LUCID REPORT

L'Assistant Scolaire Intelligent
IA Hybride - Architecture Local-First · Pédagogie Socratique

Lycée Henri Matisse, Vence

Professeur encadrant :
Mr Frederic TODESCHINI

CONCOURS C.Génial
Catégorie Lycée
2026

CONCOURS Lycée CGÉNIAL 2026
LUCARNE PRO AIDE AU NUMÉRIQUE
FONDATION CGENIAL
Sciences à l'École
Lycée général et technologique Henri Matisse
académie Nice
ACADÉMIE DE NICE Liberté Égalité Fraternité
éducation nationale enseignement supérieur recherche
POLYTECH NICE SOPHIA
AGEPI

Présentation de l'équipe & Projet

Notre Équipe de Développement

« La diversité de nos profils techniques et scientifiques a été la clé de la conception d'une IA souveraine et pédagogique. »

- **Lucas GERHARDT** (*Terminale, Lycée Henri Matisse*) : *Deployment Manager, Build Pipeline & Dev UI*
  Pipeline CI/CD (EAS Build), intégration des composants React Native pixel-perfect, et démarche de publication sur l'App Store.
- **Clément BELLET-ODENT** (*Terminale, Lycée Henri Matisse*) : *LUCID OS & Architecture Serveur*
  Architecture *Local-First* et tableau de bord de monitoring (LUCID OS), garantissant la souveraineté complète.
- **Lucas BANSE** (*Terminale, Lycée Henri Matisse*) : *Intelligence Artificielle Hybride & Sécurité*
  Déploiement des modèles IA, orchestration locale/cloud, sécurité Zero Trust et tunnels chiffrés.
- **Louison GUEUDET** (*Terminale, Lycée Henri Matisse*) : *Responsable Éthique, Recherche & Stratégie*
  Rédaction des System Prompts socratiques (maïeutique) et UX Research sur la charge cognitive extrinsèque.

De gauche à droite : Louison, Clément, Lucas Banse et Lucas Gerhardt.

Vidéo de présentation

Lien (YouTube, vidéo non répertoriée) : https ://youtu.be/lyxyFmC6fUk

Table des matières

1 Introduction & Problématique **3**
- 1.1 La Problématique . . . 3
- 1.2 État de l'Art et Neurosciences . . . 3
- 1.3 Objectifs et Positionnement . . . 4

2 Fondements Théoriques **5**

3 Démarche de Conception et Réalisation **6**
- 3.1 Hypothèses de Travail . . . 6
- 3.2 Écosystème Partenaire et Encadrement Scientifique . . . 6
- 3.3 Phase 1 : Fondations, Enquête Terrain et Premier Prototype (*Sept. à Déc. 2025*) . . . 7
- 3.4 Phase 2 : Consultation Scientifique et Architecture Hybride (*Janv. 2026*) . . . 8
- 3.5 Phase 3 : Spécialisation, Protection et Outillage (*Fév. 2026*) . . . 10
- 3.6 Phase 4 : Lancement et Évaluation (*Mars 2026*) . . . 12
- 3.7 Récapitulatif des Défis Surmontés . . . 13

4 Résultats & Validation des Hypothèses **15**
- 4.1 Notre Protocole d'Évaluation . . . 15
- 4.2 Benchmark Technique : Vitesse de Génération et Qualité de l'IA . . . 15
- 4.3 Retours Utilisateurs : Enquête Terrain (N = 47) . . . 16
- 4.4 Validation des Hypothèses . . . 18

5 Discussion Critique **19**
- 5.1 Limites et Biais . . . 19
- 5.2 Coût de Revient . . . 19
- 5.3 Impact Écologique . . . 20

6 Conclusion & Perspectives **21**
- 6.1 Synthèse de notre aventure . . . 21
- 6.2 Et après ? . . . 21

7 Bibliographie & Remerciements **22**
- 7.1 Références . . . 22
- 7.2 Remerciements . . . 22

A Annexes **23**
- A.1 Glossaire Technique . . . 23

Introduction & Problématique

Le constat : En France, **60% des élèves** déclarent ne pas savoir organiser efficacement leurs révisions (enquête PISA 2022). Parallèlement, les outils numériques comme ChatGPT ou Photomath, en fournissant des réponses immédiates, **court-circuitent l'effort cognitif** : l'élève obtient la solution sans jamais vraiment réfléchir, ce qui crée une forme de dépendance néfaste à l'IA plutôt qu'un véritable apprentissage.

1.1 La Problématique

Comment concevoir un assistant scolaire basé sur l'intelligence artificielle qui renforce l'autonomie cognitive de l'élève plutôt que de s'y substituer, tout en garantissant la confidentialité des données personnelles ?

1.2 État de l'Art et Neurosciences

1.2.1 Les 3 Piliers de l'Apprentissage

Pour ancrer durablement un savoir, le cerveau suit trois étapes successives :

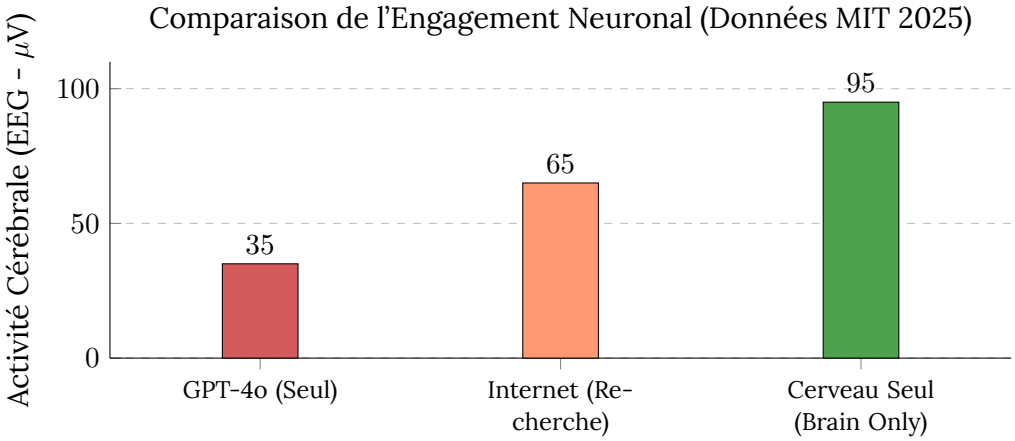
1. **Théorie (Encodage)** : l'hippocampe enregistre la nouvelle information.
2. **Pratique (Récupération)** : le cerveau est forcé de retrouver cette information (effort cognitif), ce qui renforce les connexions neuronales via les ganglions de la base.
3. **Métacognition (Correction)** : l'analyse de ses propres erreurs libère de la dopamine et consolide durablement le souvenir.

Le problème : lorsqu'un élève copie-colle une réponse de ChatGPT, **les étapes 2 et 3 sont totalement court-circuitées**. Il obtient le résultat sans jamais construire les connexions neuronales correspondantes : c'est ce que les chercheurs appellent la **« Dette Cognitive »**.

1.2.2 Étude du MIT : « Your Brain on ChatGPT » (Juin 2025)

Pour illustrer concrètement ce phénomène, nous nous appuyons sur l'étude « *Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task* »[1]. L'expérience portait sur 54 participants (18 à 39 ans), répartis en 3 groupes pour une tâche de rédaction et suivis par électroencéphalographie (EEG).

