Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

Évaluation de la cohérence des modèles vision-langage pour la tâche de question-réponse visuelle

Khanh-An C. Quan[1,2] * Camille Guinaudeau[3] Shin'ichi Satoh[4]
(1) University of Information Technology, VNU-HCM, Ho Chi Minh City, Vietnam
(2) Vietnam National University, Ho Chi Minh City, Vietnam
(3) Université Paris-Saclay / LISN-CNRS, Orsay, France
(4) National Institute of Informatics, Tokyo, Japan
anqck@uit.edu.vn, guinaudeau@lisn.fr, satoh@nii.ac.jp

RÉSUMÉ

La tâche de question-réponse visuelle constitue un défi nécessitant une compréhension approfondie de l'image et de la question associée. Bien que les modèles vision-language récents obtiennent des performances élevées, ils révèlent des comportements incohérents lorsque des modifications mineures sont appliquées à la question. Dans cet article, nous proposons deux protocoles d'évaluation centrés sur la compréhension des questions : la permutation des choix et la reformulation des questions. Ces approches permettent de tester la robustesse et la cohérence des modèles au-delà de la seule mesure d'exactitude. Nos résultats expérimentaux montrent que LLaVA et MM-CoT produisent fréquemment des réponses incohérentes malgré une haute exactitude. Cette étude met en évidence les limites des métriques traditionnelles et propose un cadre pour évaluer de manière plus fine la compréhension des questions par les modèles multimodaux.

ABSTRACT

Evaluating VQA Models' Consistency

The Visual Question Answering (VQA) task poses a significant challenge, requiring a deep understanding of both the image and the associated question. Although recent vision-language models achieve high performance, they exhibit inconsistent behaviors when minor modifications are applied to the question. In this paper, we propose two evaluation protocols focused on question understanding : answer choice permutation and question rephrasing. These approaches enable the evaluation of model robustness and consistency beyond standard accuracy metrics. Our experimental results show that LLaVA and MM-CoT frequently produce inconsistent answers despite high accuracy. This study highlights the limitations of traditional evaluation metrics and introduces a framework for more fine-grained assessment of question understanding in multimodal models.

MOTS-CLÉS : Métrique d'évaluation, LLMs multimodaux, cohérence.

KEYWORDS: Evaluation metrics, Multimodal LLMs, Consistency.

ARTICLE ACCEPTÉ À : The 31st International Conference on Multimedia Modeling, MMM 2025.
URL : https://dl.acm.org/doi/10.1007/978-981-96-2071-5_29

∗. This work was conducted during Khanh-An C. Quan's internship at the National Institute of Informatics, Tokyo, Japan.

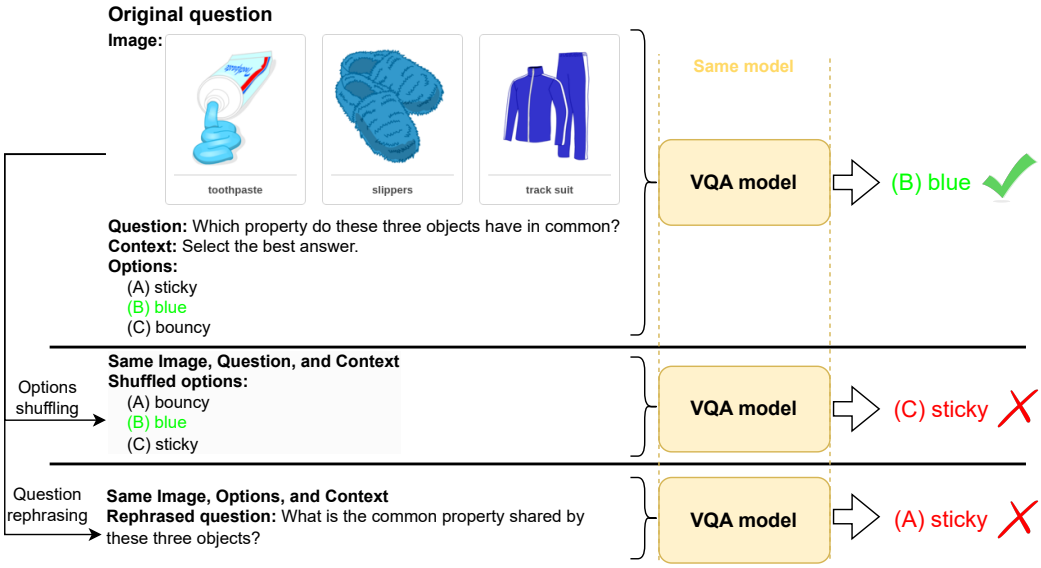

FIGURE 1 – Illustration des résultats incohérents produits par les approches VQA actuelles pour une même paire image-question, en modifiant uniquement l'ordre des choix ou en reformulant la question.

