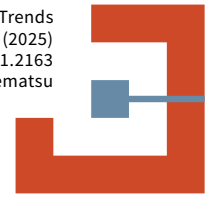
Trends

JALTCALL Trends
ISSN 2982-3749, 1, 2163 (2025)
https://doi.org/10.29140/jct.v1n1.2163



Castledown

GPT-based simulation of oral Q&A to support students attending first conference

Mayuko Aiba
The University of Tokyo
aiba@gavo.t.u-tokyo.ac.jp

Daisuke Saito
The University of Tokyo
dsk_saito@gavo.t.u-tokyo.ac.jp

Nobuaki Minematsu
The University of Tokyo
mine@gavo.t.u-tokyo.ac.jp

Students often need good practice not only for English oral presentation but also for participating in oral Q&A sessions, particularly when presenting their research for the first time at international conferences. This study aims to support those students in improving their oral skills in Q&A by having ChatGPT read the students' papers and ask questions about the papers. We explored six different configurations for question generation, focusing on two key factors: (1) whether to explicitly upload reference papers to ChatGPT, and (2) the scope of question generation – whether questions should be generated by focusing on individual keywords, on individual sections, or on the paper as a whole. Eight master's students, who had recently attended their first conferences, and eight Japanese teachers of English evaluated the questions generated for the students' papers. Surprisingly, both groups favored the questions generated by the simplest configuration, where reference papers were not explicitly uploaded, and no scope was specified. Based on these findings, ChatGPT in the simplest configuration was integrated with a speech interface, enabling oral interaction between ChatGPT and students. After using this system, all participants expressed a strong preference for utilizing it in preparation for future conferences.

Keywords: ChatGPT, oral communication skills, subjective assessment, question generation, English for Academic Purposes (EAP)

Introduction

In today's global academic environment, proficiency in English is essential, especially for those intending to participate in international conferences. They must be able to read and write academic papers, deliver presentations, and handle Q&A sessions in English. Among these tasks, preparing for Q&A sessions is particularly challenging for non-native speakers.

Typically, while students can prepare their presentation slides independently, effective Q&A practice requires interlocutors who thoroughly understand the research content. Ideally, mentors would fulfill this role, but their time is often limited. Although English teachers can help with general communication skills, they might find it difficult to ask field-specific technical questions relevant to the student's presentation.

Recent studies have explored the use of Large Language Models (LLMs) for generating educational content, including academic questions and their answers. For example, Wang et al. (2022) proposed generating high-quality questions by providing a topic and explanatory text to LLMs, which is used practically in classroom settings.

In addition, growing attention has been paid to the role of prompt engineering in optimizing interactions with LLM-based chatbots. Although prompts do not follow a strict syntax like programming languages, slight changes in phrasing can significantly affect the output quality. Studies have introduced effective prompting strategies, including meta-language framing (e.g., "When I say X, I mean Y") and persona-based prompts (e.g., "Act as persona X"), demonstrating their potential to improve the relevance and accuracy of generated content (White et al., 2023).

Accordingly, this study addresses the following research question: How can generative AI be effectively used to support non-native English speakers in preparing for oral Q&A sessions at academic conferences?

To explore this question, we developed a system using ChatGPT (OpenAI, n.d.) that simulates realistic Q&A scenarios without requiring the constant presence of a mentor. In the early stages of development, initial experiments on prompt engineering and question generation were conducted through text-based interactions. The system also incorporates speech technology to facilitate oral practice, which is crucial for building confidence in live conference settings. While existing commercial tools combine LLMs with speech interfaces for general English conversation practice (Speak, n.d.), our project uniquely adapted this technology to practice for Q&A sessions that take place immediately after presentations at international conferences.

Method

Experimental setup

In this study, we used a customized ChatGPT, utilizing "GPTs" feature of ChatGPT Plus, to create a chatbot designed specifically for practicing Q&A in academic presentations. The prompt given to GPTs is shown in Table 1.

