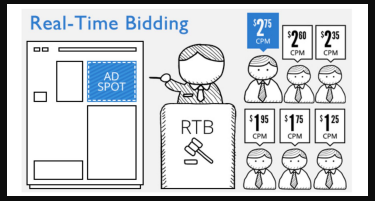
Gladia

Lost in Transcription

Whisper Roulette

WHO AM I

- CEO @Gladia (Speech-to-text)
- VP Data, AI and Quantum @OVHcloud
- CTO @Auchan Retail International
- Business Angel (40+ companies)
- Projects:
  - AI / ML:
    - Predictive maintenance 500K servers
    - Anti-Fraud & Security
    - Real-Time Bidding
    - Audio // NLP // Video // Translation // LLM
  - Other:
    - Big Data (a lot)
    - Quantum Computing (a lot)
- What I love to share: Science // Hardware



Gladia

Gladia support team for Granola

✓ **Dedicated Engineering Manager**
Technical support and guidance

✓ **Account Manager**
All account-related needs

✓ **Design Partnering Program**
Collaborate on solutions uniquely shaped for Granola use cases and business model

✓ **Direct Access to Founders**

Sami
Engineering Manager

Ludovic
Account Manager

Fred
Head of Product

Nicolas
Product Manager

Jean-Louis
Founder & CEO

Jonathan
Founder & CTO


Gladia
granola

Powering 120K+ developers around the globe developing a robust AI platform designed enhance sales performance or call-center operations (CX)

Gladia

Accuracy made for the modern Enterprise

~ 1% word-error rate

Performance tailored for high-velocity teams

~ 1mn for 1 hour audio file

Pricing in-line with business plans

avg. 30% less than main providers (Google STT, AWS Transcribe)

recall · Livestorm · Circleback. · Attention · clap · Calendly · CARV

Gladia

Whisper-Zero

Enjoy the best version of Whisper at scale
with no limitations, errors, and hallucinations

Multilingual Model

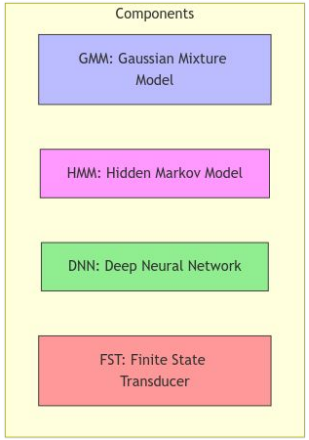
10-15% less WER

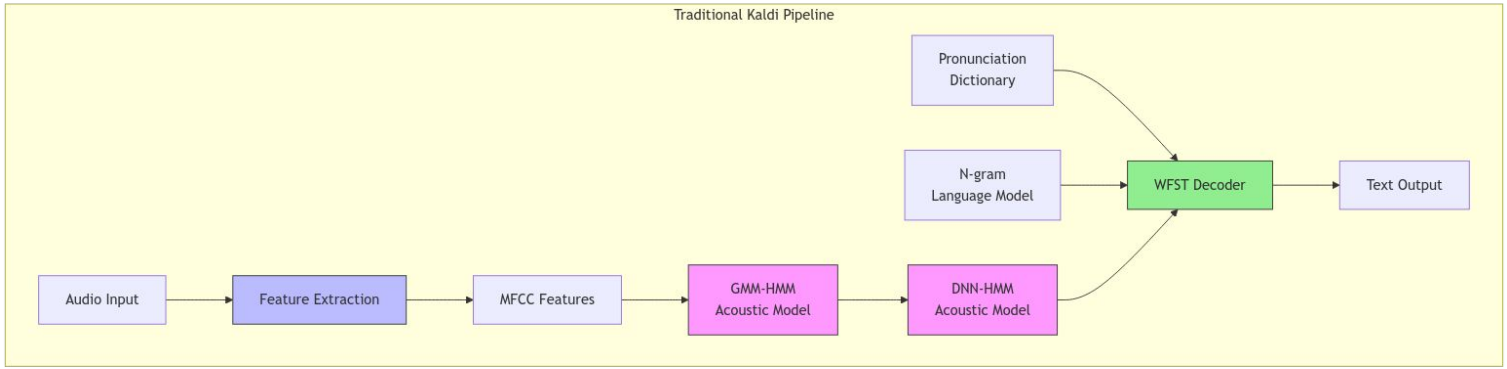
Made for real-life audio

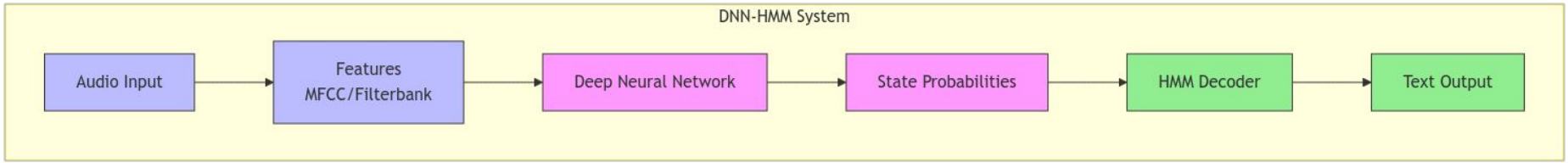
REAL TIME - LATENCY / ACCURACY

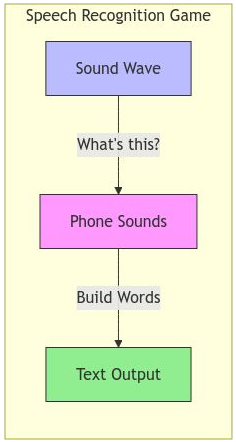
- ✓ Gladia's **real-time** and **multilingual** transcription boasts an industry-leading **latency** of **under 300 milliseconds**.
- ✓ We provide the same high level of **accuracy** in **real-time** as in **batch** processing, regardless of language.

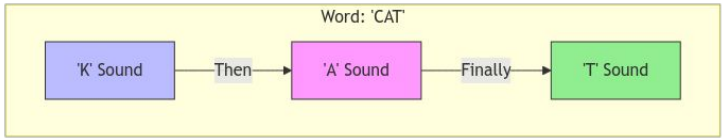
ASR / STT / speech to text traditional

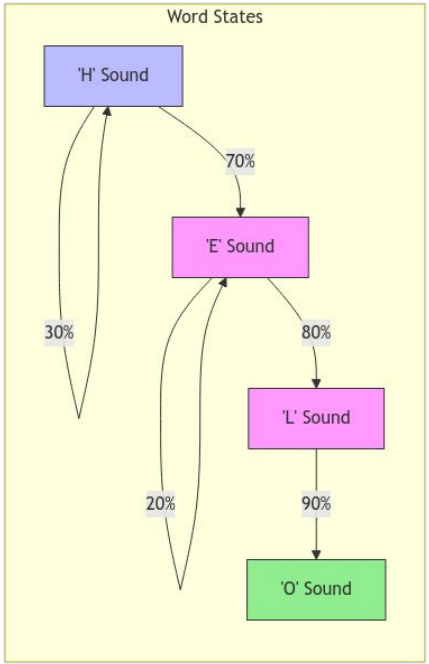


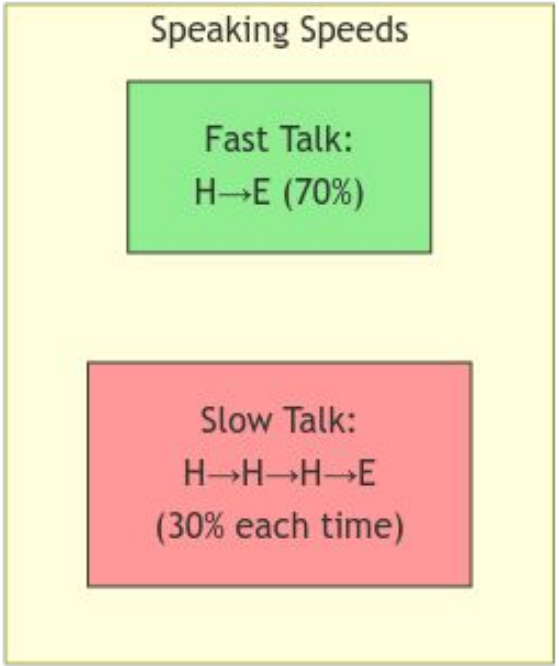


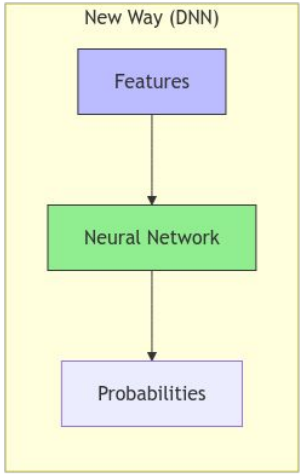


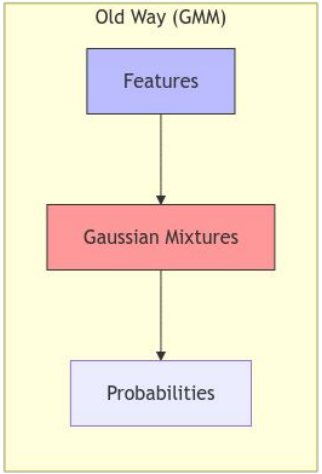

Speech Recognition Game
Sound Wave
What's this?
Phone Sounds
Build Words
Text Output

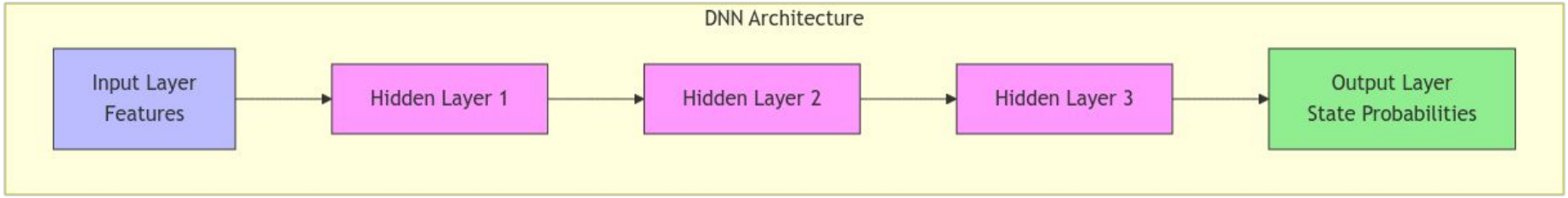

Word: 'CAT'
'K' Sound
Then
'A' Sound
Finally
'T' Sound

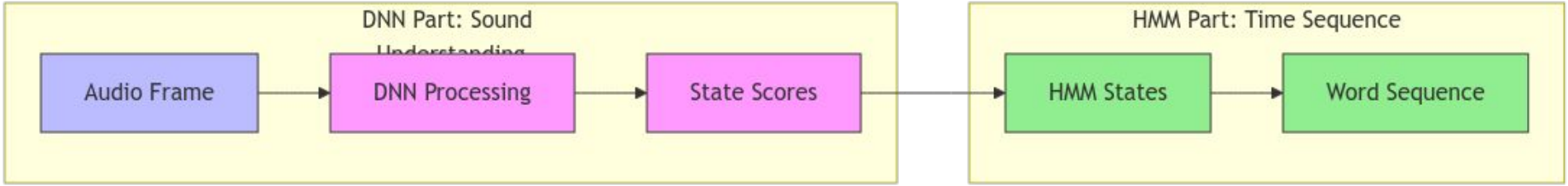

Word States
'H' Sound
70%
'E' Sound
30%
80%
'L' Sound
20%
90%
'O' Sound


Speaking Speeds
Fast Talk:
H→E (70%)
Slow Talk:
H→H→H→E
(30% each time)

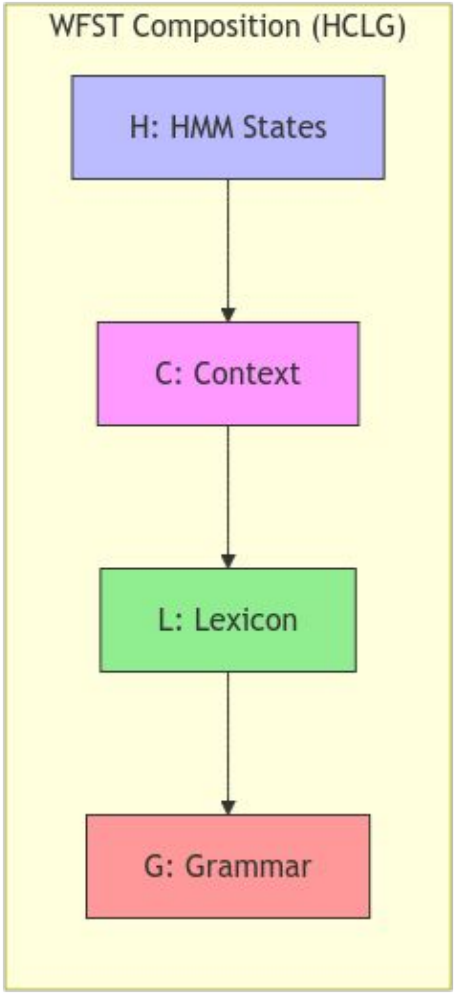

New Way (DNN)
Features
Neural Network
Probabilities

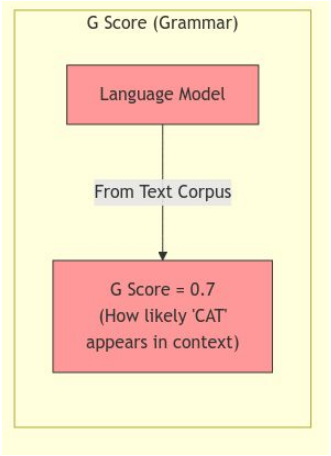

Old Way (GMM)
Features
Gaussian Mixtures
Probabilities

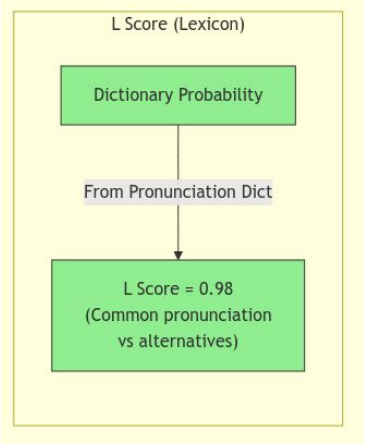

DNN Architecture
Input Layer
Features
Hidden Layer 1
Hidden Layer 2
Hidden Layer 3
Output Layer
State Probabilities

