Conversationally

AI-BASED CONVERSATIONAL TUTOR

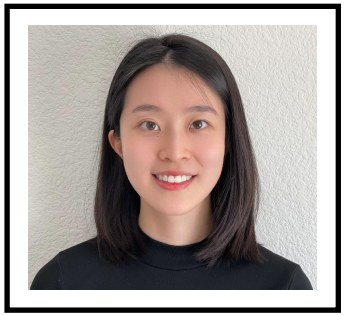
Fall 2023 MIDS Capstone Project

by Aastha, Isabel, Jess, Mon, Ram

Our Team

Aastha Khanna

Engineering/ML Ops



Isabel Chan

Data Engineering

Jess Matthews

Product Management/UX

Mon Young

Project Management

Ram Senthamarai

Data Science

Agenda

- Market Potential
- Why Conversationally?
- Demo
- Models
  - Model Overview
  - Model 1: Keep the Conversation on Topic
  - Model 2: Identify Grammatical Errors
  - Model 3: Generate a Scaffolding Response

Market Potential

Growing Demand In Language Learning

2023 Statistics

$400M (revenue) Duolingo, 50M paid subscribers

$250M (revenue) Babbel, 10M paid subscribers

- ~92M Language Learners in the US
- 10% adoption rate =
- $1B in revenue/year at $10 a month.

2028 Projection

$190B 18.3% CAGR* in language market.

*CAGR: Compound Annual Growth Rate

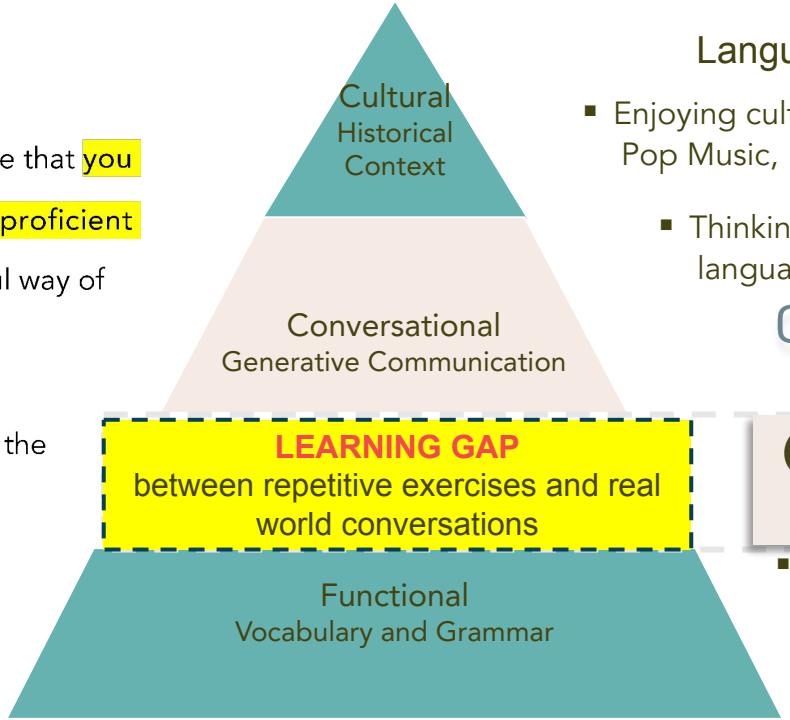
2028 is a Pitchbook market estimate. 92M is a bottoms up estimate based on numbers in each user profile, see slide 11. Revenue from publicaly shared company information.

Problem Space

Transition to Conversation Fluency

"Duolingo gets criticism from some that you cannot learn enough to become proficient in a language…but it's a wonderful way of getting people started."

Duolingo: Teaching Languages to the Masses, Harvard Business Review



Language Learning Goals

- Enjoying cultural works such as Movies, Pop Music, Literature
- Thinking and responding in the target language

