

Les modèles multimodaux peuvent-ils aider à l'interprétation de cartes ? Une étude exploratoire avec GPT-4o

Edith Galy Ahmed Moubtahij Azur Handan Marc Queudot

CRIM, 101 – 405, avenue Ogilvy, H3N 1M3, Montréal (Québec), Canada

edith.galy@crim.ca, ahmed.moubtahij@crim.ca, azur.handan@crim.ca,
marc.queudot@crim.ca

RÉSUMÉ

Cet article explore l'utilisation des modèles de langage multimodaux, en particulier GPT-4o, pour l'interprétation automatisée de cartes de risque d'inondation. Un prototype a été développé afin de permettre à des utilisateurs non-experts de poser des questions en langage naturel et d'obtenir des réponses ancrées sur des données géospatiales visuelles. Un jeu de données ad hoc a été constitué pour évaluer la capacité du modèle à répondre à des questions fermées, selon différentes stratégies de génération. Malgré certaines améliorations grâce à l'usage de schémas de génération structurée et de raisonnements intermédiaires, les résultats révèlent une forte tendance aux hallucinations et des performances insuffisantes pour une application en contexte critique. Cette étude met en évidence les limites actuelles des modèles multimodaux pour l'analyse cartographique, et souligne la nécessité de recherches fondamentales et de corpus plus étendus pour fiabiliser ces approches.

ABSTRACT

This paper investigates the use of multimodal language models, specifically GPT-4o, for the automated interpretation of flood risk maps. We developed a prototype allowing non-expert users to interact in natural language and receive answers grounded in geospatial visual data. A dedicated dataset was constructed to evaluate the model's performance on closed-ended questions using various generation strategies. While structured generation and intermediate reasoning techniques yielded modest improvements, the results reveal frequent hallucinations and performance levels that remain inadequate for high-stakes applications. This exploratory study highlights current limitations of multimodal models in cartographic analysis and emphasizes the need for foundational research and larger-scale datasets to build reliable and trustworthy systems in this domain.

MOTS-CLÉS : Modèles de langage multimodaux, Interprétation de cartes, Risque d'inondation, Question-réponse visuelle (VQA), Hallucinations, Données géospatiales, Communication de crise climatique, Génération structurée, GPT-4o.

KEYWORDS: Multimodal language models, Map interpretation, Flood risk assessment, Visual Question Answering (VQA), hallucinations, Geospatial data, Climate crisis communication, Structured generation, GPT-4o.

1 Introduction

Les agences gouvernementales et les médias sont généralement responsables de la préparation et de la centralisation de l'information en vue de répondre efficacement aux catastrophes environnementales et de protéger les communautés. Cependant, les retards de communication, les lacunes dans la préparation et l'absence d'outils adaptés aux non-spécialistes restent fréquents, même dans des contextes où l'information est disponible. Des exemples tristement célèbres, meurtriers et coûteux de réponses tardives incluent l'ouragan Katrina (2005), le cyclone Nargis (2008), les incendies de forêt en Russie (2010) ou encore les feux de brousse en Australie (2019–2020). Malgré l'amélioration des capacités de prévision, les outils permettant une interprétation compréhensible et rapide des données par le grand public restent limités.

Alors que les catastrophes liées au climat deviennent plus fréquentes et intenses, la capacité à transmettre rapidement des informations claires, fiables et visuellement ancrées devient un enjeu critique pour la gestion des risques. Ce besoin est d'autant plus pressant pour les communautés non-expertes, souvent éloignées des systèmes d'alerte professionnels.

Les modèles de langage de grande taille (LLMs) ont démontré des performances remarquables dans un large éventail de tâches, y compris dans des domaines pour lesquels ils n'avaient pas été spécifiquement entraînés. Toutefois, l'exploitation de données géospatiales cartographiques par ces modèles reste largement inexplorée, en particulier dans des formats visuels bruts tels que les cartes.

