الجـمهورية الجـزائرية الديمقراطـية الشـعبية
PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
وزارة التعليم العـالي و البحـث العـلمي
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH
جـامعة عمار ثـليجي بالأغـواط
UNIVERSITY OF Ammar Telidji LAGHOUAT

كـلية العلـوم
FACULTY OF SCIENCES
قـسـم الإعـلام الآلـي
DEPARTMENT OF COMPUTER SCIENCE

Master Thesis

Field : Mathematics and Computer Science

Option : Computer Science

Specialization : Decision Support Systems

Submitted by :

DELASSI Khaled Bachir

ZEGGANE Lakhdar

Theme

Visual Question Answering
Support to Arabic Pedagogical Tool

Jury members :

Pr.	Ouinten Youcef	Prof.	(University of Laghouat)	President
Dr.	Bensaad Lahcen	MCB	(University of Laghouat)	Examiner
Dr.	Belabbaci Amel	MCB	(University of Laghouat)	Examiner
Pr.	Hadda Cherroun	Prof.	(University of Laghouat)	Advisor

Academic Year 2023/2024

This work is dedicated to our beloved families, whose unwavering support and encouragement have been a guiding light throughout this journey. Their belief in us has been a source of strength and inspiration. To our dear friends, thank you for standing by us, cheering us on, and lifting our spirits during challenging times. Your friendship means the world to us.

We extend our heartfelt gratitude to **Ammar Telidji University** for providing us with the knowledge, resources, and opportunities that have shaped our academic journey and contributed to our growth.

Lastly, we express our gratitude to ourselves for the dedication, perseverance, and passion we have poured into this endeavor. This accomplishment is a testament to our hard work and commitment to excellence.

This work is dedicated to all of you who have been part of our story. Thank you for believing in us.

Acknowledgement

We would like to express our deepest gratitude to everyone who contributed to the completion of this work.

First and foremost, we are profoundly grateful to our research advisor, **Professor Hadda Cherroun**. Her unwavering guidance, invaluable advice, and constant support have been the pillars of this project. Her deep knowledge and enthusiasm for the subject have not only enlightened us but have also been an endless source of inspiration and motivation.

We also wish to extend our heartfelt thanks to our dear colleagues and friends, **Djamel Eddine BRIK** and **Abdelnour AOUISSI**. Your help, stimulating discussions, and moral support have been indispensable. Your camaraderie has made this journey not only bearable but truly enjoyable, transforming challenges into opportunities for growth.

A special thanks to all members of the **DEPARTMENT OF COMPUTER SCIENCE** for providing us with the necessary resources and a nurturing environment to accomplish this research. Your collective effort and support have been fundamental to our success.

We are profoundly grateful to our families for their endless love, patience, and unwavering support. Your constant encouragement has been the driving force that kept us motivated and focused, even during the most challenging times. Your belief in us has been our beacon of hope.

Finally, we would like to express our sincere appreciation to everyone who, directly or indirectly, contributed to the completion of this work. Your support, encouragement, and belief in us have been invaluable. Thank you for being part of this journey and for helping us reach this milestone.

Abstract

The Arabic language serves as a highly valuable strategic global language due to its widespread use. However, learning Arabic as a second language is quite a challenging task due to the specificity of Arabic, which encompasses many major challenges: Semitic language with complex grammar, high dialectal variation, rich phonetic specter within a vast vocabulary that includes many words with similar meanings. Faced to the high scarcity of dedicated pedagogical tools, we address this problem by designing an end-to-end web-based pedagogical tool leveraged by Deep Learning.

In fact, our convivial and ergonomic web-based tool, is based on constructivism learning model where the learning process is made through active learning. In our tool, we offer a high quality pedagogical activities for Arabic Learners such as real-life vision quizzes, interactive image-based questions, and language learning activities. These activities are designed to enhance the learning experience by integrating visual and language processing tasks, providing a comprehensive approach to language education. Within the tool, we deployed and harnessed at least two recent AI deep learning based models:

- Text-To-Text Transfer Transformer which is a large language model. This later is the basis for translation and Vision Question Generation
- Vision Language Pre-training which is a large-scale generalized model that has been trained on multiple tasks using a large volume of data and then fine-tuned on tasks like Vision question answering.

This Model is deployed for Visual Question Answering. In addition, we have designed and implemented the tool front-end while relaying on SOTA technologies React.js, React Query, Tailwind CSS,

In order to measure the performance of our AI-based pedagogical tool, we conducted a human evaluation using 150 vision quizzes. The results show that our tool is very suitable for Arabic Learning with an accuracy of 80%.

Keywords : Pedagogical Tool, Constructivism, Deep Learning, Visual Question Answering, Visual Language Pretraining, Transformers.

ملـخـــص

تُعد اللغة العربية لغة استراتيجية عالمية ذات قيمة عالية نظراً لاستخدامها الواسع. ومع ذلك، فإن تعلم اللغة العربية كلغة ثانية يُعد مهمة صعبة بعض الشيء بسبب خصوصية اللغة العربية التي تتضمن العديد من التحديات الكبرى: كونها لغة سامية ذات قواعد نحوية معقدة، تنوع لهجات كبير، غنى صوتي ضمن مفردات واسعة تشمل العديد من الكلمات ذات المعاني المتشابهة.

في مواجهة النقص الكبير في الأدوات التعليمية المخصصة، نحن نعالج هذه المشكلة من خلال تصميم أداة تعليمية عبر الويب شاملة تعتمد على التعلم العميق.

في الواقع، تعتمد أداتنا عبر الويب، السهلة الاستخدام والمريحة، على نموذج التعلم البُنيَوِي حيث يتم التعلم من خلال التعلم النِشط. في أداتنا، نقدم أنشطة تعليمية عالية الجودة للمتعلمين باللغة العربية مثل الاختبارات البصرية الواقعية، الأسئلة التفاعلية المُعتمدة على الصور، وأنشطة تعلم اللغة. تم تصميم هذه الأنشطة لتعزيز تجربة التعلم من خلال دمج المهام البصرية ومعالجة اللغة، مما يوفر نهجاً شاملاً لتعليم اللغة . ضمن الأداة، نشرنا واستفدنا على الأقل من نموذجين حديثين يعتمدان على التعلم العميق:

- Text-To-Text Transfer Transformer وهو نموذج لغوي كبير. هذا الأخير هو الأساس للترجمة و Vision Question Generation
- Vision Language Pre-training وهو نموذج عام واسع النطاق تم تدريبه على مهام متعددة باستخدام حجم كبير من البيانات ثم تم تحسينه على مهام مثل Visual Question Answering. هذا النموذج يستخدم للإجابة على الأسئلة البصرية.

بالإضافة إلى ذلك، قمنا بتصميم وتنفيذ الواجهة الأمامية للأداة باستخدام أحدث التقنيات مثل React.js, React Query, Tailwind CSS.

من أجل قياس أداء أداة التعليم القائمة على الذكاء الاصطناعي، قمنا بإجراء تقييم بشري لأداء الأداة باستخدام 150 اختبار رؤية. أظهرت النتائج أن أداتنا مناسبة جداً لتعلم اللغة العربية بدقة تصل إلى 80٪.

الكلمات المفتاحية: أداة تربوية، Constructivism, Deep Learning, Visual Question Answering, Visual Language Pretraining, Transformers.

Résumé

La langue arabe est une langue stratégique mondiale très précieuse en raison de son utilisation répandue. Cependant, apprendre l'arabe comme une seconde langue est une tâche un peu difficile en raison des spécificités de l'arabe qui comprend beaucoup de grands défis : une langue sémitique avec une grammaire complexe, une grande variation dialectale, un spectre phonétique riche avec un vocabulaire vaste incluant de nombreux mots avec des significations similaires.

Face à la grande rareté des outils pédagogiques qui lui sont dédiés, nous abordons ce problème en concevant un outil pédagogique complet basé sur le Web et soutenu par le Deep Learning.

En fait, notre outil Web convivial et ergonomique est basé sur le modèle d'apprentissage constructiviste où le processus d'apprentissage se fait par l'apprentissage actif. Dans notre outil, nous offrons des activités pédagogiques de haute qualité pour les apprenants de l'arabe, telles que des Quiz visuels réels, des questions interactives basées sur des images, et des activités d'apprentissage de la langue. Ces activités sont conçues pour améliorer l'expérience d'apprentissage en intégrant des tâches de traitement visuel et de langage, offrant une approche globale de l'éducation linguistique.

Dans la conception et la réalisation de l'outil, nous avons déployé et exploité au moins deux modèles récents basés sur le Deep Learning :

- Text-To-Text Transfer Transformer qui est un grand modèle de langue. Ce dernier est la base pour la traduction et Vision Question Generation
- Vision Language Pre-training qui est un modèle généralisé à grande échelle qui a été entraîné sur plusieurs tâches utilisant un grand volume de données, puis affiné sur des tâches comme Visual Question Answering.

Ce modèle est utilisé pour Visual Question Answering.

De plus, nous avons conçu et mis en œuvre l'interface avant de l'outil en utilisant les technologies les plus récentes comme React.js, React Query, Tailwind CSS. Afin de mesurer les performances de notre outil pédagogique basé sur l'IA, nous avons effectué une évaluation humaine en utilisant 150 quiz de vision. Les résultats montrent que notre outil est très adapté pour l'apprentissage de l'arabe avec une précision de 80%.

Mots-clés : Outil pédagogique, Constructivisme, Deep Learning, Visual Question Answering, Pré-entraînement de langage visuel, Transformers.

Contents

List of Figures

List of Tables

Introduction

Context

Language learning is essential and important now more than at any other time in the world since globalization started. It not only facilitates communication across cultures but also opens up numerous educational and professional opportunities. To be proficient in a language, you must develop your listening [2] [3], writing [4] [5], reading [6], and speaking [7] skills. This is why the use of pedagogical approaches and tools is crucial to enrich the learning process for a language learner. These approaches can range from Traditional Pedagogy (Lecture-Based Teaching, Rote Learning, Didactic Instruction) to modern Pedagogy (Blended Learning, Flipped Classroom, Problem-Based Learning (PBL), Game-Based Learning, etc.) [8] using traditional text-books or modern digital applications. This is especially true because of the advances made in the various fields of AI and as a result, many institutions are striving to incorporate AI into educational tools to deliver tailor-made adaptive learning techniques. This opens up a whole new world of possibilities for language learning, as technology progresses and presents more methods to incorporate into actual learning.

Problematic

Arabic is one of the most used languages in the world and it is classed number fifth as the most used language in the world (Arabic is spoken by 274 million people in the world) [9]. Despite the fact that Arabic is one of the most significant languages culturally and historically, it is considered as a very low-ressourced language (materials, documents, and sources) which makes its learning harder compared to English, Mandarin, Spanish, and Hindi. For this reasons, it is important to develop new and effective applications of artificial intelligence education for Arabic language learners and ensure they get an efficient resource to learn from.
However, current AI technology lacks the number of good AI-based applications used to

learn languages effectively. Most existing applications are either not AI-driven or cater to very basic levels of language proficiency1.1. Additionally, many of these applications are primarily focused on widely spoken languages like English and Chinese, leaving less commonly taught languages, such as Arabic, underrepresented. The rarity of Arabic language learning tools, particularly those leveraging AI, is a significant issue. This means that currently there is a need for advanced and intelligent learning tools that can support learners in a more complex way than has been possible in the past. Additionally, the current tools do not have flexibility in handling the requirements of language learning especially in languages in which resources are scarce.

Objectives and Contribution

Our research is part of the sustainable project *Smart-Arabic-Learning*, which aims to harness recent advancements in Deep Learning models and Active Learning paradigms. These enhancements are designed to improve pedagogical tools that facilitate Arabic learning for non-native learners through the application and deployment of advanced AI technologies

The Project functionalities are illustrated in Figure. 0.1. Its grantees are essentially Arabic Language Learners.

The goals of the project are to monitor auto-training through high quality pedagogical activities, profil analysis, the main ones are :

- Creation of many learning contexts from real world situation,
- improving the learning outcomes,
- Providing learner by customized learning context,
- Providing remediation processes,
- Providing learning feedbacks,
- Deployment of optimized Deep Learning Models that can boost the nature and the diversities of the offered Pedagogical Activities.

Our contribution in this substantial project has many goals . In a nutshell, the following are the objectives of the work :

- To explore Deep Learning approaches and models to use them in our pedagogical tool, in order to offer **varied, several and unlimited** questions.

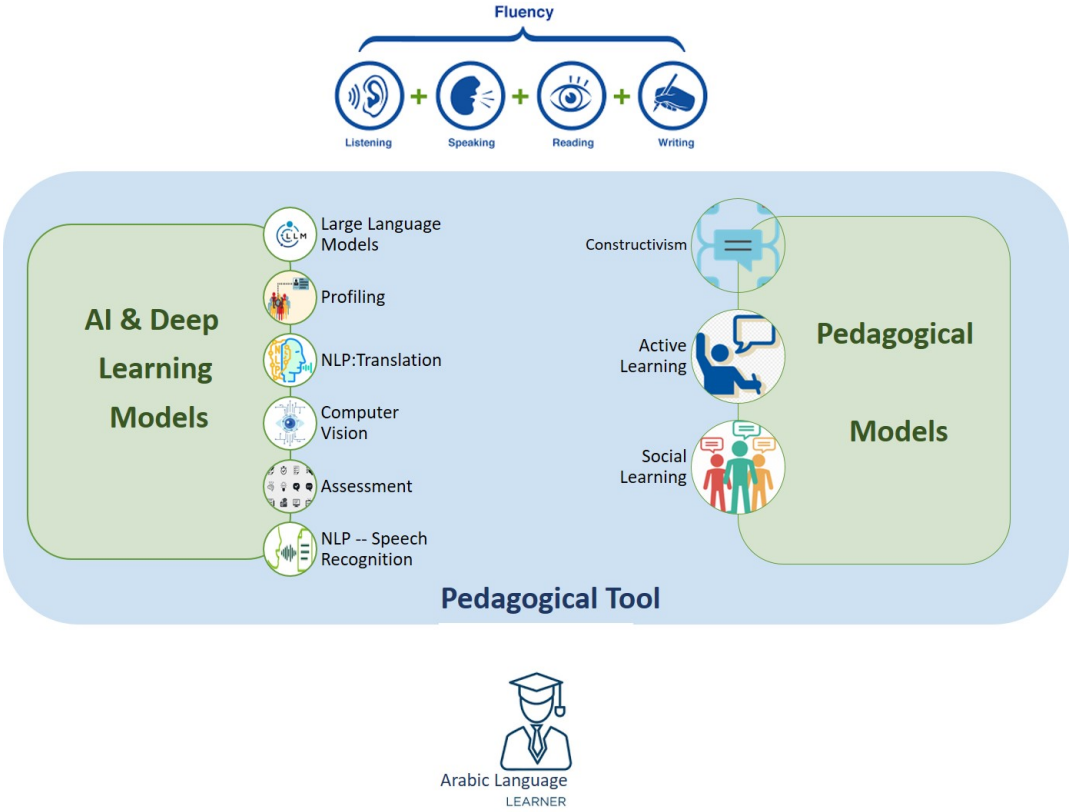

Figure 0.1: Smart Pedagogical Tool Project

- To design and develop a Web application that can be used in learning the Arabic language by assisting students in attaining a variety of learning goals, such as improved grammar, pronunciation, and vocabulary.
- To evaluate our pedagogical tool and show its effectiveness in helping students learn the Arabic language.

Theses Structure

This thesis is structured as follows :

- **Chapter 1** covers generalities, including how to learn a language by mastering the four skills : speaking, writing, listening, and reading methods and comparing language learning tools. Then we discuss natural language processing (NLP) by understanding NLP and the best ways to use NLP in education, along with an overview of Visual Question Answering (VQA) and Visual Question Generation (VQG).

- **Chapter 2** covers related work, beginning with traditional approaches to Visual Question Answering (VQA) including featurization phases, joint processing, and answer generation. It then delves into contemporary methodologies such as attention mechanisms, transformers, and Vision Language Pretraining (VLP), followed by an overview of relevant datasets. Next, we analyze various architectures for Visual Question Answering (VQA) and Visual Question Generation (VQG), focusing on their application in pedagogical tools.

- **Chapter 3** outlines our contributions, including an overview of our approach, the proposed ARABIC-EDU-VQA architecture, VQG and VQA systems, and system diagrams such as use case, class, and sequence diagrams.

- **Chapter 4**, our focus shifts to web app development, where we elaborate on the environment and tools utilized. We provide a detailed exploration of the architecture of the Visual Question Answering Web App, including its front-end components such as sign-up, log-in, dashboard, ImageIQ, quiz, and user profile. Additionally, we discuss the back-end services, AI model integration, and evaluation of our pedagogical tool.

Chapitre 1

Generalities

Contents

Learning a language is a multifaceted journey that involves various skills and strategies. This chapter will delve into the complexities of language acquisition, discussing

the fundamental components and approaches that contribute to successful language learning. We will introduce key concepts such as Natural Language Processing (NLP) and Visual Question Answering (VQA), setting the stage for further exploration in subsequent sections.

1.1 How to Learn a Language ?

Learning a language is a transformative journey that opens doors to new cultures, opportunities, and connections. In this introduction, we embark on an exploration of the art and science of language acquisition.