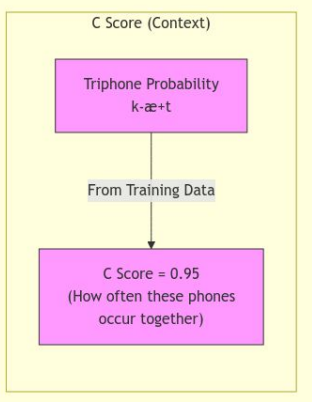

DNN Part: Sound
Audio Frame
DNN Processing
State Scores
HMM Part: Time Sequence
HMM States
Word Sequence

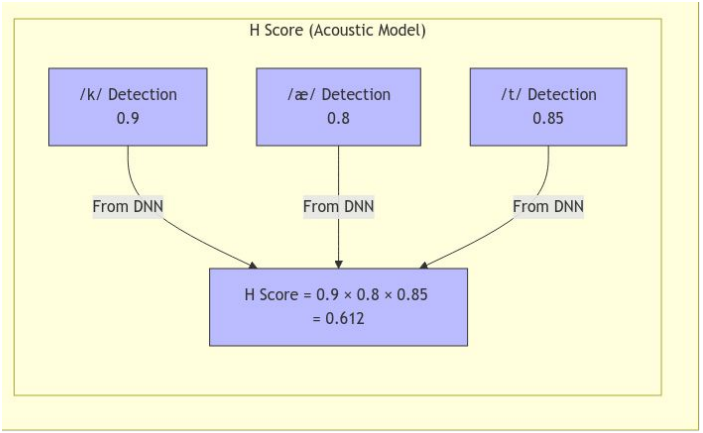
WFST: Weighted Finite State Transducer



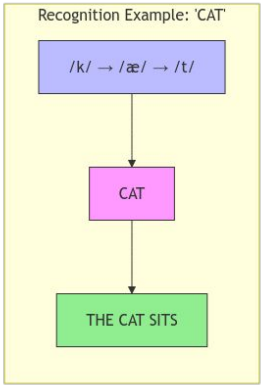


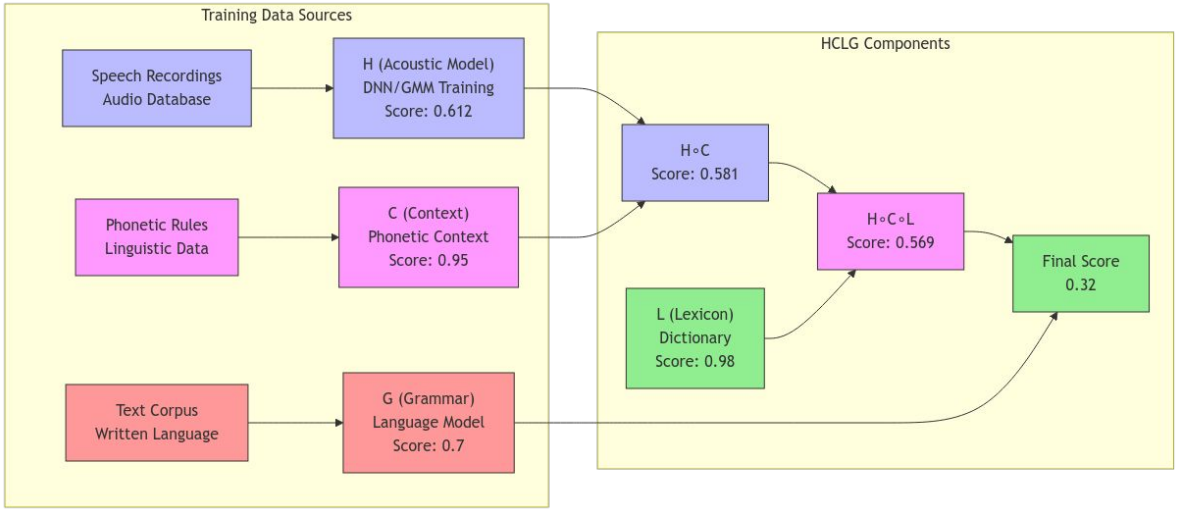




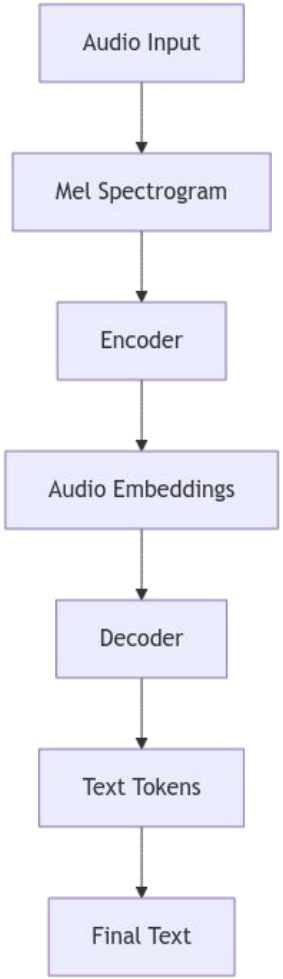


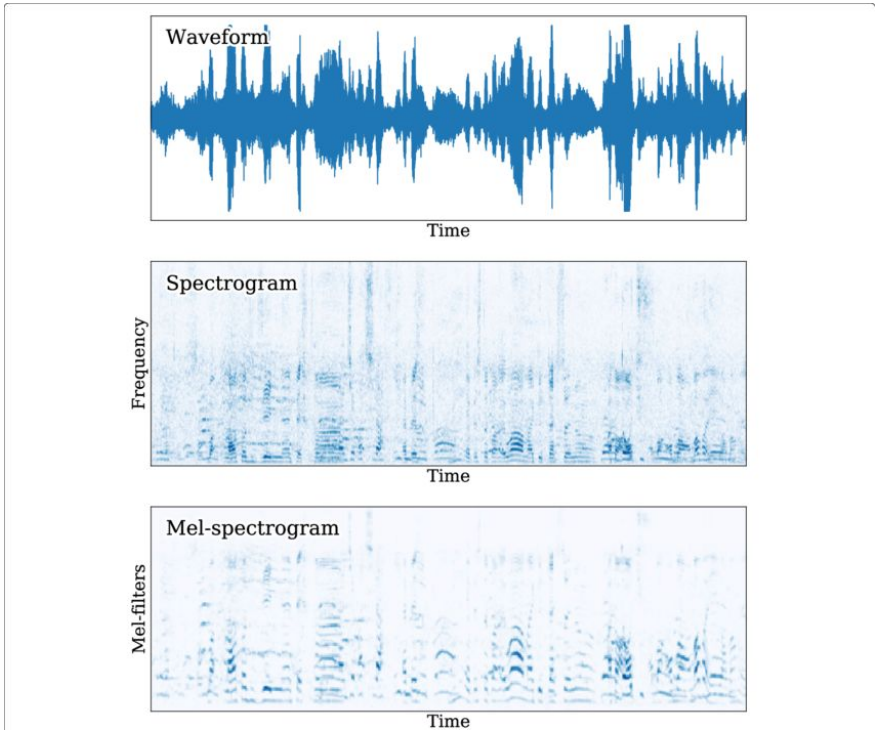
H(idden MM) C(ontext Deps) L(exicon) G(rammar)

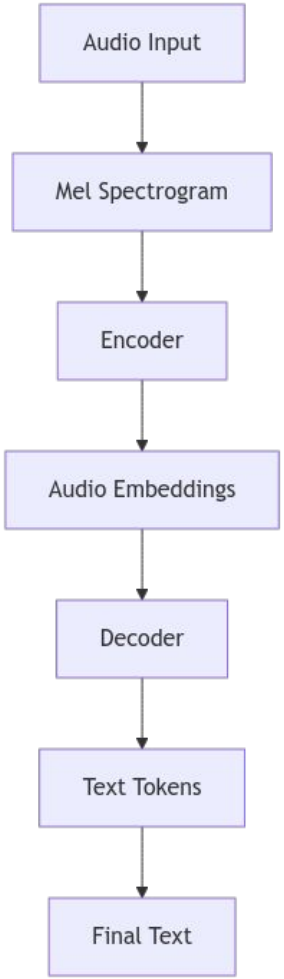




ASR / STT / speech to text Generative

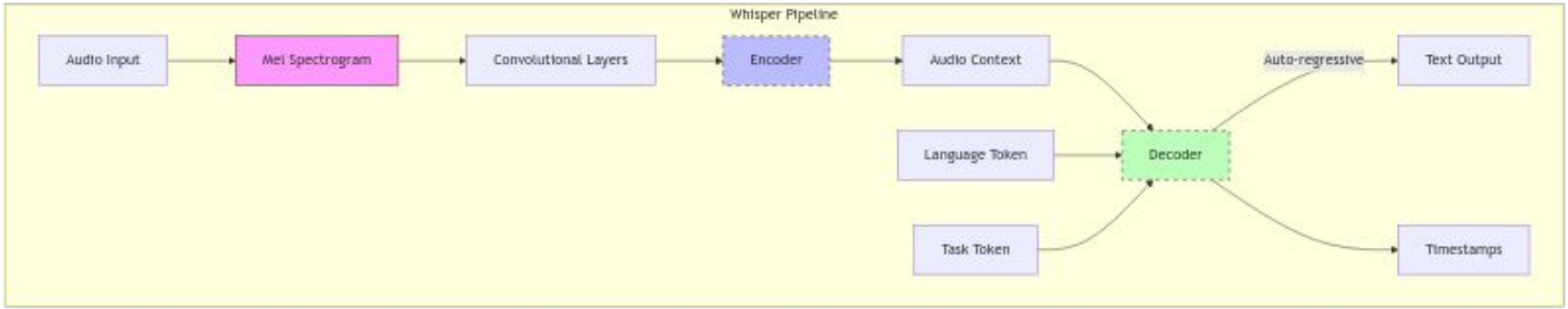




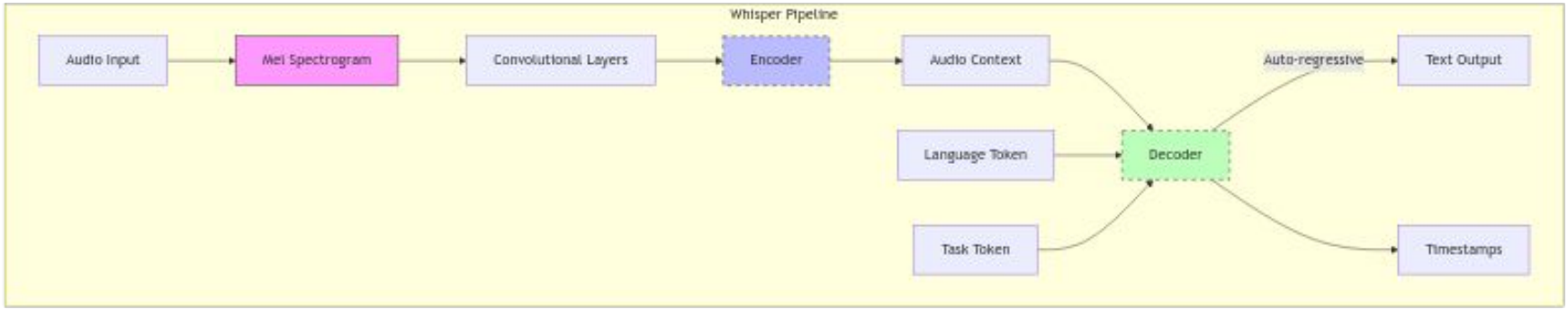

Audio Input
Mel Spectrogram
Encoder
Audio Embeddings
Decoder
Text Tokens
Final Text

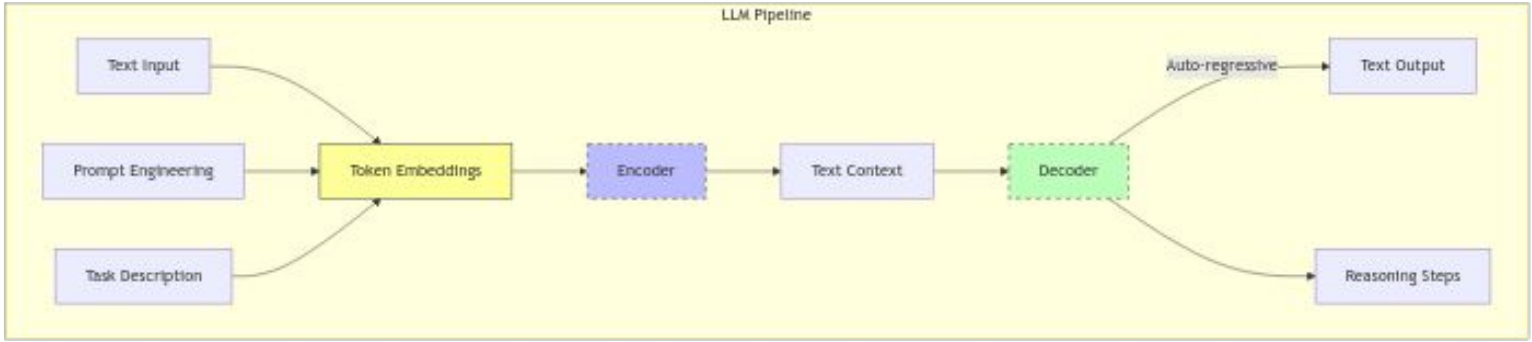

Key Components
Input Processing
Transformer Blocks
Output Generation


Key Components
Input Processing
Transformer Blocks
Output Generation

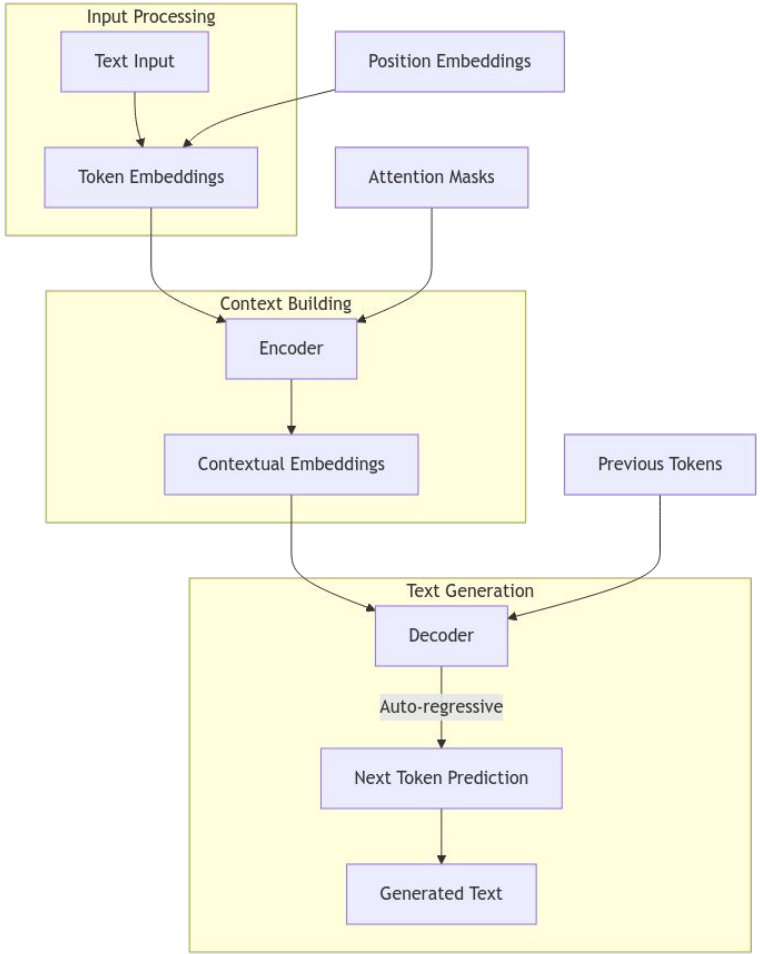

Whisper Pipeline
Audio Input
Mel Spectrogram
Convolutional Layers
Encoder
Audio Context
Auto-regressive
Text Output
Language Token
Decoder
Task Token
Timestamps