L'émergence de modèles multimodaux, tels que GPT-4o (Hurst *et al.*, 2024), capables d'intégrer simultanément des entrées textuelles et visuelles, offre une opportunité nouvelle d'interagir avec ces documents. Ce travail propose d'explorer les capacités réelles de ce type de modèle pour interpréter des cartes de risque d'inondation, dans une optique d'assistance au grand public ou aux acteurs de la gestion de crise.

Nous avons ainsi développé un prototype (disponible à l'adresse <https://ogc-demo.crim.ca/>) permettant une interaction en langage naturel avec des cartes issues de données officielles, et construit un jeu de données spécifique pour évaluer la performance du modèle dans un cadre contrôlé. Nous mettons en évidence les limites actuelles de ces approches, en particulier la fréquence des hallucinations et les difficultés à ancrer les réponses dans les éléments visuels réels. Nos résultats suggèrent que, malgré certaines améliorations techniques possibles, ces modèles restent à ce jour insuffisamment matures pour une utilisation dans des contextes sensibles.

2 Prototype développé

Le prototype développé vise à explorer le potentiel des modèles génératifs multimodaux dans l'analyse de données cartographiques, en particulier les cartes de risque d'inondation. L'objectif principal est de permettre à un utilisateur non-expert d'interagir en langage naturel avec un assistant virtuel qui s'appuie sur des sources de données géospatiales pour fournir des réponses contextualisées et vérifiables. Contrairement aux systèmes classiques basés sur des données structurées, ce prototype repose sur l'analyse d'images. Cela ouvre la possibilité d'interroger tout type de carte dans un contexte similaire.

Le fonctionnement du prototype repose sur les étapes suivantes :

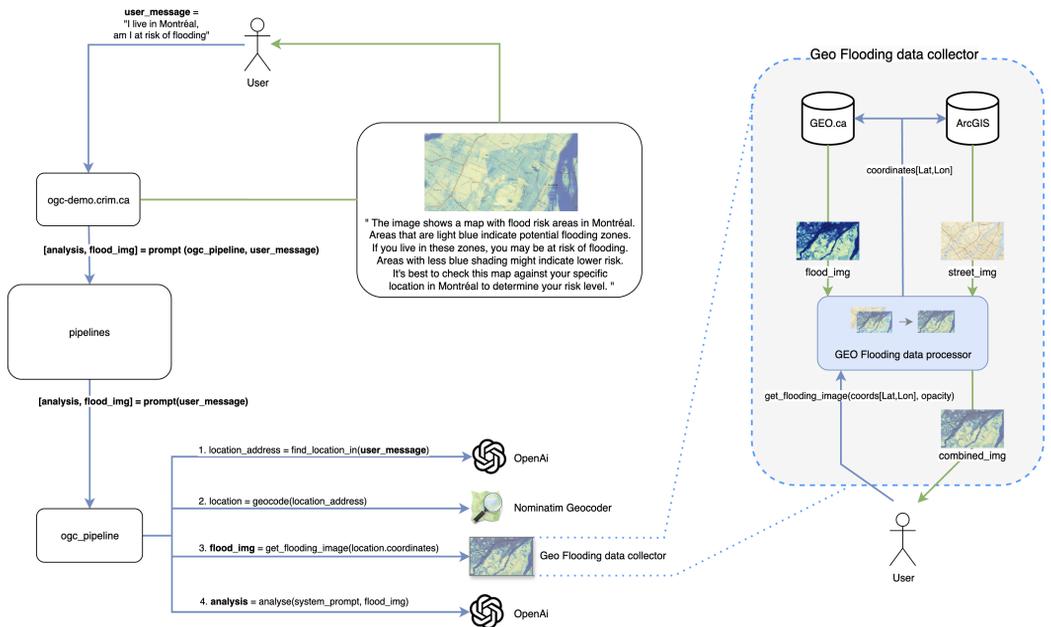


FIGURE 1 – Flux d'utilisation du prototype

1. **Identification de l'emplacement** : un premier appel à GPT-4o-mini sert à détecter la mention d'un emplacement géographique dans la requête utilisateur.
2. **Géocodage** : l'adresse est convertie en coordonnées GPS via un service basé sur Nominatim¹.
3. **Récupération d'image** : le système génère l'image de la zone à risque en interrogeant la source de données de carte. Le pipeline exact est décrit à la section suivante.
4. **Réponse ancrée basée sur l'information présente sur l'image** : le prompt initial de l'utilisateur est enrichi par l'image pour produire une réponse sans reposer sur les connaissances implicites du modèle.

