Jeux de données en français pour l'affinage et l'évaluation de modèles de langue génératifs dans le domaine des mathématiques

Liam Duignan[1] Asma Graiess[1] Matteo van Ypersele[2]
Jérôme Deshayes-Chossart[1] Olivier Ferret[1]
(1) Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
(2) LINAGORA, Toulouse, France

{liam.duignan,asma.graiess,jerome.deshayes-chossart,olivier.ferret}@cea.fr
mvanypersele@linagora.com

RÉSUMÉ

Que ce soit pour le post-entraînement des grands modèles de langue génératifs ou leur évaluation, les jeux de données de référence pour une tâche ou un domaine cible constituent des ressources essentielles pour le développement de ces modèles. La focalisation récente sur le raisonnement mathématique a ainsi donné lieu à la création d'un nombre important de jeux de données dans ce domaine. Néanmoins, la plupart d'entre eux sont en anglais et ceux disponibles pour d'autres langues résultent souvent d'une traduction à partir de l'anglais. Or, des études ont montré que même pour les mathématiques, les spécificités linguistiques et culturelles ont une influence notable sur les résultats des modèles, d'où l'intérêt de jeux de données natifs. Dans cet article, nous proposons ainsi d'exploiter deux sources de problèmes mathématiques en français afin de produire à la fois des données d'évaluation, sous la forme de questionnaires à choix multiples, et des données exploitables pour le post-entraînement. Nous présentons aussi les résultats de l'évaluation de différents modèles de référence sur ces données, montrant à la fois une même hiérarchie de ces modèles pour le français et l'anglais et l'importance du format d'évaluation sur les résultats. Nos données sont accessibles ici.

ABSTRACT

French dataset for fine-tuning and evaluating large language models for Mathematics.

Whether for the post-training of large language models (LLMs) or their evaluation, benchmark datasets tailored to a specific task or target domain constitute essential resources for model development. The recent focus on mathematical reasoning has consequently led to the creation of a significant number of datasets within this domain. However, the majority of these resources are in English, and those available for other languages frequently result from automated translations from English. Yet, research has demonstrated that even in mathematics, linguistic and cultural specificities exert a significant influence on model performance, thereby underscoring the importance of native datasets. In this article, we propose to exploit two sources of mathematical problems in French to produce both evaluation data, in the form of multiple-choice questions, and data suitable for post-training. We also present evaluation results for various reference models on these data, demonstrating that while the ranking of these models is comparable across French and English, the format of the evaluation significantly influences the results. Our datasets are available here.

MOTS-CLÉS : Benchmark natif en français - Évaluation des LLMs en mathématiques.

KEYWORDS: Native French Benchmark - Evaluation of LLMs in Mathematics.

1 Introduction

Le raisonnement mathématique est considéré depuis longtemps comme une forme avancée d'intelligence et son automatisation a fait l'objet de travaux remontant au moins aux années 1960 (Feigenbaum & Feldman, 1963). Le développement rapide des grands modèles de langue neuronaux (LLM) et de leurs capacités impressionnantes, en particulier par le biais de l'affinage des instructions (*instruction-tuning*), a remis sur le devant de la scène la possibilité de développer des modèles ayant des capacités d'intelligence artificielle générale (IAG), se concrétisant notamment par la possibilité de réaliser des raisonnements mathématiques. Ainsi qu'en témoignent plusieurs articles de synthèse récents (Liu *et al.*, 2025; Wang *et al.*, 2026; Yan *et al.*, 2025), ce domaine de recherche est de ce fait très actif, étant par ailleurs très connecté à la problématique, plus générale en termes de raisonnement, des chaînes de pensée (*chains-of-thought*) (Wei *et al.*, 2022). Compte tenu de la méthodologie de l'affinage des instructions, les données sous-tendant cet affinage ont une importance cruciale. Par ailleurs, le développement de ces modèles est étroitement lié à la possibilité de les évaluer. Disposer de jeux de données concernant le domaine des mathématiques, que ce soit pour le développement de modèles de raisonnement mathématique ou pour l'évaluation de ces derniers, est central et a donc fait l'objet de nombreux travaux et de nombreuses propositions, ce que caractérisent les inventaires qui ont pu en être faits (Liu *et al.*, 2025; Wang *et al.*, 2026).

Sans reprendre ces inventaires, qui doivent être étendus régulièrement, il est possible de mentionner plusieurs benchmarks de référence pour évaluer la résolution de problèmes mathématiques. L'un des plus connus est sans doute GSM8K (Cobbe *et al.*, 2021), centré sur des problèmes d'arithmétique de niveau élémentaire/collège. MATH (Hendrycks *et al.*, 2021) cible des exercices plus avancés issus de compétitions de niveau secondaire. Plus récemment, CompMath-MCQ (Raimondi *et al.*, 2026a) a étendu cette évaluation à des mathématiques de niveau supérieur. Enfin, MathNet (Alshammari *et al.*, 2026) introduit MathNET-Solve, un corpus de plus de 30 000 problèmes d'olympiades mathématiques, couvrant 47 pays, 17 langues et 143 compétitions, accompagnés de solutions rédigées par des experts. Toutefois, cette ressource reste majoritairement anglophone : plus de 74 % du corpus est en anglais, tandis que le français n'y occupe qu'une place marginale (environ 2,8 %). Ces ressources contribuent largement à l'évaluation du raisonnement mathématique des LLM mais elles restent majoritairement centrées sur l'anglais.

Face à cette prédominance anglophone, plusieurs travaux ont poussé à étendre l'évaluation des LLM à d'autres langues. L'approche la plus courante est de traduire les benchmarks existants : MGSM (Shi *et al.*, 2022a) propose une traduction manuelle de GSM8K dans dix langues, tandis que EU20-GSM8K (Thellmann *et al.*, 2024a) en propose une version traduite automatiquement couvrant vingt langues européennes. Ces ressources permettent de comparer les performances multilingues des modèles, mais elles reposent sur des énoncés initialement conçus en anglais. Elles ne reflètent donc pas nécessairement les conventions pédagogiques, la terminologie, les notations ou la progression curriculaire propres à chaque système éducatif. En ce sens, Karim *et al.* (2025) montrent que le raisonnement mathématique des LLM n'est pas culturellement neutre et que la performance peut varier selon le contexte culturel de l'énoncé. Par ailleurs, Peter *et al.* (2025) mettent en évidence, dans MGSM, plusieurs erreurs de traduction ainsi que des problèmes de formulation susceptibles d'affecter l'interprétation de certains énoncés, malgré le fait que le benchmark ait été traduit par des humains.

Les limites de ces approches par traduction soulignent l'intérêt de disposer de benchmarks éducatifs natifs dans d'autres langues que l'anglais. Dans le cas présent, nous proposons de telles données pour le français, langue pour laquelle il existe encore peu de jeux de données exploitables pour les LLM dans le domaine des mathématiques. Plus précisément, les contributions de cet article sont les suivantes :

— deux jeux de données pour l'évaluation des capacités mathématiques en français des LLM au travers de questionnaires à choix multiples (QCM) ;
— l'évaluation d'un ensemble de LLM de référence sur ces deux jeux de données pour rendre compte de leurs capacités mathématiques en français ;
— deux jeux de données constitués de problèmes mathématiques utilisables pour l'affinage des instructions.

2 État de l'art

Nous nous focalisons ici plus spécifiquement sur les jeux de données éducatifs natifs dans des langues autres que l'anglais, incluant en particulier des données mathématiques. Dans cette perspective, EXAMS (Hardalov *et al.*, 2020) est un benchmark multilingue construit à partir d'examens officiels de fin de lycée dans 16 langues et couvrant 24 disciplines. Bien qu'il ne soit pas spécifiquement centré sur les mathématiques, il met en évidence l'intérêt d'utiliser des évaluations scolaires natives, rédigées par des experts, plutôt que des traductions de benchmarks existants. Dans la même logique, MultiNRC (Fabbri *et al.*, 2025) est un benchmark natif de raisonnement multilingue composé de plus de 1 000 questions rédigées par des locuteurs natifs en français, espagnol et chinois. Les auteurs montrent que les performances des LLM restent limitées sur ce type de raisonnement multilingue natif et que les modèles obtiennent souvent de meilleurs résultats sur les versions anglaises équivalentes que sur les énoncés originaux, en particulier pour les questions mathématiques.

