

SuperGPQA-HCE-FR : un corpus spécialisé en français pour le domaine hydraulique et le génie civil

Markarit Vartampetian¹ Diandra Fabre¹ Philippe Mulhem¹ Sylvain Joubert¹
Didier Schwab¹

(1) Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble
prenom.nom@univ-grenoble-alpes.fr

RÉSUMÉ

Dans cet article, nous présentons SuperGPQA-HCE-FR, une adaptation française d'un sous-ensemble du benchmark SuperGPQA axé sur les domaines de l'ingénierie hydraulique et du génie civil. Il comprend 285 questions à choix multiples conçues pour évaluer et spécialiser des modèles de langue multilingues de grande taille (LLMs) sur des tâches techniques. La traduction réalisée automatiquement est ensuite évaluée par des experts des domaines. Enfin, nous présentons les premiers résultats sur des modèles Instruct généralistes multilingues en comparant les performances du corpus original en anglais à celles du corpus traduit en français.

ABSTRACT

SuperGPQA-HCE-FR : A Specialized French Corpus for Hydraulic and Civil Engineering

We present SuperGPQA-HCE-FR, a French adaptation of a subset of the SuperGPQA benchmark focusing on hydraulic engineering and civil engineering. The dataset includes 285 multiple-choice questions designed to assess and specialize multilingual large language models (LLMs) for technical tasks. Questions were automatically translated and post-edited by domain experts. We report baseline results for multilingual instruction-tuned models, comparing performances on the original English subdataset and the French version.

MOTS-CLÉS : LLM, corpus spécialisé, français, évaluation, ingénierie hydraulique, génie civil.

KEYWORDS: LLM, specialized dataset, French, evaluation, hydraulic engineering, civil engineering.

ARTICLE : **Soumis à EvalLLM2025.**

1 Introduction

Les corpus textuels jouent un rôle fondamental dans de nombreuses applications en traitement automatique du langage naturel (TALN). Ils constituent une base essentielle pour le développement des grands modèles de langue (*Large Language Models*; LLM), en fournissant des contextes d'usage représentatifs nécessaires à leur création, leur adaptation fine ainsi qu'à l'évaluation de leurs performances dans des domaines et tâches variés. Ces modèles de langue ont connu un essor important ces dernières années, en particulier dans des applications d'assistance au quotidien dans les milieux professionnels, comme le métier d'ingénieur. Dans cette étude, nous nous intéressons à deux secteurs

d'ingénierie : l'ingénierie hydraulique et le génie civil. Le domaine de l'hydraulique étudie les propriétés mécaniques des liquides et des fluides, et son ingénierie peut couvrir la conception de barrages comme la valorisation énergétique, les calculs de débit dans des conduites, la microfluidique ou les systèmes de pompage (Prasuhn, 1987; Chanson, 2007). Le domaine du génie civil couvre quant à lui l'ensemble des techniques de construction civiles, de la conception à la réalisation d'ouvrages d'infrastructures (ponts, tunnels) ainsi que leur réhabilitation (Chen & Liew, 2002). L'accès à des données textuelles pertinentes à ces domaines, pour le préentraînement comme pour l'évaluation, se révèle particulièrement limité et difficile, notamment par des contraintes liées à la confidentialité et la sensibilité des procédures opérationnelles internes. Cette absence de ressources spécialisées limite la capacité des modèles de langues génératifs — préentraînés sur de vastes corpus de données « généralistes » — à produire des textes précis ou à comprendre les subtilités de ces domaines, qui possèdent un vocabulaire propre et des connaissances techniques et spécialisées, souvent absents des données d'entraînement « standard ».

À notre connaissance, aucun corpus disponible à ce jour ne couvre de manière systématique et ciblée les domaines de l'ingénierie hydraulique et le génie civil pour la langue française. Face à ces limites, nous proposons une adaptation en français du corpus en anglais SuperGPQA (M-A-P Team *et al.*, 2025), conçu pour traiter spécifiquement des contextes techniques y compris les domaines de l'ingénierie hydraulique et du génie civil. Dans cet article, nous traduisons le corpus en français puis procédons à une post-édition par des experts humains du domaine afin d'assurer la qualité terminologique et conceptuelle du contenu. Enfin, nous évaluons les capacités de modèles de langue génératifs à répondre à des questions à choix multiples portant sur ces deux domaines. Ainsi, à travers cette démarche, nous poursuivons un double objectif : 1. fournir une ressource d'évaluation pour le français et sur ces domaines spécialisés ; 2. contribuer à une meilleure compréhension des limites des modèles de langues génératifs en contexte technique et industriel en anglais mais aussi en français, langue peu représentée dans les benchmarks spécialisés industriels.

L'article est organisé comme suit : après une présentation de l'état de l'art dans la section 2, nous présentons notre méthodologie puis détaillons notre configuration expérimentale. Nous consacrons, enfin, la section 6 à la présentation des résultats et leur analyse.

2 État de l'art

2.1 Évaluation de Grands Modèles de Langue

L'évaluation des grands modèles de langue est essentielle pour comprendre et qualifier leurs capacités et leurs limites (Holtzman *et al.*, 2023; Chang *et al.*, 2024). Bien qu'une présentation exhaustive des méthodes existantes dépasse le cadre de cet article, ces approches peuvent être regroupées en trois catégories principales : l'évaluation humaine, l'évaluation par des grands modèles de langue eux-mêmes, et l'évaluation à l'aide de corpus d'évaluation (benchmarks).

L'évaluation humaine constitue traditionnellement le standard de référence pour l'évaluation des modèles de langues (Gkatzia & Mahamood, 2015; van der Lee *et al.*, 2018; Hashimoto *et al.*, 2019). En particulier, le retour humain peut être utilisé pour améliorer à la fois les performances des modèles (Kreutzer *et al.*, 2018; Ziegler *et al.*, 2020; Stiennon *et al.*, 2020; Ouyang *et al.*, 2022) ou servir en tant qu'indicateur de la qualité des sorties générées par les grands modèles de langue (Thomson & Reiter, 2020; Khashabi *et al.*, 2022; Gehrmann *et al.*, 2021; Zellers *et al.*, 2021; Zheng *et al.*, 2023). Il permet également de recueillir des jugements nuancés sur la qualité linguistique, la cohérence, la véracité, ou encore la pertinence des réponses dans des domaines généraux ou spécialisés. Cette approche d'évaluation est particulièrement adaptée aux tâches de génération ouvertes - c'est-à-dire

où il n'existe pas strictement une réponse correcte unique - ou à celles nécessitant une expertise spécifique. Une difficulté inhérente à une évaluation humaine est qu'elle est coûteuse en temps et en ressources humaines. D'autre part, une telle évaluation est également sensible à de nombreux facteurs contextuels, tels que les biais des évaluateurs ou les conditions d'application (Nichols & Maner, 2008; Henrich *et al.*, 2010; Schoch *et al.*, 2020; Casal & Kessler, 2023; Hattori *et al.*, 2024; Joshi, 2025), ce qui la rend difficilement reproductible (Belz & Thomson, 2023).

Une approche alternative à l'évaluation humaine a récemment émergé : l'utilisation des LLM en tant qu'évaluateurs ou juges. Lorsqu'ils sont guidés par des *prompts* « soigneusement » conçus et élaborés, les LLM démontrent une capacité à produire des évaluations proches de celles des humains, avec une corrélation relativement élevée entre les deux (Zheng *et al.*, 2023; Chiang & Lee, 2023; Dubois *et al.*, 2023; Liu *et al.*, 2023; Sottana *et al.*, 2023; Hada *et al.*, 2024; Fu *et al.*, 2024). Cependant, plusieurs travaux mettent en évidence des limites des LLM, telles que la génération d'hallucinations (Cao *et al.*, 2022; Rawte *et al.*, 2023; Farquhar *et al.*, 2024), l'apparition de biais (Li *et al.*, 2024a; Gallegos *et al.*, 2024; Koo *et al.*, 2024; Panickssery *et al.*, 2024), ainsi qu'un manque de robustesse (Zhu *et al.*, 2024; Li *et al.*, 2024b; Raina *et al.*, 2024) et de connaissances spécialisées. En outre, d'autres limites importantes subsistent, notamment la nature *closed-source* de certains modèles (p. ex. GPT-4), leur manque de transparence, leur coût élevé ainsi que leur impact environnemental. Alors que les méthodes d'évaluation basées sur les LLM sont récentes et spécifiques au domaine du TALN, l'évaluation comparative s'appuie quant à elle sur des méthodes standardisées, reposant sur l'utilisation de corpus d'évaluation, ou benchmarks, largement employés pour comparer les performances des systèmes. Ces corpus servent à mesurer de manière systématique les performances des modèles de langue à travers un ensemble de tâches variées, en domaine général ou spécialisé, comme le raisonnement, la génération de langage naturel, couvrant plusieurs sous-tâches telles que la génération automatique de résumés ou de dialogues, la traduction automatique et la tâche de question-réponse. Les corpus d'évaluation jouent également un rôle fondamental dans le développement et l'adaptation des modèles, en simulant des scénarios proches et des contextes d'usage représentatifs des usages réels. Leur utilisation permet d'établir une base comparative rigoureuse et reproductible, orientant ainsi les décisions liées au développement et au déploiement des modèles (Ali *et al.*, 2024). Enfin, l'évaluation des modèles sur certains benchmarks est considérée comme une exigence fondamentale, dans la mesure où elle constitue une condition importante voire nécessaire à la validation expérimentale de leurs performances et à leur « reconnaissance » par la communauté. Parmi ces benchmarks, les formats à choix multiples (QCM) jouent un rôle particulier, car ils permettent de réaliser une évaluation contrôlée et reproductible des sorties des grands modèles de langue. La section suivante présente les benchmarks et corpus existants principaux pour l'évaluation par question-réponse, en mettant l'accent notamment sur les ressources disponibles en français.