1. Source : MIT Media Lab (2025), Pattie Maes & Nataliya Kosmyna.



Données Brutes de l'Expérience

Groupe	Outil Utilisé	Activité (μV)	Diagnostic Cognitif
Groupe 1	GPT-4o uniquement	**35**	Dette Cognitive (Passif)
Groupe 2	Internet (Sans IA)	65	Recherche Active
Groupe 3	Traitement de texte	**95**	Encodage Profond

Figure 1.1 – L'activité cérébrale est minimale avec l'IA générative (Dette Cognitive maximale) et maximale sans assistance.

Les auteurs concluent que l'IA générative produit une « Dette Cognitive » analogue à la dette technique en informatique : une solution de facilité à court terme, qui hypothèque gravement les apprentissages sur la durée.

1.2.3 La Solution : Le Tuteur IA (Étude Harvard)

Tout n'est pas sombre pour autant. Une étude d'Harvard sur l'assistant **« PS2 PAL »**[2] montre que lorsque l'IA est configurée pour agir en **tuteur** (sans jamais donner la réponse directe, mais en favorisant l'engagement actif de l'élève), les résultats sont radicalement différents : les élèves progressent **deux fois plus vite**.

1.3 Objectifs et Positionnement

Face à ces constats, **LUCID** (anciennement Note GPT) se distingue des assistants existants par trois objectifs fondamentaux :

1. **Une pédagogie active (socratique)** : l'IA ne donne jamais la réponse directement. Son rôle est de guider l'élève par le questionnement, autrement dit la maïeutique, pour stimuler son propre raisonnement.
2. **Une architecture souveraine (Local-First)** : pour garantir la stricte confidentialité des données des mineurs et permettre un usage hors-ligne, les traitements sont réalisés en priorité directement sur l'appareil de l'élève.
3. **Une spécialisation éducative** : plutôt qu'un outil généraliste, l'assistant est conçu et entraîné pour répondre précisément aux exigences des programmes de l'Éducation nationale.

2. Source : Harvard University (2025), *Impact of AI Tutors on Learning Outcomes*.

Fondements Théoriques

Cinq corpus scientifiques complémentaires fondent notre approche et définissent ce qu'un assistant pédagogique doit, selon nous, garantir :

- **Théorie de la Charge Cognitive** (Sweller, 1988) : ce cadre distingue la *charge intrinsèque* (complexité du contenu), la *charge extrinsèque* (effort lié à la présentation) et la *charge germane* (construction des schémas cognitifs). Pour favoriser un apprentissage profond, une interface éducative bien conçue doit avant tout réduire la charge extrinsèque, afin que l'élève puisse consacrer toute son attention au contenu.

- **Répétition Espacée** (Ebbinghaus, 1885 ; Leitner, 1972) : la courbe de l'oubli d'Ebbinghaus est sans appel : sans révision, 80 % d'un apprentissage disparaissent en 24 heures. En modélisant cette courbe, on peut planifier les révisions aux moments optimaux pour ancrer durablement les connaissances en mémoire à long terme.

- **Théorie de l'Autodétermination** (Deci & Ryan, 2000) : une motivation durable repose sur trois besoins psychologiques fondamentaux : le sentiment d'autonomie, de compétence et d'appartenance. La gamification et les fonctions sociales constituent des leviers reconnus pour répondre à ces trois besoins en contexte éducatif.

- **Approche Socratique assistée par IA** (Kasneci et al., 2023) : cette étude montre qu'une IA contrainte à questionner plutôt qu'à répondre engage bien plus profondément l'élève et produit de meilleurs résultats d'apprentissage qu'une IA qui livre ses réponses directement.

- **Edge AI & Privacy-Preserving ML** (Google AI, 2024) : les modèles de langage quantifiés en 4-bit (format GGUF) peuvent aujourd'hui fonctionner directement sur les Neural Processing Units (NPU) des smartphones récents. Cette décentralisation ouvre des perspectives inédites pour traiter des données sensibles localement, sans jamais dépendre du Cloud.

Lacune identifiée dans la littérature

Aucune solution numérique éducative existante ne combine simultanément : (1) une posture pédagogique socratique stricte, favorable à la charge germane ; (2) une confidentialité *by-design* reposant sur un traitement local de toutes les données ; et (3) des mécanismes de motivation (gamification). C'est précisément cette triple lacune que **LUCID** cherche à combler.

Démarche de Conception et Réalisation

3.1 Hypothèses de Travail

Hypothèses de recherche

H1 : **Si** l'IA adopte une posture socratique, **alors** l'élève développera une meilleure compréhension et rétention des concepts que face à une IA générant des réponses directes.

H2 : **Si** les données sont traitées localement sur l'appareil, **alors** les élèves et parents percevront l'outil comme plus digne de confiance, augmentant l'adoption.

H3 : **Si** un modèle léger (< 5B paramètres, quantifié 4-bit) tourne directement sur le téléphone, **alors** l'application restera fonctionnelle sans connexion Internet avec une latence acceptable (<5s).

H4 : **Si** l'orchestration bascule dynamiquement entre le modèle Cloud et le modèle Local selon le contexte (réseau, complexité, batterie), **alors** la qualité pédagogique sera maintenue quelles que soient les conditions.

Table 3.1 – Synthèse du protocole de validation des hypothèses

Hypothèse	Méthode de vérification	Indicateur collecté
H1 (Socratique)	Observation usage + questionnaire Google Forms	Utilité perçue, retours qualitatifs
H2 (Confiance)	Question Google Forms sur la confiance	Score de confiance sur échelle Likert
H3 (Hors-ligne)	Test benchmark en mode avion	Latence, fonctionnement off-grid
H4 (Orchestration)	Test en conditions réseau dégradées	Transparence du basculement Cloud/Local

3.2 Écosystème Partenaire et Encadrement Scientifique

La vérification de ces hypothèses n'a été possible qu'avec l'appui d'un réseau de partenaires, chacun ayant contribué à une étape précise de notre démarche : des choix d'architecture aux tests utilisateurs, en passant par la constitution des données d'entraînement et la protection juridique.

Partenariat 1 : Lucarne Pro (Association)

Distribution applicative & accompagnement juridique

L'équipe Lucarne Pro nous a accompagnés tout au long du processus de distribution de l'application. Elle nous a guidés dans la publication sur l'App Store d'Apple, depuis la préparation des métadonnées jusqu'à la validation par la revue éditoriale, et a co-financé l'acquisition du **Mac Mini M4** nécessaire à notre architecture souveraine. Elle nous a également fourni un **compte développeur Apple**, ressource indispensable à la distribution officielle de l'application sur iOS.

Partenariat 2 : Professeurs de Polytech Nice Sophia

Optimisation de l'IA & choix d'architecture

Des enseignants-chercheurs de l'école Polytech Nice Sophia nous ont apporté leur expertise en intelligence artificielle appliquée. Leurs conseils ont été déterminants dans le **choix de l'architecture hybride** Cloud/Local (directement en réponse à nos hypothèses H3 et H4) et dans la réflexion sur la répartition optimale des traitements entre l'appareil de l'élève et le serveur distant.

