DOCUMENT 3 : RESEARCH GAP

AIM²P — Identification des quatre verrous scientifiques non résolus

Phase : Phase 0 (cadrage scientifique) · Projet : AIM²P Initiative · Version : 0.2

1. LES QUATRE VERROUS SCIENTIFIQUES

L'audit bibliographique de clôture de la phase 0 (état au 17 novembre 2026) a montré qu'aucun des quatre verrous de la version 0.1 ne tenait tel quel : GAP 2 était faux tel que rédigé, et GAP 1, GAP 3 et GAP 4 devaient être nuancés. Les formulations ci-dessous sont plus étroites, datées, et valent à notre connaissance. Les sources sont détaillées dans le document 2 (State of the Art).

> **GAP 1 : la physique de niveau recherche reste hors des bibliothèques formelles**
>
> Les efforts de preuve formelle (Lean 4, Rocq, Isabelle) restent dominés par les mathématiques pures, les mathématiques de compétition et l'informatique théorique. La physique théorique y est un domaine minoritaire mais en croissance rapide : depuis 2024, la bibliothèque communautaire Physlib (leanprover-community ; issue de HepLean → PhysLean, puis fusionnée avec Lean-QuantumInfo) couvre la mécanique lagrangienne et hamiltonienne, la mécanique quantique, la relativité, les unités, la QFT et l'information quantique (environ 450 000 lignes en novembre 2026, dont plus de la moitié dans un niveau à relecture allégée). Les années 2025–2026 ont vu apparaître les premiers prouveurs et bancs d'essai LLM pour la physique en Lean (Ax-Prover, Lean4PHYS, PhysProver, FormalPhysics, AxQM), ainsi qu'une erreur non triviale dans un article de physique très cité, mise en évidence par formalisation (stabilité du modèle à deux doublets de Higgs, 2026). En revanche, à notre connaissance (novembre 2026), la complexité de Krylov, la récursion de Lanczos et ses coefficients, la croissance d'opérateurs et l'holographie n'ont aucun équivalent dans les bibliothèques formelles de référence (Mathlib, Physlib) ; nous n'avons identifié qu'un dépôt isolé et non relu consacré à la dynamique de Krylov en dimension finie. Les dérivations de niveau recherche, par opposition aux énoncés de manuel, restent largement non formalisées.

Ce dépôt, lanzarini-krylov-lean, est mono-auteur, assisté par IA et sans licence ouverte ; il formalise la complexité de Krylov pour un hamiltonien de Jacobi à quatre sites et déclare lui-même n'avoir pas établi de théorème général identifiant sa base à la base de Lanczos. Sources : Physlib ; Mathlib4 ; arXiv:2603.08139. Limite : le Zulip Lean et le Lean Reservoir n'ont pas été fouillés directement, et l'Archive of Formal Proofs n'a pas été interrogée par mot-clé.

> **GAP 2 : pas encore d'architecture ouverte et généraliste qui mesure ses propres critiques**
>
> Le couplage LLM–calcul formel (raisonnement intégrant des outils : ToRA, Qwen-Math TIR…) et le couplage LLM–Lean (Lean Copilot, Ax-Prover, Numina-Lean-Agent…) sont aujourd'hui matures. Il existe aussi des ponts calcul formel–Lean où le système de calcul formel sert d'oracle non fiable dont Lean vérifie les résultats (polyrith, LeanSage, LeanLink, tactiques à certificats pour le raisonnement polynomial). Quelques systèmes récents combinent les trois, mais pour des *domaines spécifiques* : inégalités olympiques (LIPS, ICLR 2025 : GPT-4o + tactiques fondées sur SymPy + Lean 4), inégalités polynomiales par sommes de carrés (NSPI, ICML 2026), géométrie euclidienne (MechGeo, 2026 : certificats Singular/SymPy vérifiés par le noyau de Lean). Des architectures plus générales sont apparues en 2026 : dans Moxia, le LLM se borne à reformuler l'énoncé pour un pipeline déterministe de calcul formel qui résout le problème, Lean certifiant les dérivations qui entrent dans un fragment de Mathlib ; SPL est un langage déclaratif qui orchestre appels au LLM et outils symboliques (SymPy, SageMath, Lean). À notre connaissance (novembre 2026), il n'existe pas encore d'architecture ouverte et généraliste dans laquelle (i) le calcul symbolique exact (SymPy) joue systématiquement le rôle de critique des étapes intermédiaires du LLM, (ii) Lean 4 certifie les résultats finaux et, lorsqu'un certificat vérifiable existe, les calculs symboliques eux-mêmes, (iii) la fiabilité du critique symbolique est mesurée plutôt que présumée, et (iv) les retours des deux outils pilotent l'autocorrection du LLM, avec une mesure par ablation de la contribution de chaque composant. AIM²P propose de construire et d'évaluer une telle architecture.

