AI in the Classroom: Examining the Feasibility of AI-Generated Questions in Educational Settings

Omar Zeghouani
psyoz1@nottingham.ac.uk
University of Nottingham
Nottingham, United Kingdom

Zawar Ali
psyza1@nottingham.ac.uk
University of Nottingham
Nottingham, United Kingdom

William Simson van Dijkhuizen
psyws2@nottingham.ac.uk
University of Nottingham
Nottingham, United Kingdom

Jia Wei Hong
psyjh22@nottingham.ac.uk
University of Nottingham
Nottingham, United Kingdom

Jeremie Clos
jeremie.clos@nottingham.ac.uk
University of Nottingham
Nottingham, United Kingdom

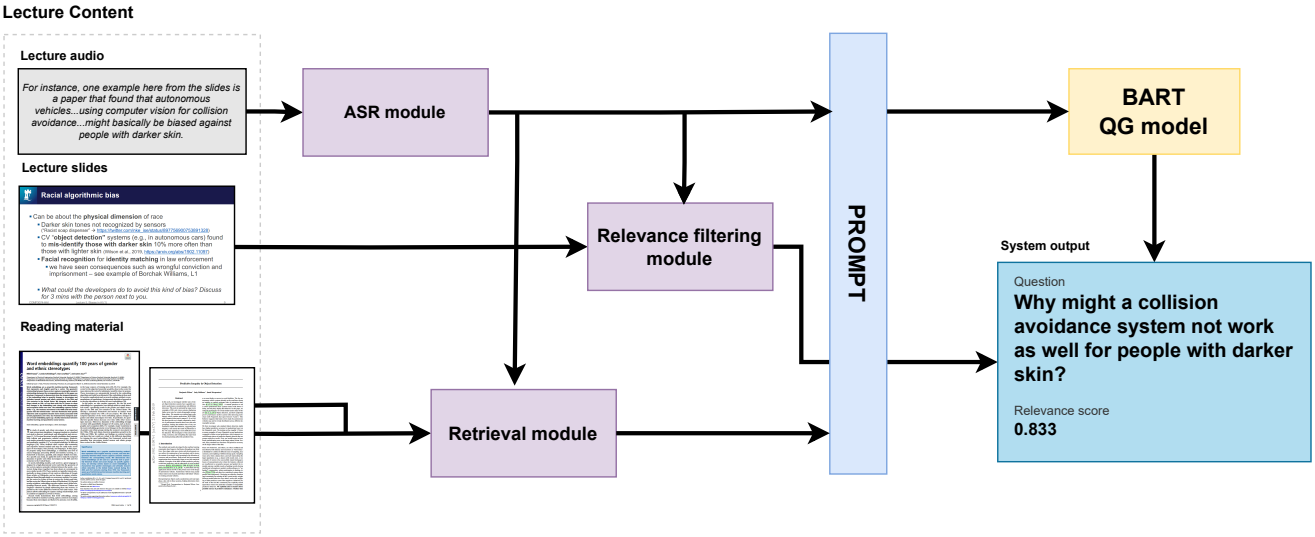

Figure 1: Architecture of our question generation system *ALINet*, which generates questions using lecture content.

Abstract

Educators face ever-growing time constraints, leading to poor work-life balance and a negative impact on work quality. Through their language generation capabilities, large language models offer an interesting avenue to ease this academic workload, allowing both students and lecturers to generate educational content. In this work, we leverage the latest developments in automatic speech recognition, natural language generation, retrieval-augmented generation, and multimodal models to design the Augmented Lecture Integration Network (ALINet), a system capable of producing a diverse range of high-quality assessment questions from lecture content. We inform the design of our system through a series of automated experiments using public datasets and evaluate it with a user study conducted on students and educators. Our results indicate a generally positive perception of the system's performance, particularly in generating natural and clear questions relevant to the taught content, demonstrating its potential as a valuable resource in educational settings. This project lays the foundation for future research in multimodal educational question generation and is available for reuse in our public repository.