1.1.1 Setting Your Language Learning Goals

It is important to set concrete and realistic language learning objectives to support learners in their language acquisition process. This process involves several key steps :

- **Defining Clear Objectives**

  Establishing goals will help enhance learner direction when acquiring a new language. It is a strategic approach to comprehend early on what learners aim to achieve with their language proficiency. Objectives should be SMART : Specific, Measurable, Achievable, Relevant, and Time-bound (see Figure 1.1). For example, saying "I want to learn Spanish" lacks specificity compared to "I aim to achieve fluency in spoken Spanish within six months." [10].

Figure 1.1: SMART Goals Image [1]

[1] https://www.recbound.com/recruitment-agency-blog/example-smart-goals-for-recruiters

- **Long-term vs Short-term Goals**

  There is a need to determine whether the language learning goals are long-term or short-term. General goals are the final outcomes that a student may hope to achieve, such as reaching proficiency in language use and passing a proficiency test. In this context, short-term goals are smaller, more easily achievable and lead to the achievement of long-term goals. For example, a one-week goal could be to learn 50 new vocabulary words in a week or to read a chapter of a language textbook in one month, until it becomes a long-term goal like learning 1500 new vocabulary in 1-month [11].

- **Creating a Realistic Study Plan**

  After establishing goals in linguistics, the next step involves developing a study plan to achieve these objectives. The study plan outlines the systematic steps required to achieve the learning targets, including determining the daily or weekly study hours, selecting appropriate materials and resources, and employing effective learning strategies. When formulating the study plan, one should consider factors such as the current proficiency level, available study time, preferred learning modalities, and external commitments that may impact study schedules.

1.1.2 Building a Strong Foundation

Establishing a solid foundation is essential for success in language learning. This foundation serves as the groundwork upon which more advanced language skills can be developed. Several key components contribute to building this strong foundation :

- **Mastering Basic Vocabulary**

  Another very important facet in the development of language learning skills is the mastery of vocabulary. These are words or phrases that are primary units of communication. Special attention : learning common nouns, verbs, adjectives, and adverbs used in everyday contexts. Let it start with the fundamental words associated with the greets, the numbers, the colors, the family members, the food, and the activities of the day [12].

- **Learning Essential Grammar Structures**

  Understanding essential grammar structures is fundamental to constructing meaningful sentences and expressing one's thoughts accurately in the target language. Begin by familiarizing oneself with basic grammar concepts such as sentence structure Conjugation, noun-adjective agreement, and word order are essential aspects.

Enhance proficiency with these structures through exercises, drills, and example sentences. As proficiency develops, gradually broaden understanding of more intricate grammar rules and sentence patterns, prioritizing mastery of foundational structures initially [13].

- **Practicing Pronunciation and Intonation**

  Effective communication involves correct pronunciation and appropriate intonation. Dedicate time to refining pronunciation skills. Engage with native speakers, emulate their speech, and observe connections between sounds, stress patterns, and language rhythm. Participate in exercises like shadowing, dialogue imitation, and tongue twisters to develop accurate pronunciation consistent with your accent. Enhance communication proficiency through focused accent practice and intonation training [3].

1.1.3 Developing Listening Skills

Listening skills are crucial for effective communication and language proficiency. Here are some strategies to enhance one's ability to understand spoken language :

- **Active Listening Exercises**

  Active listening exercises use a wide variety of listening activities during which students are actively involved and focused on listening tasks. These exercises often involve listening to audio recordings or live interactions and actively responding or reflecting on the content. Examples of active listening exercises include listening to dialogues and summarizing key points, transcribing spoken sentences, shadowing native speakers to mimic pronunciation and intonation, and participating in language exchange sessions or conversation practice with peers [2].

- **Utilizing Podcasts and Audio-books**

  Podcasts and audiobooks are valuable resources for improving listening skills in a foreign language. Podcast helps learners to turn to different authentic spoken texts such as interviews, discussions, speeches, stories, and news. Audio books offer longer-form content that can help learners practice sustained listening comprehension [14].

- **Watching Movies and TV Shows with Subtitles**

  Watching movies and TV shows with subtitles is a popular method for improving listening skills while also reinforcing vocabulary and comprehension. Learners can start by watching content in their target language with subtitles in their native

language to aid understanding. As their listening proficiency improves, they can gradually switch to subtitles in the target language or even watch without subtitles to challenge themselves further. Additionally, learners can use subtitles as a tool for vocabulary acquisition by pausing the video to look up unfamiliar words or phrases and practicing pronunciation by repeating dialogue aloud [15].

1.1.4 Improving Reading Comprehension

The ability to read is an essential language skill. Here are some strategies to enhance one's ability to understand written texts :

- **Reading Techniques**

  Engage in active reading techniques to interact with the text actively. These include scanning the text beforehand to get an idea of what is coming, paraphrasing while reading to comprehend what the text has, or summing up the ideas captured from the text post reading [6].

- **Vocabulary Expansion**

  Enhance one's vocabulary to improve reading comprehension. Develop dictionary skills to determine the meanings of unfamiliar words and maintain a vocabulary resource for newly learned words. Explore vocabulary-building tools such as flashcards or language learning applications to strengthen word knowledge [16].

- **Reading Aloud**

  Practice reading a text aloud for better fluency and understanding. Reading aloud compels individuals to pay closer attention to the text and helps reinforce pronunciation and intonation patterns. It can also aid in the retention of information by engaging multiple senses simultaneously. [17].

1.1.5 Enhancing Writing Proficiency

Writing proficiency in a foreign language can be significantly improved through various strategies like :

- **Daily Writing Practice**

  Consistent practice is key to enhancing writing skills. Allocate time each day to write in the target language. It is advisable to begin with only a few paragraphs of writing followed by slightly larger segments and proceed with progressively longer

assignments. Engage also enables learners to memorize new words, grammatical constructs, and structure sentences that eventually boost learners' fluency levels [18].

- **Feedback and Revision**

  Receive feedback on your writing from teachers, language exchange partners, or online communities. Constructive criticism helps identify strengths and weaknesses and also underscores the importance of correcting mistakes for personal improvement. Incorporate feedback into your revisions to further refine your writing skills [4] [5].

- **Diverse Writing Tasks**

  It is important to engage in a wide range of writing assignments to further improve different aspects of writing. Begin writing different kinds of texts, such as essays, e-mails, letters, stories, or summaries. Each genre of writing possesses unique vocabulary patterns, structures, and conventions, providing essential foundations for studying and writing in diverse settings [19].

1.1.6 Tech-Driven Language Learning

Nowadays technology offers and facilitates learning languages. this is achieved by using :

- **Utilizing Language Learning tools**

  Language learning device applications are one of the most used ways of practicing the subject in the comfort of the learners' homes and offices. The focus of these tools is generally on vocabulary review and games, grammar exercises and practice, speaking practice, and interactive lessons. Sometimes they use gamification in the form of incentives and monitoring measures that help to make sure that the learners are involved and challenged. There are lots of popular language learning applications that include : **Duolingo, Babble, Rosetta Stone,** and **Memrise**. They are easy to implement and use, and with proper selection of applications, you can create a custom curriculum that can be used as a compensatory language program and which can be personalized according to the training style of the learner.

- **Exploring Online Resources and Courses**

  Internet is filled with many websites, blogs, forums, and online courses to assist learners. Whether one wants to learn the short or long course, they can check in portals like Coursera, Udemy, and edX, which offer to learn the desired language via their specialists from all over the world. Moreover, numerous tools are available on

the Internet that are specifically designed to assist individuals in language learning including **FluentU** and **iTalki**, which offer interactive videos and audio records, language exchange, and personal tutors.

- **Incorporating Language Learning Software**

  Language learning software provides comprehensive tools and resources for improving language skills. These software programs often feature interactive lessons, multimedia content, speech recognition technology, and personalized learning plans. Examples include **Rosetta Stone, Pimsleur, Fluenz**, and **Transparent Language**.

Finally, one can say that technology offers a number of opportunities that may be useful for learners that learn a new language. This can leverage technology regarding language learning tools that make learning easier, more interactive to the learner as well as more effective. Technology has been successfully used to start learning another language.

1.1.7 A Glance of Learning Language Applications

In today's globalized world, communication in several languages has become a valuable feature. In light of the increasing demand, there are currently a variety of applications and websites that offer various approaches and learning tools for a second or third language These digital platforms allow learners the convenience of learning at their own time and convenience, offering them a great variety of languages as well as usage of interactive and engaging methods of learning. Table 1.1 presents a comparative view of some of the most popular language learning applications and websites. In this comparative study, we have addressed many criteria :

- *Date of the first release :* Examines when the application was initially launched.
- *The based technique of teaching language :* In this criterion, we focus on whether the application is based on AI, the subject of our investigation topic.
- *The targeted language levels :* in fact, the language learning process (or goals) is generally divided into many levels such as beginners and advanced. The most commonly used levels in language learning range from 1 to 5.
- *The Cost :* whether the application is free, paid, or free with paid tools.
- In the *Methods* column, we provide more details on the pedagogical activities offered to the learner and the deployed learning techniques.

- the last column is dedicated to *the covered language* by the application.

Tool	Description	First Release	AI based	Levels	Cost	Methods	Supported Languages
Duolingo	A language-learning platform that offers interactive exercises and lessons to help people learn new languages for free.	was officially launched on June 19, 2012	Yes - Continually adjusting lesson difficulty to match individual progress	Levels 1-5	-Free -Super Duolingo is $13 per month or $84 annually	- Translate with word puzzle - Transcribe audio in foreign language - Transcribe audio in native language - Pronounce word - Multiple Choice	There are 43 languages including arabic
Kaleela	is an arabic app for any non-native speaker or writer to learn arabic	was officially launched on 2021	no - Continually simple quiz	Levels A1-C2	-Free	- Pronounce word to answering- Multiple Choice	There is only one language to learn "Arabic".
Rosetta Stone	A language-learning software that uses immersive methods to teach users new languages through visual and auditory cues, without translation.	in 1992 with the first version of its CD-ROM language learning software	Yes - TruAccent speech recognition technology.	Levels 1-3	Paid lifetime subscrip-tion $199	Immersion-Based Learning Method	There are 24 languages including arabic
Pimsleur	An audio-based language learning method that focuses on teaching language through listening and repeating exercises, gradually increasing in complexity.	The Pimsleur app in 2018	Yes - new AI-based pronunciation feedback	Levels 1-3	Paid $20 a month	an audio based course that presents phrases in the target language first, and then in your mother tongue for you to translate into that language.	There are 59 languages including arabic
Busuu	A language-learning app that offers interactive lessons, exercises, and opportunities to practice with native speakers through a social networking feature.	in 2008 by N. Bernhard and A. Hilti.	Yes - AI algorithms to personalize lesson plans and review exercises.	Levels 1-4	- Free - Premium From $3,49 per level $14,99 full course	- Revision - Word puzzle - Word matching - Multiple Choice - Dialogues - Writing texts	There are 14 languages including arabic

Table 1.1: Overview of tools specialized in teaching languages

1.2 Natural Language Processing

The concept of humans and machines working together smoothly has become a common phenomenon. Ever wondered how virtual assistants like Siri, Cortana, or Bixby decipher commands effortlessly ? Or marveled at how spell checkers detect errors that might have been missed ? The reason behind these marvels is an important field called Natural Language Processing (NLP), which bridges human language and intelligent machines. [20].

1.2.1 Understanding NLP

NLP delves into the realms of understanding, manipulating, and generating natural language by machines, positioning itself at the intriguing intersection of computer science and linguistics. At its core, NLP empowers machines to directly engage with humans, enabling a wide array of applications across various domains [20].

1.2.2 Natural Language Processing Applications

There are different methods of utilizing NLP for working with text data. The fact that there are not many systems using NLP is just downright weird considering that text is everywhere. Here are some ways NLP is used :

- **Information Retrieval :** This means finding information in the text. Some systems, like those by Liddy and Strzalkowski, use NLP to do this better [21].
- **Information Extraction (IE) :** This involves finding and tagging important information in text, like names of people or places. This information can be used for different things, like answering questions or analyzing data [21].
- **Question-Answering :** Instead of just showing documents that might have answers, NLP can directly give answers or parts of answers to questions [21].
- **Summarization :** NLP can take big pieces of text and make shorter summaries that still capture the main ideas [21].
- **Machine Translation :** NLP has been used for a long time in systems that translate languages. Some systems look at individual words, while others analyze text more deeply [21].
- **Dialogue Systems :** These are like talking to a computer, perhaps in household appliances. Right now, they focus on basic language stuff, but using all levels of language processing could make them even better in the future [21].

- **E-learning & Distant learning :** NLP is by excellence a powerful support to learning and education. By leveraging natural language processing, educational tools can be developed to enhance comprehension, personalize learning experiences, and provide real-time feedback [21]. More details are in the next section.

1.2.3 What are the most effective ways to use NLP in education

NLP has the power to change some of the important aspects of education. Here are some of the most effective ways to use NLP in education :

- **Adaptive Learning**

  This learning way is used to create personalized learning systems that can tailor the content of lectures, their duration, and level of difficulty based on the response of students to the previous material. For instance, NLP can assist in generating tailored exercises and quizzes, providing immediate feedback, and recommending relevant resources like videos and articles [22].

- **Chat-bots and Virtual Assistants**

  Chatbots and assistants are NLP tools that engage students and teachers in natural language. These tools offer support, guidance, and motivation, answering questions, providing reminders, and assisting with administrative tasks. By employing NLP-driven chat-bots, education becomes more interactive, with real-time assistance and feedback for both students and teachers [22].

- **Text Analysis and Summarization**

  NLP can also be used to analyze and summarize large volumes of text, helping students and teachers extract key insights efficiently. Text analysis tools identify themes, topics, and sentiments from diverse sources while summarizing tools condense texts into concise summaries. These tools enhance reading and writing skills, improve comprehension, and save time by presenting essential information in a clear and coherent manner [22].

- **Natural Language Generation**

  In education, natural language generation tools enable the creation of original and engaging texts using natural language. By utilizing data, keywords, images, or audio inputs, these tools produce relevant and creative texts such as captions, summaries, and stories. They aid in improving vocabulary, grammar, and expression, encouraging imagination and enhancing communication skills for both students and teachers [22].

1.2.4 Evaluating Language Models in NLP

Machine learning pertains to the development of algorithms designed to derive valuable insights. It places a strong emphasis on persistent application within constantly evolving scenarios, highlighting the importance of adapting, retraining, and refining algorithms based on previous experiences [23]. A language model is a type of machine learning model trained to predict subsequent words or characters in texts. Language models are pivotal in natural language processing (NLP) [24]. Their performance is measured through different metrics such as perplexity, and cross entropy. This approach represents the comparison of models and evaluates the models' effectiveness in performing NLP operations.

There are two primary evaluation approaches : Intrinsic evaluation and Extrinsic evaluation. The intrinsic evaluation entails the ability of the model to acquire information on the linguistic features of a language while the extrinsic evaluation centers on how the model effectively applies to the linguistic aspects [25]. The model comparison may be done as follows : Every model used for the assignment is subjected to the overall data ; the training and the testing data sets are partitioned ; and the performance of each model is achieved through the ability to fit the data used for testing. The model that assigns higher probabilities to the test set is generally considered superior. However, it's essential to prevent test sentences from contaminating the training set to avoid biased evaluations and inaccuracies in metrics like perplexity. By ensuring rigorous evaluation protocols, researchers can effectively assess the capabilities of language models in NLP applications.

1.3 Visual Question Answering

In the recent past, there have been developments in the area of education that have provided incredible solutions to enhance the ease of teaching and learning different aspects of education. Visual Question Answering (VQA) is one of those emerging technologies.

1.3.1 Definition and Architecture of VQA System

Visual Question Answering (VQA) is an interdisciplinary field of research that lies with computer vision and NLP focuses on the development of AI models that can answer questions related to visual content. These questions are typically posed in the original or natural language (e.g., English, Arabic) and relate to images or videos. Deep learning methods come into play when analyzing visual and textual data that explains the extraction of image/ video features and meaning from questions to enable intelligent generation [26].

In the world of artificial intelligence, the convergence of Computer Vision, NLP, and

the basic understanding of the person or we can say "common sense" reasoning has paved the way for transformative advancements in the field of Visual Question Answering (VQA). The capability of machines to interpret inquiries about images or videos articulated in natural language resides at the confluence of several advanced fields, including computer vision, natural language processing, and artificial intelligence. This interdisciplinary approach enables the extraction of semantic information from visual data and the generation of accurate responses to textual queries. Figure 1.2 provides a visual representation of the intricate relationship between these domains.

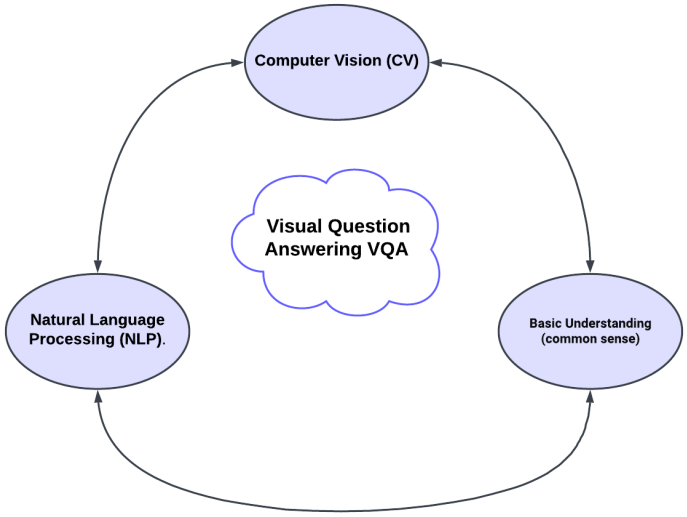

Figure 1.2: Computer Vision, VQA, NLP, and Common Sense relationship [1]

1.3.2 Types of VQA

The major types of VQA can be broadly classified along the lines of the type of question asked, the response type demanded, or the way in which answers are created. in what follows we provide an overview of the key types of VQA :

- **Open-ended VQA**

  This task requires a model to generate a free-form text answer to a given question based on an image (Figure 1.3). Unlike traditional Visual Question Answering (VQA) tasks that limit answers to a predefined set of choices, this approach demands a more sophisticated understanding as the model must comprehend the visual content and produce an appropriate response in natural language [27]. Figure 1.3 provides an example of this process.