Table 1. Prompts given to GPTs to generate questions on the uploaded paper

Role and Goal	Thesis Explorer is tailored for analyzing engineering theses, formulating diverse questions on algorithm effectiveness, data analysis, experimental design, and results relevance. It handles technical and conceptual queries adeptly.
Constraints	The GPT will generate a broad spectrum of questions, covering both intricate technicalities and wider conceptual aspects, without limitations on the complexity or nature of inquiries.
Guidelines	Questions are to be prefaced with relevant context, providing a foundation for each inquiry. The GPT can also offer insights or suggestions related to the question. This approach aids in deepening the user's understanding of the thesis content.
Clarification	If the thesis content is unclear or requires additional details for effective questioning, the GPT will request clarification.
Personalization	Maintaining a scholarly and academic tone, Thesis Explorer adapts its questioning style to the user's responses and preferences, ensuring a tailored and engaging dialogue.

We conducted our study using eight papers, one from each of eight master's students in engineering, who had recently attended their first international conferences. We generated questions on their papers to see how well our system could replicate real Q&A sessions.

We tested six configurations for generating questions. The six configurations consist of the following elements:

1. Whether reference papers are explicitly uploaded to ChatGPT in advance as knowledge .
2. Whether questions are generated by focusing on individual keywords, individual sections, or the entire paper.

The six configurations are labeled as A-F, shown in Table 2.

Table 2. Six configurations for question generation

References	Section	Keyword	Not specifying
Uploaded	A	B	C
Not uploaded	D	E	F

Evaluation procedure

Two groups of participants, the eight students and eight Japanese teachers of English, were involved in evaluating the system. The students rated the questions generated from their own papers on a scale from 1 to 5 based on five criteria shown in Table 3.

Table 3. Evaluation criteria for the students

Criteria	Description
Relevance	Is the question relevant enough to the topic of your paper?
Clarity	Is the question clear enough to understand?
Specificity	Is the question specific enough to answer? If the question is too general or too vague, you may have trouble in answering the question.
Inspiration	Did the question inspire you? Did the question open your eyes?
Predictability	You may have made presentation rehearsals in advance. Is the question expected enough?

The students also checked whether the answers to these questions were present in their papers. If the answers are absent, the questions will be good and insightful ones to the students. Each student evaluated approximately 60 questions.

The teachers evaluated questions generated for all eight papers, focusing on how *likely* the teachers were to ask such questions in their own presentation class. Each teacher evaluated 471 questions in total. Given that English teachers may not always have in-depth knowledge of engineering, their own questions tend to be more general. In contrast, the nature of questions posed by ChatGPT might reflect a higher level of technical specificity. This contrast motivated us to investigate how AI-generated questions might differ from those raised by human teachers in their class for academic presentation.

After reviewing the questions, the students and teachers were asked to allocate 100 points across the configurations based on their preference. While individual question scores reflect the quality of each question, this allocation captures holistic preferences at the configuration level, offering a complementary view of the participants' judgments. During the evaluation, the participants were informed about which configuration (A to F) was used for each question, but they were not informed of what each configuration represented.

System development

After the evaluation, as shown in Figure 1, the best-performing configuration of ChatGPT was integrated with Automatic Speech Recognition (ASR) and Text-to-Speech (TTS), enabling spoken interactions between ChatGPT and the student. ChatGPT's questions were delivered orally, and students' spoken answers are transcribed and sent to ChatGPT. Because ChatGPT can handle papers from any domain, our system serves as a fully domain-independent Q&A practice tool.

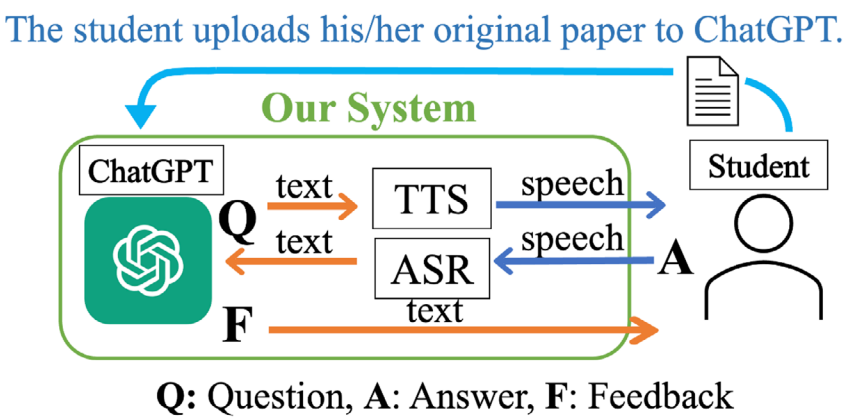


Figure 1. The flow of our system for oral Q&A practice

To better simulate real Q&A situations at academic conferences, the default text interface was hidden. Turn-taking was visualized using alternating icons, as shown in Figure 2. Students can control the timing and pacing of the interaction through on-screen buttons, allowing for more flexible and learner-friendly interaction.