CCS Concepts

• **Computing methodologies** → **Natural language processing**; • **Human-centered computing** → **Interactive systems and tools**.

Keywords

Educational Question Generation, Large Language Models, Generative AI

ACM Reference Format:
Omar Zeghouani, Zawar Ali, William Simson van Dijkhuizen, Jia Wei Hong, and Jeremie Clos. 2024. AI in the Classroom: Examining the Feasibility of

TAS '24, September 16–18, 2024, Austin, TX, USA

ACM ISBN 979-8-4007-0989-0/24/09
https://doi.org/10.1145/3686038.3686652

AI-Generated Questions in Educational Settings. In *Second International Symposium on Trustworthy Autonomous Systems (TAS '24), September 16–18, 2024, Austin, TX, USA*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3686038.3686652

1 Introduction

Academics must balance various tasks such as publishing, teaching, securing funding, and supporting their institution [34]. On average, higher education staff work 50 to 55 hours per week [25], leaving them limited time to create additional material such as practice questions and examination papers for students.

There is a mistaken belief amongst students that merely attending and passively absorbing lecture content equates to a genuine grasp of the subject matter. However, when students are confronted with questions that require the application or recall of concepts explained during the lectures, they realise that they do not have a full understanding of the content. Without questions to actively test students, this can lead to a false sense of confidence in a student's ability to perform under examination conditions [26]. Creating their own questions to prepare for examinations poses challenges for students. Misunderstanding lecture content can lead to inaccurate questions; in addition, students tend to focus on material they understand, introducing bias. Hence, study resources should be objective and sourced from taught content [11]. Studies have illustrated that examination of learned content, as opposed to mere recall, stands as a more effective study method, resulting in enhanced information retention and in turn better performance [38]. When surveyed, approximately 83.9% of students attested to the utility of past examination papers and practice questions when it comes to revision strategies [5]. These findings collectively highlight the importance of the availability of such educational materials.

Developing an AI system that generates questions from lecture content will save lecturers time, enhance student learning, and provide students with a more reliable self-assessment tool. In this study, our aim is to answer the following research question: **In the context of higher education, can AI be used to reliably generate assessment questions from lecture content?** To answer this question, we build ALINet, an open source application[1].

2 Related Work

Automatic question generation (AQG) has found applications in various real-world scenarios, including customer service chatbots [10], news diversity analysis [16] and increasingly, educational tutoring systems [2, 15]. While traditional approaches relied on rules and templates to generate questions [23], the advent of deep neural networks has led to a shift towards using transformer-based models [4, 9, 13, 24, 28].

The question generation task typically involves automatically generating questions given a text document as input. Pre-trained language models such as T5 or BART [20, 32] have not been explicitly trained to perform the question generation task; therefore, we can use existing question-answering (QA) datasets by using the context as the source and the question as the target. Datasets such as SQuAD 1.1, NarrativeQA and AdversarialQA [1, 12, 33] are most commonly used for training, with results being automatically evaluated by machine translation metrics such as BLEU, ROGUE or BERTScore [22, 30, 39]. Since our system is designed for educational purposes, QA datasets such as SciQ, RACE and LearningQ [4, 18, 37] will be particularly useful for producing and evaluating questions that test higher-order cognitive skills.

Lectures are typically distributed as videos with supplementary reading material. This multimodality poses a complexity issue for question generation. Generating questions from a video input requires an automatic speech recognition (ASR) system. Similarly to question generation models, ASR has seen a shift from traditional methods such as hidden Markov models [7] to transformer-based models such as Whisper [31]. Once a lecture video has been transcribed, it can be used as input for the question generation model. However, not all sections of a lecture may be pertinent to the learning objectives; therefore, it is necessary to filter out these irrelevant sections. To solve this problem, Wang et al. [36] make use of the original lecture slides to align the questions generated with the learning objectives of the lecture. Additionally, lecturers may briefly reference supplementary reading material during their lectures, which can cause the QG model to "hallucinate", producing incorrect questions. Lewis et al. [21] introduced RAG, which allows language models to retrieve information from an external database, helping to mitigate the problem of "hallucination".

