GazeQ-GPT: Gaze-Driven Question Generation for Personalized Learning from Short Educational Videos

Benedict Leung
Ontario Tech University
Oshawa, Canada
benedict.leung1@ontariotechu.net

Mariana Shimabukuro
Ontario Tech University
Oshawa, Canada
mariana.shimabukuro@ontariotechu.ca

Matthew Chan
Ontario Tech University
Oshawa, Canada
matthew.chan@ontariotechu.ca

Christopher Collins
Ontario Tech University
Oshawa, Canada
christopher.collins@ontariotechu.ca

Figure 1: GazeQ-GPT uses a gaze-driven interest model to personalize question generation: (a) Using the interface with gaze to watch (b) a video with subtitles highlighting phrases/collocations. (c) Further details of key phrases are located in the marginal gloss activated by gaze. (d) Questions generated by the interest model are based on fixations on words in the subtitle and gloss.

Abstract

Effective comprehension is essential for learning and understanding new material. However, human-generated questions often fail to cater to individual learners' needs and interests. We propose a novel approach that leverages a gaze-driven interest model and a Large Language Model (LLM) to generate personalized comprehension questions automatically for short (~10 min) educational video content. Our interest model scores each word in a subtitle. The top-scoring words are then used to generate questions using an LLM. Additionally, our system provides marginal help by offering phrase definitions (glosses) in subtitles, further facilitating learning. These methods are integrated into a prototype system, GazeQ-GPT, automatically focusing learning material on specific content that interests or challenges them, promoting more personalized learning. A user study (N = 40) shows that GazeQ-GPT prioritizes words in the fixated gloss and rewatched subtitles with higher ratings toward glossed videos. Compared to ChatGPT, GazeQ-GPT achieves higher question diversity while maintaining quality, indicating its potential to improve personalized learning experiences through dynamic content adaptation.

GI '25, May 26–29, 2025, Kelowna, BC, Canada

Keywords

eye tracking, question generation, LLM, interest model, collocation, complex word identification

ACM Reference Format:

Benedict Leung, Mariana Shimabukuro, Matthew Chan, and Christopher Collins. 2025. GazeQ-GPT: Gaze-Driven Question Generation for Personalized Learning from Short Educational Videos. In *Graphics Interface (GI '25), May 26–29, 2025, Kelowna, BC, Canada.* ACM, New York, NY, USA, 13 pages.

1 Introduction

Personalization is integral to modern user interfaces, both implicit (e.g., targeted ads, content recommendations) and explicit (e.g., UI layout changes, content requests). In education, personalization enhances motivation by tailoring content to individual learners, though it doesn't directly impact learning outcomes [31, 46]. As learning needs evolve with instant access to information, such as informal learning through educational videos, personalization accelerates and deepens learning by adapting to skills, interests, and needs [22, 28, 45].

However, informal learning through educational videos (e.g., YouTube videos) has limited engagement with the learners. Furthermore, self-directed learners do not have a way of tracking or assessing their comprehension effectively, which is an important aspect of informal learning [9]. Self-directed learners who watch videos over five minutes watch in a non-linear manner (rewatching or skipping portions) [29]. Responding to these aspects of video learning, our work aims to generate and personalize comprehension quizzes on video content using implicit gaze interactions with subtitles and, in turn, increase content engagement with learners.

Multiple-choice quizzes are a standard method to measure and aid learners' comprehension of educational content. However, they generally lack personalization, as questions are generated on generally important sentences [27, 44]. Designing comprehension quizzes takes a lot of care to ensure every choice (correct answer and distractors) relates to the topic and there is only one correct answer. Past research has generated multiple-choice questions on plain text and documents [4, 27, 44, 55, 62, 65, 69]. Yet, these works do not consider the individual learner's needs nor tackle multimedia content, such as educational videos and subtitles. Fortunately, this is where Large Language Models (LLMs) thrive, such as OpenAI's GPT-4 model [41], to assist in learning. Using its language capabilities, we can generate multiple-choice questions through well-engineered prompts.

Another challenging aspect of educational videos that may hinder the learners' comprehension is the presence of technical terms or jargon. On the other hand, glossed reading has been shown in text reading and subtitled videos to significantly increase the learning of new words compared to non-glossed reading [34, 71]. A gloss is a brief explanatory note or definition used to clarify the meaning of a term or phrase in a text.

To assist the learner's comprehension of educational video content, our proposed approach includes **(1) collocation detection algorithm to detect technical jargon** where **(2) gaze-triggered glosses to display said jargon**, which may be challenging to understand and impact overall comprehension of the new material and **(3) a gaze-driven interest model to personalize multiple-choice quiz questions for short educational videos**.

Traditional collocation/multi-word detection methods typically focus on fixed-length word pairs or manual extraction, and no definitions for detected collocations have been provided. Thus, our novel approach uses GPT-4 to detect and define jargon to be displayed by gloss (Figure 1a and b). In addition, we developed a gaze-driven interest model to score words in the subtitle or gloss, using attention as a proxy for interest. Similar gaze-driven models have been in different contexts, including differentiating language learners' proficiency levels and assisting with information visualization tasks [6, 37, 64, 67]. In our proposed system, GazeQ-GPT, we combined these techniques, including the gaze interest scores, to guide the personalized question generation by focusing on specific subtitles and words (Figure 1c).

In a user study, we compare GazeQ-GPT with and without gloss and have participants answer comprehension questions generated by ChatGPT with and without the gaze-driven interest model to characterize the benefits of LLMs and eye-tracking in personalizing the learner's experience. While both ChatGPT and GazeQ-GPT's questions were perceived as helpful to learners, the evaluation shows the variance across participants in the GazeQ-GPT questions and the prioritization of rewatched subtitles and fixated glosses, indicating personalization. When comparing question sets within GazeQ-GPT and ChatGPT, ChatGPT showed only slight variance, with questions repeated. In contrast, GazeQ-GPT exhibited high variance while providing meaningful guidance without losing question quality. Furthermore, glossed videos were rated more highly than those without gloss regarding usability.

2 Related Work

We were inspired by previous attempts at automatic question generation and existing research on eye-tracking to model user interest to guide linguistic processes.

2.1 Question Generation

Quality question generation is crucial for evaluating learner knowledge and fostering self-motivated learning. However, creating suitable questions can be labour-intensive, leading to significant research in automatic question generation (AQG) to reduce this burden. Kurdi et al. [27] noted that existing AQG methods often produce simplistic questions targeting lower levels of learner ability. Pan et al. [44] echoed this, adding that personalized question generation remains underexplored and suggesting that modelling user state and awareness could enhance personalization. This aligns with the review of Kurdi et al., which highlighted that current approaches generate all possible questions or analyze important sentences without considering individual needs. One notable attempt to model the user state is by Syed et al. [56], where they improved long-term learning outcomes by creating personalized quizzes using gaze tracking. They generated questions using skimmed and focused reading behaviour. However, their work only focuses on articles. Fixation behaviour for multimedia content, such as subtitled videos, differs from static texts. Furthermore, it is unclear whether the language style (linguistic registry) difference between writing and speaking will affect question quality. Our approach extends to

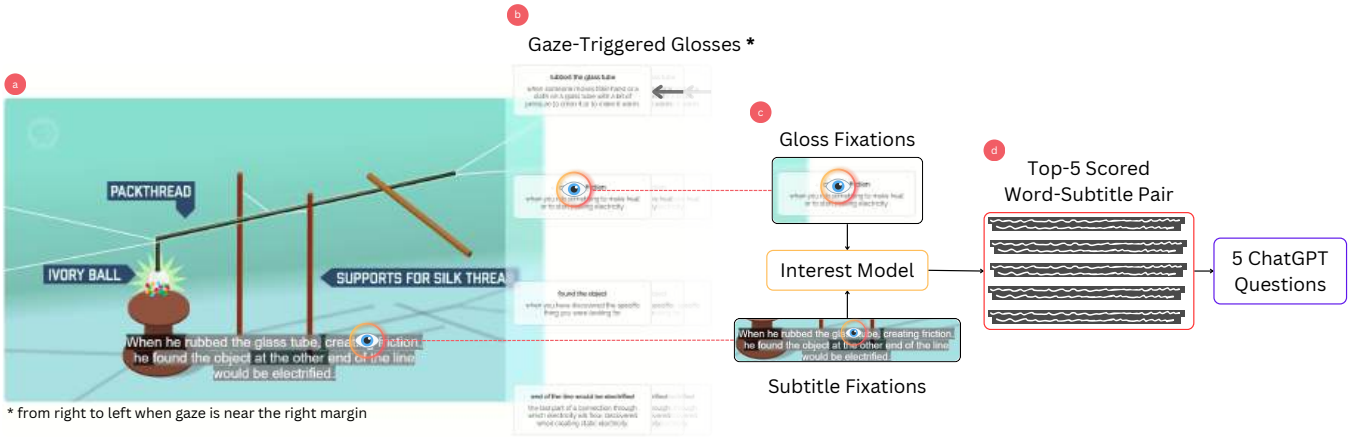

Figure 2: An overview of the implementation of GazeQ-GPT. (a) The GazeQ-GPT interface shows the video with subtitles highlighting collocations. (b) Gaze-triggered glosses will display said collocations, which are initially hidden in the right margin of the screen. The gloss will trigger once the gaze is near the right margin. (c) The interest model takes gloss and subtitle fixations to score subtitles. (d) The top-5 scored word-subtitle pair is used for context to produce 5 questions from ChatGPT.

model the user's state, as suggested by Kurdi et al. [27], using gaze for subtitled videos following the technique of Syed et al. [56] to generate personalized quizzes based on gaze patterns.