Figure 2. System interface during oral Q&A practice. Left: User speaking. Right: User listening to system output.

The system also provided feedback, including conversation transcripts, ASR confidence scores, word count analysis, and proofreading of student responses. A confidence score is a numerical value calculate by the ASR system that indicates how reliable the transcription from the ASR is. The score is defined mathematically as posterior probability of each word in the transcription given acoustic signals. The proofreading function used ChatGPT to refine the transcripts generated by ASR. Initially, we were concerned that errors in the ASR output might lead to inadequate or misleading corrections. However, when the results of ChatGPT's proofreading were presented to the students, they appreciated ChatGPT's responses highly, indicating that transcription errors did not significantly affect the overall quality of the corrected output, and that the proofreading function worked effectively in practice.

Results

We aggregated the evaluation results from both student and teacher participants, covering all questions generated under the six configurations.

Table 4 presents students' ratings of the questions generated for their own papers, based on the five evaluation criteria.

Table 4. Average student ratings for each criterion and each configuration on a five-point scale (1 is very poor and 5 is very good)

	Relevance	Clarity	Specificity	Inspiration	Predictability
A	3.90	3.96	3.69	2.29	2.61
B	3.97	4.01	3.69	3.18	2.87
C	3.29	4.03	3.97	2.58	2.43
D	4.45	4.37	3.88	3.00	3.03
E	4.30	4.27	3.85	3.28	3.10
F	4.68	4.70	4.35	3.45	3.60

Table 5 summarizes the teachers' evaluations of how likely they would be to ask each generated question. Responses were classified into five options: Likely (L), Rather likely (RL), Cannot judge (C), Rather unlikely (RU), and Unlikely (U). The table also shows L+RL (positive judgments) and U+RU (negative judgments).

Table 5. Percentage of questions that English teachers judged as likely or unlikely to ask (%)

	L	RL	L+RL	C	U+RU	RU	U
A	18.6	29.8	48.4	3.28	48.4	27.0	21.4
B	19.4	28.2	47.6	6.23	46.2	25.8	20.3
C	23.4	21.9	45.2	3.90	50.9	23.2	27.7
D	23.1	34.5	57.6	5.15	37.3	23.6	13.7
E	16.9	37.6	54.5	6.67	38.8	20.5	18.3
F	31.2	36.4	67.6	7.62	24.8	13.6	11.2

Table 6 shows the percentage of questions that were judged by the students to have no direct answers in the paper – defined here as insightful questions. These are shown both for all questions, and specifically for those rated with a Relevance score of 3 or higher.

Table 6. Percentage of questions that students judged to be insightful (%)

	Uploaded			Not uploaded		
	A	B	C	D	E	F
All	46.8	46.2	54.5	29.9	46.7	40.0
Relevance ≥3	35.2	35.4	31.7	28.0	39.6	40.0

Lastly, Table 7 summarizes the overall preferences of both students and teachers across the six configurations. These preferences were gathered via a point allocation task, in which each participant distributed a total of 100 points among the six configurations based on perceived usefulness.

Table 7. Average points allocated by participants to each configuration (out of 100), reflecting overall preference

	Uploaded			Not uploaded		
	A	**B**	**C**	**D**	**E**	**F**
Students	10.0	15.0	9.1	18.4	21.0	26.5
Teachers	10.6	13.7	14.1	17.7	17.4	26.5

Discussion

This study set out to explore how generative AI, specifically ChatGPT, can be effectively used to support non-native English speakers in preparing for oral Q&A sessions at academic conferences. Our findings suggest that the most effective configuration may not require complex prompting or uploading reference materials, but rather a simpler setup that enables ChatGPT to generate less specific questions without additional input.

