# OPENAI/ WHISPER

## SPEECH-TO-TEXT GENERATIVE AI

PRESENTED BY YVES YANG

Illustration Artist: Ruby Chen

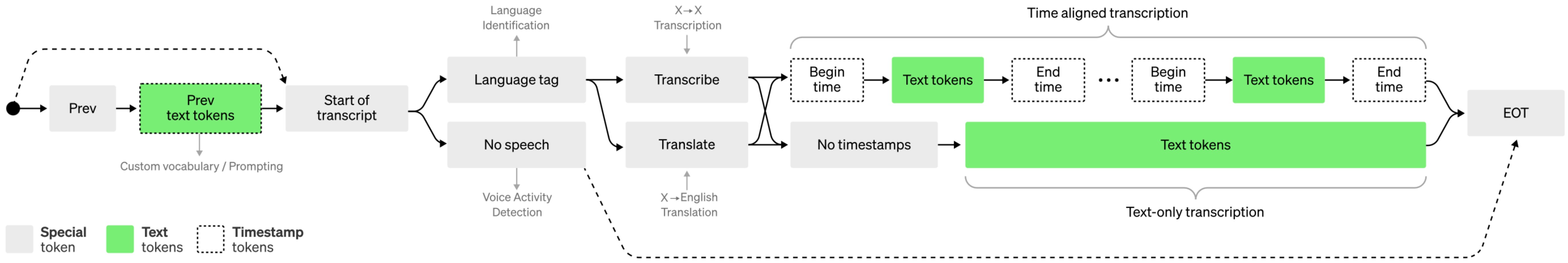
# INTRO

- Trained on 680,000 hours web data
- End-to-end approach, implemented as an encoder-decoder Transformer
    1. Input audio is split into 30-second chunks, converted, and then passed into an encoder
    2. A decoder is trained to predict the corresponding text caption
    3. Intermixed with special tokens that direct the model to perform tasks such as language identification, phrase-level timestamps, multilingual speech transcription, and to-English speech translation
- Differentiation: other existing approaches frequently use smaller, more closely paired audio-text training datasets or use broad but unsupervised audio pre-training

HTTPS://OPENAI.COM/RESEARCH/WHISPER

# DEMO

https://openai.com/research/whisper



Language Identification

X→X Transcription

Time aligned transcription

Prev

Prev text tokens

Start of transcript

Language tag

No speech

Transcribe

Translate

Begin time

Text tokens

End time

...