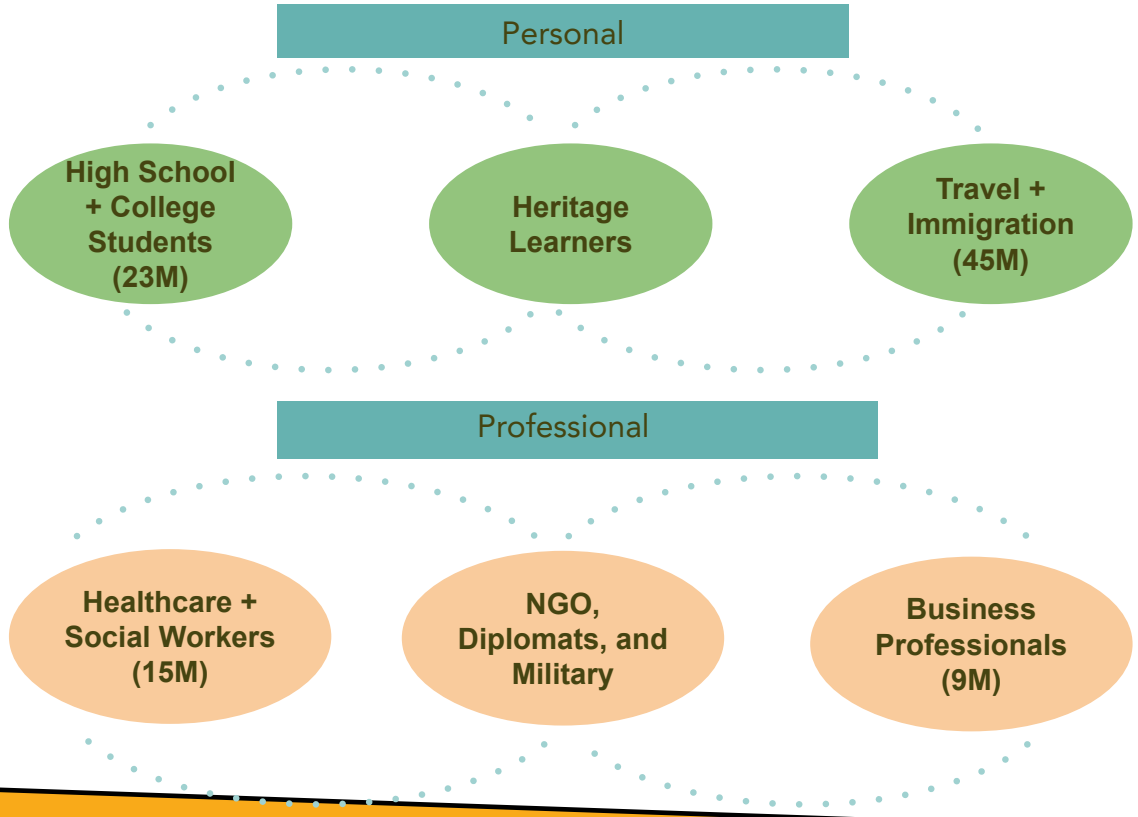
italki HelloTalk Langotalk

Conversationally
AI-BASED CONVERSATIONAL TUTOR

- Memorization through repetition and games

duolingo Rosetta Stone

Target Users



MVP Target Persona:

- *Fluent English speakers with personal language learning goals.*
- *Beginner Spanish learners who are uncomfortable holding conversations.*

Why Conversationally?

NCEDS: High School and Beyond; CBO: The Foreign-Born Population, the U.S. Economy, and the Federal Budget; BLS: Healthcare Workers
Zippia: Licensed Social Worker Demographics in the US; America Council on the Teaching of Foreign Languages: Making Languages our Business:
1 in 4 business rely a lot on worker's foreign language skills * 24M professional population with language skill needs

Why Conversationally?

How We Bridge the Learning Gap

Capabilities

- **Chatbot** with **SMS** user interface**.**
- **Curated microlessons.**
- **Immediate feedback** based on **proven pedagogy.**

Features

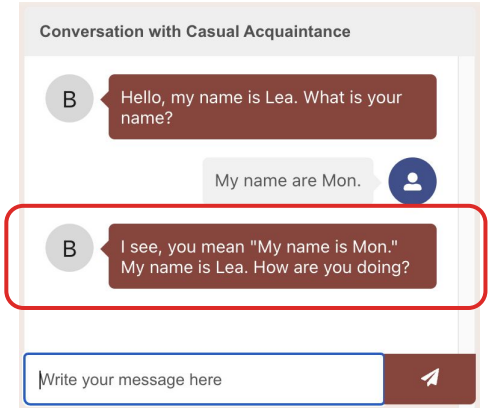
- **Grammatical error detection** with explanations.
- Cue-based feedback using **scaffolding learning**.
- Natural language responses with the learner.

Conversationally Uses Scaffolding Learning

We provide feedback using proven teaching methods.

"The scaffolding learning method works well for beginner conversation learners. It assists students with *encouraging feedback* and *subtle hints*, [maintaining] a *smooth conversation flow* and enhancing overall *satisfaction*."

– Stanford Education Professor Lee Dennig



Demo

Talk to us at conversationally.app

Demo

Images in seconds 4-15 generated using Dall-E, and voices generated using AWS Polly.

Promising Early User Feedback

8 user testing sessions conducted 12/6/23 - 12/9/23.