1 Introduction

La tâche de question réponse visuelle (*Visual Question Answering* - VQA) est une tâche complexe qui nécessite une compréhension approfondie de l'image fournie afin de répondre à une question donnée. En particulier, le modèle doit être capable d'analyser divers éléments visuels, notamment la reconnaissance d'instances, la lecture de texte, la compréhension des caractéristiques visuelles des objets, ainsi que le raisonnement à partir des données visuelles pour produire une réponse. Par ailleurs, l'intégration de différentes modalités de données, telles que les images et le texte, ajoute une complexité supplémentaire, car le modèle doit comprendre et exploiter les relations entre ces modalités.

Les jeux de données pour le VQA dans le domaine scientifique (Kembhavi *et al.*, 2017; Sampat *et al.*, 2020a; Lu *et al.*, 2022) ont été utilisés pour évaluer les capacités de raisonnement multi-étapes ainsi que l'interprétabilité des systèmes d'intelligence artificielle. Pour répondre aux questions dans ce domaine, le modèle doit non seulement comprendre des contenus multimodaux, mais également mobiliser des connaissances externes afin de déterminer la réponse correcte. Parmi les jeux de données récents dédiés au domaine scientifique, ScienceQA (Lu *et al.*, 2022) constitue un jeu de données à grande échelle composé de questions scientifiques à choix multiples, multimodales, accompagnées d'explications et couvrant un large éventail de domaines.

Les grands modèles de langage (LLMs) ont récemment démontré des performances remarquables dans de nombreuses tâches de traitement automatique du langage naturel (Touvron *et al.*, 2023; OpenAI, 2023). De plus, ils ont montré leur capacité à résoudre des problèmes de raisonnement complexes grâce aux processus de chaîne de pensée (*Chain-of-Thought*, CoT), en s'appuyant sur un nombre limité d'exemples de démonstration (Brown *et al.*, 2020). Dans cette perspective, Multimodal-CoT (MM-CoT) combine des données textuelles et visuelles selon une approche en deux étapes, en dissociant la génération du raisonnement de l'étape d'inférence de la réponse.

Bien que les LLMs et les VLMs obtiennent des résultats remarquables sur ScienceQA, ces modèles peuvent produire des sorties incohérentes. En particulier, une simple modification de l'ordre des choix

peut conduire ces modèles à fournir des réponses différentes pour une même question et une même image. La Figure 1 illustre cette incohérence dans les prédictions des modèles VQA actuels lorsqu'on conserve la même paire image-question tout en modifiant l'ordre des choix ou en reformulant la question.

Dans cet article, nous proposons deux approches qui modifient la paire image-question sans en altérer le sens, afin de mieux comprendre la manière dont les modèles VQA interprètent les questions. Dans la première approche, nous évaluons chaque question du jeu de données en considérant toutes les permutations possibles des choix, plutôt que leur ordre initial. Dans la seconde approche, nous reformulons chaque question sous différentes formes, puis évaluons les modèles VQA à partir de ces reformulations. En principe, un modèle VQA devrait produire les mêmes raisonnements et les mêmes réponses pour une question donnée, indépendamment de l'ordre des choix ou de la formulation de la question. Nous montrons que s'appuyer uniquement sur la métrique d'exactitude pour évaluer les modèles VQA est insuffisant. En effet, même pour des exemples présentant une exactitude élevée, les modèles peuvent produire des réponses incohérentes selon les deux approches d'évaluation proposées. Pour pallier cette limitation, nous introduisons deux métriques : Consistency across Choice Variations (CaCV) et Consistency across Question Variations (CaQV). Nous évaluons les performances de deux approches VQA récents, LLaVA (Liu *et al.*, 2023) et MM-CoT (Zhang *et al.*, 2024), sur le jeu de données ScienceQA (Lu *et al.*, 2022). Nos contributions peuvent être résumées en quatre points :

— nous introduisons deux approches qui apportent de légères modifications à la paire image-question sans en changer le sens, afin de mieux analyser le comportement des modèles VQA : l'évaluation par permutation des choix et l'évaluation par reformulation des questions ;
— nous proposons deux métriques pour mesurer la cohérence des modèles VQA : CaCV et CaQV. Nous comparons également ces métriques avec l'exactitude, en mettant en évidence les limites de cette dernière lorsqu'elle est utilisée seule ;
— nous menons des expérimentations sur deux approches VQA, LLaVA (Liu *et al.*, 2023) et MM-CoT (Zhang *et al.*, 2024), et montrons qu'elles atteignent respectivement 89.07% et 94.12% de CaCV, ainsi que 87.48% et 91.77% de CaQV sur le jeu de données ScienceQA (Lu *et al.*, 2022) ;
— enfin, nous analysons les caractéristiques des échantillons incohérents afin de mieux comprendre les limites actuelles des modèles.