Begin time

Text tokens

End time

No timestamps

Text tokens

EOT

Custom vocabulary / Prompting

Voice Activity Detection

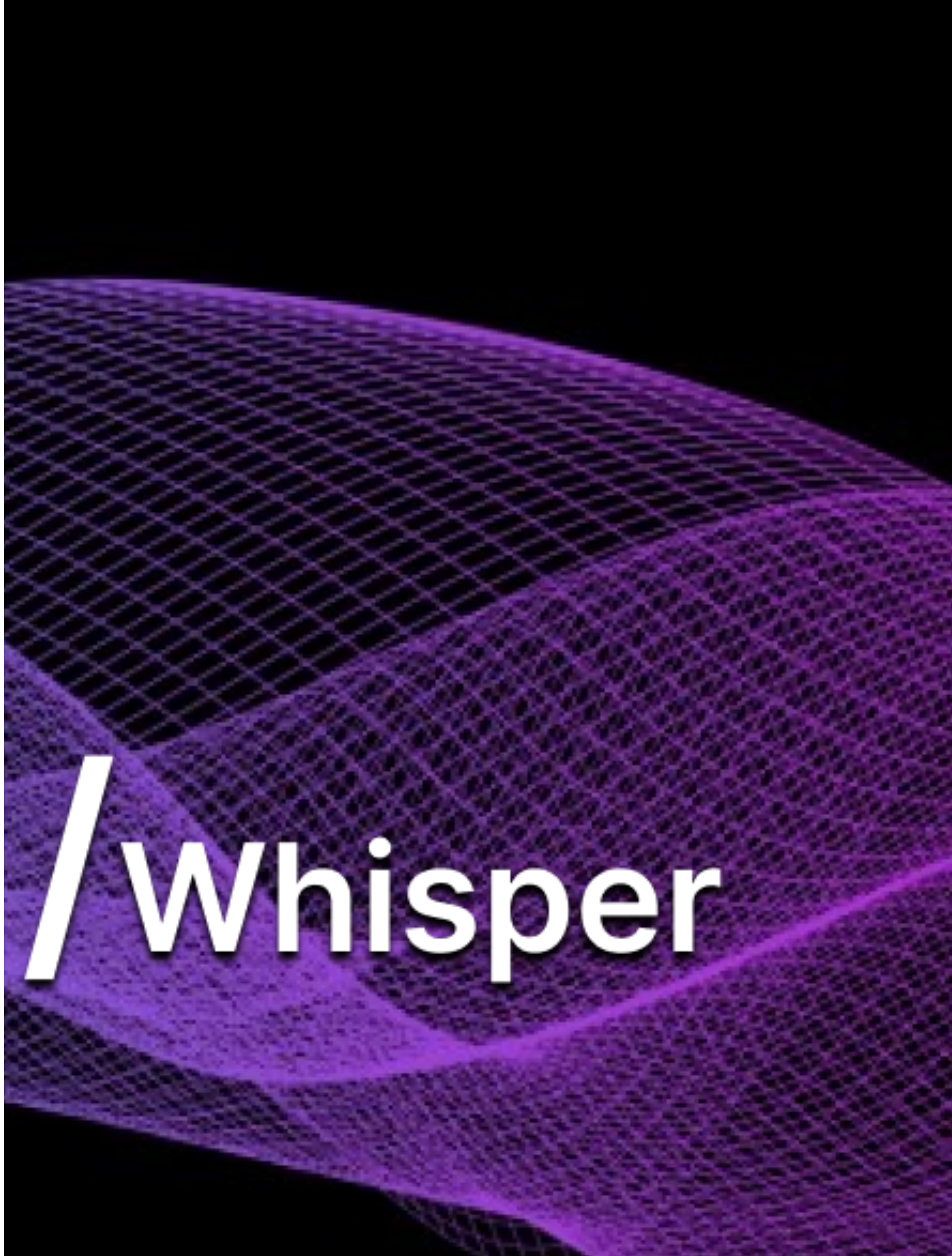
X→English Translation

Text-only transcription

⬜ Special token

🟩 Text tokens