Transformers [63] have been used to generate higher quality questions [4, 62, 65]. One prominent transformer-based architecture is OpenAI's GPT model. Prior research has explored using ChatGPT for question generation [65, 70], but these studies offer limited detail on prompt engineering, relying on simplistic prompts that may compromise question quality. Additionally, as previously discussed, prior works leveraging ChatGPT do not offer personalized question generation. In contrast, our approach, GazeQ-GPT, harnesses ChatGPT's broad knowledge base to generate domain-specific questions guided by our gaze-driven interest model.

2.2 Eye Tracking for Linguistic Processing in Subtitles

Eye tracking allows researchers to study the types of eye movement, such as fixations (points at which people pause, ranging from 150ms to 300ms [61]) and saccades (lengths between these fixations). There are two assumptions when measuring reaction times: (1) longer fixation duration and more fixations indicate greater processing effort, and shorter fixations and/or skipping indicate less processing effort. (2) What is being fixated is what is being considered [47], with much literature demonstrating the relationship between eye movements and attention, cognitive state, decision making and memory [15].

Our work focuses on eye tracking during video viewing with subtitles. Previous research primarily explores how gaze data, such as fixation count and duration, relates to subtitle processing for language learning [6, 37, 67]. Findings indicate that gaze patterns can reveal learners' language backgrounds and familiarity with content. For instance, Muñoz [38] showed that beginners tend to skip subtitles less than more advanced learners. Machine learning models have been used to predict English proficiency from gaze features [14, 35, 74]. Related work, such as SubMe [21], uses gaze patterns to classify learner skill levels and generate lists of difficult words with personalized translations and definitions.

On the other hand, fixations on video content depend on an individual's behaviour. For example, subtitles facilitate comprehension regardless of cognitive abilities, eye movement strategies or age, while only video content depends on these individual factors [18, 78]. Furthermore, there is a high correlation between subtitle reading and performance compared to fixations on video content [26, 39], suggesting individuals' tendency to process subtitles is consistent. People instinctively start reading subtitles as soon as they appear, even if they have little experience with this information without a trade-off between text and image processing [18]. This tendency is stronger when the subtitles are informative, specifically when the language of the soundtrack is unfamiliar, there is minimal overlap between the written text and the images, and the subtitles provide valuable information [17, 18].

Based on previous findings, this work presents a gaze-driven interest model based on only subtitles. Subtitle reading behaviour is consistent between individuals, while fixations on video content depend on the visuals and individual factors such as cognitive abilities and eye movement strategies, which can be unstable. Our approach advances this by triggering real-time glosses and personalized quizzes from educational videos based on gaze patterns.

3 GazeQ-GPT: Methodology

Our work supports learners by providing gaze-driven glosses for technical terms and personalized multiple-choice questions during video watching. To achieve this with low latency, GazeQ-GPT first analyzes subtitles to detect complex words and phrases, then generates glosses for each using GPT-4. Gaze patterns are used to trigger gloss display and build an interest model, which is used with GPT-4 to generate personalized questions at the end of the video. Figure 2 shows the implementation overview for GazeQ-GPT; the details are described in this section.

3.1 Complex Word Identification

Complex word identification (CWI) identifies the complexity of a given word or multi-word expression. Previous results used supervised learning and feature building to create CWI models [43, 73], but they are not easy to extend to different languages without prior setup and feature engineering [75]. Fixed general complexity scales are challenging to interpret as they aggregate an arbitrary number of absolute binary complexity judgments to give a continuous value [53]. Individual characteristics also impact lexical complexity. For example, a lawyer reading a physics article may struggle with technical jargon more than a physicist. Personalized approaches have considered demographics including language proficiency, native language, race, job, age, and education [30, 57, 58, 76]. LLMs have also recently been used to identify complex words [54, 59]. In this work, we leverage LLMs to tackle personalized CWI.

Our LLM prompt is inspired by two forms of complexity [40]. *Absolute* complexity refers to objective linguistic properties, such as the number of morphemes, the presence of derivational affixes, or having multiple meanings. *Relative* refers to individual experiences or psycholinguistic factors, such as acquisition difficulty or level of familiarity. Inspired by these properties, we made a checklist for ChatGPT to determine the word's personalized complexity, incorporating the **target audience** (e.g., 'undergraduate student'). ChatGPT outputs a score (1–5), which becomes an input to the interest model (subsection 3.4). The prompt goes as follows:

```
Here is some information to analyze the word's complexity:
1. Words having multiple meanings are more complex.
2. The word's higher cognitive load or demand is more
complex.
3. Higher acquisition difficulty of the word is more
complex.
4. Rarer words are more complex.

Consider that the person reading this word is [audience].
```

3.2 Collocation Detection

Technical jargon in educational videos can reduce accessibility and cause confusion. Previous work has offered one-word definitions [19, 21, 77], but these often fail to capture the nuances of multi-word phrases/collocations, groups of words that form a semantic unit. While experts can extract and define such phrases, manual processing is labor-intensive [2, 3, 10, 11]. Past collocation detection methods [25] (statistical, rule-based, and hybrid) typically focus on fixed-length word pairs [16], and none have provided definitions for detected collocations.

We leveraged ChatGPT (`gpt-4-turbo`) in an iterative approach when detecting collocations. Our approach starts by tokenizing the subtitle, then iteratively combines consecutive tokens to form phrases using ChatGPT, starting from the first token. To confirm if a candidate sequence forms a valid phrase, we asked ChatGPT in the form of a true/false question [23]:

```
"[token sequence]" is a phrase (Context: [context])
A) True
B) False
```

For every detected phrase, we have ChatGPT provide three simple definitions informed by the context of the complete subtitle: the definition of the whole phrase and definitions for the two most important words in the collocation.

3.3 Gloss

A gloss is a brief explanatory note or definition used to clarify the meaning of a term or phrase in a text. Glossed reading leads to significantly greater learning of words in contrast to non-glossed reading [71]. Glosses can be noninteractive (inserted at a specific place, e.g., margin) and interactive (an action required to activate, e.g. hyperlink) [71]. Previous works used interlinear and hyperlinked glosses in subtitles to define one word [19, 21, 60]. Our work extends this to collocations and phrases. As showing too much help in the subtitle area may overwhelm users, we take an interactive approach, placing gaze-activated glosses in the margin.

3.4 Gaze-Driven Interest Model

When reading text, people often fixate on complex and less commonly used words for longer when processing them visually [13, 48]. Thus, to model which words/subtitles the user is fixated on, we start with an interest model developed for data visualization [64]:

$$S_{ij} = \frac{n_i}{n_{all}} \times (log(n_i) + 1) \times \sqrt{\Delta t} \times \frac{1}{d_j + 1} \tag{1}$$

where S_{ij} is the interest score for each object j in group i. The inputs to the score are: the number of objects in group i (n_i), the total number of objects visited by the fixation (n_{all}), the time elapsed (Δt) and distance between object j and the fixation point. We adapt to score interest in a word in a subtitle as follows:

$$S_{s_i w_j f_k} = \frac{1}{n_{all}} \times \sqrt{\Delta t} \times \frac{1}{d_j + 1} \times c(w_j) \tag{2}$$

where $S_{s_i w_j f_k}$ is the score for word j in subtitle i for fixation k. Here, $n_i = 1$ as the interest group contains only one word. Resulting constant terms are dropped. A factor for the complexity of the word ($c(w_j)$) is added as described in subsection 3.1.