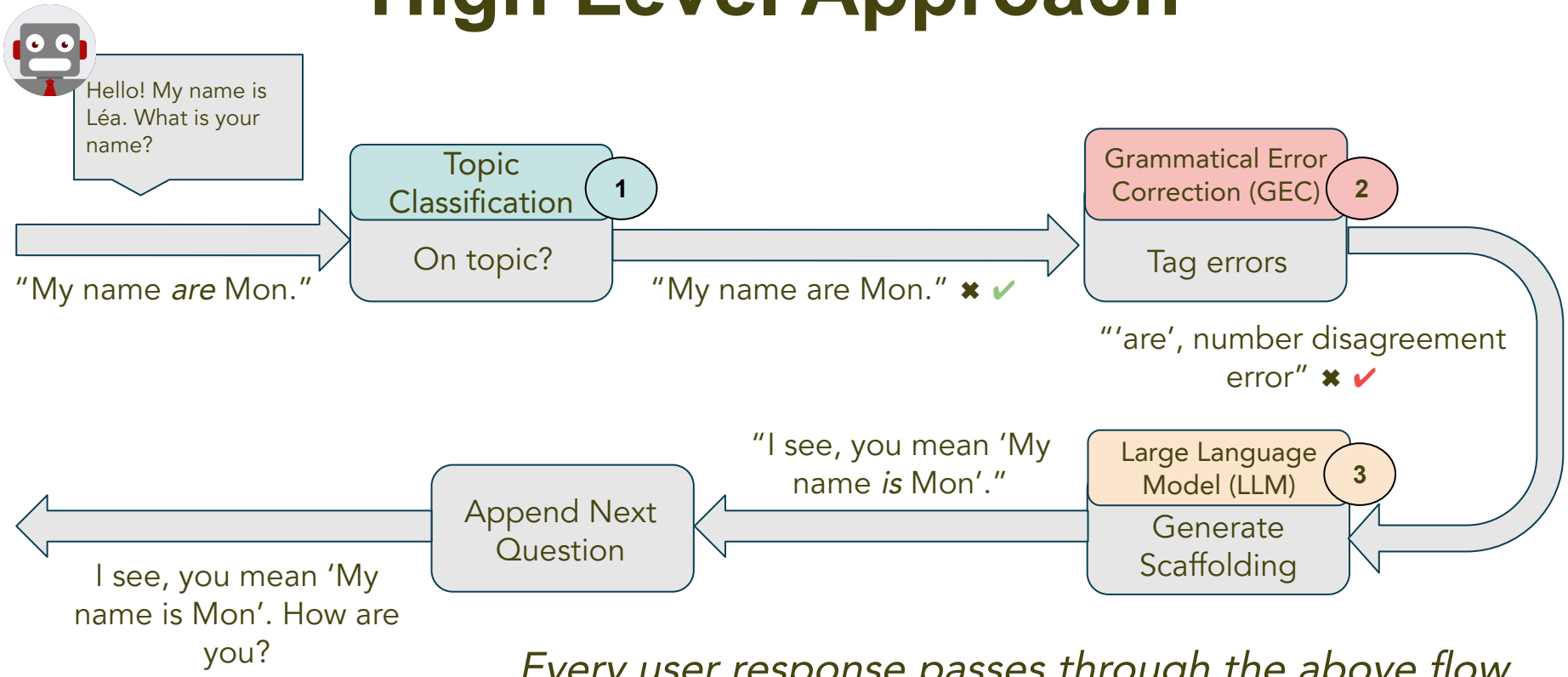
Are you able to hold a brief conversation in Spanish?	
Pre-session	**Post-session**
3 out of 8 users said they could hold a brief conversation in Spanish.	**Learning Gain**: **7 out of 8** able to hold a brief conversation in Spanish. **Positive Experience:** **7 out of 8** would continue using Conversationally.

Models Overview

Complex Sets of Tasks Require Ensemble Methods

Task/Model 1	Keep the Conversation on Topic
Task/Model 2	Identify Grammatical Errors
Task/Model 3	Generate a Natural Language Scaffolded Response

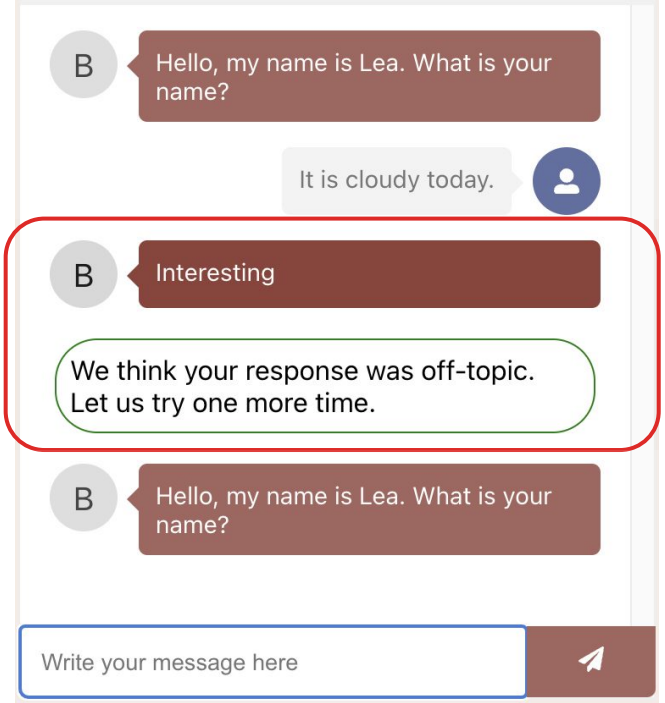
High Level Approach



Every user response passes through the above flow.