Ce principe d'ancrage visuel s'inspire des approches RAG (Retrieval-Augmented Generation)(Lewis et al., 2020), en remplaçant les documents textuels par des images comme source de vérité contextuelle. L'utilisateur peut alors vérifier lui-même la validité de la réponse. Nous avons remarqué empiriquement que le système reconnaissait bien des lieux communs au Québec (nous avons restreint le scope à la portée de la carte disponible). Certaines formulations communément utilisées "à l'intersection des rues X et Y" ne fonctionnent pas pour l'instant et demanderaient des stratégies plus spécifiques.

2.1 Collecte et traitement d'images géospatiales

Le composant de collecte de données géospatiales a été développé pour automatiser l'extraction d'images à partir des cartes de susceptibilité d'inondations (McGrath & Gohl, 2023) disponibles sur

1. Service de géocoding open-source Nominatim : <https://nominatim.org/>

le portail `geo.ca`. Si la requête utilisateur mentionne un emplacement géographique, la description de ce lieu est extraite et convertie en coordonnées GPS à l'aide d'un service de geocoding. À partir de ces coordonnées, la source de données est interrogée pour récupérer trois types d'images :

- une carte de rues issue d'ArcGIS Online², obtenue par assemblage de tuiles correspondant à la zone d'intérêt;
- une image extraite de la couche cartographique *Flood Susceptibility Index 2015*, centrée sur les mêmes coordonnées;
- une image combinée, superposant les deux précédentes avec un niveau d'opacité configurable (typiquement 50%), permettant une bonne lisibilité tout en conservant l'information sur les zones à risque.

La génération de l'image suit un format standardisé de 1024x512 pixels afin d'assurer une visualisation cohérente et réutilisable dans les modules d'analyse subséquents. La figure 2 illustre un exemple de ces 3 images pour l'île de Montréal.

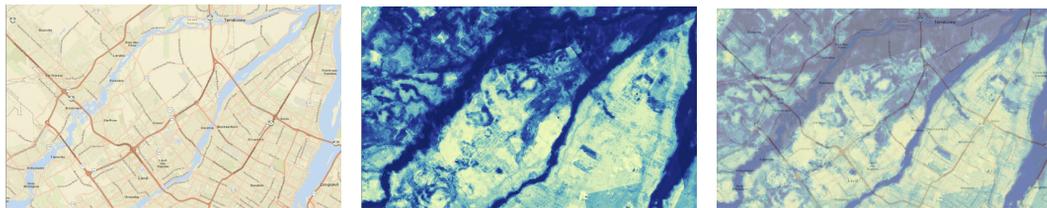


FIGURE 2 – Trois images pour un même emplacement : la carte civile, la visualisation de susceptibilité d'inondation et une combinaison des deux par superposition

2.2 Architecture et interface

Le prototype est accessible via une interface web (`ogc-demo.crim.ca`) qui agit comme médiateur entre l'utilisateur final et les différents composants du système. Une API permet également d'accéder directement aux modules de génération d'image et d'analyse.

La Figure 1 illustre l'architecture générale du système, du traitement de la requête jusqu'à la génération de réponse. La Figure 3 montre une vue exemple de l'interface conversationnelle intégrée au démonstrateur.

