based on Large Language Models [50].

T Carreira-Munich, et al.—*Proceedings*, 2024—<https://ieeexplore.ieee.org> (accessed on 2 June 2026)

Explores the use of generative AI, specifically Large Language Models like GPT-4, to assist in obtaining correct executable simulation models, which are a form of formal specification.

- Uses DEVS Formalism: Adopts the Discrete Event System Specification (DEVS), a sound modeling and simulation formalism, as a suitable candidate for exploring the limits of LLM capabilities to generate correct generalized simulation models.
- Prototype Tool Development: Introduces a methodology and tool, DEVS Copilot, which transforms natural language descriptions of systems into a DEVS executable

simulation model using AI, succeeding at producing correct DEVS simulations in a case study.

Cited by 5

DEVS (Discrete Event System Specification) is a modular and hierarchical formalism for modeling and analyzing systems, including discrete event systems, e.g., as described by state transition tables. Although DEVS was originally formulated in 1976, it does not seem to have been widely adopted, especially in the formal methods community. This 2024 conference paper presents a prototype tool, DEVS Copilot, for transforming a natural language description into a formal model, using GenAI. In practice, this is likely to need an interactive approach in general. The paper is not very well cited, so the tool looks more like a prototype example rather than one that will be adopted more generally.

Summary:

The publications returned by this research question are certainly more relevant to recent research. In particular, several are based on the Lean 4 functional language for theorem proving. This looks like an important focus for current and probably future research in this area.

3.3. Question 3: What Are the Latest Papers Investigating the Use of Generative AI to Support Formal Methods?

Finally, it was realized that Google Scholar Labs was not necessarily returning the latest publications in this fast-moving area. Earlier publications may have more citations, potentially biasing Google to return these rather than the latest research. This third research question aimed to concentrate results on the latest research papers in this area.

Publication 3.1

[PDF] <https://arxiv.org> (accessed on 2 June 2026)

A new era in software security: Towards self-healing software via large language models and formal verification [51].

N Tihanyi, Y Charalambous, et al.—Proceedings, 2025—<https://ieeexplore.ieee.org> (accessed on 2 June 2026)

Presents a novel approach integrating Large Language Models (LLMs) with Formal Verification for automatic software vulnerability repair, using Bounded Model Checking (BMC) to identify vulnerabilities and extract counterexamples for the LLM.

- **Introduces ESBMC-AI Framework:** Introduces the ESBMC-AI framework as a proof of concept, leveraging the Efficient SMT-based Context-Bounded Model Checker (ESBMC) and a pre-trained transformer model to detect and fix errors in C programs.
- **Evaluates Automatic Repair Capabilities:** Evaluates the approach on 50,000 C programs and demonstrates its capability to automate the detection and repair of issues such as buffer overflow, arithmetic overflow, and pointer dereference failures with high accuracy.

Cited by 123

This 2025 conference paper presents a novel approach that is already very well cited, even though it is presented as a proof of concept. Model checking is a widely used approach for the verification of hardware and software systems with liveness and safety requirements. Bounded model checking combines model checking with satisfiability solving [52], using a SAT solver computer program to solve the Boolean satisfiability problem (SAT) to find an interpretation (assignment of a meaning) that satisfies a given Boolean formula. In this paper, SMT (Satisfiability Modulo Theories), which generalizes SAT to cover numbers, data structures, etc., is used to determine whether a mathematical formula is satisfiable. This paper specifically uses the C programming language. The approach is worth monitoring for developments in the future. See also the related *publication 3.3* below.

Publication 3.2

[PDF] <https://arxiv.org> (accessed on 2 June 2026)

Kimina-Prover Preview: Towards large formal reasoning models with reinforcement learning [49].

H Wang, M Unsal, X Lin, M Baksys, et al.—*arXiv* preprint, 2025—<https://arxiv.org> (accessed on 2 June 2026)

Introduces Kimina-Prover Preview, a large language model utilizing a novel reasoning-driven exploration paradigm for formal theorem proving, trained with a large-scale reinforcement learning (RL) pipeline from Qwen2.5-72B.

- Lean 4 Proof Generation: Demonstrates strong performance in Lean 4 proof generation by employing a structured reasoning pattern, which allows the model to emulate human problem-solving strategies and iteratively refine proof steps.
- Formal Reasoning Pattern: Presents a formal reasoning pattern designed to align informal mathematical reasoning data with its translation to formal proofs, addressing a critical challenge for LLMs in this domain.

Cited by 74

This is the same paper as that discussed in *publication 2.8*. The Google Scholar Labs summary is slightly different in content, perhaps due to the different research question being posed. With this research question, it is rated even more highly as *publication 3.2*. Being based on the popular Lean 4 language and already highly cited even though it was published in 2025, the Kimina-Prover could be an important development and worth monitoring for the future.

Publication 3.3

[PDF] <https://acm.org> (accessed on 2 June 2026)

The FormAI dataset: Generative AI in software security through the lens of formal verification [53].

N Tihanyi, T Bisztray, R Jain, et al.—*Proceedings*, 2023—<https://dl.acm.org> (accessed on 2 June 2026)

Introduces the FormAI dataset, which consists of 112,000 AI-generated C programs classified for vulnerabilities using a formal verification method.

- Formal Verification Method: Employs the Efficient SMT-based Bounded Model Checker (ESBMC) for formal verification, which uses model checking, abstract interpretation, constraint programming, and satisfiability modulo theories to definitively detect vulnerabilities.
- AI Code Generation: Details the process of using Large Language Models (LLMs), specifically GPT-3.5-turbo, and a dynamic zero-shot prompting technique to generate diverse C programs with varying complexity.

Cited by 100

The first author is the same as for the paper returned under *publication 3.1* [51]. This paper is an earlier 2023 conference paper. Although this paper is well cited, most likely, *publication 3.1* is best for monitoring the latest research in this area.

Publication 3.4

[PDF] <https://acm.org> (accessed on 2 June 2026)

Proof automation with large language models [54].

M Lu, B Delaware, T Zhang—*Proceedings*, 2024—<https://dl.acm.org> (accessed on 2 June 2026)

Proposes PALM, a novel proof automation approach that combines Large Language Models (LLMs) and symbolic methods in a generate-then-repair pipeline to create formal proofs.

- **Evaluates PALM Performance:** Evaluates the PALM system on a large dataset of over 10,000 theorems, demonstrating significant improvement over existing state-of-the-art methods.
- **Identifies LLM Errors:** Identifies common errors made by LLMs, such as GPT-3.5, when generating formal proofs, noting that they often succeed with high-level structure but fail with low-level details.

Cited by 36

This 2024 conference paper proposes a new proof automation approach, dubbed PALM, to use LLMs in the generation of formal proofs for theorems, using the Coq theorem prover (renamed “Rocq” in 2025, <https://rocq-prover.org> (accessed on 2 June 2026)) and GPT-3.5. It notes that GPT-3.5 can generate a valid high-level structure for proofs, but makes mistakes at lower levels. Thus, the approach allows iterative correction of the lower levels. The paper claims to outperform other state-of-the-art approaches to proof generation. The approach looks promising, although the number of citations could be higher. This may be due to the use of Coq rather than Lean; the latter seems to have a greater research focus at present. Note that the paper is also available on the *arXiv* archive (<https://doi.org/10.48550/arXiv.2409.14274>), no doubt for speed of dissemination and easy open access.

Publication 3.5

[PDF] <https://neurips.cc> (accessed on 2 June 2026)

Autoformalization with large language models [55].

Y Wu, AQ Jiang, W Li, et al.—Proceedings, 2022—<https://proceedings.neurips.cc> (accessed on 2 June 2026)

Explores the use of Large Language Models (LLMs) for autoformalization, the process of translating natural language mathematics into formal specifications and proofs.

