DOCUMENT 1 : RESEARCH QUESTION

AIM²P — Formulation de la question centrale et des hypothèses expérimentales

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

---

1. QUESTION SCIENTIFIQUE CENTRALE

La question pivot à laquelle répondra l'initiative AIM²P s'énonce comme suit :

> **Question principale**
>
> « Dans quelle mesure une boucle neuro-symbolique ouverte, où un LLM raisonne, SymPy calcule et critique, et Lean 4 certifie, augmente-t-elle, à budget de calcul égal, la part de solutions correctes et intégralement vérifiées, et réduit-elle la part d'affirmations de succès non vérifiées, par rapport au même LLM utilisé seul ou filtré par le même vérificateur, sur des problèmes de mathématiques et de physique mathématique (cas d'étude : dynamique d'opérateurs en dimension finie et complexité de Krylov) ? »

Une *affirmation de succès non vérifiée* est une réponse qui se déclare « résolue » ou « vérifiée » alors que le vérificateur la rejette. C'est la définition opérationnelle de l'hallucination retenue par AIM²P : elle se mesure en croisant la déclaration du système et le verdict du vérificateur.

> **Pourquoi la formulation de la version 0.1 a été abandonnée**
>
> La version 0.1 visait à « réduire le taux d'hallucination à zéro ». Cet objectif n'est ni atteignable ni mesurable. Le noyau de Lean ne certifie que l'énoncé formel, pas sa fidélité au problème posé ; les benchmarks Lean les plus utilisés contiennent eux-mêmes des énoncés défectueux ; et les LLM utilisés comme juges détectent mal les défauts sémantiques, avec une précision de 0,24 à 0,30 (arXiv:2606.29493). La question porte désormais sur des proportions mesurables, comparées à des baselines à budget égal.

2. SOUS-QUESTIONS DE RECHERCHE (SQ)

- **SQ1 (interface et fidélité) :** Comment passer d'un raisonnement informel (langage naturel enrichi, LaTeX) à des énoncés et à des tactiques Lean 4 sans dérive d'énoncé ? Dans quels cas un résultat de SymPy peut-il devenir un certificat vérifiable par Lean, plutôt qu'un simple indice ?
- **SQ2 (boucle de rétroaction et erreurs) :** Comment les messages d'erreur de Lean 4 et les verdicts de SymPy, réinjectés dans le contexte du LLM, guident-ils une boucle de révision (agentique, éventuellement arborescente) vers une preuve certifiée, et à quel coût par itération ?
- **SQ3 (physique mathématique et invariants) :** Comment formaliser les objets de la dynamique d'opérateurs en dimension finie (liouvillien, récursion de Lanczos, complexité de Krylov, hamiltoniens quantiques), qui combinent dérivation analytique et structures algébriques ? Quelles erreurs physiques un vérificateur formel détecte-t-il mieux qu'un juge LLM ou que SymPy ?

SQ2 n'impose plus de recherche arborescente de type MCTS : l'état de l'art est aujourd'hui dominé par des boucles agentiques qui réinjectent le texte des erreurs, et l'usage du compilateur comme simple récompense

scalaire expose au piratage de récompense (arXiv:2608.28639). La recherche arborescente reste une option parmi d'autres.

3. HYPOTHÈSES EXPÉRIMENTALES (H)

Les trois hypothèses sont formulées pour être falsifiables sur **AMP-Bench-mini v1.0** : 40 items de niveaux N1 à N5, gelés avec leur correcteur avant toute exécution. La composition du banc et le protocole complet sont fixés dans la Research Proposal (document 04).

3.1 Définitions et systèmes comparés

- **Réussite vérifiée :** réponse équivalente à la clé auditée (SymPy et test numérique), ou preuve acceptée par le noyau de Lean contre l'énoncé gelé, avec les seuls axiomes standard (`propext`, `Quot.sound`, `Classical.choice`).
- **Itération :** un cycle complet « compilation Lean → message d'erreur (et verdict SymPy) → révision ».
- **Lemme intermédiaire :** lemme produit par une étape de décomposition fixée, c'est-à-dire un plan de preuve figé avant le début de la boucle.

Code	Système	Rôle
B0	Même LLM seul, un essai	Baseline de H1a
B2	Même LLM, 5 échantillons indépendants, sélection par le même vérificateur SymPy/Lean	Contrôle de H1b : sépare la valeur de l'itération de celle de la vérification
B3	Boucle AIM²P complète (LLM + SymPy + Lean 4), 5 tours de révision au plus	Système évalué

Les autres systèmes de comparaison (échantillonnage avec oracle, ablations sans SymPy et sans Lean) sont décrits dans la Research Proposal.

3.2 H1 : gain de la boucle à budget égal

- **H1a :** la boucle B3 atteint un taux de réussite vérifiée supérieur d'au moins **20 points de pourcentage absolus** à celui du même LLM utilisé seul en un essai (B0).
- **H1b :** B3 dépasse aussi le même LLM disposant de cinq essais indépendants filtrés par le même vérificateur (B2), à budget de tokens comparable.
- **Critère de décision :** chaque sous-hypothèse est soutenue si la borne inférieure de l'intervalle de confiance à 95 % de la différence appariée (problème par problème) est strictement positive et, pour H1a, si l'estimation ponctuelle atteint 20 points. Elle est réfutée si l'intervalle se situe entièrement sous son seuil (20 points pour H1a, 0 pour H1b). Entre les deux, elle est déclarée non concluante. Le gain relatif en pourcentage n'est rapporté qu'à titre descriptif.