3 Methodology

In this section, we describe the proposed Question Generation system. The architecture of our proposed system (*ALINet*) is shown in Figure 1. As shown in the figure, *ALINet* takes the lecture content as a multi-modal input. The audio is transcribed into chunks using the Distil-Whisper ASR model. Gandhi et al. [8] distilled the Whisper model [31] into a smaller variant that is 6 times faster and supports parallel long-form transcription, making it well suited for processing lecture videos. Following a similar approach to Wang et al. [36], we perform relevance filtering by calculating a similarity score between each segment of the transcript and its corresponding lecture slides, allowing us to omit questions below a specified threshold. As mentioned in Section 2, lecturers may reference supplementary reading material during their lectures. The missing context can result in the QG model hallucinating, producing incorrect questions. Our RAG setup supplements the source text with relevant information from the lecture's reading material to aid the question generation task.

QA datasets often lack an educational focus. To solve this, we created our own training dataset composed of SQuAD, AdverserialQA, NarrativeQA and SciQ [1, 12, 33, 37]. Nielsen et al. [29] proposed a question taxonomy for the purposes of educational assessment by aligning with popular educational frameworks such as Bloom's Taxonomy of Educational Objectives [3]. The taxonomy categorises questions into description, recall, method, and explanation. Our training dataset was balanced to ensure an equal distribution of questions. We also observed that 24.6% of the questions in the training dataset were "ambiguous". A question was considered ambiguous if it contained pronouns or demonstratives and did not contain a proper noun. For example:

[1] publicly available at https://github.com/ram02z/alinet

- Why did the researchers do ***that***?
- How did ***he*** improve the research?

We carried out co-reference resolution using GPT-4 to fix the ambiguous questions. Kulshreshtha et al. [14] showed that back-training achieves lower test error than self-training for question generation. To improve robustness to ASR errors, we introduced ASR noise to our training dataset as synthetic data for back-training.

To assess the performance of our system, we conducted an automatic evaluation and a human evaluation. The automatic evaluation provides an estimate to the quality and effectiveness of the generated questions. We use BERTScore as our evaluation metric as Zhang et al. [39] showed that it correlated highly with human judgment compared to other metrics like BLEU or ROGUE [22, 30]. We evaluated on the MRQA 2019 and Spoken-SQuAD [6, 19] testing sets to assess our model's reading comprehension and robustness to ASR noise. We fine-tuned BART-base [20] for the question generation task on SquAD 1.1 [33] and our training dataset. The results are shown in Table 1.

Model	F1 Score		Precision		Recall	
	MRQA	S-SQuAD	MRQA	S-SQuAD	MRQA	S-SQuAD
Baseline	0.681	0.603	0.691	0.595	0.675	0.615
ALINet	0.654	0.628	0.649	0.627	0.662	0.632

Table 1: The mean evaluation scores on the MRQA 2019 testing set and Spoken-SQuAD WER54 testing set.

However, it is important to note that while automatic evaluation is useful, it may not capture all aspects of question quality, such as relevance or naturalness. In line with the methodology of Nguyen et al. [28], we recruited university lecturers (N=9), who are experts in their relevant fields, to assess the quality of the generated questions. Furthermore, we also recruited university students (N=10) to study the difference in perspective when it comes to question quality. We built upon the criteria defined by Laban et al. [17] to assess the quality of educational assessment questions. Our resulting evaluation criteria encompass the following aspects:

- **Naturalness**: Evaluating whether the structure of the questions is fluent.
- **Answerability**: Evaluating whether the question is answerable given the context.
- **Unambiguity**: Evaluating whether the question leaves no room for interpretations or misunderstandings.
- **Relevance**: Assessing the appropriateness of the questions given the original lecture content.

