GRADUATION PROJECT

TEAM MEMBERS

Abdelmanem Mohamed
Mohamed Mahmoud
Salema Essam
Dina Omar
Aya Hani

TABLE OF CONTENTS

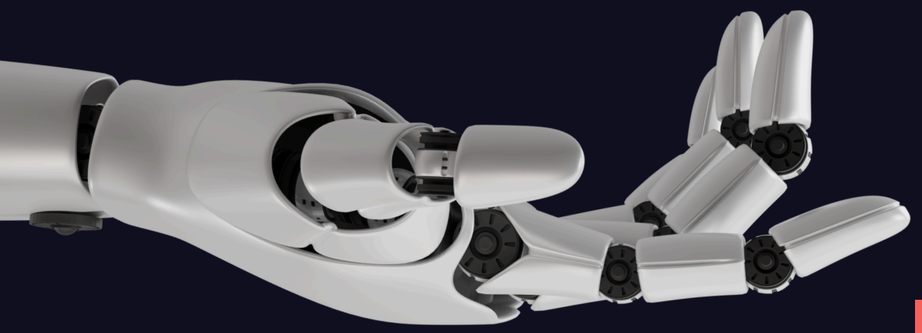
PROJECT OVERVIEW

- Interactive AI pipeline to generate Arabic captions and questions from user-uploaded images.
- Integrates multiple models for image captioning, question generation, translation, text-to-speech, and image manipulation.
- Provides a seamless user experience through a Gradio interface for image upload, caption generation, and Arabic question answering.

Models

- **Translation:** facebook/mbart-large-50-many-to-many-mmt
- **Image captioning:** Salesforce/blip-image-captioning-large
- **Visual question answering:** Salesforce/blip-vqa-base
- **Question generation:** microsoft/Phi-3.5-mini-instruct

Pipeline WorkFlow



STEP 06
VQA

Answer question That was asked by the user

STEP 05
Question input

Suggest some question on the user about image or let user ask question he want about image

STEP 04
Text to speech

Converts the Arabic text to speech

STEP 03
Translation

Translate generated caption into Arabic

STEP 02
Image captioning

Generate an English caption that describe the image

STEP 01
User Input

User Uploads Image

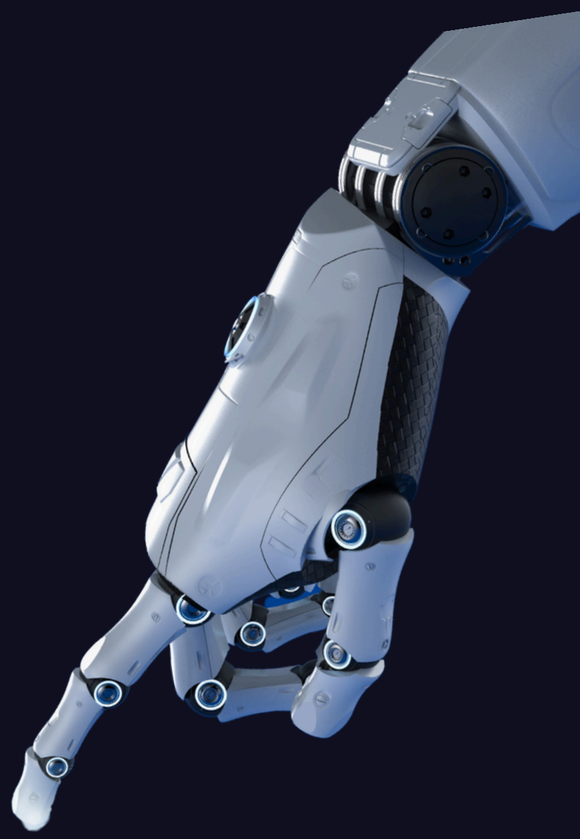
RESULTS

01

Automated pipeline that generates image captions in Arabic and delivers both text and audio outputs and answer any question about given image

02

Modularity and Scalability: The project architecture is modular, allowing for easy future integration of additional languages, models, or features such as further image modifications or customization options.



CHALLENGES

High Computational Requirements for Model loading & Inference

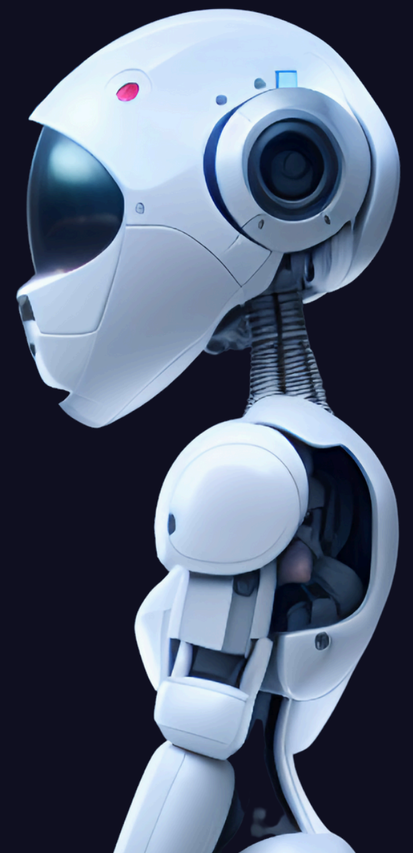
Used cloud services and tried to find small models that doesn’t need much computation power and same time doesn’t sacrifice accuracy that much

Generating Arabic Speech from Captions

Most of model in this task still struggle to output a natural and correct speech of the text so we used the gtts library which let us interact with google text to speech API

Handling Translations Between Arabic and English

As speech to text most open source model is not the best and good LLMs are above our resource limit so we search multiple model until we found mbart by Facebook which provide us a bidirectional translation which makes our pipeline simpler and open chances for supporting more lang in futrue



FUTURE IMPROVMENT

USING VLLM

By using a state of the are vison language model this will improve the understanding of the image by the model making us able to give better image description and better answers for asked question about image

INTEGRATING STABLE DIFFUSION

We can add stable diffusion to our pipeline which will help in giving user to recreate his image in a different environment of with any changes he need

SUPPORT MORE LANGUUAGES

We can add more languages as (Spanish,French,German,etc..) so that our pipeline is easy to use to most of the people and overcome any language barrier

ALLOW DIFFRENT INPUT OPTIONS

So our model need an image and after that user can interact with the model and ask question about image here we will add option to ask model using voice command so it make the pipeline more dynamic


THANK
YOU!
Project on GitHub