Key Components
Input Processing
Transformer Blocks
Output Generation

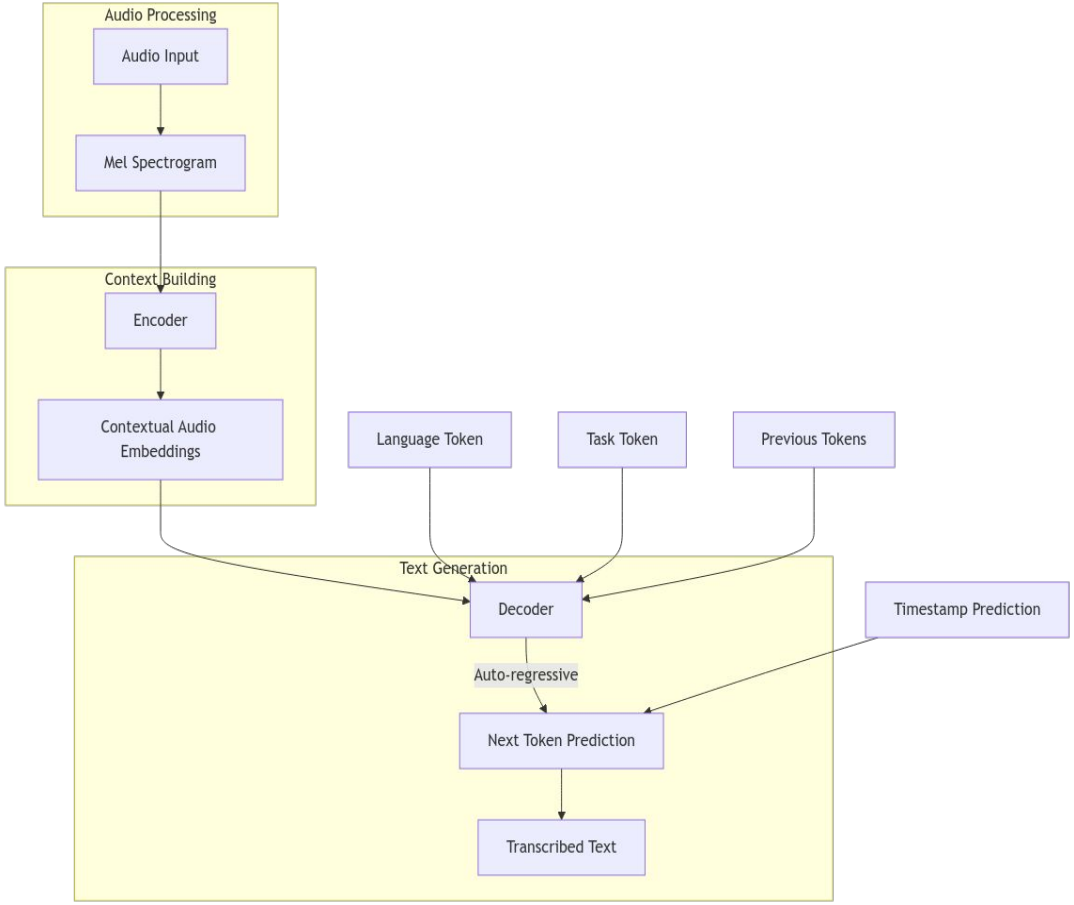

Whisper Pipeline
Audio Input
Mel Spectrogram
Convolutional Layers
Encoder
Audio Context
Auto-regressive
Text Output
Language Token
Decoder
Task Token
Timestamps


LLM Pipeline
Text Input
Auto-regressive
Text Output
Prompt Engineering
Token Embeddings
Encoder
Text Context
Decoder
Task Description
Reasoning Steps

LLM Architecture



whisper Architecture



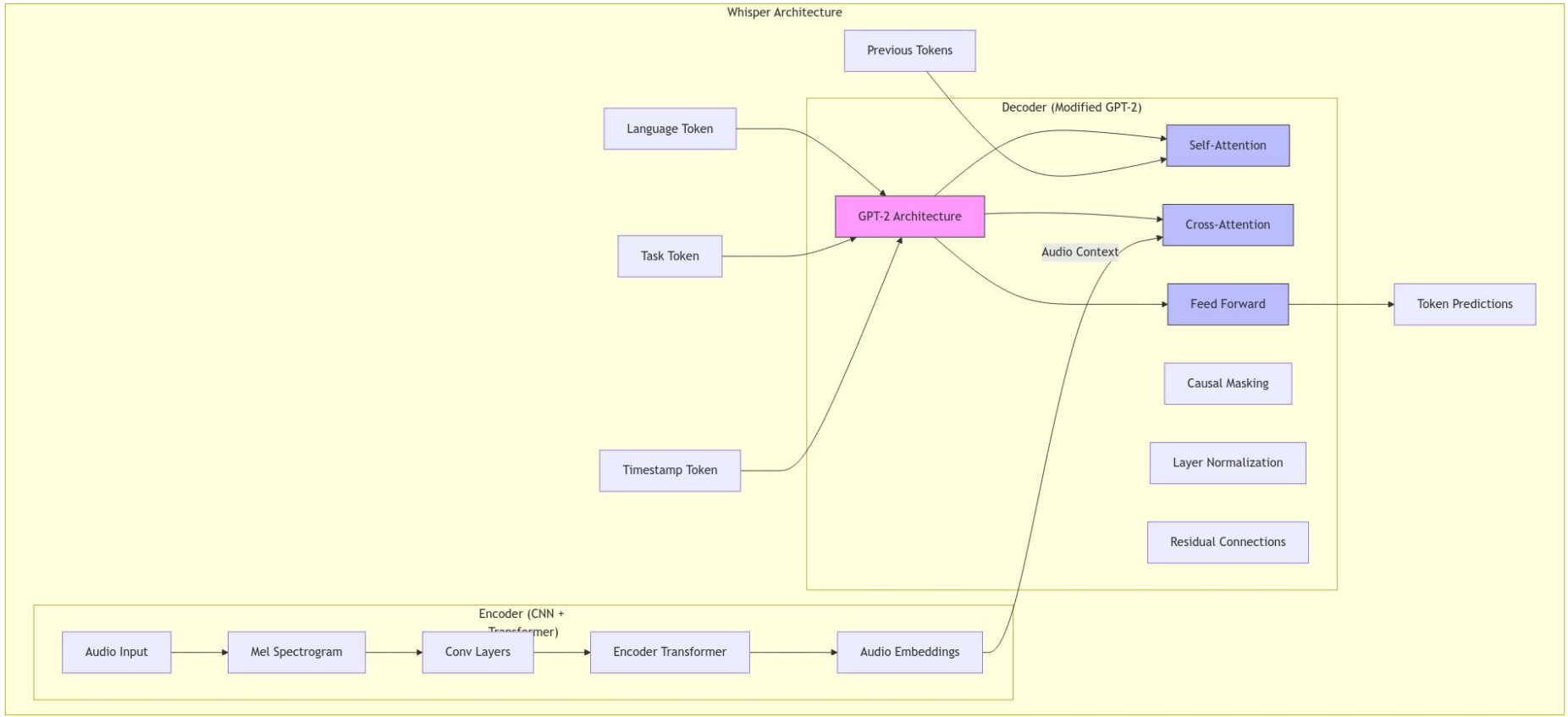
LLM Architecture

```
# LLM Output
[
    "Hello",
    "world",
    "!"
]
```

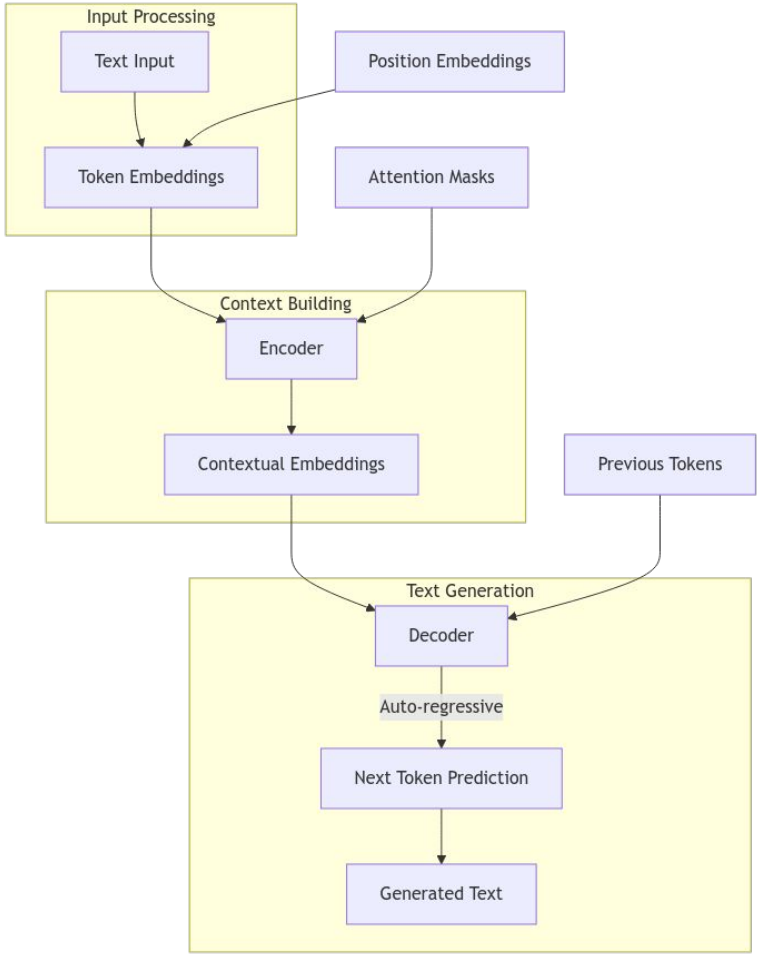
whisper Architecture

```
# Whisper Output Example
[
    "<|startoftranscript|>",  # Special token
    "<|en|>",                 # Language token
    "Hello",                  # Text token
    "<|0.0|>",                # Timestamp token
    "world",                  # Text token
    "<|1.5|>"                 # Timestamp token
]
```

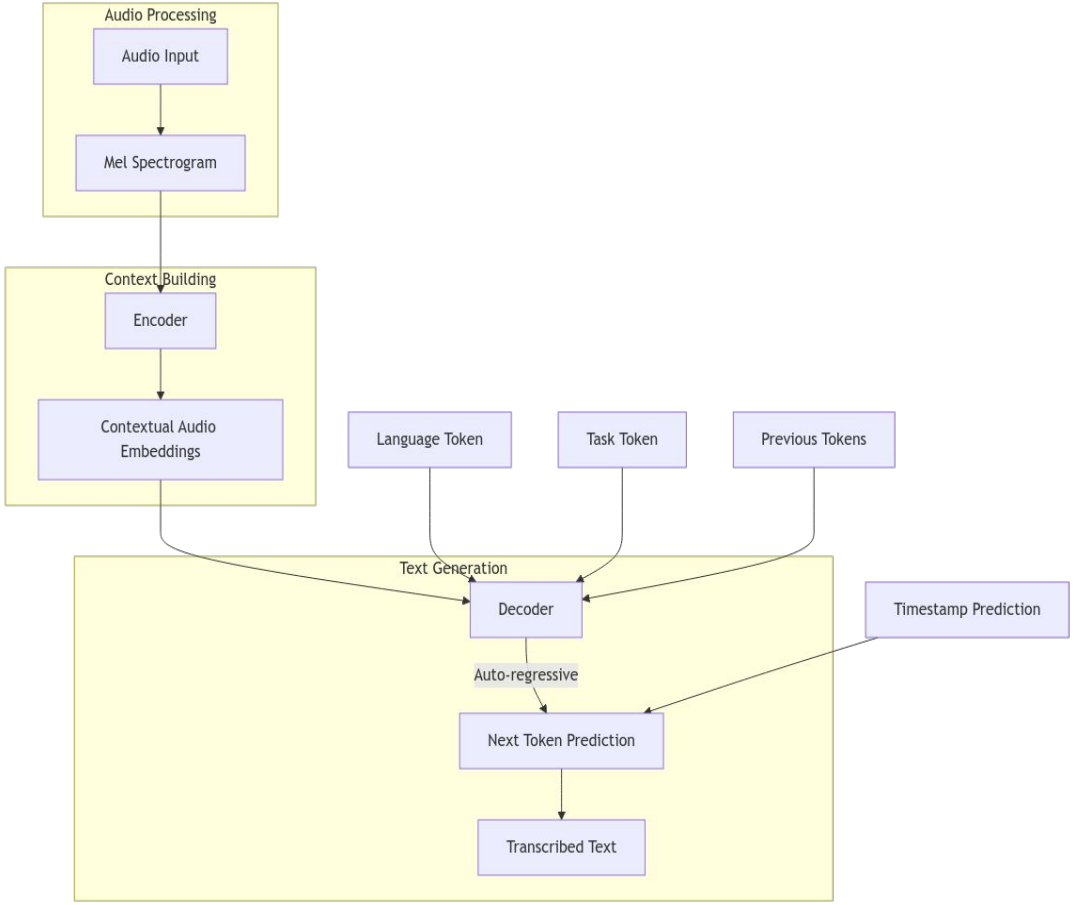
```
# Typical token sequence
sequence = [
    "<|startoftranscript|>",  # Start marker
    "<|en|>",                 # Language
    "<|transcribe|>",         # Task
    "<|0.00|>",               # Initial timestamp
    "Hello world",            # Text content
    "<|2.56|>",               # More timestamps
    "<|endoftranscript|>",    # End of current transcript
    "<|endoftext|>",          # Complete end marker
]
```