2 Travaux connexes

Évaluation des MLLMs Avec les progrès des MLLMs, de nombreux benchmarks ont été proposés pour évaluer leurs capacités de compréhension, tels que (Fu *et al.*, 2024; Liu *et al.*, 2024; Yin *et al.*, 2024; Xu *et al.*, 2023; Li *et al.*, 2023). Des benchmarks récents, comme MME (Fu *et al.*, 2024), MMBench (Liu *et al.*, 2024) et SEED-Bench (Li *et al.*, 2023), évaluent les capacités de compréhension des MLLMs à travers des questions à choix multiples couvrant différentes dimensions de compétences. Li *et al.* (Li *et al.*, 2023) montrent que la plupart des MLLMs présentent encore des performances limitées sur des tâches nécessitant une compréhension fine au niveau des instances.

Raisonnement par chaîne de pensée (Chain-of-Thought) Les LLMs ont récemment démontré des résultats impressionnants en utilisant des techniques de prompting de type Chain-of-Thought (CoT) (Kojima *et al.*, 2022; Wei *et al.*, 2022). Plus précisément, les méthodes CoT incitent le LLM à produire une chaîne de raisonnement étape par étape pour résoudre un problème. Il existe deux principaux mécanismes pour effectuer un raisonnement CoT avec les LLMs : Zero-Shot-CoT et Few-Shot-CoT.

Kojima *et al.* (Kojima *et al.*, 2022) montrent que les LLMs peuvent effectuer du Zero-Shot-CoT en ajoutant simplement une instruction telle que « Let's think step by step » à la question. Dans le Few-Shot-CoT, les modèles apprennent le raisonnement à partir de quelques exemples illustrant un processus de raisonnement étape par étape. Des travaux récents indiquent que des modèles de langage fine-tunés peuvent induire un raisonnement CoT dans des modèles plus petits (Magister *et al.*, 2023; Ho *et al.*, 2023; Hsieh *et al.*, 2023).

VQA dans le domaine scientifique Résoudre des problèmes scientifiques est une tâche difficile nécessitant qu'un système comprenne non seulement des informations multimodales, mais aussi qu'il explique comment résoudre les questions. De nombreux benchmarks ont été proposés pour la VQA dans le domaine scientifique, tels que AI2D (Kembhavi *et al.*, 2016), DVQA (Kafle *et al.*, 2018), VLQA (Sampat *et al.*, 2020b), FOODWEDS (Krishnamurthy *et al.*, 2016), et ScienceQA (Lu *et al.*, 2022). Parmi ces jeux de données, ScienceQA (Lu *et al.*, 2022) introduit le raisonnement dans la tâche VQA, établissant une référence pour l'analyse multimodale avec chaîne de pensée. Il contient environ 21 000 questions multimodales à choix multiples couvrant un large éventail de sujets scientifiques, accompagnées de réponses annotées, de supports pédagogiques et d'explications. Il existe de nombreux travaux récents sur ce problème, mais deux grandes directions se dégagent : exploiter les capacités des LLMs ou entraîner des modèles vision-langage. En utilisant le prompting CoT, Lu *et al.* (Lu *et al.*, 2022) montrent qu'un modèle GPT-3 en few-shot peut améliorer les performances de raisonnement sur ScienceQA et produire des explications pertinentes. Cependant, comme GPT-3 est unimodal (texte uniquement), des modèles de captioning sont nécessaires pour convertir les informations visuelles en texte. Cela peut entraîner une perte d'information importante pour des images complexes. Pour pallier ce problème, LLaVA (Liu *et al.*, 2023) propose d'intégrer directement les informations visuelles dans le LLM et obtient des résultats remarquables sur ScienceQA (Lu *et al.*, 2022). De son côté, MM-CoT (Zhang *et al.*, 2024) propose une architecture en deux étapes séparant le raisonnement de la réponse. Comparé à LLaVA, MM-CoT offre des performances similaires avec un coût computationnel nettement inférieur. Récemment, T-SciQ (Wang *et al.*, 2024) montre que la combinaison de MM-CoT et du raisonnement des LLMs peut encore améliorer les performances en VQA.

3 Méthodologie

Dans cette étude, nous nous concentrons sur la tâche de VQA (Antol *et al.*, 2015), qui consiste à produire une réponse en exploitant à la fois les informations contenues dans la question et l'image associée. Plus précisément, considérons un jeu de données VQA composé de k échantillons $\{X, Y\}$, où X représente les entrées multimodales et Y les réponses de référence correspondantes. L'entrée multimodale X peut être notée $X = \langle T, I \rangle$, où T désigne le contenu textuel et I le contenu visuel associé à la question donnée. Le contenu textuel T peut être décomposé en $T = \langle Q, C, M \rangle$, où Q représente la question, C le contexte, et $M = (m_1, \ldots, m_k)$ la liste des réponses possibles, avec k le nombre de choix pour la question donnée. Il est important de noter que la liste des choix M dans l'entrée des modèles VQA actuels est une liste ordonnée.

