Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

Stimuler la Pensée Étudiante avec l'AQG : Vers une Génération Automatique de Questions de Type Étudiant

Abdelbassat Labeche[1, 2] Sébastien Fournier[1]
(1) LIS Laboratory, Aix-Marseille Université, France
(2) École Supérieure d'Informatique (ESI-SBA), Algérie
a.labeche@esi-sba.dz, sebastien.fournier@univ-amu.fr

RÉSUMÉ

Les systèmes de génération automatique de questions (AQG) sont largement utilisés dans les contextes éducatifs pour évaluer les connaissances. Ces systèmes se concentrent presque exclusivement sur des questions de type enseignant, structurées et factuelles. Cet article propose une approche novatrice, le Student-AQG, qui vise à simuler des questions spontanées qu'un étudiant réel pourrait poser, reflétant ses incompréhensions, sa curiosité ou ses besoins d'approfondissement. En nous appuyant sur les travaux récents en génération de questions autonomes (Mulla & Gharpure, 2023), nous concevons un système modulaire basé sur des LLMs guidés par du prompt engineering, tenant compte du profil cognitif de l'apprenant. Nous décrivons une stratégie d'évaluation combinant des métriques automatiques et des annotations humaines sur la fluidité, la pertinence et la valeur pédagogique. Ce travail vise à aider les élèves à formuler des questions, développant ainsi leur pensée critique, une compétence essentielle souvent négligée à cause du faible questionnement spontané observé en classe (Chin & Osborne, 2008; Raj *et al.*, 2022).

ABSTRACT

Stimulating Student Thinking with AQG : Towards Automatic Generation of Student-Like Questions

Automatic question generation (AQG) systems are widely used in educational settings to support knowledge assessment. Most existing systems focus almost exclusively on teacher-like questions that are structured and factual. This paper proposes a novel approach, Student-AQG, which aims to simulate spontaneous questions a real student might ask—revealing confusion, curiosity, or the need for clarification. Building on recent research in question generation (Mulla & Gharpure, 2023), we design a modular system that leverages large language models guided by prompt engineering, while adapting to the learner's cognitive profile. We also describe an evaluation framework combining automatic metrics and human annotations, focusing on fluency, relevance, and pedagogical value. This research seeks to assist students in formulating their own questions, enhancing their critical thinking—a key competency often overlooked due to the lack of spontaneous questioning in classrooms (Chin & Osborne, 2008; Raj *et al.*, 2022).

MOTS-CLÉS : Génération automatique de questions, Modèles de langage, Apprentissage actif, Prompt engineering, Pensée critique.

KEYWORDS: Automatic question generation, Language models, Active learning, Prompt engineering, Critical thinking.

ARTICLE : **Accepté à** IA-ÉDU@CORIA-TALN 2025.

Actes de l'atelier Intelligence Artificielle générative et ÉDUcation :
Enjeux, Défis et Perspectives de Recherche 2025 (IA-ÉDU), pages 126–132.

1 Introduction

Ces dernières années, les grands modèles de langage (LLMs) comme ChatGPT, LLaMA ou Mistral ont transformé l'interaction entre intelligence artificielle et apprentissage humain. Pourtant, en contexte scolaire, les étudiants restent souvent passifs et posent rarement des questions en classe, par manque de confiance ou par crainte du jugement. Il devient donc essentiel de les accompagner dans la formulation de questions, afin de les rendre progressivement autonomes dans cette pratique réflexive.

Un usage prometteur des LLMs en éducation est la génération automatique de questions (AQG) à partir d'un texte source. Les travaux de (Mulla & Gharpure, 2023) montrent que la plupart des systèmes AQG se limitent à des questions factuelles, négligeant les aspects métacognitifs. Or, comme l'ont établi (Ikuta & Maruno, 2004), les étudiants qui posent des questions montrent une compréhension plus profonde des concepts. Notre projet s'inscrit dans cette perspective, en proposant un système original centré sur l'élève. Ce **Student-AQG system** vise à simuler, à l'aide d'un LLM, la capacité d'un étudiant réel à poser des questions ouvertes, imparfaites, mais révélatrices de sa curiosité ou de ses doutes.