En italien, Puccetti *et al.* (2025) introduisent les Invalsi Benchmarks, un ensemble de trois benchmarks dont Invalsi MATE et Olimpiadi MATE, qui ont pour but l'évaluation de la compréhension mathématique des LLM, respectivement sur des évaluations nationales standardisées italiennes et des exercices plus complexes inspirés de compétitions mathématiques, couvrant tous les niveaux scolaires pré-universitaires, soit des élèves d'environ 6 à 18 ans.

En espagnol, Salido *et al.* (2025) proposent une évaluation bilingue (anglais/espagnol) sur 1 003 questions à choix multiples issues d'examens d'entrée à l'université (UNED-ACCESS) incluant plusieurs matières, y compris les mathématiques et les mathématiques appliquées aux sciences sociales, en mettant l'accent sur l'écart de performance entre langues surtout pour les petits modèles. Perez *et al.* (2025) proposent également un benchmark natif de raisonnement mathématique de niveau universitaire, composé de 105 problèmes originaux rédigés directement en espagnol.

En chinois, Xu *et al.* (2024) ont constitué SuperCLUE-Math6, un benchmark natif de raisonnement mathématique composé de 2 144 questions (1 072 paires uniques) en contexte chinois, conçu comme une version enrichie de GSM8K. Le benchmark met l'accent sur le raisonnement multi-étapes et multi-tours et vise explicitement à combler le manque de ressources d'évaluation en raisonnement mathématique pour les modèles en chinois.

Au-delà de l'évaluation, certains travaux s'intéressent également à l'adaptation de modèles sur des données non anglophones, notamment via l'affinage des instructions ou le supervised fine-tuning (SFT). Dans ce contexte, Lasbordes & Gad (2026) proposent *Luth-SFT*, un corpus de 570 000 paires d'instruction-réponse en français. Les auteurs montrent que ce corpus améliore les performances en connaissances générales, en suivi d'instructions et en raisonnement mathématique. Le sous-ensemble Scholar est construit à partir de ressources académiques françaises, notamment des sujets et corrigés du baccalauréat et des classes préparatoires, couvrant un large éventail de disciplines, avec une prédominance des mathématiques.

3 Constitution des jeux de données

Dans ce qui suit, nous commençons par décrire, pour chacun de nos deux jeux de données, les données source dont nous sommes partis, puis la construction, à partir de ces données source, de données d'évaluation (*MCQ) prenant la forme de questionnaires à choix multiples, pour finir par la constitution de données utilisables pour l'affinage de LLM (*QA).

3.1 MathALÉA

MathALÉA (CoopMaths, 2023) est un générateur d'exercices de mathématiques aligné sur le programme actuel de mathématiques en France allant du CM1 à la terminale. Le projet est porté par l'association Coopmaths et développé de manière continue par une communauté de professeurs de mathématiques. Dans MathALÉA, un exercice est un programme génératif paramétré décrivant une famille d'instances partageant la même structure pédagogique. Chaque exercice définit un espace de paramètres, des contraintes de validité, ainsi que des procédures permettant de générer l'énoncé, de calculer la solution attendue et de produire la correction. Lors de l'exécution, le moteur instancie l'exercice par tirage pseudo-aléatoire de paramètres, éventuellement contrôlé par une graine afin d'assurer la reproductibilité, puis génère une instance satisfaisant les contraintes mathématiques et didactiques spécifiées. Ce modèle permet de produire automatiquement de multiples variantes d'un même type d'exercice tout en garantissant la cohérence du contenu généré. Les exercices proposés couvrent une large diversité de thèmes : arithmétique, algèbre, géométrie, fonctions, probabilités, statistiques, etc. L'ensemble des contenus générés sont sous licence CC-BY-SA 4.0.

Pour procéder à la collecte des exercices mathématiques avec MathALÉA, nous avons déployé localement l'outil de génération MathALÉA, disponible librement sur la Forge des communs numériques éducatifs [1]. Cet outil permet de générer aléatoirement des exercices avec leur solution correspondante à travers les différents types d'exercice. Les types d'exercice sont dénotés par un identifiant (`type_id` dans ce qui suit). Pour un tel `type_id`, un ou plusieurs templates écrits par des professeurs de mathématiques forment la base de ce que l'on peut générer comme exercice. Les changements entre les instances d'exercice concernent typiquement les noms, les constantes ou les équations elles-mêmes. Dans l'objectif d'obtenir une base de données assez conséquente, du même ordre de grandeur que des jeux de données de référence en anglais comme OpenMathInstruct-1 (Toshniwal *et al.*, 2024) ou NuminaMath (LI *et al.*, 2024), nous avons généré de l'ordre de 1,3 million de paires question/solution, dont 1 915 `type_id` pour l'ensemble des niveaux dans MathALÉA. Les fichiers obtenus, au format TeX, ont été nettoyés et filtrés. Les commandes LaTeX telles que

1. https://forge.apps.education.fr/coopmaths/mathalea

`\begin{minipage}` ou `\color{...}` et tout élément inutile à la compréhension des exercices ont été enlevés à l'aide d'expressions régulières. En outre, les exercices comprenant des supports visuels ou non textuels (e.g. des graphiques tikz) ont été supprimés complètement. Les exercices de type question à choix multiples (QCM/MCQ) ont été extraits séparément, ce qui nous donne finalement les jeux de données MathAléaMCQ et MathAléaQA.

3.1.1 MathAléaMCQ

La partie QCM de MathALÉA, qui ne représente qu'une petite partie de ce que l'on peut générer via l'outil, nous a servi à la construction d'un jeu de données d'évaluation. Après nettoyage et filtrage, nous avons obtenu 7 786 exemples de type QCM, avec 117 `type_id` uniques, représentant 22 catégories (cf. tableau 5). Ce jeu de données ne couvre que cinq niveaux, de la cinquième à la terminale (sans la seconde). Cependant, on remarque un déséquilibre entre le nombre d'exemples par niveau, avec à peu près 90 % pour les niveaux première et troisième (cf. figure 5). Chaque question comprend entre deux et cinq choix, 1 948 questions (25 %) en ayant trois et 5 755 questions (74 %), quatre. Un exemple d'une question du jeu de données MathAléaMCQ est donné à la figure 1.

Question. Le cinquième d'un tiers correspond à la fraction :

(A) $\frac{1}{8}$

(B) $\frac{3}{5}$

(C) $\frac{5}{3}$

(D) $\frac{1}{15}$

Solution. Le cinquième d'un tiers est égal à $\frac{1}{5} \times \frac{1}{3}$ soit $\frac{1}{15}$. La bonne réponse est **D**.

FIGURE 1 – Exemple de question (niveau première, calcul numérique et algébrique) dans MathAléaMCQ.

3.1.2 MathAléaQA

Le reste des exemples au format simple question/solution constitue ce que nous appelons MathAléaQA, un jeu de données bien plus grand et plus adapté à l'affinage ou d'autres méthodes de post-entraînement des modèles. L'ensemble brut ainsi formé présente cependant un problème de déséquilibre de la distribution du nombre d'exercices générés selon le `type_id`. Le nombre d'exemples par `type_id` va ainsi de moins de 100 jusqu'à plus de 5 000. Cette variation s'explique par la taille de l'espace de paramètres et la sévérité des contraintes de validité propres à chaque type d'exercice : certains n'admettent qu'un faible nombre d'instances distinctes, là où d'autres en autorisent plusieurs milliers. Pour contrôler ce déséquilibre, nous limitons le nombre d'exemples par `type_id` à 1 000

(cf. figure 6). Cette valeur a été choisie afin de préserver la taille visée du jeu de données tout en réduisant la prédominance de certains types d'exercices. Un seuil nettement inférieur pourrait être envisagé afin de favoriser davantage la diversité. Après ce dernier filtrage, nous obtenons 342 718 exemples pour 1 157 `type_id` et 8 niveaux, du CM1 à terminale (+ une catégorie "autre" pour les exercices qui ne sont pas associés à un niveau spécifique ; cf. figure 7 en annexe).