3.1 Évaluation par permutation des choix

En théorie, un système de Visual Question Answering (VQA) devrait produire une réponse identique pour une même question associée à un même ensemble de choix, indépendamment de l'ordre

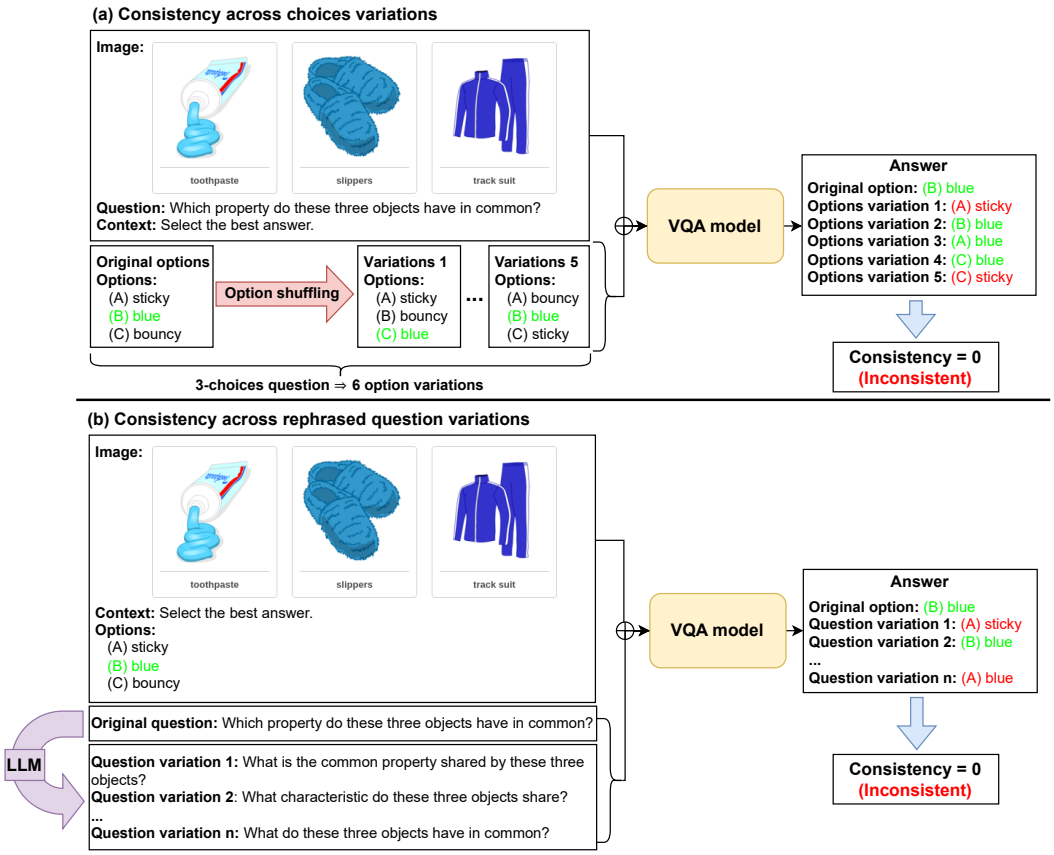

FIGURE 2 – Illustration des approches proposées pour évaluer la cohérence des modèles VQA dans le domaine scientifique.

dans lequel ces choix sont présentés. Afin d'évaluer cette propriété, nous introduisons la métrique *Consistency across all Choice Variations* (CaCV), qui mesure la cohérence des prédictions d'un modèle face aux permutations de l'ordre des choix de réponse.

Pour ce faire, nous considérons l'ensemble des permutations possibles des choix associés à chaque question. Plus précisément, étant donnée une liste de choix $M = (m_1, \ldots, m_k)$, nous construisons l'ensemble (M^*) contenant toutes les permutations de cette liste. Pour chaque permutation $M' \in M^*$, nous comparons la réponse prédite par le modèle à celle obtenue avec l'ordre original des choix. Un score binaire est alors attribué : il vaut 1 lorsque la réponse prédite est identique à celle obtenue avec l'ordre initial, et 0 lorsqu'elle diffère.

La valeur globale de la métrique CaCV sur un jeu de données est obtenue en calculant le pourcentage moyen de réponses cohérentes sur l'ensemble des échantillons et de leurs permutations. Une valeur élevée de CaCV indique que les prédictions du modèle demeurent globalement invariantes aux permutations des choix, tandis qu'une valeur faible révèle une dépendance importante à leur ordre de présentation.

3.2 Évaluation par reformulation des questions

En complément de l'évaluation fondée sur les permutations des choix de réponse, nous proposons une seconde métrique, appelée *Consistency across Question Variations* (CaQV), visant à mesurer la

robustesse des modèles VQA face à des reformulations linguistiques d'une même question. L'objectif est de vérifier si les prédictions du modèle reposent sur la compréhension sémantique du contenu de la question plutôt que sur sa formulation exacte.

