Towards an AI to Win Ghana's National Science and Maths Quiz

George Boateng [1, 2] Jonathan Abrefa Mensah [2] Kevin Takyi Yeboah [2]
William Edor [2] Andrew Kojo Mensah-Onumah [2] Naafi Dasana Ibrahim [2]
Nana Sam Yeboah [2]

[1]ETH Zurich [2]NSMQ AI, Kwame AI Inc.

Background

- There is a lack of enough qualified teachers in Sub-Saharan Africa.
- Only **65%** of Sub-Saharan Africa teachers have the minimum required skills.
- 15 million more teachers are needed to reach UNESCO goals by 2030.

Motivation and Goal

Motivation: Augment limited number of qualified teachers in Africa by creating an AI teaching assistant for teachers.

Goal: Develop **NSMQ AI**, an AI to compete live and win in Ghana's National Science and Maths Quiz (NSMQ).

Why NSMQ?

- NSMQ is a live contest that covers Biology, Physics, Chemistry and Maths.
- Provides a robust benchmark for an AI teaching assistant in an African context

Technical Challenges: The project involves complex tasks such as speech-to-text, text-to-speech, question-answering, and human-computer interaction.

NSMQ AI Teams: The project has 6 teams that work together collaboratively towards accomplishing the goal: Data Curation, Data Preprocessing, Web App, Speech-to-Text, Question Answering, and Text-to-Speech. Focus on Riddles round for debut in October 2023 ahead of NSMQ 2023.

Data Curation

- Generates Google sheet with links to YouTube / Facebook videos of contests, and relevant information (e.g., year, contest, schools competing, total scores of each school, etc.).
- Ongoing efforts to annotate timestamps for clues in each riddle across contests

Challenges: Insufficient data sources online

Data Preprocessing

- Focuses on automatic cleaning, transforming, and preparing datasets for the NSMQ AI project.

Major Tasks

1. Automate NSMQ video downloads, delineating start and end of riddles and cropping of riddles.
2. Automate extraction of HTML of resource books and parse into JSON format.

Tech Stack

1. Python
2. Open AI Whisper
3. Google Colab

Web App

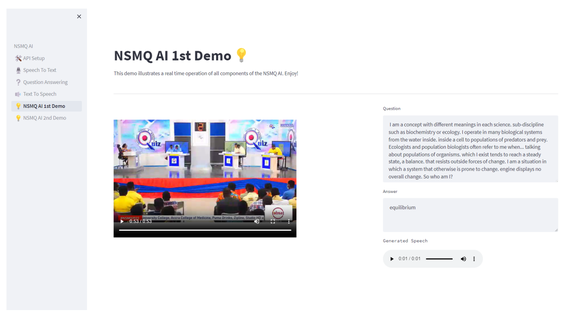

Figure 1. Screenshot of NSMQ AI web app

Tech stack

1. Front end
   - Python
   - Streamlit
2. Back end
   - FastAPI in Google Colab

- Aims to create an integrated app showcasing the machine learning subsystems of NSMQ AI.
- The app offers demo and live quiz modes, acting as the coordination layer for NSMQ AI.
- Facilitates user interactions and communicates with ML inference servers via API calls.

Speech-to-Text (STT)

Goal: Provide a robust, fast transcription for Ghanaian-accented English in mathematical and scientific contexts.

Model: Whisper

Datasets: 3 audios (each 15 secs long) and transcript from the NSMQ competition with Ghanaian accents.

Evaluation Metrics

- **WER**: Measures speech intelligibility
- **Latency**: Measures model inference speed

Table 1. Evaluation results from Pre-Trained Whisper models

Model size	Mean WER/%	Mean Latency/s
Whisper tiny	31.11	2.88
Whisper tiny.en	30.29	1.77
Whisper base	**29.70**	1.33
Whisper base.en	30.69	1.06
Whisper small	**29.51**	0.97
Whisper small.en	31.29	**0.88**
Whisper medium	30.92	**0.92**
Whisper medium.en	31.61	0.94
Whisper large	31.11	1.06

- Fine-tuning Whisper model on Ghanaian accented speech is needed to improve results.