Les deux précédents les plus proches sont LIPS, le plus proche par la pile logicielle (arXiv:2502.13834), et MechGeo, le plus proche par le mécanisme (arXiv:2608.02295) ; NSPI (arXiv:2605.15445) et Aristotle (arXiv:2510.01346) complètent le tableau. Moxia (arXiv:2606.00671) est le précédent généraliste le plus proche, mais le calcul formel y remplace le raisonnement du LLM au lieu de le critiquer, et il n'y a pas de boucle d'auto-correction ; SPL (arXiv:2607.07727) fournit une couche d'orchestration, évaluée sur la correction vérifiée des réponses. À l'inverse, les grands agents Lean généralistes (Ax-Prover, Numina-Lean-Agent, AIProver) ne listent aucun outil de calcul formel. La version 0.1 affirmait qu'aucune architecture n'orchestrait LLM, SymPy et Lean 4 : cette affirmation est retirée.

> **GAP 3 : l'auto-correction guidée par Lean n'a pas encore rencontré le calcul formel en physique mathématique**
>
> L'auto-correction guidée par le compilateur est devenue le paradigme dominant de la preuve automatique en Lean 4 : révision de preuves complètes à partir des messages d'erreur (Goedel-Prover-V2, OProver), décomposition en lemmes ou en plan de preuve avec réparation itérative (Seed-Prover 1.5, Hilbert, Goedel-Architect), boucles agentiques d'outils LSP/MCP autour de LLM généralistes (AxProverBase, Numina-Lean-Agent, Ax-Prover, Aleph). Ces systèmes dépassent 99 % sur miniF2F et atteignent sur PutnamBench de 88,8 % (système fondé sur un modèle à poids ouverts) à 99,4 % (annonce d'un fournisseur, non reproduite). Des évaluations en physique existent depuis octobre 2025 (QuantumTheorems, LeanPhysBench, PhysProver, AxQM). À notre connaissance (novembre 2026), quatre manques subsistent néanmoins : (i) ces évaluations en physique sont récentes, de petite taille (134 à 1 019 tâches) et centrées sur l'information quantique et la physique de premier cycle, tandis que la physique mathématique à forte composante d'analyse (théorie spectrale, dynamique d'opérateurs) n'est pas couverte ; (ii) aucun pipeline publié ne couple un système de calcul formel à la réparation guidée par Lean pour des dérivations de physique mathématique ; (iii) les courbes de certification par lemme en fonction du nombre d'itérations sont rarement publiées, et l'effet du type d'erreur sur la réparabilité commence seulement à être mesuré ; (iv) les systèmes les plus performants sont fermés et coûteux, et la validité de leurs preuves exige un audit au niveau du noyau (`sorryAx` , altération de l'énoncé, erreurs de formalisation).

Sources : arXiv:2508.03613 ; arXiv:2605.17283 ; arXiv:2602.24273 ; arXiv:2510.12787 ; arXiv:2510.26094 ; arXiv:2609.05157 ; arXiv:2608.28639 ; arXiv:2610.26942. La version 0.1 décrivait l'auto-correction comme une « réévaluation à l'aveugle » et la disait évaluée sur des benchmarks de mathématiques pures : ces deux affirmations ne sont plus exactes pour Lean. La première reste vraie des LLM généralistes utilisés sans outils.

> **GAP 4 : le raisonnement formellement vérifié n'a pas encore de foyer en Afrique**
>
> À notre connaissance (état des lieux de novembre 2026), aucune plateforme ni aucun benchmark open source dédié au raisonnement mathématique ou scientifique *formellement vérifié* (assistants de preuve tels que Lean 4, Rocq ou Isabelle) n'est développé et maintenu depuis le continent africain, et aucun événement de la communauté Lean ne s'y est tenu. Des benchmarks ouverts portés par des équipes africaines existent en revanche pour le raisonnement *informel* en langues africaines, par exemple AfriMGSM (IrokoBench, NAACL 2025) et Uhura (questions scientifiques). Le continent dispose aussi de communautés solides de formation et de recherche en IA (AIMS et son master AI for Science, Deep Learning Indaba, Masakhane, Data Science Africa). AIM²P vise à relier ces forces à l'écosystème des méthodes formelles, où la participation africaine reste limitée.