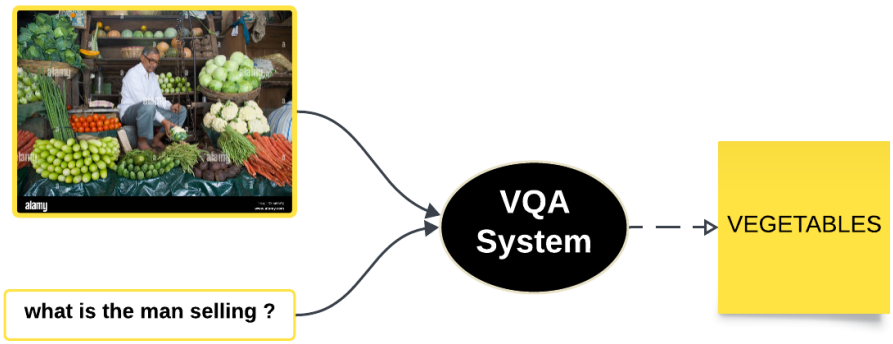

Figure 1.3: illustration of Open-ended VQA process

- **Multiple-Choice VQA**

  This task involves presenting a model with a question about an image along with several predefined answer options. The model's objective is to select the correct answer from these options (Figure 1.4). This multiple-choice format simplifies evaluation by providing clear criteria for correctness and can be less challenging for the model since it only needs to choose from a set of pre-selected answers rather than generating a full paragraph response. The multiple-choice QA format is extensively applied in Visual Question Answering (VQA) tasks due to its structured approach, which facilitates the measurement of system performance. Specifically, it allows for straightforward comparison of accuracy and other performance metrics across different models and datasets [28]. Figure 1.4 provides an illustrative example of this process, showcasing how the model interacts with the predefined options to determine the correct answer.

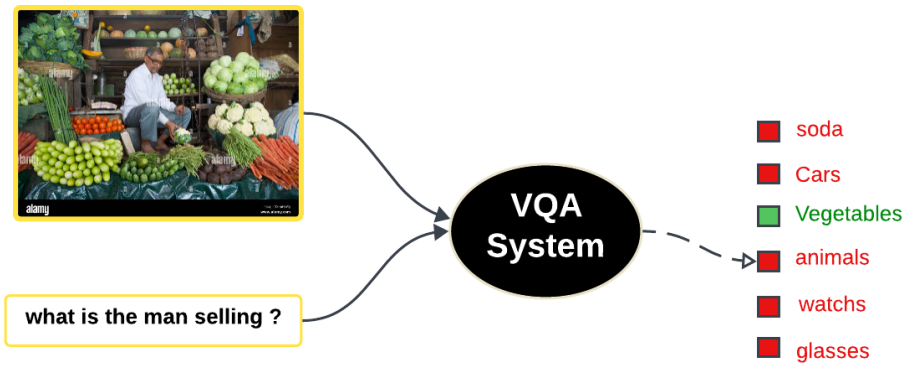

Figure 1.4: illustration of Multiple-Choice VQA process

- **Binary VQA**

  Involves questions that can be answered with a simple "yes" or "no". These types of questions often focus on the presence, absence, or verification of objects or attributes within an image, making the problem more constrained compared to open-ended or multiple-choice VQA [29]. In Figure 1.5, there is an illustrative example.

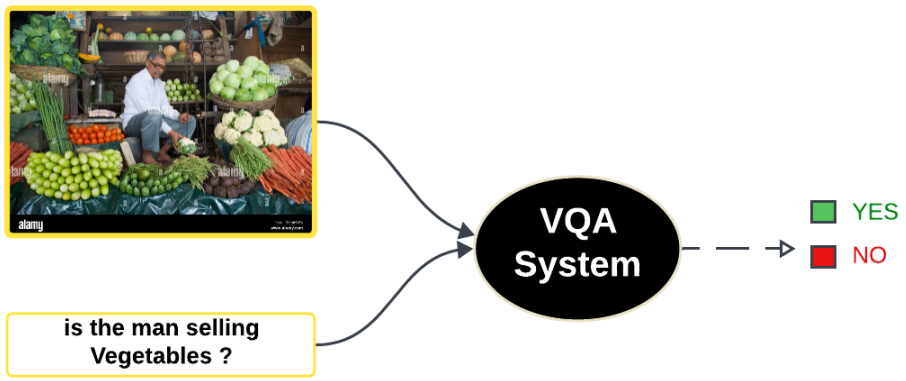

Figure 1.5: illustration of Binary VQA process

- **Visual Dialog**

  Visual Dialog involves an agent answering a sequence of questions about an image in a coherent manner. The task is different from all other VQA tasks where only a single question and answer concerning an image are considered a challenge, for Visual Dialog, the context of the conversation and the entire history of the exchange is important creating an additional challenge. The goal is to maintain a meaningful dialog that correctly references and relates to the given image [30]. In Figure 1.6, there is an example of this phenomenon.

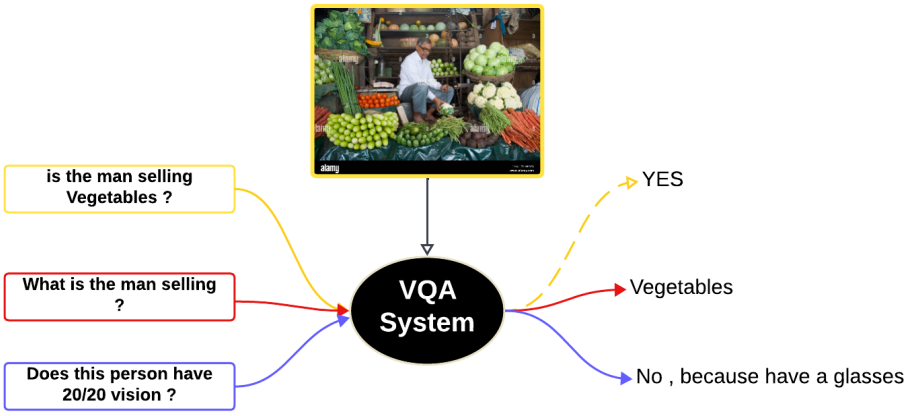

Figure 1.6: illustration of Visual Dialog process

1.4 Visual Question Generation

Visual Question Generation (VQG) is a task that has recently become an interesting research area. The task of VQG involves generating meaningful questions based on the input image. It is a multi-modal problem involving image understanding and natural language generation, especially using deep learning methods. VQG can be considered as a complementary task of VQA [31].

Visual Question Generation (VQG) refers to the automatic creation of questions from visual inputs. This task requires advanced knowledge of computer vision and NLP to create informative questions about the visual content. The questions that are generated can either belong to a declarative and referential type or belong to an inductive and relational type. VQG systems aim to enhance various applications, such as educational tools, interactive storytelling, and improving machine understanding of visual scenes [31].

1.5 Conclusion

For the purpose of our study, in this chapter we have presented an overview of the key aspects of language learning, such as speaking, listening, writing, and reading. after we describe a comparative analysis of language learning applications, identifying their respective advantages and limitations. The chapter also introduced Natural Language Processing (NLP) and its educational applications, followed by an exploration of Visual Question Answering (VQA) and their types, and then Visual Question Generation (VQG). This sets the stage for understanding the integration of advanced technologies in language education and our contribution in this domain.

Chapitre 2

Related Work

Contents

This chapter is dedicated to a state-of-the-art review related to the approaches and architectures that we will deploy in our AI-based pedagogical Tool. These approaches are namely VQA, VQG, and related Datasets.

In order to design VQA systems, two categories of solutions are raised. Traditional machine learning approaches and modern deep learning approaches.

2.1 Traditional Approaches of VQA

Traditional approaches predominantly utilized both visual and textual encoders to extract features from multimodal inputs followed by some form of fusion strategy to combine the encodings. The combined result is subsequently forwarded to either a classifier or a generator, contingent upon whether the task of generating answers is regarded as a classification issue or a generative one. Traditional approaches of VQA typically involve three phases 2.1 :

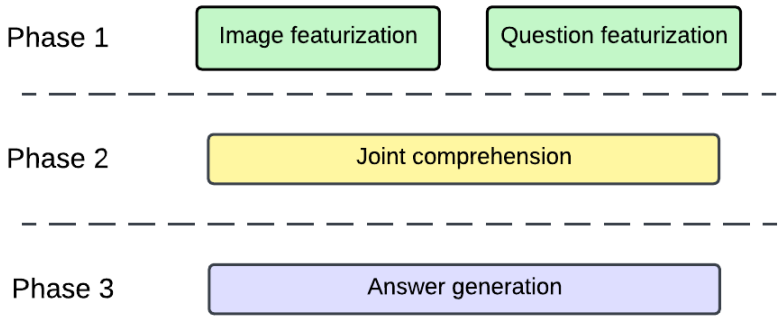

Figure 2.1: General VQA phases [1]

In what follows, we detail the functionality of each phase while reviewing the related aspects.

Phase 1 : Featurization

Featurization is the process of extracting useful and exploitable key information from data :

- **Image Featurization**

  In this phase, the input image is processed to extract meaningful visual features. This can be done using pre-trained convolutional neural networks (CNNs) such as ResNet or VGG, which encode the image into a fixed-size feature vector capturing its visual content [1].

- **Question Featurization**

  Similarly, the input question is transformed into a structured representation that can be understood by the model. This often involves tokenizing the question into words and then mapping each word to a high-dimensional vector representation using techniques like :

- **Count-based Methods**
  - **One-hot encoding :** This is one of the simplest approaches where each word in the vocabulary is represented as a binary vector with a single high value (1) and all other values low (0). However, this method does not capture semantic similarity between words [1].
  - **Co-occurrence matrix :** This method captures the frequency with which words appear together within a specific context window. Words that often appear together will have similar vectors, providing a more nuanced representation of word meanings compared to one-hot encoding [1].

- **Prediction-based Methods**
  - **Neural Network-based Embeddings :** These involve training neural networks to learn word representations. Words are represented in such a way that words with similar meanings are close to each other in the vector space [32].
  - **Word2Vec :** A well-known method developed by Mikolov et al., which uses a shallow neural network to predict words in a context (Skip-gram) or predict context words for a given word (CBOW) [33] [1].
- **Recent Trends in Text Embedding for VQA**
  - **Deep learning techniques :** Recent advancements have seen the adoption of deep learning models such as convolutional neural networks (CNNs), long short-term memory (LSTM) networks, and gated recurrent units (GRUs). CNNs are effective for extracting local features from text, while LSTMs and GRUs are capable of capturing long-term dependencies in sequences, which is crucial for understanding the context in questions.
- **VQA Question Embeddings**
  - **LSTM Models :** Long Short-Term Memory networks (a type of RNN) are frequently used because they are capable of learning order dependencies in sequence prediction problems. They can capture the sequential nature of questions effectively.
  - **Tree-LSTM Networks :** An advanced version of LSTMs that models the syntactic structure of sentences, helping in better understanding and reasoning.

* **Semantic Tree Representations :** These methods aim to capture linguistic structures and improve reasoning abilities by representing sentences as tree structures.

Phase 2 : Joint Comprehension

In this phase, the featured image and question representations are combined to form a joint representation that captures the relationship between the visual and textual inputs. This joint representation is typically achieved through fusion techniques such as concatenation, element-wise multiplication, or attention mechanisms. The goal is to enable the model to understand both the content of the image and the semantics of the question simultaneously [1].

Phase 3 : Answer Generation

Finally, based on the joint representation formed in the previous phase, the model predicts the answer to the given question. This can be done using various techniques such as multi-layer perceptrons (MLPs), recurrent neural networks (RNNs), or transformer architectures.VQA can be considered either a generative problem if the model must generate a natural language response, or a classification problem if the model must select an answer from a predetermined set of options. When treated as a classifier or discriminative model, a single-word answer is usually generated. However, generative models often give longer responses [1].

2.2 Modern Approaches of VQA

Recent approaches in VQA have transitioned from the conventional joint encoding strategy to the utilization of Vision Language Pre-training (VLP) frameworks, which incorporate transformer architectures. These models are initially trained on diverse tasks using extensive image-text paired datasets, followed by fine-tuning for specific downstream tasks such as VQA. Even if traditional approaches form the basis of VQA, current methods are based on several advanced techniques that can provide several benefits such as performance improvement, model scalability, and reliability. The field of advanced VQA techniques is very vast. hereafter, we discuss some of these advanced methods :

2.2.1 Attention Mechanism

A general weakness of most models using the traditional method is that a generic (image-wide) feature is substituted for the visual input. This can introduce unrelated or superfluous information to the prediction phase. Transformer architecture addresses this issue by using local soft weights on image features where the model is trained on more specific regions of an image [34].

There are several ways to implement attention, the mechanism is basically based on creating an attention vector by correlation of categories followed by some kind of normalization. The whole process is usually called a single layer of attention. Depending on the architecture, models can be the focus of the enterprise in one or more layers. Some types of multi-layered attention may depend on the attention of previous layers. Attention mechanisms are defined into single-hop and multi-hop attention based on the number of layers of attention. The classifications of attention can be further understood through their types such as visual attention, textual attention and joint attention [35].

2.2.2 Transformer

Recently, Transformer models have demonstrated their capabilities, achieving outstanding results across a wide range of linguistic tasks, such as text classification, machine translation, and question answering. These models have set new benchmarks in performance and versatility. Similar to RNNs, there are various forms of transformers, including BERT, which excels in understanding context, GPT-1 to GPT-4, known for generating coherent text, RoBERTa, which enhances BERT's pretraining, and T5, designed for multiple text generation tasks.

Two main concepts have advanced traditional transformer models :

- The first concept is self-attention, which captures long-term dependencies better than recurrent models.
- The next main concept is first training on a big set of data, then adjusting on a small labeled dataset for the specific task like VQA.

Dosovitskiy et al [36]. created the Vision Transformer (ViT), a transformer-based alternative that gained popularity for extracting visual features. Instead of traditional convolutional networks, ViT leverages transformer architecture to process image data. Transformer-based templates have been widely used to integrate visual and textual components. Additionally, transformer-based models were extensively used for combining visual and textual modalities. The structures varied based on the number of transformer

layers or the type of attention utilized. CLIP framework [37] aligns vision and language modalities and has been widely used in zero-shot VQA techniques.

2.2.3 Vision Language Pre-training (VLP)

Recently, there has been a widely researched field Vision-and-Language (VL) which combines Computer Vision and Natural Language Processing. Prompted by the remarkable advancements of pretrained language models in NLP like BERT , RoBERTa , T5, and GPT-3, the process of VL pretraining for tasks such as Image Captioning, VQA, and Visual Reasoning has raised. The objective of pre-training is to train a large-scale global model across various tasks using a large amount of data. The resulting pretrained model will be further fine-tuned for specific tasks. The most common methods for pre-training are Masked Language Modeling (MLM), Masked Vision Modeling (MVM), and Vision-language Matching (VLM) [35, 38].

2.3 Visual Questions Answering Architectures

A wide variety of approaches are present in the literature guide to different architectures. Traditional approaches often use joint embedding architectures where two input streams are processed and merged independently of each other. Instead, modern VLP architectures include transformer-based encoders, fusers, and decoders. VQA architectures can be classified into three distinct categories the traditional CNN-RNN-based, the subsequent CNN-BERT-based, and the VLP-based architecture :

2.3.1 CNN-RNN-based Architecture

Deep learning was first based on two major types of models, CNN-RNN, in which the CNN network was used for visual features extraction, and the RNN network for text representation. CNNs like GoogLeNet, VGGNet and the deep ResNet were impactful on the images because they had few convolutional layers. LSTM and GRU-based RNNs were suitable for sequential data, mitigating vanishing gradient problems and learning text data efficiently. The combination of CNNs and RNNs allowed for parallel data analysis in space and time and produced better performance of numerous tasks such as video categorizing and image tagging where CNN analyzed spatial data and RNN analyzed temporal data [35].

2.3.2 CNN-BERT-based Architecture

The VQA methodologies were transformed by a BERT which means Bidirectional Encoder Representations, a variation of the transformer architecture. In the subsequent CNN-BERT-based architectures, BERT became the primary choice for textual encoding along with a later variant of CNN for visual encoding. To combine multimodal input streams, derivatives of transformers such as multimodal transformers, crossmodal transformers (CMTs), and BERTs were used. The CMTs used both bi-directional cross-attention and self-attention. Cross-attention calculates attentional weights based on inputs from one category and comparison with inputs from the other category [35].