Whisper Architecture
Previous Tokens
Language Token
Task Token
Timestamp Token
Decoder (Modified GPT-2)
GPT-2 Architecture
Self-Attention
Cross-Attention
Audio Context
Feed Forward
Token Predictions
Causal Masking
Layer Normalization
Residual Connections
Encoder (CNN + Transformer)
Audio Input
Mel Spectrogram
Conv Layers
Encoder Transformer
Audio Embeddings

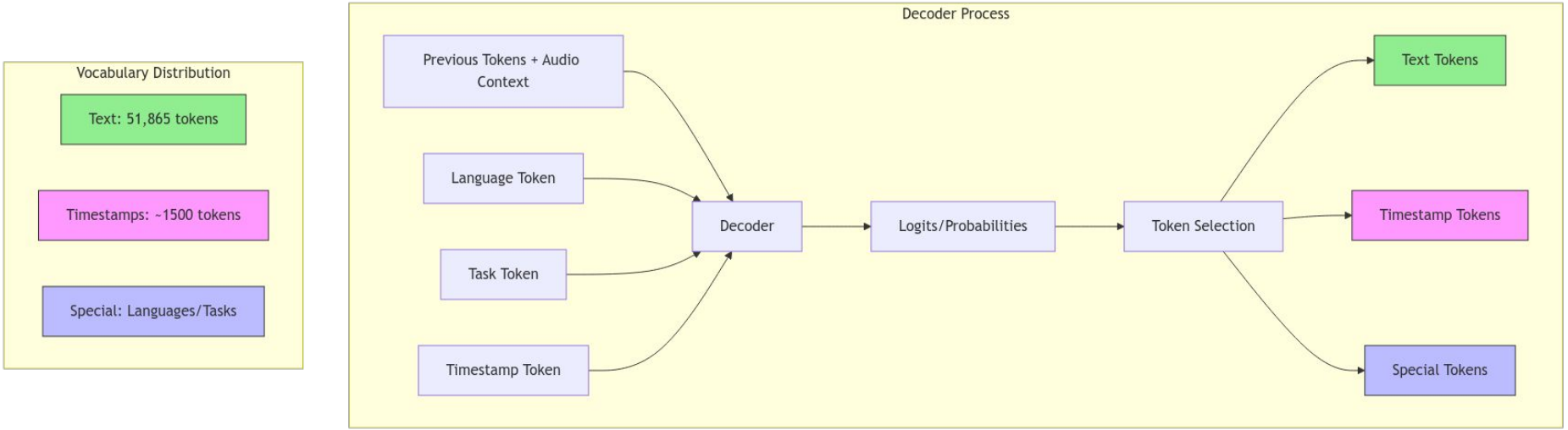
LLM Architecture



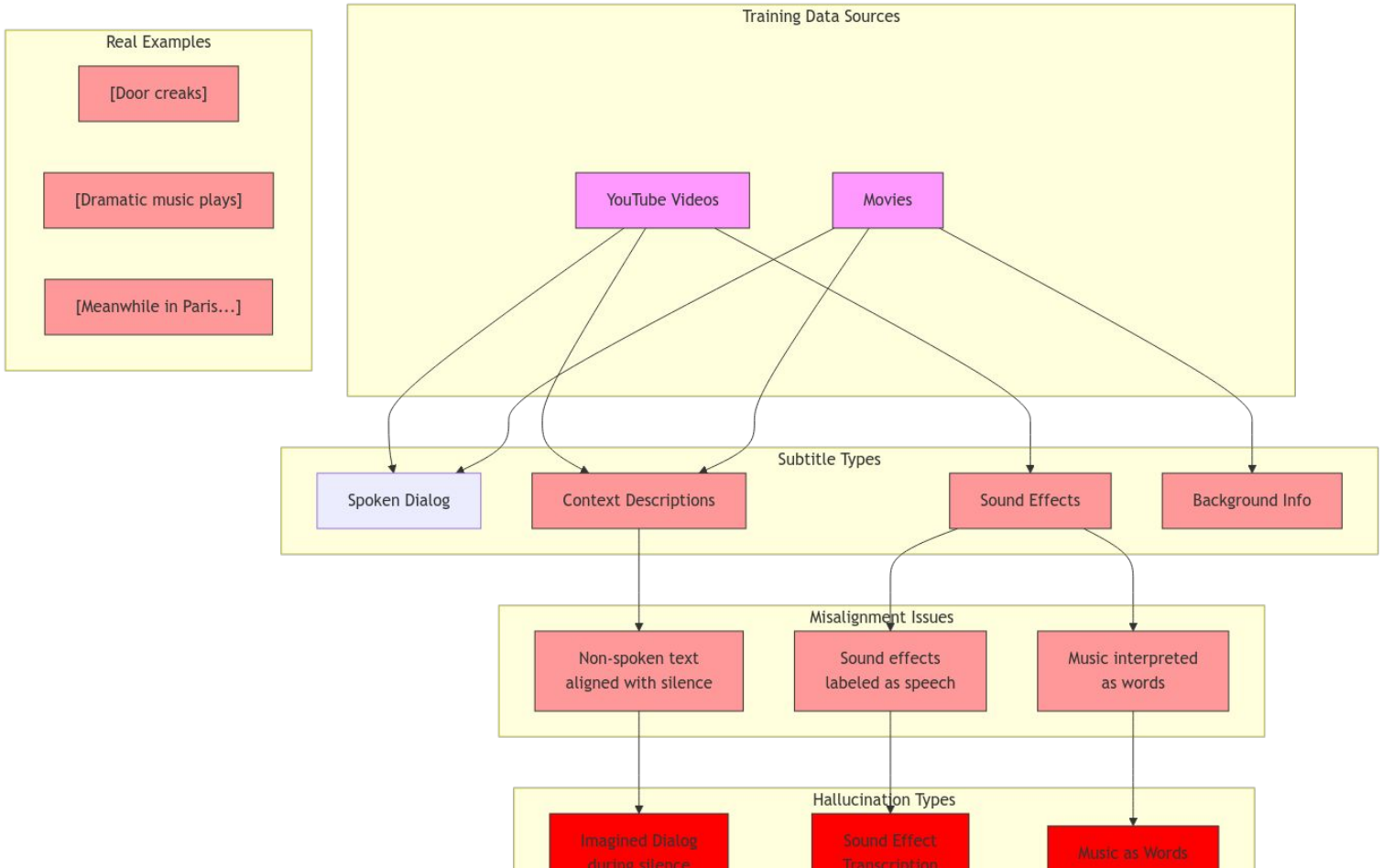
whisper Architecture



Whisper decoder



Training issues

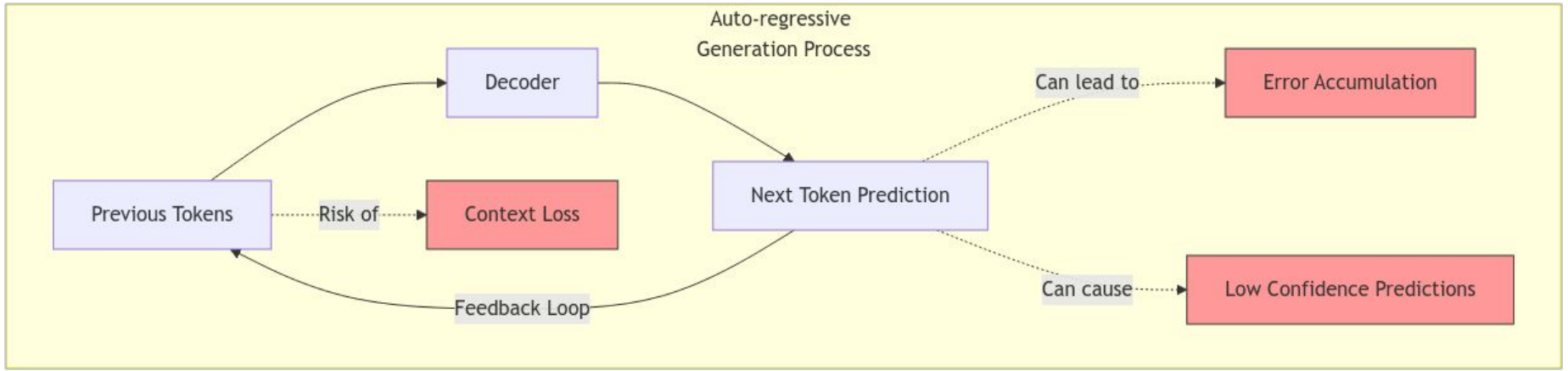


Training issues

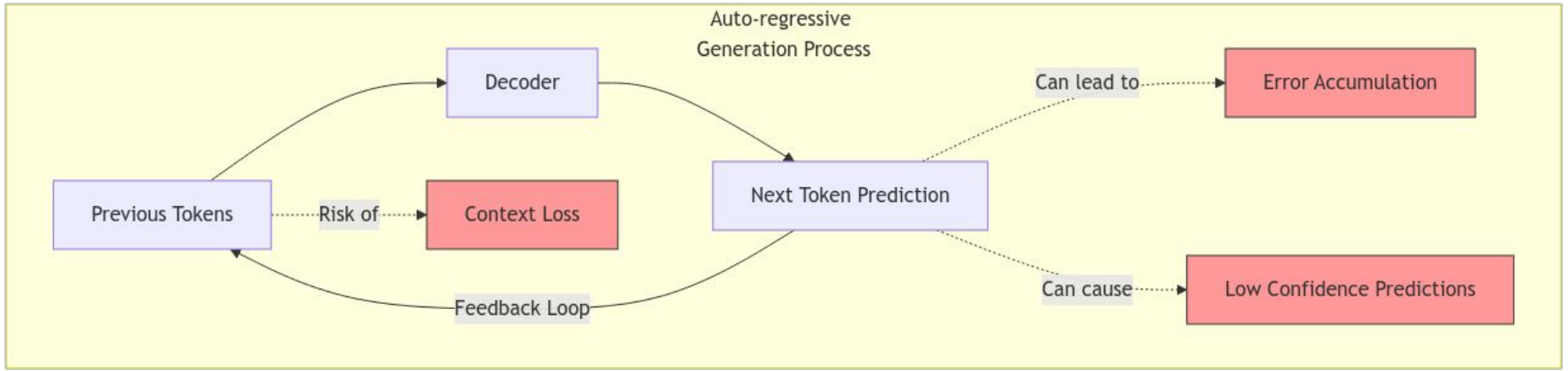
*During early development and evaluation we observed that Whisper models had a tendency to transcribe plausible but almost always incorrect guesses for the names of speakers. T**his happens because many transcripts in the pre-training dataset include the name of the person who is speaking, encouraging the model to try to predict them, but this information is only rarely inferable from only the most recent 30 seconds of audio context.** To avoid this, **we fine-tune Whisper models briefly on the subset of transcripts that do not include speaker annotations which removes this behavior**.*