Sources : événements de la communauté Lean ; IrokoBench, arXiv:2406.03368 ; Uhura, arXiv:2412.00948 ; AIMS AI for Science. La clause de la version 0.1 selon laquelle ce manque « limite la participation des chercheurs locaux » était une hypothèse causale non démontrée : elle est retirée. Limite : la géographie des contributeurs de Mathlib et du Zulip Lean n'a pas été mesurée.

Synthèse : ce qui tombe, ce qui reste

Verrou	Verdict de l'audit	Ce qui tombe	Ce qui reste défendable
GAP 1, physique formelle	À nuancer	« HepLean devenu PhysLean puis Physlib » (c'est une fusion) ; « centrées sur les hautes énergies et l'information quantique » ; « marginale » au sens de quasi absente	Krylov, Lanczos, croissance d'opérateurs et holographie absents de Mathlib et Physlib ; dérivations de niveau recherche peu formalisées
GAP 2, triade	Faux tel que rédigé	« LLM seul, LLM + Lean ou LLM + SymPy » ; « aucune architecture unifiée »	Pas d'architecture ouverte et généraliste où SymPy critique les étapes, où Lean certifie aussi les calculs symboliques et où des ablations mesurent chaque composant
GAP 3, auto-correction	À nuancer	« Réévaluation à l'aveugle » (pour Lean) ; « évaluées sur des benchmarks de mathématiques pures » ; DeepSeek-Prover-V1.5 comme référence la plus proche	Pas de couplage calcul formel + réparation Lean en physique mathématique ; pas de courbes de certification par lemme ; systèmes de pointe fermés
GAP 4, Afrique	À nuancer	« Raisonnement scientifique » au sens large ; la clause causale sur la participation des chercheurs locaux	Aucune plateforme ni benchmark de raisonnement *formellement vérifié* porté depuis l'Afrique ; aucun événement Lean en Afrique

2. POSITIONNEMENT DE L'INITIATIVE AIM²P

> **Positionnement**
>
> AIM²P ne revendique ni la première physique formelle, ni la première combinaison LLM + calcul formel + Lean, ni la première boucle d'auto-correction guidée par Lean. Le programme se place à l'intersection de quatre manques précis : une boucle ouverte et généraliste dont le critique symbolique est mesuré et dont les calculs sont certifiés quand c'est possible ; une évaluation de cette boucle sur la physique mathématique, à commencer par la dynamique d'opérateurs en dimension finie et la complexité de Krylov ; des courbes publiques de certification par lemme ; et un foyer africain pour le raisonnement formellement vérifié.

La contribution attendue est donc moins un score qu'un ensemble de **mesures**. À notre connaissance (novembre 2026), aucune des trois suivantes n'a été publiée :

- le taux d'erreur d'un critique SymPy, et la part de ces erreurs que Lean rattrape ;
- la courbe de certification par lemme en fonction du nombre d'itérations, sur des énoncés de physique mathématique ;
- le taux de détection d'erreurs physiques injectées par un vérificateur formel, comparé à un juge LLM et à SymPy.

Ces mesures sont réalisables sur un petit banc de 40 items et portent sur la fiabilité de la vérification plutôt que sur un score de concours. S'y ajoute un artefact modeste et durable : un premier noyau relu de Lanczos et de Krylov en dimension finie (tridiagonalité, terminaison, unitarité), à proposer en amont à PhyslibAlpha plutôt qu'à garder dans un dépôt isolé.

> **Avant toute publication**
>
> Ces verrous sont datés du 17 novembre 2026. Des préprints du domaine paraissent chaque semaine : une recherche arXiv finale (cs.AI, cs.LO, cs.SC) sur « SymPy + Lean », « Krylov + Lean » et « physics + Lean 4 » doit précéder toute publication de ce document.

HISTORIQUE DES VERSIONS

- **v0.1 :** version initiale de la phase 0.
- **v0.2 (17 novembre 2026) :** mise à jour après l'audit bibliographique de clôture de P0. GAP 2 reformulé (des systèmes LLM + calcul formel + Lean existent pour des domaines spécifiques) ; GAP 1, GAP 3 et GAP 4 nuancés et datés ; sources ajoutées ; positionnement réécrit autour de mesures précises plutôt que d'une revendication de première. Intègre la recherche arXiv de vérification finale du 17 novembre 2026 (notamment Moxia et SPL).