2.3.3 VLP-based Architecture

Two propositions introduce the structure of the VLP model. These propositions emerge from two perspectives. Initially, regarding how the *fusion of multiple modalities* is done ; it can be observed whether it involves single-stream fusion or dualstream fusion. Subsequently, in terms of *the overall architectural design*, it can be analyzed whether it is encoder-only or encoder-decoder [38].

The single-stream architecture refers to the text and visual features being concatenated together and then fed into a single transformer block. The single-stream structure utilizes merged attention to fuse multi-modal inputs. The dual-stream architecture refers to the text and visual features are not concatenated together but sent to two different transformer blocks independently. These two transformer blocks do not share parameters. To increase the effectiveness of the model cross-attention is used to enable cross-modal interaction [35].

Viewed from another perspective, VLP models employ the encoder-only architecture, into which the cross-modality embedding vectors are directly fed into an output layer to generate the outcomes. On the other hand, other types of VLP models benefit from a transformer-encoder-decoder framework where the cross-modal embeddings by first fed to the decoder layer and then to an output layer to produce the responses [38].

2.4 Datasets

The domain of Visual Question Answering (VQA) is characterized by the presence of numerous extensive datasets that have been curated over time. These datasets typically consist of triples comprising an image, a corresponding question, and the correct answer. The task of constructing a VQA dataset can be traced back to the period before the widespread adoption of deep learning techniques in the realm of VQA. Table 2.1 reports a

comprehensive examination of multiple publicly accessible datasets along with a detailed analysis of their individual characteristics.

Nature of Images	Dataset	structure	Size	More Details	Disadvantage	Language
Natural Images	DAQUAR	Derived from NYU-Depth v2	1449 RGBD images	12468 question-answer pairs, synthetic and human-generated	Questions are restricted to templates and answers are limited to classes	English
	COCO-QA	Utilizes COCO images	123,287 images	Questions classified by nature, vast range of queries	Answers are limited to a single word only	English
	FM-IQA	Utilizes COCO images	158,392 images	Questions and answers are human-generated	Visual turing test-based manual evaluation is unscalable	Chinese
	VQA v2	Natural Images from MS-COCO	204,721 images	Diverse questions including binary and multiple-choice	Lacks questions on general knowledge,Insufficient reasoning-based questions	English
	Visual Genome	Natural Images from MS-COCO and YFCC100M	108,249 images	Includes 1,773,258 question-answer ,scene graphs, diverse answers	Difficult to evaluate long answers,Inherently too large	English
	VAQA	Utilizes MS-COCO	5000 real-world images	2712 unique questions, and two answers 'yes' and 'no', resulting in 137,888 IQA triplets.	can be used just for binary classification problems.	Arabic
Clipart Images	VQA-balanced	Composed of clipart images	15,623 images	Balanced distribution of question types, answer types, and image attributes	Only yes/no questions	English
	VQA-abstract Scenes	Composed of clipart images	50,000 images	Questions and answers focus on abstract visual concepts and reasoning	-	English
knowledge Base Enhanced	KB-VQA	Natural images from MS COCO	700 images	Images, questions, answers, and relevant facts from knowledge bases	Small scale dataset	English
	FVQA	Natural images from MS COCO val set and ImageNet	2190 images	Integrates external facts and commonsense reasoning	Long training time	English

Table 2.1: Some VQA Datasets

The development of image datasets has played a pivotal role in the advancement of computer vision, providing essential resources for training and evaluating models. In Figure 2.2, we summarize all of these dataset and more in a timeline containing a collection of datasets from 2015 to 2023.

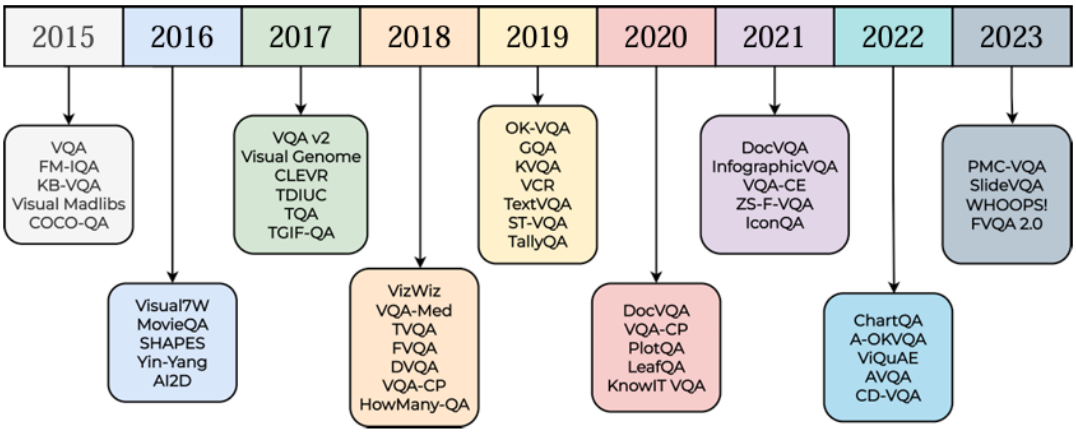

Figure 2.2: Timeline of popular VQA datasets. [1]

2.5 VQA Approaches for Pedagogical tools

The sector of education bears the responsibility of cultivating the scholars of the forthcoming generation. Artificial intelligence integration in education will be discussed and how it can help increase the availability and extend educational options. However, it should be stated that this particular area of development is impressively sprawling and can cover a wide range of activities, extending from teaching the basics to young learners all the way to offering professional development to people of diverse occupations. The process of question-answering serves as a fundamental method utilized to strengthen and assess the acquired knowledge. Systems for visual question answering have the capability to serve as valuable educational aids for students, enthusiasts, as well as professionals.

He et al [39] have proposed the use of an AI Robot system in achieving the Preschooler's education objectives including metacognition tutoring and geometrical thinking training, although it incorporates Contextual teaching which is the learning that takes place in real life. The main function of this module is to understand the objects involved in the scene from the image captured by the robot using Faster R-CNN and then produce several questions about the qualities and colors of the objects detected. They use there approach

[1] from survey of Department of Computer Science and Engineering, Techno International New Town, Kolkata, India

was successful in capturing the interest of both parents and kids while boosting the kids' curiosity.

Sophia and Jacob [40] presented during the pandemic a Student Chatbot software integrated with a user website that helps students in education by handling students' queries through defined intents The chatbot system uses the Recurrent Neural Network (RNN) to deal with Textual part and the Convolutional Neural Network (CNN) to handle the Visual part.

Dipali Koshti and all [35] propose a different method for the VQA system for education that looks to make greater use of the semantic knowledge extracted from the image. they created their dataset using educational images where each data point is made up of an image, a question that corresponds to it, a valid response, and a fact that supports it. each fact is then represented by a triple (S, V,O) where S represents the subject while V represents the verb and lastly O is the object. Visual Features are extricated using pre-trained ResNet and Textual Features using a pre-trained BERT model. Gyeong-Geon Lee and Xiaoming Zhai [41] proposed a VQA system for education using GPT-4.

2.6 Architecture, Approach and Datasets of VQG

As mentioned in the previous chapter, VQG is a compelling issue that has garnered recent attention. The task of question generation is often perceived as more cognitively demanding than question answering, requiring a profound comprehension of the material followed by articulating it in clear and precise language to facilitate unambiguous responses. Visual question generation (VQG) constitutes a task involving the creation of questions based on images, thus the execution of two pivotal sub-tasks consecutively : (1) interpreting the image and (2) formulating a coherent sequence of words constituting valid questions. Further comprehension of the image requires successful object detection, classification, and labeling, as well as identifying relationships between objects, understanding the scene, and classifying the scene, among other tasks The observed methods for addressing this problem include creating an image representation through Convolutional Neural Networks (CNN) or similar architectures. Then, either RNN networks or other related networks are applied to generate word sequences [42].

Questions related to visuals can be classified into three distinct categories. The initial category pertains to questions that are entirely rooted in the image itself, where both the questions and their corresponding responses can be deduced solely from the information available within the image. The second category involves questions based on common sense, requiring additional external common sense knowledge in conjunction with the in-

formation present in the image to answer the questions. Lastly, the third category consists of questions relying on worldly knowledge, which cannot be formulated or addressed solely based on the provided image [42].

VQA datasets could be used to train VQG models. Most of the currently available VQG techniques rely on pre-existing VQA datasets. Most of the VQA datasets have questions that are human generated and obtained through crowdsourcing platforms, including AMT(Amazon Mechanical Turk). While some of the datasets described in the Datasets section are not yet used in VQG, they can be. There are some datasets specifically for VQG, such as VQG Flickr-5000, VQG COCO-5000, and VQG Bing-5000, CRIC. However, we have not described these datasets [42].

2.7 Discussion and Conclusion

In conclusion, the prospective utilizations of VQA present promising prospects for engineers and business innovators. VQA has undergone significant transformations through the introduction of deep-learning-based structures, transformer structures, and presently, Large Language Models (LLMs) and generative Artificial Intelligence (AI). Despite the progress in Visual Question Answering (VQA) research facilitated by varied datasets and sophisticated models, there are several persistent challenges.

- **Arabic VQA Disregarded** Most of the VQA literature worked with English QA pairs and disregarded the performance of models in other languages, particularly in the context of Arabic language VQA.

- **Lack of Arabic VQA datasets** There is a scarcity of Arabic VQA datasets, which limits the training and evaluation of VQA models in this language. Sarah M. Kamel et al. [43] generated the first Visual Arabic Question Answering (VAQA) dataset to address this gap.

- **VAQA datasets are more binary question** The Visual Arabic Question Answering (VAQA) dataset, generated by Sarah M. Kamel et al. [43], primarily consists of binary questions. This dataset was fully automatically generated, providing a valuable resource for training and evaluating VQA models in the Arabic language. The automation process ensures consistency and scalability, though it may also introduce certain biases inherent in automated data generation methods. The dataset consists of almost 138k Image-Question-Answer (IQA) triplets and is specialized in just yes/no questions about real-world images which can be used just for binary classification problems.

- **Single Question vs Set Question** While the conventional approach of utilizing a single image and single question has been widely accepted across various fields, alternative setups such as image pair and image-set question answering have received minimal attention. The scope of counting-based inquiries, as observed in some studies, is confined to a solitary image, leaving the issue of multi-image counting unanswered.

- **Everyday scenes images vs. Specialized images** the majority of the aforementioned research focuses on utilizing benchmark datasets such as VQA, COCO-QA, and DAQAUR, which consist of images depicting everyday scenes. However, there is limited research on using specialized images or contexts for question answering.

- **Specialized domains datasets** There remains a notable difficulty in implementing VQA in specialized domains like Medicine, Satellite imagery, or Education, primarily because of the insufficient availability of domain-specific datasets.

Chapitre 3

Pedagogical Tool Design

Contents

Learning, speaking, and understanding the Arabic language is of great importance skills, not only in the Arab world but also in the global area, because the Arabic language is one of the most popular languages in the world. The field of visual question answering (VQA) has seen significant developments with the introduction of deep learning-based architectures, transformer architectures, and, more recently, large language models (LLMs) and generative artificial intelligence (AI). These innovations have transformed the quality of VQA, offering promising prospects for engineers and business innovators. However, most advances in VQA research have mostly focused on English-language datasets, leaving a significant gap in performance and applicability for non-English languages. This gap is particularly evident for the Arabic language, which poses unique challenges in creating VQA datasets and models.

Additionally, finding suitable images to ask appropriate questions about them poses a challenge, especially from an Islamic perspective. Generating images is also problematic, as the available sources are either paid, costing £9 or more, or are incomprehensible and complex.

This work seeks not only to advance the theoretical understanding of Visual Question Answering (VQA), but also to provide practical solutions that can be applied to Study and develop what we have achieved, learning the Arabic language, improving understanding of vocabulary, and distinguishing between objects in the picture

3.1 Overview of the Targeted Pedagogical Tool

Our proposed pedagogical Tool, baptized Arabic-EDU and based on VQA involves a multi-step process that integrates advanced image processing and Natural Language Generation techniques to accurately answer questions about visual content. Here is a succinct overview of our approach :

3.1.1 The Proposed ARABIC-EDU-VQA Architecture

We propose this model of pedagogical tool to explain our architecture for Visual Question Answering that combines Visual Question Answering (VQA) and Visual Question Generation (VQG) components. This pedagogical tool aims at active learning by providing the learner with a set of pedagogical activities. The targeted ones are based on VQA system by the integration of the two components and benefits. We provide an overview of their architecture and its components in Figure. 3.1 :

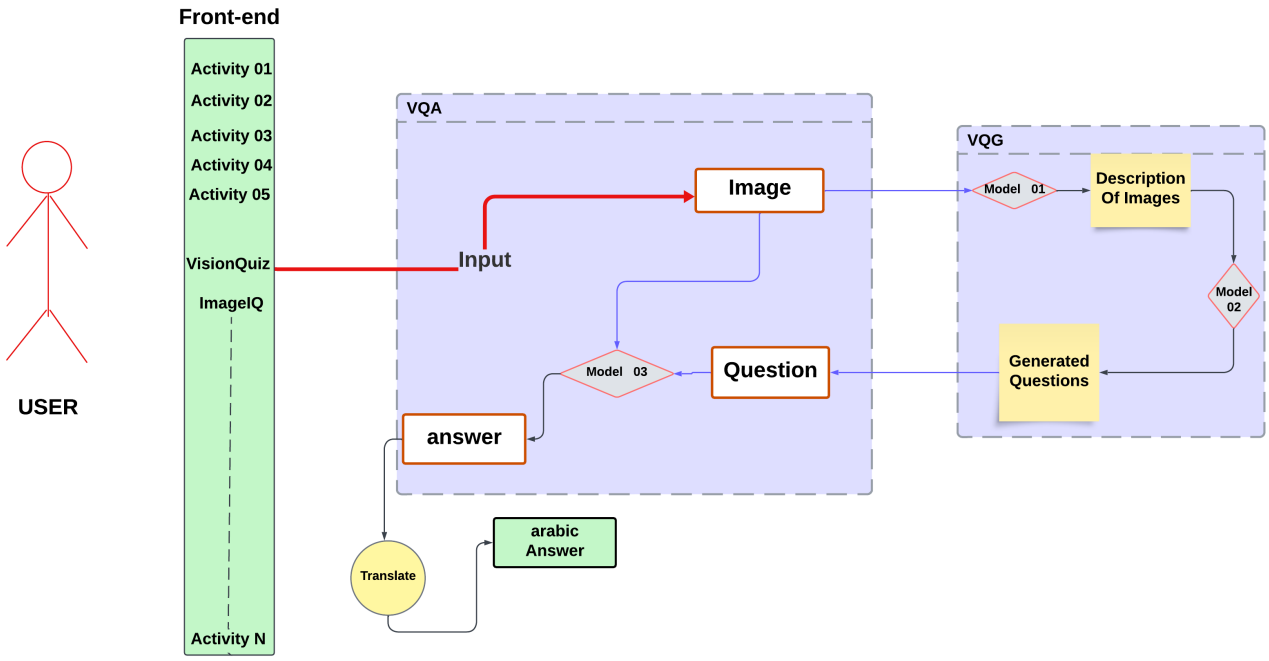

Figure 3.1: The Pedagogical Tool Architecture

When the pedagogical activity **"Quiz"** or **"ImageIQ"** is launched, an image is uploaded, we deploy a classic deep model to obtain important features and work with a text description. Next, a language generation model creates diverse, contextually relevant questions from these descriptions. Our question generation module synthesizes pertinent questions, forming the basis for the VQA process. A fine-tuned VQA model combines visual and textual information to provide accurate answers. Finally, our system provides multi-choice quiz based on model outputs and textual analysis translated into Arabic for the user to solve this Quiz.

3.1.2 Our VQG

This system is responsible for generating questions from the image, it contains two parts **Description of Image** and **Generated Questions** and we use two models in Figure 3.2 :

- **Model 01 :** Model 01 is responsible for generating descriptions of images. The theoretical basis for this model involves translating visual data into textual descriptions. This process relies on deep learning techniques. small vision-language model is perfect for this task since its pedagogical tool helps to learn Arabic we need simple descriptions of images.

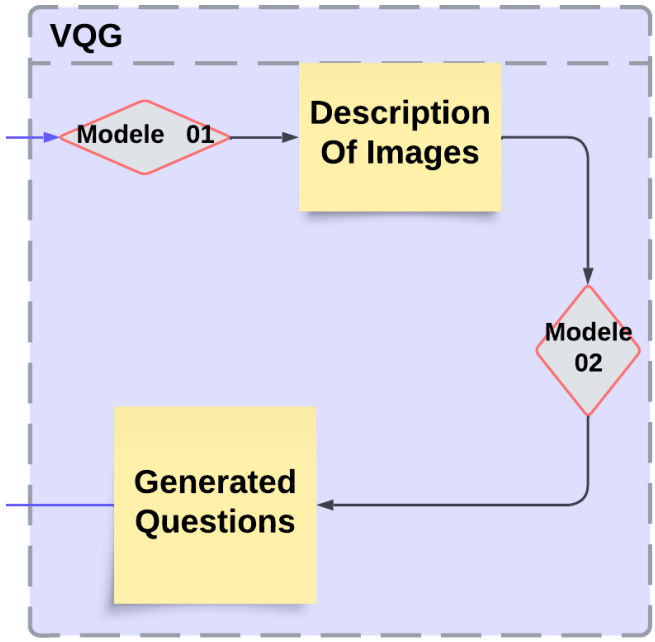

Figure 3.2: The architecture of the Proposed VQG

- **Description of image :** model 01 creates clear descriptions of uploaded images. It identifies key objects, scenes, and spatial relationships, and then turns this information into text. These descriptions summarize the image content, help generate relevant questions, and improve the accuracy of the VQA system.
- **Model 02 :** We used this model to generate questions based on image descriptions. This powerful model generates diverse and relevant questions for the user to choose from.
- **Generated Questions :** Our module creates a variety of questions related to the content of uploaded images using the **model 02**. These questions form the basis for accurate and meaningful interactions in the VQA process.