Pourquoi un seuil absolu et une baseline filtrée. Le seuil de la version 0.1 (« plus de 40 % », en relatif) était indétectable. Sur 40 problèmes, l'intervalle de Wilson à 95 % autour d'une exactitude de 50 % va de 35,2 % à 64,8 %. Un gain relatif de 40 % sur une baseline à 30 % ne représente qu'environ 5 problèmes de plus ; or, avec un test du signe exact, même 5 problèmes gagnés contre 0 perdu donnent p = 0,0625. Il faut 6 contre 0, 8 contre 1 ou 10 contre 2 pour conclure : quand peu de problèmes basculent dans l'autre sens, le plus petit écart détectable est de l'ordre de 8 problèmes nets, soit environ 20 points (calculs effectués lors du cadrage). Par ailleurs, une fois

le coût compté, l'auto-réparation n'aide que modestement (Olausson et al., arXiv:2306.09896). Si la boucle bat B0 mais pas B2, le gain vient de la vérification et de l'échantillonnage, pas de l'itération : ce serait un résultat légitime, mais différent de H1.

3.3 H2 : courbe de certification des lemmes

- **Énoncé :** pour les lemmes intermédiaires des items N3 à N5, avec un LLM, un contexte et un budget figés, au moins **70 % des lemmes sont certifiés en au plus 5 itérations** (noyau de Lean, axiomes standard uniquement, énoncé inchangé).
- **Ce qui est publié :** la courbe cumulée succès(k) pour k = 0 à 10, comparée à un rééchantillonnage indépendant de même coût, et ventilée par type d'erreur Lean (identifiant inconnu, incompatibilité de types, classes de types, délai dépassé).
- **Résultat secondaire :** la version conditionnelle (« parmi les lemmes finalement certifiés, au moins 70 % le sont en 5 itérations au plus ») est rapportée à part. Elle est presque certainement vraie, donc peu informative.

Ce que dit la littérature. Les données publiées soutiennent la *forme* de H2 : sur 1 990 théorèmes LeanDojo, 84 à 96 % du gain total de réparation arrive dès le premier tour (arXiv:2610.26942). Son *niveau* absolu n'a en revanche aucun précédent. Les taux de succès inconditionnels publiés restent bas, et certaines erreurs ne se réparent presque jamais : incompatibilités de types (1,8 % de réparations réussies), échecs de classes de types (2,0 %), délais dépassés (0,8 %). Or les lemmes de physique mathématique, chargés de coercions réel/complexe et de structures riches en classes de types, risquent de tomber dans ces catégories. À notre connaissance (novembre 2026), aucun article ne publie de courbe de réparation par lemme sur des énoncés de physique : c'est la contribution que H2 peut apporter.

3.4 H3 : détection et correction d'erreurs physiques par injection contrôlée

Protocole. Sur un corpus de dérivations correctes de physique mathématique (items N5 et mécanique hamiltonienne), on injecte des erreurs contrôlées et documentées : inversion de signe, terme de bord supprimé, condition aux limites erronée, terme non invariant, unités incohérentes, erreur d'indexation ou de normalisation. On compare les taux de détection et de faux positifs de trois critiques : un juge LLM sans accès aux artefacts du générateur, une vérification pas à pas par SymPy, et Lean 4. Les énoncés formels de référence sont relus par un humain avant toute mesure.

- **H3a (typage) :** les erreurs dimensionnelles et d'invariance sont rejetées dès l'élaboration lorsque les grandeurs sont typées (module `Units` de Physlib, indices tensoriels).
- **H3b (obligations de preuve) :** les hypothèses manquantes, les cas limites ou dégénérés et les conditions aux limites erronées (au sens d'une frontière de domaine de paramètres comme au sens d'une équation aux dérivées partielles) apparaissent comme des buts non prouvables ou des contre-exemples formels, avec un taux de détection supérieur à celui des deux autres critiques.
- **H3c (correction, exploratoire) :** la boucle propose un énoncé ou une dérivation corrigés, acceptés par Lean et jugés fidèles par un relecteur humain.

Pourquoi scinder H3. L'hypothèse dispose désormais d'un précédent direct : en formalisant en Lean un article de 2006 très cité sur la stabilité du potentiel du modèle à deux doublets de Higgs, Tooby-Smith a montré que la condition présentée comme nécessaire et suffisante n'est pas suffisante, avec un contre-exemple formellement vérifié (arXiv:2603.08139). La *correction* automatique est en revanche beaucoup moins étayée que la *détection*, et la dérive sémantique reste la principale menace : si le LLM formalise le mauvais énoncé, Lean certifie la mauvaise chose. D'où le statut exploratoire de H3c et la relecture humaine des énoncés de référence.

4. PRÉ-ENREGISTREMENT

Hypothèses pré-enregistrées

Les hypothèses H1 à H3 sont pré-enregistrées : elles sont fixées et horodatées publiquement avant toute mesure, avec les items d'AMP-Bench-mini, le correcteur, les identifiants et dates des modèles, les paramètres de décodage et les versions des outils. Les critères de décision et le protocole statistique (nombre de répétitions, tests appariés, intervalles) sont fixés dans la Research Proposal (document 04). Sur 40 items, AIM²P ne peut établir que des effets d'environ 20 points : les résultats par niveau seront rapportés à titre descriptif, et un résultat négatif ou non concluant sera publié comme tel.

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. Question principale reformulée (abandon de l'objectif « taux d'hallucination nul », comparaison à budget égal) ; SQ1 étendue à la fidélité des énoncés et aux certificats SymPy, SQ2 élargie aux boucles de révision agentiques, SQ3 recentrée sur la dynamique d'opérateurs en dimension finie et sur la détection d'erreurs ; H1 à H3 rendues falsifiables (seuil absolu et test apparié, baseline filtrée à budget égal, définitions opérationnelles, protocole d'injection d'erreurs) ; ajout du pré-enregistrement.