For fixations in marginal glosses, an alternate formulation is used, as it is assumed the user will look at all words in one collocation at each fixation. Thus the distance term in eq. 1 is dropped as a constant, and the number of words (n_i) and total number of words visited by the fixation (n_{all}) will be the same, leading to:

$$S_{g_i w_j f_k} = (log(n_{all}) + 1) \times \sqrt{\Delta t} \times c(w_j) \tag{3}$$

Each word will have a combined score using these two formulas. Finally, the score of the subtitle is calculated by summing all word scores in the subtitle for each fixation and normalizing by the duration of the subtitle:

$$S_{s_i} = \frac{\sum_{w_j \in S_i} \sum_{k=0}^{K} S_{s_i w_j f_k} + S_{g_i w_j f_k}}{t(s_i)} \tag{4}$$

For each fixation, the group n_{all} is made of words in the paracentral vision region. This region changes based on the user's head distance from the screen, so a fixed-size circle around the gazepoint would be inaccurate. Thus, a formula to compute this region depends on the distance between the user and the screen (d) in cm,

the resolution of the screen (*PPI*) with a constant conversion factor from inches to cm, and the angular size of the paracentral vision (α) [64]. In this case, the paracentral's angular size is 5°. All objects within the circular region around the gaze point will be selected for the interest model. The radius of the region is defined by:

$$r = d \times tan(\alpha) \times 0.393701 \times PPI \quad (5)$$

3.5 Automatic Question Generation

We used multiple-choice questions for our post-test. Xiao et al. [70] found that ChatGPT-generated questions were too simplistic (e.g., "What is something?") and had easy distractors, due to straightforward prompts. To address these issues, we set three AQG requirements: (1) target content with the highest interest score (subsection 3.4) to focus on challenging areas for learners, (2) ensure choices explain a concept or idea to avoid vague questions, and (3) ensure distractors are related to the correct answer to avoid irrelevant options. The criteria prompt is as follows:

```
Here are the criteria for the question:
1. The question must have the word: "[word]".
2. All choices should explain a concept or an idea in
a sentence about 15 words long without giving away the
answer.
3. All incorrect choices must be from the video and
related to the correct choice.
4. All choices should have a similar number of words.
```

ChatGPT generates questions with four choices (one correct answer and three distractors) for target subtitles, including five seconds before and after, with the complete subtitle file in its knowledge base for context.

3.6 Implementation

A web application implemented GazeQ-GPT using ReactJS [36] and ElectronJS [20] to display the videos and glosses and communicate with an eye tracker. The official OpenAI API library [42] was used for executing prompts. The full prompts are in the appendix (subsection A.2).

3.6.1 Eye Tracking. We used a Tobii Eye Tracker 5, downsampled to 33Hz. It also tracks head position. To address gaze jitter and microsaccades, we applied 1€ smoothing [8]. For classifying gaze into saccades and fixations, we followed Lobão-Neto et al. [33], who recommended context-specific parameters (img: $\sigma = 0.01$, $\lambda = 0.95$, $\alpha = 0.095$, $\beta = 1.0$; vid: $\sigma = 0.01$, $\lambda = 0.95$, $\alpha = 0.025$, $\beta = 1.0$). To adjust for sampling rate differences, we adjusted the moment transition (α) to 0.175 for video/subtitle fixations and 0.665 for marginal gloss (img). Fixations are analyzed to extract words from the paracentral region and calculate the interest score as detailed in subsection 3.4.

3.6.2 Collocations in Marginal Gloss. The subtitle highlights available collocations (Figure 2a), in which the marginal gloss displayed definitions of said collocations. Initially, the marginal gloss is hidden off-screen (right side). To show the gloss, the gaze must shift near the right margin of the screen (Figure 2b). Once the gloss has been activated, the video will pause.

3.6.3 Question Generation Process. Once the video ends, GazeQ-GPT administers a five-question multiple-choice comprehension quiz. The system selects the top five subtitles scored by the interest model and selects the highest-scored word as the target word (Figure 2c). For each (word, subtitle) pair, a multiple-choice question with four choices is generated (Figure 2d). Explanations are also generated to give feedback on why each choice is incorrect or correct. Once all questions are generated, the order of the questions and choices are randomized. Throughout the process, the complete subtitle file is also in the knowledge base via file retrieval. We used `gpt-4o` due to its fast generation speed.

4 User Study

To evaluate whether GazeQ-GPT can automatically generate effective questions for video content using subtitles and assess if gloss can enhance comprehension on videos, we conducted a user study to measure the helpfulness of marginal gloss for jargon and explore the personalization and helpfulness of questions contrasting with (GazeQ-GPT) and without (ChatGPT) the interest model using only subtitles.

4.1 Participants and Apparatus

We recruited 40 participants using recruitment posters and a mass email sent to students. Individuals reported their highest or current degree: a Bachelor's degree ($N = 25$), a Master's degree ($N = 9$), a PhD ($N = 2$), or a high school diploma or lower ($N = 4$). On a 5-point scale from 1-"never" to 5-"all the time," they reported the frequency of how often they enable subtitles as $Mdn = 4$ and their strategies for catching up with video content. Strategies reported are rewinding parts of the video ($N = 26$), reading subtitles ($N = 14$), taking notes and searching concepts on the internet ($N = 4$), and no strategies ($N = 2$) (e.g., binge-watching).

The study took place on campus, and participants were in person. They used a desktop with a 24-inch monitor, keyboard, and mouse, running the web application locally. The participant's screen, interactions, and tool logs were all recorded. Participants received the equivalent of $20 CAD.

4.2 Design

The study followed a mixed design with two conditions. GLOSS (without gloss vs. with gloss) is a between-subject condition where half the participants were exposed to gloss. QUESTION TYPE (GazeQ-GPT vs. ChatGPT) is a within-subject condition where all were exposed to both sets of questions. Two educational videos (~9min) ("The History of Chemical Engineering" and "The History of Electrical Engineering" from CrashCourse[1][2]) were used. Video and question orders were counterbalanced — half of the participants watched video one first, and half completed the ChatGPT questions first. Participants were unaware of the difference in the question generation process.

[1] https://www.youtube.com/watch?v=aRKyJRAxjpM
[2] https://www.youtube.com/watch?v=3nB1Ntku06w

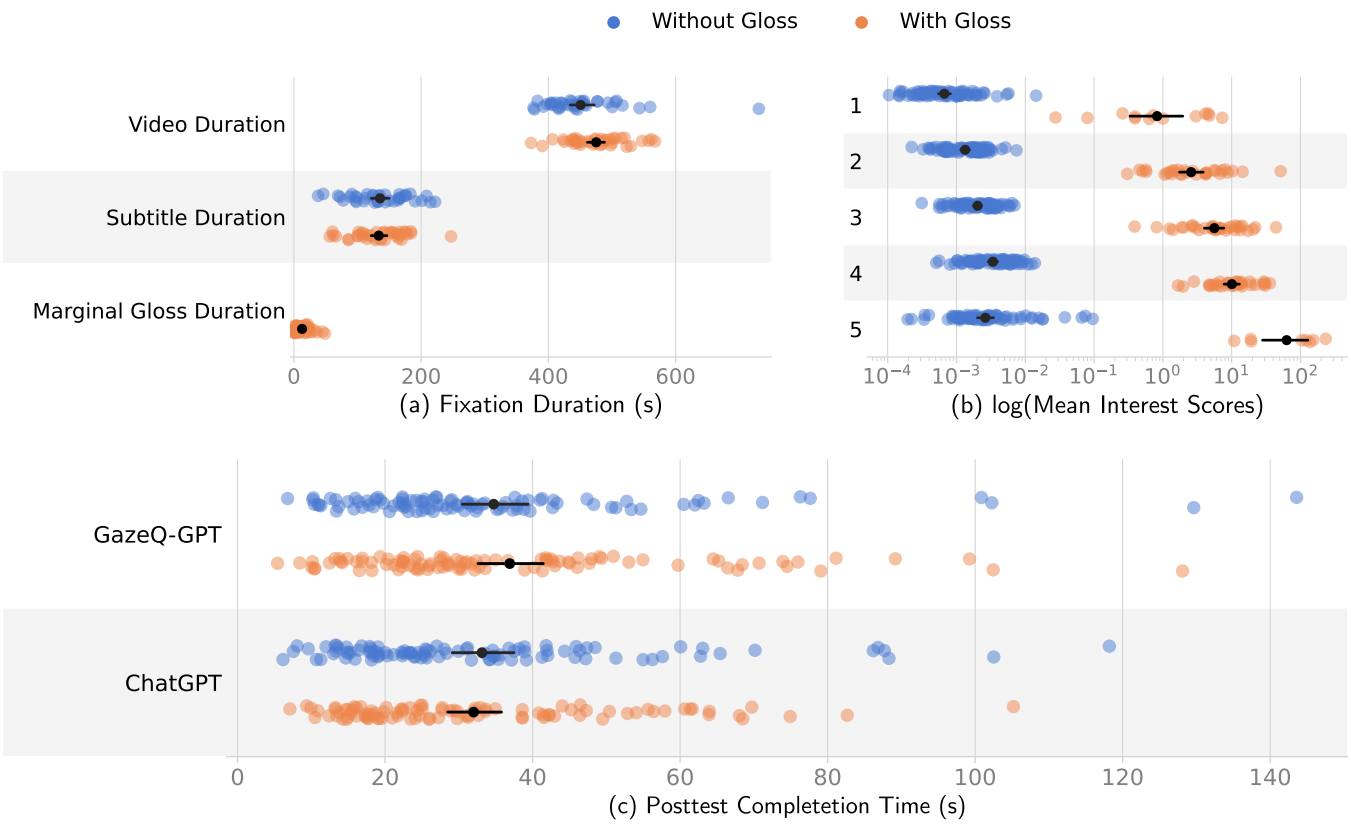

Figure 3: Data for the question generation and testing: (a) fixation durations in screen regions, (b) subtitle and gloss interest scores grouped by word complexity, (c) post-test completion time for each question. Black points represent means and 95% CI.