Model 1: Keep the Conversation On Topic

Model 1: Keep the Conversation On Topic



Why	Keep user on track with the conversation flow to - reduce LLM hallucination - keep learner focused on learning objective
Approach	Calculate **cosine similarity** from sentence embeddings to identify the semantic meaning
Model	**Sentence Transformer** (multi-qa-MiniLM-L6-cos-v1)
Metrics	**F1 Score, Accuracy**
Dataset	**Custom created 240 Q&A** pairs according to practice scenarios and annotated as on or off topic
Challenges	No available datasets

Evaluation

BOT: What is your name?

The goal is to find a **decision threshold** that maximize **Accuracy and F1**

Dataset - Input	Cos Similarity	*Decision Threshold*	True Label	Predicted Labels (0.4)	Predicted Labels(0.3)
My name is Mon.	0.92		1	1	1
I are Mon.	0.49	<0.4	1	1	1
My friend's name is Mon.	0.45		0	0	1
I goes by Mon.	0.32	<0.3	1	0	1
I like this class.	0.11		0	0	0

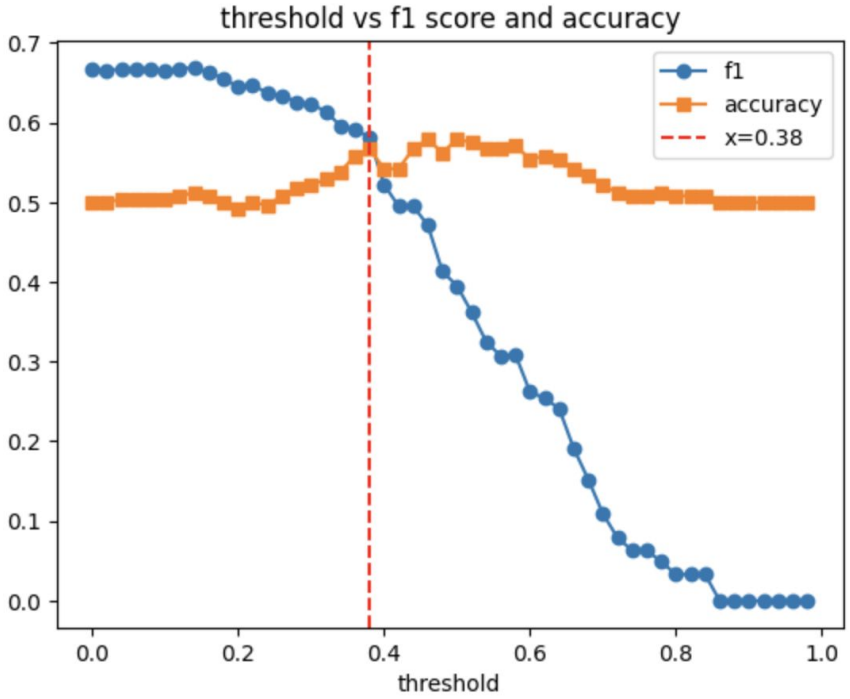
Evaluation

Accuracy: 100%
F1: 66%

Accuracy: 75%
F1: 80%

Different decision threshold has tradeoff between Accuracy and F1

Evaluation Result

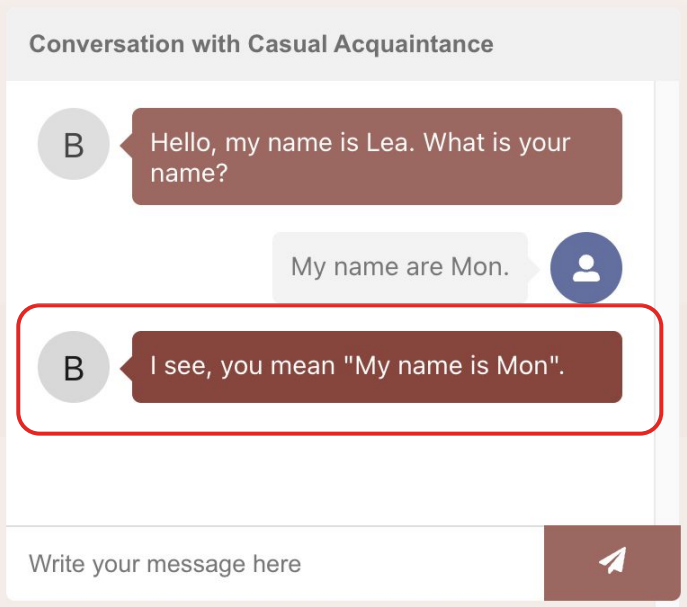


Observation

- Baseline for accuracy is 0.5 and accuracy is the highest when the decision threshold between 0.38 to 0.6
- F1 decreases in a faster rate starting at 0.38 because the recall is dropping in a faster rate compare to the increasing in precision.
- Similarity Decision Threshold: 0.38
  - with the highest accuracy and F1 score (accuracy: 0.567, F1: 0.581)