https://arxiv.org/pdf/2212.04356

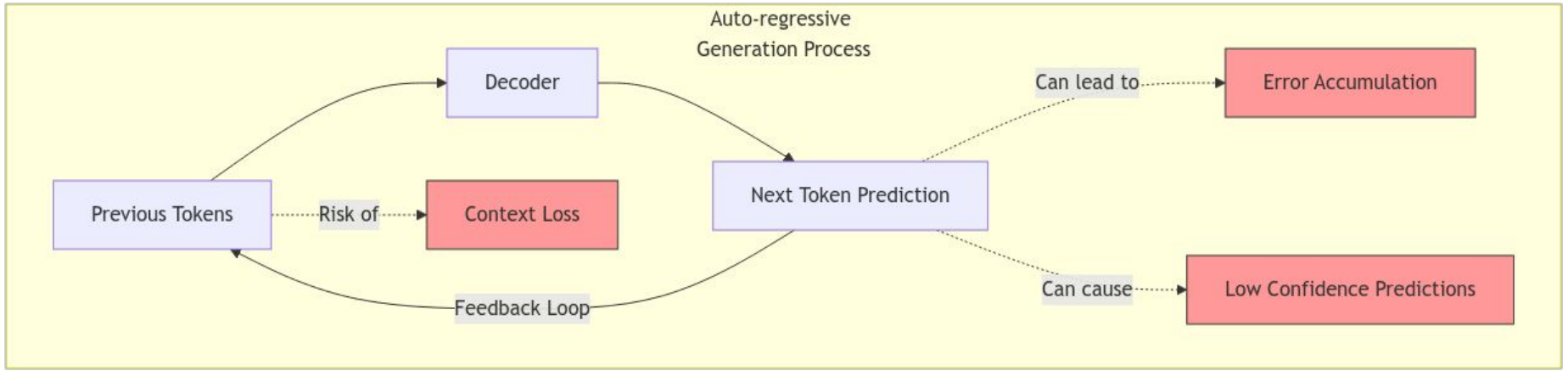
hallucinations



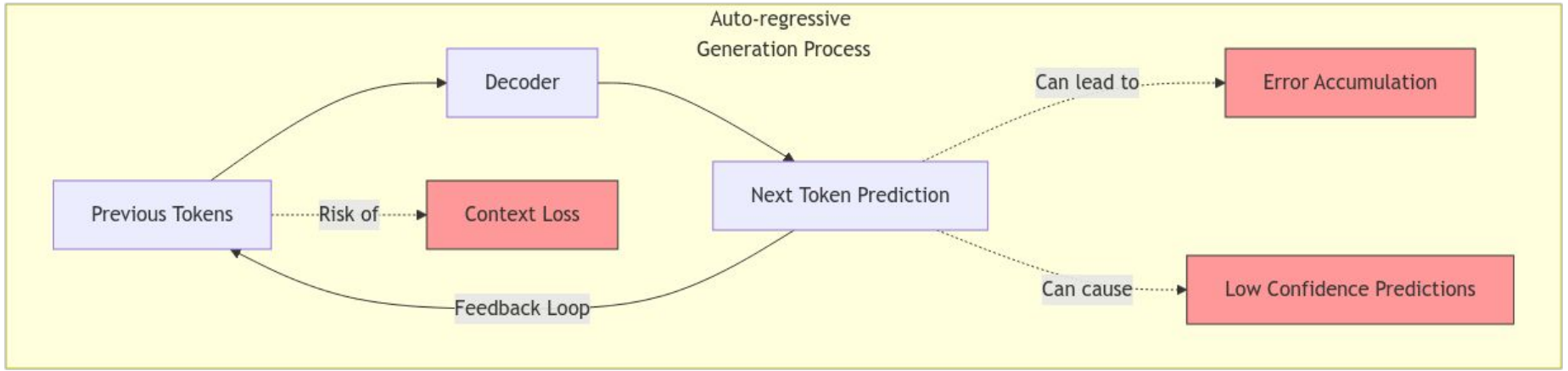
hallucinations



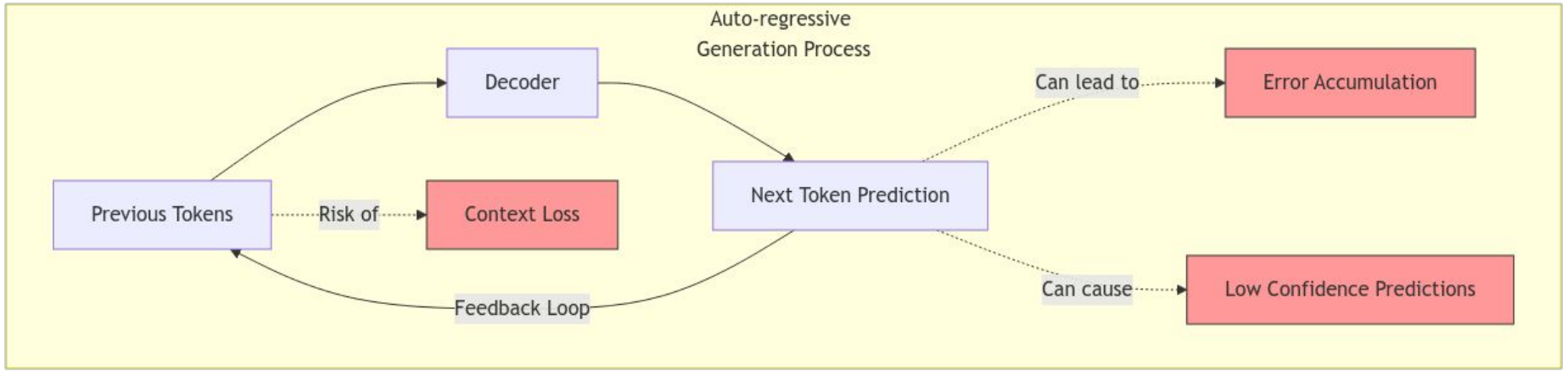
hallucinations



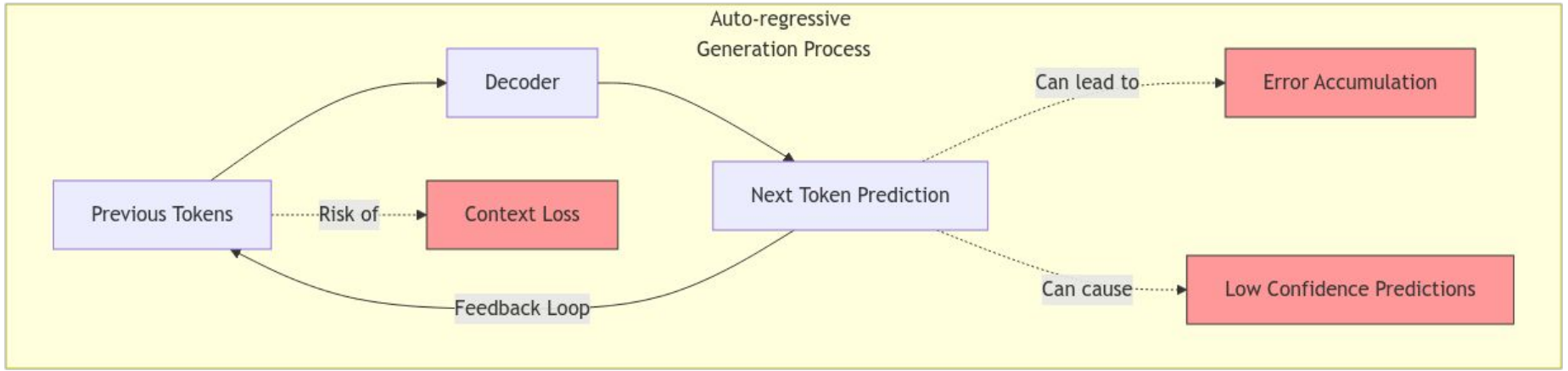
hallucinations



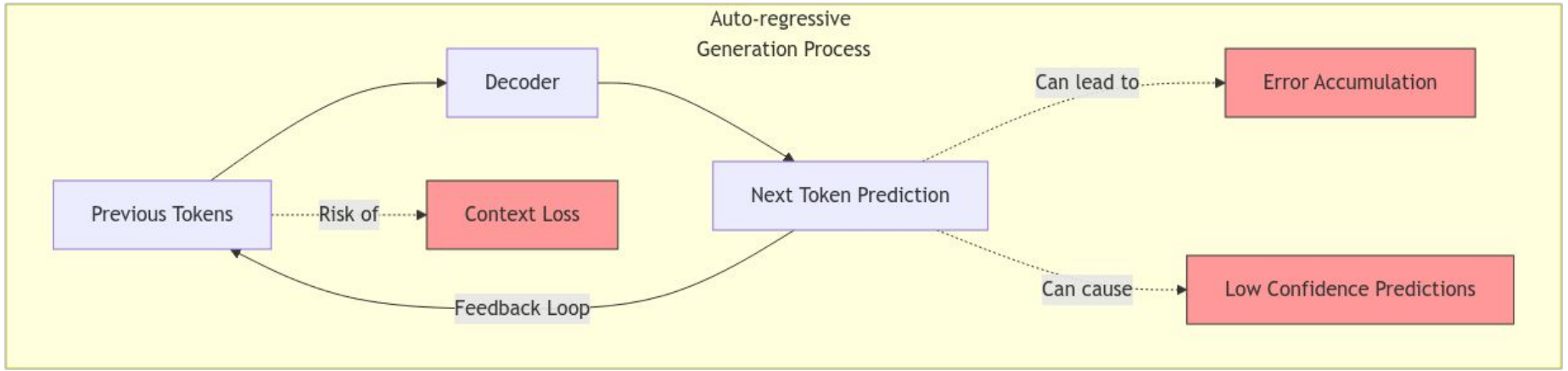
hallucinations



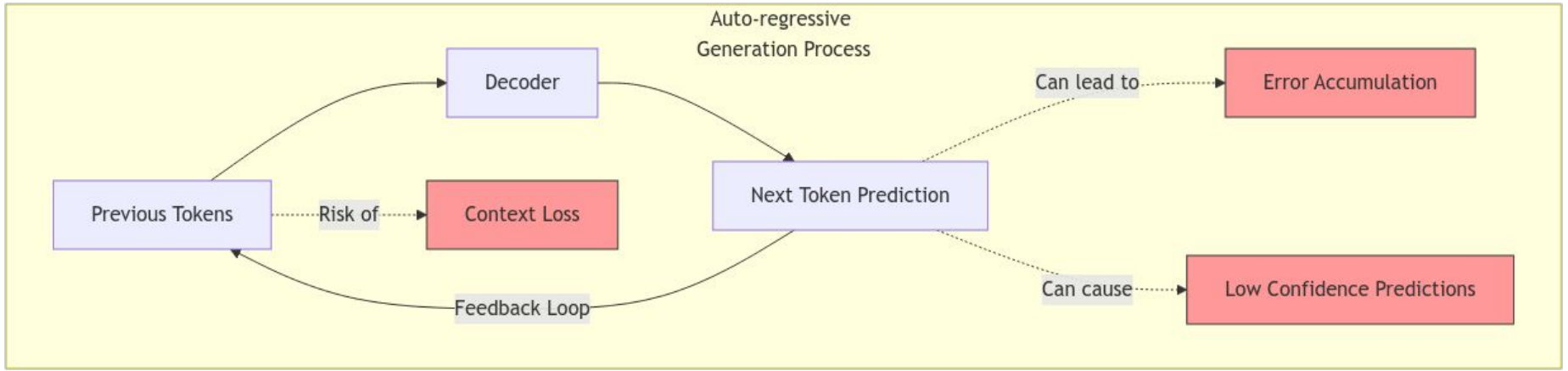
hallucinations



hallucinations



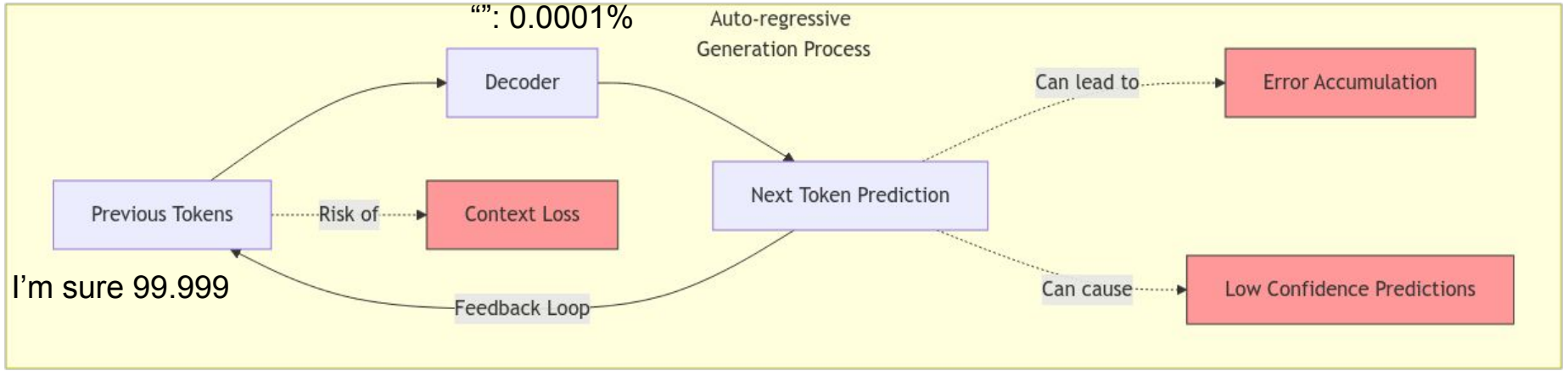
hallucinations



repetitions

I’m sure 99.999%

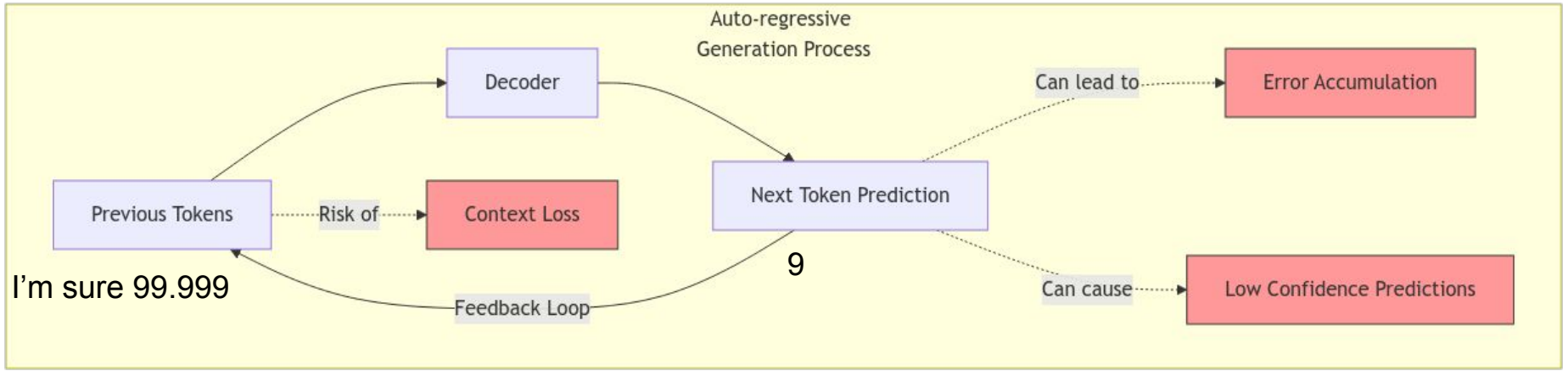
“9”: 95%
“5”: 3%
“0”: 1%
“”: 0.0001%



I’m sure 99.999

repetitions

I’m sure 99.999%