4.3 Procedure and Tasks

Introduction: After completing the consent form and the demographics questionnaire, the participant was told to evaluate the questions after watching the video.

Tutorial: For participants watching with gloss, a video demonstrates how to activate the gloss and must activate it at least once before proceeding.

Watch Videos: The participants watched the whole video and can rewind it using the progress bar or arrow keys to rewatch parts. Participants with gloss can also activate the gloss at any time. Once the video ended, the participants answered five multiple-choice questions. A 25-second delay was added to ChatGPT questions to mock the generation process of GazeQ-GPT questions. Participants must answer each question correctly and rate it (like or dislike).

Questionnaires: After each video, the participant completed a questionnaire and 5-point scale statements. The questionnaire was identical for both methods. After both videos, a final questionnaire, including a System Usability Scale [7], is administered.

5 Results

All 40 participants watched, answered, and rated all questions without abandoning. We analyzed how our interest model drove the question-generation process by detailing the fixation duration and interest scores. Finally, questionnaires (5-point statements and usability) are detailed below. Results are visualized in Figure 3 and Figure 4.

Fixation Duration. Participants fixated on the video more than the subtitles, and the marginal gloss the least. Participants without gloss fixated on the video less (M=7m31s; 95% CI: [7m14s, 7m51s]) compared to participants with gloss (M=7m55s; 95% CI: [7m41s, 8m10s]). There was no difference in subtitle fixation.

Interest Scores. A trend of greater interest for more complex words was observed across both subtitles and gloss.

Post-test Completion Time. Using ART ANOVA [68], there was a significant main effect of QUESTION TYPE ($F(1, 349) = 4.85, p = 0.049, \eta_p^2 = 0.011$) where participants took an average of 4s (95% CI: [-0.2s, 7s]) longer completing GazeQ-GPT questions.

Binary Question Ratings. Participants rated each question on a binary scale (e.g. "like" and "dislike"). 175 ChatGPT questions and 171 GazeQ-GPT questions were liked out of 200 ratings each. Using McNemar's test with continuity correction, there was no significant effect ($\chi^2 = 13.0, p = 0.22, \phi = 0.25$).

Usability. On the System Usability Scale, with gloss was rated 88 (95% CI: [82.66, 92.17]), typically considered to be "Excellent" [1], whereas without gloss was rated 80 (95% CI: [75.42, 84.50]), between

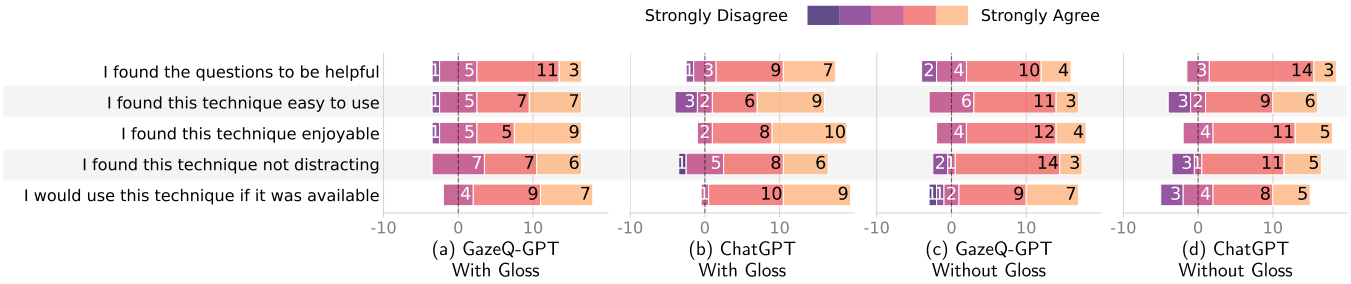

Figure 4: Ratings for 5-point statements for (a) GazeQ-GPT with gloss, (b) ChatGPT with gloss, (c) GazeQ-GPT without gloss, and (d) ChatGPT without gloss.

"Good" and "Excellent." Using the Mann-Whitney U test, GLOSS significantly affected usability ($U = 0.45, p = 0.019, r = 0.45$).

Questionnaires. Participants rated all five 5-point statements: "I found the questions to be helpful" ($IQR = (3-4), (4-5), (3-4), (4-4)$), "I found this technique easy to use" ($IQR = (3-5), (3.75-5), (3-4), (3.75-5)$), "I found this technique enjoyable" ($IQR = (3-5), (4-5), (4-4), (4-4.25)$), "I found this technique not distracting" ($IQR = (3-5), (3-5), (4-4), (4-4.25)$), "I would use this technique if it was available" ($IQR = (4-5), (4-5), (4-5), (3-4.25)$) for GazeQ-GPT with gloss, ChatGPT with gloss, GazeQ-GPT without gloss and ChatGPT without gloss, respectively. All statements have a $Mdn = 4$ except for "I found this technique enjoyable" for ChatGPT questions with gloss with a $Mdn = 4.5$. Using ART ANOVA, there were no significant effects ($p > 0.05$). For participants with gloss, they rated "I found the marginal gloss helpful" with $Mdn = 4$ ($IQR = (4-5)$) with 9 rated "Strongly agree", 7 rated "Agree", 3 rated "Neutral" and 1 rated "Disagree".

5.1 Participant's Comments

This section will discuss the results and describe the themes in the participants' comments.

Marginal Gloss Usability. Participants' comments on the marginal gloss were favoured. For example, it helped them understand the video better (e.g., *"I really liked the definitions sidebar because it helped me understand more complex words and then understand the whole concept being taught in the video. It was also really helpful because I did not have to use my mouse, and using my eyes made it much quicker."* – P4 and *'I found the marginal definitions so useful because some terms were technical for someone who is not familiar with the field."* – P12). They also favoured the ease of accessing the marginal gloss via gaze (e.g., *"I liked that the words with definitions provided in the side panel were highlighted so I knew when to look to the right if I needed it."* – P20).

Question Quality. Participants were asked to comment on the quality of the questions. We replaced the "first set" or "second set" with their associated question set (i.e., GazeQ-GPT or ChatGPT) for easy reading. The questions helped them understand the content and gave great explanations (e.g. *"I thought the [GazeQ-GPT] questions were relevant to the important topics of the video."* – P1) on why the option is incorrect/correct (e.g., *"Both sets of questions were challenging but gave great explanations as to what the correct responses were."* – P19 and *"During the [post-test], I liked how you were given feedback if the answer that you selected was incorrect."* – P4). Some thought it would be a good knowledge check and a way to review the content afterward (e.g., *"I think this method of teaching and learning will help both the teacher and student to know how the students are grasping the knowledge from the course content and how well do they know the course respectively."* – P14).

However, there were conflicting comments on how well-written the questions were. For example, the questions were straightforward (e.g., *"[GazeQ-GPT] questions were clear"* – P16, *"The questions didn't hurt my brain to read."* – P30, *"I thought the quality of the questions was pretty good because it helps to see how much information you retained from the videos."* – P37 and *"[GazeQ-GPT] questions related more closely to the video, without being overly detailed to the point where the information can't be remembered."* – P40) or hard to read (e.g., *"I think some of the [ChatGPT] questions are hard to understand."* – P33). Some mentioned inconsistent quality of the distractors (e.g., *"Some of the options in the test were not related to the question"* – P16).

6 Discussion

The study's main takeaway is that GazeQ-GPT implicitly models user interest by prioritizing fixated words in the marginal gloss and rewatched subtitles. Both question types were suitable, but GazeQ-GPT has more variance in the questions generated while providing meaningful guidance for the LLM than ChatGPT. Marginal gloss helped participants understand the video better and was not distracting. Overall, GazeQ-GPT can generate various questions based on meaningful viewing behaviours and subtitles for educational videos, while marginal gloss improves video comprehension. The details are outlined below.

Marginal gloss helped understand the video better. The comments about the marginal gloss were positive. They agreed the collocation definitions helped them understand the concepts being taught. They also liked how the gaze-triggered gloss was easy and faster to use than the mouse. Note that all collocations were highlighted in the subtitle. Still, there were no significant levels of distraction compared to the no gloss condition, suggesting they hold value in understanding technical jargon to understand the concepts better. Furthermore, the use of gloss positively impacted the SUS score, providing additional evidence to support this finding.