2.2 Corpus d'évaluation : le format QCM

Le format QCM est largement utilisé pour l'évaluation des modèles de langue (Rein *et al.*, 2023), car il permet de simuler des contextes proches des examens humains, mais aussi d'évaluer les réponses des modèles de manière rapide et contrôlée (p. ex. options limitées et prédéfinies). Néanmoins, si des corpus de questions/réponses existent en français (Asai *et al.* (2018), (d'Hoffschmidt *et al.*, 2020), (Keraron *et al.*, 2020), (Béchet *et al.*, 2022), (Faysse *et al.*, 2024)), la majorité d'entre eux se concentre sur des thématiques ou des connaissances générales et encyclopédiques. D'autres ressources ciblent des domaines spécifiques, tels que Autogestion (Antoine *et al.*, 2022) pour les sciences sociales, FrenchMedMCQA (Labrak *et al.*, 2022) pour la pharmacie, LLeQA (Louis *et al.*, 2024) pour le droit ou encore CQuAE (Gerald *et al.*, 2024) pour l'éducation. Bien que précieuses, les ressources

existantes restent centrées sur des contenus généraux et sont peu adaptées à l'évaluation des capacités des modèles dans des domaines spécialisés. Malgré les efforts de la communauté francophone, à notre connaissance, aucun corpus n'est publiquement disponible pour ces deux domaines — une lacune importante à la fois pour le développement et l'évaluation de modèles adaptés à ces contextes spécifiques. Une limitation similaire est observée dans les corpus anglophones. Par exemple, MMLU (Hendrycks *et al.*, 2021) et MMLU-Pro (Wang *et al.*, 2024) couvrent une large variété de disciplines académiques, tandis que GPQA (Rein *et al.*, 2023) se concentre sur la biologie, la physique et la chimie. Cependant, aucune de ces ressources ne traite spécifiquement du génie civil ou de l'ingénierie hydraulique. Plus récemment, SuperGPQA (M-A-P Team *et al.*, 2025) a adopté une approche plus étendue couvrant 285 disciplines ; y compris l'ingénierie hydraulique et le génie civil. Par ailleurs, sa date de publication récente contribue à réduire le risque associé de contamination des données (*data contamination*), un phénomène désignant l'intégration accidentelle ou intentionnelle de corpus d'évaluation dans les jeux de corpus d'évaluation dans les jeux de données d'entraînement, pouvant aboutir à une surestimation artificielle des performances (Brown *et al.*, 2020; Magar & Schwartz, 2022; Balloccu *et al.*, 2024). Afin de faciliter la comparaison des performances de différents LLM sur des benchmarks, des cadres d'évaluation unifiés ont été développés, tels que LM Eval Harness (Gao *et al.*, 2024)¹ et OpenCompass (Contributors, 2023) permettent une évaluation holistique des modèles en raisonnement, en génération de texte ou en connaissances spécialisées. D'autres initiatives, comme BIG-bench (Beyond the Imitation Game Benchmark) (Srivastava *et al.*, 2023), HELM (Holistic Evaluation of Language Models) (Liang *et al.*, 2023) ou BUFFET (Asai *et al.*, 2024), visent à agréger un large éventail de tâches pour une évaluation plus globale. Cependant, l'absence de données pertinentes pour certains domaines spécialisés, tels que l'ingénierie hydraulique, demeure une difficulté récurrente. De plus, de manière générale, l'ensemble de ces cadres d'évaluation holistiques est majoritairement centré sur l'anglais.

2.3 Typologies des questions

La typologie des questions constitue un élément central dans les travaux en recherche d'information et en question-réponse. De nombreuses taxonomies ont été proposées pour classifier les questions selon le type d'information recherchée et la forme de la réponse attendue. Ces classifications s'appuient sur plusieurs dimensions complémentaires, allant de la nature de la tâche à accomplir jusqu'aux indices linguistiques présents dans la formulation de la requête. Les corpus d'évaluation pour les modèles de langue reposent sur des requêtes de type QCM (comme SuperGPQA) qui exploitent principalement des tâches à objectif spécifique. La réponse correcte correspond à l'atteinte d'un but précis, défini et concret, par la sélection d'une option parmi un ensemble limité et prédéfini. Ce format impose un chemin unique vers la réponse et permet d'évaluer les performances des modèles de manière rapide, contrôlée et reproductible. Dans SuperGPQA, une classification des requêtes à gros grains indique si une requête est computationnelle (c'est-à-dire, elle nécessite des calculs), notée Q-COMP dans la suite, ou non, notée Q-NON-COMP. Il est possible de raffiner une telle classification en utilisant des indices linguistiques pour la classification des questions (Li & Roth, 2006; Zhang & Lee, 2003; Tayyar Madabushi & Lee, 2016; Gerald *et al.*, 2024). Les principaux marqueurs linguistiques couramment exploités sont : les mots interrogatifs (« qui », « quoi », « où », « quand », « comment », « pourquoi », etc.) indiquant le type de réponse attendu (factuelle, définitionnelle, procédurale, causale, etc.); la structure syntaxique de la question (présence de subordonnées, inversion sujet-verbe,

1. À noter que certaines ressources, comme LM Eval Harness (Gao *et al.*, 2024), proposent bien des questions en hydrologie et en génie civil, mais celles-ci sont disponibles en coréen, ce qui limite leur portée pour l'évaluation de modèles dans des contextes francophones.

syntagmes nominaux ciblés) ou encore des patrons linguistiques récurrents (« Quelle est la définition de... » pour une question définitionnelle, « Comment procéder à... » pour une question procédurale, etc.). Nous allons utiliser certains de ces éléments syntaxiques pour définir des sous-classes pour les requêtes non-computationnelles de SuperGPQA de notre benchmark.

3 Cadre de référence : le corpus SuperGPQA

Dans cette section, nous présentons le corpus SuperGPQA (M-A-P Team *et al.*, 2025), que nous utilisons comme point de départ pour proposer une version adaptée en langue française, axée sur les domaines du génie civil et de l'ingénierie hydraulique. En particulier, SuperGPQA constitue un jeu de données en anglais² conçu pour évaluer les capacités de raisonnement et de compréhension des grands modèles de langue. Il contient 24.000 questions à choix multiples couvrant 285 disciplines, telles que l'ingénierie, les sciences naturelles, la médecine, les sciences humaines et les sciences humaines et sociales. Ce corpus a été construit à partir de sources validées par des experts, et annoté via un processus collaboratif combinant des contributions experts humains et des approches de filtrages automatiques (à base de règles et par des grands modèles de langue).

3.1 Format et métadonnées

Les paires question-réponse sont disponibles dans un fichier json. Chaque entrée, illustrée en figure 2, comprend un identifiant unique (*uuid*), l'énoncé de la question (*question*), une liste à dix options de réponses (*options*), la réponse correcte sous forme textuelle (*answer*), ainsi que la lettre associée à cette réponse dans la liste des options (*answer_letter*). D'autres métadonnées sont également incluses, telles que le domaine (*discipline*), le secteur (*field*) et le sous-secteur (*subfield*), le niveau de difficulté (*difficulty*), ainsi qu'un indicateur précisant s'il s'agit ou non d'une question de calcul (*is_calculation*). Les tableaux en figure 1 présentent la répartition des paires question-réponse issues du corpus original, regroupées par sous-domaine au sein des deux disciplines ciblées³. Pour chaque sous-domaine, sont indiqués le nombre total de questions (*NB-Q*) ainsi que leur catégorisation selon la métadonnée, distinguant les questions computationnelles (*Q-COMP*) des questions non computationnelles (*Q-NON-COMP*). SuperGPQA est formé de requêtes accompagnées d'un ensemble de réponses possibles (*options*), parmi lesquelles une seule est correcte.

4 SuperGPQA-HCE-FR

Dans cette section, nous présentons SuperGPQA-HCE-FR, notre adaptation française d'un sous-ensemble du corpus SuperGPQA, ciblant les domaines de l'ingénierie hydraulique et du génie civil (HCE : *Hydraulic & Civil Engineering*). Nous décrivons la méthodologie suivie pour sa construction et la sélection des paires de questions-réponses, leur traduction automatique, la post-édition manuelle des textes traduits, ainsi que des statistiques descriptives des paires retenues.

2. Certaines questions initialement rédigées dans d'autres langues, notamment le chinois, ont été traduites vers l'anglais avec une adaptation au registre académique. Notons qu'actuellement les métadonnées disponibles ne permettent pas d'identifier la langue source des questions traduites ni les éventuelles adaptations de registre appliquées.

3. Les domaines, secteurs et de sous-secteurs sont traduits en français pour des raisons de présentation dans l'article. Cependant, les métadonnées originales sont conservées en anglais dans notre version française afin de faciliter la traçabilité et la reproduction des résultats. Nous précisons en annexe les équivalences entre les libellés anglais et français.

Ingénierie Hydraulique (218)			
Sous-domaine	NB-Q	Q-COMP	Q-NON-COMP
Hydraulique et hydrologie	136	72	64
Aménagement hydraulique et hydroélectricité	82	40	42
Somme-NB-Q	218	112	106

Génie Civil (358)			
Sous-domaine	NB-Q	Q-COMP	Q-NON-COMP
Génie des ponts et tunnels	51	3	48
Géotechnique	179	129	50
Ingénierie des structures	57	32	25
Génie des infrastructures urbaines	71	5	66
Somme-NB-Q	358	169	189

FIGURE 1 – Répartition des questions Q-COMP et Q-NON-COMP dans les domaines de l’ingénierie hydraulique et du génie civil ainsi que leurs sous-domaines respectifs.

4.1 Méthodologie

La constitution du corpus SuperGPQA-HCE-FR repose dans un premier temps sur l’extraction d’un sous-ensemble ciblé du corpus source, focalisé sur les domaines de l’ingénierie hydraulique et du génie civil grâce à la métadonnée du domaine (*discipline*). Dans un deuxième temps, un filtrage a été appliqué afin de ne retenir que les questions non computationnelles (*is_calculation*), notre objectif étant de construire un corpus d’évaluation adapté aux tâches de compréhension et de génération du langage naturel indépendamment des compétences arithmétiques ou algorithmiques des LLM. Nous avons ensuite procédé à une traduction automatique de l’anglais vers le français des paires question-réponse en utilisant l’API d’OpenAI et le modèle GPT-4o (gpt-4o-2024-08-06). Nous avons adopté une approche en *two-shot*, afin de préserver la structure des données (formatage, ordre des éléments, etc.) ainsi que la conformité du fichier json généré. Le prompt employé pour cette étape est fourni en [annexe](#). Nous avons enfin demandé à deux ingénieurs experts de chaque domaine d’évaluer la qualité et la pertinence de ces traductions, selon les critères suivants : précision du vocabulaire et fluidité de la traduction. En raison du nombre limité de questions à notre disposition, nous n’en avons exclu aucune sur la base de critères géographiques ou culturels.