Pour ce faire, nous générons plusieurs reformulations de chaque question à l'aide d'un grand modèle de langage (LLM). Étant donnée une question originale Q, nous construisons un ensemble de questions reformulées $Q_{\text{rephrase}} = (Q_1, \ldots, Q_n)$, où n désigne le nombre de variantes générées. Dans cette étude, nous utilisons ChatGPT-3.5 avec l'instruction « Rephrase this question into n different forms » afin de produire cinq reformulations pour chaque question. Un exemple de ces reformulations est présenté dans la Figure 2. La qualité et la fidélité sémantique des reformulations générées ont également été vérifiées manuellement.

À l'instar de la métrique CaCV, la métrique CaQV évalue la cohérence des prédictions du modèle en comparant les réponses obtenues pour chaque reformulation à celle produite à partir de la question originale. Pour chaque question reformulée , un score binaire est attribué : il vaut 1 lorsque la réponse prédite est identique à celle obtenue avec la question originale, et 0 lorsqu'elle diffère.

La valeur globale de CaQV est obtenue en calculant le pourcentage moyen de réponses cohérentes sur l'ensemble des reformulations et des échantillons du jeu de données. Une valeur élevée indique que les prédictions du modèle demeurent largement invariantes aux reformulations des questions, tandis qu'une valeur faible révèle une forte dépendance à leur formulation linguistique.

4 Expérimentation

4.1 Configuration expérimentale

jeu de données Nous utilisons le jeu de données ScienceQA (Lu *et al.*, 2022) pour l'évaluation et l'analyse. Ce jeu de données multimodal de questions scientifiques à choix multiples comprend $21\,000$ questions réparties sur trois disciplines, couvrant 26 thèmes, 127 catégories et 379 compétences distinctes. Le jeu de données est divisé en ensembles d'entraînement, de validation et de test, contenant respectivement $12\,726$, $4\,241$ et $4\,241$ échantillons. Dans cet article, nous nous concentrons sur les questions comportant 2, 3 ou 4 choix dans l'ensemble de test de ScienceQA, soit $2\,228$ questions à 2 choix, 971 questions à 3 choix et $1\,004$ questions à 4 choix.

Métriques Pour évaluer les performances des modèles, nous utilisons la métrique d'*exactitude* (*accuracy*). Étant donné que les réponses sont formulées sous forme de choix multiples (QCM), une prédiction est considérée comme correcte uniquement lorsque le choix sélectionné par le modèle correspond exactement à la réponse de référence annotée dans le jeu de données. Dans l'évaluation par permutation des choix (partie gauche du Tableau reftab :Overall), l'exactitude de chaque échantillon est calculée en moyennant l'exactitude sur toutes les variations de choix générées par le modèle évalué. De manière similaire, dans l'évaluation par reformulation des questions (partie gauche du tableau), l'exactitude de chaque échantillon est mesurée en moyennant l'exactitude sur toutes les reformulations de la question. Nous utilisons également les métriques proposées *CaCV* et *CaQV* pour les évaluations par permutation des choix et par reformulation des questions, respectivement.

Méthodes VQA comparées Dans cet article, nous utilisons deux modèles VQA comme référence : LLaVA (Liu *et al.*, 2023) et MM-CoT (Zhang *et al.*, 2024). Pour le modèle LLaVA, nous utilisons la version pré-entraînée sur ScienceQA avec 13 milliards de paramètres et fixons la température à

TABLE 1 – Pourcentage global de cohérence (CaCV et CaQV) et d'exactitude (Exact.) pour différents types de questions pour les modèles LLaVA (Liu *et al.*, 2023) et MM-CoT (Zhang *et al.*, 2024), avec les approches de permutation des choix (à gauche, en blanc) et de reformulation des questions (à droite, en bleu) sur ScienceQA (Lu *et al.*, 2022). Les meilleurs résultats sont indiqués en **gras**.

	Permutation				**Reformulation**			
	LLaVA		**MM-CoT**		**LLaVA**		**MM-CoT**	
Type de question	**CaCV**	**Exact.**	**CaCV**	**Exact.**	**CaQV**	**Exact.**	**CaQV**	**Exact.**
2-choix	95,37	**92,93**	**99,64**	92,14	85,18	**91,15**	**89,99**	90,78
3-choix	74,15	85,35	**88,36**	**86,06**	87,12	84,74	**92,79**	**86,03**
4-choix	89,54	**94,33**	**91,53**	92,80	92,92	93,71	**94,72**	**97,94**
Global	89,07	**91,53**	**94,12**	90,96	87,48	90,28	**91,77**	**91,39**

TABLE 2 – Résultats obtenus dans deux types de situations. À gauche, l'exactitude pour les exemples cohérents (Coh.) vs incohérents (Inc.). À droite, la cohérence (CaCV ou CaQV) pour les questions avec image (Img.) et sans image (W/o).