Comme on peut le noter au niveau de l'exemple de la figure 2, la solution des exercices prend la forme d'un développement en langage naturel au sein duquel la réponse à la question n'est pas identifiée en tant que telle. Or, disposer de la réponse précise à la question posée est indéniablement un plus dans la perspective du post-entraînement d'un modèle comme on a pu le voir avec un modèle tel que DeepSeek (Guo *et al.*, 2025). Pour enrichir notre jeu de données, nous avons donc effectué une extraction des réponses précises à partir des solutions à l'aide du modèle `Qwen3-4B-Instruct-2507` (Yang *et al.*, 2025). Nous avons utilisé 50 exemples pour mettre au point le prompt d'extraction des réponses, avec un résultat final de 100 % de réussite sur cet ensemble (cf. figure 8 en annexe pour le prompt utilisé). La dernière étape dans la préparation de MathAléaQA a été de diviser les données en jeux d'entraînement, de test et de validation. Pour éviter toute fuite de template entre le jeu d'entraînement et celui de test, nous nous sommes assurés de ne pas avoir d'intersection entre les ensembles de `type_id` destinés à chaque sous-jeu. Le jeu d'entraînement représente environ 95 % des données avec 1 100 `type_id` uniques tandis que le jeu de test représente les 5 % restants, avec 57 `type_id`. Le jeu de validation est construit à partir de 1 % des exemples du jeu d'entraînement et partage le même ensemble de `type_id`.

Question. (t_n) est une suite géométrique de raison $q = -0{,}8$ et de premier terme $t_1 = -5$. Calculer t_{10}. Donner la valeur arrondie au dixième.

Solution. La suite (t_n) est géométrique de raison $q = -0{,}8$ et de premier terme $t_1 = -5$. On en déduit que pour tout $n \in \mathbb{N}^*$, $t_n = t_1 \times q^{n-1}$ Ainsi, $t_n = -5 \times (-0{,}8)^{n-1}$

$$t_{10} = -5 \times (-0{,}8)^9$$

$t_{10} \simeq 0{,}7$.

Réponse : $0{,}7$

FIGURE 2 – Exemple de question (niveau première) dans MathAléaQA avec réponse extraite.

3.2 Exo7

Le site Exo7 contient une collection de cours et exercices destinés aux étudiants en licence de mathématiques. Les contenus sont rédigés au format TeX par des enseignants des universités de Lille, Versailles et Toulouse et disponibles sur GitHub sous licence ouverte non commerciale : Creative Commons BY-NC-SA 4.0 FR. Comme pour MathALÉA, la ressource dispose d'exercices de type QCM ainsi que des exercices avec solution en langage naturel. Afin de créer des jeux de données exploitables, un ensemble spécifique de traitements est appliqué à chaque ensemble d'exercices.

3.2.1 Exo7MCQ

Comme pour MathAléaMCQ, l'objectif est de produire un jeu d'évaluation pour LLM à partir des QCM d'Exo7. Après un simple filtrage des questions contenant des figures, nous avons obtenu un ensemble de 946 QCM se répartissant en 28 catégories, comme les équations différentielles, la logique ou les espaces vectoriels (cf. tableau 6). Une question peut comporter entre deux et six choix, sachant qu'environ 94 % en ont quatre. Une caractéristique spécifique de ces QCM est la possibilité d'avoir plusieurs bonnes réponses possibles pour une question donnée (1,8 en moyenne) (cf. figure 3). L'exemple donné ici est typique de ce point de vue, de sorte que l'on peut considérer Exo7MCQ comme particulièrement adapté à l'évaluation de la capacité des LLM à déterminer la véracité des propositions mathématiques.

Question. Quelles sont les affirmations vraies ?

(A) x^3 est une primitive de $3x^2 + 3$.

(B) $x^3 + 3$ est une primitive de $3x^2$.

(C) $\ln(x^2 + 1)$ est une primitive de $\frac{1}{x^2+1}$.

(D) $\sqrt{x}$ est une primitive de $\frac{1}{2\sqrt{x}}$ (sur $]0, +\infty[$).

Réponses correctes : B, D

FIGURE 3 – Exemple de question dans Exo7MCQ avec plusieurs bonnes réponses.

3.2.2 Exo7QA

Comme dans le cas de MathALÉA, les exercices hors QCM d'Exo7 représentent la majorité des exercices. Nous avons extrait dans un premier temps 3 472 exercices comprenant une correction auxquels nous avons appliqué plusieurs étapes de traitement. La première étape a consisté à nettoyer les éléments de LaTeX inutiles à la compréhension. Les exercices comprenant plusieurs sous-questions ont ensuite été divisés en exemples séparés avec leur contexte partagé ajouté en en-tête. Nous avons filtré aussi les exercices intégrant des images ou des figures. Comme pour MathAléaQA, nous avons par ailleurs effectué une extraction des réponses précises avec un LLM. À l'issue de tests préliminaires, nous avons opté pour un plus gros modèle que dans le cas de MathAléaQA, `Qwen3-30B-A3B-Instruct-2507`, nécessaire pour gérer la complexité des solutions d'Exo7. Le modèle détermine aussi si la réponse extraite est *exacte* ou *descriptive*. Pour qu'une réponse soit exacte, elle ne doit pas être trop longue (> 80 caractères sans chiffres) et correspondre à une valeur ou expression mathématique précise. Les exemples sans réponse exacte sont par la suite filtrés. Parallèlement, en nous inspirant de l'approche d'OpenMathReasoning (Moshkov *et al.*, 2025), nous avons utilisé le LLM pour reformuler les exercices de type preuve en questions avec réponse vérifiable (cf. figure 9 en annexe). La dernière étape de traitement a consisté à diviser les exemples en jeux d'entraînement, de test et de validation avec les proportions respectives de 90 %, 5 % et 5 %. Lors de ce découpage, nous nous sommes assurés de l'absence de fuite entre les jeux pour les exercices multi-parties divisés. Nous obtenons 2 653 exemples dans le jeu de données final dont un exemple est donné à la figure 4.

Énoncé. Écrire sous la forme $a + ib$ les nombres complexes suivants : Nombre de module 2 et d'argument $\pi/3$.

Solution. $z_1 = 2e^{i\frac{\pi}{3}} = 2(\cos\frac{\pi}{3} + i\sin\frac{\pi}{3}) = 2(\frac{1}{2} + i\frac{\sqrt{3}}{2}) = 1 + i\sqrt{3}$.

Réponse extraite : $1 + i\sqrt{3}$

FIGURE 4 – Exemple de question dans Exo7QA avec réponse extraite.

3.3 Synthèse

Les deux sources de données exploitées, MathALÉA et Exo7, nous ont ainsi permis de produire quatre jeux de données, deux pour l'évaluation des capacités mathématiques des LLM en français et deux pour le post-entraînement de ces modèles dans le même domaine. Le tableau 1 résume la taille de ces différents jeux de données.

Jeu de données	# exemples
MathAléaMCQ	7 786
MathAléaQA	342 718
Exo7MCQ	946
Exo7QA	2 653

TABLE 1 – Taille des quatre jeux de données après filtrage.

4 Méthodologie d'évaluation

Pour évaluer des modèles sur les jeux de données MathAléaMCQ et Exo7MCQ, nous avons choisi le cadre d'évaluation LightEval (Habib *et al.*, 2023), qui permet d'évaluer des LLM sur un très large ensemble de benchmarks standard et d'ajouter facilement un nouveau jeu de données à l'aide d'un seul script.