Both question types are suitable using only subtitles. Based on the 5-point statements, comments and binary ratings, ChatGPT and

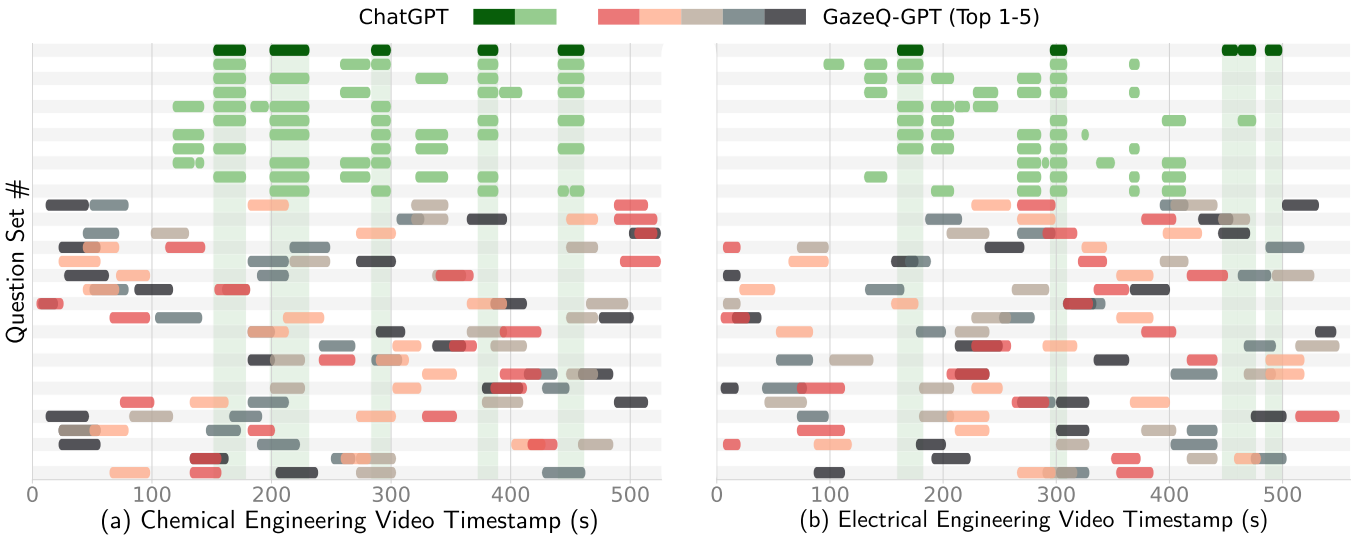

Figure 5: Top-5 scoring subtitles scored by the interest model for (a) chemical engineering and (b) electrical engineering videos. The green rows represent what ChatGPT thought was important, and the first row represents the questions used in the study. Each subsequent row represents a question set for a participant. It shows that target subtitles varied between participants, while ChatGPT showed little variance. Green verticals highlight alignment to the timestamps used in the ChatGPT condition.

Table 1: Sample question set from the most similar timestamps between two participants. Bolded words are the target words used in the prompt.

P3	What was the key point of the development of the **point**-contact transistor in 1947?
	Thomas Edison tried to **discredit** the push for alternating current (AC) by:
	In 1968, American engineer **Marcian** Hoff contributed significantly to computing by developing which of the following advancements?
	In the context of the video, how did Samuel **Morse** utilize his morse code development to advance telecommunication in the United States?
	In the context of early electrical engineering, what was a primary purpose of the Gramme **dynamo** developed by Zénobe-Théophile Gramme?
P38	What was the primary **work** of early computers before World War II?
	What strategy did Thomas Edison use in the War of **Currents** to discredit alternating currents (AC)?
	How did Samuel Morse utilize the **electromagnet** in the development of the telegraph?
	Which statement best describes how sound is **reproduced** in Alexander Graham Bell's telephone invention?
	Which statement best describes the evolution of electrical **engineering** in relation to computers?

GazeQ-GPT were perceived positively in learning the content. Previous AQG works rely on articles and texts. Still, whether subtitles or video scripts can provide enough information to generate meaningful questions is unclear. The study shows subtitles can provide enough information and ignore noise to provide helpful questions based on the 5-point statements. For example, the difference in register between writing and speaking and subtitle timestamps presented in the file does not affect the quality of the question. GazeQ-GPT further changes and shrinks the context by extracting the subtitle, but ultimately, it still provides the same quality as ChatGPT. This suggests subtitle content can provide enough context to generate viable questions for educational videos.

GazeQ-GPT prioritizes fixated words in the marginal gloss and rewatched subtitles. Our interest model implicitly guides the question generation process to generate questions based on fixated marginal gloss despite having a total average fixated time of 13s. It is evident in the interest scores, as none of the confidence intervals overlap with the non-gloss scores. This could be due to reading the definition resulting in longer fixation duration and, as a result, an increase in interest score.

The primary strategy reported was rewinding parts of the video (N = 26). Thus, we analyzed whether our interest model can implicitly model this behaviour in the question generation. 13 participants rewound parts of the video. Out of the 13 participants, the model targeted 81.5% (95% CI: [65.6%, 95.4%]) of the rewatched subtitles.

GazeQ-GPT produces a variety of questions between participants, indicating personalization. Our interest model extracted different subtitles and prompt words between participants, resulting in personalized questions. For example, the questions varied when extracting the most similar timestamps between two participants on the same video (Table 1). Figure 5 shows the different contexts used for the question generation for all participants. Our interest

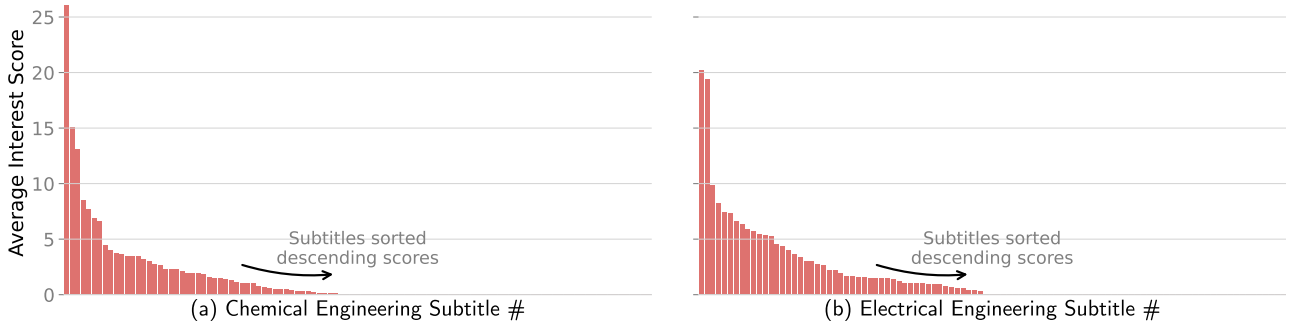

Figure 6: Average interest score across participants for each subtitle for (a) chemical engineering and (b) electrical engineering video. The red bars are sorted in ascending order. The exponential distribution indicates some subtitles had more interest than others, indicating meaningful guidance to the LLM.

model can extract many contexts at a time and, thus, create multiple meaningful questions of interest (e.g., fixated gloss or rewatched subtitles) in parallel.

To further investigate the variation of AQG, we compared the variance between ChatGPT and GazeQ-GPT, generating an additional 10 ChatGPT question sets. The green rows in Figure 5 represent the timestamps extracted by ChatGPT, where the first row is the question set used in the study. It shows a slight variance between the question sets compared to GazeQ-GPT. In some cases, questions can duplicate within the question set even with the previous history of questions and will need to be re-generated, requiring human verification. Comparing the generation process, our method can generate questions in parallel without the additional help ChatGPT requires and without sacrificing question quality.

GazeQ-GPT provides meaningful guidance to the LLM without AQG quality loss. Our method provides a greater variance in the questions generated. To verify whether our interest model selected subtitles randomly, we plotted (Figure 6) the distribution of the average interest score for each subtitle across participants. If the selection were random, the expected distribution would be constant. However, Figure 6 shows that the average interest scores follow an exponential distribution. Most of the subtitles have an interest score close to zero despite participants fixating on an average of 84% of the subtitles for at least 500ms. The same trend can be observed for both videos. Additionally, the subtitles with the highest average score contain valuable information. For example, the highest average score for the subtitle for electrical engineering is "This is called arc lighting," and the lowest is "Well, it can be dangerous." This higher variance in questions could also explain the longer post-test completion time for GazeQ-GPT, where participants must recall secondary topics. Furthermore, there was no significant difference in the questions' quality based on the comments, binary ratings and 5-point statements, suggesting that guidance by fixations on subtitles does not affect the AQG quality. Overall, the interest model can adapt to different viewing behaviours and focus on which segments of the video (subtitle) to generate questions about. GazeQ-GPT questions can be found in the supplementary material.

7 Limitations and Future Work

The results could be LLM dependent. We used GPT-4o for both ChatGPT and GazeQ-GPT, but it can still produce hallucinations, such as irrelevant definitions. This rare issue can be resolved by regenerating the question without significantly affecting the overall experience, as indicated by 5-point statements. Additionally, question readability varied among participants. To address this, we could over-generate and rank questions by difficulty, response time, and readability [5, 50, 51, 72].