- **Translating Math Problems:** Shows that LLMs can translate a significant portion of mathematical competition problems perfectly to formal specifications in Isabelle/HOL, indicating their capability in generating formal language from natural language.
- **Improving Theorem Prover:** Demonstrates the usefulness of autoformalization by improving a neural theorem prover, resulting in a new state-of-the-art result on the MiniF2F theorem proving benchmark.

Cited by 311

This 2022 conference paper is highly cited. It looks at transforming natural language mathematical descriptions into complete formal specifications with associated proofs. The paper uses the well-established Isabelle/HOL theorem prover [56]. Related developments include the DeepIsaHOL project (<https://github.com/yonoteam/DeepIsaHOL> (accessed on 2 June 2026)), using reinforcement learning (RL), machine learning paradigm where an autonomous agent learns to make decisions by trial and error, to improve proof-automation in theorem proving using the Isabelle proof assistant.

Publication 3.6

[PDF] <https://hal.science> (accessed on 2 June 2026)

Applications of AI to study of finite algebraic structures and automated theorem proving [57].

B Shminke—Thesis, 2023—<https://theses.hal.science> (accessed on 2 June 2026)

Proposes a generative neural network architecture for creating finite models of algebraic structures, drawing inspiration from image generation models like Generative Adversarial Networks (GANs) and autoencoders.

- **Semigroup Generation Package:** Contributes a Python package for generating small-sized finite semigroups, serving as a reference implementation of the proposed generative method.

- Reinforcement Learning for Provers: Designs a general architecture for guiding saturation provers using reinforcement learning algorithms, which is a key component of automated theorem proving (a formal method).

Cited by 0

Unusually, compared to most of the publications included, this is a PhD research thesis rather than a journal or conference paper. It dates from 2023, which, in this fast-moving area, may already be dated. It has no citations on Google Scholar, which also raises questions over its inclusion and influence. As introduced under **publication 3.5**, it uses reinforcement learning. Without associated peer-reviewed papers, this research may have little impact in practice.

Publication 3.7

[PDF] <https://arxiv.org> (accessed on 2 June 2026)

Specify what? Enhancing neural specification synthesis by symbolic methods [58].

G Granberry, W Ahrendt, M Johansson—Proceedings, 2024—Springer

Investigates how combinations of Large Language Models (LLMs) and symbolic analyses can synthesize specifications for C programs in the ACSL language.

- Augmenting LLMs with Formal Tools: Augments LLM prompts with outputs from formal methods tools, specifically Pathcrawler and EVA from the Frama-C ecosystem, to produce C program annotations.
- Impact of Symbolic Analysis: Demonstrates that integrating symbolic analysis enhances the quality of annotations, making them more context-aware and attuned to runtime errors, and enabling inference of program intent rather than just behavior.

Cited by 11

This 2024 conference paper is not very well cited. The ACSL language is the ANSI/ISO C Specification Language, a behavioral specification language for C programs. The paper investigates the use of LLMs in synthesizing C program specifications in ACSL. It uses tools from Frama-C (<https://frama-c.com> (accessed on 2 June 2026)), an open-source, extensible, and collaborative platform for source-code analysis of C programs. Given the low number of citations, it is not clear if this approach will be widely adopted in practice.

Publication 3.8

[PDF] <https://arxiv.org> (accessed on 2 June 2026)

Generative AI augmented induction-based formal verification [59].

A Kumar, DN Gadde—Proceedings, 2024—<https://ieeexplore.ieee.org> (accessed on 2 June 2026)

Demonstrates how Generative AI can be used in induction-based formal verification to increase the verification throughput.

- Automating Assertion Generation: Explores the potential of GenAI to automate assertion generation, which helps resolve the complexity and labor-intensive nature of formal verification.
- GenAI-Based Verification Flow: Presents a GenAI-based flow that generates helper assertions and generates lemmas in the event of an inductive step failure during formal verification.

Cited by 7

This 2024 conference paper claims to demonstrate the use of GenAI to improve the efficiency of formal verification. The subject area certainly looks very relevant. However, given the low number of citations, its influence may have a low impact in practice.

Publication 3.9

[PDF] <https://arxiv.org> (accessed on 2 June 2026)

Autoformalization in the Era of Large Language Models: A Survey [60].

K Weng, L Du, S Li, W Lu, H Sun, et al.—*arXiv* preprint, 2025—<https://arxiv.org> (accessed on 2 June 2026)

Examines how autoformalization, driven by Large Language Models (LLMs), is applied across various mathematical domains and levels of difficulty, relating to automated theorem proving.

- **Autoformalization Workflow Analysis:** Analyzes the end-to-end workflow of autoformalization, including data preprocessing, model design, and evaluation protocols for converting informal math to formal representations.
- **Verifiability of LLM Outputs:** Explores the emerging role of autoformalization in enhancing the verifiability and trustworthiness of LLM-generated outputs, bridging the gap between human intuition and machine-verifiable reasoning.

Cited by 11

This is a 2025 survey paper, which can be a useful starting point for investigating further research. Survey papers can receive a large number of citations if they are influential, but this has not yet happened with this paper.

Publication 3.10

[PDF] <https://arxiv.org> (accessed on 2 June 2026)

Vulnerability detection: from formal verification to large language models and hybrid approaches: a comprehensive overview [61].

N Tihanyi, T Bisztray, MA Ferrag, et al.—*arXiv* preprint, 2025—<https://arxiv.org> (accessed on 2 June 2026)

Presents a comprehensive study of state-of-the-art software testing and verification, focusing on classical formal methods, Large Language Model (LLM)-based analysis, and emerging hybrid techniques.

- **Enhancing Verification Scalability:** Highlights the potential of hybrid systems, which combine formal rigor with LLM-driven insights, to enhance the effectiveness and scalability of software verification.
- **LLM-Assisted Formal Verification:** Details how integrating LLMs with formal methods and program analysis tools is a promising direction for improving software vulnerability detection by reducing the manual burden through automatic generation of formal assertions/invariants.

Cited by 6

The first author is the same as for the paper returned under *publication 3.1* [51] and the first two authors are the same as the paper under *publication 3.3* [53]. Like *publication 3.9*, this is a survey paper on the *arXiv* archive, but despite this, it has not yet been widely cited. Since the Google Scholar Labs search, the paper has been formally published in 2026 [62], so at least, it has now been peer reviewed.

Summary:

The publications for this research question are at least all recent (one from 2022, two from 2023, three from 2024, and four from 2025). An issue is that some are not well cited. On the other hand, some of these recent papers are very well cited already. It is these that are already influential and may lead to further worthwhile developments.

4. Statistical Summary and Analysis

This section summarizes what the Google Scholar Labs search surfaced, first as a set of returned query records and then as a set of coded publications. The purpose is not only to count the publications, but also to examine the structure of the returned literature: whether the search process favors recent work, what publication types dominate, which primary technical direction is most prominent, and which AI4FM techniques appear to be driving the current trend.

It should be noted that the sample size in this study is relatively small, so statistical analysis may be subject to “noise”. For a better statistical analysis, a larger sample would be needed. That said, as well as a statistical investigation, the authors were also interested in exploring the practical aspects of using Google Scholar Labs in supporting research, and these issues are discussed in Section 5.

4.1. Search-Output Characteristics

Figure 1 summarizes the publication years for the 30 query-result records returned by the three research questions. Since the yearly distribution is relatively sparse, the records are grouped into four time periods. This is a count of returned records rather than a count of unique publications, so repeated appearances of the same publication in different query results are retained.

The time distribution shows that the returned literature is strongly concentrated in recent years. Half of the records are from 2024–2025, and 22 of the 30 records are from 2022–2025. The subfigures also reflect the progressive nature of the three research questions. The first broad question returns a mixture of historical and recent publications, including three records from before 2015. By contrast, the second and third questions, which focus more directly on recent GenAI support for formal methods, return seven records each from 2024–2025. This supports the view that the relationship between GenAI and formal methods is a rapidly developing research area.