En accord avec les pratiques standard de la littérature (Brown *et al.*, 2020; Biderman *et al.*, 2024), nous utilisons des scores de log-vraisemblance pour effectuer les évaluations de QCM en nous appuyant sur les trois formulations incluses dans LightEval : le format *multiple choice* (MCF), le format *cloze* (CF) et un format hybride (Fourrier *et al.*, 2025).

Nous implémentons également une évaluation générative, qui permet aux modèles instruits de raisonner avant de répondre. Balepur *et al.* (2025) argumentent en faveur de l'utilisation de formats génératifs pour mieux mesurer les connaissances des LLM. Cette approche apporte une dimension de robustesse vis-à-vis du format de réponse préféré des modèles, qui fait souvent défaut aux évaluations automatiques des LLM à l'aide de benchmarks QCM.

Notation. Soit q, une question de MathAléaMCQ ou Exo7MCQ, $\mathcal{C} = (c_1, \ldots, c_n)$, les choix, $\ell_i \in \{\texttt{A}, \texttt{B}, \ldots\}$, l'étiquette du i-ième choix et i^*, l'indice de la bonne réponse. Pour un modèle de paramètres θ, nous notons $\log p_\theta(t \mid s)$ la log-vraisemblance conditionnelle attribuée à une cible candidate t (une lettre ou une séquence de tokens représentant un choix) étant donné un contexte s (le prompt), et $|c_i|$ la longueur du choix c_i en caractères. Pour chaque formulation F, nous calculons le score $\mathrm{logL}_i^{(F)}$ par choix. La réponse prédite par le modèle est :

$$\hat{i}_F \;=\; \arg\max_{i \in [n]} \mathrm{logL}_i^{(F)} \tag{1}$$

Lorsque la cible est la réponse entière (formulations CF et hybride), nous appliquons une normalisation par longueur en divisant par $|c_i|$ pour réduire le biais lié à la longueur des choix :

$$\hat{i}_F \;=\; \arg\max_{i \in [n]} \frac{\mathrm{logL}_i^{(F)}}{|c_i|}, \quad F \in \{\mathrm{CF}, \mathrm{Hybrid}\} \tag{2}$$

Pour Exo7MCQ, où chaque question peut avoir jusqu'à quatre bonnes réponses, nous utilisons la même métrique que **TruthfulQA-MC** (Lin *et al.*, 2022) en calculant les scores de vraisemblance par choix mais en remplaçant la formule 1 par la masse des probabilités que le modèle donne aux choix de référence, comme détaillé en Section 4.1.

4.1 Format *multiple choice* – MCF

Dans cette formulation, le prompt donné en entrée au modèle comprend la question *et* les choix étiquetés (A, B, …). Pour chaque choix c_i, la cible est le token correspondant à son étiquette ℓ_i :

$$\mathrm{logL}_i^{(\mathrm{MCF})} \;=\; \log p_\theta\big(\ell_i \,\big|\, q, \mathcal{C}\big) \tag{3}$$

Pour MathAléaMCQ, la réponse prédite du modèle est calculée via la formule 1. Pour Exo7MCQ, comme plusieurs choix peuvent être corrects, nous calculons la masse de probabilité attribuée à l'ensemble des choix corrects :

$$\mathrm{score}_{\mathrm{Exo7}}^{(\mathrm{MCF})} \;=\; \sum_{i:\, y_i = 1} \frac{\exp\big(\log p_\theta(\ell_i \mid q, \mathcal{C})\big)}{\sum_j \exp\big(\log p_\theta(\ell_j \mid q, \mathcal{C})\big)} \tag{4}$$

où $y \in \{0,1\}^n$ est le vecteur *multi-hot* des indices de référence où $y_i = 1$ si et seulement si le choix i est correct.

4.2 Format hybride

Dans cette formulation, le prompt est identique à celui de MCF mais le score de vraisemblance est calculé sur la réponse entière (tous les tokens) du choix :

$$\mathrm{logL}_i^{(\mathrm{Hybrid})} \;=\; \log p_\theta\big(c_i \,\big|\, q, \mathcal{C}\big) \tag{5}$$

Pour MathAléaMCQ, la réponse prédite du modèle est calculée via la formule normalisée 2. Pour Exo7MCQ, nous appliquons la normalisation par la longueur avant de calculer la masse des probabilités sur les choix corrects :

$$\text{score}_{\text{Exo7}}^{(\text{Hybrid})} = \sum_{i:\, y_i=1} \frac{\exp\left(\frac{\log p_\theta(c_i|q,\mathcal{C})}{|c_i|}\right)}{\sum_j \exp\left(\frac{\log p_\theta(c_j|q,\mathcal{C})}{|c_j|}\right)} \tag{6}$$

4.3 Format cloze – CF

Dans cette formulation, les choix sont absents du prompt et les scores calculés via la vraisemblance de chaque réponse entière comme continuation immédiate de la question :

$$\log \text{L}_i^{(\text{CF})} = \log p_\theta\left(c_i \,\middle|\, q\right) \tag{7}$$

Pour MathAléaMCQ, la réponse prédite du modèle est calculée via la formule normalisée 2. Vu les caractéristiques des questions ouvertes/génériques dans Exo7MCQ, nous n'utilisons pas cette formulation pour ce benchmark.

4.4 Évaluation générative

Dans cette formulation, nous demandons au modèle de générer une réponse à la question en terminant par `Réponse : <LETTRE>`. La lettre prédite $\hat{\ell}$ est ensuite extraite de sa génération $\hat{r}$ par un extracteur à base d'expressions régulières Ex :

$$\hat{\ell} = \text{Ex}(\hat{r}) \tag{8}$$

La réponse prédite est correcte lorsque $\hat{\ell} = \ell_{i^*}$. L'extracteur tient compte également des réponses générées dans un environnement `\boxed{}`.

Pour Exo7MCQ, le modèle est supposé générer un ensemble de lettres séparées par des virgules (e.g. `Réponse : A, B`). L'extracteur renvoie un ensemble $\hat{S} \subseteq \{\text{A}, \text{B}, \dots\}$ et nous calculons le score F1 entre $\hat{S}$ et l'ensemble de référence $S^* = \{\ell_i : y_i = 1\}$.

4.5 Sélection de modèles

Dans une optique de transparence et de reproductibilité, nous effectuons une évaluation sur cinq LLM populaires, à poids ouverts (*open-weight*) et de taille restant compacte :

- `Llama3.1-8B-Instruct` (Grattafiori *et al.*, 2024) ;
- `EuroLLM-9B-Instruct` (Martins *et al.*, 2025) ;
- `Mistral-7B-Instruct-v0.3` (Jiang *et al.*, 2023) ;
- `Qwen2.5-7B-Instruct` (Team, 2024) ;
- `Qwen2.5-Math-7B-Instruct` (Yang *et al.*, 2024) ;

Modèle	MathAléaMCQ				Exo7MCQ			GSM8K
	MCF	Hybride	CF	Générative	MCF	Hybride	Générative	
Qwen2.5-7B-Instruct	37,3	47,8	51,6	**95,0**	48,6	**51,2**	78,7	82,6
Qwen2.5-Math-7B-Instruct	31,9	**49,8**	**51,8**	83,0	46,2	48,8	**83,0**	**95,1**
EuroLLM-9B-Instruct	30,2	36,8	42,5	55,3	44,5	47,3	46,3	65,0
Llama-3.1-8B-Instruct	31,1	40,5	48,0	75,9	46,5	48,0	54,0	78,2
Mistral-7B-Instruct-v0.3	**38,8**	41,8	49,1	44,5	**49,5**	48,5	49,4	49,2
Lucie-7B-Instruct-v1.1	31,6	32,3	36,3	22,7	46,4	46,5	44,7	44,8

TABLE 2 – Performances (x100) sur MathAléaMCQ et Exo7MCQ ainsi que sur GSM8K (0-shot) comme référence anglaise. Pour Exo7MCQ, seule la mesure F1 est donnée ici, les résultats de la mesure EM (Exact Match) étant présentés en annexe dans le tableau 8.