⬚ Timestamp tokens



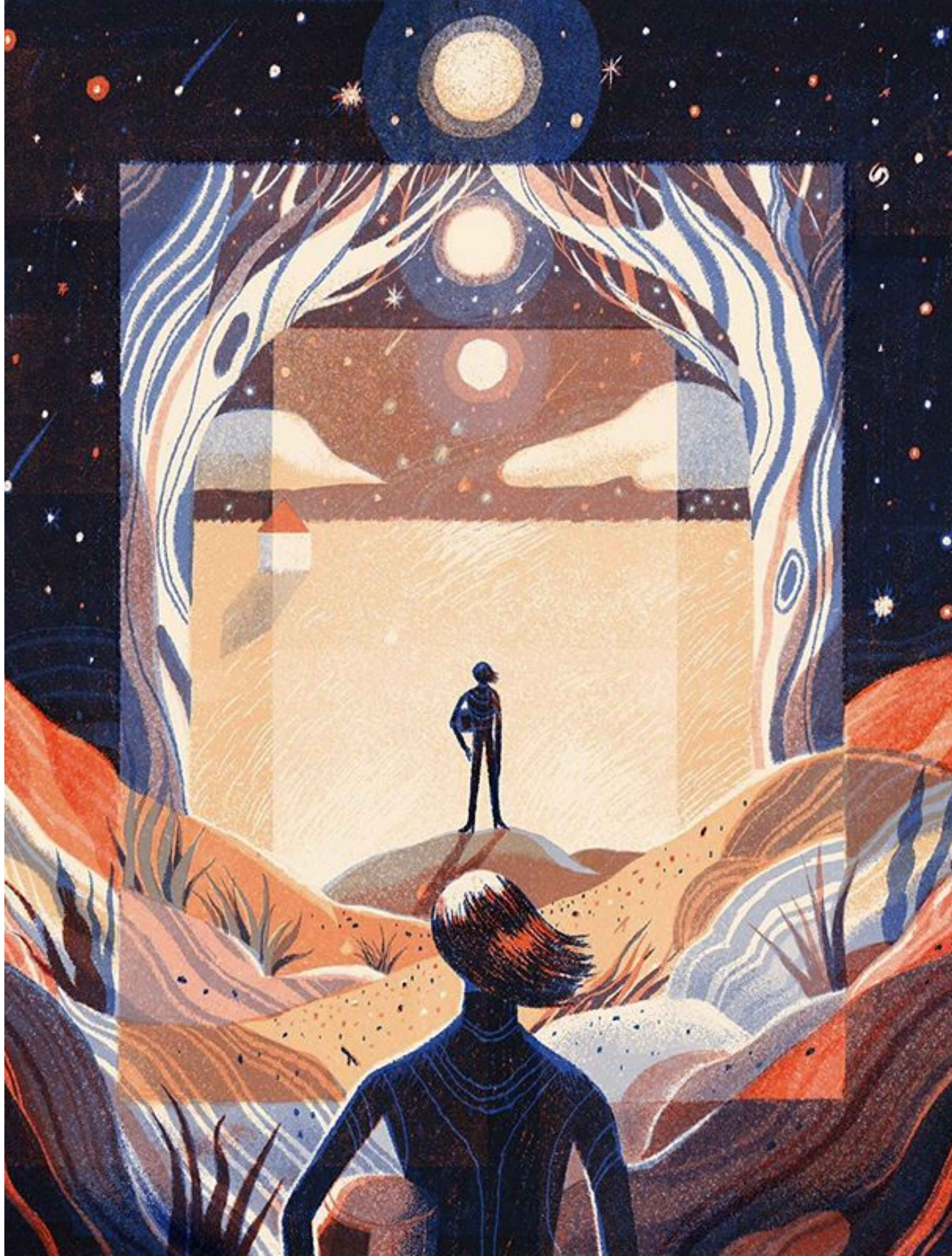
# TESTING

Recorded a paragraph from the book “The Three-Body Problem”

Around 100 words, 25 seconds

Audios in English, Spanish, Mandarin, and Chinese Dialect (all with and w/o background noise)