3 Évaluation de la maturité des modèles de langage sur leur capacité à interpréter des cartes

3.1 Méthode

Nous évaluons les capacités d'interprétation cartographique des modèles de langage multimodaux à travers une tâche de type *Visual Question Answering (VQA)* (Antol *et al.*, 2015). L'objectif est de déterminer dans quelle mesure un modèle est capable de répondre correctement à des questions

2. <https://www.arcgis.com/index.html>

I live in Montréal,
am I at risk of flooding

OI OIG-Flood Susceptibility

The image shows a map with flood risk areas in Montréal. Areas that are light blue indicate potential flooding zones. If you live in these zones, you may be at risk of flooding. Areas with less blue shading might indicate lower risk. It's best to check this map against your specific location in Montréal to determine your risk level.

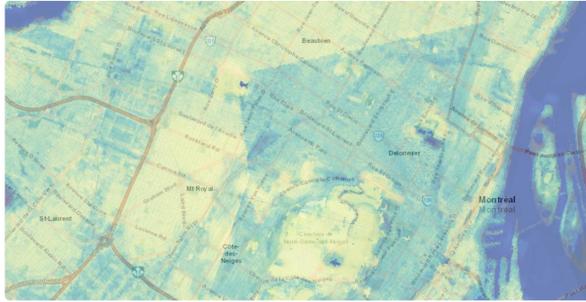


FIGURE 3 – Exemple d'interaction avec l'agent conversationnel

portant sur le contenu d'une carte, en s'appuyant uniquement sur les informations visuelles qu'elle contient. La fidélité au contexte visuel fourni est cruciale afin de limiter le potentiel d'hallucinations et permettre une vérification par l'utilisateur. Pour cette raison, on jugera négativement une réponse du modèle qui ne se baserait pas sur l'image. Les questions adoptent un format fermé — les réponses possibles étant « oui », « non » ou « en partie » — ce qui permet une évaluation plus objective et systématique des performances.

Les corpus VQA existants intègrent rarement des informations géospatiales précises, et incluent encore plus rarement des représentations cartographiques. Les jeux de données les plus proches identifiés — tels que MapQA (Chang *et al.*, 2022), qui propose des questions basées sur une carte très simplifiée des États-Unis, ou FloodNet (Rahneemoonfar *et al.*, 2021), axé sur l'analyse d'images aériennes dans un contexte d'inondation — ne couvrent ni la diversité ni le niveau de précision requis pour notre étude. Pour ces raisons, nous avons constitué notre propre jeu de données à partir d'une carte de susceptibilité aux inondations au Canada (McGrath & Gohl, 2023).

Les expériences sont menées en utilisant le modèle GPT-4o (Hurst *et al.*, 2024) d'OpenAI et comparent un ensemble de techniques d'adaptation sur notre jeu de données. Nous considérons la performance sur ces questions fermées comme un proxy simplifié de la performance générale d'interprétation de cartes avec des modèles multimodaux - si cette performance est trop basse dans ce cas "facile", la technologie n'est pas assez mature pour traiter le cas général.

3.2 Construction du jeu de données

Pour évaluer la capacité du modèle multimodal GPT-4o à interpréter des cartes de risque d'inondation, nous avons conçu un jeu de données original, centré sur une tâche de type Visual Question Answering (VQA), orientée vers des questions fermées.

Le jeu de données final comprend 88 paires question-réponse en français³, générées de manière

3. Disponible avec le code de l'application et d'expériences à l'adresse <https://github.com/crim-ca/>

Question	Les zones au nord de Montréal, comme Laval, présentent-elles un risque d'inondation plus élevé que celles situées au sud, comme Saint-Lambert ?
Réponse obtenue	Oui
Explication	Dans le quadrant Nord-Ouest, Laval montre une prédominance de bleu clair et foncé, indiquant un risque élevé d'inondation, alors que Saint-Lambert, dans le quadrant Sud-Est, montre moins de bleu, indiquant un risque moindre.
Réponse attendue	Non
Analyse	Saint-Lambert est bleu foncé et Laval a du jaune et un peu de bleu clair.