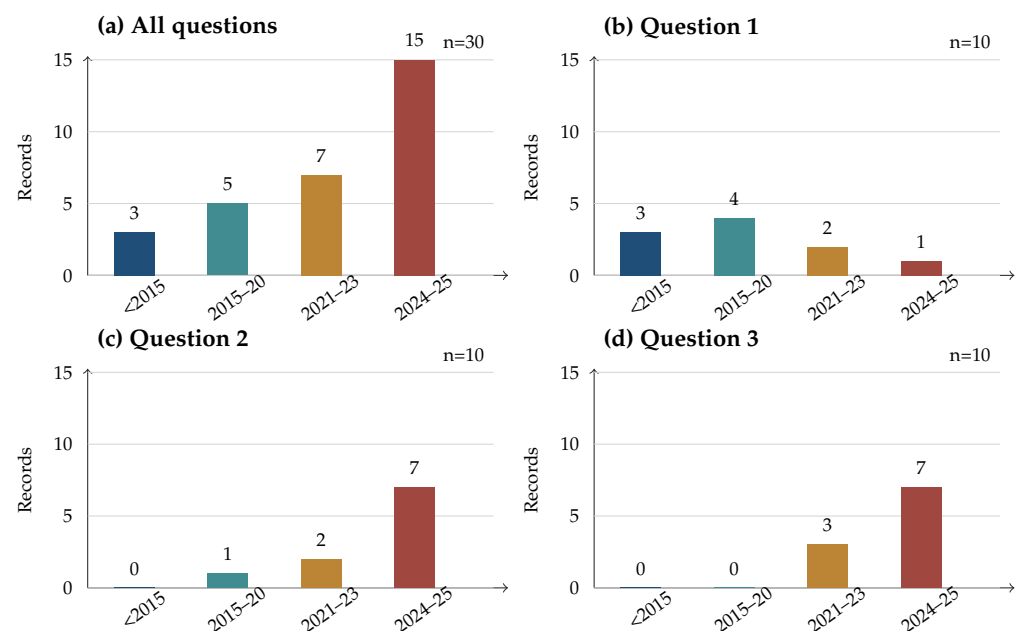


Figure 1. Publication-period distribution of the 30 query-result records returned by Google Scholar Labs: (a) all questions; (b) Question 1; (c) Question 2; (d) Question 3.

Figure 2 summarizes the publication types for the same 30 query-result records.

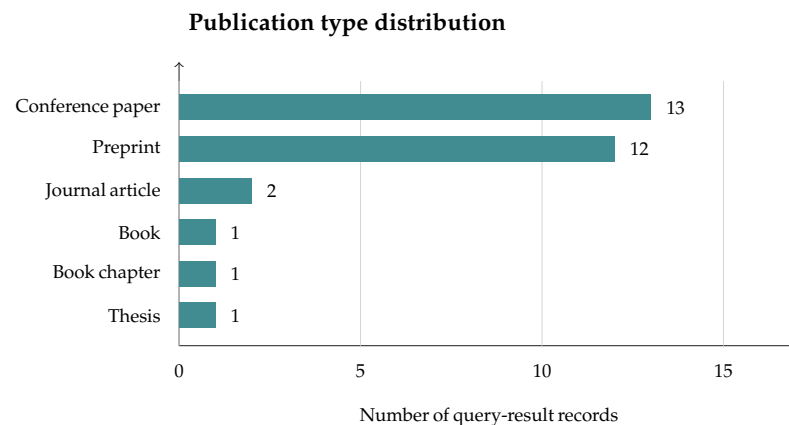


Figure 2. Publication-type distribution of the 30 query-result records returned by Google Scholar Labs.

The publication-type distribution is also revealing: conference papers and preprints dominate the returned results, with 13 and 12 records respectively. This reflects the fast-moving nature of the area, where work often appears first in conference venues or on preprint servers before slower journal publication.

4.2. Inter-Coder Agreement and Adjudication

The 30 query-result records were deduplicated into 28 unique publications before manual coding. The two independent coders completed blind coding before resolving the remaining differences through adjudication. Before adjudication, the coders agreed on the primary technical direction for 22 of the 28 unique publications, giving a percentage agreement of 78.6% and Cohen's kappa of 0.417. This level of agreement is best understood as moderate rather than conclusive. It is also affected by the uneven category distribution, since many returned publications were coded as *AI4FM*. The primary-direction distribution in Section 4.3 should therefore be interpreted as a cautious descriptive summary rather than as a definitive classification of the field. By contrast, the second-level *AI4FM* technique distribution in Section 4.4 is more stable, because agreement was stronger once both coders had identified publications as *AI4FM*.

The disagreement pattern is nevertheless informative. Most primary-direction disagreements concern borderline cases where AI techniques, formal verification, tool support, and assurance of AI systems are all present, making it difficult to decide which direction is dominant. These cases illustrate the fuzzy boundary between *AI4FM*, *FM4AI*, and *HYBRID* work in recent literature.

Agreement was stronger for the second-level *AI4FM* technique code. For the 19 publications that both coders coded as *AI4FM*, the coders agreed on the main *AI4FM* technique in 16 cases, giving a percentage agreement of 84.2% and Cohen's kappa of 0.784. This suggests that once the primary direction has been agreed, the more specific characterization of *AI4FM* work is comparatively stable. The remaining technique differences mainly concern whether a paper should be coded by its formal-method target, such as model checking or specification synthesis, or by its implementation form, such as tool or system development.

4.3. Primary Technical Directions

Figure 3 shows the technical direction distribution based on the adjudicated codes for the 28 unique publications.

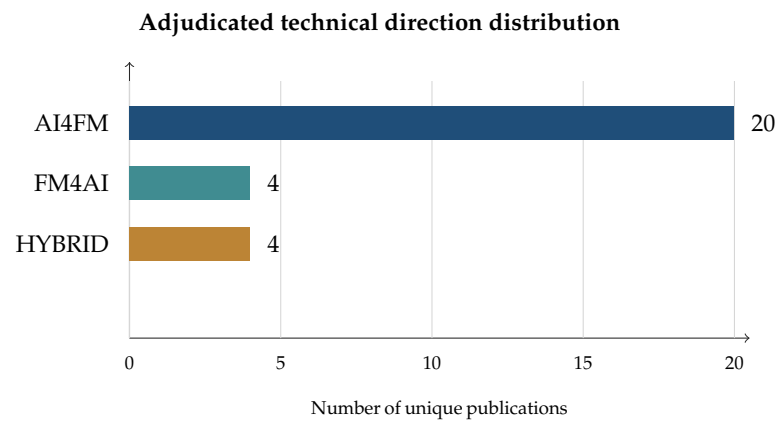


Figure 3. Adjudicated technical direction distribution for the 28 unique publications.

The adjudicated distribution shows that the largest group is *AI4FM*, with 20 of the 28 unique publications. This is followed by four *FM4AI* publications and four *HYBRID* publications. This indicates that the returned literature is weighted toward the use of AI techniques to support formal methods. This distribution is partly shaped by the progressive design of the three research questions, since the later questions focused more directly on GenAI support for formal methods. However, it also reflects the broader character of the returned literature: many recent publications frame AI not only as an object of formal analysis, but also as a means of supporting formal reasoning, modeling, proof search, and specification-related tasks.

4.4. *AI4FM* Technique Distribution

Figure 4 gives a further breakdown of the 20 adjudicated *AI4FM* publications. As specified in the coding scheme, this further technique analysis is only applied to *AI4FM* records.

Within the adjudicated *AI4FM* group, theorem proving is the dominant technique, accounting for nine of the 20 publications. Autoformalization is the next most common category, with three publications. Specification synthesis, model checking, and formal verification each appear in two publications, while dataset or benchmark construction and tool or system development each occur once. This pattern is consistent with recent interest in Large Language Models for proof generation, proof assistant interaction, and the translation of informal mathematical or system descriptions into formal representations.

