

Marina Castillo Ariza

```
getchar()
while (getchar() != '\n')
    continue;
```

# CLBot

Your trusted steed to dive into the  
world of Computational Linguistics!



# 1. It's getting meta up in here: a bot to learn Computational Linguistics

## Project Goal:

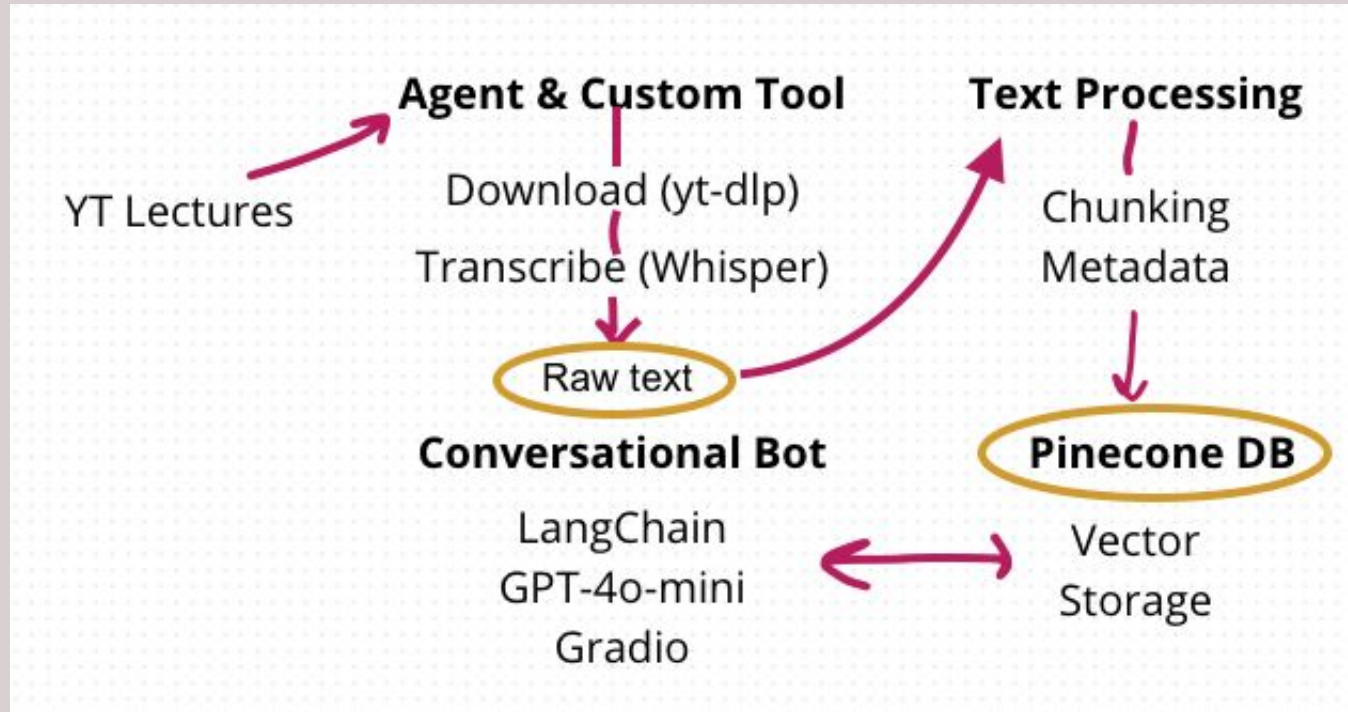
Create a **supportive tool** for students learning CL:

- Retrieves relevant information from actual university lectures
- Provides answers with **source verification**
- Highlights **key concepts** for better understanding

## Highlights:

- Reliable answers from **verified course content**
- Connects answers to specific lecture
- Makes complex CL topics more **approachable**

## 2. Project Architecture: From Lectures to Learning 🏗️



### 3. Dataset: high quality information

Source: First four lectures on Computational Linguistics in Stanford University.