	MathAléaMCQ	**Exo7MCQ**
MCF – hybride	0,40	0,58
MCF – CF	0,49	NA
MCF – générative	0,13	0,20
MCF – GSM8K	-0,16	-0,16
générative – GSM8K	0,92	0,83

TABLE 3 – Corrélation entre les différentes formulations au sein de chaque jeu de données (coefficient de Pearson).

— `Lucie-7B-Instruct-v1.1` (Gouvert *et al.*, 2025).

Nous évaluons également ces modèles sur GSM8K pour avoir une référence générale de leur performance en mathématiques en anglais, la mesure d'évaluation étant comparable à celle utilisée pour MathAléaMCQ en mode génératif. Toutes les évaluations s'effectuent en *zero-shot* avec une température égale à 0. Conformément au comportement par défaut de LightEval, nous appliquons le modèle d'instruction (*chat template*) de chaque modèle aux prompts.

5 Résultats et discussion

Les résultats de nos expériences pour les différentes formulations retenues sont présentés dans le tableau 2 et la corrélation de ces résultats pour les différents modèles est évaluée dans le tableau 3. Une observation remarquable est que la corrélation entre le classement des modèles en MCF et leur

	MathAléaMCQ – Exo7MCQ
MCF	0,96
hybride	0,83
générative	0,83

TABLE 4 – Corrélation entre les deux jeux de données pour chaque formulation (coefficient de Pearson).

capacité à résoudre des exercices mathématiques de manière générative est faible. Cela est notamment illustré par le fait que `Mistral-7B-Instruct-v0.3` obtient les meilleurs scores MCF tandis qu'il se situe assez bas dans les classements des évaluations génératives. D'autre part, l'éventail des scores MCF (MathAléaMCQ : 8,6 points ; Exo7MCQ : 5 points) reste bien inférieur à celui des scores en génération (MathAléaMCQ : 72,3 points ; Exo7MCQ : 38,3 points). La formulation MCF n'est donc pas une mesure fiable des capacités mathématiques des modèles instruits. Cette différence peut s'expliquer par le fait que les modèles sont affinés pour un usage conversationnel et ont tendance à détailler leur raisonnement avant de fournir la réponse finale.

Concernant l'évaluation générative, nous observons une corrélation plus marquée avec le classement donné par GSM8K. L'éventail des scores est bien plus proche de celui de GSM8K (50,3 points). `Qwen2.5-Math-7B-Instruct` arrive en tête pour les formulations hybride et CF de MathAléaMCQ et l'évaluation générative d'Exo7MCQ. Cela confirme qu'un modèle adapté au raisonnement mathématique en anglais et chinois peut obtenir de bonnes performances sur des exercices de mathématiques en français. Sa performance reste stable entre MathAléaMCQ et Exo7MCQ, ce qui suggère un post-entraînement axé sur les mathématiques complexes et de niveau supérieur. On peut toutefois noter que l'adaptation aux mathématiques du modèle `Qwen2.5-Math-7B-Instruct` ne lui permet pas de dépasser systématiquement le modèle `Qwen2.5-7B-Instruct`, son alter ego non adapté. Nous observons également qu'`EuroLLM-9B-Instruct`, un modèle plus gros et centré sur les langues européennes comme le français, n'obtient pas de résultats compétitifs sur nos benchmarks, ce qui suggère un post-entraînement moins axé sur le raisonnement mathématique.

Bien que le tableau 4 montre une corrélation importante entre les résultats des deux jeux de données pour les différentes formulations, ceux-ci se caractérisent également par une certaine diversité. On observe ainsi qu'Exo7MCQ est bien plus difficile pour certains modèles en évaluation générative (`Qwen2.5-7B-Instruct`, `EuroLLM-9B-Instruct` et `Llama-3.1-8B-Instruct`) que MathAléaMCQ. Cette difficulté s'explique non seulement par son contenu plus avancé mais aussi par le fait que le modèle évalué doit renvoyer l'ensemble des propositions correctes pour obtenir le maximum des points possibles. À l'inverse, `Mistral-7B-Instruct-v0.3` présente des performances relativement homogènes au travers des trois évaluations génératives.

Comme MathAléaMCQ est dominé par les niveaux troisième et première, nous avons également effectué une évaluation par niveau pour avoir une vision plus précise des résultats en les agrégeant selon une macro-moyenne, les résultats du tableau 2 étant des micro-moyennes (cf. tableau 7 en annexe). Cette analyse permet de constater en particulier que `Qwen2.5-7B-Instruct` est très nettement le modèle le plus stable entre macro et micro-moyennes pour toutes les formulations et que CF est à l'inverse la formulation largement la moins stable entre macro et micro-moyennes pour tous les modèles, la formulation hybride étant la plus stable.

6 Limitations

Le travail que nous présentons ici comporte un certain nombre de limitations. Un certain nombre d'entre elles sont associées à notre méthodologie d'évaluation. Ainsi, concernant le format génératif, notre métrique repose sur une extraction heuristique de la réponse. Même si le modèle a correctement raisonné dans sa solution générée, s'il ne respecte pas le format spécifié dans le prompt, sa réponse ne sera pas considérée comme correcte. Une façon d'apporter plus de flexibilité à ce niveau pourrait être d'adopter une approche fondée sur le paradigme *LLM-as-a-judge*, toutefois plus coûteuse du point de

vue calculatoire. En termes de modèles testés, nous n'avons pas inclus comme référence les modèles à poids fermés dans nos évaluations. Les prendre en compte donnerait certainement une meilleure vision de l'état de l'art actuel en raisonnement mathématique en français, en particulier sur le plan des performances hautes.

Concernant les jeux de données *QA destinés à l'affinage, des limites potentielles associées à l'utilisation d'un LLM pour l'extraction des réponses existent également. Une évaluation humaine plus approfondie avec le calcul d'un score d'accord inter-annotateur pourrait être intéressante afin d'éliminer les hallucinations dans les solutions de référence.

7 Conclusion et perspectives

Dans ce travail, nous avons présenté deux jeux de données en français natif pour l'évaluation des LLM dans le domaine des mathématiques (MathAléaMCQ et Exo7MCQ) et deux jeux de données destinés à un usage de post-entraînement (MathAléaQA et Exo7QA). Toutes les données viennent de ressources éducatives libres, mises à disposition sous licence ouverte *Creative Commons*. Nous avons détaillé les étapes de traitement, de la collecte des données à leur nettoyage et filtrage afin d'obtenir des ressources exploitables. Nous avons également exploré les différentes formulations possibles pour évaluer des LLM sur les benchmarks de type QCM. Sur la base de ces formats, nous avons effectué une évaluation de cinq LLM à poids ouverts de taille 7B-9B. Nous avons observé un écart important en performance entre les différentes formulations pour un même modèle. Suite à cette évaluation, nous avons discuté de l'utilité de favoriser une évaluation générative pour les modèles instruits, en constatant une corrélation entre leur capacité à raisonner en français et en anglais.

Par la suite, nous comptons exploiter nos jeux de données MathAléaQA et Exo7QA pour mener des expériences d'affinage supervisé. En particulier, nous souhaiterions étudier dans quelle mesure ces jeux de données permettraient d'améliorer les performances des moins bons modèles pour les mathématiques en français et de les hisser au niveau des meilleurs modèles dans ce domaine.

Remerciements

Ces travaux ont été financièrement soutenus par le projet FRANCE 2030 OpenLLM-France. Ils ont été réalisés grâce au supercalculateur Factory-IA, financé par le Conseil Régional d'Île-de-France, et ont également bénéficié d'un accès aux moyens de calcul de l'IDRIS au travers de l'allocation de ressources AD011016978 attribuée par GENCI.