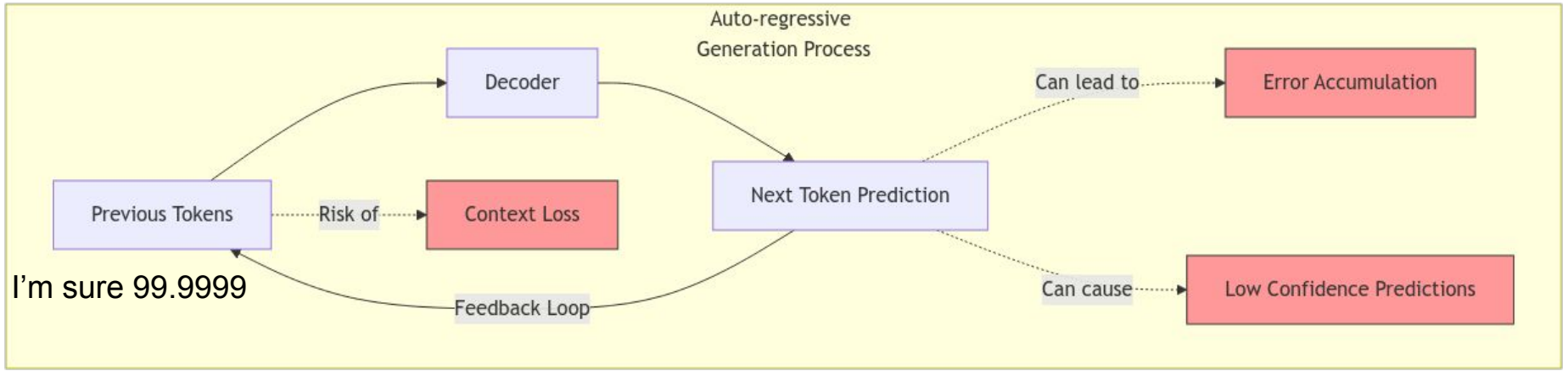


I’m sure 99.999

9

repetitions

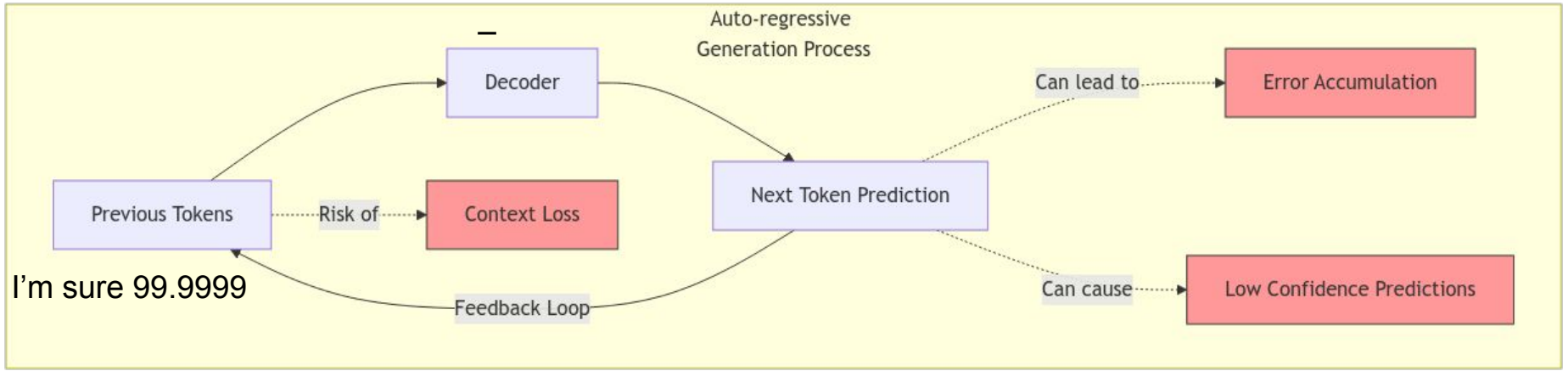
I’m sure 99.999%



I’m sure 99.9999

repetitions

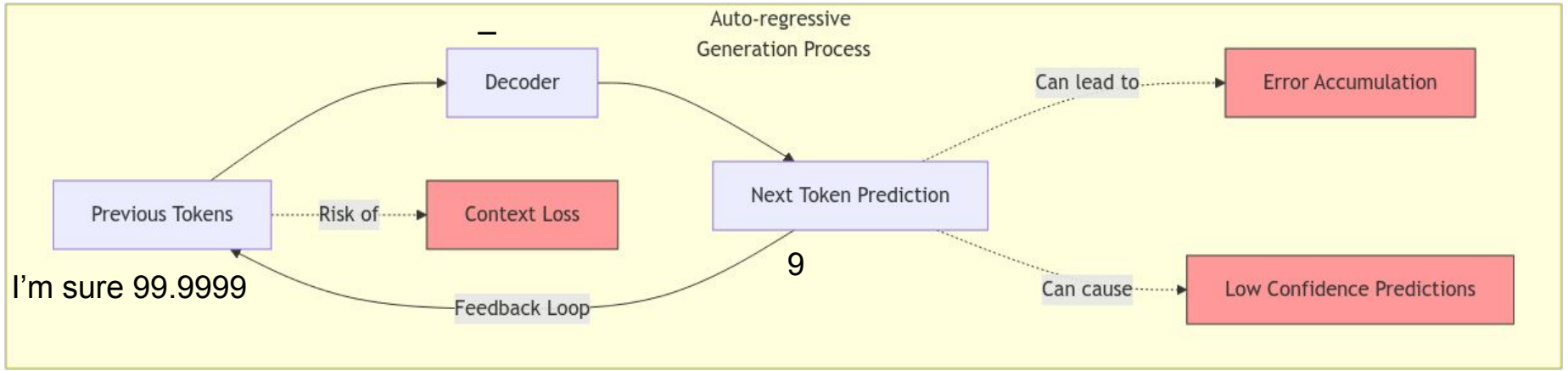
I’m sure 99.999%



I’m sure 99.9999

repetitions

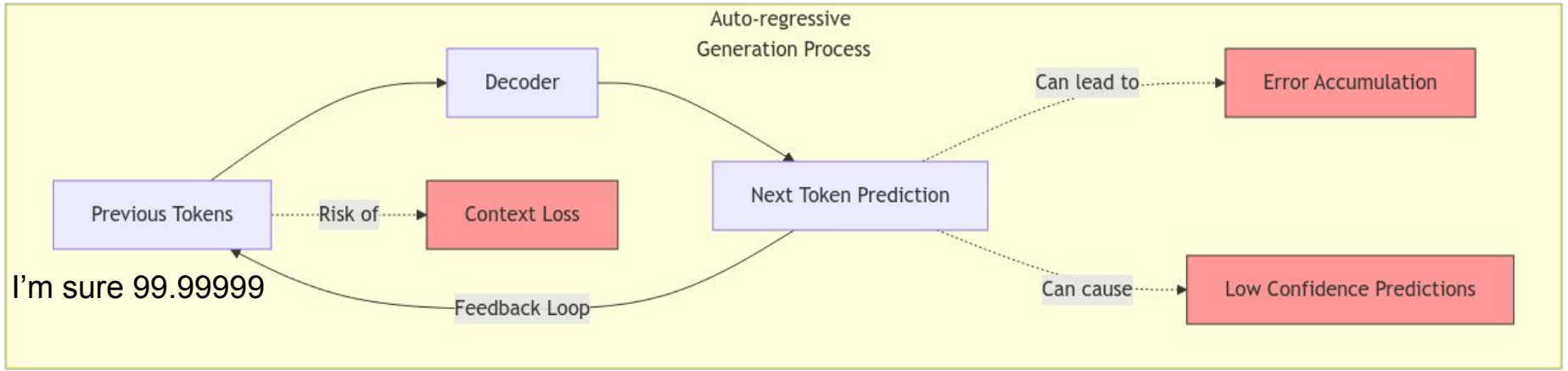
I’m sure 99.999%



I’m sure 99.9999

repetitions

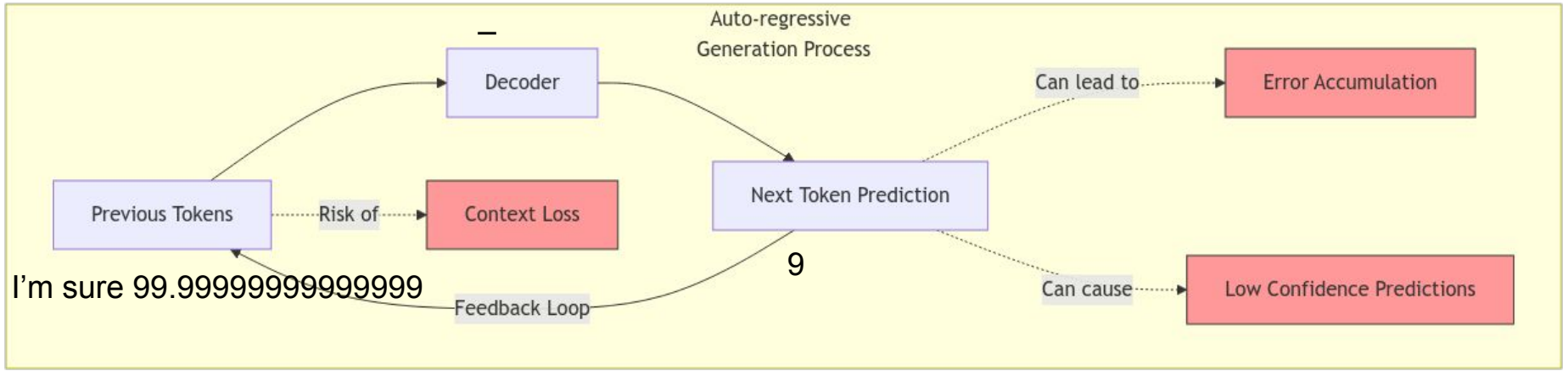
I’m sure 99.999%



I’m sure 99.99999

repetitions

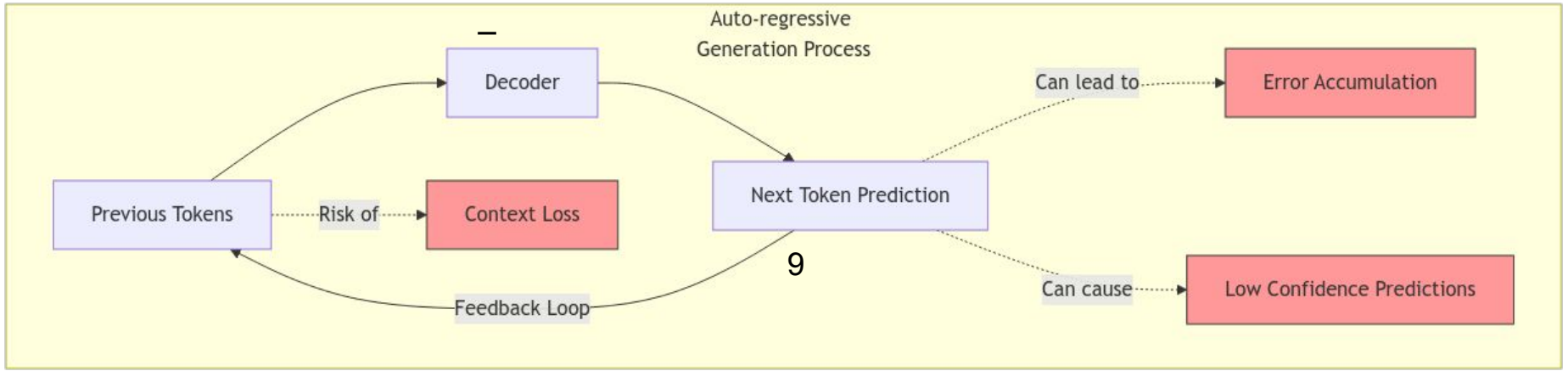
I’m sure 99.999%



I’m sure 99.99999999999999

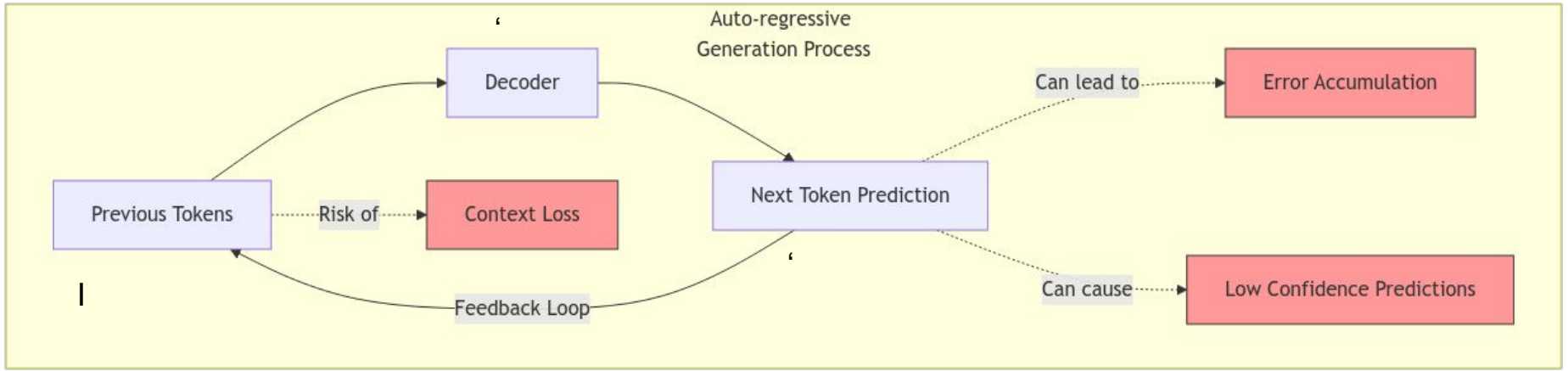
don’t be too silent

I’m <500ms> sure!