Processing Pipeline:

1. YouTube lecture videos
2. Whisper transcription
3. Text chunking (1500 tokens, 200 overlap)
4. OpenAI Ada-002 embeddings
5. Metadata enrichment with:
  - Lecture numbers and titles
  - Key concepts
  - Topic categorization

## 4. Conversational bot: Core Components



### 1. Retrieval Chain

- ConversationalRetrievalChain
- k=8 documents per query
- Strict source verification

### 2. Memory System

- ConversationBufferWindowMemory
- Chat history tracking

### 3. Custom Prompt Engineering

- Helpful teaching assistant persona
- Only answers from lecture content
- Always references source material

### 4. Response Formatting

- Source Attribution
- Key Concepts Highlighting

## 5. Tools & Integrations

### 1. Vector Database (Pinecone)

- Storage & retrieval of lecture embeddings
- Metadata management
- Real-time querying

### 2. Custom Tool Development

BatchDownloadAndTranscribe

- YouTube downloading (yt-dlp)
- Audio processing
- Whisper transcription
- File management

### 3. Models

- OpenAI embeddings (Ada-002)
- GPT-4o-mini for response generation
- Gradio for web interface

## 6. Evaluation & Quality Assurance

### 1. LangSmith

- Retrieval from Pinecone
- Memory
- Model adequacy

### 2. Manual Inspection

- Quality of answers
- Clear format: text and interface

### 3. Improvements:

- Final version: strict retrieval information
- Enhanced prompt engineering
- Increased retrieval to 8 documents

## 7. Conclusions & Future Path

### Next Steps:

- 🎯 Add more course content
- 💡 Fine-tune model: improve latency
- ✨ Improve the interactions:
  - Interactive Quizzes
  - More proactive assistant

### A formative journey

- Creating the bot and refactoring
- Checking answers & refreshing concepts

Ask me anything about these topics!

Your Question	Assistant Response
what am I gonna do without my IH friends now that the bootcamp is over?!	I can't find specific information

**Send** **Clear**



Thank you so much for your  
attention and you support!



# Demo: memory test



## Computational Linguistics Teaching Assistant

Welcome! I'm your CS teaching assistant, trained on Stanford's lectures. I can help you understand:

- 📖 Word Vectors and Embeddings
- 🧠 Neural Networks in NLP
- 🌐 GloVe and Word Representations

Ask me anything about these topics!

### Your Question

What is cross entropy?

Send

Clear

### Assistant Response

The lectures discuss two types of evaluations for word vectors: intrinsic and extrinsic evaluations. Intrinsic evaluations focus on assessing the quality of word vectors based on their properties, such as cosine similarity between vectors and how well they capture human intuitions about word similarity. Extrinsic evaluations involve testing the performance of word vectors in downstream tasks, like machine translation or sentiment analysis, to determine if improvements in the word vectors lead to better overall system performance. The lectures emphasize the importance of isolating changes to ensure that any observed improvements can be attributed to the specific modifications made to the word vectors.

#### 📖 Sources:

- Lecture 1.0: Natural Language Processing with Deep Learning
- Lecture 2.0: Word Vector Representations: word2vec
- Lecture 3.0: GloVe: Global Vectors for Word Representation

#### 🔑 Key Concepts:

- CBOW, Evaluation, GloVe, NLP, Skip-gram, Word Embeddings, Word Vectors, Word2Vec



Not sure where to start?

☰ Try these questions



What is cross entropy?



How do neural networks work in NLP?



What is GloVe and what problem does it solve?

# Demo: memory test



## Computational Linguistics Teaching Assistant

Welcome! I'm your CS teaching assistant, trained on Stanford's lectures. I can help you understand:

- 📖 Word Vectors and Embeddings
- 🧠 Neural Networks in NLP
- 🌐 GloVe and Word Representations

Ask me anything about these topics!

### Your Question

How is it used in neural networks?