3.1.3 Our VQA

This module is responsible for **answering questions** from **the image**, we marge two models :

- **Image :** When a picture is inputted into the system, the image is saved in a database. that can easily be retrieved for feature extraction, descriptions, and other questions.

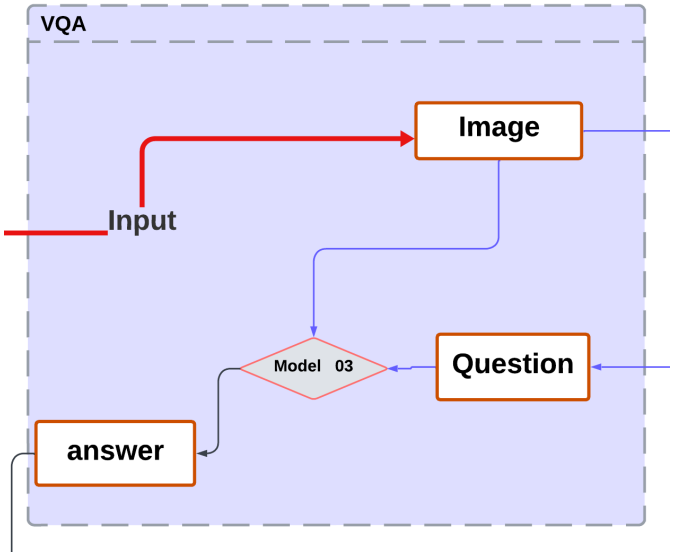

Figure 3.3: The architecture of the Proposed VQA

- **Question :** Our system then chooses a random question from the many questions in the generated question, and the user can select the question he likes or understands. When he chooses the question, it's processed in the next model to give the final result.
- **Model 03 :** This Proposed ARABIC-EDU-VQA system is used to help language learners learn Arabic via activities, one of these activities is a Quiz game to help the user test his knowledge of Arabic words, we give the user images and questions to answer and list of answers which represent quiz options, example Figure. 3.4 :

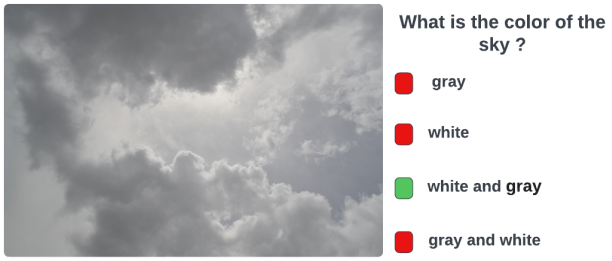

Figure 3.4: Illustration of a VisionQuiz Proposed by ARABIC-EDU-VQA

In the related work chapter, we mention that there are two types of answer generation models : one that generates a natural language response and another that selects an answer from a predetermined set of options. In the first type of model we get an accurate response but when we get several responses from the model, these

are arrays of responses provided for the user to select from. The responses are often similar or have close values, making it difficult for the user to choose the correct one. With the second type of model, we can get a range of different values, but a less precise answer.
to solve this problem we merge the two types of models to get accurate answer and distinct values.

- **Answer :** In our Visual Question Answering (VQA) system, the combination of the uploaded image and the selected question from the user and generated questions generates accurate answers. The system integrates the image and question using model 03, to give multiple choices.

3.1.4 Translate to Arabic

Finally, we provide the array of answers obtained from model 03 to another Arabic translation model so as to expand the ability of the VQA system to provide to the user Arab responses to select from and also translate questions into Arabic so as to enrich his knowledge [44]. This added functionality is also due to the luck of accurate Arabic-VQA and Arabic-VQG.

3.2 Design of our System

Our ARABIC-EDU-VQA pedagogical tool is an end-to-end system. The system is composed of several integrated components that work together smoothly. Using UML, in the following diagrams, we describe the key components of the system and relation between models :

3.2.1 Use Case Diagram

The use case diagram provides an overview of the interactions between users and the system, showcasing the various functionalities available Figure. 3.5 :

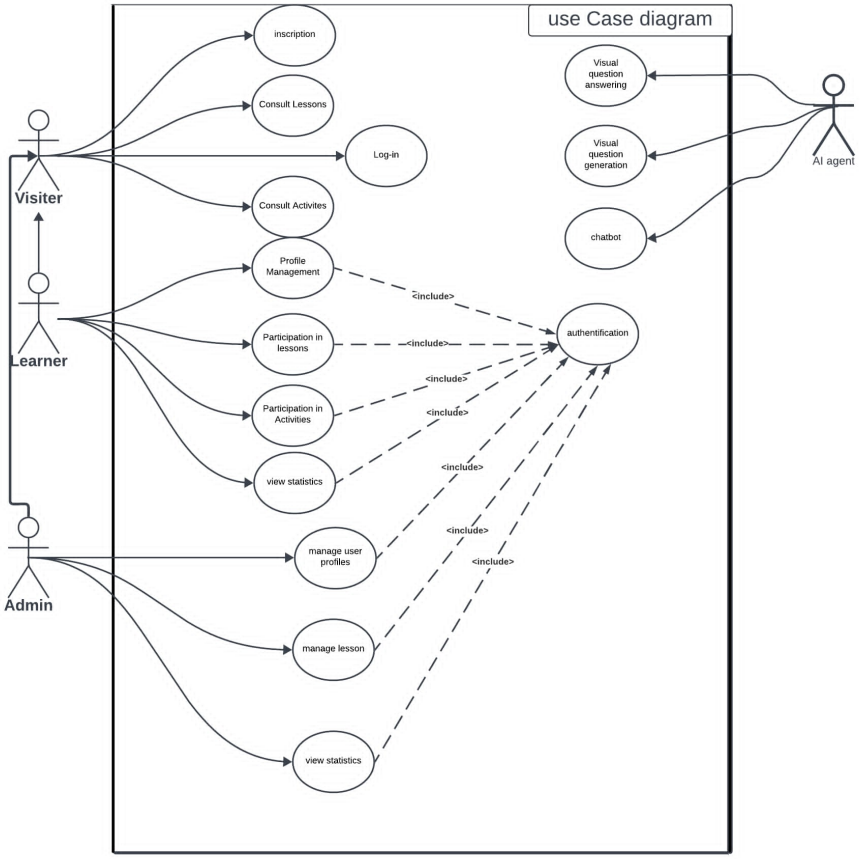

Figure 3.5: Use case diagram

3.2.2 Class Diagram

The class diagram outlines the structure of our pedagogical tool, highlighting the main classes, their attributes, methods, and relationships.

The three groups in red indicate that we have not yet developed in these areas **(person, courses, section)** in Figure 3.6 :

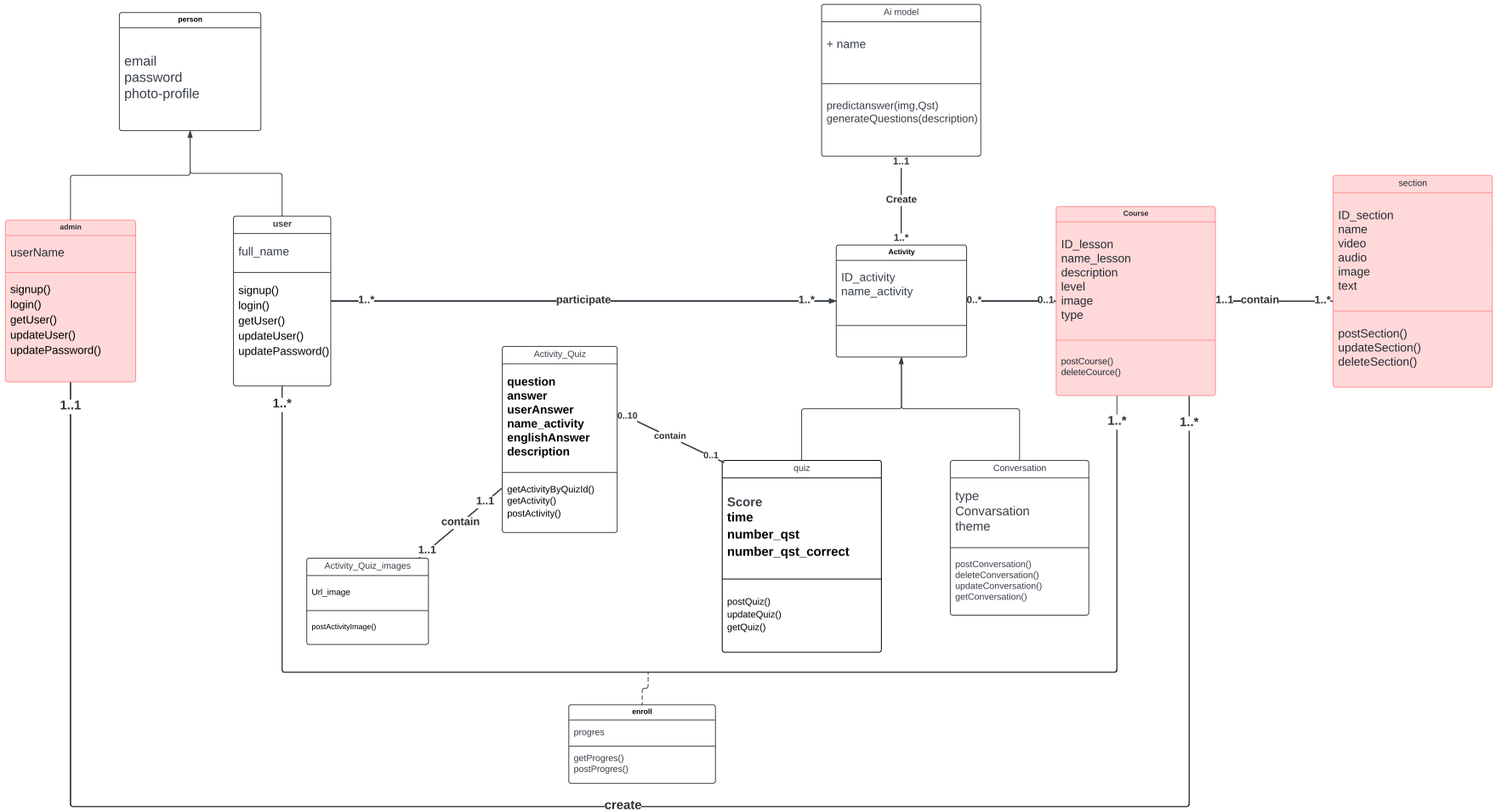

Figure 3.6: Class Diagram

3.2.3 Sequence Diagram

A sequence diagram is a Unified Modeling Language (UML) diagram that illustrates the sequence of messages between objects in an interaction. A sequence diagram consists of a group of objects that are represented by lifelines and the messages that they exchange over time during the interaction.

The sequence diagrams detail the interactions between objects in the system for specific functionalities, illustrating the sequence of messages exchanged.

- Sequence diagram associated with the scenario ("quiz game") in Figure. 3.7.
  Important : The imageService is an external component used for handling images in the pedagogical tool.

- Sequence diagram associated with the scenario ("ImageIQ game") in Figure. 3.8.

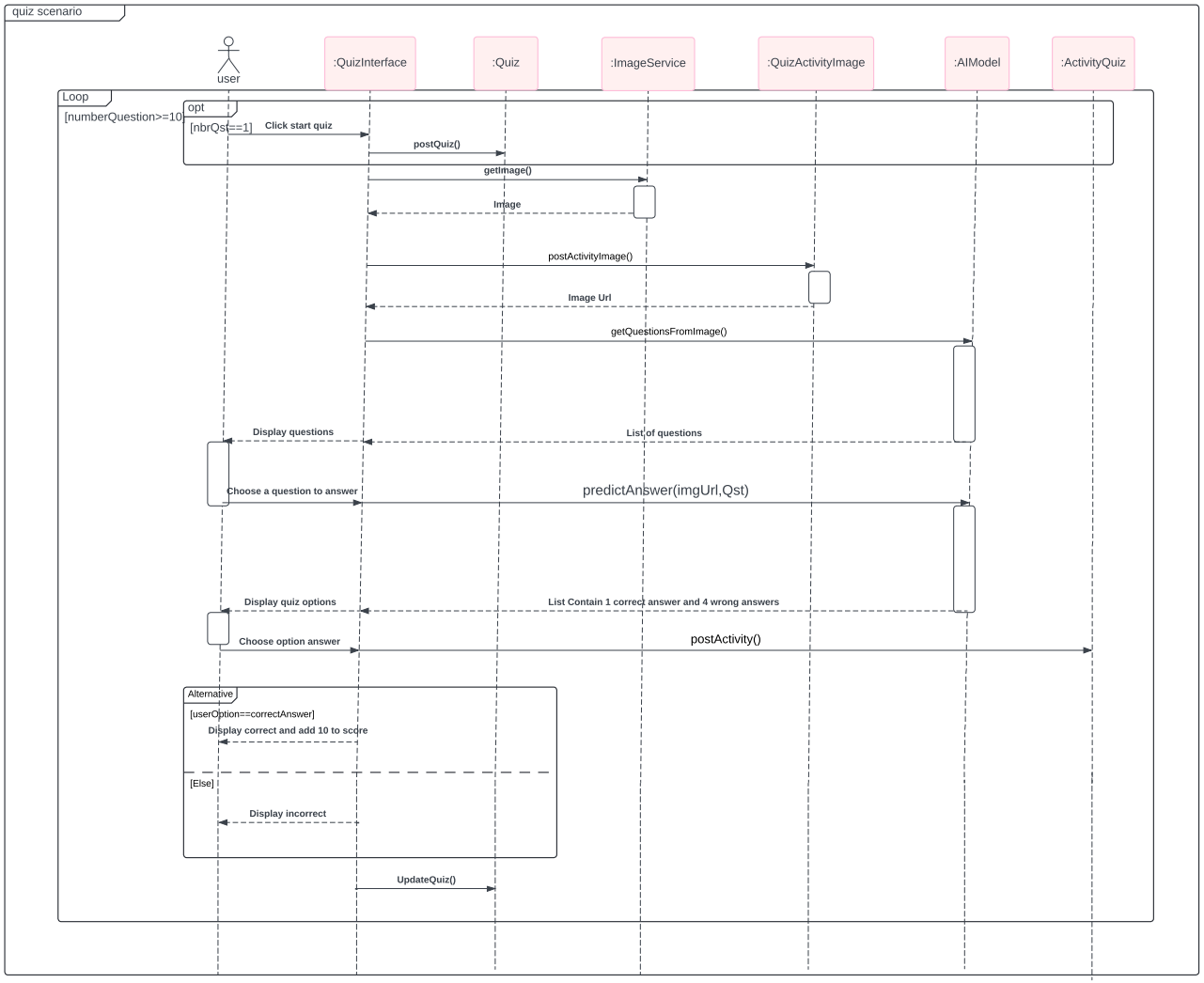

Figure 3.7: Quiz game scenario sequence diagram

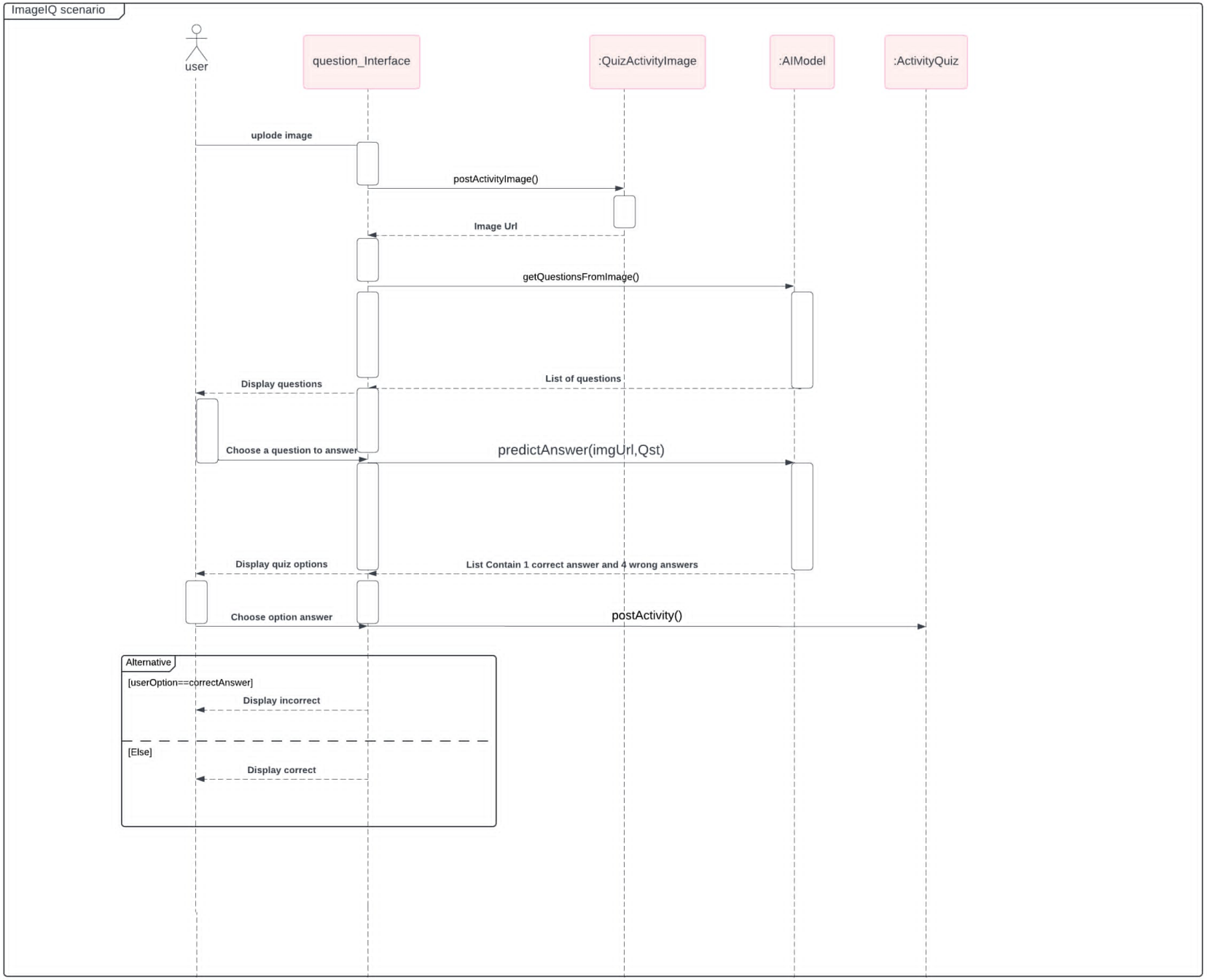

Figure 3.8: ImageIQ game scenario sequence diagram

3.3 Conclusion

In this Chapter, we presented a detailed overview of our ARABIC-EDU-VQA system design. We outlined our system model of VQA in the name of **ARABIC-EDU-VQA** architecture, and we explain their sides as Visual Question Generation (VQG) and Visual Question Answering (VQA), highlighting their features and functionalities. Additionally, we provided comprehensive system diagrams, including the use case, class, and sequence diagrams, to illustrate the structural and operational aspects of our system front-end. This chapter lays the groundwork for understanding the technical implementation and design choices behind our innovative educational tool.