TESTED ON REPLICATE HTTPS://REPLICATE.COM/OPENAI/WHISPER



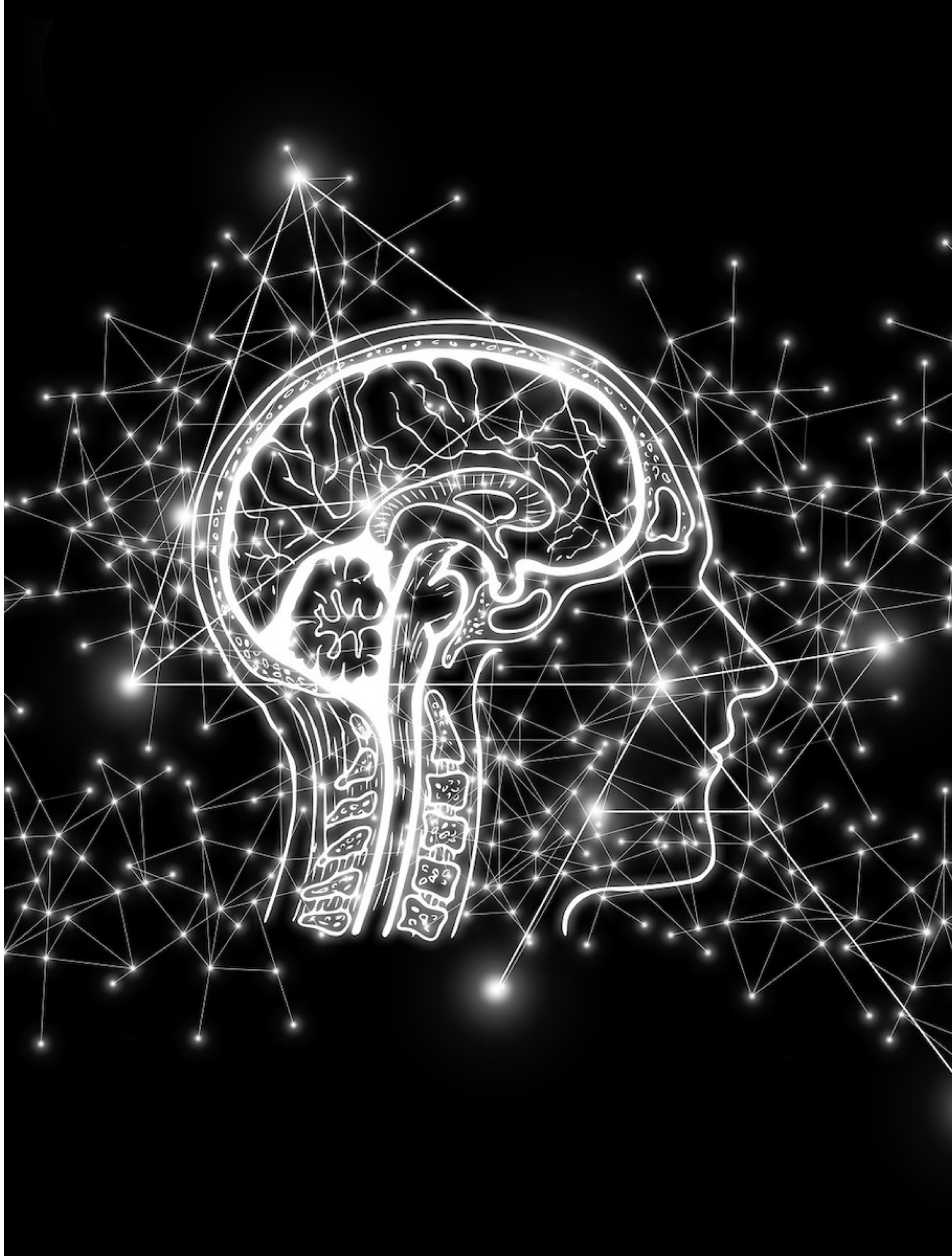
| Language | Content | Generated Output | Correctiveness | Time |
|---|---|---|---|---|
| English | The history of humanity is a history of the development of the understanding of intelligent life about itself. The development of human history has never been smooth. It is full of pain, detours, and sometimes it even progresses in the wrong direction. But the overall trend of the development of human history is a process of accumulating wisdom, of understanding oneself, and of getting closer and closer to the truth. | The history of humanity is a history of the development of the understanding of intelligent life about itself. The development of human history has never been smooth. It is full of pain, detours, and sometimes it even progresses in the wrong direction. But the overall trend of the development of human history is a process of accumulating wisdom, of understanding oneself, and of getting closer and closer to the truth. | **100%** | **5.11s** |
| English | Same but with Another Soundtrack of Noise | Same | **100%** | **5.93s** |
| Spanish | La historia de la humanidad es una historia del desarrollo del entendimiento de la vida inteligente sobre sí misma. El desarrollo de la historia humana nunca ha sido suave. Está lleno de dolor, desvíos y a veces incluso avanza en la dirección equivocada. Pero la tendencia general del desarrollo de la historia humana