4.2 Analyse qualitative et quantitative du corpus

Étant donné que le corpus suit un format de type QCM, nous avons étudié la présence des mots ou des expressions déclencheurs, notamment les mots interrogatifs, afin d’orienter une première classification en grandes catégories (Factuelle, Définition, Explication, Question à trou). Nous calculons ensuite le nombre de mots et la longueur de chaque question, et présentons les résultats par type de question. La répartition des questions de notre corpus selon ces quatre catégories est présentée en figures 3 et 4.

Ces observations suggèrent que la longueur des questions constitue un indice discriminant partiel, pouvant être utilisé en complément des marqueurs lexicaux et syntaxiques pour affiner cette classification. Certains types de questions sont facilement identifiables à l’aide de mots-clés, mais d’autres nécessitent une analyse plus fine, et la précision de la classification peut varier selon les catégories.

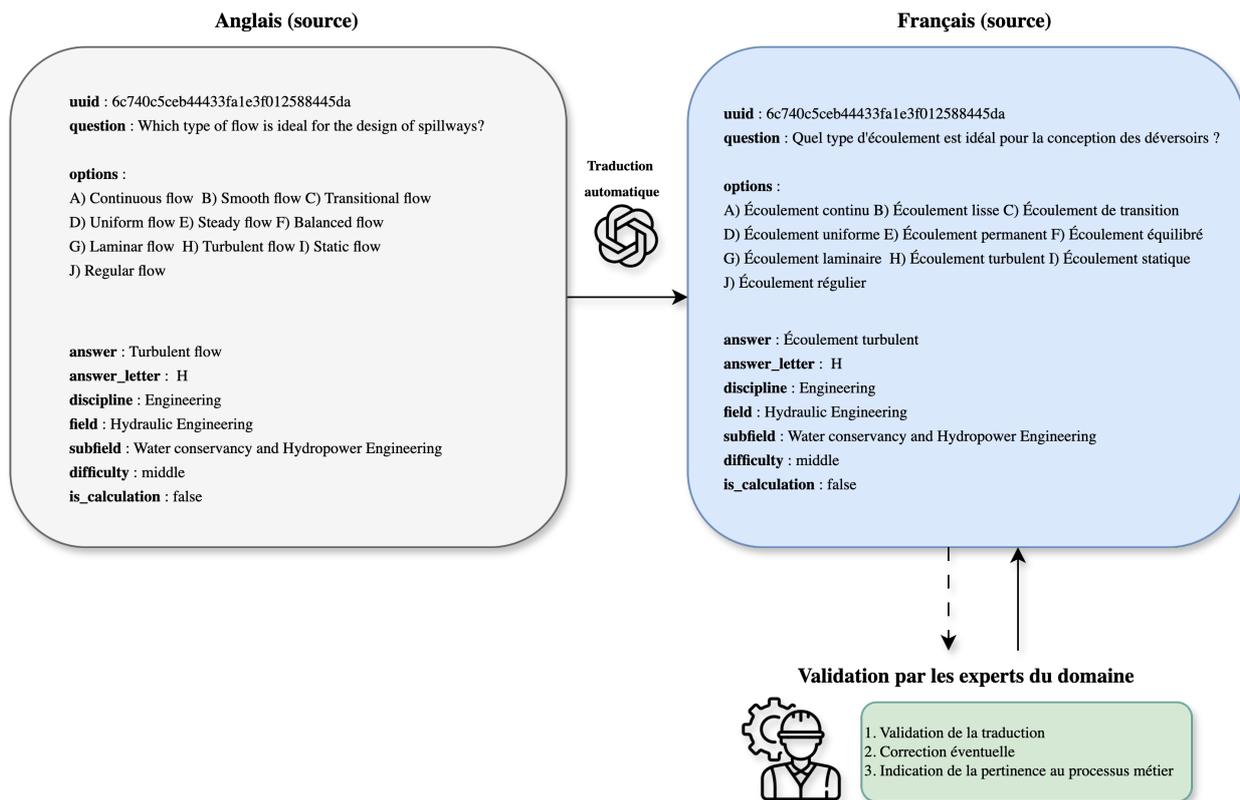


FIGURE 2 – Exemple tiré du corpus initial en anglais. Processus de traduction automatique en français avec GPT-4o, suivi d’une validation humaine en trois étapes : 1. validation de la traduction ; 2. correction éventuelle ; 3. indication de la pertinence de la question par rapport au processus métier de l’évaluateur.

L’objectif à terme est d’adapter cette taxonomie en collaboration avec des experts du domaine, afin de mieux refléter les besoins spécifiques et les caractéristiques propres aux questions en ingénierie hydraulique et en génie civil, comme cela existe déjà dans le domaine médical (Roberts *et al.*, 2014; Ben Abacha & Demner-Fushman, 2019).

5 Configuration expérimentale

Nous évaluons la capacité de cinq modèles de langue décodeurs à générer la réponse correcte aux questions de SuperGPQA-HCE-FR, avec deux types d’expériences : *zero-shot* et *five-shot*.

Choix de modèles Nous évaluons les modèles présentés dans le tableau 5. Les LLM d’un nombre de paramètres proche de 7-8 milliards de paramètres constituent aujourd’hui un standard et un bon compromis entre performances et taille. Nous avons aussi sélectionné Qwen2.5-1.5B pour comparer avec CroissantLLM, ainsi que Qwen2.5-14B qui nous permet d’étudier l’impact de la taille du modèle sur la résolution de QCM en ingénierie hydraulique et en génie civil. Les modèles sélectionnés sont tous open source. En outre, étant donné que notre expérimentation repose sur un *prompt* précis et une contrainte de formatage des réponses, nous avons sélectionné des modèles de type *Instruct*. Ces modèles - dérivés de modèles dits de *base*- ont été spécifiquement affinés sur des jeux de triplets (instruction, entrée, réponse attendue) et sont donc mieux adaptés pour réaliser des tâches explicitement décrites et formulées via un prompt. À l’inverse, les modèles de *base* sont entraînés sur

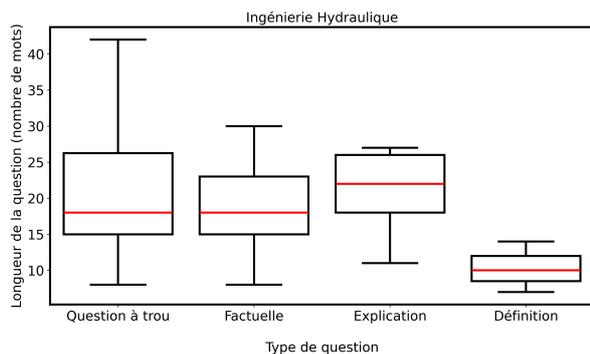


FIGURE 3 – Ingénierie hydraulique. Distribution de la longueur des questions (en mots) par type.

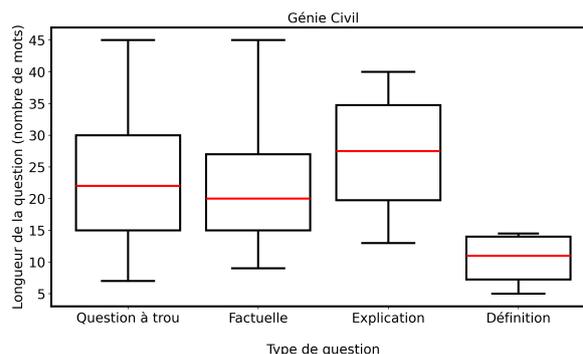


FIGURE 4 – Génie civil. Distribution de la longueur des questions (en mots) par type.

des données non annotées selon un objectif de modélisation de la langue (prédiction du *token* suivant). Par conséquent, ils deviennent plus susceptibles à produire des sorties dégénérées (p. ex. répétitions excessives, incohérences sémantiques, mélanges de langues) (Welleck *et al.*, 2019; Holtzman *et al.*, 2020; Riley & Chiang, 2022; Li *et al.*, 2023).

Modèle	Nb paramètres	Open source	Open data
CroissantLLM-Chat (Faysse <i>et al.</i> , 2025)	1,3B	✓	✓
Qwen2.5-Instruct (Qwen <i>et al.</i> , 2025)	1,5B, 7B, 14B	✓	
Mistral-7B-Instruct (Jiang <i>et al.</i> , 2023)	7B	✓	
Llama-3.1-8B-Instruct (Grattafiori <i>et al.</i> , 2024)	8B	✓	

Prompt et paramètres de génération Malgré la diversité potentielle des formats de sortie selon les modèles, nous conservons la version initiale et uniformisée des *prompts* pour l’ensemble des expériences (cf. annexe) pour une comparaison homogène et cohérente. Par ailleurs, pour toutes les expérimentations, nous avons fixé la température à 0 puisqu’elle ne dégrade pas les performances pour la tâche (Renze, 2024), et assure la reproductibilité des expérimentations.

Mesure d’exactitude Dans (M-A-P Team *et al.*, 2025), l’évaluation repose sur une extraction d’options choisies par le modèle (de A à J), à partir de textes « libres » contenant de formulations variées. L’approche s’appuie sur des expressions régulières appliquées d’abord à la fin de la réponse, reposant sur l’hypothèse que le modèle conclut sa réponse généralement par une affirmation explicite de leur choix d’après les instructions données dans le *prompt*. Si aucune réponse n’est détectée, l’analyse est ensuite étendue à l’ensemble du texte pour repérer une éventuelle mention plus implicite dans le corps de la réponse. Dans les deux cas, un ensemble de patrons est défini pour capturer différentes formulations typiques, telles que : [...] « La bonne réponse est donc H) Centrale Hydroélectrique de Barrage. »; « Ainsi, la réponse correcte est H. »; « Réponse : H »; « La réponse correcte est : H) Centrale Hydroélectrique de Barrage. ». La métrique utilisée est *exact match*, évaluant une réponse comme correcte uniquement si la lettre extraite correspond exactement à la lettre attendue.