Chapitre 4

Pedagogical Tool Implementation

Contents

In this chapter, we describe the essential components involved in developing our proposed pedagogical tool. This entails designing a graphical user interface through which users can post images, get questions, and get reliable answers from the VQA system. We will cover the back-end infrastructure required to support these models, and the front-end design principles that ensure a seamless user experience.

The various web technologies and frameworks that can be leveraged to build a robust VQA platform, including Tailwind CSS, JavaScript, and popular web development frameworks like React.js and React query.

Let's us now start detailing how we created our website that bridges the gap between cutting-edge VQA technology and how you can learn a language.

4.1 Pedagogical Web-based Architecture

To design our end-to-end web-based Pedagogical tool, there is a need for a solid architecture that is able to interface the front-end interface with the back-end services and the machine learning systems. Figure 4.1 shows the over all design. It shows the interconnection between the front-end and the back-end. This latter is divided into two parts the database and AI-models :

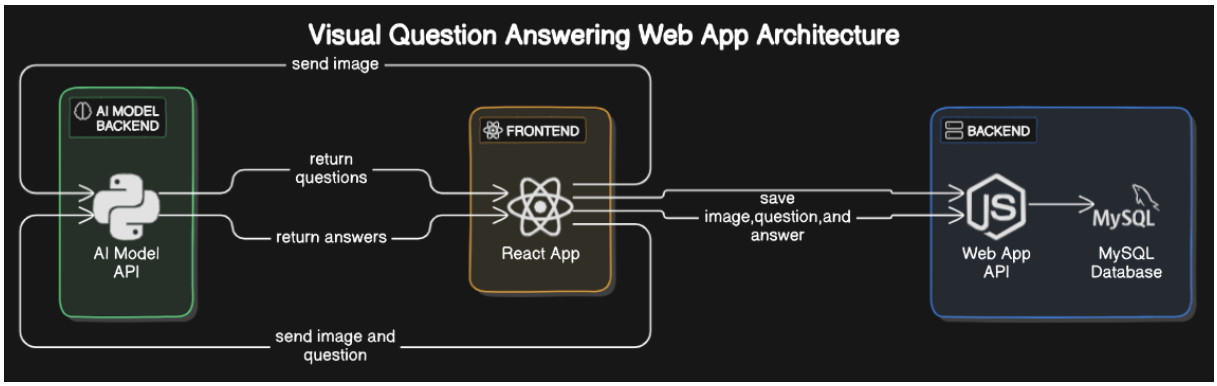

Figure 4.1: Front-end, Back-end services of Our Pedagogical Tool

4.2 Development Environment and Tools

For developing our pedagogical tool, we used some technologies and many tools that allowed us to smoothly implement it, which are mentioned in this table 4.1 :

Component	**Tools**	Description
Software tools	Visual Studio : Xaamps :	-A versatile, open-source code editor developed by Microsoft, known for its extensibility and support for various programming languages. -A free and open-source cross-platform web server solution stack package, including Apache, MySQL, PHP, and Perl, used for local development and testing.
Hardware	CPU GPU RAM Storage	Intel Core i7-12th Gen NVIDIA GeForce RTX 4050 16 GB 2TB
Languages and Libraries	javascript : python : PyTorch :	-A versatile language for creating interactive web content. -A readable, versatile programming language used across various domains. -A machine learning library based on the Torch library, used for applications such as computer vision and natural language processing.

Table 4.1: Development Environment and Tools : Overview of Software and Platforms for Programming and Application Development of our website.

4.3 Front-end

Front-end is the outside face of the web application from which users engage with the system and participate in activities by entering their commands. For the front-end design we used 4.2 :

Frontend	**React.js**	A JavaScript library for building user interfaces with reusable components.
	React Query	A library for managing server state in React applications, providing tools for data fetching, caching, and synchronization.
	Tailwind CSS	A utility-first CSS framework for rapidly building custom user interfaces.

Table 4.2: Core Libraries and Tools for Front-end Development

We start with the **LOGO design**, the adopted Logo is **Kalima** 4.2 which symbolizes the term "word" in the Arabic language. it is a simple logo for learning the first word. Let us just recall that our tool targeted learners with some knowledge of Arabic, not beginners.

Figure 4.2: Logo symbolizing the "word" in Arabic

4.3.1 Sign-up and Log-in

- On the Sign-up page, you have two options. The first option is to fill in your personal information and open a new ‘sign up’ account. The second option is to go to the ’Log-in’ page and enter your email address and password if you have an account (see Figure 4.3).

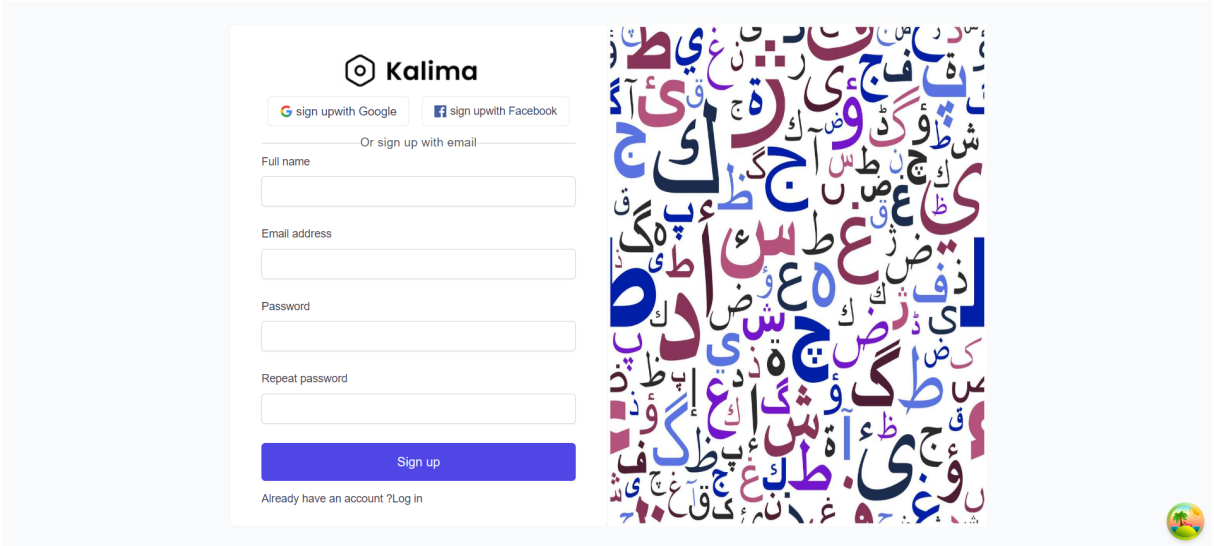

Figure 4.3: Sign-up page in Kalima Pedagogical Tool

- Log-in on this page. All you have to do is enter your registered email and password. if you don't have you must go to the 'Sign-up' page (See Screenshot in Figure 4.4).

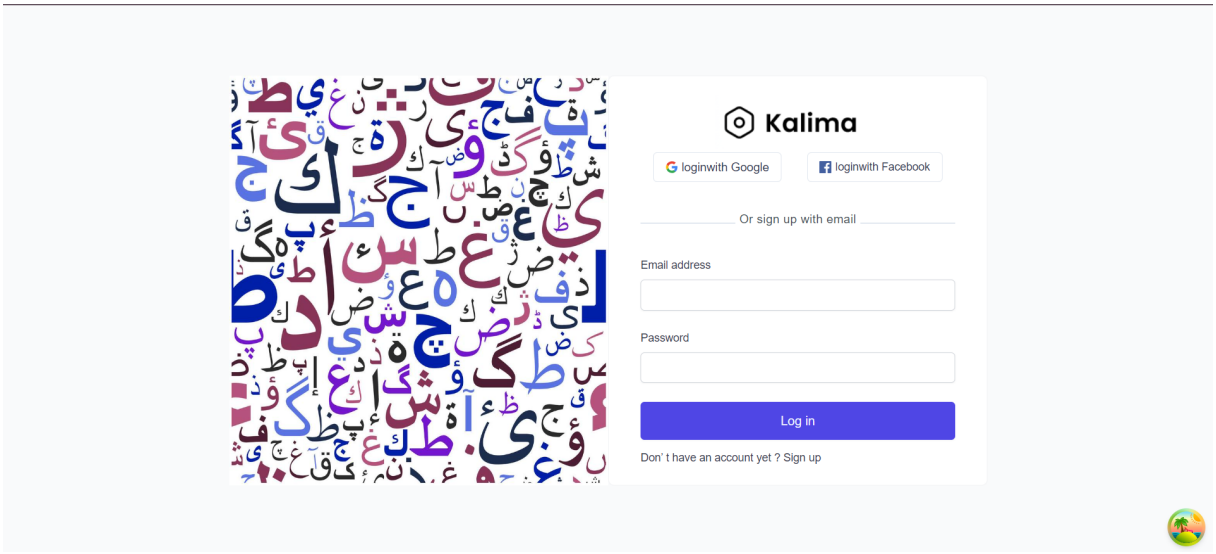

Figure 4.4: Log-in page in Kalima Pedagogical Tool

4.3.2 Dashboard

Our comprehensive dashboard is designed to optimize the learner learning experience and streamline his interactions with our platform. this page is not yet finished, in the future, we can add statistics and more details about the user and their development. Here's what each section offers like in Figure 4.5 :

- **account :** This section contains user information.

- **Home :** The home section is currently undergoing development and will soon include statistical bars to track your progress.
- **Courses :** Dive into our diverse range of courses tailored to suit the learner interests and skill levels. this section is currently undergoing development.
- **ImageIQ :**Test the learner Arabic knowledge with various questions on an image from his choice.
- **Quiz :** Test the learner Arabic knowledge with engaging quizzes on various subjects.
- **Chatbot :** Get instant assistance and personalized recommendations from our intelligent chatbot for learning a language.
- **Help :** Access comprehensive guides, troubleshooting tips, and support resources for a seamless learning experience.

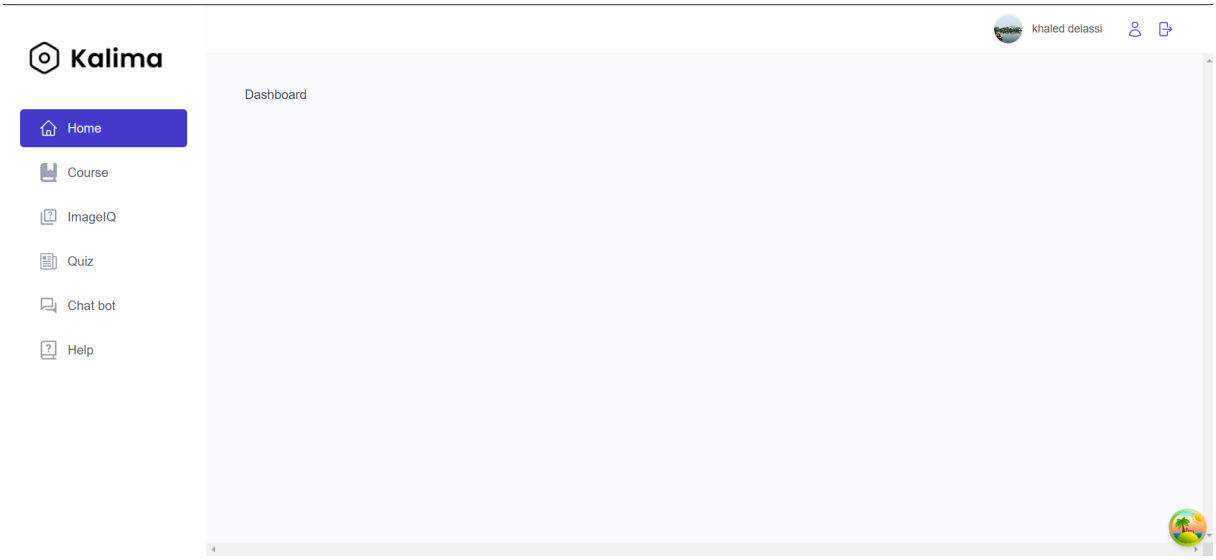

Figure 4.5: Home page in Kalima web application

4.3.3 ImageIQ

ImageIQ is a Quiz Activity where the interface allows a learner to upload an image, and the system will generate a question about it along with five possible answers in Arabic you can see the page in Figure 4.6. Upon selecting an answer, the system provides a feedback :

if the chosen answer is correct then it blinks green, else if wrong the selected answer blinks red with the correct one blinking green like Figure 4.7. This instant feedback mechanism enhances the learning experience by visually indicating the correctness of the response. Additionally, if users don't understand the question they have the option to

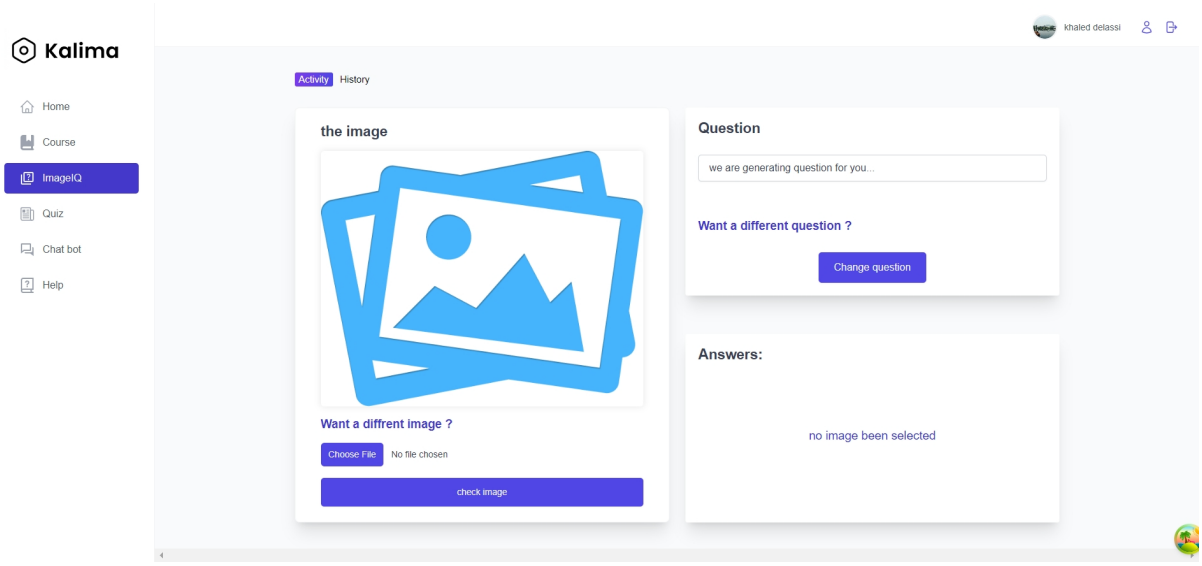

Figure 4.6: ImageIQ Quiz interface in Kalima web application

change the question by clicking the "Change Question" button, allowing for a dynamic and engaging way to practice and improve their visual and language comprehension skills.