TABLE 1 – Un exemple du jeu de données de questions fermées et de l'analyse associée

semi-automatique à partir de cartes issues de l'indice de susceptibilité aux inondations au Canada (McGrath & Gohl, 2023). Chaque paire a été validée manuellement par deux annotateurs.

Le processus de création s'est déroulé en trois étapes. Dans un premier temps, des zones ont été sélectionnées sur la carte du Canada (la couverture du modèle de susceptibilité aux inondations) en cherchant la variété des exemples. Les emplacements sélectionnés sont des zones habitées ou proches d'habitations. Par la suite, le modèle GPT-4o a été sollicité pour formuler des questions précises sur ces cartes, en mettant l'accent sur l'identification de lieux spécifiques tels que des noms de rue, des intersections ou des quartiers clairement définis. Nous avons obtenu de cette manière une plus grande diversité de questions que celles produites par nos annotateurs humains dans une première passe, et avons donc poursuivi avec cette approche.

Plusieurs prompts ont été utilisés pour générer ces paires de question - réponse, un exemple peut être trouvé en annexe A. Les questions fermées ont été conservées pour ce jeu de données, et les autres conservées pour d'autres explorations.

Dans un second temps, les réponses ont été générées par le modèle sous la forme d'une réponse courte associée à une explication. Enfin, une phase de validation manuelle a permis de corriger les réponses produites par le modèle. Deux annotateurs ont été impliqués dans cette tâche, chacun se répartissant une série de paires Q/R. Dix paires ont été annotées en double, permettant d'obtenir un accord inter-annotateur Cohen's Kappa de 0.828. La table 1 illustre une question type de ce jeu de données.

À l'issue de cette phase, 13 paires ont été retirées en raison d'erreurs typiques du modèle, notamment :

- la mention de lieux non visibles sur la carte fournie (ex. une île inexistante) ;
- une erreur de localisation manifeste (ex. un lieu annoncé au sud-est qui se situe en réalité au nord-ouest) ;
- des formulations trop vagues, faisant appel à des connaissances implicites non visualisables (ex. "la zone industrielle", "les infrastructures majeures").

Ce processus a mis en lumière les premières limitations du modèle pour la compréhension cartographique, dès la phase de génération du corpus lui-même. Il a également révélé la difficulté intrinsèque de la tâche : même pour un humain, interpréter les dégradés de couleur représentant la probabilité d'inondation, et localiser précisément les zones mentionnées, reste un exercice délicat sans connaissance préalable du territoire. Le processus d'annotation, et particulièrement de vérification

de l'absence d'hallucinations dans la réponse ou les explications, est particulièrement laborieux, ce qui a limité notre capacité à produire une grande quantité d'exemples de qualité. Bien que de taille modeste, ce jeu de données constitue une première base pour l'évaluation de modèles multimodaux dans un contexte visuel géospatial.

3.3 Mesure quantitative de la réponse aux questions fermées

Nous avons appliqué diverses techniques de génération pour répondre aux questions du jeu de données et avons évalué ces réponses par rapport aux réponses validées manuellement (le "ground truth"). À chaque fois l'image associée à la question est fournie à l'assistant.

Les techniques de génération utilisées sont les suivantes :

- **Génération directe** : prédiction directe de la réponse en utilisant la question et l'image associée comme entrée.
- **COT (*Chain Of Thought*, (Wei *et al.*, 2023))** : génération par complétion de réponse à partir de l'instruction "réfléchissons étape par étape".
- **Génération structurée** : génération avec garantie d'adhérence au format de réponse décrit par un schéma JSON.
- **Génération structurée avec COT** : combinaison de la génération structurée et du raisonnement ; une première étape de raisonnement permet au modèle d'assimiler le problème dans l'espace des tokens, améliorant potentiellement sa classification.
- **Génération structurée avec repli** : si la génération directe ne contient pas le format attendu ("oui", "non" ou "en partie"), tentative d'extraction de la réponse à partir du texte généré en forçant la structure (méthode employée par (Moubtahij *et al.*, 2024)).
- **Génération structurée avec repli et COT** : combinaison des trois techniques précédentes.