Références

ALSHAMMARI S., WEN K., ZAINAL A., HAMILTON M., SAFAEI N., ALBARAKATI S., FREEMAN W. T. & TORRALBA A. (2026). MathNet : A Global Multimodal Benchmark for Mathematical Reasoning and Retrieval. In *The Fourteenth International Conference on Learning Representations*.

BALEPUR N., RUDINGER R. & BOYD-GRABER J. L. (2025). Which of These Best Describes Multiple Choice Evaluation with LLMs ? A) Forced B) Flawed C) Fixable D) All of the Above. In W. CHE, J. NABENDE, E. SHUTOVA & M. T. PILEHVAR, Éds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 3394–3418, Vienna, Austria : Association for Computational Linguistics. DOI : 10.18653/v1/2025.acl-long.169.

BIDERMAN S., SCHOELKOPF H., SUTAWIKA L., GAO L., TOW J., ABBASI B., AJI A. F., AMMANAMANCHI P. S., BLACK S., CLIVE J., DIPOFI A., ETXANIZ J., FATTORI B., FORDE J. Z., FOSTER C., HSU J., JAISWAL M., LEE W. Y., LI H., LOVERING C., MUENNIGHOFF N., PAVLICK E., PHANG J., SKOWRON A., TAN S., TANG X., WANG K. A., WINATA G. I., YVON F. & ZOU A. (2024). Lessons from the trenches on reproducible evaluation of language models.

BROWN T. B., MANN B., RYDER N., SUBBIAH M., KAPLAN J., DHARIWAL P., NEELAKANTAN A., SHYAM P., SASTRY G., ASKELL A., AGARWAL S., HERBERT-VOSS A., KRUEGER G., HENIGHAN T., CHILD R., RAMESH A., ZIEGLER D. M., WU J., WINTER C., HESSE C., CHEN M., SIGLER E., LITWIN M., GRAY S., CHESS B., CLARK J., BERNER C., MCCANDLISH S., RADFORD A., SUTSKEVER I. & AMODEI D. (2020). Language models are few-shot learners.

COBBE K., KOSARAJU V., BAVARIAN M., CHEN M., JUN H., KAISER L., PLAPPERT M., TWOREK J., HILTON J., NAKANO R. *et al.* (2021). Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

COOPMATHS (2023). MathALEA : Générateur d'exercices de mathématiques. Accessed : 2026-01-29.

DU W., TOSHNIWAL S., KISACANIN B., MAHDAVI S., MOSHKOV I., ARMSTRONG G., GE S., MINASYAN E., CHEN F. & GITMAN I. (2025). Nemotron-math : Efficient long-context distillation of mathematical reasoning from multi-mode supervision.

FABBRI A. R., MARES D., FLORES J., MANKIKAR M., HERNANDEZ E., LEE D., LIU B. & XING C. (2025). MultiNRC : A Challenging and Native Multilingual Reasoning Evaluation Benchmark for LLMs. *arXiv preprint arXiv:2507.17476*.

FEIGENBAUM E. A. & FELDMAN J. (1963). *Computers and Thought*. McGraw-Hill.

FOURRIER C., FRERE T., PENEDO G. & WOLF T. (2025). The LLM Evaluation Guidebook.

GOUVERT O., HUNTER J., LOURADOUR J., CERISARA C., DUFRAISSE E., SY Y., RIVIÈRE L., LORRÉ J.-P. & COMMUNITY O.-F. (2025). The lucie-7b llm and the lucie training dataset : Open resources for multilingual language generation.

GRATTAFIORI A., DUBEY A., JAUHRI A., PANDEY A., KADIAN A., AL-DAHLE A., LETMAN A., MATHUR A., SCHELTEN A., VAUGHAN A. *et al.* (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

GUO D., YANG D., ZHANG H., SONG J., WANG P., ZHU Q., XU R., ZHANG R., MA S. & BI X. E. A. (2025). Deepseek-r1 : Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

HABIB N., FOURRIER C., KYDLÍČEK H., WOLF T. & TUNSTALL L. (2023). LightEval : A lightweight framework for LLM evaluation.

HARDALOV M., MIHAYLOV T., ZLATKOVA D., DINKOV Y., KOYCHEV I. & NAKOV P. (2020). EXAMS : A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 5427–5444.

HENDRYCKS D., BURNS C., KADAVATH S., ARORA A., BASART S., TANG E., SONG D. & STEINHARDT J. (2021). Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

JIANG A. Q., SABLAYROLLES A., MENSCH A., BAMFORD C., CHAPLOT D. S., DE LAS CASAS D., BRESSAND F., LENGYEL G., LAMPLE G., SAULNIER L., LAVAUD L. R., LACHAUX M.-A., STOCK P., SCAO T. L., LAVRIL T., WANG T., LACROIX T. & SAYED W. E. (2023). Mistral 7b.

JU H. & DONG B. (2026). AI for Mathematics : Progress, Challenges, and Prospects. *arXiv preprint arXiv:2601.13209*.

KARIM A., KARIM A., LOHANA B., KEON M., SINGH J. & SATTAR A. (2025). Lost in Cultural Translation : Do LLMs Struggle with Math Across Cultural Contexts ? *arXiv preprint arXiv:2503.18018*.

LASBORDES M. & GAD S. (2026). Luth : Efficient french specialization for small language models and cross-lingual transfer. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 4 : Student Research Workshop)*, p. 48–59.

LI J., BEECHING E., TUNSTALL L., LIPKIN B., SOLETSKYI R., HUANG S. C., RASUL K., YU L., JIANG A., SHEN Z., QIN Z., DONG B., ZHOU L., FLEUREAU Y., LAMPLE G. & POLU S. (2024). NuminaMath. [https://huggingface.co/AI-MO/NuminaMath-CoT](https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf).

LIN S., HILTON J. & EVANS O. (2022). TruthfulQA : Measuring How Models Mimic Human Falsehoods. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 3214–3252, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.229.

LIU W., HU H., ZHOU J., DING Y., LI J., ZENG J., HE M., CHEN Q., JIANG B., ZHOU A. & HE L. (2025). Mathematical Language Models : A Survey. *ACM Computing Survey*, **58**(6), 146 :1–146 :37. DOI : 10.1145/3773985.

MARTINS P. H., ALVES J., FERNANDES P., GUERREIRO N. M., REI R., FARAJIAN A., KLIMASZEWSKI M., ALVES D. M., POMBAL J., BOIZARD N. *et al.* (2025). Eurollm-9b : Technical report. *arXiv preprint arXiv:2506.04079*.

MOSHKOV I., HANLEY D., SOROKIN I., TOSHNIWAL S., HENKEL C., SCHIFFERER B., DU W. & GITMAN I. (2025). Aimo-2 winning solution : Building state-of-the-art mathematical reasoning models with openmathreasoning dataset.

PEREZ M. A. P., OROZCO B. L., SOTO J. T. C., HERNANDEZ M. B., GONZALEZ M. A. A. & MALAGON S. (2025). Ai4math : A native spanish benchmark for university-level mathematical reasoning in large language models. *arXiv preprint arXiv:2505.18978*.

PETER J.-T., VILAR D., DOMHAN T., MALKIN D. & FREITAG M. (2025). Mind the Gap... or Not ? How Translation Errors and Evaluation Details Skew Multilingual Results. *arXiv preprint arXiv:2511.05162*.

PUCCETTI G., CASSESE M. & ESULI A. (2025). The invalsi benchmarks : measuring the linguistic and mathematical understanding of large language models in Italian. In *Proceedings of the 31st International Conference on Computational Linguistics*, p. 6782–6797.

RAIMONDI B., PIVI F., EVANGELISTA D. & GABBRIELLI M. (2026a). The CompMath-MCQ Dataset : Are LLMs Ready for Higher-Level Math ? *arXiv preprint arXiv:2603.03334*.