(a) selected answer is incorrect **(b)** selected answer is correct

Figure 4.7: Two illustrations about correct and incorrect Answer

Another option, the ImageIQ page includes a "History" in Figure 4.8, where Learners can view their previously uploaded images and the corresponding questions and answers. This learning trace allows learners to review their past interactions and learn from their mistakes, reinforcing their knowledge and understanding over time.

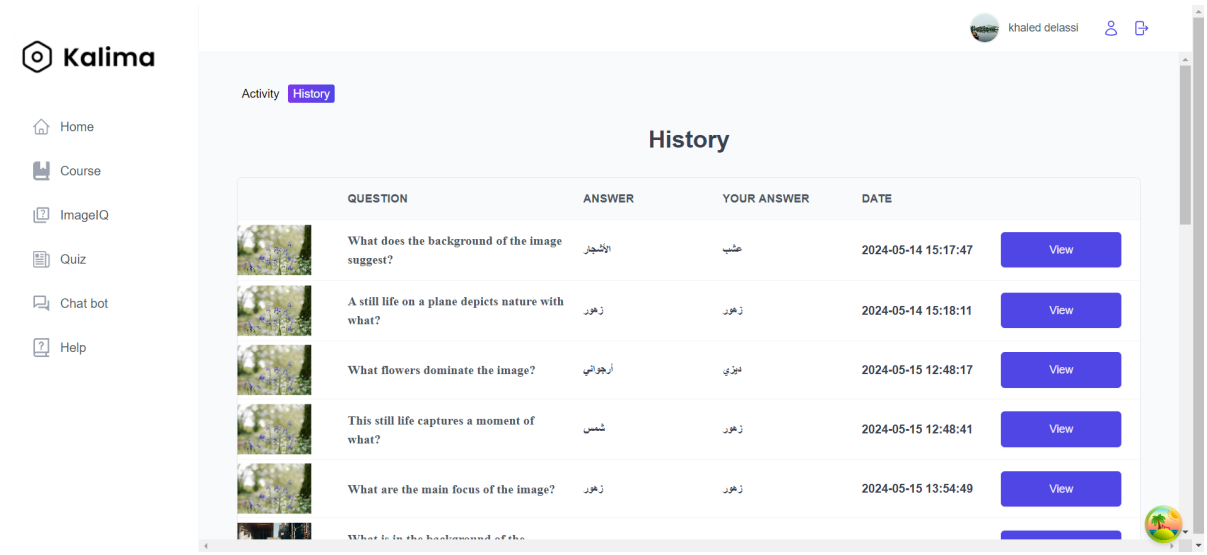

Figure 4.8: History and Activity Reporting for ImageIQ page

4.3.4 Quiz

The Quiz page is an engaging feature where users can start a timed quiz by clicking the "Start" button in Figure 4.9.

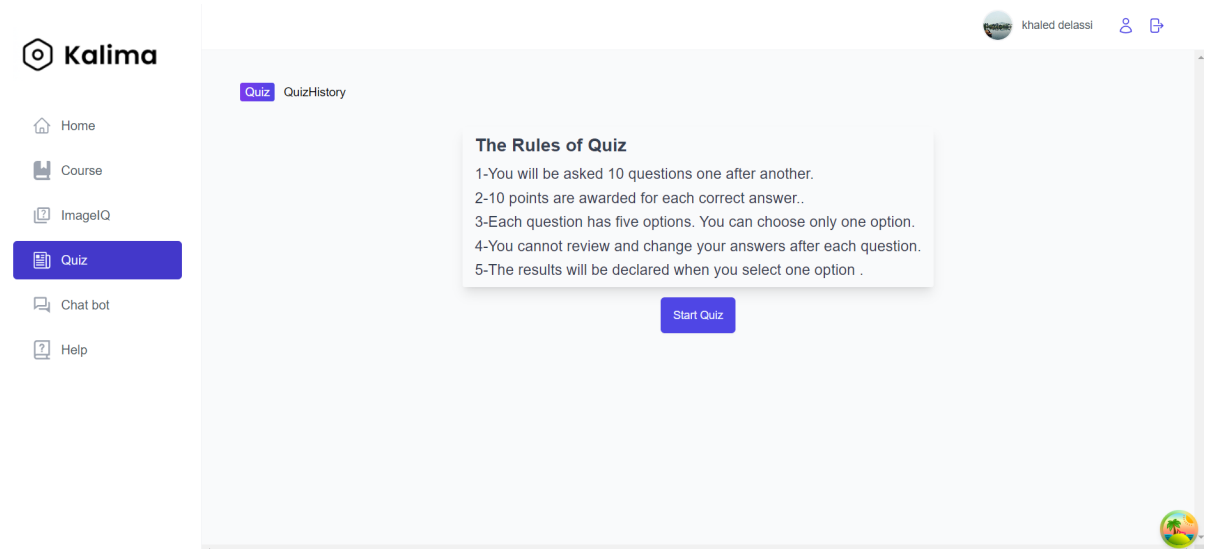

Figure 4.9: Quiz page in Kalima web application

Once the game begins, a 10-minute countdown starts and the system presents a random image along with a question and five possible answers in Arabic, similar to the ImageIQ page. Upon answering, the system immediately provides a new image, question, and set of answers, continuing this process for a total of 10 questions you can see more details in Figure 4.10.

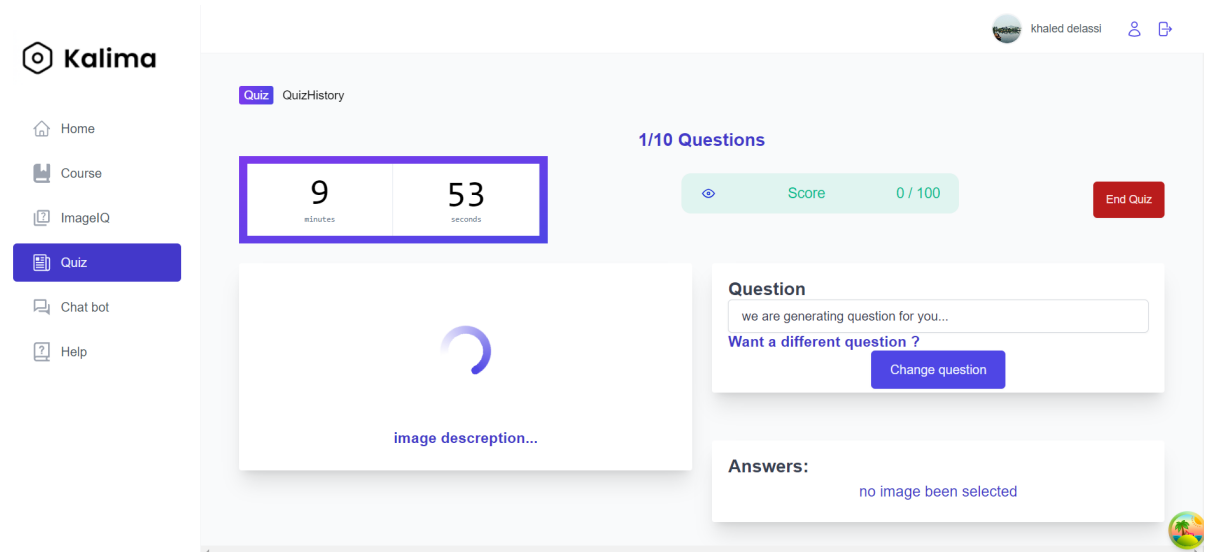

Figure 4.10: Quiz round page in Kalima web application

If the user gets all the questions right like in Figure 4.11 in the time provided for answering the questions, the user passes the quiz. But if the case timer gets to **00 :00** before all questions are answered, then that game is over and the quiz stopped. At the end of the quiz or when time runs out, the system provides a mark summarizing the user's performance in Figure 4.12. This dynamic setup tests users' quick thinking and comprehension skills, providing a fun and educational challenge.

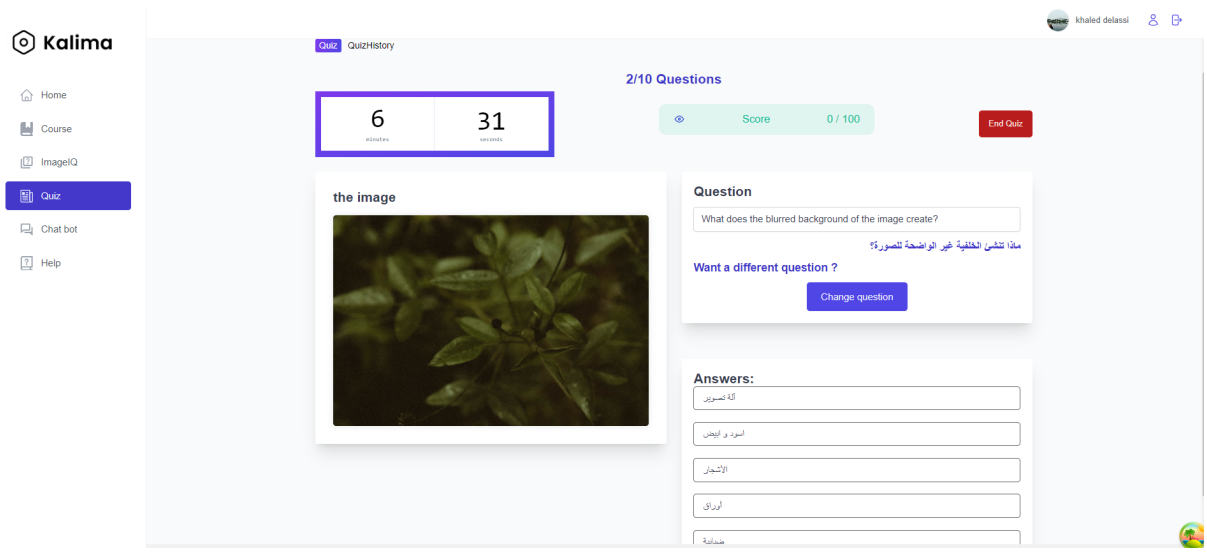

Figure 4.11: Illustration Quiz round page

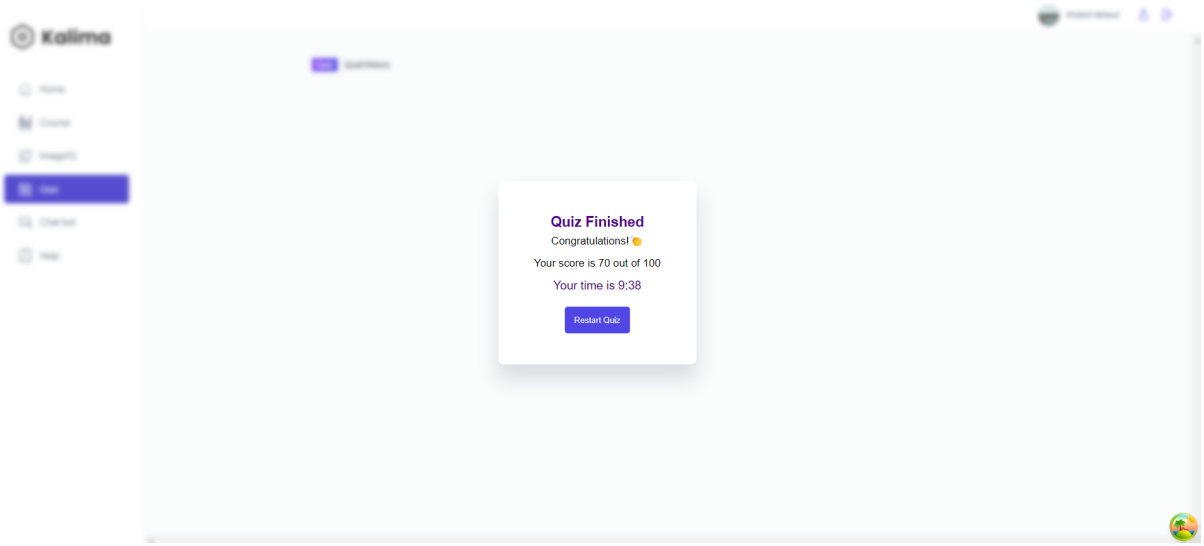

Figure 4.12: the end of Quiz

Additionally, the Quiz page includes a "History" shown in Figure 4.13, where users can view all their previous quizzes and review the questions and answers in Figure 4.14. This means that one can be aware of his or her progress, and weaknesses, and make the learning experience more comprehensive and effective.

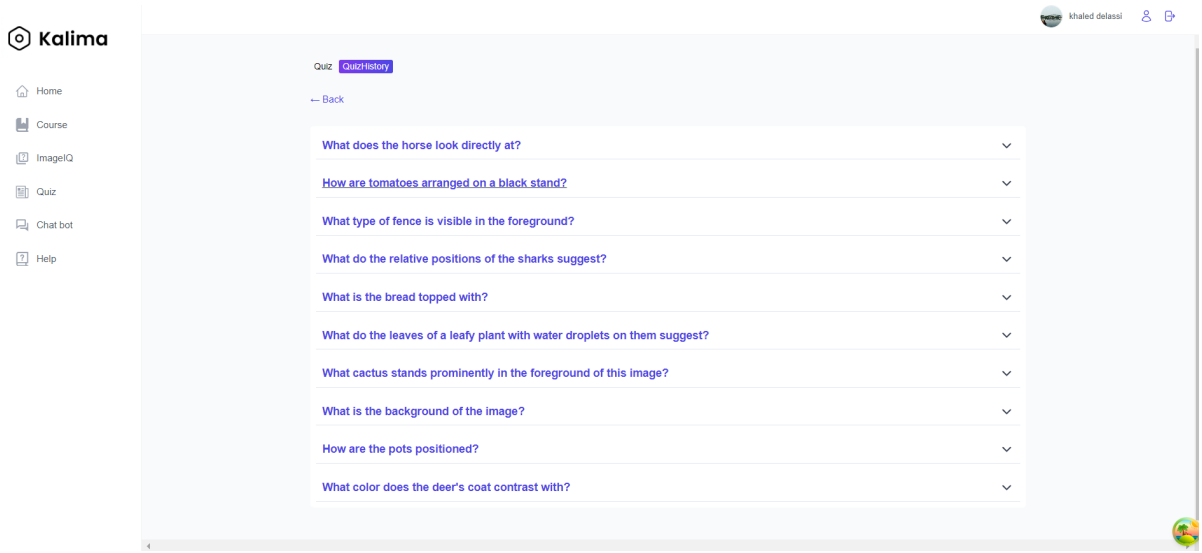

Figure 4.13: Quiz history page in Kalima web application

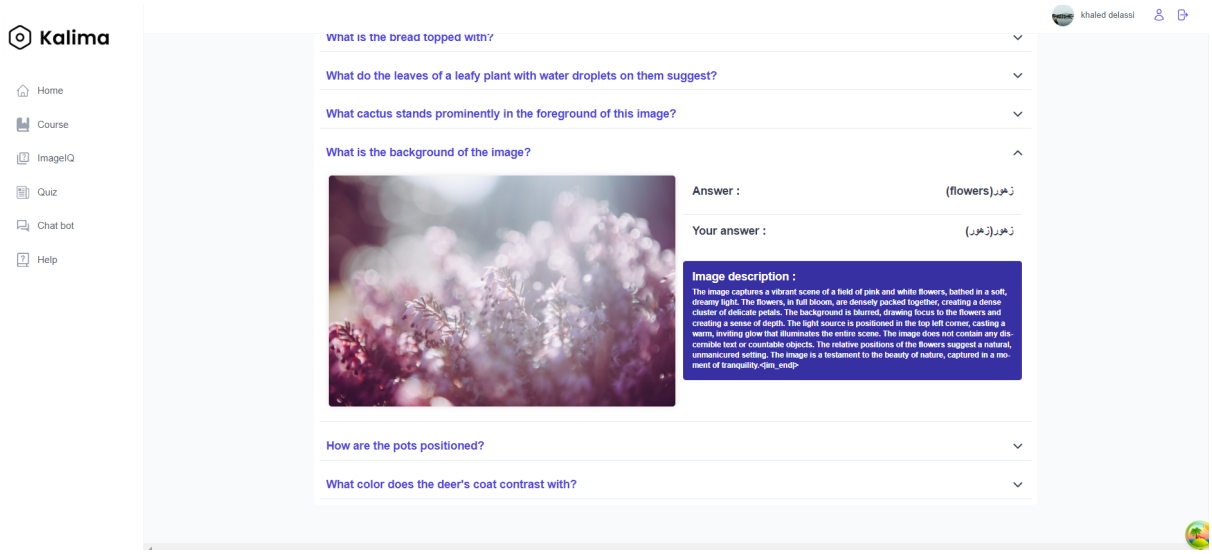

Figure 4.14: Quiz round history details page in Kalima web application

4.3.5 Profil of User

The User Profile page provides a personalized space where users can manage and update their account information you can see it in Figure 4.15. On this page, users can modify their email address and name, change their profile picture, and update their password to ensure their account remains secure.