Ces techniques ont été implémentées à l'aide de modules DSPY⁴ afin de permettre de faire abstraction du modèle choisi dans le prompt.

Les mesures d'évaluation utilisées sont la précision, le rappel et le score F1. Une réponse est considérée correcte si elle correspond exactement à la réponse attendue (à l'exception de la casse, qui est ignorée). Dans le tableau 2, on remarque premièrement les deux premières lignes : pour ces deux techniques sans génération structurée, le texte libre ne contenait généralement pas l'une des réponses attendues (soit par refus de répondre, ou des réponses détournées ou formulées de manière différentes). La génération structurée pour les autres techniques élimine ce problème. L'expérience combinant toutes les approches améliore significativement la performance dans cette situation contrôlée, sans pour autant atteindre des scores très élevés. Comme déjà décrit dans les sections précédentes, l'apparition fréquente d'hallucinations (mention de lieux non présents sur l'image ou contradictoires à l'information présentée mais "intuitives") semble limiter la performance qu'il est possible d'obtenir dans ce contexte. Nous avons aussi analysé les explications associées à chaque réponse courte, et constaté à de nombreuses reprises une réponse contradictoire par rapport à l'explication.

Les résultats pour chaque classe (ou réponse attendue) sont présentés dans la table 3, où on peut remarquer une forte disparité de performance entre les réponses attendues. Pour la technique de

4. <https://dspy.ai/>

Mode	Précision	Rappel	F1-score
Génération directe*	NA	NA	NA
Chain-of-Thought*	NA	NA	NA
Génération structurée	56	53	50
Génération structurée + Chain-of-Thought	54	49	46
Repli vers génération structurée	62	60	59
Repli vers génération structurée + Chain-of-Thought	66	67	64

TABLE 2 – Résultats de questions fermées (moyenne pondérée de correspondance exacte)

génération structurée avec repli et Chain of Thought par exemple, la précision pour les réponses "oui" et "non" est respectivement de 69 et 73%, contre 44% pour la réponse "en partie". Les situations pour lesquelles la réponse attendue est "en partie" sont souvent plus nuancées et donc plus difficiles que les autres, ce qui peut expliquer ce résultat. Bien sûr, nous rappelons la faible taille du jeu de données qui limite la capacité de généraliser complètement ces observations.

Mode	Précision			Rappel			F1		
	oui	non	en partie	oui	non	en partie	oui	non	en partie
Directe*	NA			NA			NA		
CoT*	NA			NA			NA		
Structurée	63	57	29	73	16	54	67	25	38
Structurée + CoT	56	71	17	73	20	23	63	31	19
Repli	64	75	33	80	36	38	71	49	36
Repli + CoT	69	73	44	91	44	31	78	55	36

TABLE 3 – Détail des performances de questions fermées : correspondance exacte par réponse attendue

3.4 Questions ouvertes

En plus des questions fermées, nous avons testé le système avec un petit jeu de questions - réponses ouvertes. La figure 4 présente un exemple de question ouverte de ce type et de réponse produite. On peut remarquer dans cet exemple la difficulté de déterminer le bon niveau de précision et de détails dans la réponse. Les difficultés rencontrées lors de la création du jeu de données sont amplifiées dans ce cas avec des questions ouvertes. Les hallucinations sont plus difficiles à détecter que pour les questions fermées et le jugement de la justesse des réponses est beaucoup plus subjectif. De fait, il a été difficile de construire un tel jeu de données (pour produire les réponses).

Sans surprise, s'il est difficile de produire des réponses dans ce contexte pour un humain, la tâche s'est avérée encore plus complexe pour le système. Nous avons eu plus de succès en demandant d'abord au modèle si la carte fournie lui permet de répondre à la question plutôt que de le faire répondre directement à la question. Nous avons aussi constaté, comme pour les questions fermées, la présence d'hallucinations géographiques (lieu inexistant sur la carte) et le recours à des connaissances implicites plutôt qu'à une réponse basée sur la carte pour répondre aux questions.