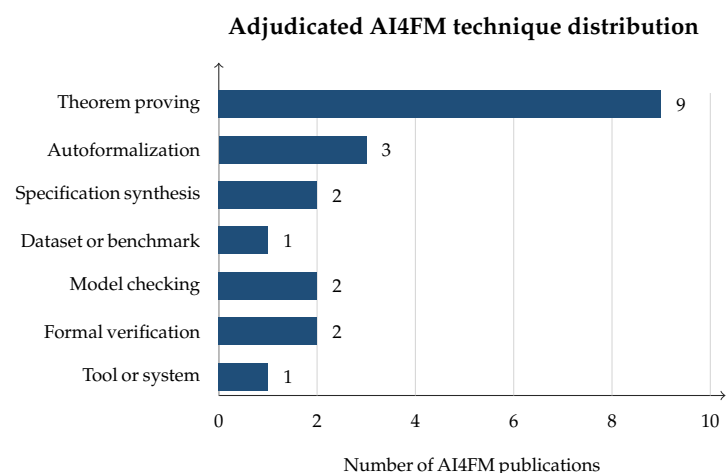


Figure 4. Adjudicated technique distribution within the *AI4FM* category.

4.5. Emerging Trends and Underrepresented Areas

Taken together, the search-output characteristics and the adjudicated coding distribution suggest a coherent picture. The returned literature is recent, often appears as conference papers or preprints, and is concentrated around AI-assisted support for formal methods. Within *AI4FM*, the most visible technical areas are theorem proving, autoformalization, specification synthesis, model checking, and formal verification.

The adjudicated distribution also reveals notable absences in the returned set. Although formal methods include a broad range of techniques, including abstract interpretation, model-based testing, runtime verification, refinement-based development, and industrial formal specification, these areas are not prominent in the returned *AI4FM* results. Model checking appears in only two adjudicated *AI4FM* publications, and several traditional branches of formal methods do not appear as major categories at all. Instead, the current GenAI–formal methods intersection produced by Google Scholar Labs appears to concentrate on language-facing formal tasks, especially theorem proving, proof assistant interaction, and autoformalization, rather than being evenly distributed across the full spectrum of formal methods.

This pattern should be interpreted cautiously. It should not be read as evidence that these other branches are inactive in the wider research literature. Rather, it is a feature of the returned set and of the particular AI-assisted search process studied here. It may reflect the search tool and query design, the citation-weighted ranking behavior of Google Scholar Labs, and the current visibility of AI-related work in those formal-methods subfields.

The adjudication process is therefore not merely a technical cleanup step. It also helps identify conceptual boundary cases in the field. The resolved disagreements show that some papers do not sit cleanly in a single direction of influence: they may use formal methods to analyze AI systems, use AI to scale or guide formal verification, and contribute tool support at the same time. These cases are particularly useful for refining the interpretation of *AI4FM*, *FM4AI*, and *HYBRID*.

4.6. Lightweight Search-Result Reproducibility

Because Google Scholar Labs is an AI-assisted search interface, its returned results should not be assumed to be fully stable over time. To examine this issue without turning the paper into a full systematic review, we treated repeated query rounds as a lightweight reproducibility check. The same three research questions were submitted again, and the top ten returned records for each question were compared with the initial search round reported in Section 3. Round 2, returned on 22 May 2026, and Round 3, returned on 28 May 2026, each produced 30 query-result records and 28 unique publications after deduplication.

Using normalized publication titles, the three query rounds produced 60 unique publications in total. A total of 5 publications appeared in all three rounds, 14 appeared in exactly two rounds, and 41 appeared in only one round. The pairwise overlap was highest between Rounds 1 and 2, with 12 shared unique publications; Rounds 1 and 3 shared 7, and Rounds 2 and 3 shared 10. The publications appearing in all three rounds include work on autoformalization, proof automation, generative language modeling for theorem proving, Goedel-Prover, and GenAI-augmented induction-based formal verification. Figure 5 summarizes this three-round overlap.

The three-round comparison shows that the repeated searches surface a small stable core of publications, but also a large set of publications that appear in only one round. The methodological implications of this variation are discussed in Section 5.

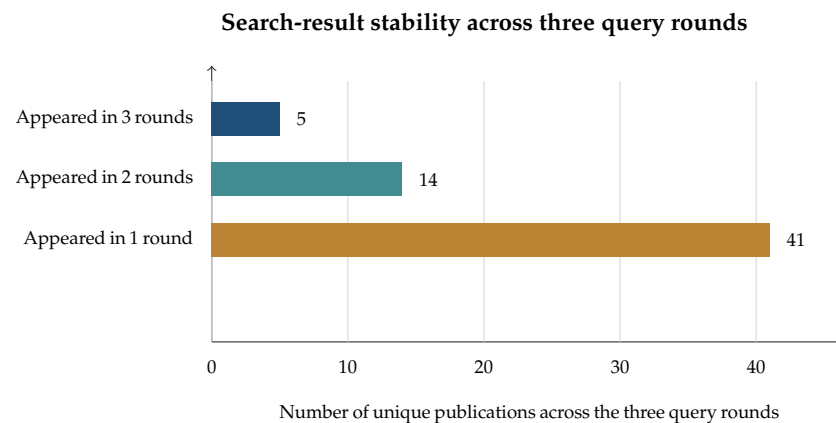


Figure 5. Three-round overlap summary for repeated Google Scholar Labs searches, based on normalized publication titles.

5. Discussion

5.1. The GenAI Survey

The use of Google Scholar Labs to answer the research questions in Section 3 provides some useful results, but these still need to be interpreted by an expert in the field. The main benefit is speeding up the process of surveying existing literature. However, some results may be less relevant, as determined by detailed inspection of the results. Research questions can be refined very quickly by assessing the results and then making a more precise enquiry depending on the issues found in the returned publications. In this paper, this has been done with three research questions, but the process could easily be continued at speed and interactively until enough relevant results are obtained.

The study as it stands relies on the top ten Google Scholar Labs results for each research question. It should be noted that this is an experiment, and the paper would be over-long if more than this was included. In any case, further results tend to have fewer citations and are thus of less relevance. Since Google Scholar Labs uses a proprietary ranking mechanism, search results may change over time, affecting reproducibility. In any case, GenAI uses a statistical approach including probabilistic aspects. Thus, even posing the same research question again immediately is likely to give different results, at least in the wording of the summaries. The selected publications are less likely to change since this is more likely to be a standard Google Scholar selection. Naturally, Google Scholar Labs may change at any time, and this is not under the control of the users. In addition, and probably more importantly, this is a fast-moving research area. Over time, there will be new papers, and the number of citations to papers will increase.

We have used citation counts to help in evaluating the importance of publications. Citation counts are (we believe it can be assumed from the results, in a similar way to the PageRank algorithm [63] for web pages with standard Google searches) used by Google Scholar Labs (and Google Scholar itself) to aid in the sorting of returned results. Certainly, this is a good indication of the importance of a research publication, just as links to a web page are an indication of the importance of that web page, as used in the PankRank algorithm. Of course, as will links to web pages, recent publications will have less time to accumulate citations. This is where the judgement of researchers using such tools is important. GenAI is useful as an aid in research; it is not a replacement for human expertise.

In this paper, we have introduced the categories *AI4FM*, *FM4AI*, and *HYBRID* for different types of publication. The judgement was made by the two authors independently initially, using their expertise in this research area. Differences in judgement were then assessed, and a final result was agreed upon by the two authors where needed. Of course, there is naturally some subjectivity involved, especially for borderline cases. A similar issue

of subjectivity arises in assessing the impact and relevance of cited publications. This is inevitable, but they are based on the expertise of the authors in the field of formal methods, with four decades of relevant experience in the case of the first author.