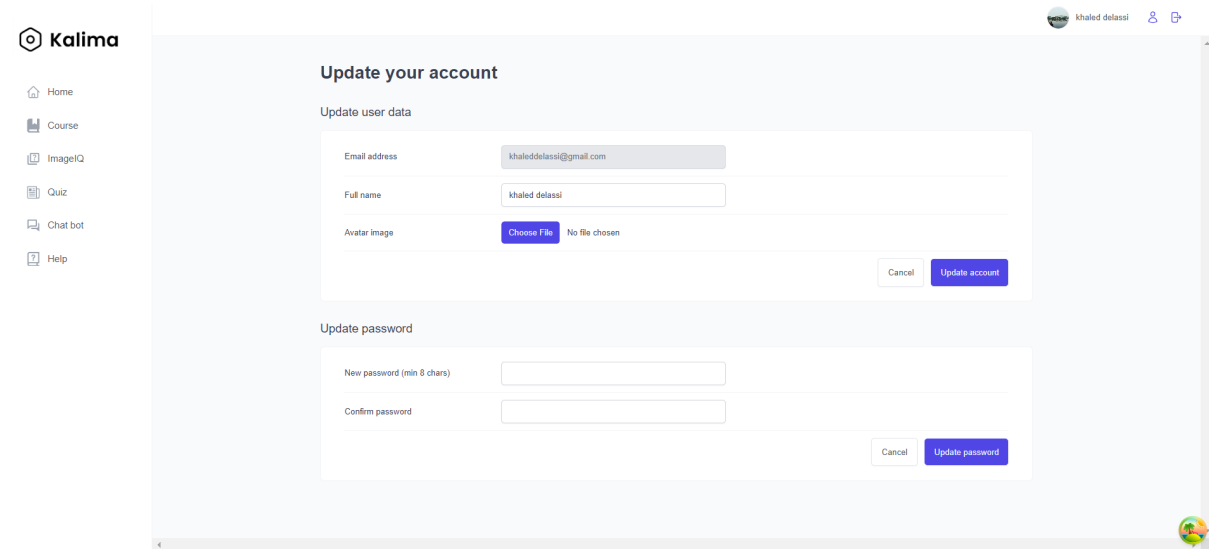

Figure 4.15: Profile page in Kalima web application

4.4 Back-end Services

The back-end for the Visual Question Answering (VQA) web application architecture, as depicted in Figure 4.1, is an essential component that manages data storage and pro-

cessing. we used in our development a set of languages tools and DBMS reported in Table 4.3 :

Backend	**Node.js**	A JavaScript runtime for building fast, scalable server-side applications.
	Express.js	A minimal and flexible Node.js web application framework.
	Sequelize	A promise-based ORM for Node.js, simplifying database interactions.
	MySQL	MySQL is an open-source relational database management system.

Table 4.3: Deployed Tools for the back-end development

The backend is implemented using Node.js and Sequelize ORM to interact with a MySQL database to store and retrieve users, images, questions, and answers.

In this pedagogical tool, the backend receives requests from the React frontend application, which include images, questions, and corresponding answers.

4.5 AI models backend

For the environment, we used the tools described in Table 4.4.

Environment	**Miniconda**	A minimal installer for Conda, providing a lightweight way to install the Conda package manager and Python, allowing for easy management of environments and packages.
	d2l package	an educational deep learning framework that simplifies learning .
AI Backend	**FastAPI**	A modern, high-performance Python framework for building APIs.

Table 4.4: Environment for AI Development

The AI backend for the pedagogical tool web-based architecture consists of three primary models, each serving a distinct function.

- The first model, we used **UForm-Gen2-dpo** [45] is a small but highly effective generative vision-language model dichotomized for tasks such as Image Captioning and VQA. This particular model is efficiently trained on preference datasets such as the VLFeedback and LLaVA-Human-Preference-10K which help improve the quality and relevance of the generated outputs based on the human preference [45]. This

model is dedicated to image description, generating detailed textual descriptions from input images.

- The second model, **potsawee/t5-large-generation-squad-QuestionAnswer** [46]. The "t5-large fine-tuned to SQuAD for Generating Question+Answer" refers to a variant of the T5 (Text-To-Text Transfer Transformer) model, specifically adapted through fine-tuning on the SQuAD (Stanford Question Answering Dataset). This configuration enables the model to input textual contexts, such as news articles, and generate corresponding questions followed by extractive answers. This model focuses on question generation, formulating relevant questions based on the provided images and their descriptions.

- The third model is a combination of :

  - **Salesforce/blip-image-captioning-large** [47] which is a VLP framework that can be flexibly applied to both vision language understanding and generation tasks. BLIP effectively exploits the noisy web data by bootstrapping the captioning, where a captioner generates synthetic captions and a filter removes the noisy ones. It achieves state-of-the-art results on a variety of visual language tasks, including picture-text retrieval. BLIP also shows strong generalization when applied directly to video language tasks in a zero-shot fashion.
  - **dandelin/vilt-b32-finetuned-vqa** [48] which is a Vision-and-Language Transformer (ViLT) model that has been fine-tuned on the VQAv2 dataset for visual question answering (VQA). It is a powerful and versatile model that can be used to answer a wide range of questions about images, from simple factual queries to more complex open-ended questions. The dandelin/vilt-b32-finetuned-vqa is the state-of-the-art Model. This model is responsible for visual question answering. This combination processes the image and the user's question to provide accurate and contextually appropriate answers, leveraging the strengths of both models for enhanced performance.

These models are accessible via the FastAPI framework, which facilitates seamless communication between the front-end and the back-end. For instance, to obtain an image description, a client can make an HTTP request to the endpoint description_image. This request leads to the description_image(image_path), where image_path concerns the location of the image to be described. The function processes the image through the description model and returns a textual description.

4.6 Pedagogical Tool Evaluation

This section presents an evaluation of our application's performance, accuracy, and usability. We use various metrics and benchmarks to analyze its effectiveness.

4.6.1 Experiment

We conducted a thorough evaluation of our application using three participants test subjects : participant A (age 22), participant B (age 24), and participant C (age 45). Each participant answered a total of 50 questions, divided into 5 quizzes with 10 questions each. The results were categorized into two cases for analysis :

- **Case 1** : Both the human and the Pedagogical Tool are correct.
- **Case 2** : The human is correct, but the Pedagogical Tool is not.

The accuracy statistics for each participant are summarized in the table 4.5 below :

Case	Human A	Human B	Human C	Average
Case 1 : Human and Pedagogical Tool Correct	76%	77%	88%	80.33%
Case 2 : Human Correct, Pedagogical Tool Incorrect	24%	23%	12%	19.66%

Table 4.5: Accuracy Statistics of Human and Pedagogical Tool Performance

4.6.2 Discussion

The evaluation of our pedagogical tool using three participants across various age groups provides valuable insights into its performance, accuracy, and usability in a pedagogical context.

Case 1 Analysis : The results indicate that the pedagogical tool and human participants collectively achieved high accuracy rates in Case 1, where both correctly answered the visual questions. Participant C, in particular, demonstrated the highest combined accuracy of 88%, suggesting strong alignment between pedagogical tool responses and human judgments. This outcome underscores the effectiveness of our pedagogical tool in providing accurate and reliable answers that align with human understanding.

Case 2 Analysis : In Case 2, where the human participants answered correctly while the application did not, the accuracy rates varied but remained noticeably lower than in Case 1. Participant A had the highest discrepancy, with 24% of questions answered correctly by the human but incorrectly by the pedagogical tool. This discrepancy highlights

areas where the pedagogical tool may struggle, such as nuanced or ambiguous questions or the complex questions generated with our model requiring deeper contextual understanding beyond visual recognition. the pedagogical tool may also struggle in cases where the image is too complex.

4.7 Conclusion

This chapter detailed our pedagogical tool web application architecture, including a user-friendly front-end with features like sign-up, dashboard, and ImageIQ for visual question answering. The back-end ensures data integrity, while integrated AI models enhance functionality. Evaluation highlighted strong accuracy and usability for Users.

Conclusion and Perspectives

Despite its importance Arabic language is badly served in terms of pedagogical tools that eases its learning as asecond language (L2).

Aiming to fill this gap, we have proposed a Pedagogical AI/Web-based tool for learning Arabic Language within the realm of e-learning and distant learning. Our pedagogical tool advocates the constructivism learning model that helps learners enhance their comprehension experience by integrating visual content with intelligent questioning systems through a question-generator model. This pedagogical tool is based on the idea that people actively construct or make their own knowledge.

Our tool relies on two components. The first is a VQG system which is designed to generate relevant questions based on the visual content provided. the second one is VQA focuses on providing accurate answers to the questions generated by the VQG system.

Before delving into our approaches, we provided an introduction highlighting the effective ways to learn a language and a glance at real-world apps that are exciting in the market, then NLP and its usage in the education field.

At the end, we presented the ARABIC-EDU-VQA tool design, highlighting VQG and VQA models and a solution to deal with Arabic using a translation model. Lastly, we presented the implementation of our tool, covering its web-based architecture, the front-end of ImageIQ, and quizzes. The back-end infrastructure, AI model integration, and evaluation methodologies are meticulously described through experimental setups and insightful discussions. The results show that the performances of our AI-based pedagogical tool is very suitable with an accuracy near of 80.33% compared to human evaluation.

Future Developments

The developed tool is the first version of an AI-based system designed to enhance Arabic language learning. Several areas for future development have been identified to enhance the capabilities of future versions of our pedagogical tool :

- **Dataset :** For further improvement of our pedagogical tool we need to create an Arabic dataset dedicated to education and learning the Arabic language and fine-tunning our system ;

- **Improved Accuracy :** To further refine our pedagogical tool, we need to improve the VQG system to get accurate questions to help language learners enrich their vocabulary. We, also, need to strengthen the VQA system to be able to answer questions about complex images ;
- **Advanced Features :** Incorporating features such as real-time feedback, adaptive learning pathways based on VQA interactions, and deeper contextual understanding can significantly improve the VQA system ;
- **Integration of other Technologies :** Integrating VQA with other educational technologies, such as augmented reality (AR) and virtual reality (VR), can create immersive learning environments that enhance student engagement and learning outcomes.

Incorporating these perspectives into future work can make learning languages especially Arabic more interactive, accessible, and engaging. Continued research and development will be essential to fully harness the power of AI in educational settings and address the challenges identified in our work.

Bibliographie

[1] Sruthy Manmadhan and Binsu C. Kovoor. Visual question answering : a state-of-the-art review. *Springer Nature*, 2020.

[2] S. E. Spataro and J. Bloch. "can you repeat that ?" teaching active listening in management education. *Journal of Management Education*, 42(2) :168–198, 2018.

[3] Kenneth Beare. English pronunciation practice. https://www.thoughtco.com/english-pronunciation-practice-1212076, 2024. Accessed : 2024-06-07.

[4] D. M. Murray. *Write to Learn*. Cengage Learning, 2009.

[5] Feedback and revision in second language writing : Contextual, teacher, and student variables. *Unknown Journal*, August 2006.

[6] Learning English with Oxford. Reading techniques for english learners. https://learningenglishwithoxford.com/2023/08/31/reading-techniques-for-english-learners/, 2023. Accessed : 2024-06-07.

[7] Cultural Vistas. How to start speaking another language. https://culturalvistas.org/impact-learning/news-stories/how-to-start-speaking-another-language, 2024. Accessed : 2024-06-07.

[8] Tanmayi Nagale and Anand Khandare. Enhancing interactive learning : A comparative survey on ict tools and pedagogical techniques for student engagement. *Educational Administration : Theory and Practice*, 30(4) :6914–6919, 2024.

[9] Morocco World News. Arabic fifth most spoken language in the world. https://www.moroccoworldnews.com/2021/08/343835/arabic-fifth-most-spoken-language-in-the-world, 2021. Accessed : 2024-06-07.

[10] Talkpal. Set realistic language learning goals with these tips. https://talkpal.ai/set-realistic-language-learning-goals-with-these-tips/, 2024. Accessed : 2024-06-07.

[11] Together Platform. Short-term and long-term goals examples : A goal setting guide. https://www.togetherplatform.com/blog/short-term-and-long-term-goals-examples-a-goal-setting-guide, 2024. Accessed : 2024-06-07.

[12] Eliza Comodromos and Paul Langan. *Mastering Vocabulary Skills*. Townsend Press, 2019. Reading level : 12+.

[13] Arina Kravchenko. Typical grammar structures for each level. https://grade-university.com/blog/typical-grammar-structures-for-each-level, 2024. Accessed : 2024-06-07.

[14] Nazlı GÜNDÜZ. Contributions of e-audiobooks and podcast to efl listening classes. *Selçuk Üniversitesi Edebiyat Fakültesi Dergisi*, (21) :249–259, 2006.

[15] Language Mentoring. How to use subtitles for language learning. https://www.languagementoring.com/subtitles/, 2024. Accessed : 2024-06-07.

[16] Rikke L Bundgaard-Nielsen, Catherine T Best, Christian Kroos, and Michael D Tyler. Second language learners' vocabulary expansion is associated with improved second language vowel intelligibility. *Applied Psycholinguistics*, 33(3) :643–664, 2012.

[17] EnglishRadar. English learning tips : Advantages of reading aloud. https://www.englishradar.com/study-ideas/english-learning-tips-advantages-of-reading-aloud/, 2024. Accessed : 2024-06-07.

[18] J. C. Richards. *Communicative Language Teaching Today*. Cambridge University Press, 2006.

[19] J. Harmer. *How to Teach Writing*. Longman, 2004.

[20] Le Journal du Net. Natural language processing (nlp), n.d. Accessed : 2024-06-06.

[21] N. Baker Gillis. Sexism in the judiciary : The importance of bias definition in nlp and in our courts. In M. Costa-jussa, H. Gonen, C. Hardmeier, and K. Webster, editors, *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 45–54, Online, 2021. Association for Computational Linguistics.

[22] What are the most effective ways to use NLP in education ? — linkedin.com. https://www.linkedin.com/advice/0/what-most-effective-ways-use-nlp-education-jtijf#:~:text=One%20of%20the%20most%20effective,needs%2C%20strengths%2C%20and%20weaknesses. [Accessed 08-06-2024].

[23] Machine Learning : Bridging Between Business and Data Science — altexsoft.com. https://www.altexsoft.com/whitepapers/machine-learning-bridging-between-business-and-data-science/. [Accessed 08-06-2024].

[24] Language Models Explained — altexsoft.com. https://www.altexsoft.com/blog/language-models-gpt/. [Accessed 08-06-2024].

[25] Sudheer Madala. Evaluating Language Models in NLP - Scaler Topics — scaler.com. https://www.scaler.com/topics/nlp/language-models-in-nlp/. [Accessed 09-06-2024].

[26] Antol, S. and Agrawal, A. and Lu, J. and Mitchell, M. and Batra, D. and Lawrence Zitnick, C. and Parikh, D. Vqa : Visual question answering. *The Journal*, Vol(Num) :552–574, 2015.

[27] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome : Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1) :32–73, 2017.

[28] Peter Clark. Vanilla VQA, 2021. Accessed : 2024-06-07.

[29] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa : Visual question answering. *arXiv preprint arXiv :1612.00837*, 2016.

[30] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. Visual dialog, 2017. Accessed : 2024-06-07.

[31] Ruqing Zhang, Jiafeng Guo, Lu Chen, Yixing Fan, and Xueqi Cheng. A review on question generation from natural language text. 40(1), sep 2021.

[32] Shane Settle and Karen Livescu. Discriminative acoustic word embeddings : Tecurrent neural network-based approaches. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 503–510, 2016.

[33] Yoav Goldberg and Omer Levy. Word2Vec Explained : deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv :1402.3722*, 2014.

[34] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. Visual question answering : A survey of methods and datasets. *Computer Vision and Image Understanding*, 163 :21–40, 2017.

[35] Md Farhan Ishmam, Md Sakib Hossain Shovon, MF Mridha, and Nilanjan Dey. From image to language : A critical analysis of visual question answering (vqa) approaches, challenges, and opportunities. *Information Fusion*, page 102270, 2024.

[36] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words : Transformers for image recognition at scale. *arXiv preprint arXiv :2010.11929*, 2020.

[37] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[38] Fei-Long Chen, Du-Zhen Zhang, Ming-Lun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. Vlp : A survey on vision-language pre-training. *Machine Intelligence Research*, 20(1) :38–56, 2023.

[39] Bin He, Meng Xia, Xinguo Yu, Pengpeng Jian, Hao Meng, and Zhanwen Chen. An educational robot system of visual question answering for preschoolers. In *2017 2nd international conference on robotics and automation engineering (ICRAE)*, pages 441–445. IEEE, 2017.

[40] J Jinu Sophia and T Prem Jacob. Edubot-a chatbot for education in covid-19 pandemic and vqabot comparison. In *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pages 1707–1714. IEEE, 2021.

[41] Gyeong-Geon Lee and Xiaoming Zhai. Realizing visual question answering for education : Gpt-4v as a multimodal ai. *arXiv preprint arXiv :2405.07163*, 2024.

[42] Charulata Patil and Manasi Patwardhan. Visual question generation : The state of the art. *ACM Computing Surveys (CSUR)*, 53(3) :1–22, 2020.

[43] Sarah kamel, Shimaa Hassan, and Lamiaa Elrefaei. Vaqa : Visual arabic question answering. *Arabian Journal for Science and Engineering*, 48, 03 2023.

[44] Aidan Welch. google-translate-api. https://github.com/AidanWelch/google-translate-api#readme, 2024. Accessed : 2024-05-26.

[45] Unum Cloud. Uform-gen2-dpo. https://huggingface.co/unum-cloud/uform-gen2-dpo, 2024. Accessed : 2024-05-26.

[46] Potsawee Manakul. t5-large-generation-squad-questionanswer. https://huggingface.co/potsawee/t5-large-generation-squad-QuestionAnswer, 2023. Accessed : 2024-05-26.

[47] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip : Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022.

[48] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt : Vision-and-language transformer without convolution or region supervision, 2021.