Question : Quelles sont les zones principales de Montréal les plus à risque d'inondation (en bleu foncé) ?

Réponse :

- Collège Reine-Marie entre la 125 et Boulevard Saint-Michel
- entre Parc Extension et Outremont
- au pied du Mont Royal au nord-ouest

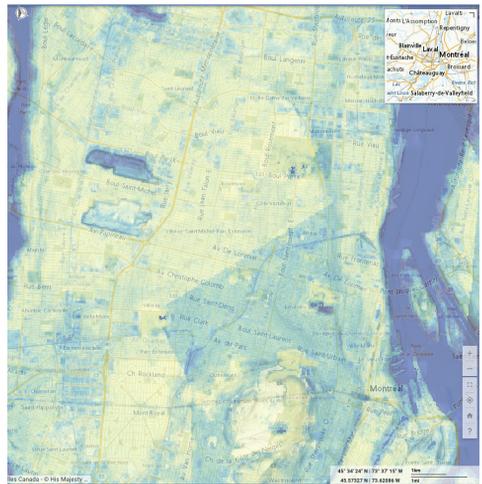


FIGURE 4 – Exemple de question ouverte avec réponse géolocalisée et carte de référence

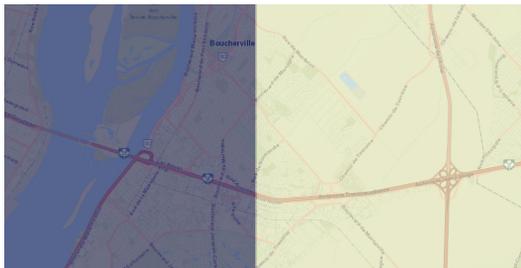


FIGURE 5 – Procédé expérimental pour vérifier l'utilisation de l'information visuelle de la carte : un calque avec des probabilités d'inondations fausses et contre-intuitives

3.5 Détection d'hallucinations

Après avoir constaté une grande quantité d'hallucinations comme décrit plus haut, nous avons voulu vérifier que le modèle exploitait bien l'information présente explicitement sur la carte, plutôt que des informations apprises pendant son entraînement. Pour ce faire, nous avons appliqué systématiquement un calque avec des portions bleues et jaunes sur des zones sélectionnées, sans aucun lien avec la réalité. Le modèle devrait théoriquement ajuster ses réponses en tenant compte de ces calques bleus. Si, en revanche, les réponses restent identiques, cela indiquerait que le modèle repose davantage sur ses connaissances implicites du lieu. La figure 5 illustre cette configuration expérimentale.

Pour les exemples où nous avons testé ce procédé, nous constatons que le modèle change bien sa réponse lorsque le calque de probabilité d'inondation est inversée. Cette observation - si elle tient dans le cas général - permet de réfuter l'hypothèse selon laquelle le modèle n'exploite absolument pas le calque de probabilité d'inondation et ne répondrait que par des connaissances apprises lors du pré-entraînement.

La génération de calques plus précis et subtils que celui de la figure 5 est plus délicate mais mériterait d’être explorée dans de travaux futurs.

4 Conclusion et travaux futurs

Ce projet a permis d’explorer le potentiel des modèles génératifs multimodaux dans un contexte d’analyse cartographique à fort enjeu sociétal : l’évaluation des risques d’inondation. En exposant des données géospatiales sous forme d’images (notamment des cartes de susceptibilité aux inondations) à un modèle multimodal, l’outil développé offre une interface conversationnelle capable de formuler des réponses contextualisées, tout en laissant à l’utilisateur la possibilité de les vérifier visuellement.