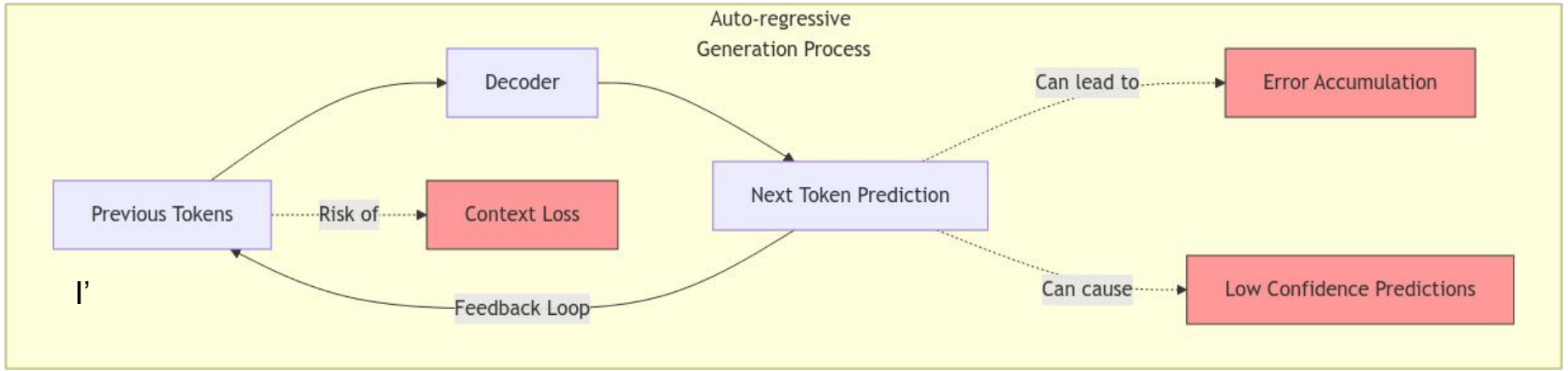
don’t be too silent

I’m <500ms> sure!



don’t be too silent

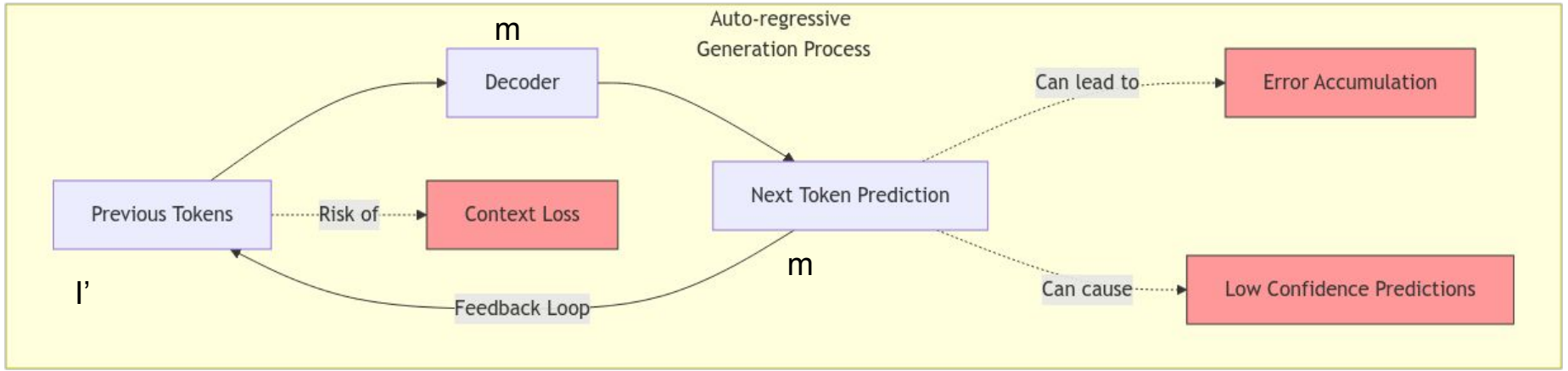
I’m <500ms> sure!



I’

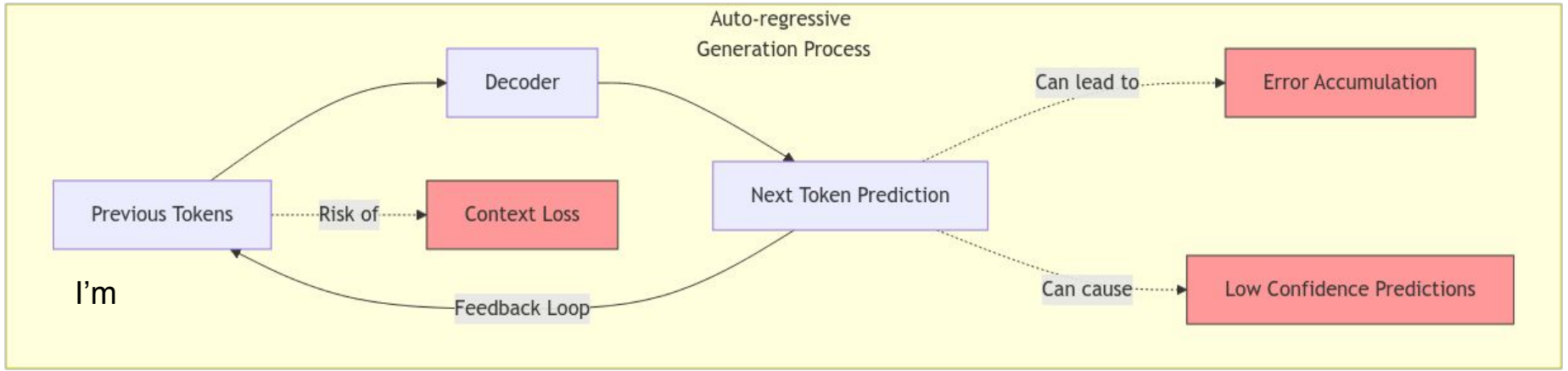
don’t be too silent

I’m <500ms> sure!



don’t be too silent

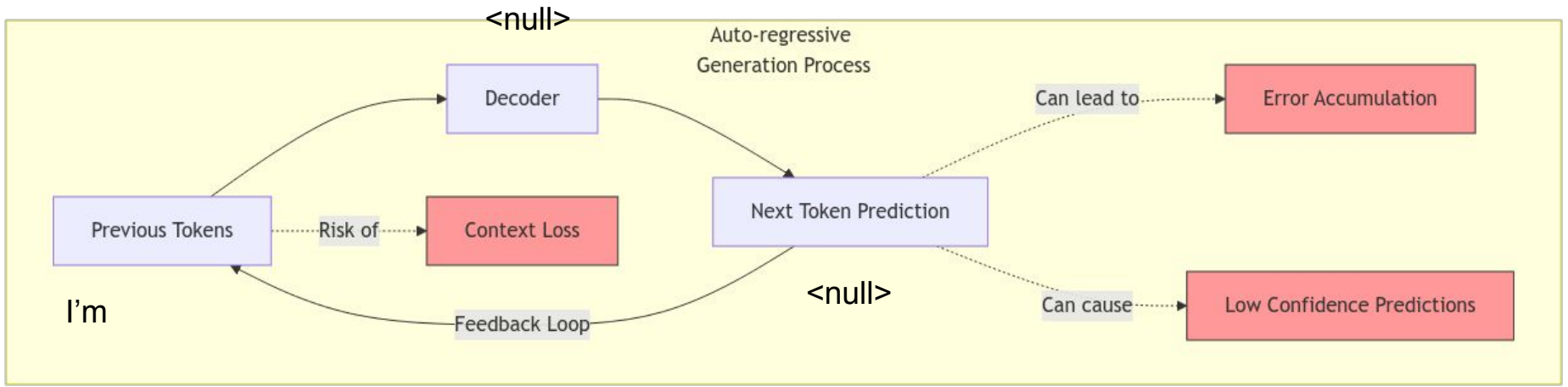
I’m <500ms> sure!



I’m

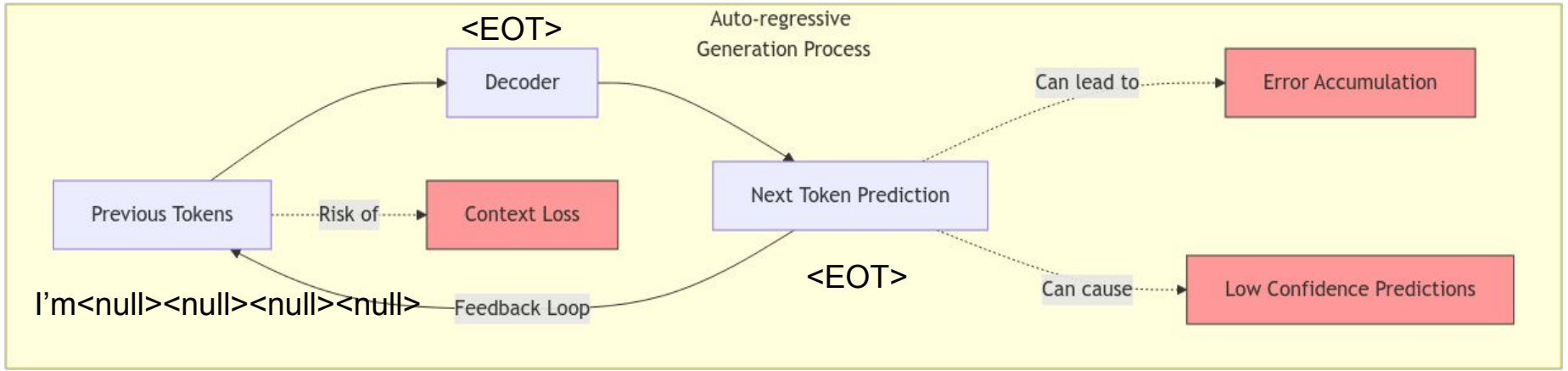
don’t be too silent

I’m <500ms> sure!



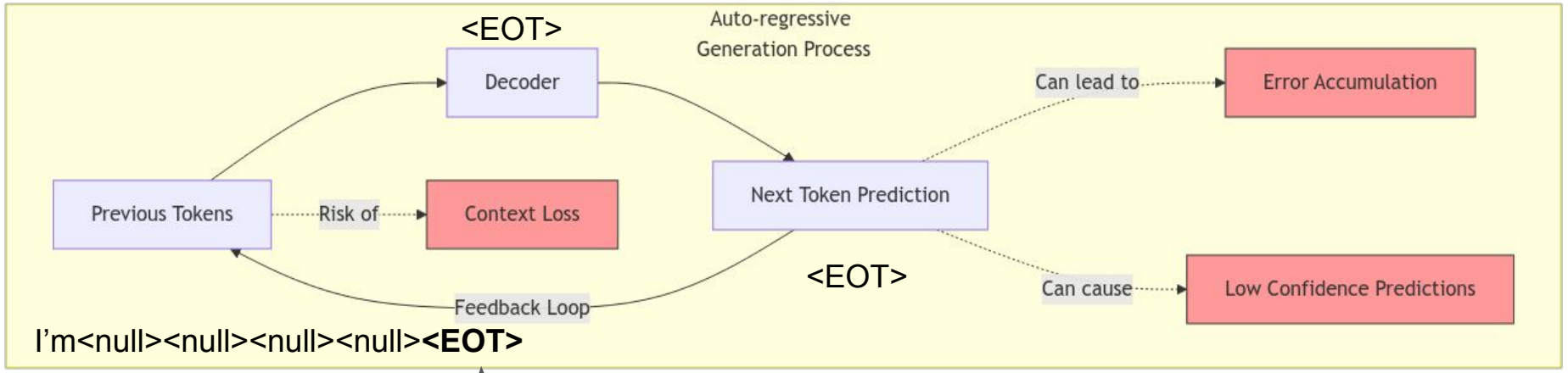
don’t be too silent

I’m <500ms> sure!



don’t be too silent

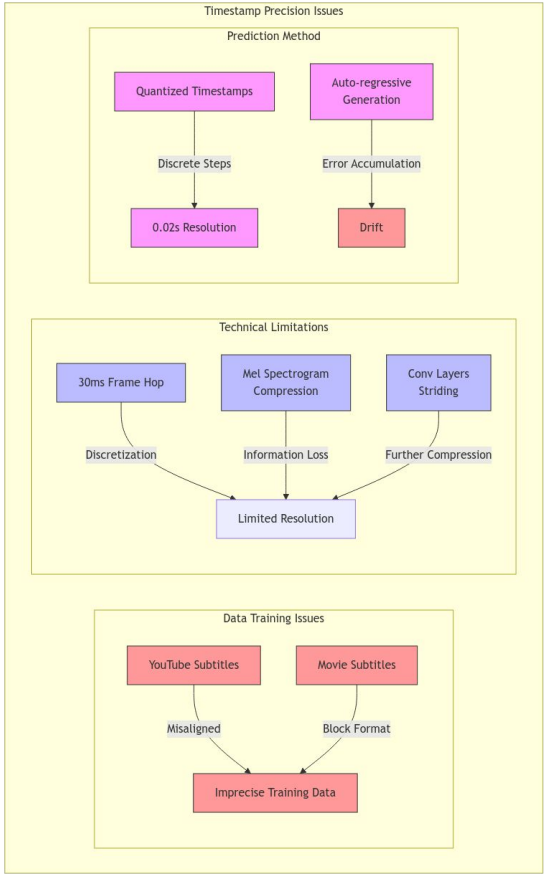
I’m <500ms> sure!



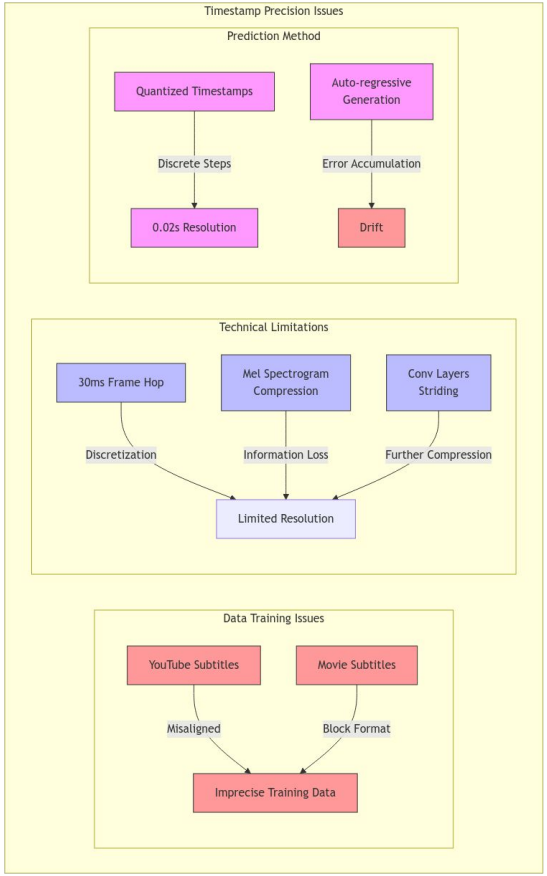
break

I’m <500ms> sure!