A major point of the paper is to demonstrate the use of Google Scholar Labs for literature exploration and summarization, based on an area in which the authors have some knowledge, so that we can assess its effectiveness against that of human evaluation. From our experience, we believe that the approach, combined with expert human judgement, has the potential to be applied in any (mainly scientific) field where Google Scholar has a large enough corpus of publications and where the importance of publications can be assessed to some extent by the citation counts (taking into account the age of the publication too, of course). This applies to many scientific fields, but is less relevant in the humanities, where books and reviews of books can be far more important.

Naturally, the AI-generated summaries cannot be taken at face value since GenAI has the possibility of “hallucinating” [64]. Thus, manual checking is necessary by an expert in the field. This is still very important with GenAI. The technology is improving rapidly, but the authors believe that GenAI output results should always be checked by human researchers, and never used without such verification. In this experimental study, we have only analyzed highly ranked search results. In real use of GenAI for aiding literature surveys, it would be expected to examine the results more deeply. In particular, emerging or less-cited research areas may need more consideration in practice and this is a matter of expert academic judgement. Of course, the results are a snapshot of a moving target. Re-examining results a year later is very likely to look different in fast-moving fields like the one considered in this paper.

The inclusion and exclusion criteria used to select publications from the search results was entirely done automatically by Google Scholar Labs in this experimental study. The inclusion criterion in this study was simply the top ten results for each research question. Of course, in a real literature survey, there should be more human selection using academic judgement. That said, the higher results are likely to be more important in general. It will be a matter of human decision-making as to how far to continue the search.

A notable feature of the publications in the survey is the number of papers available through the *arXiv* online archive. This allows rapid publication by researchers, with checking for appropriateness, but not formal peer reviewing. The developments in AI are so fast that authors appear to prefer quick publication of their latest research. Even many of the papers that have been formally published also have versions that are freely available on *arXiv*. Additionally, of course, articles on *arXiv* can be peer-reviewed and published subsequently (e.g., as is the case for *publication 3.10*). Without peer reviewing, extra caution should be applied in consideration of publications on *arXiv*. However, if they do have high citation counts, that is a form of approval by peer researchers. It is difficult to fake that to a great enough extent in the case of highly cited publications, which are, of course, the ones of more interest in general anyway.

Although there is some overlap of results returned by the three research questions, this is remarkably low, and has been commented on in the results where it did occur. Whether this low overlap is a good or bad aspect is difficult to determine, but perhaps makes the results from multiple related research questions more useful by avoiding repetition.

A useful feature of Google Scholar is the number of citations by other publications in the Google Scholar database. This helps to indicate the impact of a publication, bearing in mind the number of years since it appeared. This has been commented on for many of the answers in the results. Even some of the recent publications are already highly cited, indicating their rapid influence. Older papers, even if well cited, are less likely to be relevant for the research area under investigation.

Google Scholar has the ability to generate BibTeX-format records for citations, convenient for use with the L^AT_EX document preparation system, widely used by scientists in writing papers, including the present paper. However, the BibTeX records are not very complete, missing out book series information, publisher locations, etc. Most importantly, DOIs (Digital Object Identifiers) are not included for publications. These provide reliable and immutable online access to publications, unlike URLs (Uniform Resource Locators), which can vary at any time and may be short-lived. Therefore, although this paper used the Google Scholar generation of BibTeX records for the surveyed publications initially, significant manual effort was needed to add further information, especially DOIs where available.

5.2. Search-Result Stability

The repeated-query comparison in Section 4 suggests partial stability rather than exact reproducibility. Across the three query rounds, 90 returned records corresponded to 60 unique publications. Only five publications appeared in all three rounds, while 14 appeared in exactly two rounds, and 41 appeared in only one round. This distribution indicates that Google Scholar Labs does not behave like a fixed database query whose top results can simply be rerun and expected to reappear in the same form.

The overlap is nevertheless meaningful. The publications that appeared in all three rounds are concentrated around central *AI4FM* themes: autoformalization, proof automation, generative language modeling for theorem proving, Goedel-Prover, and GenAI-augmented formal verification. These repeatedly returned papers can be treated as a stable core of the AI-assisted search output. By contrast, the one-round publications broaden the map, often pointing to adjacent or more specialized topics such as formal specification generation, requirements formalization, theorem-generation benchmarks, hardware verification, and formal verification of AI-generated artifacts.

This pattern has methodological consequences. For an AI-assisted scholarly search tool, one query round is useful for rapid exploration, but should not be treated as a fully reproducible evidence-based. Repeated querying provides a more informative picture: publications that recur across rounds help identify the stable centre of the returned literature, while publications that appear in only one round help reveal the breadth and instability of the search space. In this sense, the variation is not merely a limitation; it is also a diagnostic feature of an exploratory AI-assisted mapping process. The results should therefore be interpreted as a structured snapshot of a fast-moving field rather than as the output of a conventional exhaustive literature search.

For completeness, the three rounds of interrogating Google Scholar Labs were undertaken on 21 February, 22 May, and 28 May 2026. Exactly the same three research questions were supplied, in order, one after another. That said, precise replication would be difficult. Reproducibility cannot be guaranteed with possible changes in Google Scholar Labs outside the control of users, new publications being added to the Google Scholar corpus regularly (most likely almost daily), and new citations being added to the corpus too. In addition, GenAI is inherently probabilistic, meaning that the resulting summaries are likely to vary in their precise wording, even if rerun immediately with the same question. That said, reproducibility in the traditional sense is less of an issue with this type of survey, which is essentially a snapshot of the current state of the art, which will always need some interpretation by a human researcher to help ensure its validity.

5.3. Future Directions

As we look to the future (2026 and beyond), the landscape of formal methods has the potential to converge with the recent rapid advancements in Artificial Intelligence (AI). Traditionally, there has been a dichotomy:

- **Formal Methods:** Rigorous, “correct by construction”, based on precise language and proof assistants.
- **Machine Learning/AI:** Based on large learned models, statistical probability, and often “unreliable” or “hallucinating” outputs [64].

However, we are seeing the emergence of hybrid approaches. Just as deep neural networks have revolutionized pattern recognition, AI is now being applied to mathematical discovery and program proof [65,66].

Recent research highlights the potential for AI to guide search in theorem proving. For example, systems like *Lean* (a proof assistant and programming language) are being integrated with AI tools to suggest outline proofs, which are then rigorously checked by the formal system [67,68]. This addresses the “unreliability” of LLMs by using the formal system as a verifier of the AI’s “intuition”.

New concepts have emerged and are likely to develop further. “Vibe coding” is a new software development approach with the assistance of AI in which the software developer provides a desired programming project or task in a natural language prompt to a GenAI system with the aim of generating the required source code automatically. The term was the Collins Dictionary word of the year in 2025 [69]. Despite its informal introduction, it has rapidly been adopted in scientific literature as well (e.g., see [70]). The equivalent term when applying AI to the generation of formal program proofs, or indeed using AI to aid mathematics in general, could be “vibe proving”.

Publications in *Nature* regarding mathematical discoveries from program search [71] and the ideas of the Fields Medalist Terence Tao [66] suggest we are entering a new and exciting era where AI fills in the steps of a proof, easing the use of formal verification in an interactive approach to proving both mathematics in general and programs in particular. The Leiden Declaration on Artificial Intelligence and Mathematics (<https://leidendeclaration.ai> (accessed on 2 June 2026)), issued in 2026 [72], concerning the reliance on AI-generated proofs, may be relevant to proofs in the application of formal methods within software engineering as well. More generally, will AI help with a scientific renaissance or restrict the research areas that will be investigated in the future [73]?