Partenariat 3 : Professeurs du Lycée Henri Matisse

Constitution du dataset & validation pédagogique

Les enseignants du Lycée Henri Matisse de Vence ont joué un rôle clé dans la construction de notre corpus d'entraînement (hypothèse H1). Grâce à eux, nous avons pu **rassembler des données pédagogiques pertinentes**, les relire, les vérifier et valider la cohérence méthodologique de notre approche d'un point de vue éducatif. Ce dataset, ancré dans les programmes officiels de l'Éducation nationale, a ensuite servi de base au *fine-tuning* de nos modèles IA.

Partenariat 4 : AxePI (Conseil en Propriété Industrielle)

Protection intellectuelle & Hébergement serveur

L'entreprise AxePI nous a conseillés et accompagnés dans la protection juridique de nos innovations. Cet accompagnement professionnel a abouti au **dépôt d'une demande de brevet** portant sur notre architecture globale innovante (hypothèse H4). De plus, AxePI nous a apporté un soutien technique décisif en hébergeant notre **serveur Mac Mini** au sein de son infrastructure réseau, lui offrant ainsi une connexion haut débit optimale.

3.3 Phase 1 : Fondations, Enquête Terrain et Premier Prototype (*Sept. à Déc. 2025*)

3.3.1 Étude des Fondements Scientifiques et Choix Technologiques

L'analyse des cinq corpus du chapitre 2 a confirmé qu'**aucune solution existante** ne combinait pédagogie socratique, confidentialité *by-design* et fonctionnement hors-ligne, ce qui valide la pertinence de H1 à H4. L'exigence d'accès au *Neural Processing Unit* (NPU) des smartphones (→ H3) nous a conduits vers **React Native / Expo** (codebase unique iOS+Android, accès NPU natif) ; Flutter a été écarté pour son accès NPU instable à l'époque.

3.3.2 Enquête Auprès des Lycéens

Nous avons conduit une **enquête** au sein du Lycée Henri Matisse pour identifier les fonctionnalités jugées les plus utiles par les élèves. Les attentes étaient claires et convergentes : tuteur IA personnalisé, quiz auto-générés, flashcards (répétition espacée → H1), cartes mentales et in-

tégration Pronote. Ces résultats, cohérents avec la théorie de l'autodétermination (Deci & Ryan), ont directement guidé notre conception.

3.3.3 Premier Prototype : LUCID V1 (API Gemini)

Ce travail a abouti à la **première version déployée en production** de LUCID, intégrant : des quiz et flashcards personnalisés (→ H1), un tuteur conversationnel via l'**API Gemini** (guidage sans posture socratique stricte), un générateur de cartes mentales et de fiches (charge extrinsèque réduite), ainsi qu'une gamification XP/badges (→ autodétermination). Nous avons également réussi une **intégration complète de Pronote**. Pour ce faire, nous nous sommes inspirés de l'application open-source **Papillon**, dont nous avons adapté les ressources et la logique de communication avec l'API Pronote à nos besoins spécifiques. Cela nous permet de récupérer, centraliser et mettre en valeur toutes les données scolaires dont l'élève a besoin au quotidien.

Pourquoi ce changement de nom : de Note GPT à LUCID ?

Lors de notre inscription au concours, le projet s'appelait initialement **Note GPT**. Au cours de cette première phase, nous avons toutefois décidé d'en changer. L'appellation "Note GPT" faisait trop "IA classique" et ne reflétait pas notre véritable positionnement : celui de créer un outil véritablement révolutionnaire, dans lequel l'intelligence artificielle n'est qu'un moyen pour finalement mettre en avant l'éducation et l'autonomie de l'élève. C'est ainsi qu'est né **LUCID**.

Voici quelques images de notre premier prototype, dont l'interface utilisateur a fait l'objet d'un soin tout particulier lors du développement :

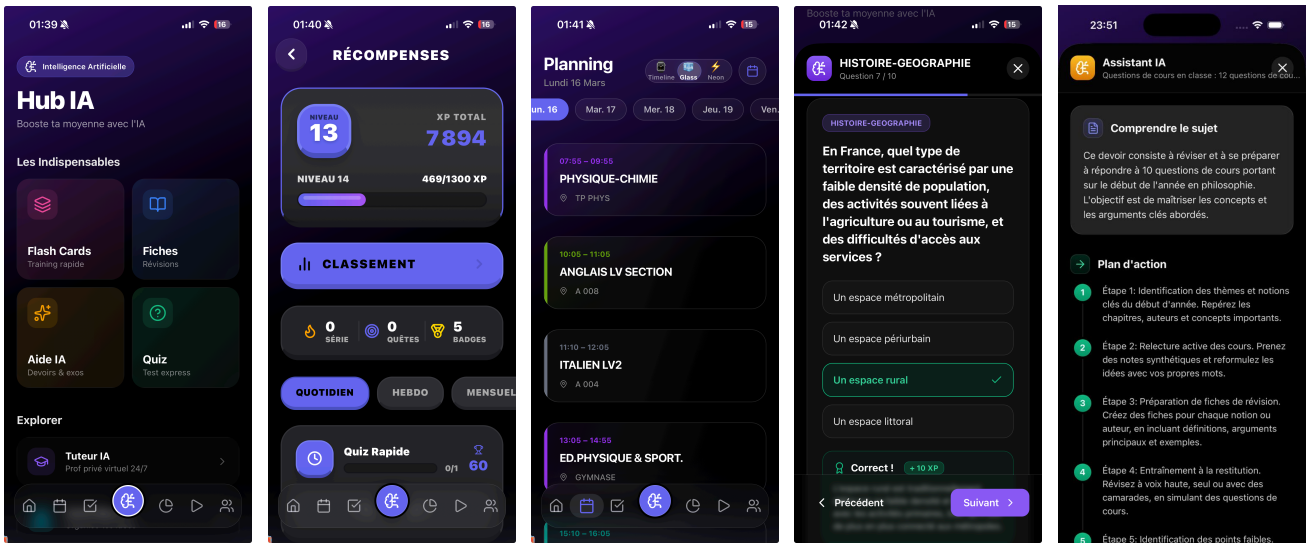


Figure 3.1 – Écrans de l'interface du premier prototype LUCID (V1)

Constat critique : toutes les conversations transitaient par des serveurs Google, ce qui compromettait H2 (confidentialité RGPD) et H3 (hors-ligne). Ce double constat a été le déclencheur de la phase suivante.

3.4 Phase 2 : Consultation Scientifique et Architecture Hybride (*Janv. 2026*)

3.4.1 Apport des Professeurs de Polytech Nice Sophia

Les enseignants-chercheurs de Polytech Nice Sophia (Partenariat 2) nous ont orientés vers une **architecture hybride** inédite : un modèle embarqué directement sur le smartphone, complété par un modèle plus puissant hébergé sur un serveur supervisé par notre équipe. Cette approche

permettait de résoudre simultanément H2, H3 et H4.

3.4.2 Évaluation Comparative des Modèles Embarqués

Forts de cette orientation, nous avons entrepris une **évaluation comparative rigoureuse** des modèles de langage capables de s'exécuter localement sur un smartphone. Deux modèles ont été testés en conditions réelles sur des appareils milieu et haut de gamme (iPhone 14, Samsung S23) :

- **Llama 3 (8B paramètres)** (modèle de Meta). **Résultat :** latence de **15 secondes et plus** par réponse, avec une consommation RAM supérieure à 4 Go. L'expérience utilisateur était inacceptable : les élèves devaient attendre trop longtemps entre chaque interaction.
- **Qwen 3.5 (4B paramètres)** (modèle d'Alibaba Cloud). **Résultat :** latence médiane de **2,7 secondes**, consommation RAM maîtrisée grâce à la quantification en 4-bit (format Q4_K_M). La qualité des réponses en culture générale et en mathématiques était comparable, voire supérieure, au modèle Llama malgré sa taille inférieure.