Participants' prior knowledge could affect usability. Only two participants are in electrical engineering. Therefore, it is unclear whether the marginal gloss would have any effect if they have prior knowledge of technical jargon or in the field. Future works should consider curating advanced marginal content for users with prior knowledge. Most participants were also undergraduates, future works should consider simpler topics (e.g. Grade 1 math) or different demographics.

Video duration. The chosen videos for GazeQ-GPT were around 10 minutes long. Thus, the knowledge base of the subtitle file in GPT-4o is relatively small. Longer videos could affect GPT-4o's retrieval performance and, thus, question generation. This could be mitigated by subdividing the video and subtitles. For longer videos, an interest score threshold could be used to generate questions during playback, rather than queuing to the end.

Different video domains, such as language learning or training videos, should also be considered. Language learning videos may require different question-generation prompts, focusing more on vocabulary, grammar, or comprehension, while training videos might emphasize procedural knowledge and application.

Alternative applications. Our interest model only considers gloss and subtitle fixation duration. Another approach is to use the video content to drive question generation with multimodal models considering visual frames content and saliency (i.e., detecting a diagram or schematic), or apply the approach to reading text documents.

8 Conclusion

We have described GazeQ-GPT, which implements a method for a personalizing question-generation process in a subtitled video context by leveraging a gaze-driven interest model and LLMs. Our

user study found that GazeQ-GPT's gaze-triggered marginal gloss improves usability as it provides brief explanations and definitions for technical jargons to improve the learner's understanding of complex terms and concepts encountered in the videos. Furthermore, the study results comparing GazeQ-GPT and ChatGPT sets of questions, showed that our interest model successfully implicitly models the learner's behaviour, guiding the questions to generate based on fixated gloss and rewatched subtitles. Both question types were found to be helpful in a video context. However, GazeQ-GPT produces a variety of questions for learners and personalizes the question generation on the fly.

References

[1] Philip T. Kortum Aaron Bangor and James T. Miller. 2008. An Empirical Evaluation of the System Usability Scale. *International Journal of Human–Computer Interaction* 24, 6 (2008), 574–594. https://doi.org/10.1080/10447310802205776 arXiv:https://doi.org/10.1080/10447310802205776

[2] Ali Farhan AbuSeileek. 2011. Hypermedia annotation presentation: The effect of location and type on the EFL learners' achievement in reading comprehension and vocabulary acquisition. *Computers & Education* 57, 1 (2011), 1281–1291. https://doi.org/10.1016/j.compedu.2011.01.011

[3] Ali Farhan Munify AbuSeileek. 2013. Hypermedia Annotation Presentation: Learners' Preferences and Effect on EFL Reading Comprehension and Vocabulary Acquisition. *CALICO Journal* 25, 2 (Jan. 2013), 260–275. https://doi.org/10.1558/cj.v25i2.260-275

[4] Nischal Ashok Kumar, Nigel Fernandez, Zichao Wang, and Andrew Lan. 2023. Improving Reading Comprehension Question Generation with Data Augmentation and Overgenerate-and-rank. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, Ekaterina Kochmar, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Nitin Madnani, Anaïs Tack, Victoria Yaneva, Zheng Yuan, and Torsten Zesch (Eds.). Association for Computational Linguistics, Toronto, Canada, 247–259. https://doi.org/10.18653/v1/2023.bea-1.22

[5] Nischal Ashok Kumar and Andrew Lan. 2024. Improving Socratic Question Generation using Data Augmentation and Preference Optimization. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, Ekaterina Kochmar, Marie Bexte, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Anaïs Tack, Victoria Yaneva, and Zheng Yuan (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 108–118. https://aclanthology.org/2024.bea-1.10

[6] Marie-Josée Bisson, Walter J. B. Van Heuven, Kathy Conklin, and Richard J. Tunney. 2014. Processing of native and foreign language subtitles in films: An eye tracking study. *Applied Psycholinguistics* 35, 2 (2014), 399–418. https://doi.org/10.1017/S0142716412000434

[7] John Brooke. 1995. SUS: A quick and dirty usability scale. *Usability Eval. Ind.* 189 (11 1995).

[8] Géry Casiez, Nicolas Roussel, and Daniel Vogel. 2012. 1 € Filter: A Simple Speed-Based Low-Pass Filter for Noisy Input in Interactive Systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 2527–2530. https://doi.org/10.1145/2207676.2208639

[9] Rimika Chaudhury and Parmit K Chilana. 2024. Designing Visual and Interactive Self-Monitoring Interventions to Facilitate Learning: Insights from Informal Learners and Experts. *IEEE Transactions on Visualization and Computer Graphics* (2024).

[10] I-Jung Chen. 2016. Hypertext glosses for foreign language reading comprehension and vocabulary acquisition: effects of assessment methods. *Computer Assisted Language Learning* 29, 2 (2016), 413–426. https://doi.org/10.1080/09588221.2014.983935 arXiv:https://doi.org/10.1080/09588221.2014.983935

[11] I-Jung Chen and Jung-Chuan Yen. 2013. Hypertext annotation: Effects of presentation formats and learner proficiency on reading comprehension and vocabulary learning in foreign languages. *Computers & Education* 63 (2013), 416–423. https://doi.org/10.1016/j.compedu.2013.01.005

[12] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sashank Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. PaLM: scaling language modeling with pathways. *J. Mach. Learn. Res.* 24, 1, Article 240 (Jan. 2023), 113 pages.

[13] Charles Clifton, Fernanda Ferreira, John M. Henderson, Albrecht W. Inhoff, Simon P. Liversedge, Erik D. Reichle, and Elizabeth R. Schotter. 2016. Eye movements in reading and information processing: Keith Rayner's 40year legacy. *Journal of Memory and Language* 86 (2016), 1–19. https://doi.org/10.1016/j.jml.2015.07.004

[14] Leana Copeland, Tom Gedeon, and Balapuwaduge Mendis. 2014. Predicting reading comprehension scores from eye movements using artificial neural networks and fuzzy output error. *Artificial Intelligence Research* 3 (07 2014). https://doi.org/10.5430/air.v3n3p35

[15] Brendan David-John, Candace Peacock, Ting Zhang, T. Scott Murdison, Hrvoje Benko, and Tanya R. Jonker. 2021. Towards gaze-based prediction of the intent to interact in virtual reality. In *ACM Symposium on Eye Tracking Research and Applications* (Virtual Event, Germany) *(ETRA '21 Short Papers)*. Association for Computing Machinery, New York, NY, USA, Article 2, 7 pages. https://doi.org/10.1145/3448018.3458008

[16] Anca Dinu, Liviu Dinu, and Ionut Sorodoc. 2014. Aggregation methods for efficient collocation detection. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), Reykjavik, Iceland, 4041–4045. http://www.lrec-conf.org/proceedings/lrec2014/pdf/1184_Paper.pdf

[17] Géry d'Ydewalle and Wim De Bruycker. 2007. Eye movements of children and adults while reading television subtitles. *European psychologist* 12, 3 (2007), 196–205.

[18] Marco Porta Elisa Perego, Fabio Del Missier and Mauro Mosconi. 2010. The Cognitive Effectiveness of Subtitle Processing. *Media Psychology* 13, 3 (2010), 243–272. https://doi.org/10.1080/15213269.2010.502873 arXiv:https://doi.org/10.1080/15213269.2010.502873

[19] Isabeau Fievez, Maribel Montero Perez, Frederik Cornillie, and Piet Desmet. 2023. Promoting incidental vocabulary learning through watching a French Netflix series with glossed captions. *Computer Assisted Language Learning* 36, 1-2 (2023), 26–51. https://doi.org/10.1080/09588221.2021.1899244 arXiv:https://doi.org/10.1080/09588221.2021.1899244

[20] OpenJS Foundation. 2014. Electron – Build cross-platform desktop apps with JavaScript, HTML, and CSS. https://electronjs.org

[21] Katsuya Fujii and Jun Rekimoto. 2019. SubMe: An Interactive Subtitle System with English Skill Estimation Using Eye Tracking. In *Proceedings of the 10th Augmented Human International Conference 2019* (Reims, France) *(AH2019)*. Association for Computing Machinery, New York, NY, USA, Article 23, 9 pages. https://doi.org/10.1145/3311823.3311865

[22] P. Grant and D. Basye. 2014. *Personalized Learning: A Guide for Engaging Students with Technology*. International Society for Tech in Ed. https://books.google.ca/books?id=96apCgAAQBAJ

[23] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language Models (Mostly) Know What They Know. arXiv:2207.05221 [cs.CL]

[24] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, Article 1613, 15 pages.

[25] Olga Kolesnikova. 2016. Survey of word co-occurrence measures for collocation detection. *Computación y Sistemas* 20, 3 (2016), 327–344.

[26] Jan-Louis Kruger and Faans Steyn. 2013. Subtitles and Eye Tracking: Reading and Performance. *Reading Research Quarterly* 49 (10 2013). https://doi.org/10.1002/rrq.59

[27] Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education* 30 (2020), 121–204.