6 Résultats

L’évaluation proposée respecte la méthodologie suivante : 1. génération de réponses pour chaque question ; 2. calcul des métriques pour chaque réponse (dans notre cas, l’exactitude) 3. et agrégation

des résultats pour obtenir un score moyen par modèle, par langue et par domaine visé.

Impact du n -shot Les performances observées en *zero-shot* et *five-shot* sont résumées dans le tableau 2. Nous constatons que les performances en *zero-shot* surpassent généralement celles en *five-shot* en français, dans les deux domaines. Cette différence est moins évidente pour l’anglais.

Modèle	Hydrau-EN	Hydrau-FR	Civil-EN	Civil-FR
CroissantLLM-Chat-five-shot	10,38%	5,66%	14,29%	8,47%
CroissantLLM-Chat-zero-shot	4,72%	0,94%	6,36%	4,76%
Llama-3.1-8B-Instruct-five-shot	23,58%	19,81%	22,75%	20,63%
Llama-3.1-8B-Instruct-zero-shot	30,19%	26,42%	22,22%	23,81%
Mistral-7B-Instruct-five-shot	25,47%	24,53%	17,99%	16,93%
Mistral-7B-Instruct-zero-shot	24,53%	31,13%	15,34%	21,69%
Qwen2.5-1.5B-Instruct-five-shot	23,58%	18,87%	17,99%	14,29%
Qwen2.5-1.5B-Instruct-zero-shot	19,81%	21,70%	22,22%	21,69%
Qwen2.5-7B-Instruct-five-shot	30,19%	34,91%	29,10%	22,22%
Qwen2.5-7B-Instruct-zero-shot	33,96%	37,74%	30,16%	22,75%
Qwen2.5-14B-Instruct-five-shot	31,13%	34,91%	33,33%	27,51%
Qwen2.5-14B-Instruct-zero-shot	36,79%	30,19%	37,04%	27,51%

TABLE 1 – Performances des modèles sur les domaines de l’ingénierie hydraulique (Hydrau) et le génie civil (Civil) en anglais (EN) et en français (FR)

Ainsi, en hydraulique et en français, Llama-3.1-8B-Instruct atteint 30.19% en *zero-shot*, contre 23.58% en *five-shot*. Il en est de même pour le modèle Mistral-7B-Instruct (31.13% *versus* 24.53%). Nous faisons l’hypothèse que l’ajout d’exemples dans le *prompt* surcharge le contexte et dilue les instructions, rendant le format de réponse attendu moins clair, en particulier lorsque les formulations divergent légèrement. Pour tester cette hypothèse, des expérimentations complémentaires en *two-shot* ont été menées (cf. [annexe](#)), mais elles se révèlent souvent moins performantes que les deux autres configurations. Cela suggère que la seule longueur du prompt n’explique pas les écarts observés. Une explication plus plausible réside dans le fait que les prompts few-shot, uniformisés ne sont pas toujours adaptés au format d’instruction propre à chaque LLM. Cette observation ouvre des perspectives pour l’adaptation des prompts à ces formats ou l’usage de méthodes de génération structurée (balisage, formats normalisés, décodage contraint) afin d’optimiser l’extraction des réponses, souvent limitées à une seule lettre.

Impact de la taille du modèle Nous observons que l’augmentation de la taille du modèle au sein d’une même famille (ex. Qwen-1.5B, Qwen-7B, Qwen-14B) est corrélée à une amélioration des performances (resp. 19.81%, 33.96%, 36.79% pour l’ingénierie hydraulique) ([Kaplan et al., 2020](#)), indiquant que les modèles de plus grande taille sont capables de mieux traiter des questions spécifiques aux domaines ciblés, même si les résultats globaux restent relativement faibles. Notons également que certains modèles de taille équivalente (7–8 milliards de paramètres) présentent des performances hétérogènes. Ainsi, Qwen-2.5-7B-Instruct et Llama-3.1-8B-Instruct surpassent globalement Mistral-7B. Cette disparité suggère que d’autres facteurs, tels que la qualité des données d’entraînement, la présence de données francophones, ou encore la pertinence des contenus plus proches de ces domaines spécialisés semblent contribuer de manière déterminante aux performances observées. Pour les modèles de plus petite taille (1.3B–1.5B), les performances sont plus faibles, en *zero* comme en *five-shot*. Cela illustre les limites de ces modèles pour les tâches de raisonnement à choix multiples

ainsi que leurs connaissances en contexte spécialisé. Par exemple, CroissantLLM-Chat (entraîné sur l’anglais et le français), présente des performances très faibles, suggérant une couverture encore plus restreinte des domaines cibles ou un entraînement moins orienté vers le raisonnement.

Analyse qualitative des erreurs et des formats de réponse Nous avons également mené une analyse qualitative des réponses générées afin de mieux comprendre les limites des modèles et les types d’erreurs fréquemment observés. Malgré un *prompt* explicite contraignant le format de réponse (lettre unique précédée de « *Answer :* »), les modèles présentent une grande variabilité dans la structuration de leur sortie. Parmi les écarts les plus fréquents, on observe : 1. des réponses formulées sous forme de phrases complètes (ex. «La bonne réponse est (B)»); 2. l’ajout d’un raisonnement avant l’annonce de la réponse; 3. l’usage de parenthèses ou de numérotation non conformes; 4. des duplications d’options. Certains modèles comme Mistral-7B-Instruct présentent une plus grande variabilité dans leurs réponses, en développant un raisonnement ou en reformulant la question. D’autres, comme CroissantLLM-Chat, favorisent une formulation directe de l’option jugée correcte, suivie d’une répétition de segments de l’énoncé, des options ou de la production de texte hors sujet, sans justification explicite liée au choix effectué. Qwen-2.5 7B et 14B respectent globalement le format attendu. Ces observations mettent en évidence des disparités inter-modèles dans la gestion des consignes, certains produisant des réponses concises et plus conformes au format demandé, d’autres adoptant des stratégies discursives plus libres ou imprécises.

Par rapport aux résultats rapportés sur SuperGPQA pour l’anglais sur l’ensemble des questions (computationnelles ou non), nous constatons que les meilleures performances obtenues en *zero-shot* par Qwen-2.5-14B-Instruct sont relativement proches, confirmant que notre adaptation du corpus en français dans ces domaines spécialisés constitue un corpus d’évaluation exigeant mais pertinent pour évaluer les capacités de raisonnement des LLM dans ces deux domaines industriels.

7 Conclusion

Dans cet article, nous avons présenté SuperGPQA-HCE-FR, une adaptation en français d’un sous-ensemble du corpus SuperGPQA, centrée sur les domaines de l’ingénierie hydraulique et du génie civil. Cette ressource comprend 285 questions non computationnelles à choix multiples, traduites automatiquement et validées par des experts du domaine. Nos expérimentations montrent que, malgré des performances encore limitées, certains modèles multilingues généralistes parviennent à traiter des questions techniques complexes. Les résultats varient en fonction de l’architecture et la taille des modèles ainsi que du paramétrage des configurations (*zero-shot* et *five-shot*), mais demeurent relativement homogènes entre les deux domaines et les langues cibles. À court terme, nous prévoyons d’enrichir le corpus en augmentant le nombre et la diversité des questions, afin de mieux équilibrer les types de questions et analyser plus finement les performances des modèles selon ces typologies. À plus long terme, nous souhaitons explorer l’impact de la spécialisation des LLM (open source et/ou open data) sur ces domaines (*fine-tuning*, *RAG*) - en étendant aussi le benchmark à d’autres tâches de génération du texte - afin d’améliorer leur capacité à répondre de manière fiable et pertinente dans des contextes professionnels et industriels spécialisés.

8 Remerciements

Ce travail est réalisé dans le cadre de la Chaire AugmentIA dirigée par Didier Schwab et portée par la Fondation Grenoble INP grâce au mécénat du Groupe Artelia. Cette chaire bénéficie également d’une aide de l’État gérée par l’Agence Nationale de la Recherche au titre de France 2030 portant la référence ANR-23-IACL-0006 (MIAI Cluster).

Références

- ALI M., RAO P., MAI Y. & XIE B. (2024). Using benchmarking infrastructure to evaluate llm performance on cs concept inventories : Challenges, opportunities, and critiques. In *Proceedings of the 2024 ACM Conference on International Computing Education Research-Volume 1*, p. 452–468.
- ANTOINE E., AUGUSTE J., BECHET F. & DAMNATI G. (2022). Génération de question à partir d'analyse sémantique pour l'adaptation non supervisée de modèles de compréhension de documents (question generation from semantic analysis for unsupervised adaptation of document understanding models). In Y. ESTÈVE, T. JIMÉNEZ, T. PARCOLLET & M. ZANON BOITO, Édts., *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, p. 104–115, Avignon, France : ATALA.
- ASAI A., ERIGUCHI A., HASHIMOTO K. & TSURUOKA Y. (2018). Multilingual extractive reading comprehension by runtime machine translation.
- ASAI A., KUDUGUNTA S., YU X., BLEVINS T., GONEN H., REID M., TSVETKOV Y., RUDER S. & HAJISHIRZI H. (2024). BUFFET : Benchmarking large language models for few-shot cross-lingual transfer. In K. DUH, H. GOMEZ & S. BETHARD, Édts., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (Volume 1 : Long Papers)*, p. 1771–1800, Mexico City, Mexico : Association for Computational Linguistics. DOI : [10.18653/v1/2024.naacl-long.100](https://doi.org/10.18653/v1/2024.naacl-long.100).
- BALLOCCU S., SCHMIDTOVÁ P., LANGO M. & DUSEK O. (2024). Leak, cheat, repeat : Data contamination and evaluation malpractices in closed-source LLMs. In Y. GRAHAM & M. PURVER, Édts., *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 67–93, St. Julian's, Malta : Association for Computational Linguistics.
- BÉCHET F., ROBERT L., ROJAS-BARAHONA L. & DAMNATI G. (2022). Calor-Dial : a corpus for Conversational Question Answering on French encyclopedic documents. In *CIRCLE (Joint Conference of the Information Retrieval Communities in Europe)*, Samatan, France. HAL : [hal-03714189](https://hal.archives-ouvertes.fr/hal-03714189).
- BELZ A. & THOMSON C. (2023). The 2023 ReproNLP shared task on reproducibility of evaluations in NLP : Overview and results. In A. BELZ, M. POPOVIĆ, E. REITER, C. THOMSON & J. SEDOC, Édts., *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, p. 35–48, Varna, Bulgaria : INCOMA Ltd., Shoumen, Bulgaria.
- BEN ABACHA A. & DEMNER-FUSHMAN D. (2019). A question-entailment approach to question answering. *BMC bioinformatics*, **20**, 1–23.
- BROWN T., MANN B., RYDER N., SUBBIAH M., KAPLAN J. D., DHARIWAL P., NEELAKANTAN A., SHYAM P., SASTRY G., ASKELL A. *et al.* (2020). Language models are few-shot learners. *Advances in neural information processing systems*, **33**, 1877–1901.
- CAO M., DONG Y. & CHEUNG J. (2022). Hallucinated but factual ! inspecting the factuality of hallucinations in abstractive summarization. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Édts., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 3340–3354, Dublin, Ireland : Association for Computational Linguistics. DOI : [10.18653/v1/2022.acl-long.236](https://doi.org/10.18653/v1/2022.acl-long.236).
- CASAL J. E. & KESSLER M. (2023). Can linguists distinguish between chatgpt/ai and human writing ? : A study of research ethics and academic publishing. *Research Methods in Applied Linguistics*, **2**(3), 100068. DOI : <https://doi.org/10.1016/j.rmal.2023.100068>.