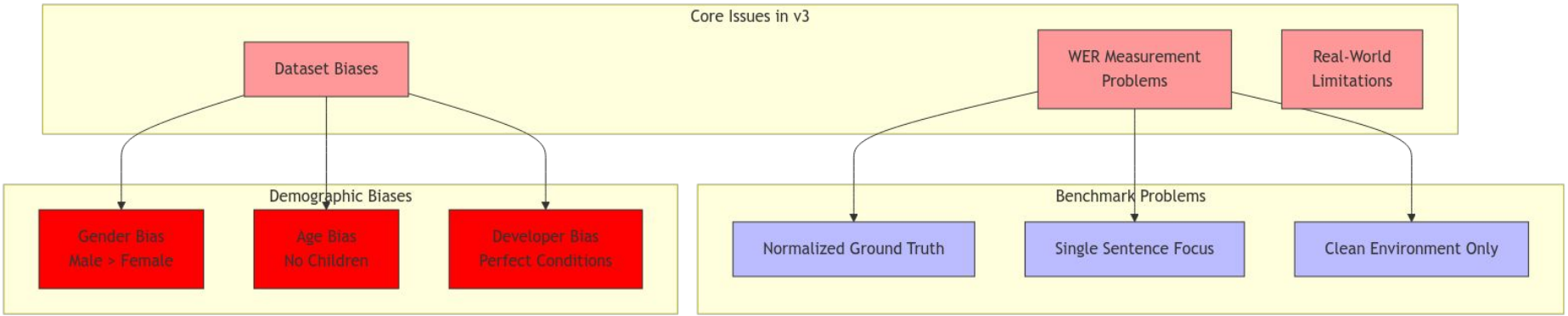
timestamp drifting / precision



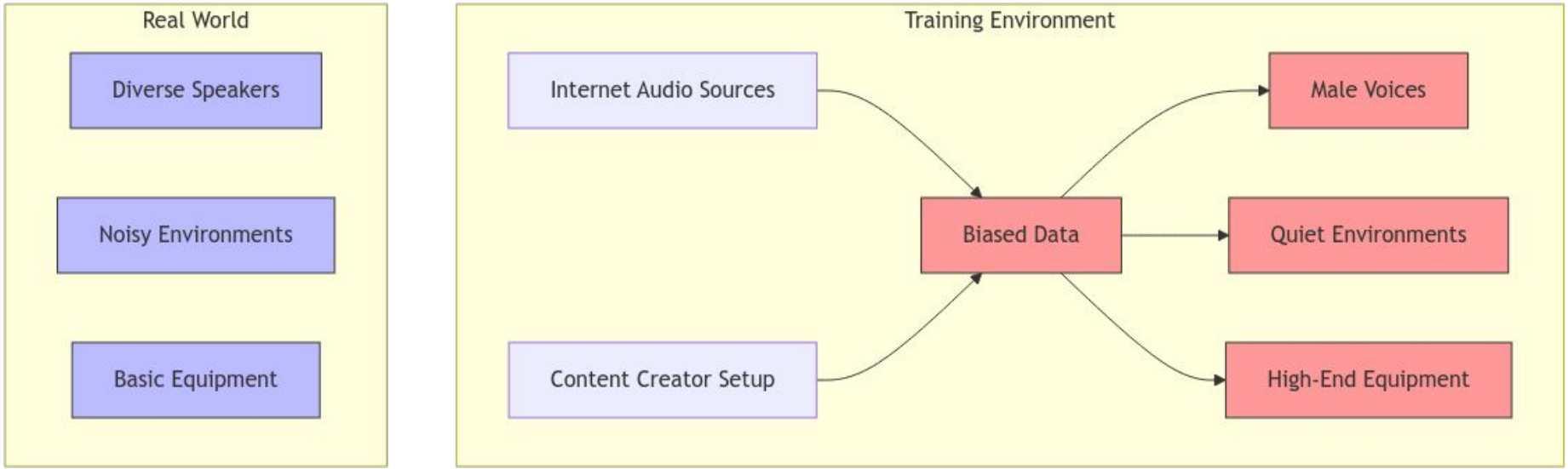
timestamp drifting / precision



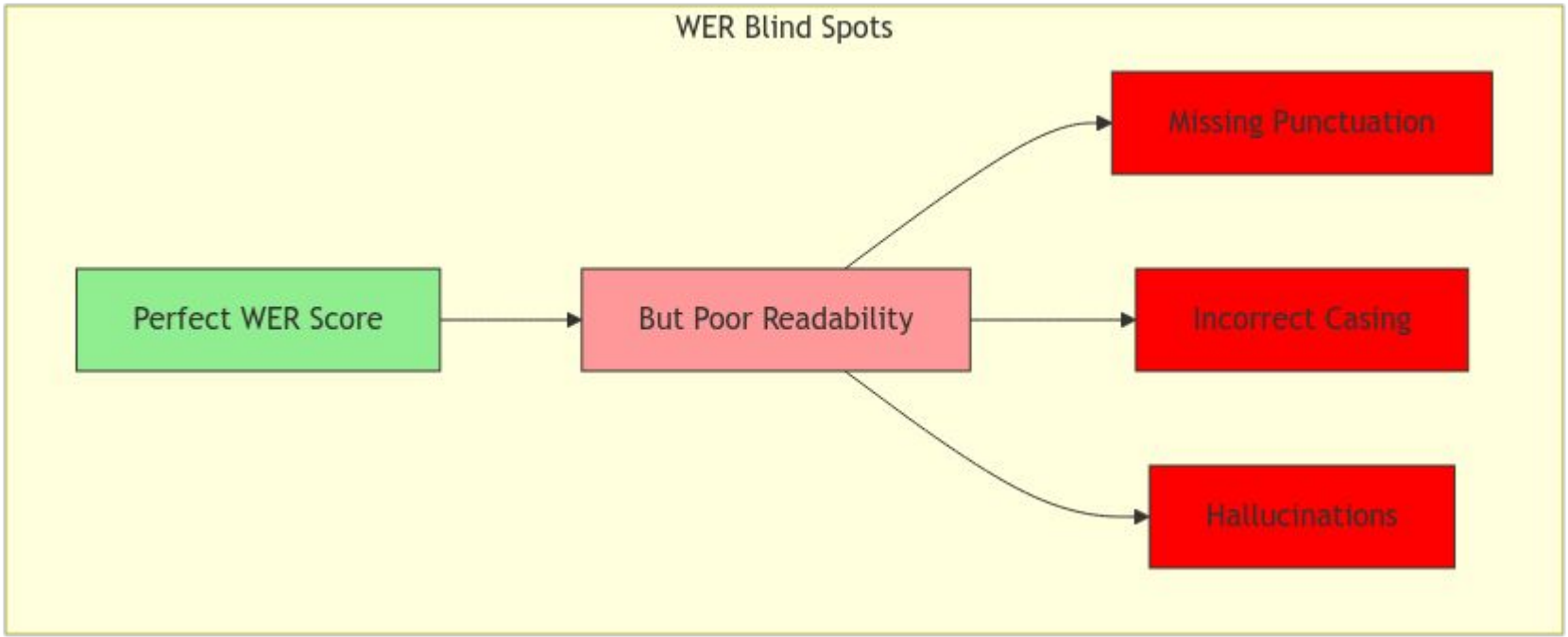
Whisper v2/3: The Hidden Issues
training biaises



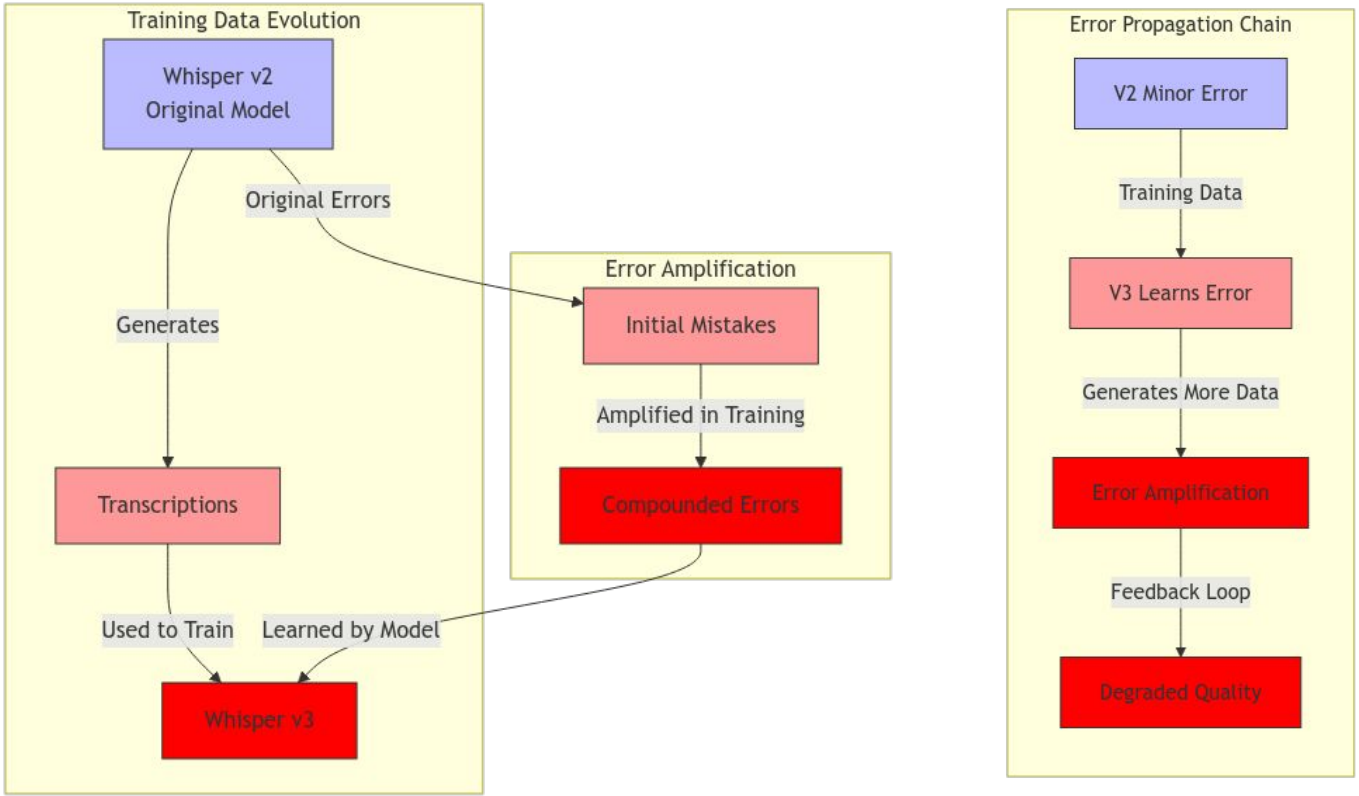
Whisper v2/3: The Hidden Issues
training biaises



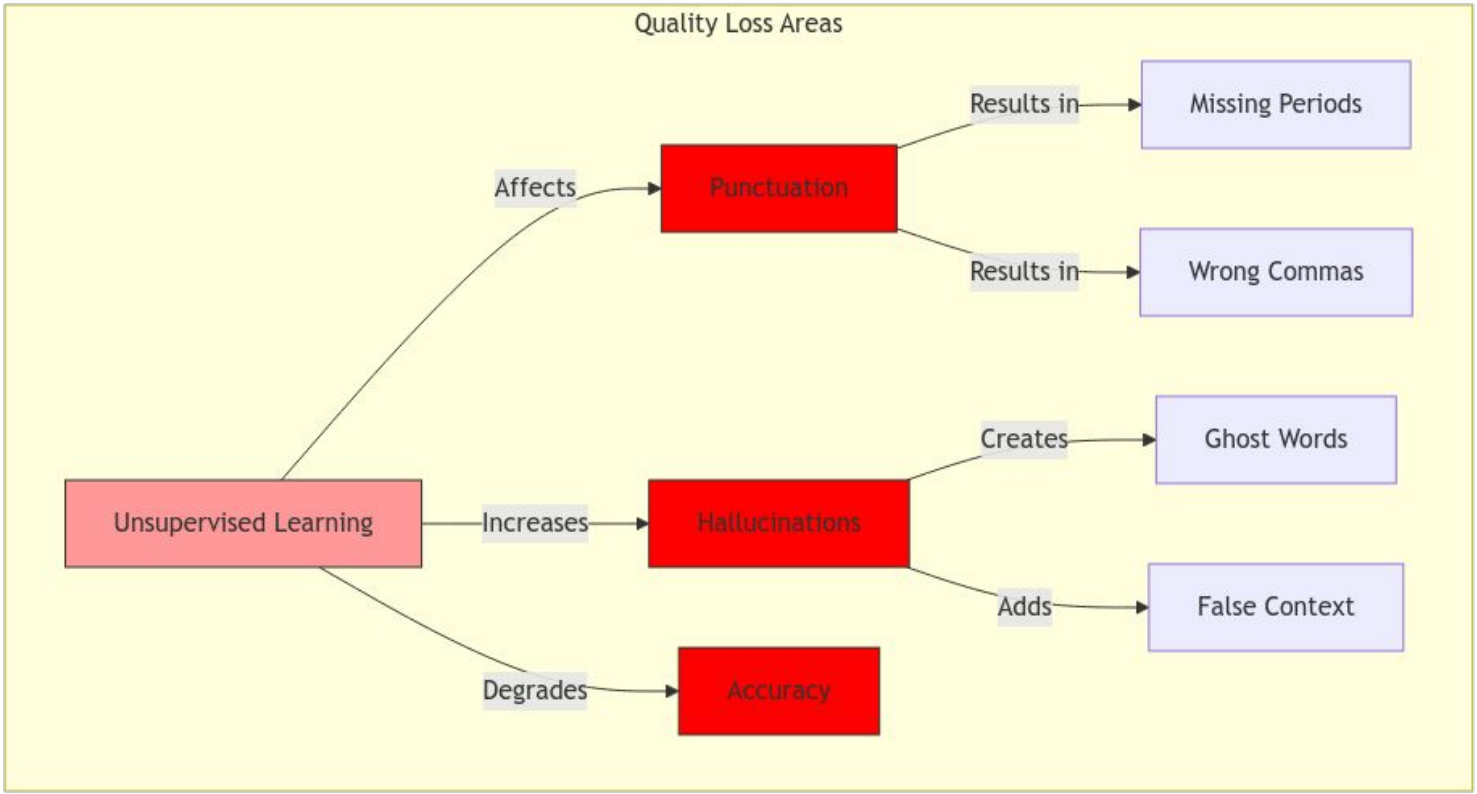
Whisper v2/3: The Hidden Issues
WER is the wrong metrics to optimize/evaluate



Whisper v3: The Hidden Issues
Error cascade effect



Whisper v3: The Hidden Issues
Error cascade effect



Thank you

@gladiaio
@jilijeanlouis

https://gladia.io