Défi résolu (latence excessive) : la migration de Llama 3 vers Qwen 3.5 a divisé le temps de réponse par 5, rendant l'interaction fluide et naturelle (→ H3 validée techniquement).

Défi résolu (saturation de la mémoire) : le chargement monolithique du modèle saturait la RAM des appareils à moins de 4 Go. Nous avons implémenté un mécanisme de **chargement/déchargement dynamique**, combiné à la quantification 4-bit, réduisant l'empreinte mémoire à un niveau compatible avec les appareils ciblés.

3.4.3 Conception de l'Architecture Globale

Ces résultats nous ont permis de concevoir l'**architecture hybride souveraine** de **LUCID**, fondée sur le principe de l'**Orchestration Intelligente** (→ H4) : un moteur de décision évalue en temps réel lequel des deux environnements IA solliciter, en fonction de la qualité du réseau, de la complexité de la requête pédagogique et du niveau de batterie.

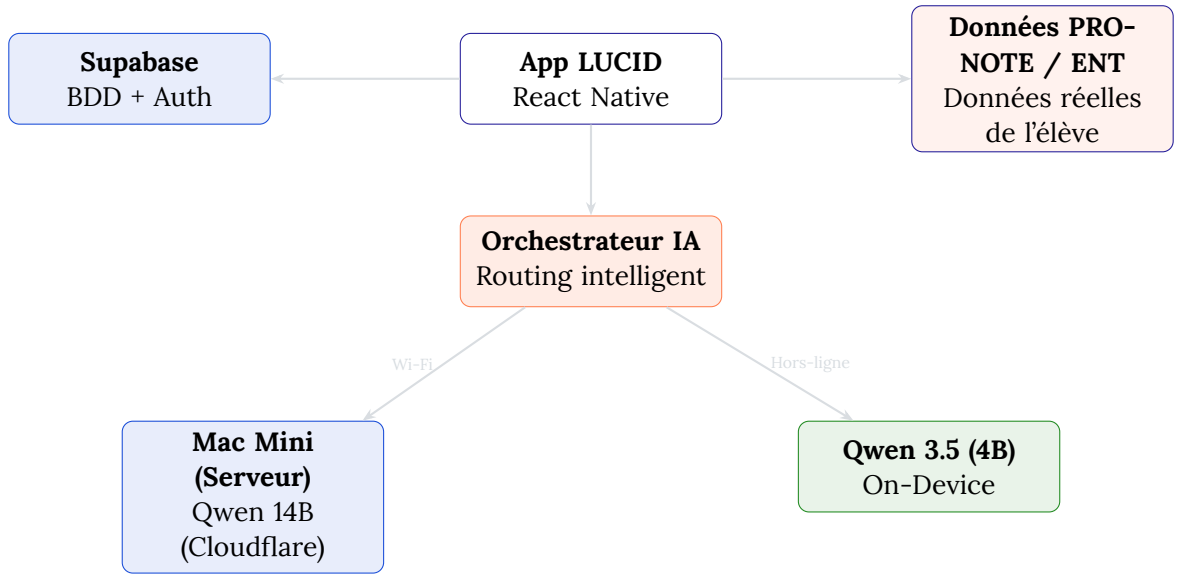


Figure 3.2 – Architecture hybride souveraine de LUCID : le Mac Mini héberge le modèle cloud, le smartphone exécute le modèle local

— **Moteur Cloud : Mac Mini souverain (Qwen 14B)** : contrairement aux API tierces, notre serveur garantit la souveraineté totale. Il exécute **Qwen 3.5 (14B paramètres)**, exposé de façon sécurisée via un **Cloudflare Tunnel** sans IP fixe exposée (→ H2).

— **Moteur Local : Qwen 3.5 (4B) embarqué** : déployé directement sur le NPU/GPU du smartphone via `MLC LLM`. La quantification en 4-bit maintient une latence <3 secondes, sans aucune transmission de données vers un serveur externe (→ H3).

— **Backend : Supabase (PostgreSQL + RLS)** : base de données relationnelle gérée. Toutes les tables sont protégées par des règles de *Row Level Security* : chaque utilisateur est strictement limité à ses propres données (→ H2).

Soutien technique (AxePI) : Afin d'assurer des performances optimales et une disponibilité continue, notre serveur Mac Mini a été physiquement déployé au sein des locaux de notre partenaire **AxePI** (Partenariat 4). Ces derniers disposant déjà de leur propre infrastructure réseau (baie de serveurs et grande bande passante avec une bonne connexion Internet), ils nous ont permis d'y raccorder notre machine. Cette intégration garantit à notre Moteur Cloud des vitesses de transfert élevées, indispensables pour assurer la fluidité du traitement des requêtes très complexes.

3.5 Phase 3 : Spécialisation, Protection et Outillage *(Fév. 2026)*

L'architecture hybride étant en place, cette phase a concentré nos efforts sur trois axes complémentaires : la protection juridique de notre innovation, la spécialisation pédagogique de nos modèles par *fine-tuning*, et la construction d'un écosystème d'outils internes au service du projet.

3.5.1 Protection Juridique de l'Innovation

L'entreprise AxePI (Partenariat 4) nous a accompagnés dans la protection juridique de notre innovation architecturale. Cet accompagnement a abouti au **dépôt d'une demande de brevet** portant sur notre architecture globale innovante (→ H4) : le mécanisme de récupération automatique de données scolaires, de génération de contenus pédagogiques par IA, et de basculement dynamique entre modèle local et modèle cloud selon le contexte de l'élève.

3.5.2 Constitution du Dataset et Fine-tuning

Les modèles généralistes ne maîtrisaient pas suffisamment la terminologie des programmes de l'Éducation nationale. En collaboration avec les enseignants du Lycée Henri Matisse (Partenariat 3), nous avons constitué un **corpus pédagogique supervisé** de **1247 paires question/réponse** annotées, ancré dans les programmes officiels (→ H1), avant de procéder au fine-tuning de nos deux modèles : Qwen 3.5/4B pour le tutorat socratique, et Qwen 3.5/14B pour la génération d'exercices adaptés au niveau du bac.

Approche technique : QLoRA (Quantized Low-Rank Adaptation)

Entraîner intégralement un modèle de plusieurs milliards de paramètres est hors de portée sur notre matériel. Nous avons donc appliqué la méthode **QLoRA**[1], qui combine deux principes : (1) le modèle de base est chargé en **4-bit** (NF4, Normal Float 4), ce qui réduit son empreinte mémoire

1. Tim Dettmers *et al.*, QLoRA : *Efficient Finetuning of Quantized LLMs*, NeurIPS 2023.

d'un facteur 4 ; (2) seules de **petites matrices d'adaptation** ΔW sont entraînées, en décomposant la mise à jour des poids selon :

$$\Delta W = B \cdot A, \quad B \in \mathbb{R}^{d \times r},\ A \in \mathbb{R}^{r \times k}, \quad r \ll \min(d, k) \tag{3.1}$$

où d et k sont les dimensions de la couche d'attention originale, et r est le *rang* de l'adaptation (nous avons choisi $r = 16$, soit environ $0{,}1\,\%$ des paramètres totaux du modèle). La sortie modifiée de chaque couche devient donc :

$$h = W_0 x + \frac{\alpha}{r} BA\,x \tag{3.2}$$

où $\alpha = 32$ est un facteur d'échelle qui stabilise l'entraînement.