Model 2: Identify Grammatical Errors

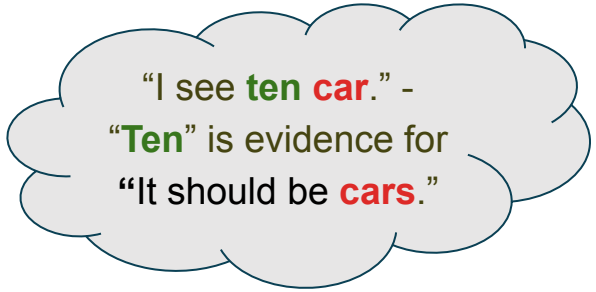
Model 2: Identify Grammatical Errors

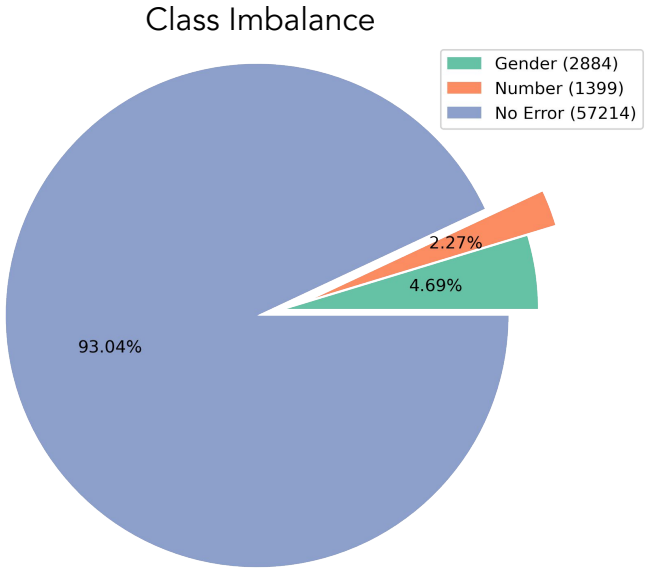


Why	• Keeps LLM responses predictable and on-topic. • Makes language learning easier.
Approach	**Classify each word into one of 3 classes** - No error, Number error, Gender error.
Model	Fine tuned **Beto**, a Spanish BERT model
Metrics	**Macro Average F1 Score**, Accuracy
Loss	**Weighted Categorical Cross Entropy Loss**
Dataset	Fine tuning on **COWS-L2H**
Challenges	Class imbalance, Feature engineering - evidence words

COWS-L2H Dataset

- Dataset has
  - 1725 student essays in Spanish
  - Multiple errors annotated
- What we did
  - Extracted 5223 sentences with one error in each
  - 93% of words do not have any error
  - Feature engineering using ChatGPT for **Evidence** words





Classification Report

The model does great with words without errors but not as good on the other two classes. Below are the metrics on the test dataset of 260 sentences.

- Overall macro avg F1 score is 0.90
- Word Level Test Accuracy of 98% and sentence level accuracy of 80%
- Class level F1-scores:
  - Correct words: 0.99
  - Gender mismatch: 0.89
  - Number mismatch: 0.83
- Kappa Score: 0.87

	Precision	Recall	F1 Score	Support
Gender error	0.88	0.89	**0.89**	155
Number error	0.81	0.84	**0.83**	68
No Error	0.99	0.99	**0.99**	3121
Micro Avg			0.98	3344
Macro Avg	0.90	0.91	**0.90**	3344
Weighted Avg	0.98	0.98	0.98	3344