L'objectif est de stimuler l'engagement cognitif et d'encourager la pensée critique par la formulation de questions pertinentes, adaptées à leur âge et à leur niveau. Contrairement aux approches classiques centrées sur l'évaluation, notre système cherche à reproduire la dynamique naturelle d'un élève face à un contenu éducatif, et à l'accompagner vers une autonomie dans la création de ses propres questions.

2 Problématique

La génération automatique de questions est un domaine actif en traitement automatique des langues (TAL), mais la majorité des approches restent centrées sur l'enseignant (Bulathwela *et al.*, 2023). Si ces méthodes sont efficaces pour créer des QCM ou des quiz, elles n'encouragent pas toujours l'engagement cognitif ni la pensée critique des apprenants.

À l'inverse, un système Student-AQG devrait produire des questions traduisant une curiosité réelle, une compréhension partielle ou un besoin d'approfondissement (Chin & Osborne, 2008). Ces questions ne servent pas seulement à évaluer, mais initient un dialogue avec l'enseignant, stimulant un apprentissage plus interactif et réflexif. Le tableau 1 illustre les différences clés entre les deux approches.

TABLE 1 – Comparaison Teacher-AQG vs Student-AQG

Élément	Teacher-AQG	Student-AQG
Finalité	Évaluer la compréhension	Simuler le questionnement naturel
Style	Structuré, factuel	Curieux, confus, exploratoire
Utilisation	Enseignants, quiz	Élèves, auto-apprentissage
Données	SQuAD, RACE	Forums, prompts simulés

Cependant, guider un LLM à adopter une posture « étudiante » reste un défi majeur. Il faut concevoir des prompts générant des formulations naturelles mais pertinentes (Zamfirescu-Pereira *et al.*, 2023), éviter les questions trop simples ou complexes, évaluer leur valeur pédagogique, et limiter les biais cognitifs ou culturels.

3 État de l'art

Cette section présente brièvement les principales approches de génération automatique de questions, en s'appuyant sur les méthodologies recensées par (Mulla & Gharpure, 2023), des systèmes à base de règles aux modèles neuronaux actuels (Lopez *et al.*, 2021).

3.1 Approches fondées sur des règles et méthodes neuronales supervisées

Les premières méthodes utilisent des structures syntaxiques et des patrons linguistiques pour transformer des phrases en questions. Elles exploitent des arbres syntaxiques, des patrons sémantiques ou des règles heuristiques issues de corpus, mais manquent de souplesse et de généralisation hors domaine.

Avec des corpus comme SQuAD ou WikiAnswers, des architectures Seq2Seq avec attention ont permis de générer automatiquement des questions à partir de textes et réponses. Des modèles comme T5, BART ou des réseaux LSTM ont été entraînés pour produire des questions proches de celles humaines, évaluées par des métriques telles que BLEU, METEOR ou ROUGE.

3.2 Transformers et prompting sans supervision

Des modèles récents tels que GPT-3, LLaMA ou BERT, utilisés en zero-shot ou few-shot via le prompt engineering, génèrent des questions pertinentes sans entraînement dédié. Ces approches s'appuient sur la contextualisation des prompts, la reformulation par rôle ou le raisonnement en chaîne pour simuler un questionnement plus naturel. S'inscrivant dans cette troisième approche, les recherches récentes incluent des travaux sur la génération de questions critiques visant à stimuler la pensée analytique (Figueras & Agerri, 2025), la génération de questions motivées par la curiosité simulant le questionnement d'un apprenant découvrant un nouveau concept (Javaji & Zhu, 2024), et l'utilisation des techniques de prompting pour générer des questions éducatives, notamment avec le jeu de données EduProbe qui privilégie les questions commençant par "Pourquoi" et "Comment" (Maity *et al.*, 2024).

Par rapport aux méthodes précédentes, ces modèles offrent une grande souplesse, mais posent des défis en termes de variabilité, de contrôle du style et d'évaluation de la pertinence des questions.

4 Méthodologie

Notre méthodologie repose sur la conception d'un système capable de simuler le questionnement naturel d'un élève à partir d'un texte pédagogique. Ce système, appelé Student-AQG, mobilise des modèles de langage avancés et des techniques modernes de prompt engineering pour générer des questions pertinentes, révélatrices de la curiosité ou des incompréhensions d'un apprenant.