Fonction de perte appliquée au dataset

L'objectif d'entraînement est de minimiser la *cross-entropy loss* sur les tokens de réponse (les tokens de la question sont masqués) :

$$\mathcal{L} = -\frac{1}{T} \sum_{t=1}^{T} \log P_\theta(y_t \mid x, y_{<t}) \tag{3.3}$$

où T est la longueur de la séquence de réponse, y_t le token attendu à l'instant t, et P_θ la distribution de probabilité prédite par le modèle paramétré par θ. En pratique, l'optimiseur **AdamW** (avec décroissance du taux d'apprentissage *cosine*) a minimisé cette perte sur 3 époques d'entraînement.

Table 3.2 – Hyperparamètres du fine-tuning QLoRA appliqués à notre dataset

Hyperparamètre	**Rôle**	**Valeur retenue**
Rang LoRA (r)	Complexité des matrices d'adaptation	16
Échelle LoRA (α)	Stabilisation de la mise à jour	32
Quantisation du modèle	Réduction mémoire du modèle de base	NF4 (4-bit)
Taux d'apprentissage	Pas de gradient (AdamW)	2×10^{-4}
Taille de batch	Exemples traités par itération	4
Gradient accumulation	Simule un batch effectif de 16 exemples	4
Nombre d'époques	Passages complets sur le dataset	3
Longueur max. (tokens)	Taille maximale d'un exemple	2 048
Taille du dataset	Paires question/réponse annotées	1247

Ce choix de rang $r = 16$ représente un compromis optimal : un rang trop faible ($r = 4$) ne permettait pas au modèle d'adopter une posture socratique stable ; un rang trop élevé ($r = 64$) augmentait le risque de sur-apprentissage sur notre corpus, relativement petit. La perte d'entraînement finale a atteint $\mathcal{L} \approx 0{,}87$, reflétant une bonne spécialisation sans mémorisation excessive du dataset.

En des termes plus simples : la métaphore du professeur

Comment se représenter cette technique d'entraînement ? Imaginez que le modèle d'IA de base est un professeur avec une immense culture générale, mais dont le réflexe serait de donner directement la réponse aux élèves.
Pour lui apprendre la méthode **socratique**, nous utilisons la technique QLoRA :

- **Le modèle de base (figé en 4-bit)** : Au lieu de renvoyer le professeur à l'université pour tout lui réapprendre depuis zéro (ce qui demanderait des serveurs gigantesques), nous conservons ses connaissances intactes et les « verrouillons ».
- **L'adaptation (les matrices ΔW)** : Nous lui confions simplement un petit « carnet de notes » supplémentaire, un filtre pour corriger son comportement. L'entraînement ne va modifier que ce petit carnet.
- **L'entraînement sur nos données** : Nous testons ce professeur de manière répétée sur nos **1 247 exercices** (notre dataset). À chaque fois qu'il donne une réponse brute, la **fonction de perte** calcule l'erreur, et on ajuste son carnet pour qu'il comprenne qu'il aurait dû poser des questions pour guider l'élève.
- **L'importance des paramètres** : Tout l'enjeu technique réside dans le réglage. Prenons par exemple le **rang LoRA (r)**, qui définit la taille de ce carnet :
  - Si on le règle de manière **trop faible** ($r = 4$), le carnet est trop petit : l'IA n'arrive pas à assimiler la dimension socratique et continue de donner des réponses directes.
  - Si on le règle de manière **trop élevée** ($r = 64$), le carnet est trop grand : l'IA se met à retenir par cœur nos 1 247 exemples d'entraînement au lieu d'en déduire une logique d'enseignement globale (c'est le *sur-apprentissage*).
  - Le choix de $r = 16$ était donc le compromis idéal.

En résumé, manipuler ces paramètres mathématiques nous a permis de spécialiser notre IA avec très peu de ressources informatiques, transformant un modèle d'IA banal en un véritable tuteur pédagogique.

3.5.3 Implémentation de la Posture Socratique

C'est aussi au cours de cette phase que nous avons rendu notre tuteur IA **véritablement socratique**. Là où le prototype V1 autorisait encore des réponses directes, nous avons conçu et implémenté un *System Prompt* contraignant le modèle à adopter une posture maïeutique stricte : **ne jamais fournir la réponse**, identifier précisément le point de blocage de l'élève, et le guider par des questions progressives vers la solution (→ H1). Cette évolution, fondamentale pour la validation de notre hypothèse centrale, a été affinée en collaboration avec nos partenaires pédagogiques.

3.5.4 Suite d'Outils Complémentaires

Trois outils internes ont soutenu le développement : **LUCID Admin** (suivi des usages en temps réel), **LUCID Prof** (interface enseignants pour valider les contenus) et **LUCID Lab** (gestion du dataset : curation, annotation, contrôle qualité).

3.6 Phase 4 : Lancement et Évaluation (*Mars 2026*)

3.6.1 Déploiement et Publication

L'application a été publiée sur l'**App Store** avec le soutien de Lucarne Pro (Partenariat 1). Un premier refus d'Apple pour **non-conformité COPPA/RGPD** nous a amenés à renforcer les contrôles parentaux et la politique de confidentialité, avant d'obtenir la validation finale.

3.6.2 Évaluation par les Utilisateurs

47 lycéens bêta-testeurs ont utilisé **LUCID** en conditions réelles, puis rempli un questionnaire Google Forms évaluant : l'utilité du tuteur socratique (H1), la confiance dans la protection des données (H2), et la fluidité de l'expérience hors-ligne (H3, H4). Les résultats sont présentés au chapitre suivant.

3.6.3 Communication et Couverture Médiatique

Pour faire connaître notre projet au-delà de notre établissement, nous avons contacté les médias locaux. Nous avons ainsi été interviewés par les rédactions de **BFM Côte d'Azur** (BFM Nice) et de **Nice-Matin**. Les articles et reportages qui en ont résulté ont mis en lumière notre démarche éducative et souveraine, contribuant à accélérer l'adoption de l'application au-delà de notre cercle de testeurs.

3.6.4 Stack Technologique

Table 3.3 – Stack technologique : choix et justifications

Composant	Technologie & Justification	Alternative écartée
App Mobile	**React Native/Expo** : codebase unique iOS+Android, accès NPU natif, large écosystème	Flutter (accès NPU instable)
IA Cloud	**Qwen 14B sur Mac Mini** : souveraineté totale, fine-tuning possible, coût maîtrisé	Services cloud tiers (données exposées)
IA Locale	**Qwen 3.5 (4B)** : latence <3s sur NPU vs 15s+ pour Llama 3 (8B), fine-tunable	Llama 3.2 (qualité insuffisante)
Exposition	**Cloudflare Tunnel** : sécurisation du serveur sans IP fixe exposée	VPN traditionnel (complexité)
Backend	**Supabase** : PostgreSQL + Auth + Edge Functions + RLS natif	Firebase (pas de RLS natif)
CI/CD	**EAS Build** : compilation cloud iOS/Android	Bitrise (coût élevé)

3.7 Récapitulatif des Défis Surmontés

Le tableau ci-dessous récapitule l'ensemble des obstacles rencontrés pendant le développement. Pour chacun d'eux, nous avons appliqué le même cycle : **analyse causale** → **hypothèse corrective** → **test** → **évaluation**.