Certaines des directions explorées sont prometteuses. 1) Premièrement, on remarque que dans un domaine restreint et avec des sorties bien définies, il est possible d’améliorer significativement la fiabilité des réponses grâce à des techniques de génération structurée et l’intégration d’outils tiers pour le géocodage et la récupération d’images en temps réel. 2) Ensuite, l’approche proposée repose sur un principe essentiel : donner à l’utilisateur la possibilité de vérifier à tout moment la véracité des affirmations du système, en confrontant la réponse produite avec la carte source. Ce principe de transparence est particulièrement important dans les contextes à fort enjeu, où la confiance dans l’outil est primordiale.

La généralisation à des questions ouvertes et dans un contexte plus général reste cependant difficile à évaluer. Nos observations nous indiquent aussi que les hallucinations sont un frein important au développement et à la mise en place de systèmes fiables pour l’interprétation de cartes.

Ce travail constitue un premier constat sur le manque de maturité des techniques actuelles pour l’analyse cartographique interactive à l’aide de modèles multimodaux. Il met en lumière le besoin de développer des techniques fondationnelles pour améliorer la compréhension des cartes avec les modèles multimodaux, et de développer des corpus de plus grande taille avant d’entreprendre un développement dans un contexte applicatif.

Références

- ANTOL S., AGRAWAL A., LU J., MITCHELL M., BATRA D., ZITNICK C. L. & PARIKH D. (2015). VQA : Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.
- CHANG S., PALZER D., LI J., FOSLER-LUSSIER E. & XIAO N. (2022). MapQA : A dataset for question answering on choropleth maps. In *NeurIPS 2022 First Table Representation Workshop*.
- HURST A., LERER A., GOUCHER A. P., PERELMAN A., RAMESH A., CLARK A., OSTROW A., WELIHINDA A., HAYES A., RADFORD A. *et al.* (2024). Gpt-4o system card. *arXiv preprint arXiv :2410.21276*.
- LEWIS P., PEREZ E., PIKTUS A., PETRONI F., KARPUKHIN V., GOYAL N., KÜTTLER H., LEWIS M., YIH W.-T., ROCKTÄSCHEL T. *et al.* (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, **33**, 9459–9474.
- MCGRATH H. & GOHL P. N. (2023). Prediction and classification of flood susceptibility based on historic record in a large, diverse, and data sparse country. *Environmental Sciences Proceedings*, **25**(1), 18.

MOUBTAHIJ A., CUMMINGS C.-W., HANDAN A., GALY E. & CHARTON E. (2024). Participation du crim à defst 2024 : Utilisation de petits modèles de langue pour des qcms dans le domaine médical. In *Actes du Défi Fouille de Textes@ TALN 2024*, p. 11–22.

RAHNEMOONFAR M., CHOWDHURY T., SARKAR A., VARSHNEY D., YARI M. & MURPHY R. R. (2021). Floodnet : A high resolution aerial imagery dataset for post flood scene understanding. *IEEE Access*, **9**, 89644–89654.

WEI J., WANG X., SCHUURMANS D., BOSMA M., ICHTER B., XIA F., CHI E., LE Q. & ZHOU D. (2023). Chain-of-thought prompting elicits reasoning in large language models.

Annexes

A Prompts utilisés

À partir de la carte fournie qui représente les zones inondables (en fonction d'un dégradé de couleur : bleu foncé étant le plus inondable et jaune le moins inondable), génère des questions plausibles à poser à propos de cette carte. Voici les contraintes : 1-Les questions doivent mener à des réponses facilement vérifiables binaires (oui ou non) ou catégoriques (des zones clairement identifiables, rues, etc non ambiguës). 2- Suffixe chaque réponse par le(s) quadrant(s) (Nord Ouest, Nord Est, Sud Est, Sud Ouest) des régions concernées. Je veux 10 occurrences (5 binaires et 5 catégoriques), à générer dans un fichier excel avec 3 colonnes : question, réponses, explication.

FIGURE 6 – Exemple de prompt de génération de paires de questions fermées et leur réponses associées sur des extraits de cartes