Question Answering (QA)

Goal: Develop a QA system that answers NSMQ riddles swiftly and accurately

Dataset Preparation

- 1144 riddle-answer pairs, divided into train, test, and dev splits (60:20:20).
- CSV files contain "Clue 1" to "Clue 9" for riddle clues, "Answer" for the correct answer, and "Answer 1" to "Answer 4" for alternative answers.

Models

- Extractive Models:
  - DistilBERT
  - SciBERT
- Generative Model:
  - Falcon-7b-Instruct
  - GPT 3.5

- All clues were concatenated and used to generate top three contexts for extractive models. Clues were then passed as questions to the concatenated top three contexts.
- Concatenated clues were passed as inputs in a crafted prompt template to the generative models.

Evaluation Metrics

- **Exact Match (EM)**: For exact match with ground truth
- **Fuzzy Match (FM)**: For answer overlap with ground truth

Table 2. Evaluation results for QA Models

Model	Exact Match	Fuzzy Match
Falcon-7b-Instruct	**23.14%**	**35.81%**
DistilBERT	8.3%	14.85%
SciBERT	0.0%	6.11%
GPT-3.5	64.19%	75.54%

- Initial evaluation showed the generative model outperforming extractive models but underperforming ChatGPT.
- There's potential for performance improvement through fine-tuning the model on the dataset.

Text-to-Speech (TTS)

Goal: Develop a TTS system that can synthesize answers from a QA system into audible speech with a Ghanaian accent.

Models

- Baseline Models:
  - Glow TTS + Multiband Melgan
  - Tacotron 2 + Multiband Melgan
  - Fastpitch + Hifigan v2
- Fine-tuned Model: VITS

Datasets

- **George Boateng**: Podcasts presentations
- **Prof Elsie Kaufmann**: 20 minutes TEDx talk

Evaluation Metrics

- **WER**: Measures speech intelligibility
- **Automatic MOS**: Measures 'naturalness' of speech on a scale from 1 (bad) to 5 (excellent)
- **Latency**: Measures model inference speed

Table 3. Evaluation results for TTS system

	Evaluation 1			Evaluation 2		
Model	Mean latency / s	Mean WER (%)	Mean MOS	Mean latency / s	Mean WER (%)	Mean MOS
Fast pitch/Hifiganv2	0.46	**3.60**	2.98	0.45	**3.69**	2.98
Tacotron2/Multiband Melgan	0.45	4.14	2.99	0.54	4.14	2.99
Glow TTs/Multiban Melgan	**0.26**	6.32	2.99	**0.26**	6.32	2.99
Vits_Elsie	3.88	35.76	3.38	4.03	5.40	3.20
Vits_Jojo	5.15	88.47	**3.39**	5.26	63.07	**3.198**

- Evaluation 1: Synthesized 30 scientific and mathematical speeches from past NSMQ questions
- Evaluation 2: Synthesized 30 conversational speech
- Vits-Elsie model performs better in terms of WER compared to Vits-Jojo
- Overall, fine-tuned models achieve high MOS scores, suggesting human-like sound, but struggle with scientific and mathematical text synthesis

Next Steps

Data Curation

- Get audio segments of riddles and contestant performance for the last 5 years

Data Preprocessing

- Parse comprehensive corpus of science textbooks
- Refine HTML extraction validation process

Web App

- Include live mode with contest stream and AI attempts
- Make app accessible on cloud
- Support demo and live quiz modes with media data and user input

Speech-to-Text

- Automatically detect start of riddle reading
- Fine-tune model on past NSMQ audios and transcripts

Question Answering

- Improve model accuracy and latency
- Generate confidence scores during reading for informed early attempts

Text-to-Speech

- Enhance scientific and mathematical speech synthesis