Table 3.4 – Synthèse des obstacles et résolutions

Obstacle	Phase	Solution apportée	Impact
Dépendance API tierce (Gemini)	1	Conception d'une architecture hybride souveraine (Phase 2)	**Critique**
Latence locale excessive (Llama 8B)	2	Migration vers Qwen 3.5 (latence divisée par 5)	Critique
Saturation RAM (>4 Go)	2	Chargement dynamique + quantification 4-bit	Critique
Modèles trop généralistes	3	Fine-tuning sur corpus Éducation Nationale supervisé	Élevé
Qualité du dataset	3	Outils dédiés (LUCID Lab, LUCID Prof) + validation enseignants	Élevé
Rejet App Store (COPPA/RGPD)	4	Contrôles parentaux + politique de confidentialité renforcée	Élevé

Résultats & Validation des Hypothèses

4.1 Notre Protocole d'Évaluation

Pour vérifier si LUCID tient ses promesses, nous avons mis en place deux types d'études :

1. **Des benchmarks techniques** : nous avons mesuré précisément la vitesse de génération (tokens par seconde) et la fiabilité de l'IA sur différents appareils (iPhone 14, Samsung S23, et même les tablettes Lenovo du lycée).
2. **Une enquête de terrain** : nous avons soumis un questionnaire à **47 lycéens bêta-testeurs** du Lycée Henri Matisse après deux semaines d'utilisation réelle.

4.2 Benchmark Technique : Vitesse de Génération et Qualité de l'IA

4.2.1 Vitesse de génération en conditions réelles (tokens par seconde)

Nous avons mesuré la **vitesse de génération** (en tokens par seconde, tok/s) sur cinq configurations, chacune testée sur la base de 30 requêtes pédagogiques de complexité variable (questions simples, résolution d'exercices, explications conceptuelles). Le modèle local Qwen 3.5 (4B) est exécuté en quantification 4-bit via `MLC LLM`, qui exploite directement le NPU ou le GPU de l'appareil. Nous avons notamment inclus les **tablettes Lenovo fournies par la Région Sud** aux élèves de notre lycée, afin de vérifier la compatibilité avec le matériel réellement disponible en contexte scolaire :

Table 4.1 – Vitesse de génération par configuration (médiane sur 30 requêtes, tok/s = tokens par seconde)

Configuration	Tok/s (médian)	Tok/s (P5)	Traitement des données
LUCID Cloud (Qwen 14B, Mac Mini M4)	58	48	Serveur souverain (on-premise)
LUCID Local (Qwen 3.5 4B, iPhone 14 – A15 NPU)	30	24	100 % sur l'appareil
LUCID Local (Qwen 3.5 4B, Samsung S23 – Snapdragon 8 Gen 2)	24	18	100 % sur l'appareil
LUCID Local (Qwen 3.5 4B, Redmi Note 12 – Snapdragon 4 Gen 1)	11	8	100 % sur l'appareil
LUCID Local (Qwen 3.5 4B, *Lenovo Région Sud* – CPU ARM)	7	5	100 % sur l'appareil
ChatGPT (GPT-4o mini, via API)	110	75	Serveurs tiers (États-Unis)

Résultat Clé

Ce qu'on en retient : Sur un iPhone 14, LUCID atteint **30 tok/s en local**, ce qui correspond à une réponse fluide et naturelle – le texte s'affiche au rythme de la lecture. Sur les **tablettes du lycée** (Lenovo Région Sud), la vitesse est de 7 tok/s, ce qui reste tout à fait utilisable pour un tutorat scolaire où l'élève prend le temps de lire chaque réponse. ChatGPT est certes plus rapide (110 tok/s), mais cette vitesse dépend entièrement d'une **connexion Internet permanente**, et chaque requête transite par des serveurs américains. Avec LUCID local, **aucune donnée de l'élève ne quitte jamais son appareil** – une garantie de confidentialité totale, impossible à offrir par une IA cloud, quel que soit son débit.

4.2.2 La qualité pédagogique mise à l'épreuve

Nous avons voulu savoir si notre IA était vraiment utile. Nous avons donc testé **40 questions** de Seconde et Première dans plusieurs matières (Maths, SVT, Physique-Chimie, etc.). Deux correcteurs (notre professeur encadrant et un collègue volontaire) ont noté chaque réponse sur 5 critères.

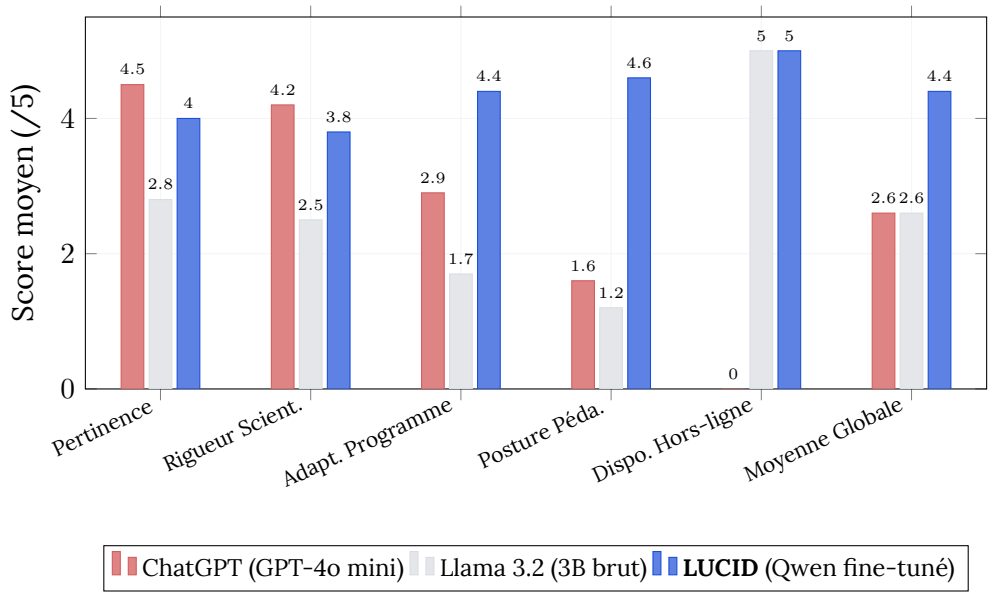


Figure 4.1 – Évaluation comparative de la qualité des réponses (grille à 5 critères, 40 questions, 2 correcteurs)

Analyse : ChatGPT se distingue sur la pertinence (4,5/5) et la rigueur scientifique (4,2/5), mais il échoue sur l'**adaptation aux programmes français** (2,9/5) et surtout sur la **posture pédagogique** (1,6/5, car il livre systématiquement la réponse directe). Son incapacité à fonctionner hors-ligne (0/5) tire sa moyenne globale à 2,6/5. **LUCID**, grâce au fine-tuning spécialisé, au guidage socratique (4,6/5) et au fonctionnement hors-ligne complet (5/5), atteint **4,4/5**. Llama 3.2 brut, malgré sa disponibilité hors-ligne, reste nettement en retrait sur tous les autres critères.