es un proceso de acumulación de sabiduría, de comprenderse a uno mismo y de acercarse cada vez más a la verdad. | La historia de la humanidad es una historia del desarrollo del entendimiento de la vida inteligente sobre sí misma. El desarrollo de la historia humana nunca ha sido suave. Está lleno de dolor, desvíos y a veces incluso avanza en la dirección equivocada. Pero la tendencia general del desarrollo de la historia humana es un proceso de acumulación de sabiduría, de comprenderse a uno mismo y de acercarse cada vez más a la verdad. | **100%** | **8.09s** |
| Spanish | Same but with Another Soundtrack of Noise | Same | **100%** | **8.23s** |
| Mandarin | 人类的历史就是一部发展中的智慧生命对自己的认识史。而人类历史的发展，从来都不是一帆风顺，它充满了痛苦，歧路重重，甚至有时候是错误的方向上前进。但人类历史发展的总趋势，却是积累智慧的过程，也是认清自己的过程，是越来越接近真理的过程。 | 人类的历史就是一部发展中的智慧生命对自己的认识史而人类历史的发展从来都不是一帆风顺它充满了痛苦,歧路重重,甚至有时候是错误的方向上前进但人类历史发展的总趋势却是积累智慧的过程,也是认清自己的过程,是越来越接近真理的过程 | **99.05%** | **9.95s** |
| Mandarin | Same but with Another Soundtrack of Noise | Same | **99.05%** | **11.13s** |
| Chinese Dialect | 人类的历史就是一部发展中的智慧生命对自己的认识史。而人类历史的发展，从来都不是一帆风顺，它充满了痛苦，歧路重重，甚至有时候是错误的方向上前进。但人类历史发展的总趋势，却是积累智慧的过程，也是认清自己的过程，是越来越接近真理的过程。 | 人类的历史就是一部法战中的智慧生命对自己的认识而人类历史的法战从来都不是一帆风顺,它充满了痛苦其路重重,甚至有时候是错误的方向上前进但人类历史法战的纵曲师却是积累智慧的过程也是人情子际的过程,是越来越洁净整理的过程 | **82.69%** | **9.62s** |
| Chinese Dialect | Same but with Another Soundtrack of Noise | 人类的历史就是一部法战中的智慧生命对自己的认识而人类历史的法战从来都不是一帆风顺,它充满了痛苦其路重重,甚至有时候是错误的方向上前进但人类历史法战的纵曲师却是积累智慧的过程也是人情子际的过程,是越来越洁净整理的过程 | **80.77%** | **10.03s** |