4.1 Vue d'ensemble du système

Inspiré par le cadre multi-agents proposé dans (Sun *et al.*, 2024), notre système combine plusieurs modules où le LLM alterne entre rôles génératifs et évaluatifs. Le pipeline comprend quatre grandes étapes successives : (1) le prétraitement linguistique, (2) l'extraction de concepts clés, (3) la génération

de questions via un LLM guidé par des prompts optimisés (Lemeš, 2024), et enfin (4) le filtrage et l'évaluation des questions générées.

Le système prend en compte le **profil cognitif de l'élève**, notamment son âge et son niveau scolaire, pour ajuster la formulation, la complexité et l'intention pédagogique de la question. Cette personnalisation vise à renforcer l'engagement et la pertinence des interactions.

4.2 Architecture technique

L'implémentation repose sur une architecture modulaire HuggingFace. Elle comprend un pipeline linguistique, une couche d'abstraction entre modèles, et des connecteurs d'évaluation. Notre étude compare des modèles de tailles variées pour évaluer le compromis entre qualité et ressources.

TABLE 2 – Modèles envisagés pour le Student-AQG

Modèle	Taille	Type
GPT-4	∼1T	Propriétaire
Qwen	14B	Open source
Mistral	7B	Open source
Gemma	2B	Open source

4.3 Ingénierie des prompts

Le guidage du LLM repose sur une ingénierie fine des prompts, adaptée au profil cognitif ciblé. Comme l'ont montré (Zamfirescu-Pereira *et al.*, 2023), la formulation des instructions joue un rôle clé dans la qualité des réponses. Nous nous appuyons sur plusieurs stratégies, dont (Scaria *et al.*, 2024).

Nous utilisons un gabarit dynamique où l'élève est simulé comme locuteur. Par exemple :

```
Joue le rôle d'un(e) étudiant(e) de niveau[niveau] en[domaine].Tu viens
d'apprendre le concept suivant:[concept],mais certains points restent
flous. Instructions : 1. Identifie un aspect difficile ou intéressant
2. Formuleune question à poser à ton enseignant
3. (Avancé uniquement) Expliqueton raisonnement
```

Trois techniques principales enrichissent la génération : a) **L'exemple guidé (Few-shot)** : le modèle reçoit 2 ou 3 exemples authentiques de questions d'élèves ; b) **Le raisonnement en chaîne (Chain-of-Thought)** : selon (Wei *et al.*, 2022), cette méthode aide le modèle à formuler sa pensée séquentiellement, améliorant cohérence et plausibilité, surtout pour les niveaux avancés ; c) **La reformulation par rôle** : le modèle adopte l'identité d'un étudiant curieux face à un nouveau concept, favorisant une formulation plus naturelle et spontanée.

Ces techniques génèrent des questions authentiques. Pour un cours d'IHM, le système produit : « Quelle est la différence entre le système sensoriel et le système cognitif dans l'interface homme machine ? » — imitant le questionnement spontané étudiant.

4.4 Extension : aide à la formulation de questions

En complément de la génération automatique de questions, nous envisageons d'intégrer un module interactif destiné à **aider les étudiants à formuler leurs questions**. Ce module proposera des *indices*, des *pistes de réflexion* ou des *mots-clés extraits du texte* pour guider l'élève dans sa réflexion.

L'objectif est de favoriser une démarche active où l'étudiant participe au processus de questionnement, ce qui renforce l'appropriation des connaissances et développe la pensée critique. Cette extension se positionne comme un outil pédagogique interactif et complémentaire au système Student-AQG.

4.5 Réduction des biais dans la génération de questions

Les grands modèles de langage (LLMs) peuvent reproduire ou amplifier des biais présents dans leurs données d'entraînement, notamment sur le genre, l'origine culturelle ou les stéréotypes implicites (Navigli *et al.*, 2023). Dans un contexte éducatif, ces biais peuvent compromettre la neutralité des questions générées, en introduisant des formulations inadéquates ou non inclusives.