Just as AI experienced “winters” before its current boom, formal methods have seen cycles of hype and disillusionment. The integration with GenAI may trigger a “Formal Methods Spring”. Tools like *DeepSeekMath* [74] and *LeanProgress* [67] indicate that the barrier to entry for formal proofs is lowering. We may soon see a unification where engineers calculate and prove properties of software with the same ease that civil engineers calculate loads, aided by intelligent assistants.

6. Conclusions

This paper, while not providing a systematic review of formal methods and Artificial Intelligence (AI), has demonstrated an AI-assisted search process to aid in producing a literature survey in a fast-moving research area, specifically using the newly established Google Scholar Labs facility. It is a fast-moving area, as illustrated in the variation of results even over a short period of time. The authors hope that the approach may be applicable for other researchers in their specific fields. Compared with established systematic literature review (SLR) guidelines, the AI-assisted review methodology presented in this paper has a major advantage in the speed of undertaking such reviews. An important limitation is the danger of GenAI hallucinations without some human checking.

With regard to the specific field studied in this paper, a significant issue when applying formal methods in practice is the upfront cost of doing so. Using GenAI to aid in the use of formal methods has the potential to reduce the time and cost when they are used for software development. Extending the idea of “vibe coding” to “vibe proving” could well be

a fruitful area of research exploration. Developments in the field of mathematics in general when using GenAI to aid in proofs look promising for formal methods in the future,

Formal methods have traveled a long path from Alan Turing’s manual checking of routines to industrial-scale use, including semi and even fully automated verification. Specifically, the authors believe that a likely future lies in the symbiotic relationship between formal rigor and AI. By leveraging AI to handle the tedious aspects of proof and specification, while retaining the mathematical certainty of formal methods for verification, it should be possible to achieve a level of software reliability previously thought impossible, and in a more efficient manner as well. Software has often been produced in an artisan manner in the past. Formal methods, with the help of AI, have the potential to provide a more professional and cost-effective approach to software engineering in the future.

7. Survey Publications

Note that some of the publications in the references have been generated by Google Scholar Labs and some are normal references included by the authors. The two were combined into a single set of references during the finalization of the paper to follow the style of *Applied Sciences*.

Author Contributions: Conceptualization, J.P.B.; Methodology, J.P.B. and S.C.; Validation, S.C.; Formal analysis, S.C.; Investigation, J.P.B. and S.C.; Resources, J.P.B.; Data curation, J.P.B. and S.C.; Writing—original draft, J.P.B. and S.C.; Writing—review & editing, J.P.B. and S.C.; Supervision, J.P.B.; Project administration, J.P.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wing, J.M. A Specifier’s Introduction to Formal Methods. *Computer* **1990**, *23*, 8–23. <https://doi.org/10.1109/2.58215>.
2. Woodcock, J.; Larsen, P.G.; Bicarregui, J.; Fitzgerald, J. Formal methods: Practice and experience. *ACM Comput. Surv. (CSUR)* **2009**, *41*, 1–36. <https://doi.org/10.1145/1592434.1592436>.
3. Turing, A.M. Checking a Large Routine. In *Report of a Conference on High Speed Automatic Calculating Machines*; Cambridge University Mathematical Laboratory: Cambridge, UK, 1949; pp. 67–69. Available online: <https://turingarchive.kings.cam.ac.uk/checking-large-routine> (accessed on 2 June 2026).
4. Morris, F.L.; Jones, C.B. An Early Program Proof by Alan Turing. *IEEE Ann. Hist. Comput.* **1984**, *6*, 139–143. <https://doi.org/10.1109/MAHC.1984.10017>.
5. Turing, A.M. Computing Machinery and Intelligence. *Mind* **1952**, *LIX*, 433–460. <https://doi.org/10.1093/mind/LIX.236.433>.
6. Google. Google Scholar Labs. Available online: https://scholar.google.com/scholar_labs/search (accessed on 25 February 2026).
7. Bowen, J.P. Formal Methods: Whence and Whither? In *Rigorous Methods in Theory and Practice*; Raschke, A., Riccobene, E., Schewe, K-D., Thalheim, B., Eds.; LNCS; Springer: Cham, Switzerland, 2026; Volume 16580, pp. 16–29. https://doi.org/10.1007/978-3-032-24494-9_2.
8. Bowen, J.P. The Z Notation: Whence the Cause and Whither the Course? In *Engineering Trustworthy Software Systems, SETSS 2014*; Liu, Z., Zhang, Z., Eds.; LNCS; Springer: Cham, Switzerland, 2016; Volume 9506, pp. 103–151. https://doi.org/10.1007/978-3-319-29628-9_3.
9. Floyd, R.W. Assigning meaning to programs. In *Symposia in Applied Mathematics*; Scharfs, S.T., Ed.; Mathematical Aspects of Computer Science; American Mathematical Society: Providence, RI, USA, 1967. https://doi.org/10.1007/978-94-011-1793-7_4.
10. Hoare, C.A.R. An axiomatic basis for computer programming. *Commun. ACM* **1969**, *12*, 576–580. <https://doi.org/10.1145/363235.363259>.