We designed a feedback form to evaluate the effectiveness of each generated question separately. We set the relevance filtering threshold to 0.5 to omit irrelevant questions. For every question, we offer video context and our evaluation criteria on the Likert scale. The experts assessed questions from a module they teach, whereas, the students assessed questions from a module that all had in common. We obtained full ethics approval from our university before starting our human evaluation and followed the research ethics procedure throughout.

4 Discussion

To answer our research question, we start by analysing the results of the automatic evaluation shown in Table 1. The baseline model, which was trained on the SQuAD 1.1 [33] dataset, outperforms the ALINet model when evaluated on the MRQA 2019 [6] dataset.

As shown in Figure 2, the MRQA 2019 and SQuAD 1.1 datasets have a similar question distribution, with over 97% of the questions classified as description or recall, whilst the ALINet dataset has an equal distribution. As a consequence, the ALINet model has a higher chance of generating a question that is semantically different, due to the balanced distribution of the training data. Therefore, the discrepancy in the question distribution can explain the poorer F1 score of the ALINet model on the MRQA 2019 testing set. However, the ALINet model significantly outperforms the baseline on the Spoken-SQuAD [19] test set, demonstrating its robustness to spoken noise.

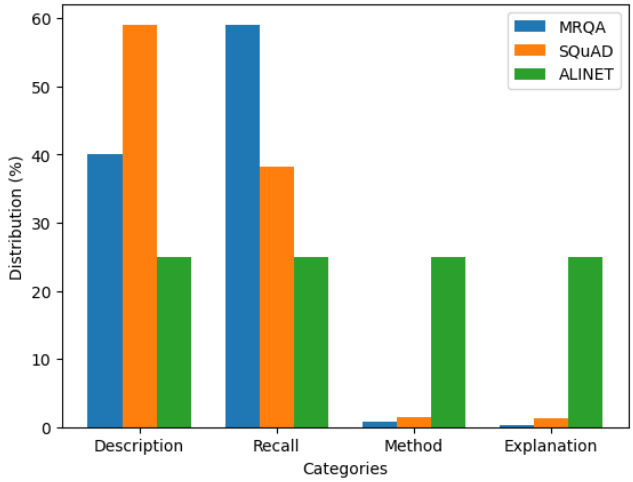

Figure 2: Question distribution between the SQuAD 1.1, MRQA 2019 and ALINet datasets

To assess the system's ability to generate assessment questions reliably and answer our research question from a user's standpoint, we gauge the perspectives of key stakeholders of the system through the results of the human evaluation. For both experts and students, we group the data by the criteria established in Section 3. By doing so, we gain a better understanding of the strengths and weaknesses of our system as evaluated by the primary stakeholders.

Tables 2 and 3 show that lecturers and students generally perceive the generated questions positively, as illustrated by the aggregated mean scores of 3.53 and 3.78, respectively. The Fleiss kappa scores among students are also impressive, ranging from 0.48 to 0.60 across evaluation criteria, illustrating a high level of agreement amongst the raters. The high Fleiss kappa scores among students reinforce the reliability of interpretations discussed later in this section.

Naturalness was regarded as the best property of the generated questions by both professors and students. Figures 3 and 4 show that students overwhelmingly approved of their fluency, with a

Criteria	Mean	Mode	Median	Std deviation
Naturalness	3.91	4.0	4.0	1.00
Answerability	3.27	4.0	4.0	1.30
Unambiguity	3.46	5.0	4.0	1.31
Relevance	3.49	4.0	4.0	1.19
All	3.53	4.0	4.0	1.23

Table 2: Expert responses

Criteria	Mean	Mode	Median	Std Deviation	Fleiss' Kappa
Naturalness	4.61	5.00	5.00	0.69	0.48
Answerability	3.00	N/A	3.00	1.51	0.59
Unambiguity	3.98	5.00	4.00	1.19	0.52
Relevance	3.54	5.00	4.00	1.38	0.60
All	3.78	5.00	4.00	1.37	0.59