As the core aim of this study was to identify which configurations are most effective for supporting Q&A practice, we began by analyzing the preferences revealed in Table 7, which summarizes participants' ratings across configurations. Contrary to our expectations, both students and teachers rated Configuration F the highest. This configuration did not specify any sections or keywords, nor did it include reference papers. Interestingly, configurations without reference materials consistently received higher ratings. Table 6 further supports this, showing that Configuration F generated the highest number of "insightful" questions – those for which answers could not be found directly in the paper.

Student responses, as presented in Table 4, indicated that Configuration F was rated highest across all criteria. Interestingly, configurations that included reference materials tended to receive lower ratings, possibly due to the increased specificity or complexity of the questions generated. Teacher evaluations followed a similar trend: Configuration F received the highest L+RL score and the lowest U+RU score, indicating that this configuration generated the most likely questions they would ask.

These results suggest that general-purpose questions, which are similar to those typically asked by teachers, are more effective for Q&A practice than highly specialized ones. This suggests that the analytical and detail-oriented questions GPT tends to ask when reference materials are uploaded may not align well with questions typically asked in international conference Q&A sessions.

To better understand these findings, we compared our results with prior research using GPT in academic peer review settings. Liang et al. (2023) found that the degree of overlap between feedback provided by GPT and that provided by human reviewers was comparable to the overlap observed between two human reviewers. This implies that GPT can be viewed as another human reviewer. However, the research contexts differ significantly between the two cases: in journal peer reviews, reviewers are expected to thoroughly read and analyze the entire paper, while Q&A at conferences typically occurs immediately after a 10–15-minute presentation based solely on slides. Thus, while detailed questions may be appropriate in peer review settings,

less specific and more accessible questions may be better suited to support students in oral Q&A situations.

Another possible interpretation of our findings is related to the evaluators themselves. The students may have lacked sufficient background knowledge to fully grasp the intent and significance of the questions generated by ChatGPT. In actual Q&A sessions, it is not uncommon for the dialogue between the questioner and the presenter to be misaligned, often due to the presenter's lack of sufficient background or related knowledge. A similar situation may have occurred in the evaluation of the questions in this study. Therefore, it is possible that if supervisors or researchers with more domain expertise had evaluated the questions, they might have preferred configurations that incorporated reference materials.

Conclusion

The results of this study show that the effectiveness of integrating LLMs with ASR and TTS offers a novel approach to Q&A practice, enabling students to prepare for English Q&A sessions independently. Both students and teachers consistently rated configuration F – where no reference papers were uploaded and no scope was specified – as the most effective for generating questions simulating Q&A sessions.

After using the system, all participating students expressed strong interest in using it again before attending their next international conferences, highlighting their satisfaction with both the quality of generated questions and the overall speaking interface.

References

Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D., Yang, X., Vodrahalli, K., He, S., Smith, D., Yin, Y., McFarland, D. A., & Zou, J. (2023). Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *arXiv*. https://arxiv.org/abs/2310.01783

OpenAI. (n.d.). *ChatGPT.* Retrieved April 6, 2025, from https://chatgpt.com/

Speak. (n.d.). *The language learning app that gets you speaking*. Retrieved September 10, 2024, from https://www.speak.com/

Wang, Z., Valdez, J., Basu Mallick, D., & Baraniuk, R. G. (2022). Towards human-like educational question generation with large language models. In *International Conference on Artificial Intelligence in Education* (pp. 153–166). Springer.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv*. https://arxiv.org/abs/2302.11382

Author biodata

Mayuko Aiba is a Masters student at the Graduate School of Engineering, The University of Tokyo. Her research focuses on the application of large language models

like ChatGPT in foreign language education, aiming to enhance learners' oral skills while reducing the burden on teachers supporting students from diverse majors.

Daisuke Saito is currently an associate professor in the Graduate School of Engineering, the University of Tokyo. He is interested in various areas of speech engineering, including voice conversion, speech synthesis, acoustic analysis, speaker recognition, and speech recognition.

Nobuaki Minematsu is a professor of Electrical Engineering and Language Education. His research interest covers speech science and technology, especially application of speech technology to support for language education. He is one of the board members of International Speech Communication Association (ISCA) and a scientific committee member of Speech and Language Technology in Education (SLaTE).