11. Hoare, C.A.R.; He, J. *Unifying Theories of Programming*; Prentice Hall, Series in Computer Science: Upper Saddle River, NJ, USA, 1998.
12. Beth, E.W. *Formal Methods: An Introduction to Symbolic Logic and to the Study of Effective Operations in Arithmetic and Logic*; D. Reidel Publishing Company: Dordrecht, The Netherlands, 1962.
13. Strachey, C. The Strachey's Philosophy. Department of Computer Science, University of Oxford. Available online: <https://www.cs.ox.ac.uk/activities/concurrency/courses/stracheys.html> (accessed on 2 March 2025).
14. Hall, J.A. Seven myths of formal methods. *IEEE Softw.* **1990**, *7*, 11–19. <https://doi.org/10.1109/52.57887>.
15. Bowen, J.P.; Hinchey, M.G. Seven more myths of formal methods. *IEEE Softw.* **1995**, *12*, 34–41. <https://doi.org/10.1109/52.391826>.
16. Bowen, J.P.; Hinchey, M.G. Ten commandments of formal methods. *IEEE Comput.* **1995**, *28*, 56–63. <https://doi.org/10.1109/2.375178>.
17. Hinchey, M.G.; Bowen, J.P. (Eds.) *Applications of Formal Methods*. In *Prentice Hall, Series in Computer Science*; Prentice Hall: Hoboken, NJ, USA, 1995.
18. Hinchey, M.G.; Bowen, J.P. (Eds.) *Industrial-Strength Formal Methods in Practice*; Springer: Berlin/Heidelberg, Germany; FACIT: Toronto, ON, Canada, 1999.
19. Frappier, M.; Habrias, H. (Eds.) *Software Specification Methods: An Overview Using a Case Study*; Springer: Berlin/Heidelberg, Germany; FACIT: Toronto, ON, Canada, 2001.
20. Boulanger, J.-L. (Ed.) *Formal Methods: Industrial Use from Model to the Code*; ISTE: Arlington, Virginia, USA; Wiley: Hoboken, NJ, USA, 2012.
21. Gnesi, S.; Margaria, T. *Formal Methods for Industrial Critical Systems: A Survey of Applications*; IEEE Computer Society Press: New York, NY, USA; Wiley: Hoboken, NJ, USA, 2012.
22. Bowen, J.P.; Stavridou, V. Safety-critical systems, formal methods and standards. *Softw. Eng. J.* **1993**, *8*, 189–209. <https://doi.org/10.1049/sej.1993.0025>.
23. Liu, Z.; Woodcock, J.; Bowen, J.P. From Logic to Programming. In *Engineering Trustworthy Software Systems, SETSS 2024*; Bowen, J.P., Gomes, C., Liu, Z., Eds.; LNCS; Springer: Singapore, 2025; Volume 15584, pp. 7–67. https://doi.org/10.1007/978-981-96-4656-2_2.
24. Brereton, P.; Kitchenham, B.A.; Budgen, D.; Turner, M.; Khalil, M. Lessons from applying the systematic literature review process within the software engineering domain. *J. Syst. Softw.* **2007**, *80*, 571–583. <https://doi.org/10.1016/j.jss.2006.07.009>.
25. Dieste, O.; Grimán, A.; Juristo, N. Developing search strategies for detecting relevant experiments. *Empir. Softw. Eng.* **2009**, *14*, 513–539. <https://doi.org/10.1007/s10664-008-9091-7>.
26. Petersen, K.; Feldt, R.; Mujtaba, S.; Mattsson, M. Systematic mapping studies in software engineering In *EASE'08: Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering*; ACM Digital Library: New York, NY, USA, 2008; pp. 68–77.
27. Petersen, K.; Vakkalanka, S.; Kuzniarz, L. Guidelines for conducting systematic mapping studies in software. *Inf. Softw. Technol.* **2015**, *64*, 1–18. <https://doi.org/10.1016/j.infsof.2015.03.007>.
28. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. <https://doi.org/10.1177/001316446002000104>.
29. de Moura, L.; Kong, S.; Avigad, J.; van Doorn, F.; von Raumer, J. The Lean Theorem Prover (System Description). In *Automated Deduction—CADE-25*; Felty, A., Middeldorp, A., Eds.; LNCS; Springer: Cham, Switzerland, 2015; Volume 9195, pp. 378–388. https://doi.org/10.1007/978-3-319-21401-6_26.
30. Ramsay, A. *Formal Methods in Artificial Intelligence*; Cambridge University Press: Cambridge, UK, 1988; Volume 6.
31. Krichen, M.; Mihoub, A.; Alzahrani, M.Y.; Adoni, W.Y.H.; Nahhal, T. Are formal methods applicable to machine learning and Artificial Intelligence? In *2022 2nd International Conference of Smart Systems and Emerging Technologies (SMARTTECH)*; IEEE: New York, NY, USA, 2022; pp. 48–53. <https://doi.org/10.1109/SMARTTECH54121.2022.00025>.
32. Szegedy, C. A promising path towards autoformalization and General Artificial Intelligence. In *International Conference on Intelligent Computer Mathematics*, LNCS; Springer: Cham, Switzerland, 2020; Volume 12236, pp. 3–20. https://doi.org/10.1007/978-3-030-53518-6_1.
33. Wang, H. Computer theorem proving and Artificial Intelligence. In *Computation, Logic, Philosophy: A Collection of Essays; Mathematics and its Application (China Series)*; Springer: Dordrecht, The Netherlands, 1990; Volume 2, pp. 63–75. https://doi.org/10.1007/978-94-009-2356-0_5.
34. Whitehead, A.N.; Russell, B. *Principia Mathematica*; Cambridge University Press: Cambridge, UK, 1910–1913.
35. Cai, C.H.; Sun, J.; Dobbie, G. Automatic B-model repair using model checking and machine learning. *Autom. Softw. Eng.* **2019**, *26*, 653–704. <https://doi.org/10.1007/s10515-019-00264-4>.
36. Abrial, J.-R. *The B-Book: Assigning Programs to Meanings*; Cambridge University Press: Cambridge, UK, 1996.
37. Sun, X.; Khedr, H.; Shoukry, Y. Formal verification of neural network controlled autonomous systems. In *22nd ACM International Conference on Hybrid Systems: Computation and Control*; ACM Digital Library: New York, NY, USA, 2019; pp. 147–156. <https://doi.org/10.1145/3302504.331180>.

38. Zhang, Y.; Cai, Y.; Zuo, X.; Luan, X.; Wang, K.; Hou, Z.; Zhang, Y.; Wei, Z.; Sun, M.; Sun, J.; et al. The fusion of Large Language Models and formal methods for trustworthy AI agents: A roadmap. *arXiv* **2024**, arXiv:2412.06512. <https://doi.org/10.48550/arXiv.2412.06512>.
39. Bauer, A.; Petković, M.; Todorovski, L. MLFMF: Data sets for machine learning for mathematical formalization. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 50730–50741. <https://doi.org/10.48550/arXiv.2310.16005>.
40. Bortolussi, L.; Milios, D.; Sanguinetti, G. Machine learning methods in statistical model checking and system design—Tutorial. In *Runtime Verification*; Bartocci, E., Majumdar, R., Eds.; LNCS; Springer: Cham, Switzerland, 2015; Volume 9333, pp. 323–341. https://doi.org/10.1007/978-3-319-23820-3_23.
41. Kuznetsov, S.O. Machine learning and formal concept analysis. In *International Conference on Formal Concept Analysis*; Eklund, P., Ed.; LNCS; Springer: Cham, Switzerland, 2004; Volume 2961, pp. 287–312. https://doi.org/10.1007/978-3-540-24651-0_25.
42. Jha, S.; Jha, S.K.; Lincoln, P.; Bastian, N.D.; Velasquez, A.; Neema, S. Dehallucinating Large Language Models using formal methods guided iterative prompting. In *2023 IEEE International Conference on Assured Autonomy (ICAA)*; IEEE: New York, NY, USA, 2023; pp. 149–152. <https://doi.org/10.1109/ICAA58325.2023.00029>.
43. Polu, S.; Sutskever, I. Generative language modeling for automated theorem proving. *arXiv* **2020**, arXiv:2009.03393. <https://doi.org/10.48550/arXiv.2009.03393>.
44. Lin, Y.; Tang, S.; Lyu, B.; Wu, J.; Lin, H.; Yang, K.; Li, J.; Xia, M.; Chen, D.; Arora, S.; et al. Goedel-Prover: A frontier model for open-source automated theorem proving. *arXiv* **2025**, arXiv:2502.07640. <https://doi.org/10.48550/arXiv.2502.07640>.
45. Lin, Y.; Tang, S.; Lyu, B.; Yang, Z.; Chung, J.H.; Zhao, H.; Jiang, L.; Geng, Y.; Ge, J.; Sun, J.; et al. Goedel-Prover-V2: Scaling formal theorem proving with scaffolded data synthesis and self-correction. *arXiv* **2025**, arXiv:2508.03613. <https://doi.org/10.48550/arXiv.2508.03613>.
46. Xin, H.; Ren, Z.; Song, J.; Shao, Z.; Zhao, W.; Wang, H.; Liu, B.; Zhang, L.; Lu, X.; Du, Q.; et al. DeepSeek-Prover-V1.5: Harnessing proof assistant feedback for reinforcement learning and Monte-Carlo tree search. *arXiv* **2024**, arXiv:2408.08152. <https://doi.org/10.48550/arXiv.2408.08152>.
47. Xin, H.; Guo, D.; Shao, Z.; Ren, Z.; Zhu, Q.; Liu, B.; Ruan, C.; Li, W.; Liang, X. DeepSeek-Prover: Advancing theorem proving in LLMs through large-scale synthetic data. *arXiv* **2024**, arXiv:2405.14333. <https://doi.org/10.48550/arXiv.2405.14333>.
48. Katz, S.M.; Corso, A.L.; Strong, C.A.; Kochenderfer, M.J. Verification of image-based neural network controllers using generative models. *J. Aerosp. Inf. Syst.* **2022**, *19*, 574–584. <https://doi.org/10.2514/1.1011071>.
49. Wang, H.; Unsal, M.; Lin, X.; Baksys, M.; Liu, J.; Santos, M.D.; Sung, F.; Vinyes, M.; Ying, Z.; Zhu, Z.; et al. Kimina-Prover preview: Towards large formal reasoning models with reinforcement learning. *arXiv* **2025**, arXiv:2504.11354. <https://doi.org/10.48550/arXiv.2504.11354>.
50. Carreira-Munich, T.; Paz-Marcolla, V.; Castro, R. DEVS Copilot: Towards Generative AI-Assisted Formal Simulation Modeling based on Large Language Models. In *2024 Winter Simulation Conference (WSC)*; IEEE: New York, NY, USA, 2024; pp. 2785–2796. <https://doi.org/10.1109/WSC63780.2024.10838994>.
51. Tihanyi, N.; Charalambous, Y.; Jain, R.; Ferrag, M.A.; Cordeiro, L.C. A new era in software security: Towards self-healing software via Large Language Models and formal verification. In *2025 IEEE/ACM International Conference on Automation of Software Test (AST)*; IEEE: New York, NY, USA, 2025; pp. 136–147. <https://doi.org/10.1109/AST66626.2025.00020>.
52. Clarke, E.; Biere, A.; Raimi, R.; Zhu, Y. Bounded Model Checking Using Satisfiability Solving. *Form. Methods Syst. Des.* **2001**, *19*, 7–34. <https://doi.org/10.1023/A:1011276507260>.
53. Tihanyi, N.; Bisztray, T.; Jain, R.; Ferrag, M.A.; Cordeiro, L.C.; Mavroeidis, V. The FormAI dataset: Generative AI in software security through the lens of formal verification. In *PROMISE 2023: Proceedings of the 19th International Conference on Predictive Models and Data Analytics in Software Engineering*; ACM Digital Library: New York, NY, USA, 2023; pp. 33–43. <https://doi.org/10.1145/3617555.3617874>.
54. Lu, M.; Delaware, B.; Zhang, T. Proof automation with Large Language Models. In *ASE'24: Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*; IEEE/ACM Digital Library: New York, NY, USA, 2024; pp. 1509–1520. <https://doi.org/10.1145/3691620.3695521>.
55. Wu, Y.; Jiang, A.Q.; Li, W.; Rabe, M.; Staats, C.; Jamnik, M.; Szegedy, C. Autoformalization with Large Language Models. In *NIPS'22: Proceedings of the 36th International Conference on Advances in Neural Information Processing Systems*; Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A., Eds.; Curran Associates: Red Hook, NY, USA, 2022; Volume 35, pp. 32353–32368. Available online: https://proceedings.neurips.cc/paper_files/paper/2022/hash/d0c6bc641a56bebee9d985b937307367-Abstract-Conference.html (accessed on 2 June 2026).
56. Nipkow, T.; Wenzel, M.; Paulson, L.C. (Eds.) *Isabelle/HOL: A Proof Assistant for Higher-Order Logic*; LNCS; Springer: Cham, Switzerland, 2002; Volume 2283. <https://doi.org/10.1007/3-540-45949-9>.
57. Shminke, B. Applications of AI to Study of Finite Algebraic Structures and Automated Theorem Proving. Ph.D. Thesis, Université Côte d’Azur, Nice, France, 2023.