Table 3: Student responses

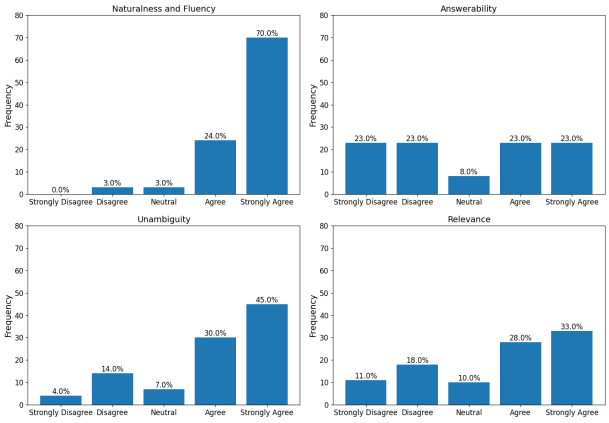

Figure 3: Distribution of student responses aggregated per criteria

dominant score of 70%, "strongly agree", however, educators appeared more reserved, leaning towards an "agree" rating. This trend persisted across the different evaluation criteria, suggesting that experts might adopt a more cautious approach to fully endorsing a question. This hesitation could stem from their role as educators, prompting them to critically assess each aspect of the question.

Unambiguity emerged as the second best property of the generated questions, as indicated by the educators who gave it the highest mode rating of 5 as seen in Table 2. This is reinforced by the data presented in Table 3 where students gave a notably high average rating of 3.98, with a corresponding mode rating of 5. Initially introduced to evaluate the effectiveness of our coreference resolution process, unambiguity addresses the clarity of questions generated by the system. Recognising question clarity as another strong attribute of our system's output underscores the effectiveness of coreference resolution on our training dataset.

The expert assessment of relevance, depicted in Figure 4, shows a distribution that is strongly centered on "agree" (4), as supported

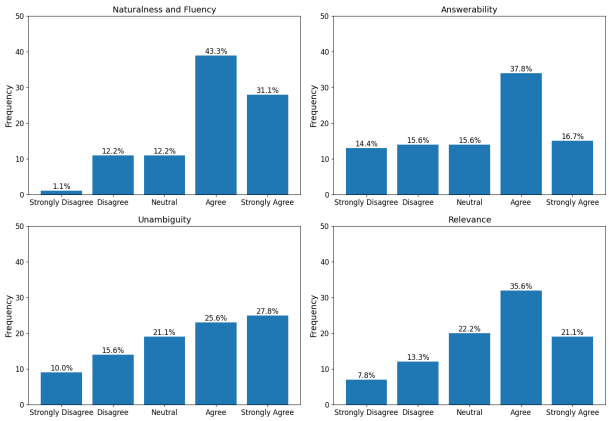

Figure 4: Distribution of expert responses aggregated Per criteria

by metrics in Table 2. To streamline the human evaluation process, we filtered the questions with a threshold of 0.5. In summary, the experts' responses indicate that this relevance filtering approach is successful. However, in the student evaluation, the relevance scores show a moderate mean of 3.54 and a large standard deviation of 1.38, indicating room for improvement. Currently, our assumption that slides serve as adequate source material may lead to questions being deemed relevant solely based on slide content alignment; however, in the case that the slides themselves are not relevant to the learning objectives, this approach fails. To address this, the system could incorporate the learning objectives from the lecture as additional input to verify the appropriateness of the question alongside existing relevance filtering methods.

The most notable weakness of our system lies in its ability to generate answerable questions, a limitation underscored by comparatively low evaluations. Experts rated this aspect with a mean of 3.27, while students assigned a lower score of 3.00, see Tables 2 and 3. Although both groups ranked question answerability as the weakest aspect, the degree of dissatisfaction varied significantly. Experts' assessments revealed a median of 4 for answerability, with a majority indicating agreement (4), suggesting a moderate level of satisfaction. In contrast, students' assessments had a median of 3 and a high Fleiss' kappa score of 0.59 indicating a high level of agreement in the dissatisfaction. Figure 3 shows that 46% of student responses expressed disagreement or strong disagreement, making it the only survey statement with a significant negative sentiment. This stark contrast between students and educators may be due to personal bias. The educators having created the content may subconsciously overlook gaps in question clarity due to their familiarity with the material, an occurrence known as the expert blind spot phenomenon, commonly observed in academia [27].