Pour limiter ces effets, plusieurs mécanismes sont prévus. D'abord, les prompts seront formulés de manière explicite et neutre, en évitant les termes marquant une identité spécifique (par exemple, "l'apprenant" au lieu de formulations genrées). Ensuite, un filtrage automatique analysera chaque question à l'aide d'un lexique de sensibilité (issu de ressources sur les biais linguistiques), couplé à une mesure de similarité sémantique via des embeddings (Sentence-BERT ou KeyBERT). Enfin, une évaluation humaine sera menée sur un échantillon représentatif : un groupe d'enseignants et d'étudiants annotera les questions selon une grille de neutralité, représentativité et absence de stéréotypes. Cette validation mesurera l'accord inter-annotateurs et permettra d'ajuster progressivement prompts ou filtres.

Cette approche vise à concilier le réalisme cognitif des questions étudiantes avec des garanties minimales d'équité et d'inclusivité.

4.6 Stratégie d'évaluation

L'évaluation du système Student-AQG repose sur une combinaison d'indicateurs automatiques et d'annotations humaines, en accord avec les recommandations de (Mulla & Gharpure, 2023). Le tableau 3 présente les principales métriques utilisées.

TABLE 3 – Métriques d'évaluation du Student-AQG

Critère	Méthode d'évaluation	Échelle
Fluidité	Score GPT-4 (cohérence grammaticale)	0–1
Pertinence contextuelle	Similarité sémantique (Sentence-BERT)	0–3
Réalisme cognitif	Annotation humaine (marqueurs de doute, confusion)	1–5
Valeur pédagogique	Jugement expert (utilité, discussion induite)	1–5
Similarité avec questions étudiantes	Similarité d'embeddings	0–1
Détection de biais	Checklist qualitative	Binaire

L'expérimentation inclura une validation humaine par annotateurs indépendants, avec mesure du niveau d'accord et améliorations itératives du processus.

4.7 Limites des jeux de données existants

Les systèmes traditionnels de génération de questions éducatives reposent principalement sur des corpus comme SQuAD ou SciQ (Bulathwela *et al.*, 2023). Bien que ces jeux de données produisent des questions factuelles bien structurées, ils ne reflètent ni la spontanéité ni les marqueurs d'incertitude du questionnement étudiant réel. Leurs questions supposent une compréhension totale du texte et visent l'évaluation, plutôt que l'expression de doutes ou de curiosité.

Comme le montre (Bulathwela *et al.*, 2023), même les modèles pré-entraînés sur des textes scientifiques (S2ORC) reproduisent ces limites. Notre approche Student-AQG contourne ces biais en misant sur le prompt engineering plutôt qu'un apprentissage supervisé, générant ainsi des questions plus authentiques et pédagogiquement pertinentes.

5 Considérations éthiques

Notre approche intègre plusieurs préoccupations éthiques essentielles à l'usage des LLMs en contexte éducatif. D'abord, la **protection des données** est assurée : aucune information personnelle ni contenu original n'est conservé. Ensuite, des mécanismes de **filtrage post-génération** sont mis en place pour détecter et limiter les biais culturels ou stéréotypes, comme mentionné précédemment. Enfin, la **transparence** est garantie : les utilisateurs sont informés que les questions proviennent d'une simulation, évitant ainsi toute confusion pédagogique.

6 Conclusion et perspectives

Le projet Student-AQG propose une approche innovante pour encourager l'apprentissage actif, en simulant le questionnement spontané d'élèves via des modèles de langage. Contrairement aux systèmes centrés sur l'enseignant, notre démarche cherche à reproduire la diversité, l'imprécision et la curiosité propres aux véritables questions étudiantes. Les travaux de (Elkins *et al.*, 2023) montrent que les LLMs, guidés par des taxonomies comme celle de Bloom, peuvent générer des questions pertinentes. Notre contribution se distingue par son attention au réalisme cognitif de l'élève simulé.

Les perspectives futures incluent l'implémentation complète du pipeline basé sur HuggingFace pour valider tous les composants. Nous visons aussi la constitution d'un jeu de données annoté de vraies questions étudiantes, inspiré de (Ikuta & Maruno, 2004), afin de mieux calibrer nos outils d'évaluation. Enfin, une évaluation à différents niveaux scolaires permettra de mesurer l'adaptabilité du système aux profils d'apprenants et la pertinence pédagogique des questions générées.