	Cohérent / Incohérent				**Avec / Sans image**			
	LLaVA		**MM-CoT**		**LLaVA**		**MM-CoT**	
	Inc.	**Coh.**	**Inc.**	**Coh.**	**Img.**	**W/o**	**Img.**	**W/o**
Permutation	51,42	96,45	45,72	93,78	86,48	91,93	92,87	95,33
Reformulation	56,56	95,10	60,16	94,19	87,95	88,97	92,59	90,04

0 afin d'assurer la reproductibilité. Pour le modèle MM-CoT, nous utilisons le plus grand modèle pré-entraîné sur ScienceQA (768 millions de paramètres), qui obtient les meilleures performances.

4.2 Résultats et analyse

Résultats globaux Le résultat global des évaluations par permutation des choix et par reformulation des questions est présenté dans le Tableau 1. Malgré des niveaux d'exactitude similaires, MM-CoT présente une cohérence plus élevée que LLaVA, avec respectivement 94,12% contre 89,07% pour la permutation des choix, et 91,77% contre 87,48% pour la reformulation des questions. Comparée à l'évaluation par permutation des choix, la reformulation des questions conduit à une cohérence plus faible pour les deux modèles. Cela peut s'expliquer par la difficulté accrue liée à la compréhension de formulations variées d'une même question. On observe également que la cohérence n'est pas liée au nombre de choix. En particulier, les questions à 3 choix présentent la cohérence la plus faible, tandis que les questions à 2 choix affichent la cohérence la plus élevée pour les deux modèles LLaVA et MM-CoT. Le Tableau 2 (droite) illustre la comparaison de cohérence entre les questions avec images et celles uniquement textuelles dans le jeu de données ScienceQA (Lu *et al.*, 2022). Lors de l'évaluation par permutation des choix, les questions comportant des images présentent une cohérence plus faible que celles sans image. Bien que la cohérence des questions avec images reste globalement stable dans l'évaluation par reformulation, une légère baisse est observée pour les questions purement textuelles.

Comparaison avec l'exactitude Le Tableau 2 (gauche) met en évidence l'exactitude des échantillons cohérents et incohérents pour les deux méthodes d'évaluation. Alors que les échantillons cohérents

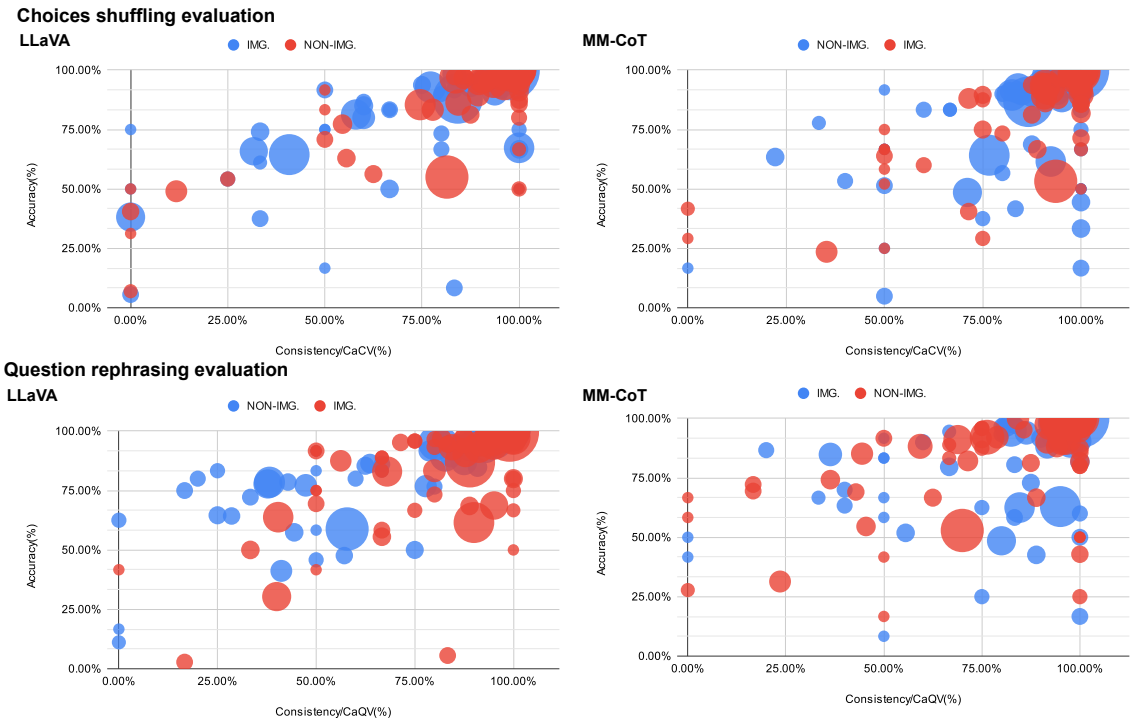

FIGURE 3 – Nuage de points représentant la relation entre cohérence et exactitude pour chaque question du jeu de données ScienceQA (la taille des cercles correspond au nombre d'exemples).