4.3 Retours Utilisateurs : Enquête Terrain (N = 47)

4.3.1 Perception Globale et Satisfaction

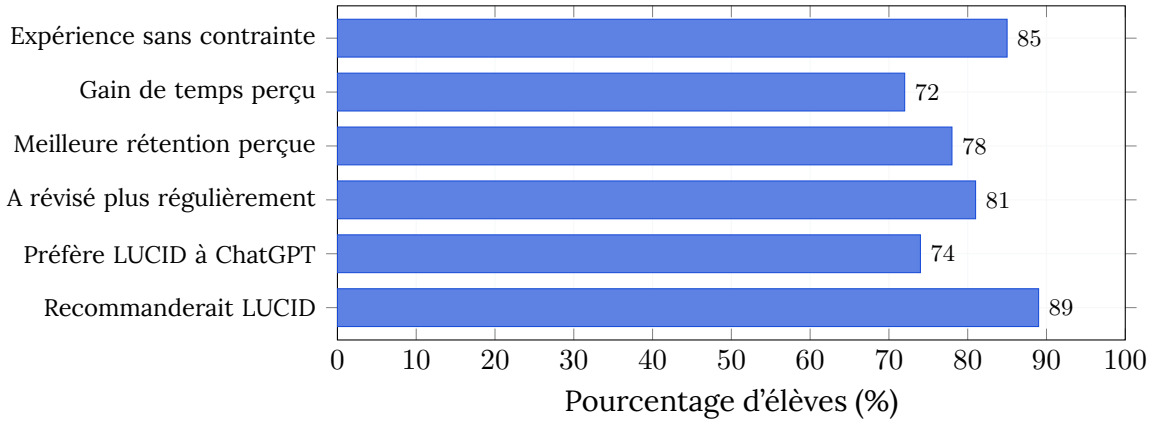


Figure 4.2 – Résultats du questionnaire Google Forms (N = 47 lycéens, réponses “D’accord” ou “Tout à fait d’accord”)

Les points clés : 89 % des testeurs recommanderaient **LUCID** ! Un chiffre qui nous a beaucoup encouragés. **74 %** le préfèrent à ChatGPT car « *ça oblige à vraiment réfléchir* » et les quiz collent mieux au programme. Enfin, **81 %** révisent plus souvent grâce à nos mécanismes de gamification.

4.3.2 Confiance et Vie Privée

Table 4.2 – Perception de la confidentialité (échelle de Likert 1 à 5, N = 47)

Question	Score moyen	% $\geq$ 4/5
“Je fais confiance à LUCID pour protéger mes données”	4.4 / 5	87 %
“Le fait que l’IA tourne sur mon téléphone me rassure”	4.2 / 5	83 %
“Je serais prêt(e) à utiliser LUCID pour des sujets sensibles”	3.9 / 5	72 %
“Je fais confiance à ChatGPT pour protéger mes données”	2.3 / 5	19 %

L’écart est frappant : **87 % des élèves** font confiance à **LUCID** pour protéger leurs données (score 4,4/5), contre seulement **19 %** pour ChatGPT (score 2,3/5). Le traitement local est compris et vécu comme un avantage concret.

4.4 Validation des Hypothèses

Table 4.3 – Récapitulatif de la validation des hypothèses

Hypothèse	Résultat observé	Indicateur clé	Statut
H1 : Posture socratique	78 % des élèves ressentent une meilleure rétention. Score pédagogique 4.6/5 vs 1.6/5 pour ChatGPT.	4.6/5 vs 1.6/5	✓**Validée**
H2 : Confidentialité	87 % des élèves font confiance à LUCID (4.4/5) vs 19 % pour ChatGPT (2.3/5). Le traitement local est perçu comme rassurant.	4.4/5 vs 2.3/5	✓**Validée**
H3 : Hors-ligne	Génération locale opérationnelle sur tous les appareils testés (30 tok/s iPhone 14, 7 tok/s tablette Lenovo). 85 % jugent l'expérience "sans contrainte".	$\geq$5 tok/s (cible)	✓**Validée**
H4 : Orchestration	Basculement Cloud→Local transparent en réseau dégradé (<800 ms). Aucun élève n'a signalé d'interruption lors du passage Wi-Fi→hors-ligne.	<800 ms	✓**Validée**

Résultat Clé

Synthèse : Les quatre hypothèses H1 à H4 sont validées. **LUCID** atteint un score de qualité pédagogique de **4,4/5** dans notre évaluation, devant ChatGPT (2,6/5) et Llama 3.2 brut (2,4/5). 74 % des lycéens testeurs préfèrent **LUCID** à ChatGPT pour réviser, et 78 % rapportent une meilleure rétention de leurs connaissances, tout en gagnant du temps (72 %) et sans contrainte technique (85 %). **La meilleure IA éducative n'est pas la plus puissante : c'est celle qui sait quand ne pas répondre.**

Discussion Critique

5.1 Limites et Biais

Bien que nos résultats soient très positifs, nous restons lucides sur les limites de notre étude. Il ne s'agit pas d'un produit fini parfait, mais d'un prototype testé dans des conditions précises.

Table 5.1 – Limites identifiées et plans d'atténuation

Limite	Description	Impact	Plan d'atténuation
Taille échantillon	Panel de 47 élèves dans un seul établissement (résultats indicatifs, validité externe limitée)	Élevé	Étude pilote multi-établissements (n>200) prévue
Biais de sélection	Élèves volontaires (potentiellement plus motivés que la moyenne)	Moyen	Recrutement aléatoire stratifié dans l'étude suivante
Taille modèles	Modèles >4B incompatibles avec smartphones <3 Go RAM	Moyen	Quantification 2-bit + modèles distillés (travaux en cours)
Couverture matières	5 matières testées (langues vivantes et arts exclus)	Moyen	Extension du corpus de System Prompts spécialisés
Qualité modèle local	~15 % moins précis que Cloud sur questions complexes	Faible	Fine-tuning sur programmes Éducation Nationale

5.2 Coût de Revient

Table 5.2 – Budget du prototype LUCID

Poste	Coût (EUR)	Note
Compte Apple Developer	0	Fourni par Lucarne Pro
Compte Google Play Developer	25	Financé par CGénial
Supabase (plan gratuit)	0	Suffisant
Cloudflare	0	Plan gratuit
Mac Mini M4 (Serveur)	600	Co-financé Lucarne PRO / CGénial
Entraînement fine-tuning	15	Financé par CGénial
TOTAL (an 1)	**640**	

5.3 Impact Écologique

Dès le début du projet, nous avons voulu que LUCID soit exemplaire sur le plan de la **sobriété numérique**. C'est un aspect qui nous tient à cœur :

- **Moins de données transportées :** en traitant l'IA sur le téléphone, on évite d'envoyer des gigaoctets de données vers des serveurs aux États-Unis. C'est une économie d'énergie invisible mais réelle.
- **Des modèles « légers » :** on utilise des modèles de 4B paramètres au lieu de modèles géants. C'est un peu comme préférer un vélo électrique à un camion pour faire ses courses : c'est plus efficace pour l'usage qu'on en a.
- **Un serveur basse consommation :** notre Mac Mini M4 ne consomme presque rien (40W max), contrairement aux énormes cartes graphiques des datacenters.
- **Donner une seconde vie aux anciens téléphones :** parce que LUCID est très optimisé, il tourne même sur des appareils qui ne sont pas de dernière génération.