We can gain insight as to why the system struggles to generate answerable questions by examining the disparity between the expected and actual distribution of question types shown in Figure 5. The system seems to demonstrate a strong preference for method and recall questions, despite this not always being the most suitable choice. For instance, questions often arise from lecture segments in which certain topics are briefly mentioned as part of a broader

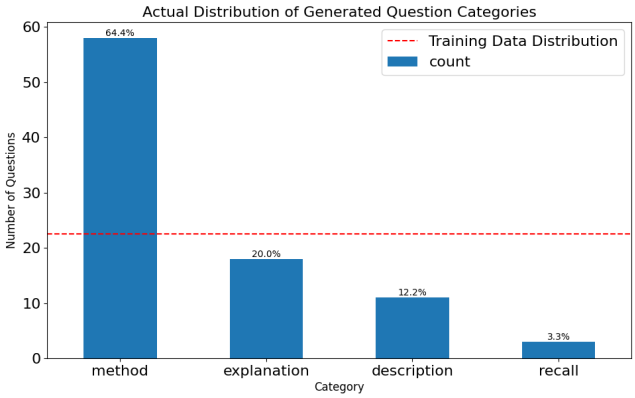

Figure 5: Distribution of questions in human evaluation forms

point. However, these topics are not the primary focus of the discussion and are only mentioned in passing. Without a mechanism to specifically target the portion of the context best aligned with the lecture's learning objectives, this leads to the model sometimes generating questions on non-pertinent topics.

5 Limitations

We have identified some key limitations ranging from the methods used to build ALINet, to contemporary issues within the broader field of question generation. Firstly, we were unable to use state-of-the-art language models in our research, which would have offered better insight into the current capabilities of AI in educational settings. While human evaluation gives us insight of our system's real-world performance, it lacks scalability. Evaluating thousands of questions incurs significant costs and time in recruiting qualified annotators. Additionally, our methods for human evaluation primarily involved recruiting academics and students from the field of Computer Science. To enhance the representativeness and generalisability of the system's performance, it would have been preferable to recruit a larger and more diverse group of participants with varied educational backgrounds.

6 Responsible Research and Innovation

To ensure that our research aligns with the principles of responsible research and innovation (RRI), we have applied the Anticipate, Reflect, Engage, Act (AREA) framework [35] to our research and development process. The purpose of ALINet is to reduce the workload of academics and improve their efficiency when creating assessment material. However, this automation may impact the academic job market, particularly affecting early-career academics and support staff. Another issue is that potential biases in the system's generation of question styles may lead to a lack of question diversity, negatively impacting students by hindering their ability to engage critically with course material and develop essential analytical skills. Lecturers who rely on the system without considering its limitations may inadvertently contribute to this issue. Lastly, inequitable access to computational resources may result in an unfair advantage for institutions and academics with greater resources. Reflecting on potential issues, we must consider whether the benefits of increased efficiency outweigh potential drawbacks such as job displacement and concerns about educational quality, and take into account how technological disparities might affect career advancement opportunities and student learning outcomes. To address these concerns, we will engage with various stakeholders. This includes involving academic staff at different career stages to discuss job security and changing roles, consulting educational technology experts to explore equitable access solutions, and gathering student feedback to understand the impact on learning experiences. Finally, we will create guidelines for the effective use of AI-generated questions, emphasising the importance of human oversight.

7 Conclusion

This research examined the feasibility of using a multi-modal QG system to automatically generate educational questions from lecture content. Through our human evaluation with university lecturers and students, we gained insight into the strengths and limitations of the system. The generated questions excelled in aspects like naturalness and clarity, but there is room for improvement in generating answerable questions that effectively assess higher-order cognitive skills. In general, our research demonstrates the effectiveness of a system like ALINet to support academics. By addressing the outlined limitations, the system's potential as a valuable tool for academics could be fully realised, alleviating their workload while promoting effective assessment and learning in higher education.