# ADVANTAGE

Adaptability: large and diverse training dataset leads to improved robustness to languages, accents and background noise

Accuracy: Whisper’s zero-shot performance is much more robust and makes 50% fewer errors than those models fine-tuned to any specific dataset.

Application: can be utilized in various areas including transcription services, voice assistants, humanoid robots…



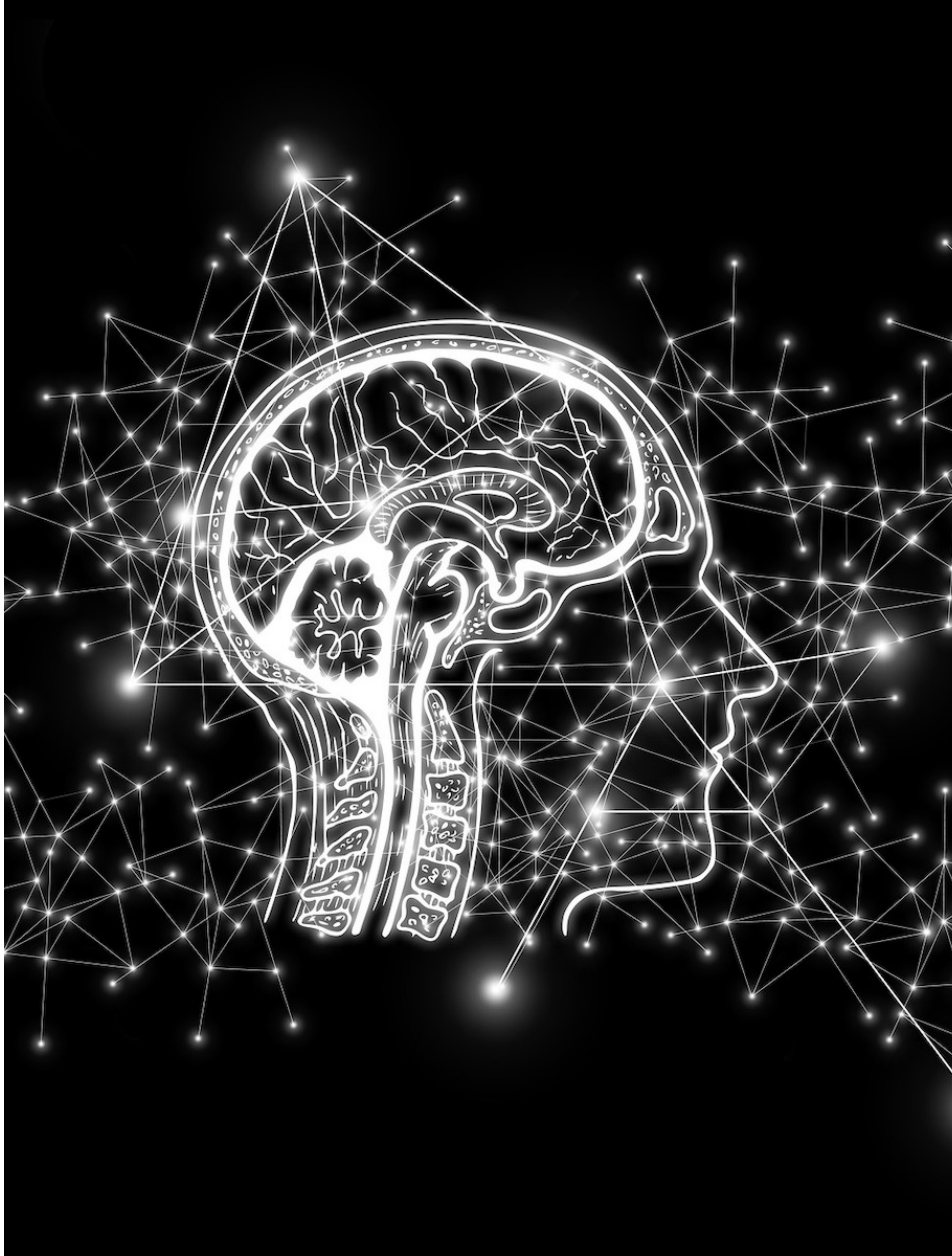
# DEFICIENCY

Efficiency: a 25 seconds audio file with around 100 words takes 8-10 seconds to process. (scalability? real-time?)

Punctuation: hight inaccuracy in punctuation especially when processing real voice recordings.

Data Privacy: audio file submission might raise concerns about user privacy and data protection if not processed locally or encrypted.

Data Quality: The performance could be significantly impacted by the quality of the audio input (such as noise, file compression, file format, file size).