→ *Ce regard critique étant posé, le chapitre suivant synthétise les apports du projet et trace les perspectives de développement à court et moyen terme.*

Conclusion & Perspectives

6.1 Synthèse de notre aventure

LUCID n'est pas qu'une simple application, c'est la preuve qu'on peut changer la façon d'apprendre :

1. **On apprend mieux** : l'approche socratique fonctionne vraiment (78 % des élèves retiennent mieux leurs cours).
2. **On est en sécurité** : vos données restent sur votre téléphone (87 % de confiance contre 19 % pour ChatGPT).
3. **On révise partout** : pas besoin d'Internet pour avoir un tuteur de qualité.
4. **C'est fluide** : le passage entre le cloud et le local se fait sans qu'on s'en rende compte.

Au final, avec une note de **4,4/5**, LUCID fait mieux que ChatGPT (2,6/5) pour les révisions. Notre conclusion est simple : **la meilleure IA n'est pas celle qui donne la réponse, c'est celle qui vous aide à la trouver**.

6.2 Et après ?

Ce premier prototype n'est que le début. Nous avons de grandes ambitions pour la suite :

- **Lancer LUCID dans toute la France** : on veut que chaque lycéen puisse utiliser cet outil pour devenir plus autonome.
- **Travailler avec l'Éducation Nationale** : on aimerait que notre approche « souveraine » et locale devienne un standard pour l'école de demain.
- **S'intégrer avec Pronote** : pour que LUCID devienne le compagnon indispensable de chaque élève, directement là où se trouvent ses notes et ses devoirs.
- **Défendre une IA éthique** : prouver qu'on peut utiliser l'IA sans sacrifier sa vie privée ou son intelligence.

Pour concrétiser ces ambitions, nous explorons actuellement de nouvelles pistes d'accompagnement. Nous sommes notamment en pleine discussion avec la branche dédiée aux startups d'**Orange Numérique**. Ce partenariat en évolution a pour objectif de nous épauler stratégiquement et techniquement afin d'accélérer le développement de LUCID, de rendre la solution accessible à tous, et de porter notre projet toujours plus loin.

Enfin, dans la dernière étape de ce projet, nous avons choisi de rédiger l'intégralité de ce rapport en LaTeX. Ce choix nous a permis de mettre en forme de manière rigoureuse l'ensemble de notre démarche, de nos résultats techniques et de nos formules mathématiques, garantissant ainsi une présentation professionnelle et structurée de notre travail.

> *« La science n'est jamais finie. LUCID est un premier pas vers une éducation où l'intelligence artificielle amplifie l'intelligence humaine au lieu de la remplacer. »*

Bibliographie & Remerciements

7.1 Références

[1] Sweller, J. (1988). Cognitive Load During Problem Solving. *Cognitive Science*, 12(2).

[2] Ebbinghaus, H. (1885). *Über das Gedächtnis*. Duncker & Humblot.

[3] Deci, E. L. & Ryan, R. M. (2000). Self-Determination Theory. *Psychological Inquiry*, 11(4).

[4] Kasneci, E. et al. (2023). ChatGPT for Good ? *Learning and Individual Differences*, 103.

[5] Maes, P. & Kosmyna, N. (2025). *Accumulation of Cognitive Debt*. MIT Media Lab.

[6] Harvard University (2025). *Impact of AI Tutors : PS2 PAL*.

[7] Deslauriers et al. (2019). Measuring actual learning. *PNAS*.

[8] Touvron, H. et al. (2024). Llama 3. *Meta AI Research*.

[9] Qwen Team, Alibaba Cloud (2024). Qwen Technical Report. *arXiv* :2407.10671.

[10] Hu, E. J. et al. (2021). LoRA : Low-Rank Adaptation of Large Language Models. *arXiv* :2106.09685.

[11] OECD (2023). PISA 2022 Results (Volume I). *OECD Publishing*.

[12] CNIL (2024). *Guide RGPD pour les développeurs*.

[13] AxePI (2026). *Demande de brevet : Architecture globale innovante*. INPI, France.

7.2 Remerciements

- **Mr Frederic TODESCHINI** : encadrant pédagogique du projet.
- **L'équipe Lucarne Pro** : accompagnement à la distribution App Store et mise à disposition du compte développeur Apple.
- **Les enseignants-chercheurs de Polytech Nice Sophia** : conseils sur l'architecture IA hybride et l'optimisation des modèles embarqués.
- **Les professeurs du Lycée Henri Matisse** : contribution à la constitution et validation du dataset pédagogique ; soutien logistique.
- **AxePI** : accompagnement juridique et dépôt de la demande de brevet, ainsi que pour l'hébergement gracieux de notre serveur sur leur infrastructure réseau.
- **La Fondation CGénial** : organisation du concours et soutien financier (subvention ayant contribué au financement du serveur).
- **Les lycéens testeurs** : retours précieux lors des phases de test utilisateur.
- **Nos familles** : soutien indéfectible tout au long du projet.

LUCID

L'assistance scolaire créée par des
élèves, pour des élèves, et leur réussite

Annexes

A.1 Glossaire Technique

Terme	Définition
API	*Application Programming Interface* : interface de communication entre logiciels.
CI/CD	*Continuous Integration / Continuous Deployment* : pipeline d'automatisation des tests, compilations et déploiements.
Cloudflare Tunnel	Service créant une connexion internet sécurisée vers un serveur local sans en exposer l'adresse IP publique.
COPPA	*Children's Online Privacy Protection Act* : loi fédérale américaine très stricte protégeant la vie privée des enfants en ligne.
Edge AI	Traitement de l'intelligence artificielle directement à la périphérie, sur l'appareil de l'utilisateur (smartphone).
EEG	*Électroencéphalographie* : méthode d'exploration de l'activité électrique du cerveau.
Fine-tuning	Paramétrage fin et ré-entraînement d'un modèle d'IA préexistant sur un jeu de données spécifique (ici éducatif).
GGUF	Format de fichier optimisé pour le chargement et l'exécution rapide de modèles d'IA locaux.
LLM	*Large Language Model* : modèle de langage de grande taille (GPT, Llama, Qwen, Gemini).
Local-First	Architecture où la majorité des données et traitements s'effectuent prioritairement sur l'appareil.
MLC LLM	*Machine Learning Compilation* : moteur universel permettant d'exécuter des LLM nativement sur mobile (NPU/GPU).
NPU	*Neural Processing Unit* : coprocesseur informatique spécialisé dans l'accélération des calculs d'intelligence artificielle.
Quantification	Technique de compression d'un modèle d'IA réduisant la précision mathématique de ses poids (ex : de FP16 vers INT4).
RGPD	*Règlement Général sur la Protection des Données* : réglementation européenne de protection de la vie privée.
RLS	*Row Level Security* : sécurisation native en base de données restreignant l'accès aux seules lignes appartenant à l'utilisateur.
Socratique	Méthode pédagogique (maïeutique) guidant l'élève par le questionnement progressif sans donner la réponse directe.
Token	Unité sémantique de base (mot ou sous-mot) traitée en entrée ou générée en sortie par un modèle de langage.
Zero Trust	Modèle de cybersécurité considérant qu'aucune requête ni entité réseau n'est digne de confiance par défaut.

Table A.1 – Glossaire des termes techniques