Références

BULATHWELA S., MUSE H. & YILMAZ E. (2023). Scalable educational question generation with pre-trained language models. In N. WANG, G. REBOLLEDO-MENDEZ, N. MATSUDA, O. C. SANTOS & V. DIMITROVA, Éds., *Artificial Intelligence in Education*, p. 327–339, Cham : Springer Nature Switzerland.

CHIN C. & OSBORNE J. (2008). Students' questions : a potential resource for teaching and learning science. *Studies in Science Education*, **44**(1), 1–39. DOI : 10.1080/03057260701828101.

ELKINS S., KOCHMAR E., SERBAN I. & CHEUNG J. C. K. (2023). How useful are educational questions generated by large language models ? In N. WANG, G. REBOLLEDO-MENDEZ, V. DIMITROVA, N. MATSUDA & O. C. SANTOS, Éds., *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, p. 536–542, Cham : Springer Nature Switzerland.

FIGUERAS B. C. & AGERRI R. (2025). Benchmarking critical questions generation : A challenging reasoning task for large language models. https://synthical.com/article/057ecc00-f754-4396-b66b-a22f7393884e.

IKUTA J. & MARUNO S. (2004). Do elementary school students come up with questions during class ? *Graduate School of Human-Environment Studies, Kyushu University*. DOI : 10.15017/3566.

JAVAJI S. R. & ZHU Z. (2024). What would you ask when you first saw $a^2 + b^2 = c^2$? evaluating llm on curiosity-driven questioning.

LEMEŠ S. (2024). Prompt engineering. In *Artificial Intelligence in Industry 4.0 : The Future That Comes True*, p. 159–170. DOI : 10.5644/PI2024.215.08.

LOPEZ L. E., CRUZ D. K., CRUZ J. C. B. & CHENG C. (2021). Simplifying paragraph-level question generation via transformer language models. In D. N. PHAM, T. THEERAMUNKONG, G. GOVERNATORI & F. LIU, Éds., *PRICAI 2021 : Trends in Artificial Intelligence*, p. 323–334, Cham : Springer International Publishing.

MAITY S., DEROY A. & SARKAR S. (2024). Harnessing the power of prompt-based techniques for generating school-level questions using large language models. In *Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '23, p. 30–39, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3632754.3632755.

MULLA N. & GHARPURE P. (2023). Automatic question generation : a review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence*, **12**(1), 1–32. DOI : 10.1007/s13748-023-00295-9.

NAVIGLI R., CONIA S. & ROSS B. (2023). Biases in large language models : Origins, inventory, and discussion. *J. Data and Information Quality*, **15**(2). DOI : 10.1145/3597307.

RAJ T., CHAUHAN P., MEHROTRA R. & SHARMA M. (2022). Importance of critical thinking in the education. *World Journal of English Language*, **12**(3), 126–135. DOI : 10.5430/wjel.v12n3p126.

SCARIA N., DHARANI CHENNA S. & SUBRAMANI D. (2024). Automated educational question generation at different bloom's skill levels using large language models : Strategies and evaluation. In A. M. OLNEY, I.-A. CHOUNTA, Z. LIU, O. C. SANTOS & I. I. BITTENCOURT, Éds., *Artificial Intelligence in Education*, p. 165–179, Cham : Springer Nature Switzerland.

SUN H., LIU Y., WU C., YAN H., TAI C., GAO X., SHANG S. & YAN R. (2024). Harnessing multi-role capabilities of large language models for open-domain question answering. In *Proceedings of the ACM Web Conference 2024*, p. 4372–4382.

WEI J., WANG X., SCHUURMANS D., BOSMA M., ICHTER B., XIA F., CHI E., LE Q. V. & ZHOU D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In S. KOYEJO, S. MOHAMED, A. AGARWAL, D. BELGRAVE, K. CHO & A. OH, Éds., *Advances in Neural Information Processing Systems*, volume 35, p. 24824–24837 : Curran Associates, Inc.

ZAMFIRESCU-PEREIRA J., WONG R. Y., HARTMANN B. & YANG Q. (2023). Why johnny can't prompt : How non-ai experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, p. 1–21. DOI : 10.1145/3544548.3581388.