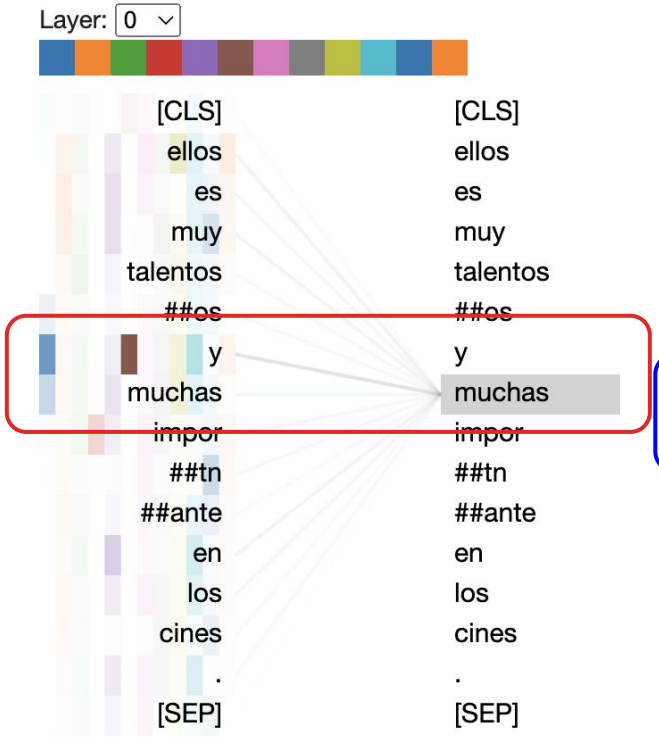
What Model Attended To

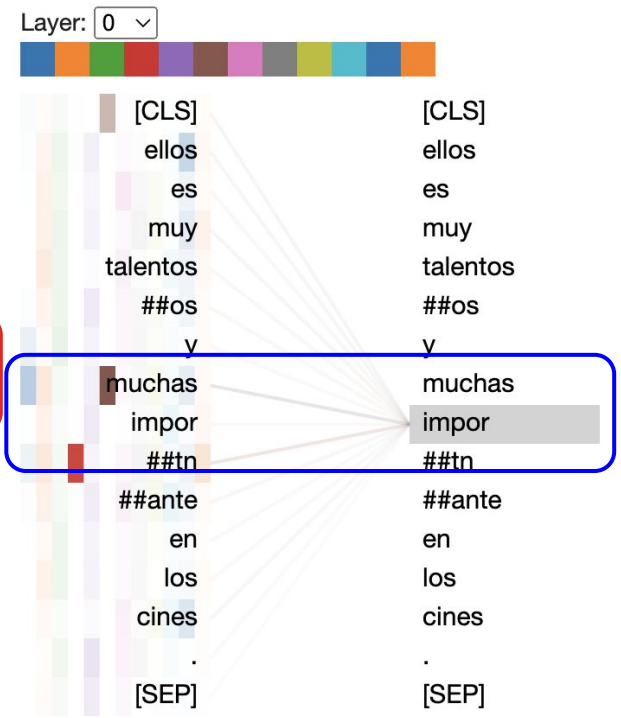
Input
Ellos… y muchas importante…

Ground Truth
Ellos… y muchas importantes...

Predicted
Ellos… y mucha importante…

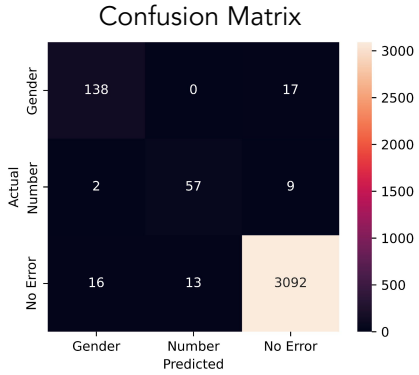
ChatGPT
“The correct form should be ‘Ellos son **muy talentosos y muy importantes** en los cines.’"