- CHANG Y., WANG X., WANG J., WU Y., YANG L., ZHU K., CHEN H., YI X., WANG C., WANG Y. *et al.* (2024). A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, **15**(3), 1–45.
- CHANSON H. (2007). Hydraulic engineering in the 21st century : Where to ? *Journal of Hydraulic Research*, **45**(3), 291–301.
- CHEN W.-F. & LIEW J. R. (2002). *The civil engineering handbook*. Crc Press.
- CHIANG C.-H. & LEE H.-Y. (2023). Can large language models be an alternative to human evaluations ? In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Édts., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 15607–15631, Toronto, Canada : Association for Computational Linguistics. DOI : [10.18653/v1/2023.acl-long.870](https://doi.org/10.18653/v1/2023.acl-long.870).
- CONTRIBUTORS O. (2023). Opencompass : A universal evaluation platform for foundation models. <https://github.com/open-compass/opencompass>.
- D’HOFFSCHMIDT M., BELBLIDIA W., HEINRICH Q., BRENDLÉ T. & VIDAL M. (2020). FQuAD : French question answering dataset. In T. COHN, Y. HE & Y. LIU, Édts., *Findings of the Association for Computational Linguistics : EMNLP 2020*, p. 1193–1208, Online : Association for Computational Linguistics. DOI : [10.18653/v1/2020.findings-emnlp.107](https://doi.org/10.18653/v1/2020.findings-emnlp.107).
- DUBOIS Y., LI X., TAORI R., ZHANG T., GULRAJANI I., BA J., GUESTRIN C., LIANG P. & HASHIMOTO T. B. (2023). AlpacaFarm : a simulation framework for methods that learn from human feedback. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS ’23*, Red Hook, NY, USA : Curran Associates Inc.
- FARQUHAR S., KOSSEN J., KUHN L. & GAL Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature*, **630**(8017), 625–630.
- FAYSSE M., FERNANDES P., GUERREIRO N., LOISON A., ALVES D., CORRO C. *et al.* (2024). CroissantLLM : A Truly Bilingual French-English Language Model. working paper or preprint.
- FAYSSE M., FERNANDES P., GUERREIRO N. M., LOISON A., ALVES D. M., CORRO C., BOIZARD N., ALVES J., REI R., MARTINS P. H., CASADEMUNT A. B., YVON F., MARTINS A. F. T., VIAUD G., HUDELLOT C. & COLOMBO P. (2025). Croissantllm : A truly bilingual french-english language model.
- FU J., NG S.-K., JIANG Z. & LIU P. (2024). GPTScore : Evaluate as you desire. In K. DUH, H. GOMEZ & S. BETHARD, Édts., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (Volume 1 : Long Papers)*, p. 6556–6576, Mexico City, Mexico : Association for Computational Linguistics. DOI : [10.18653/v1/2024.naacl-long.365](https://doi.org/10.18653/v1/2024.naacl-long.365).
- GALLEGOS I. O., ROSSI R. A., BARROW J., TANJIM M. M., KIM S., DERNONCOURT F., YU T., ZHANG R. & AHMED N. K. (2024). Bias and fairness in large language models : A survey. *Computational Linguistics*, **50**(3), 1097–1179. DOI : [10.1162/coli_a_00524](https://doi.org/10.1162/coli_a_00524).
- GAO L., TOW J., ABBASI B., BIDERMAN S., BLACK S., DIPOFI A., FOSTER C., GOLDING L., HSU J., LE NOAC’H A., LI H., MCDONELL K., MUENNIGHOFF N., OCIEPA C., PHANG J., REYNOLDS L., SCHOELKOPF H., SKOWRON A., SUTAWIKA L., TANG E., THITE A., WANG B., WANG K. & ZOU A. (2024). A framework for few-shot language model evaluation. DOI : [10.5281/zenodo.12608602](https://doi.org/10.5281/zenodo.12608602).
- GEHRMANN S., ADEWUMI T., AGGARWAL K., AMMANAMANCHI P. S., AREMU A., BOSSELUT A. *et al.* (2021). The GEM benchmark : Natural language generation, its evaluation and metrics. In A. BOSSELUT, E. DURMUS, V. P. GANGAL, S. GEHRMANN, Y. JERNITE, L. PEREZ-BELTRACHINI, S. SHAIKH & W. XU, Édts., *Proceedings of the 1st Workshop on Natural Language Generation*,

- Evaluation, and Metrics (GEM 2021)*, p. 96–120, Online : Association for Computational Linguistics. DOI : [10.18653/v1/2021.gem-1.10](https://doi.org/10.18653/v1/2021.gem-1.10).
- GERALD T., VILNAT A., ETTAYEB S., TAMAMES L. & PAROUBEK P. (2024). Introducing CQuAE : A new French contextualised question-answering corpus for the education domain. In N. CALZOLARI, M.-Y. KAN, V. HOSTE, A. LENCI, S. SAKTI & N. XUE, Édés., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, p. 9234–9244, Torino, Italia : ELRA and ICCL.
- GKATZIA D. & MAHAMOOD S. (2015). A snapshot of NLG evaluation practices 2005 - 2014. In A. BELZ, A. GATT, F. PORTET & M. PURVER, Édés., *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, p. 57–60, Brighton, UK : Association for Computational Linguistics. DOI : [10.18653/v1/W15-4708](https://doi.org/10.18653/v1/W15-4708).
- GRATTAFIORI A., DUBEY A., JAUHRI A., PANDEY A., KADIAN A., AL-DAHLE A. *et al.* (2024). The llama 3 herd of models.
- HADA R., GUMMA V., DE WYNTER A., DIDDEE H., AHMED M., CHOUDHURY M., BALI K. & SITARAM S. (2024). Are large language model-based evaluators the solution to scaling up multilingual evaluation? In Y. GRAHAM & M. PURVER, Édés., *Findings of the Association for Computational Linguistics : EACL 2024*, p. 1051–1070, St. Julian’s, Malta : Association for Computational Linguistics.
- HASHIMOTO T. B., ZHANG H. & LIANG P. (2019). Unifying human and statistical evaluation for natural language generation. In J. BURSTEIN, C. DORAN & T. SOLORIO, Édés., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 1689–1701, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : [10.18653/v1/N19-1169](https://doi.org/10.18653/v1/N19-1169).
- HATTORI E. A., YAMAKAWA M. & MIWA K. (2024). Human bias in evaluating ai product creativity. *Journal of Creativity*, **34**(2), 100087. DOI : <https://doi.org/10.1016/j.yjoc.2024.100087>.
- HENDRYCKS D., BURNS C., BASART S., ZOU A., MAZEIKA M., SONG D. & STEINHARDT J. (2021). Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.
- HENRICH J., HEINE S. J. & NORENZAYAN A. (2010). The weirdest people in the world? *Behavioral and brain sciences*, **33**(2-3), 61–83.
- HOLTZMAN A., BUYS J., DU L., FORBES M. & CHOI Y. (2020). The curious case of neural text degeneration.
- HOLTZMAN A., WEST P. & ZETTLEMOYER L. (2023). Generative models as a complex systems science : How can we make sense of large language model behavior ?
- 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.
- JOSHI N. N. (2025). Evaluating human perception and bias in AI-generated humor. In C. F. HEMPELMANN, J. RAYZ, T. DONG & T. MILLER, Édés., *Proceedings of the 1st Workshop on Computational Humor (CHum)*, p. 79–87, Online : Association for Computational Linguistics.
- KAPLAN J., MCCANDLISH S., HENIGHAN T., BROWN T. B., CHESSE B., CHILD R., GRAY S., RADFORD A., WU J. & AMODEI D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv :2001.08361*.
- KERARON R., LANCRENON G., BRAS M., ALLARY F., MOYSE G., SCIALOM T., SORIANO-MORALES E.-P. & STAIANO J. (2020). Project PIAF : Building a native French question-answering

- dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 5481–5490, Marseille, France : European Language Resources Association.
- KHASHABI D., STANOVSKY G., BRAGG J., LOURIE N., KASAI J., CHOI Y., SMITH N. A. & WELD D. (2022). GENIE : Toward reproducible and standardized human evaluation for text generation. In Y. GOLDBERG, Z. KOZAREVA & Y. ZHANG, Éds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 11444–11458, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics. DOI : [10.18653/v1/2022.emnlp-main.787](https://doi.org/10.18653/v1/2022.emnlp-main.787).
- KOO R., LEE M., RAHEJA V., PARK J. I., KIM Z. M. & KANG D. (2024). Benchmarking cognitive biases in large language models as evaluators. In L.-W. KU, A. MARTINS & V. SRIKUMAR, Éds., *Findings of the Association for Computational Linguistics : ACL 2024*, p. 517–545, Bangkok, Thailand : Association for Computational Linguistics. DOI : [10.18653/v1/2024.findings-acl.29](https://doi.org/10.18653/v1/2024.findings-acl.29).
- KREUTZER J., KHADIVI S., MATUSOV E. & RIEZLER S. (2018). Can neural machine translation be improved with user feedback? In S. BANGALORE, J. CHU-CARROLL & Y. LI, Éds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 3 (Industry Papers)*, p. 92–105, New Orleans - Louisiana : Association for Computational Linguistics. DOI : [10.18653/v1/N18-3012](https://doi.org/10.18653/v1/N18-3012).
- LABRAK Y., BAZOGE A., DUFOUR R., DAILLE B., GOURRAUD P.-A., MORIN E. & ROUVIER M. (2022). FrenchMedMCQA : A French multiple-choice question answering dataset for medical domain. In A. LAVELLI, E. HOLDERNESS, A. JIMENO YEPES, A.-L. MINARD, J. PUSTEJOVSKY & F. RINALDI, Éds., *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, p. 41–46, Abu Dhabi, United Arab Emirates (Hybrid) : Association for Computational Linguistics. DOI : [10.18653/v1/2022.louhi-1.5](https://doi.org/10.18653/v1/2022.louhi-1.5).
- LI H., LAN T., FU Z., CAI D., LIU L., COLLIER N., WATANABE T. & SU Y. (2023). Repetition in repetition out : towards understanding neural text degeneration from the data perspective. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23*, Red Hook, NY, USA : Curran Associates Inc.
- LI X. & ROTH D. (2006). Learning question classifiers : the role of semantic information. *Natural Language Engineering*, **12**(3), 229–249.
- LI Z., WANG C., MA P., WU D., WANG S., GAO C. & LIU Y. (2024a). Split and merge : Aligning position biases in LLM-based evaluators. In Y. AL-ONAIKAN, M. BANSAL & Y.-N. CHEN, Éds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, p. 11084–11108, Miami, Florida, USA : Association for Computational Linguistics. DOI : [10.18653/v1/2024.emnlp-main.621](https://doi.org/10.18653/v1/2024.emnlp-main.621).
- LI Z., XU X., SHEN T., XU C., GU J.-C., LAI Y., TAO C. & MA S. (2024b). Leveraging large language models for NLG evaluation : Advances and challenges. In Y. AL-ONAIKAN, M. BANSAL & Y.-N. CHEN, Éds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, p. 16028–16045, Miami, Florida, USA : Association for Computational Linguistics. DOI : [10.18653/v1/2024.emnlp-main.896](https://doi.org/10.18653/v1/2024.emnlp-main.896).
- LIANG P., BOMMASANI R., LEE T., TSIPRAS D., SOYLU D., YASUNAGA M. *et al.* (2023). Holistic evaluation of language models.
- LIU Y., ITER D., XU Y., WANG S., XU R. & ZHU C. (2023). G-eval : NLG evaluation using gpt-4 with better human alignment. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 2511–2522, Singapore : Association for Computational Linguistics. DOI : [10.18653/v1/2023.emnlp-main.153](https://doi.org/10.18653/v1/2023.emnlp-main.153).

- LOUIS A., VAN DIJCK G. & SPANAKIS G. (2024). Interpretable long-form legal question answering with retrieval-augmented large language models. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, p. tba, Vancouver, Canada : AAAI Press.
- M-A-P TEAM, DU X., YAO Y., MA K., WANG B., ZHENG T., ZHU K., LIU M., LIANG Y., JIN X. *et al.* (2025). SuperGPQA : Scaling llm evaluation across 285 graduate disciplines.
- MAGAR I. & SCHWARTZ R. (2022). Data contamination : From memorization to exploitation. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Édts., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 157–165, Dublin, Ireland : Association for Computational Linguistics. DOI : [10.18653/v1/2022.acl-short.18](https://doi.org/10.18653/v1/2022.acl-short.18).
- NICHOLS A. L. & MANER J. K. (2008). The good-subject effect : Investigating participant demand characteristics. *The Journal of general psychology*, **135**(2), 151–166.
- OUYANG L., WU J., JIANG X., ALMEIDA D., WAINWRIGHT C., MISHKIN P. *et al.* (2022). Training language models to follow instructions with human feedback. In S. KOYEJO, S. MOHAMED, A. AGARWAL, D. BELGRAVE, K. CHO & A. OH, Édts., *Advances in Neural Information Processing Systems*, volume 35, p. 27730–27744 : Curran Associates, Inc.
- PANICKSSERY A., BOWMAN S. R. & FENG S. (2024). Llm evaluators recognize and favor their own generations. In A. GLOBERSON, L. MACKEY, D. BELGRAVE, A. FAN, U. PAQUET, J. TOMCZAK & C. ZHANG, Édts., *Advances in Neural Information Processing Systems*, volume 37, p. 68772–68802 : Curran Associates, Inc.
- PRASUHN A. L. (1987). Fundamentals of hydraulic engineering.
- QWEN, YANG A., YANG B., ZHANG B., HUI B., ZHENG B. *et al.* (2025). Qwen2.5 technical report.
- RAINA V., LIUSIE A. & GALES M. (2024). Is LLM-as-a-judge robust ? investigating universal adversarial attacks on zero-shot LLM assessment. In Y. AL-ONAIZAN, M. BANSAL & Y.-N. CHEN, Édts., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, p. 7499–7517, Miami, Florida, USA : Association for Computational Linguistics. DOI : [10.18653/v1/2024.emnlp-main.427](https://doi.org/10.18653/v1/2024.emnlp-main.427).
- RAWTE V., CHAKRABORTY S., PATHAK A., SARKAR A., TONMOY S. T. I., CHADHA A., SHETH A. & DAS A. (2023). The troubling emergence of hallucination in large language models - an extensive definition, quantification, and prescriptive remediations. In H. BOUAMOR, J. PINO & K. BALI, Édts., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 2541–2573, Singapore : Association for Computational Linguistics. DOI : [10.18653/v1/2023.emnlp-main.155](https://doi.org/10.18653/v1/2023.emnlp-main.155).
- REIN D., HOU B. L., STICKLAND A. C., PETTY J., PANG R. Y., DIRANI J., MICHAEL J. & BOWMAN S. R. (2023). Gpqa : A graduate-level google-proof q&a benchmark.
- RENZE M. (2024). The effect of sampling temperature on problem solving in large language models. In Y. AL-ONAIZAN, M. BANSAL & Y.-N. CHEN, Édts., *Findings of the Association for Computational Linguistics : EMNLP 2024*, p. 7346–7356, Miami, Florida, USA : Association for Computational Linguistics. DOI : [10.18653/v1/2024.findings-emnlp.432](https://doi.org/10.18653/v1/2024.findings-emnlp.432).
- RILEY D. & CHIANG D. (2022). A continuum of generation tasks for investigating length bias and degenerate repetition. In J. BASTINGS, Y. BELINKOV, Y. ELAZAR, D. HUPKES, N. SAPHRA & S. WIEGREFFE, Édts., *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, p. 426–440, Abu Dhabi, United Arab Emirates (Hybrid) : Association for Computational Linguistics. DOI : [10.18653/v1/2022.blackboxnlp-1.36](https://doi.org/10.18653/v1/2022.blackboxnlp-1.36).

- ROBERTS K., KILICOGU H., FISZMAN M. & DEMNER-FUSHMAN D. (2014). Automatically classifying question types for consumer health questions. In *AMIA Annual Symposium Proceedings*, volume 2014, p. 1018.
- SCHOCH S., YANG D. & JI Y. (2020). “this is a problem, don’t you agree?” framing and bias in human evaluation for natural language generation. In S. AGARWAL, O. DUŠEK, S. GEHRMANN, D. GKATZIA, I. KONSTAS, E. VAN MILTENBURG & S. SANTHANAM, Édés., *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, p. 10–16, Online (Dublin, Ireland) : Association for Computational Linguistics.
- SOTTANA A., LIANG B., ZOU K. & YUAN Z. (2023). Evaluation metrics in the era of GPT-4 : Reliably evaluating large language models on sequence to sequence tasks. In H. BOUAMOR, J. PINO & K. BALI, Édés., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 8776–8788, Singapore : Association for Computational Linguistics. DOI : [10.18653/v1/2023.emnlp-main.543](https://doi.org/10.18653/v1/2023.emnlp-main.543).
- SRIVASTAVA A., RASTOGI A., RAO A., SHOEB A. A. M., ABID A., FISCH A., BROWN A. R., SANTORO A., GUPTA A., GARRIGA-ALONSO A., KLUSKA A., LEWKOWYCZ A., AGARWAL A., POWER A. *et al.* (2023). Beyond the imitation game : Quantifying and extrapolating the capabilities of language models.
- STIENNON N., OUYANG L., WU J., ZIEGLER D., LOWE R., VOSS C., RADFORD A., AMODEI D. & CHRISTIANO P. F. (2020). Learning to summarize with human feedback. In H. LAROCHELLE, M. RANZATO, R. HADSELL, M. BALCAN & H. LIN, Édés., *Advances in Neural Information Processing Systems*, volume 33, p. 3008–3021 : Curran Associates, Inc.
- TAYYAR MADABUSHI H. & LEE M. (2016). High accuracy rule-based question classification using question syntax and semantics. In Y. MATSUMOTO & R. PRASAD, Édés., *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics : Technical Papers*, p. 1220–1230, Osaka, Japan : The COLING 2016 Organizing Committee.
- THOMSON C. & REITER E. (2020). A gold standard methodology for evaluating accuracy in data-to-text systems. In B. DAVIS, Y. GRAHAM, J. KELLEHER & Y. SRIPADA, Édés., *Proceedings of the 13th International Conference on Natural Language Generation*, p. 158–168, Dublin, Ireland : Association for Computational Linguistics. DOI : [10.18653/v1/2020.inlg-1.22](https://doi.org/10.18653/v1/2020.inlg-1.22).
- VAN DER LEE C., VERDUIJN B., KRAHMER E. & WUBBEN S. (2018). Evaluating the text quality, human likeness and tailoring component of PASS : A Dutch data-to-text system for soccer. In E. M. BENDER, L. DERCZYNSKI & P. ISABELLE, Édés., *Proceedings of the 27th International Conference on Computational Linguistics*, p. 962–972, Santa Fe, New Mexico, USA : Association for Computational Linguistics.
- WANG Y., MA X., ZHANG G., NI Y., CHANDRA A., GUO S., REN W. *et al.* (2024). Mmlu-pro : A more robust and challenging multi-task language understanding benchmark. In A. GLOBERSON, L. MACKEY, D. BELGRAVE, A. FAN, U. PAQUET, J. TOMCZAK & C. ZHANG, Édés., *Advances in Neural Information Processing Systems*, volume 37, p. 95266–95290 : Curran Associates, Inc.
- WELLECK S., KULIKOV I., ROLLER S., DINAN E., CHO K. & WESTON J. (2019). Neural text generation with unlikelihood training. *arXiv preprint arXiv :1908.04319*.
- ZELLERS R., HOLTZMAN A., CLARK E., QIN L., FARHADI A. & CHOI Y. (2021). TuringAdvice : A generative and dynamic evaluation of language use. In K. TOUTANOVA, A. RUMSHISKY, L. ZETTLEMOYER, D. HAKKANI-TUR, I. BELTAGY, S. BETHARD, R. COTTERELL, T. CHAKRABORTY & Y. ZHOU, Édés., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 4856–4880, Online : Association for Computational Linguistics. DOI : [10.18653/v1/2021.naacl-main.386](https://doi.org/10.18653/v1/2021.naacl-main.386).