Acknowledgments

The authors are supported by the Engineering and Physical Sciences Research Council [grant number EP/V00784X/1]. Large language models were used to edit the grammar of this work, but not for content generation. No participants gave explicit consent to the sharing of their data. Therefore, we can only present the data in aggregated form which is presented throughout the paper.

References

[1] Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the AI: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics* 8 (2020), 662–678.

[2] Ayan Kumar Bhowmick, Ashish Jagmohan, Aditya Vempaty, Prasenjit Dey, Leigh Hall, Jeremy Hartman, Ravi Kokku, and Hema Maheshwari. 2023. Automating question generation from educational text. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence*. Springer, 437–450.

[3] Benjamin S Bloom, Max D Engelhart, EJ Furst, Walker H Hill, and David R Krathwohl. 1956. Handbook I: cognitive domain. *New York: David McKay* (1956), 483–498.

[4] Guanliang Chen, Jie Yang, Claudia Hauff, and Geert-Jan Houben. 2018. LearningQ: a large-scale dataset for educational question generation. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12.

[5] Simon Cross, Denise Whitelock, and Jenna Mittelmeier. 2016. Does the quality and quantity of exam revision impact on student satisfaction and performance in the exam itself?: Perspectives from undergraduate distance learners. (2016).

[6] Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. MRQA 2019 Shared Task: Evaluating Generalization in Reading Comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen (Eds.). Association for Computational Linguistics, Hong Kong, China, 1–13. https://doi.org/10.18653/v1/D19-5801

[7] Mark Gales, Steve Young, et al. 2008. The application of hidden Markov models in speech recognition. *Foundations and Trends® in Signal Processing* 1, 3 (2008), 195–304.

[8] Sanchit Gandhi, Patrick von Platen, and Alexander M Rush. 2023. Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling. *arXiv preprint arXiv:2311.00430* (2023).

[9] R. Goyal, P. Kumar, and V. P. Singh. 2023. Automated Question and Answer Generation from Texts using Text-to-Text Transformers. *Arab Journal of Science and Engineering* (2023). https://doi.org/10.1007/s13369-023-07840-7

[10] Jing Gu, Mostafa Mirshekari, Zhou Yu, and Aaron Sisto. 2021. Chaincqg: Flow-aware conversational question generation. *arXiv preprint arXiv:2102.02864* (2021).

[11] Samuel C Karpen. 2018. The Social Psychology of Biased Self-Assessment. *American Journal of Pharmaceutical Education* 82, 5 (2018), 6299.

[12] Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics* 6 (2018), 317–328.

[13] Kettip Kriangchaivech and Artit Wangperawong. 2019. Question generation by transformers. *arXiv preprint arXiv:1909.05017* (2019).

[14] Devang Kulshreshtha, Robert Belfer, Iulian Serban, and Siva Reddy. 2021. Back-Training excels Self-Training at Unsupervised Domain Adaptation of Question Generation and Passage Retrieval. In *Conference on Empirical Methods in Natural Language Processing*. https://api.semanticscholar.org/CorpusID:233295982

[15] Devang Kulshreshtha, Muhammad Shayan, Robert Belfer, Siva Reddy, Iulian Vlad Serban, and Ekaterina Kochmar. 2022. Few-shot question generation for personalized feedback in intelligent tutoring systems. In *PAIS 2022*. IOS Press, 17–30.

[16] Philippe Laban, Chien-Sheng Wu, Lidiya Murakhovs' ka, Xiang'Anthony' Chen, and Caiming Xiong. 2022. Discord questions: A computational approach to diversity analysis in news coverage. *arXiv preprint arXiv:2211.05007* (2022).