Send

Clear

### Assistant Response

In the context of neural networks, cross entropy is used as an objective function that we aim to minimize during training. The lectures explain that cross entropy quantifies the difference between the true class probabilities (represented as a one-hot vector) and the predicted probabilities from the model's softmax outputs. Specifically, the cross entropy error is calculated based on the negative log probability of the true class, which serves as a loss function in classification tasks. The goal is to maximize the probability of the correct class, which is equivalent to minimizing the negative log probability. This process involves using the softmax function to create a probability distribution from the model's outputs, and the cross entropy loss guides the training of the neural network by adjusting the weights to improve the model's predictions.

📖 Sources:

- Lecture 2.0: Word Vector Representations: word2vec
- Lecture 3.0: GloVe: Global Vectors for Word Representation
- Lecture 4.0: Word Window Classification and Neural Networks

🔑 Key Concepts:

- Backpropagation, CBOW, Classification, Evaluation, GloVe, Neural Networks, Skip-gram, Word Embeddings, Word Vectors



Not sure where to start?

☰ Try these questions



What is cross entropy?



How do neural networks work in NLP?



What is GloVe and what problem does it solve?

# Demo: memory test



## Computational Linguistics Teaching Assistant

Welcome! I'm your CS teaching assistant, trained on Stanford's lectures. I can help you understand:

- 📖 Word Vectors and Embeddings
- 🧠 Neural Networks in NLP
- 🗣️ GloVe and Word Representations

Ask me anything about these topics!

### Your Question

What other loss functions are mentioned in the neural networks lecture?

Send

Clear

### Assistant Response

The context mentions the max margin loss as another loss function in addition to cross entropy loss. It is described as a powerful and robust loss function that is often more effective than cross entropy error and softmax. This is discussed in the context of training objectives for distinguishing between true and corrupt windows in the neural network model.

#### 📖 Sources:

- Lecture 2.0: Word Vector Representations: word2vec
- Lecture 3.0: GloVe: Global Vectors for Word Representation
- Lecture 4.0: Word Window Classification and Neural Networks

#### 🔑 Key Concepts:

- Backpropagation, CBOW, Classification, Evaluation, GloVe, Neural Networks, Skip-gram, Word Embeddings, Word Vectors



Not sure where to start?

≡ Try these questions



What is cross entropy?



How do neural networks work in NLP?



What is GloVe and what problem does it solve?

# Demo: strict retrieval

## 🎓 Computational Linguistics Teaching Assistant

Welcome! I'm your CS teaching assistant, trained on Stanford's lectures. I can help you understand:

- 📖 Word Vectors and Embeddings
- 🧠 Neural Networks in NLP
- 🗄️ GloVe and Word Representations

Ask me anything about these topics!

### Your Question

What is BERT?

Send

Clear

### Assistant Response

I can't find specific information about this in the course materials.

💡 Not sure where to start?

☰ Try these questions

🤔 What is cross entropy?

🌟 How do neural networks work in NLP?

🔍 What is GloVe and what problem does it solve?

# Demo: strict retrieval

## 🎓 Computational Linguistics Teaching Assistant

Welcome! I'm your CS teaching assistant, trained on Stanford's lectures. I can help you understand:

- 📖 Word Vectors and Embeddings
- 🧠 Neural Networks in NLP
- 🗄️ GloVe and Word Representations

Ask me anything about these topics!

### Your Question

What methods for word representation ARE covered in the lectures?

Send

Clear

### Assistant Response

The lectures cover several methods for word representation, including the bag of words approach, which averages the vector representations of individual words to create a sentence vector representation. Additionally, the course introduces the word2vec model, which is a method for learning word vectors that encode word meanings. There are also mentions of more complex models such as recurrent neural networks, recursive neural networks, and convolutional neural networks, which may be explored later in the course. Overall, the focus is on how these representations can be used to capture the meaning of words and sentences in natural language processing tasks.

#### 📖 Sources:

- Lecture 1.0: Natural Language Processing with Deep Learning
- Lecture 2.0: Word Vector Representations: word2vec

#### 🔑 Key Concepts:

- CBOW, NLP, Skip-gram, Word Vectors, Word2Vec

💡 Not sure where to start?

≡ Try these questions

😬 What is cross entropy?

🌟 How do neural networks work in NLP?

🔍 What is GloVe and what problem does it solve?