ZHANG D. & LEE W. S. (2003). Question classification using support vector machines. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, p. 26–32, New York, NY, USA : Association for Computing Machinery. DOI : [10.1145/860435.860443](https://doi.org/10.1145/860435.860443).

ZHENG L., CHIANG W.-L., SHENG Y., ZHUANG S., WU Z., ZHUANG Y., LIN Z., LI Z., LI D., XING E. P., ZHANG H., GONZALEZ J. E. & STOICA I. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA : Curran Associates Inc.

ZHU K., WANG J., ZHOU J., WANG Z., CHEN H., WANG Y., YANG L., YE W., ZHANG Y., GONG N. & XIE X. (2024). Promptrobust : Towards evaluating the robustness of large language models on adversarial prompts. In *Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis*, LAMPS '24, p. 57–68, New York, NY, USA : Association for Computing Machinery. DOI : [10.1145/3689217.3690621](https://doi.org/10.1145/3689217.3690621).

ZIEGLER D. M., STIENNON N., WU J., BROWN T. B., RADFORD A., AMODEI D., CHRISTIANO P. & IRVING G. (2020). Fine-tuning language models from human preferences.

Annexes

Résultats complémentaires en *zero-shot*, *two-shot*, *five-shot*

Modèle	Hydrau-EN	Hydrau-FR	Civil-EN	Civil-FR
CroissantLLM-Chat-five-shot	10,38%	5,66%	14,29%	8,47%
CroissantLLM-Chat-two-shot	7,55%	1,89%	8,47%	3,70%
CroissantLLM-Chat-zero-shot	4,72%	0,94%	6,36%	4,76%
Llama-3.1-8B-Instruct-five-shot	23,58%	19,81%	22,75%	20,63%
Llama-3.1-8B-Instruct-two-shot	20,75%	23,58%	20,11%	19,58%
Llama-3.1-8B-Instruct-zero-shot	30,19%	26,42%	22,22%	23,81%
Mistral-7B-Instruct-five-shot	25,47%	24,53%	17,99%	16,93%
Mistral-7B-Instruct-two-shot	28,30%	20,75%	16,93%	14,81%
Mistral-7B-Instruct-zero-shot	24,53%	31,13%	15,34%	21,69%
Qwen2.5-1.5B-Instruct-five-shot	23,58%	18,87%	17,99%	14,29%
Qwen2.5-1.5B-Instruct-two-shot	27,36%	18,87%	17,46%	17,46%
Qwen2.5-1.5B-Instruct-zero-shot	19,81%	21,70%	22,22%	21,69%
Qwen2.5-7B-Instruct-five-shot	30,19%	34,91%	29,10%	22,22%
Qwen2.5-7B-Instruct-two-shot	33,02%	28,30%	27,51%	25,40%
Qwen2.5-7B-Instruct-zero-shot	33,96%	37,74%	30,16%	22,75%
Qwen2.5-14B-Instruct-five-shot	31,13%	34,91%	33,33%	27,51%
Qwen2.5-14B-Instruct-two-shot	30,19%	33,02%	37,57%	30,69%
Qwen2.5-14B-Instruct-zero-shot	36,79%	30,19%	37,04%	27,51%

TABLE 2 – Performances des modèles sur les domaines de l’ingénierie hydraulique (Hydrau) et le génie civil (Civil) en anglais (EN) et en français (FR)

Prompts pour la traduction

Prompt utilisé pour la traduction du français vers l’anglais pour l’ingénierie hydraulique :

Votre traduction doit être :

- Techniquement précise avec un usage correct de la terminologie en ingénierie hydraulique
- Fluide, idiomatique et naturelle en français et adaptée à un usage scientifique et académique
- Fidèle au sens, au ton et au style de l’original
- Jamais littérale ou mot à mot — évitez les formulations mécaniques ou maladroit

Vous devez également préserver :

- Toutes les notations mathématiques (LaTeX) : ne modifiez pas les expressions LaTeX
- Toutes les valeurs numériques telles qu’elles apparaissent dans le texte source.
- La structure et l’ordre des options
- La structure JSON globale (tous les autres champs doivent rester inchangés)

Traduisez l’objet JSON ci-dessous de l’anglais vers le français, en suivant les instructions ci-dessus.

Liste et correspondance de traduction des métadonnées

Anglais (langue source)	Français (langue cible)
Hydraulic Engineering	Ingénierie hydraulique
Hydraulics and Hydrology	Hydraulique et hydrologie
Water conservancy and Hydropower	Aménagement hydraulique et hydroélectricité
Civil Engineering	Génie civil
Bridge and Tunnel Engineering	Génie des ponts et tunnels
Geotechnical Engineering	Géotechnique
Structural Engineering	Ingénierie des structures
Urban Infrastructure Engineering	Génie des infrastructures urbaines

TABLE 3 – Correspondances de traduction de domaines, secteurs et sous-secteurs de l’anglais vers le français.

Prompts utilisés pour les expérimentations

Prompt zero-shot en français

Réponds à la question à choix multiple suivante. Il y a une seule réponse correcte. La dernière ligne de ta réponse doit être au format ‘Answer : \$LETTRE’ (sans guillemets), où LETTRE est l’une des options A, B, C, D, E, F, G, H, I ou J.
{ }

Prompt zero-shot en anglais

Answer the following multiple choice question. There is only one correct answer. The last line of your response should be in the format ‘Answer : \$LETTER’ (without quotes), where LETTER is one of A, B, C, D, E, F, G, H, I, or J.
{ }

Prompt five-shot en français

Réponds à la question à choix multiple suivante. Il y a une seule réponse correcte. La dernière ligne de ta réponse doit être au format Answer : \$LETTRE (sans guillemets), où LETTRE est l’une des options A, B, C, D, E, F, G, H, I ou J.

Question : Un télescope réfringent est composé de deux lentilles convergentes séparées par 100 cm. La lentille oculaire a une distance focale de 20 cm. Le grossissement angulaire du télescope est : A) 10 B) 40 C) 6 D) 25 E) 15 F) 50 G) 30 H) 4 I) 5 J) 20

Answer : Analysons cela étape par étape. Dans un télescope réfringent, si les deux lentilles sont convergentes, leurs foyers doivent être situés entre elles, ce qui signifie que la somme de leurs distances focales doit être égale à leur séparation. Puisque la distance focale d’une lentille est de 20

cm, celle de l'autre doit être de 80 cm. Le grossissement est le rapport de ces deux distances focales, soit 4. Answer : H.

Question : Supposons que la pupille de ton œil ait un diamètre de 5 mm et que tu disposes d'un télescope avec une ouverture de 50 cm. Combien de fois le télescope peut-il collecter plus de lumière que ton œil ? A) 1000 fois plus B) 50 fois plus C) 5000 fois plus D) 500 fois plus E) 10000 fois plus F) 20000 fois plus G) 2000 fois plus H) 100 fois plus I) 10 fois plus J) N/A

Answer : Analysons cela étape par étape. La quantité de lumière qu'un télescope peut collecter par rapport à l'œil humain est proportionnelle à l'aire de son ouverture. L'aire d'un cercle est donnée par la formule $A = \pi \left(\frac{D}{2}\right)^2$, où D est le diamètre. Par conséquent, la capacité relative de collecte de lumière se calcule comme suit :

$$\frac{\left(\frac{50 \text{ cm}}{2}\right)^2}{\left(\frac{5 \text{ mm}}{2}\right)^2} = \frac{\left(\frac{50 \text{ cm}}{0.1 \text{ cm}}\right)^2}{\left(\frac{5 \text{ mm}}{0.1 \text{ cm}}\right)^2} = \frac{500^2}{5^2} = 10000.$$

Answer : E.

Question : D'où proviennent la plupart des comètes à courte période et comment le sait-on ? A) La ceinture de Kuiper ; les comètes à courte période tendent à se situer dans le plan du Système solaire, comme la ceinture de Kuiper. B) La ceinture d'astéroïdes ; les comètes à courte période semblent venir de directions aléatoires, indiquant une distribution sphérique des comètes appelée ceinture d'astéroïdes. C) La ceinture d'astéroïdes ; les comètes à courte période tendent à se situer dans le plan du Système solaire, tout comme la ceinture d'astéroïdes. D) Le nuage d'Oort ; les comètes à courte période ont des périodes orbitales similaires à celles des astéroïdes comme Vesta et se trouvent dans le plan du Système solaire, comme le nuage d'Oort. E) Le nuage d'Oort ; les comètes à courte période semblent venir de directions aléatoires, indiquant une distribution sphérique des comètes appelée nuage d'Oort. F) Le nuage d'Oort ; les comètes à courte période tendent à se situer dans le plan du Système solaire, tout comme le nuage d'Oort. G) La ceinture d'astéroïdes ; les comètes à courte période ont des périodes orbitales similaires à celles des astéroïdes comme Vesta et se trouvent dans le plan du Système solaire, comme la ceinture d'astéroïdes. Answer : Analysons cela étape par étape. La plupart des comètes à courte période proviennent de la ceinture de Kuiper. Cela est déduit de l'observation que ces comètes suivent généralement des orbites situées dans le plan du Système solaire, ce qui correspond à la distribution des objets dans la ceinture de Kuiper. Ainsi, l'alignement de ces orbites cométaires avec le plan de l'écliptique indique leur origine dans la ceinture de Kuiper. Answer : A.