[28] Jean-Francois Lapointe, Heather Molyneaux, Irina Kondratova, and Aida Freixanet Viejo. 2016. Learning and Performance Support - Personalization Through Personal Assistant Technology. In *Learning and Collaboration Technologies*, Panayiotis Zaphiris and Andri Ioannou (Eds.). Springer International Publishing, Cham, 223–232.

[29] Kelly Morris Laurie O. Campbell, Tracey Planinz and Joshua Truitt. 2019. Investigating Undergraduate Students' Viewing Behaviors of Academic Video in Formal and Informal Settings. *College Teaching* 67, 4 (2019), 211–221. https://doi.org/10.1080/87567555.2019.1650703 arXiv:https://doi.org/10.1080/87567555.2019.1650703

[30] John Lee and Chak Yan Yeung. 2018. Personalizing Lexical Simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 224–232. https://aclanthology.org/C18-1019

[31] Joanne Leong, Pat Pataranutaporn, Valdemar Danry, Florian Perteneder, Yaoli Mao, and Pattie Maes. 2024. Putting Things into Context: Generative AI-Enabled Context Personalization for Vocabulary Learning Improves Learning Motivation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 677, 15 pages. https://doi.org/10.1145/3613904.3642393

[32] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput. Surv.* 55, 9, Article 195 (Jan. 2023), 35 pages. https://doi.org/10.1145/3560815

[33] Ruivaldo Lobão-Neto, Adrien Brilhault, Sergio Neuenschwander, and Ricardo Rios. 2022. Real-time identification of eye fixations and saccades using radial basis function networks and Markov chains. *Pattern Recognition Letters* 162 (2022), 63–70. https://doi.org/10.1016/j.patrec.2022.08.013

[34] Andrea Lofgren. 2022. Unraveling Contradictions: Which Glosses Facilitate Reading Comprehension Among ELLs, and Why? *Journal of Language Teaching and Research* (2022). https://api.semanticscholar.org/CorpusID:245661675

[35] Pascual Martínez-Gómez and Akiko Aizawa. 2014. Recognition of Understanding Level and Language Skill Using Measurements of Reading Behavior. In *Proceedings of the 19th International Conference on Intelligent User Interfaces* (Haifa, Israel) *(IUI '14)*. Association for Computing Machinery, New York, NY, USA, 95–104. https://doi.org/10.1145/2557500.2557546

[36] Meta. 2014. React – The library for web and native user interfaces. https://react.dev

[37] Maribel Montero Perez, Elke Peters, and Piet Desmet. 2015. Enhancing vocabulary learning through captioned Video: An eye-tracking study. *The Modern Language Journal* 99, 2 (2015), 308–328.

[38] Carmen Muñoz. 2017. The role of age and proficiency in subtitle reading. An eye-tracking study. *System* 67 (2017), 77–86. https://doi.org/10.1016/j.system.2017.04.015

[39] Shivsevak Negi and Ritayan Mitra. 2020. Fixation duration and the learning process: an eye tracking study with subtitled videos. *Journal of Eye Movement Research* 13 (2020). https://api.semanticscholar.org/CorpusID:229538626

[40] Kai North, Marcos Zampieri, and Matthew Shardlow. 2023. Lexical Complexity Prediction: An Overview. *Comput. Surveys* 55, 9 (jan 2023), 1–42. https://doi.org/10.1145/3557885

[41] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[42] OpenAI. 2023. OpenAI Node API Library. https://platform.openai.com

[43] Gustavo Paetzold and Lucia Specia. 2016. SemEval 2016 Task 11: Complex Word Identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, San Diego, California, 560–569. https://doi.org/10.18653/v1/S16-1085

[44] Liangming Pan, Wenqiang Lei, Tat-Seng Chua, and Min-Yen Kan. 2019. Recent Advances in Neural Question Generation. arXiv:1905.08949 [cs.CL]

[45] John F. Pane, Elizabeth D. Steiner, Matthew D. Baird, and Laura S. Hamilton. 2015. *Continued Progress: Promising Evidence on Personalized Learning*. RAND Corporation, Santa Monica, CA. https://doi.org/10.7249/RR1365

[46] Zhenhui Peng, Xingbo Wang, Qiushi Han, Junkai Zhu, Xiaojuan Ma, and Huamin Qu. 2023. Storyfier: Exploring Vocabulary Learning Support with Text Generation Models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. ACM. https://doi.org/10.1145/3586183.3606786

[47] Martin J Pickering, Steven Frisson, Brian McElree, and Matthew J Traxler. 2004. Eye movements and semantic composition. *On-line study of sentence comprehension: Eyetracking, ERPs and beyond* (2004), 33–50.

[48] Keith Rayner, Alexander Pollatsek, and Elizabeth R Schotter. 2012. Reading: word identification and eye movements. *Handbook of Psychology, Second Edition* 4 (2012).

[49] Laria Reynolds and Kyle McDonell. 2021. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 314, 7 pages. https://doi.org/10.1145/3411763.3451760

[50] Ana-Cristina Rogoz and Radu Tudor Ionescu. 2024. UnibucLLM: Harnessing LLMs for Automated Prediction of Item Difficulty and Response Time for Multiple-Choice Questions. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, Ekaterina Kochmar, Marie Bexte, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Anaïs Tack, Victoria Yaneva, and Zheng Yuan (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 493–502. https://aclanthology.org/2024.bea-1.41

[51] Alexander Scarlatos, Wanyong Feng, Andrew Lan, Simon Woodhead, and Digory Smith. 2024. Improving Automated Distractor Generation for Math Multiple-choice Questions with Overgenerate-and-rank. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, Ekaterina Kochmar, Marie Bexte, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Anaïs Tack, Victoria Yaneva, and Zheng Yuan (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 222–231. https://aclanthology.org/2024.bea-1.19

[52] Timo Schick and Hinrich Schütze. 2021. It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 2339–2352. https://doi.org/10.18653/v1/2021.naacl-main.185

[53] Matthew Shardlow, Michael Cooper, and Marcos Zampieri. 2020. CompLex — A New Corpus for Lexical Complexity Prediction from Likert Scale Data. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, Núria Gala and Rodrigo Wilkens (Eds.). European Language Resources Association, Marseille, France, 57–62. https://aclanthology.org/2020.readi-1.9/

[54] Patrycja Śliwiak and Syed Afaq Ali Shah. 2024. Text-to-text generative approach for enhanced complex word identification. *Neurocomputing* (2024), 128501.

[55] Yuni Susanti, Takenobu Tokunaga, Hitoshi Nishikawa, and Hiroyuki Obari. 2017. Evaluation of automatically generated English vocabulary questions. *Research and Practice in Technology Enhanced Learning* 12 (03 2017). https://doi.org/10.1186/s41039-017-0051-y

[56] Rohail Syed, Kevyn Collins-Thompson, Paul N. Bennett, Mengqiu Teng, Shane Williams, Dr. Wendy W. Tay, and Shamsi Iqbal. 2020. Improving Learning Outcomes with Gaze Tracking and Automatic Question Generation. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 1693–1703. https://doi.org/10.1145/3366423.3380240

[57] Anaïs Tack, Piet Desmet, Cédrick Fairon, and Thomas François. 2021-06-25. Mark My Words! On the Automated Prediction of Lexical Difficulty for Foreign Language Readers.

[58] Anaïs Tack, Thomas François, Anne-Laure Ligozat, and Cédrick Fairon. 2016. Modèles adaptatifs pour prédire automatiquement la compétence lexicale d'un apprenant de français langue étrangère (Adaptive models for automatically predicting the lexical competence of French as a foreign language learners). In *Actes de la conférence conjointe JEP-TALN-RECITAL 2016. volume 2 : TALN (Articles longs)*. AFCP - ATALA, Paris, France, 221–234. https://aclanthology.org/2016.jeptalnrecital-long.17

[59] Keren Tan, Kangyang Luo, Yunshi Lan, Zheng Yuan, and Jinlong Shu. 2024. An LLM-Enhanced Adversarial Editing System for Lexical Simplification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 1136–1146. https://aclanthology.org/2024.lrec-main.102/

[60] (Mark) Feng Teng. 2022. Vocabulary learning through videos: captions, advance-organizer strategy, and their combination. *Computer Assisted Language Learning* 35, 3 (2022), 518–550. https://doi.org/10.1080/09588221.2020.1720253 arXiv:https://doi.org/10.1080/09588221.2020.1720253

[61] Tom Tullis and Bill Albert. 2013. Chapter 7 - Behavioral and Physiological Metrics. In *Measuring the User Experience (Second Edition)* (second edition ed.), Tom Tullis and Bill Albert (Eds.). Morgan Kaufmann, Boston, 163–186. https://doi.org/10.1016/B978-0-12-415781-1.00007-8

[62] Kristiyan Vachev, Momchil Hardalov, Georgi Karadzhov, Georgi Georgiev, Ivan Koychev, and Preslav Nakov. 2022. Leaf: Multiple-Choice Question Generation. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part II* (Stavanger, Norway). Springer-Verlag, Berlin, Heidelberg, 321–328. https://doi.org/10.1007/978-3-030-99739-7_41

[63] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

[64] Feiyang Wang. 2021. *Implicit Gaze Interaction for Information Visualization*. Master's thesis. University of Ontario Institute of Technology.