RAIMONDI B., PIVI F., EVANGELISTA D. & GABBRIELLI M. (2026b). The compmath-mcq dataset : Are llms ready for higher-level math ?

SALIDO E. S., MORANTE R., GONZALO J., MARCO G., DE ALBORNOZ J. C., PLAZA L., AMIGÓ E., GARCÍA A. F., BENITO-SANTOS A., ESPINOSA A. G. *et al.* (2025). Bilingual evaluation of language models on general knowledge in university entrance exams with minimal contamination. In *Proceedings of the 31st International Conference on Computational Linguistics*, p. 6184–6200.

SHI F., SUZGUN M., FREITAG M., WANG X., SRIVATS S., VOSOUGHI S., CHUNG H. W., TAY Y., RUDER S., ZHOU D. *et al.* (2022a). Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

SHI F., SUZGUN M., FREITAG M., WANG X., SRIVATS S., VOSOUGHI S., CHUNG H. W., TAY Y., RUDER S., ZHOU D., DAS D. & WEI J. (2022b). Language models are multilingual chain-of-thought reasoners.

TEAM Q. (2024). Qwen2.5 : A Party of Foundation Models.

THELLMANN K., STADLER B., FROMM M., BUSCHHOFF J. S., JUDE A., BARTH F., LEVELING J., FLORES-HERR N., KÖHLER J., JÄKEL R. *et al.* (2024a). Towards multilingual llm evaluation for european languages. *arXiv preprint arXiv:2410.08928*.

THELLMANN K., STADLER B., FROMM M., BUSCHHOFF J. S., JUDE A., BARTH F., LEVELING J., FLORES-HERR N., KÖHLER J., JÄKEL R. & ALI M. (2024b). Towards multilingual llm evaluation for european languages.

TOSHNIWAL S., MOSHKOV I., NARENTHIRAN S., GITMAN D., JIA F. & GITMAN I. (2024). OpenMathInstruct-1 : A 1.8 Million Math Instruction Tuning Dataset. *arXiv preprint arXiv:Arxiv-2402.10176*.

WANG P.-Y., LIU T.-S., WANG C., LI Z., WANG Y., YAN S., JIA C., LIU X.-H., CHEN X., XU J. & YU Y. (2026). A Survey on Large Language Models for Mathematical Reasoning. *ACM Computing Survey*, **58**(8), 209 :1–209 :35. DOI : 10.1145/3786333.

WEI J., WANG X., SCHUURMANS D., BOSMA M., XIA F., CHI E., LE Q. V. & ZHOU D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, **35**, 24824–24837.

XU L., XUE H., ZHU L. & ZHAO K. (2024). Superclue-math6 : Graded multi-step math reasoning benchmark for LLMs in Chinese. *arXiv preprint arXiv:2401.11819*.

YAN Y., SU J., HE J., FU F., ZHENG X., LYU Y., WANG K., WANG S., WEN Q. & HU X. (2025). A Survey of Mathematical Reasoning in the Era of Multimodal Large Language Model : Benchmark, Method & Challenges. In W. CHE, J. NABENDE, E. SHUTOVA & M. T. PILEHVAR, Éds., *Findings of the Association for Computational Linguistics : ACL 2025*, p. 11798–11827, Vienna, Austria : Association for Computational Linguistics. DOI : 10.18653/v1/2025.findings-acl.614.

YANG A., LI A., YANG B., ZHANG B., HUI B., ZHENG B., YU B., GAO C., HUANG C., LV C., ZHENG C., LIU D., ZHOU F., HUANG F., HU F., GE H., WEI H., LIN H., TANG J., YANG J., TU J., ZHANG J., YANG J., YANG J., ZHOU J., ZHOU J., LIN J., DANG K., BAO K., YANG K., YU L., DENG L., LI M., XUE M., LI M., ZHANG P., WANG P., ZHU Q., MEN R., GAO R., LIU S., LUO S., LI T., TANG T., YIN W., REN X., WANG X., ZHANG X., REN X., FAN Y., SU Y., ZHANG Y., ZHANG Y., WAN Y., LIU Y., WANG Z., CUI Z., ZHANG Z., ZHOU Z. & QIU Z. (2025). Qwen3 technical report.

YANG A., ZHANG B., HUI B., GAO B., YU B., LI C., LIU D., TU J., ZHOU J., LIN J. *et al.* (2024). Qwen2. 5-math technical report : Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

A Statistiques sur les données

Catégorie	Nombre
Arithmétique	103
Calcul littéral	438
Calcul numérique et algébrique	2 664
Combinatoires et dénombrement	101
Espace	122
Évolutions et variations	642
Fonctions affines et linéaires	69
Fonctions et représentations	616
Fonctions logarithme népérien	10
Fractions	164
Généralités sur les fonctions	124
Nombres complexes algèbre	19
Nombres réels	21
Primitives et équations différentielles	73
Probabilités	343
Proportionnalité	77
Proportions et pourcentages	508
Puissances	643
Statistiques	525
Suites numériques	224
Théorème de Thalès	9
Variables aléatoires discrètes finies	288
Total	**7 786**

TABLE 5 – Distribution des questions par catégorie dans MathAléaMCQ.

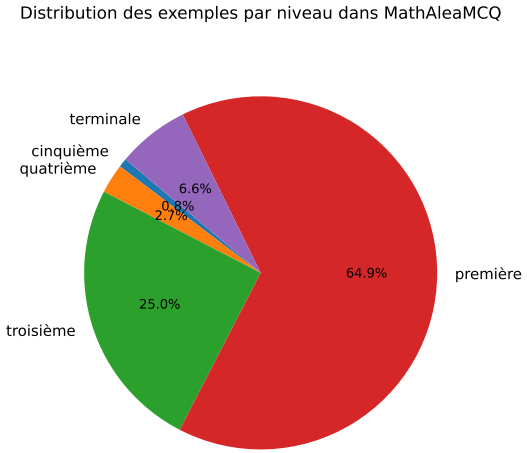

FIGURE 5 – Distribution des questions par niveau dans MathAléaMCQ.

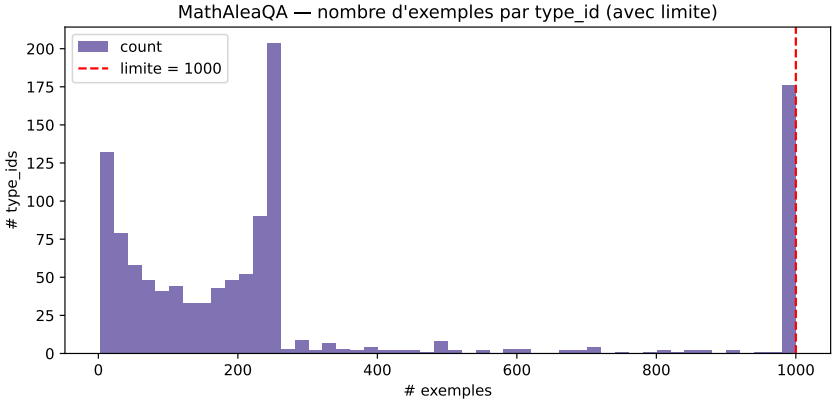

FIGURE 6 – Distribution du nombre d'exemples par `type_id` dans MathAléaQA après application d'un seuil de 1 000.

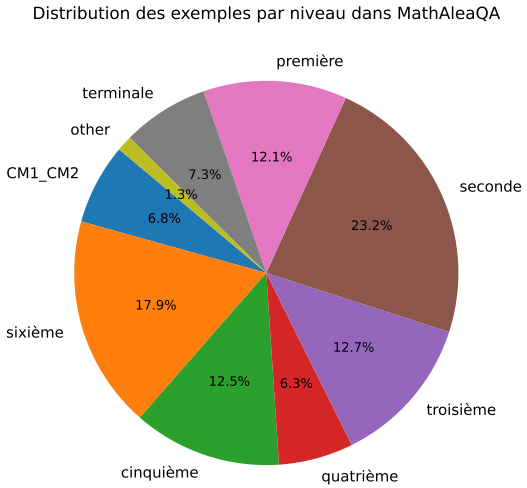

FIGURE 7 – Distribution des questions par niveau dans MathAléaQA.