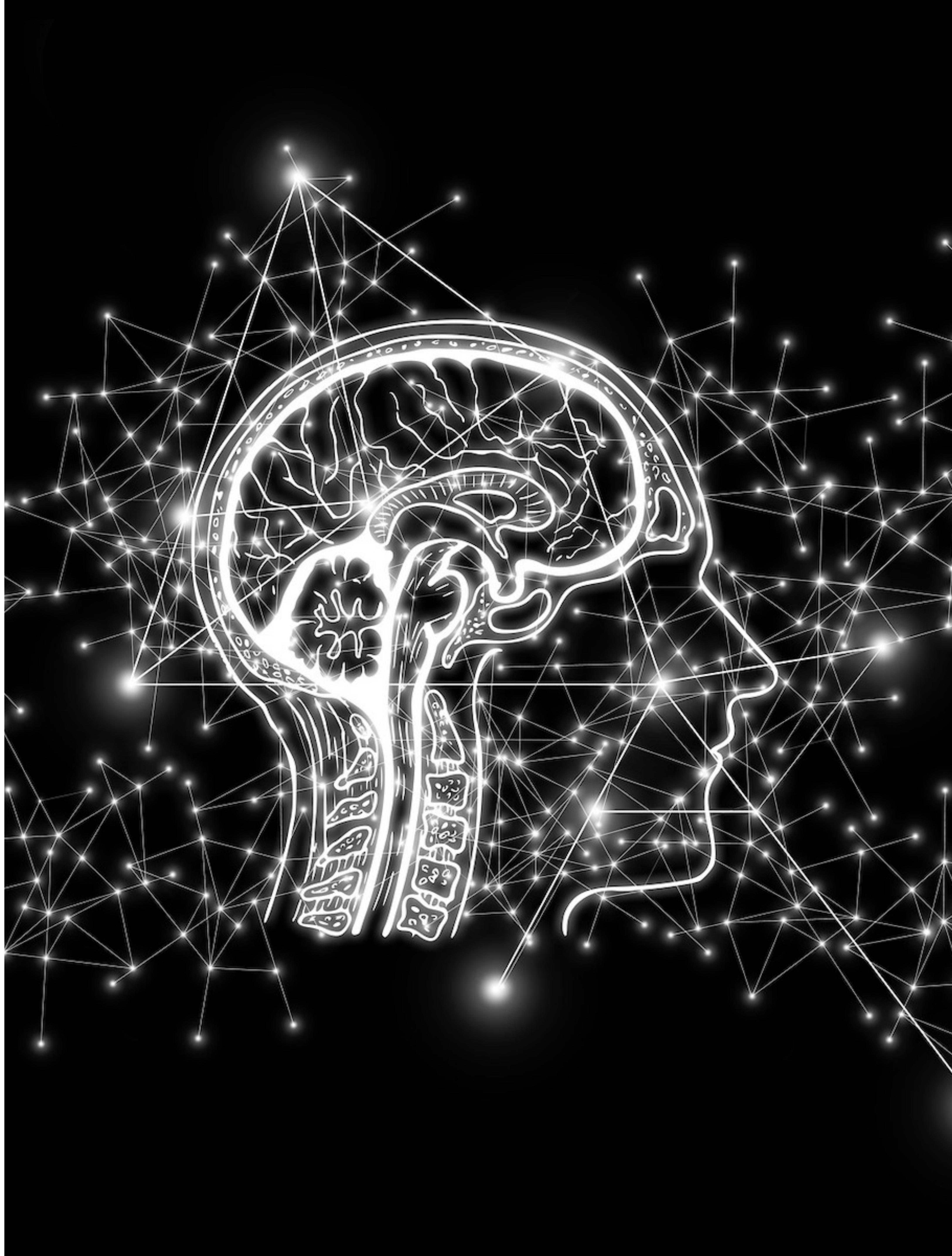
# OPINION

Whisper’s extensive language support and impressive accuracy make it a promising technology in the field of speech recognition.

More details such as punctuation accuracy, real-life dialects support, and issues such as privacy concerns should be taken into consideration.

An Integration of Text-to-Speech and Speech-to-Text could lead to a more comprehensive and versatile application.

Continuous refinement and ongoing enhancements are needed in order to reach its full potential in accessibility and efficiency for a wider user base.



# OPENAI/ WHISPER

## SPEECH-TO-TEXT GENERATIVE AI

PRESENTED BY YVES YANG

Illustration Artist: Ruby Chen