Question : Les couleurs visibles dans une bulle de savon résultent de la A) dispersion B) déviation C) réfraction D) réflexion E) interférence F) conversion en une autre fréquence G) polarisation H) absorption I) diffraction J) transmission

Answer : Analysons cela étape par étape. Les motifs colorés observés dans une bulle de savon sont causés par le phénomène d'interférence lumineuse. Cela se produit lorsque les ondes lumineuses se réfléchissent entre les deux surfaces du film savonneux, se combinant de manière constructive ou destructive selon leurs déphasages et l'épaisseur variable du film. Ces interactions entraînent des motifs colorés vibrants dus aux variations d'intensité des différentes longueurs d'onde de la lumière. Answer : E.

Question : Un four à micro-ondes est branché sur une prise de 120 V et consomme un courant de 2 ampères. À quel rythme l'énergie est-elle utilisée par le four à micro-ondes ? A) 240 W B) 120 W C) 10 W D) 480 W E) 360 W F) 200 W G) 30 W H) 150 W I) 60 W J) 300 W

Answer : Analysons cela étape par étape. La puissance, c'est-à-dire le taux d'utilisation de l'énergie dans un circuit électrique, se calcule en multipliant la tension par l'intensité du courant. Pour un four à micro-ondes branché sur une prise de 120 V et consommant un courant de 2 ampères, la puissance consommée se calcule comme suit :

$$\text{Puissance} = \text{Tension} \times \text{Courant} = 120 \text{ V} \times 2 \text{ A} = 240 \text{ W}.$$

Ainsi, le four à micro-ondes utilise de l'énergie à un rythme de 240 watts. Answer : A.

Question :

Answer : Analysons cela étape par étape.

Prompt two-shot en français

Réponds à la question à choix multiple suivante. Il y a une seule réponse correcte. La dernière ligne de ta réponse doit être au format Answer : \$LETTRE (sans guillemets), où LETTRE est l'une des options A, B, C, D, E, F, G, H, I ou J.

Question : Un télescope réfringent est composé de deux lentilles convergentes séparées par 100 cm. La lentille oculaire a une distance focale de 20 cm. Le grossissement angulaire du télescope est : A) 10 B) 40 C) 6 D) 25 E) 15 F) 50 G) 30 H) 4 I) 5 J) 20

Answer : Analysons cela étape par étape. Dans un télescope réfringent, si les deux lentilles sont convergentes, leurs foyers doivent être situés entre elles, ce qui signifie que la somme de leurs distances focales doit être égale à leur séparation. Puisque la distance focale d'une lentille est de 20 cm, celle de l'autre doit être de 80 cm. Le grossissement est le rapport de ces deux distances focales, soit 4. Answer : H.

Question : Supposons que la pupille de ton œil ait un diamètre de 5 mm et que tu disposes d'un télescope avec une ouverture de 50 cm. Combien de fois le télescope peut-il collecter plus de lumière que ton œil ? A) 1000 fois plus B) 50 fois plus C) 5000 fois plus D) 500 fois plus E) 10000 fois plus F) 20000 fois plus G) 2000 fois plus H) 100 fois plus I) 10 fois plus J) N/A

Answer : Analysons cela étape par étape. La quantité de lumière qu'un télescope peut collecter par rapport à l'œil humain est proportionnelle à l'aire de son ouverture. L'aire d'un cercle est donnée par la formule $A = \pi \left(\frac{D}{2}\right)^2$, où D est le diamètre. Par conséquent, la capacité relative de collecte de lumière se calcule comme suit :

$$\frac{\left(\frac{50 \text{ cm}}{2}\right)^2}{\left(\frac{5 \text{ mm}}{2}\right)^2} = \frac{\left(\frac{50 \text{ cm}}{0.1 \text{ cm}}\right)^2}{\left(\frac{5 \text{ mm}}{0.1 \text{ cm}}\right)^2} = \frac{500^2}{5^2} = 10000.$$

Answer : E.

Prompt five-shot en anglais

Answer the following multiple choice question. There is only one correct answer. The last line of your response should be in the format Answer : \$LETTER (without quotes), where LETTER is one of A, B, C, D, E, F, G, H, I, or J.

Question :

A refracting telescope consists of two converging lenses separated by 100 cm. The eye-piece lens has a focal length of 20 cm. The angular magnification of the telescope is :

- A) 10
- B) 40
- C) 6
- D) 25
- E) 15
- F) 50
- G) 30
- H) 4
- I) 5
- J) 20

Answer : Let's think step by step. In a refracting telescope, if both lenses are converging, the focus of both lenses must be between the two lenses, and thus the focal lengths of the two lenses must add up to their separation. Since the focal length of one lens is 20 cm, the focal length of the other must be 80 cm. The magnification is the ratio of these two focal lengths, or 4. Answer : H.

Question : Say the pupil of your eye has a diameter of 5 mm and you have a telescope with an aperture of 50 cm. How much more light can the telescope gather than your eye ? A) 1000 times more B) 50 times more C) 5000 times more D) 500 times more E) 10000 times more F) 20000 times more G) 2000 times more H) 100 times more I) 10 times more J) N/A

Answer : Let's think step by step. The amount of light a telescope can gather compared to the human eye is proportional to the area of its apertures. The area of a circle is given by the formula $A = \pi \left(\frac{D}{2}\right)^2$, where D is the diameter. Therefore, the relative light-gathering power is calculated as :

$$\frac{\left(\frac{50 \text{ cm}}{2}\right)^2}{\left(\frac{5 \text{ mm}}{2}\right)^2} = \frac{\left(\frac{50 \text{ cm}}{0.1 \text{ cm}}\right)^2}{\left(\frac{5 \text{ mm}}{0.1 \text{ cm}}\right)^2} = \frac{500^2}{5^2} = 10000.$$

Answer : E.

Question : Where do most short-period comets come from and how do we know ? A) The Kuiper belt ; short period comets tend to be in the plane of the solar system like the Kuiper belt. B) The asteroid belt ; short period comets tend to come from random directions indicating a spherical distribution of comets called the asteroid belt. C) The asteroid belt ; short period comets tend to be in the plane of the solar system just like the asteroid belt. D) The Oort cloud ; short period comets have orbital periods similar to asteroids like Vesta and are found in the plane of the solar system just like the Oort cloud. E) The Oort Cloud ; short period comets tend to come from random directions indicating a spherical distribution of comets called the Oort Cloud. F) The Oort cloud ; short period comets tend to be in the plane of the solar system just like the Oort cloud. G) The asteroid belt ; short period comets have orbital periods similar to asteroids like Vesta and are found in the plane of the solar system just like the asteroid belt. Answer : Let's think step by step. Most short-period comets originate from the Kuiper belt. This is deduced from the observation that these comets tend to follow orbits that lie in the plane of the solar system, similar to the distribution of objects in the Kuiper belt itself. Thus, the alignment of these cometary orbits with the ecliptic plane points to their Kuiper belt origin. Answer : A.

Question : Colors in a soap bubble result from light A) dispersion B) deflection C) refraction D) reflection E) interference F) converted to a different frequency G) polarization H) absorption I) diffraction J) transmission

Answer : Let's think step by step. The colorful patterns observed in a soap bubble are caused by the phenomenon of light interference. This occurs when light waves bounce between the two surfaces of the soap film, combining constructively or destructively based on their phase differences and the varying thickness of the film. These interactions result in vibrant color patterns due to variations in the intensity of different wavelengths of light. Answer : E.

Question : A microwave oven is connected to an outlet, 120 V, and draws a current of 2 amps. At what rate is energy being used by the microwave oven? A) 240 W B) 120 W C) 10 W D) 480 W E) 360 W F) 200 W G) 30 W H) 150 W I) 60 W J) 300 W

Answer : Let's think step by step. The rate of energy usage, known as power, in an electrical circuit is calculated by the product of voltage and current. For a microwave oven connected to a 120 V outlet and drawing a current of 2 amps, the power consumption can be calculated as follows :

$$\text{Power} = \text{Voltage} \times \text{Current} = 120 \text{ V} \times 2 \text{ A} = 240 \text{ W}.$$

Therefore, the microwave oven uses energy at a rate of 240 watts. Answer : A.

Question :

Answer : Let's think step by step.

Prompt two-shot en anglais

Answer the following multiple choice question. There is only one correct answer. The last line of your response should be in the format Answer : \$LETTER (without quotes), where LETTER is one of A, B, C, D, E, F, G, H, I, or J.

Question :

A refracting telescope consists of two converging lenses separated by 100 cm. The eye-piece lens has a focal length of 20 cm. The angular magnification of the telescope is :

- A) 10
- B) 40
- C) 6
- D) 25
- E) 15
- F) 50
- G) 30
- H) 4
- I) 5
- J) 20

Answer : Let's think step by step. In a refracting telescope, if both lenses are converging, the focus of both lenses must be between the two lenses, and thus the focal lengths of the two lenses must add up to their separation. Since the focal length of one lens is 20 cm, the focal length of the other must be 80 cm. The magnification is the ratio of these two focal lengths, or 4. Answer : H.

Question : Say the pupil of your eye has a diameter of 5 mm and you have a telescope with an aperture of 50 cm. How much more light can the telescope gather than your eye? A) 1000 times more B) 50 times more C) 5000 times more D) 500 times more E) 10000 times more F) 20000 times more G) 2000 times more H) 100 times more I) 10 times more J) N/A

Answer : Let's think step by step. The amount of light a telescope can gather compared to the human eye is proportional to the area of its apertures. The area of a circle is given by the formula $A = \pi \left(\frac{D}{2}\right)^2$, where D is the diameter. Therefore, the relative light-gathering power is calculated as :

$$\frac{\left(\frac{50 \text{ cm}}{2}\right)^2}{\left(\frac{5 \text{ mm}}{2}\right)^2} = \frac{\left(\frac{50 \text{ cm}}{0.1 \text{ cm}}\right)^2}{\left(\frac{5 \text{ mm}}{0.1 \text{ cm}}\right)^2} = \frac{500^2}{5^2} = 10000.$$

Answer : E.