Catégorie	**Nombre**
Équations différentielles	87
Logique, ensembles et raisonnements	56
Espaces vectoriels	56
Arithmétique	52
Primitives des fonctions réelles	50
Développements limités	47
Calculs d'intégrales	37
Calcul matriciel	34
Applications linéaires	33
Applications linéaires et matrices	30
QCM de révisions (Arnaud)	30
Polynômes – Fractions rationnelles	30
Nombres complexes	30
Limites des fonctions réelles	30
Ensembles, applications	30
Fonctions usuelles	30
Suites réelles	30
Réels	30
Logique – Raisonnement	30
Géométrie du plan	30
Dérivabilité des fonctions réelles	29
Continuité	29
Systèmes d'équations linéaires	26
Géométrie dans l'espace	24
Probabilités, événements	20
Variables discrètes	16
Courbes paramétrées	12
Variables continues	8
Total	**946**

TABLE 6 – Distribution des questions par catégorie dans Exo7MCQ.

B Prompts d'extraction de réponses

System prompt

```
You are an expert at extracting final answers from math solutions written in
     LaTeX.

Given a math solution, extract ONLY the final answer -- the conclusion or
    result that answers the original question.

Rules:
1. Return ONLY the answer, nothing else -- no explanation, no "The answer is
    "
2. Keep LaTeX formatting (e.g., \dfrac{1}{2}, x^2, \sqrt{3})
3. If there are multiple answers (e.g., two roots), return them separated by
     semicolons
4. Include units if they are part of the answer (e.g., "13 EUR", "57 km")
5. If the answer is a sentence or phrase, return it as-is
6. Do not generate any empty response; the solution will always contain a
    specific final answer to be extracted.
```

User template

```
Extract the final answer from this math problem.

Question: {question}

Solution: {solution}
```

FIGURE 8 – Prompts utilisés pour l'extraction de réponses à partir des solutions de MathAléaQA. Le *System prompt* est le contenu associé au rôle `system` tandis que le *User template* est le contenu associé au rôle `user` selon le format standard *chat-messages*.

System prompt

```
Transform the proof into a DIFFERENT question type that has a concrete, \
verifiable answer. The reformulated question must NOT use "Montrer que", \
"Démontrer que", or "Prouver que". Keep it in French, using $...$ for all
    math. \
Common transformations:
   - "Montrer que X = Y" -> "Calculer X" (answer: $Y$)
   - "Montrer que f est continue" -> "Déterminer l'ensemble de continuité de
     f"
   - "Montrer que la suite converge" -> "Calculer la limite de la suite"
   - "Montrer que E est compact" -> "La partie E est-elle compacte ?
    Justifier."
   - "Montrer que G est isomorphe à H" -> "\'A quel groupe G est-il
    isomorphe ?"
   - "Montrer qu'il existe..." -> "Déterminer..." or "Trouver..."
   If the proof cannot be meaningfully reformulated (e.g. pure existence
    proofs \
with no concrete answer), set to null instead of restating the proof.
```

FIGURE 9 – Partie du *System prompt* utilisé pour la reformulation d'exercices de type preuve avec `Qwen3-30B-A3B-Instruct-2507` pour Exo7QA.

C Détail des résultats pour MathAléaMCQ

Modèle	Formulation	5^{e}	4^{e}	3^{e}	$1^{ère}$	T^{ale}	Micro	Macro
Qwen2.5-7B-Instruct	MCF	32,2	33,5	**48,8**	27,7	**34,9**	37,3	35,4
	CF	91,5	59,8	58,3	**47,7**	54,0	51,6	62,3
	Hybride	16,9	47,4	61,4	42,2	54,8	47,8	44,5
	Générative	**98,3**	**93,3**	**97,5**	**94,1**	**93,1**	**95,0**	**95,3**
Qwen2.5-Math-7B-Instruct	MCF	42,4	33,5	35,3	31,1	23,7	31,9	33,2
	CF	98,3	**60,8**	61,2	45,5	**69,6**	**51,8**	**67,1**
	Hybride	23,7	**54,1**	**67,2**	**42,2**	**60,7**	**49,8**	**49,6**
	Générative	89,8	90,0	90,7	79,2	86,4	83,0	87,2
Llama-3.1-8B-Instruct	MCF	27,1	26,3	41,6	29,6	33,5	31,1	31,6
	CF	**100,0**	58,9	**64,4**	40,0	51,3	48,0	62,9
	Hybride	10,2	41,6	55,0	34,6	48,7	40,5	38,0
	Générative	91,5	84,7	87,8	74,0	51,9	75,9	78,0
EuroLLM-9B-Instruct	MCF	25,4	**34,0**	37,1	25,2	33,9	30,2	31,1
	CF	88,1	34,0	56,5	34,6	65,1	42,5	55,7
	Hybride	22,0	37,3	47,9	30,5	50,1	36,8	37,6
	Générative	67,8	67,9	72,1	51,6	44,4	55,3	60,8
Mistral-7B-Instruct-v0.3	MCF	**89,8**	23,9	48,0	**36,7**	20,9	**38,8**	**43,9**
	CF	79,7	42,6	62,3	43,4	49,5	49,1	55,5
	Hybride	**72,9**	41,1	49,1	39,7	28,4	41,8	46,2
	Générative	49,2	44,0	58,7	39,8	29,0	44,5	44,1
Lucie-7B-Instruct-v1.1	MCF	78,0	33,0	41,3	28,8	22,3	31,6	40,7
	CF	**100,0**	43,1	47,2	29,4	51,9	36,2	54,3
	Hybride	15,3	34,9	42,7	27,1	48,1	32,5	33,6
	Générative	25,4	21,1	21,5	24,0	20,5	22,7	22,5

TABLE 7 – Performances (x100) sur MathAléaMCQ par niveau scolaire et formulation en *zero-shot*. Les niveaux : cinquième/5^{e} (N=59), quatrième/4^{e} (209), troisième/3^{e} (1 942), première/$1^{ère}$ (5 044) et terminale/T^{ale} (507). **Micro** est la micro-moyenne calculée sur les 7 761 exemples dans l'ensemble du jeu de données (cinq exemples par sous-ensemble sont réservés pour la possibilité de faire une évaluation en **few-shot**). **Macro** est la moyenne non pondérée des cinq scores par niveau.

D Résultats pour Exo7MCQ selon les différentes mesures

Pour des raisons de place, le tableau 2 de synthèse des résultats ne donne que la mesure F1 pour la tâche Exo7MCQ mais la mesure EM (Exact Match), plus stricte, est également classiquement utilisée pour ce genre de tâches en considérant, puisque plusieurs réponses sont parfois possibles, que l'appariement doit se faire pour toutes les réponses données avec les réponses de référence pour qu'une question soit considérée comme résolue. Il faut cependant noter que les mesures F1 et EM sont presque parfaitement corrélées pour les modèles sélectionnés ici puisque la mesure de corrélation de Pearson entre les deux mesures est égale à 0,99.

Modèle	**EM**	**F1**
Qwen2.5-7B-Instruct	54,1	78,7
Qwen2.5-Math-7B-Instruct	61,4	83,0
EuroLLM-9B-Instruct	12,5	46,3
Llama-3.1-8B-Instruct	25,0	54,0
Mistral-7B-Instruct-v0.3	12,9	49,4
Lucie-7B-Instruct-v1.1	5,2	44,7

TABLE 8 – Performance (x100) des différents modèles sélectionnés sur la tâche Exo7MCQ avec les deux mesures d'évaluation possibles, EM (Exact Match) et F1.