58. Granberry, G.; Ahrendt, W.; Johansson, M. Specify what? Enhancing neural specification synthesis by symbolic methods. In *Proceedings of the Integrated Formal Methods (IFM 2024)*; Kosmatov, N., Kovács, L., Eds.; LNCS; Springer: Cham, Switzerland, 2024; Volume 15234, pp. 307–325. https://doi.org/10.1007/978-3-031-76554-4_19.
59. Kumar, A.; Gadde, D.N. Generative AI augmented induction-based formal verification. In *2024 IEEE 37th International System-on-Chip Conference (SOCC)*; IEEE: New York, NY, USA, 2024; pp. 1–2. <https://doi.org/10.1109/SOCC62300.2024.10737803>.
60. Weng, K.; Du, L.; Li, S.; Lu, W.; Sun, H.; Liu, H.; Zhang, T. Autoformalization in the Era of Large Language Models: A Survey. *arXiv* **2025**, arXiv:2505.23486. <https://doi.org/10.48550/arXiv.2505.23486>.
61. Tihanyi, N.; Bisztray, T.; Ferrag, M.A.; Cherif, B.; Dubniczky, R.A.; Jain, R.; Cordeiro, L.C. Vulnerability detection: From formal verification to Large Language Models and hybrid approaches: A comprehensive overview. *arXiv* **2025**, arXiv:2503.10784. <https://doi.org/10.48550/arXiv.2503.10784>.
62. Tihanyi, N.; Bisztray, T.; Ferrag, M.A.; Cherif, B.; Dubniczky, R.A.; Jain, R.; Cordeiro, L.C. Vulnerability detection: From formal verification to Large Language Models and hybrid approaches: A comprehensive overview. In *Adversarial Example Detection and Mitigation Using Machine Learning*; Nowroozi, E., Taheri, R., Cordeiro, L., Eds.; Springer: Cham, Switzerland, 2026; Chapter 3, pp. 33–47. https://doi.org/10.1007/978-3-031-99447-0_3.
63. Langville, A.N.; Meyer, C.D. *Google's PageRank and Beyond: The Science of Search Engine Rankings*; Princeton University Press: Princeton, NJ, USA, 2006.
64. Farquhar, S.; Kossen, J.; Kuhn, L.; Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature* **2024**, *630*, 625–630. <https://doi.org/10.1038/s41586-024-07421-0>.
65. Avigad, J. Mathematics in the Age of AI. LMS/BCS-FACS Seminar, London Mathematical Society. 6 November 2025. Available online: <https://www.lms.ac.uk/events/lms-bcs-facs-seminar-jeremy-avigad> (accessed on 2 June 2026).
66. Tao, T. The Potential for AI in Science and Mathematics. Oxford Mathematics Public Lectures, Science Museum, London. 17 July 2024. Available online: <https://www.maths.ox.ac.uk/node/68243> (accessed on 2 June 2026).
67. George, R.J.; Huang, S.; Song, P.; An kumar, A. LeanProgress: Guiding Search for Neural Theorem Proving via Proof Progress Prediction. *arXiv* **2025**, arXiv.2502.17925. <https://doi.org/10.48550/arXiv.2502.17925>.
68. Lu, J.; Emond, K.; Yang, K.; Chaudhuri, S.; Sun, W.; Chen, W. Lean Finder: Semantic Search for Mathlib That Understands User Intents. *arXiv* **2025**, arXiv.2510.15940. <https://doi.org/10.48550/arXiv.2510.15940>.
69. Cress, L. 'Vibe Coding' Named Word of the Year by Collins Dictionary. BBC News, BBC. November 2025. Available online: <https://www.bbc.co.uk/news/articles/cpd2y053nleo> (accessed on 2 June 2026).
70. Fawzy, A.; Tahir, A.; Blincoe, K. Vibe Coding in Practice: Motivations, Challenges, and a Future Outlook—A Grey Literature Review. *arXiv* **2025**, arXiv.2510.00328. <https://doi.org/10.48550/arXiv.2510.00328>.
71. Romera-Paredes, B.; Barekatin, M.; Novikov, A.; Balog, M.; Kumar, M.P.; Dupont, E.; Ruiz, F.J.; Ellenberg, J.S.; Wang, P.; Fawzi, O.; et al. Mathematical discoveries from program search and large language models. *Nature* **2024**, *625*, 468–475. <https://doi.org/10.1038/s41586-023-06924-6>.
72. Nature Editorial. Mathematicians are developing rules for AI use—Other fields should follow. *Nature* **2026**, *654*, 571. <https://doi.org/10.1038/d41586-026-01881-2>.
73. Zhang, X. Will AI spark a scientific renaissance—Or a diffuse monoculture? *Nature* **2026**, *654*, 843. <https://doi.org/10.1038/d41586-026-01954-2>.
74. Shao, Z.; Luo, Y.; Lu, C.; Ren, Z.Z.; Hu, J.; Ye, T.; Gou, Z.; Ma, S.; Zhang, X. DeepSeekMath-V2: Towards Self-Verifiable Mathematical Reasoning. *arXiv*, **2025**, arXiv.2511.22570. <https://doi.org/10.48550/arXiv.2511.22570>.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.