présentent une exactitude élevée d'environ 95% pour les deux modèles, les échantillons incohérents affichent encore une exactitude d'environ 50%. Cela suggère que le modèle peut parfois prédire la bonne réponse par hasard ; néanmoins, l'exactitude des échantillons incohérents contribue de manière significative à l'exactitude globale, tant pour les permutations de choix que pour les reformulations de questions. La Figure 3 illustre la relation entre cohérence et exactitude pour chaque question du jeu de données, en utilisant les deux approches proposées. Le graphique montre de nombreux cas où cohérence et exactitude ne sont pas alignées. Dans les cas de forte exactitude mais faible cohérence, le modèle VQA ne fonctionne pas correctement et peut produire la bonne réponse de manière aléatoire. À l'inverse, dans les cas de faible exactitude mais forte cohérence, le modèle ne comprend pas correctement la question et produit des réponses incorrectes de manière systématique. Ainsi, les métriques de cohérence proposées complètent efficacement l'exactitude, en apportant une compréhension plus approfondie du comportement des modèles VQA. Nous mesurons également l'impact de la cohérence sur l'exactitude selon quatre scénarios : choix originaux, toutes les variations de choix, meilleurs cas et pires cas. Dans le cas des choix originaux, l'exactitude est évaluée uniquement sur les choix fournis par ScienceQA. Dans le cas des variations complètes, l'exactitude est calculée sur toutes les permutations possibles. Dans le scénario des meilleurs cas, une réponse est considérée correcte si au moins une des variations est correcte. À l'inverse, dans les pires cas, une réponse est considérée incorrecte si au moins une variation est incorrecte. Dans le cas des variations complètes, l'exactitude reste similaire à celle des choix originaux. En revanche, dans les scénarios meilleur/pire cas, on observe respectivement une augmentation et une diminution significatives de l'exactitude.

Analyse de la permutation des choix Nous avons constaté que la majorité des échantillons incohérents dans l'évaluation par permutation des choix correspond à des questions comportant des images, représentant 67.32% et 64.37% pour LLaVA et MM-CoT respectivement. Nous remarquons

également que la cohérence des questions avec images est inférieure à celle des questions purement textuelles dans ce cadre d'évaluation. L'analyse des échantillons incohérents met en évidence plusieurs caractéristiques influençant la cohérence : (1) la compréhension fine des instances dans l'image (de nombreux échantillons incohérents correspondent à des questions nécessitant une compréhension détaillée de l'image) ; (2) la nécessité de connaissances spécifiques ou de raisonnement logique ; (3) des problèmes de données (certaines questions nécessitent une image qui n'est pas fournie).

Analyse de la reformulation des questions Comparée à l'évaluation par permutation des choix, l'évaluation par reformulation des questions montre des tendances similaires concernant les échantillons incohérents. Cependant, en plus de la diminution de la cohérence pour les questions purement textuelles, de nouveaux cas incohérents apparaissent également. La plupart de ces nouveaux cas nécessitent un raisonnement logique de la part du modèle.

Caractéristiques des échantillons cohérents L'analyse des échantillons cohérents issus des deux types d'évaluation montre que la majorité correspond à des questions nécessitant principalement la compréhension du texte pour déduire la réponse. Certains de ces échantillons incluent des images qui jouent un rôle illustratif mais ne sont pas essentielles pour répondre à la question. Les questions liées aux cartes, qui requièrent une compréhension globale de l'image, présentent également une forte cohérence. Bien que certaines questions nécessitant une analyse fine des détails visuels présentent une faible cohérence, la majorité des autres cas affiche de bonnes performances en termes de cohérence et d'exactitude.

5 Conclusion

Dans cet article, nous avons proposé deux nouvelles approches d'évaluation, à savoir la permutation des choix et la reformulation des questions, afin d'analyser plus finement le comportement des modèles de VQA. Nous avons introduit deux métriques dédiées à l'évaluation de la cohérence des modèles : *Consistency across all Choice Variations* (CaCV) et *Consistency across Question Variations* (CaQV). Nos expériences mettent en évidence que l'utilisation exclusive de la métrique d'exactitude pour évaluer les modèles VQA est insuffisante, et que la combinaison de mesures de précision et de cohérence permet d'obtenir une compréhension beaucoup plus complète et nuancée de leurs performances. Nous montrons notamment que les modèles VQA actuels peuvent produire des réponses incohérentes pour une même paire image-question, et ce indépendamment de leur niveau d'exactitude.