[65] Zichao Wang and Richard Baraniuk. 2023. MultiQG-TI: Towards Question Generation from Multi-modal Sources. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*. Association for Computational Linguistics, Toronto, Canada, 682–691. https://aclanthology.org/2023.bea-1.55

[66] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International*

Conference on Neural Information Processing Systems (New Orleans, LA, USA) *(NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, Article 1800, 14 pages.

[67] Paula Winke, Susan Gass, and Tetyana Sydorenko. 2013. Factors influencing the use of captions by foreign language learners: An eye-tracking study. *The Modern language journal* 97, 1 (2013), 254–275.

[68] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 143–146. https://doi.org/10.1145/1978942.1978963

[69] Changrong Xiao, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Lei Xia. 2023. Evaluating Reading Comprehension Exercises Generated by LLMs: A Showcase of ChatGPT in Education Applications. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*. Association for Computational Linguistics, Toronto, Canada, 610–625. https://aclanthology.org/2023.bea-1.52

[70] Changrong Xiao, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Lei Xia. 2023. Evaluating Reading Comprehension Exercises Generated by LLMs: A Showcase of ChatGPT in Education Applications. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*. Association for Computational Linguistics, Toronto, Canada, 610–625. https://doi.org/10.18653/v1/2023.bea-1.52

[71] Akifumi Yanagisawa, Stuart Webb, and Takumi Uchihara. 2020. How do different forms of glossing contribute to L2 vocabulary learning from reading?: A meta-regression analysis. *Studies in Second Language Acquisition* 42, 2 (2020), 411–438. https://doi.org/10.1017/S0272263119000688

[72] Victoria Yaneva, Kai North, Peter Baldwin, Le An Ha, Saed Rezayi, Yiyun Zhou, Sagnik Ray Choudhury, Polina Harik, and Brian Clauser. 2024. Findings from the First Shared Task on Automated Prediction of Difficulty and Response Time for Multiple-Choice Questions. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, Ekaterina Kochmar, Marie Bexte, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Anaïs Tack, Victoria Yaneva, and Zheng Yuan (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 470–482. https://aclanthology.org/2024.bea-1.39

[73] Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A Report on the Complex Word Identification Shared Task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, Joel Tetreault, Jill Burstein, Ekaterina Kochmar, Claudia Leacock, and Helen Yannakoudakis (Eds.). Association for Computational Linguistics, New Orleans, Louisiana, 66–78. https://doi.org/10.18653/v1/W18-0507

[74] Kazuyo Yoshimura, Koichi Kise, and Kai Kunze. 2015. The eye as the window of the language ability: Estimation of English skills by analyzing eye movement while reading documents. In *13th IAPR International Conference on Document Analysis and Recognition, ICDAR 2015 - Conference Proceedings (Proceedings of the International Conference on Document Analysis and Recognition, ICDAR)*. IEEE Computer Society, 251–255. https://doi.org/10.1109/ICDAR.2015.7333762 Publisher Copyright: © 2015 IEEE.; 13th International Conference on Document Analysis and Recognition, ICDAR 2015 ; Conference date: 23-08-2015 Through 26-08-2015.

[75] George-Eduard Zaharia, Dumitru-Clementin Cercel, and Mihai Dascalu. 2020. Cross-Lingual Transfer Learning for Complex Word Identification. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*. 384–390. https://doi.org/10.1109/ICTAI50040.2020.00067

[76] Qing Zeng, Eunjung Kim, Jon Crowell, and Tony Tse. 2005. A Text Corpora-Based Estimation of the Familiarity of Health Terminology. In *Biological and Medical Data Analysis*, José Luís Oliveira, Víctor Maojo, Fernando Martín-Sánchez, and António Sousa Pereira (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 184–192.

[77] Zixin Zhao. 2023. *Gloss Positioning for a Gaze Aware L2 Reading Aid.* Master's thesis. University of Ontario Institute of Technology.

[78] Yueyuan Zheng, Xinchen Ye, and Janet Hsiao. 2019. Does Video Content Facilitate or Impair Comprehension of Documentaries? The Effect of Cognitive Abilities and Eye Movement Strategy. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

[79] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large Language Models are Human-Level Prompt Engineers. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=92gvk82DE-

A Appendix

A.1 Collocation Detection

The pseudocode (Algorithm 1) for detecting collocation/phrases for a subtitle as described in subsection 3.2.

Algorithm 1 Collocation Detection Algorithm

```
function GetPhrases(subtitle)
    tokens ← Tokenize(subtitle)
    phrases ← [ ]
    n ← length(tokens)
    for i = 0 to n − 1 do
        phrase ← tokens[i]
        if ¬IsStopword(tokens[i]) then
            for j = i + 1 to n do
                phrase ← Concatenate(phrase, tokens[j])
                if ¬IsStopword(tokens[j]) then
                    if IsValidPhrase(phrase) then        ▷ Ask ChatGPT if token sequence is a valid phrase
                        Append(phrases, phrase)
                    else
                        break
                    end if
                end if
            end for
        end if
        i ← j − 1
    end for
    return phrases
end function
```

A.2 ChatGPT Prompts

Our proposed methods involved using ChatGPT, an LLM, due to its multi-language support as it is trained on publicly available data (e.g. internet data). You can condition ChatGPT by inputting instructions describing the task to solve various tasks, providing the LLM examples (few-shot learning) or without examples (zero-shot learning) [32, 49, 52]. Chain of thought (CoT) prompting [66] proposed a method by modifying the examples to step-by-step answers and achieved higher performance across difficult benchmarks [12]. A zero-shot approach, zero-shot-CoT, eliminates the need to hand-craft few-shot examples per task while extracting step-by-step reasoning by simply adding "Let's think step by step" to the end of the prompt [24]. Automatic instruction generation has improved the solution to "Let's work this out in a step by step way to be sure we have the right answer." [79]. We will use zero-shot-CoT prompts to maximize performance when implementing AQG and CWI. After each response, ChatGPT is asked to parse the response in JSON.

A.2.1 Complex Word Identification. The prompt for CWI (Table 2) will have ChatGPT describe the absolute and relative complexity of the target word [40], and based on its analysis, it will score the target word 1–5. The target audience is also needed for personalize complexity [30, 57, 58, 76].

Table 2: ChatGPT prompts for complex word identification. A checklist is used to guide ChatGPT in scoring the complexity of the target word.

Prompt Role	Prompt
System	You are an expert on NLP. You analyze the word's complexity using a scale of 1 to 5, with 1 being the least complex and 5 being the most complex. Here is some information to analyze the word's complexity: 1. Words having multiple meanings are more complex. 2. The word's higher cognitive load or demand is more complex. 3. Higher acquisition difficulty of the word is more complex. 4. Rarer words are more complex. Consider that the person reading this word is [*enter target audience*].
User	Word: [*enter word*]
Assistance	Let's work this out in a step by step way to be sure we have the right answer.

A.2.2 Collocation Detection. For each subtitle, each subtitle is processed as described in subsection 3.2 using ChatGPT (Table 3). Once the token sequence fails, the previous successful token sequence will be displayed back to the user via gloss.

Table 3: ChatGPT prompts for collocation detection using true/false format.

Prompt Role	Prompt
System	You are a language expert. Check if when combining two terms forms a phrase. If so, provide one-sentence definition for each term in the given context and the whole phrase so that a 12 year old can understand. Also, you must provide example sentences using the phrase.
User	"[*enter token sequence*]" is a phrase (Context: [*enter context*]) A) True B) False

A.2.3 Question Generation. The prompt has a checklist to ensure the questions are high quality and avoid simple questions (Table 4). It also gives a target word with the maximum score within the subtitle to guide the topic of the question. After the assistant gives the question, the prompt will ask for feedback for each option on why it is incorrect or correct.

Table 4: ChatGPT prompts for question generation. A checklist ensures the questions generated will be of consistent quality.

Prompt Role	Prompt
System	You are a professor making a multiple-choice test about a video. Describe your steps first.
User	Create an advanced multiple-choice question about the video given with four choices. Give the correct answer at the end of the question. Here are the criteria for the question: 1. The question must have the word: "[*enter word*]". 2. All choices should explain a concept or an idea in a sentence about 15 words long without giving away the answer. 3. All incorrect choices must be from the video and related to the correct choice. 4. All choices should have a similar number of words.
User	Video: [*enter subtitle text*]
Assistant	Let's work this out in a step by step way to be sure we have the right question that fits the criteria.
User	Explain in one sentence why option [*option letter*] is [*incorrect/correct*]. Do not use words in any of the choices. Output the explanation.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009