# VIN Charles - Mini-bilan

## Article 1 : The Unreliability of Explanations in Few-shot Prompting for Textual Reasoning

L'article présenté par Aladdin offre une perspective intéressante sur le fine-tuning des LLM incluant des explications dans les prompts. Cette thématique est d'autant plus pertinente que les LLM ont gagné en importance, surtout après l'écriture de l'article en 2022. La capacité à les utiliser efficacement et à les prompter devient donc une compétence intéressante à maîtriser.

Ce que j'ai trouvé particulièrement intéressant dans cet article, c'est la mise en lumière des limites des LLM. Par exemple, les auteurs ont noté que les explications générées par les LLM, bien que souvent correctement formulées et convaincantes, peuvent manquer de cohérence avec la réponse donnée à une question spécifique. Cette observation a fait écho à un point que j'avais relevé dans mon propre article : les utilisateurs ont tendance à ne pas s'engager suffisamment sur le plan cognitif avec les explications fournies par les IA. Combinées à une confiance excessive envers les LLM, des réponses/explications erronés pourraient valider de fausses informations.

De plus, l'article d'Aladdin m'a permis de constater que le comportement des LLM peut être assez imprévisible. La différence de résultats entre les conditions Explication-Prompt (E-P) et Prompt-Explication (P-E) illustre bien cette idée. Les auteurs soulignent la nécessité d'outils de calibration complexes et spécifiques à chaque dataset pour obtenir des explications plus cohérentes. Cette observation est d'autant plus pertinente à l'heure où une multitude d'outils "simples" pour le fine-tuning des LLM sont disponibles, suggérant que la réalité de leur utilisation est plus nuancée et complexe.

En conclusion, l'article d'Aladdin a contribué à ma compréhension des défis liés à l'utilisation des LLM. Il souligne l'importance d'un fine-tune minutieux pour obtenir des explications cohérentes aux prédictions générées par ces modèles, un enjeu nécessaire pour leur utilisation fiable et éthique.

## Article 2 : Diffusion Visual Counterfactual Explanations

L'article d'Aymeric a retenu mon attention, notamment grâce aux discussions préalables que nous avons eue, éveillant ainsi ma curiosité avant même sa présentation. L'un des aspects les plus intrigants de cet article est l'application de la méthode des "classifier guidance" dans le domaine de l'XAI. Cette technique, que j'ai découverte lors de mon projet en REDS concernant la génération d'architectures neuronales, semble également avoir du potentiel dans le contexte de l'XAI.

En écoutant les autres présentations, j'ai cru comprendre que la génération de contrefactuels dans le domaine de la classification d'image est une tâche complexe, et les résultats présentés dans ce papier m'ont impressionné. Ils sont super réalistes comme ce que les modèles de diffusion ont l'habitude de nous offrir.

Un autre aspect auquel je n'avais pas pensé est l'utilisation de contrefactuels comme outil de débogage pour détecter les caractéristiques erronées, que le modèle utilise à tors, illustré par l'exemple de l'abeille et des fleurs.

L'article aborde également une technique astucieuse : la projection du gradient sur un cône pour empêcher le modèle de converger vers des modifications non sémantiques triviales. Cette idée, bien que complexe, m'a rappelé certaines projections euclidiennes étudiées lors de mes cours issus du master M2A, l'application sur des cônes fut surprenante et rigolote.

En conclusion, malgré le coût computationnel élevé, la qualité des exemples contrefactuels présentés dans cet article justifie largement cet investissement. Cet article m'a permis de faire de nombreux liens avec d'autre cours et ce fut sympathique.