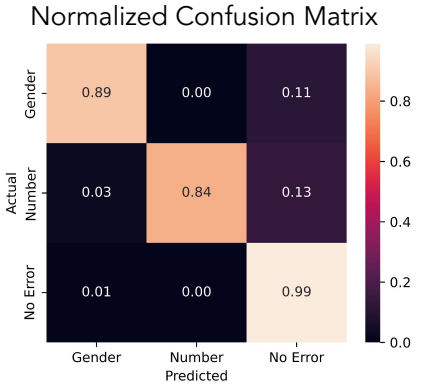




More Observations

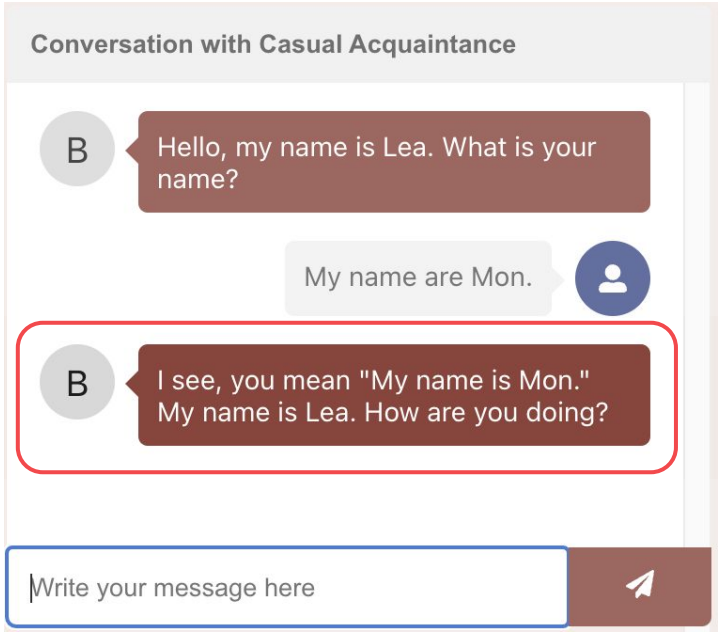
- Often model considers 'Y' (Spanish equivalent of and) as a boundary and pays less attention to other side of the sentence
- More generally, complex sentences with lot of dependencies that are farther apart confuse the model
- Sometimes it gets mixed up on the word to fix. "Mucha importante" or "muchas importantes"? It seems to fix the second word more often.
- Confusion matrix shows that model is mixing up "no error" with other two classes more often.





Model 3: Generate a Natural Language Scaffolded Response

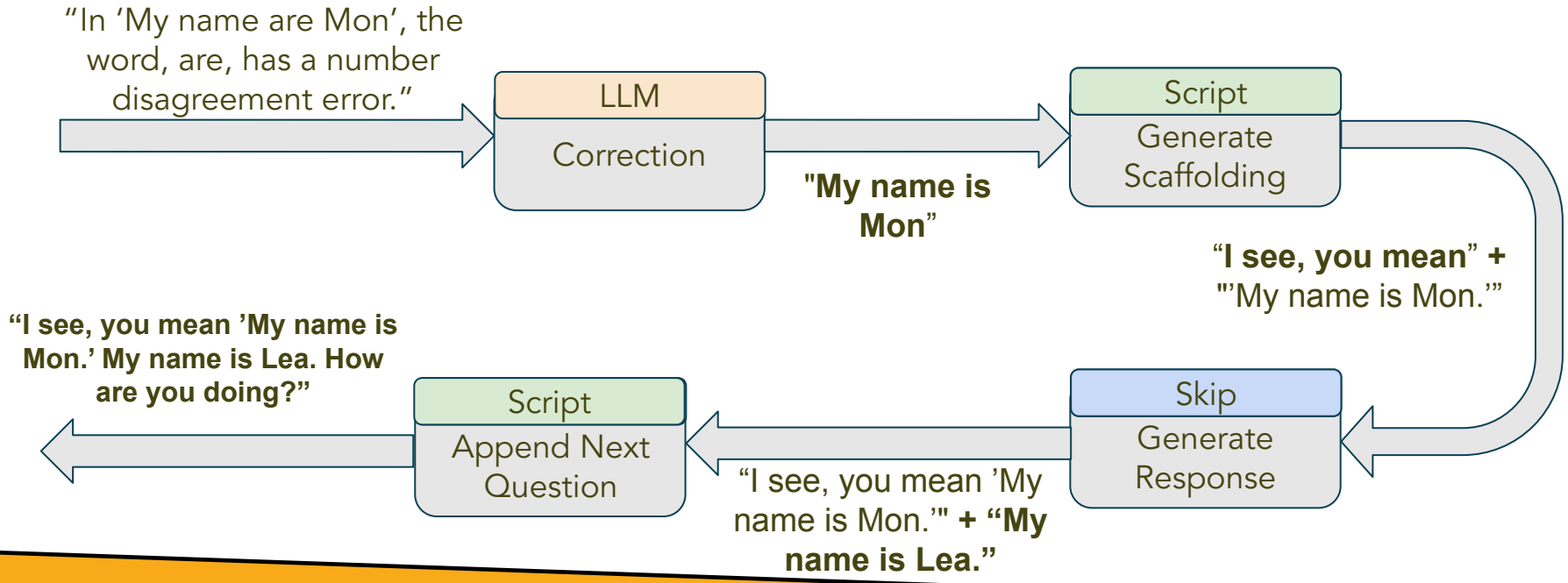
Model 3: Generate a Natural Language Scaffolded Response



Why	Utilizes user input and results from GEC to generate natural language responses with a scaffolding learning approach.
Approach	Using **prompt engineering and LLMs**
Model	Tried various LLM 7b local models ⇒ **Mistral 7b**
Metrics	Survey result
Dataset	None
Challenges	Less than ideal output from Mistral 7b with its soft instructions.