Références

ANTOL S., AGRAWAL A., LU J., MITCHELL M., BATRA D., ZITNICK C. L. & PARIKH D. (2015). VQA : Visual Question Answering. In *ICCV*.

BROWN T., MANN B., RYDER N., SUBBIAH M., KAPLAN J. D., 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., 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. In H. LAROCHELLE, M. RANZATO, R. HADSELL, M. BALCAN & H. LIN, Éds., *Advances in Neural Information Processing Systems*, volume 33, p. 1877–1901 : Curran Associates, Inc.

FU C., CHEN P., SHEN Y., QIN Y., ZHANG M., LIN X., YANG J., ZHENG X., LI K., SUN X., WU Y. & JI R. (2024). Mme : A comprehensive evaluation benchmark for multimodal large language models.

HO N., SCHMID L. & YUN S.-Y. (2023). Large language models are reasoning teachers. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *ACL*, p. 14852–14882.

HSIEH C.-Y., LI C.-L., YEH C.-K., NAKHOST H., FUJII Y., RATNER A., KRISHNA R., LEE C.-Y. & PFISTER T. (2023). Distilling step-by-step ! outperforming larger language models with less training data and smaller model sizes. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *ACL*, p. 8003–8017.

KAFLE K., PRICE B., COHEN S. & KANAN C. (2018). Dvqa : Understanding data visualizations via question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 5648–5656.

KEMBHAVI A., SALVATO M., KOLVE E., SEO M., HAJISHIRZI H. & FARHADI A. (2016). A diagram is worth a dozen images. In B. LEIBE, J. MATAS, N. SEBE & M. WELLING, Éds., *ECCV*, p. 235–251 : Springer International Publishing.

KEMBHAVI A., SEO M., SCHWENK D., CHOI J., FARHADI A. & HAJISHIRZI H. (2017). Are you smarter than a sixth grader ? textbook question answering for multimodal machine comprehension. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 5376–5384.

KOJIMA T., GU S. S., REID M., MATSUO Y. & IWASAWA Y. (2022). Large language models are zero-shot reasoners. *NeurIPS*, **35**, 22199–22213.

KRISHNAMURTHY J., TAFJORD O. & KEMBHAVI A. (2016). Semantic parsing to probabilistic programs for situated question answering. In J. SU, K. DUH & X. CARRERAS, Éds., *EMNLP*, p. 160–170.

LI B., WANG R., WANG G., GE Y., GE Y. & SHAN Y. (2023). Seed-bench : Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv :2307.16125*.

LIU H., LI C., WU Q. & LEE Y. J. (2023). Visual instruction tuning. In *NeurIPS*.

LIU Y., DUAN H., ZHANG Y., LI B., ZHANG S., ZHAO W., YUAN Y., WANG J., HE C., LIU Z., CHEN K. & LIN D. (2024). Mmbench : Is your multi-modal model an all-around player ?

LU P., MISHRA S., XIA T., QIU L., CHANG K.-W., ZHU S.-C., TAFJORD O., CLARK P. & KALYAN A. (2022). Learn to explain : Multimodal reasoning via thought chains for science question answering. In *NeurIPS*.

MAGISTER L. C., MALLINSON J., ADAMEK J., MALMI E. & SEVERYN A. (2023). Teaching small language models to reason. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 1773–1781 : ACL.

OPENAI (2023). GPT-4 technical report. *CoRR*.

SAMPAT S. K., YANG Y. & BARAL C. (2020a). Visuo-linguistic question answering (VLQA) challenge. In T. COHN, Y. HE & Y. LIU, Éds., *EMNLP*, p. 4606–4616.

SAMPAT S. K., YANG Y. & BARAL C. (2020b). Visuo-linguistic question answering (VLQA) challenge. In T. COHN, Y. HE & Y. LIU, Éds., *EMNLP*, p. 4606–4616.

TOUVRON H., LAVRIL T., IZACARD G., MARTINET X., LACHAUX M.-A., LACROIX T., ROZIÈRE B., GOYAL N., HAMBRO E., AZHAR F. *et al.* (2023). Llama : Open and efficient foundation language models. *arXiv preprint arXiv :2302.13971*.

WANG L., HU Y., HE J., XU X., LIU N., LIU H. & SHEN H. (2024). T-sciq : Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. *AAAI*, **38**, 19162–19170.

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

XU P., SHAO W., ZHANG K., GAO P., LIU S., LEI M., MENG F., HUANG S., QIAO Y. & LUO P. (2023). Lvlm-ehub : A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv :2306.09265*.

YIN Z., WANG J., CAO J., SHI Z., LIU D., LI M., HUANG X., WANG Z., SHENG L., BAI L. *et al.* (2024). Lamm : Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *Advances in Neural Information Processing Systems*, **36**.

ZHANG Z., ZHANG A., LI M., HAI ZHAO, KARYPIS G. & SMOLA A. (2024). Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*.