[17] Philippe Laban, Chien-Sheng Wu, Lidiya Murakhovs'ka, Wenhao Liu, and Caiming Xiong. 2022. Quiz Design Task: Helping Teachers Create Quizzes with Automated Question Generation. In *NAACL-HLT*. https://api.semanticscholar.org/CorpusID:248512983

[18] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding Comprehension Dataset From Examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 785–794. https://doi.org/10.18653/v1/D17-1082

[19] Chia-Hsuan Lee, Szu-Lin Wu, Chi-Liang Liu, and Hung yi Lee. 2018. Spoken SQuAD: A Study of Mitigating the Impact of Speech Recognition Errors on Listening Comprehension. In *Proc. Interspeech 2018*. 3459–3463. https://doi.org/10.21437/Interspeech.2018-1714

[20] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703

[21] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[22] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[23] David Lindberg, Fred Popowich, John Nesbit, and Phil Winne. 2013. Generating natural language questions to support learning on-line. In *Proceedings of the 14th European workshop on natural language generation*. 105–114.

[24] Luis Enrico Lopez, Diane Kathryn Cruz, Jan Christian Blaise Cruz, and Charibeth Cheng. 2020. Transformer-based end-to-end question generation. *arXiv preprint arXiv:2005.01107* 4 (2020).

[25] C. L. Cooper M. Y. Tytherleigh *, C. Webb and C. Ricketts. 2005. Occupational stress in UK higher education institutions: a comparative study of all staff categories. *Higher Education Research & Development* 24, 1 (2005), 41–61.

[26] Christopher A McKay, Juan Razo, and Adam M Persky. 2019. The Self-Assessment of Pharmacy Students: A Mixed-Methods Study. *American Journal of Pharmaceutical Education* 83, 9 (2019), 7323.

[27] Mitchell J. Nathan and Anthony Petrosino. 2003. Expert Blind Spot Among Preservice Teachers. *American Educational Research Journal* 40, 4 (2003), 905–928. https://doi.org/10.3102/00028312040004905

[28] Huy A. Nguyen, Shravya Bhat, Steven Moore, Norman Bier, and John Stamper. 2022. Towards Generalized Methods for Automatic Question Generation in Educational Domains. In *Educating for a New Future: Making Sense of Technology-Enhanced Learning Adoption*, Isabel Hilliger, Pedro J. Muñoz-Merino, Tinne De Laet, Alejandro Ortega-Arranz, and Tracie Farrell (Eds.). Springer International Publishing, Cham, 272–284.

[29] Rodney D Nielsen, Jason Buckingham, Gary Knoll, Ben Marsh, and Leysia Palen. 2008. A taxonomy of questions for question generation. In *Proceedings of the workshop on the question generation shared task and evaluation challenge*.

[30] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.

[31] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*. PMLR, 28492–28518.

[32] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67. http://jmlr.org/papers/v21/20-074.html

[33] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Jian Su, Kevin Duh, and Xavier Carreras (Eds.). Association for Computational Linguistics, Austin, Texas, 2383–2392. https://doi.org/10.18653/v1/D16-1264

[34] Kathleen Smithers, Nerida Spina, Jess Harris, and Sarah Gurr. 2023. Working every weekend: The paradox of time for insecurely employed academics. *Time & Society* 32, 1 (01 Feb 2023), 101–122.

[35] UKRI. 2023. Framework for responsible research and innovation. https://www.ukri.org/who-we-are/epsrc/our-policies-and-standards/framework-for-responsible-innovation/

[36] Hei-Chia Wang, Martinus Maslim, and Chia-Hao Kan. 2023. A question–answer generation system for an asynchronous distance learning platform. *Education and Information Technologies* (2023), 1–30.

[37] Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing Multiple Choice Science Questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, Leon Derczynski, Wei Xu, Alan Ritter, and Tim Baldwin (Eds.). Association for Computational Linguistics, Copenhagen, Denmark, 94–106. https://doi.org/10.18653/v1/W17-4413

[38] Brenda W Yang, Juan Razo, and Adam M Persky. 2019. Using Testing as a Learning Tool. *American Journal of Pharmaceutical Education* 83, 9 (2019), 7324.

[39] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. https://openreview.net/forum?id=SkeHuCVFDr