Multi-Stage to Overcome 7b LLM

Chain Small LLM Instructions, Reduce Hallucinations

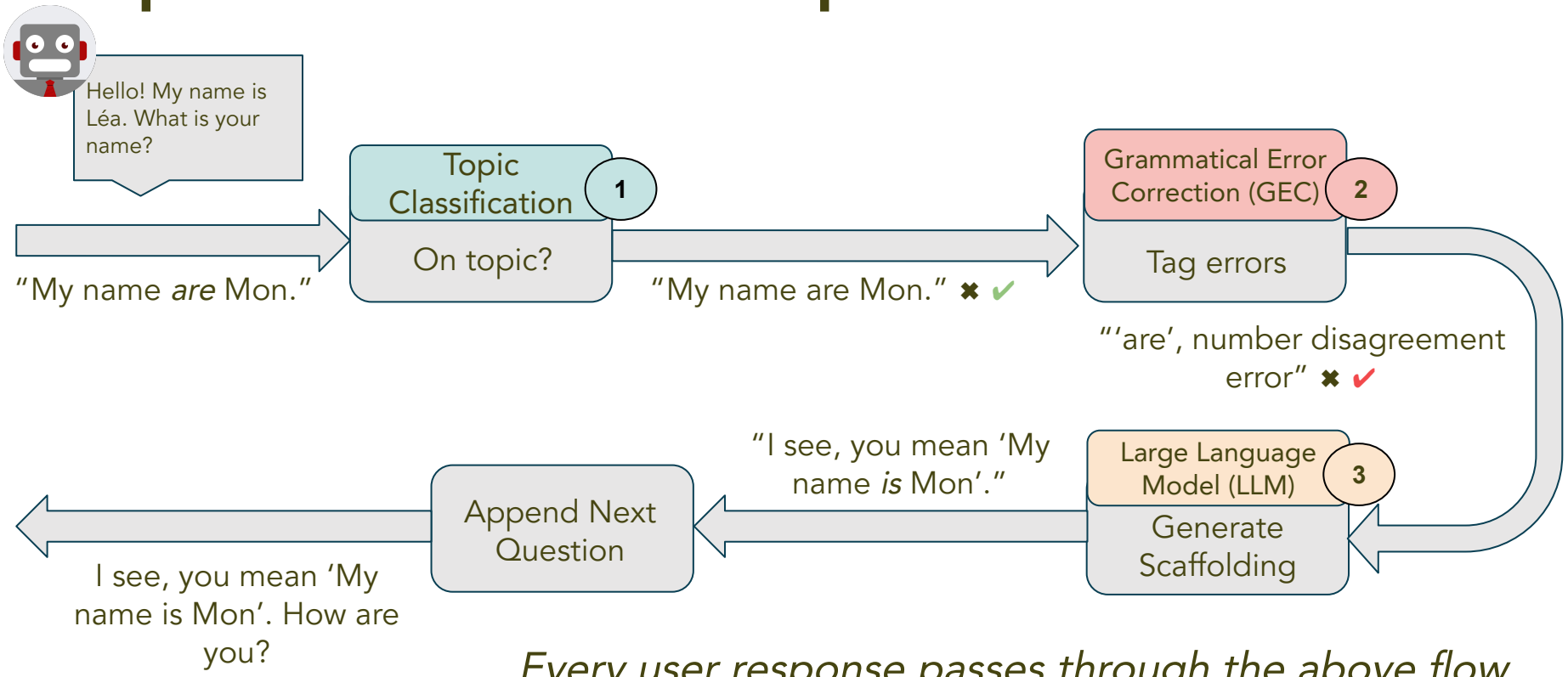


Human Evaluation of LLM Performance

A correct response must meet these three criteria		
eval criteria	performance	
correct error	56%	😐
respond ethically	100%	😊
scaffolding response	100%	😊

5 rounds, total 25 responses

Complex Sets of Tasks Require Ensemble Methods



Every user response passes through the above flow.

Next Steps

Prioritize	Additional Feature
Improve LLM responses	Accept audio input and voice output
Support more error types, like conjugation errors	Create learning curriculums with additional microlessons
Add evidence words to summary explanation	Personalize language learning goal

Bridging the Learning Gap

From Functional to Conversational

Address the fear of making mistakes
By providing positive conversational experiences early in the learning journey.

Build confidence
Sequence feedback and learning goals to increase exposure to common exchanges and delay learning edge cases.

Increase accessibility
By providing always on language tutor to support conversational practice.

ACKNOWLEDGEMENTS

Joyce Shen

UC Berkeley Professor

Kira Wetzel

UC Berkeley Professor

Lee Dennig

Linguistic Professor

Mark Butler

UC Berkeley Professor

Prabhu Narsina

W210 Teaching Assistant

Thank you

Talk to us at conversationally.app

GOAL: Generate Cue and Response

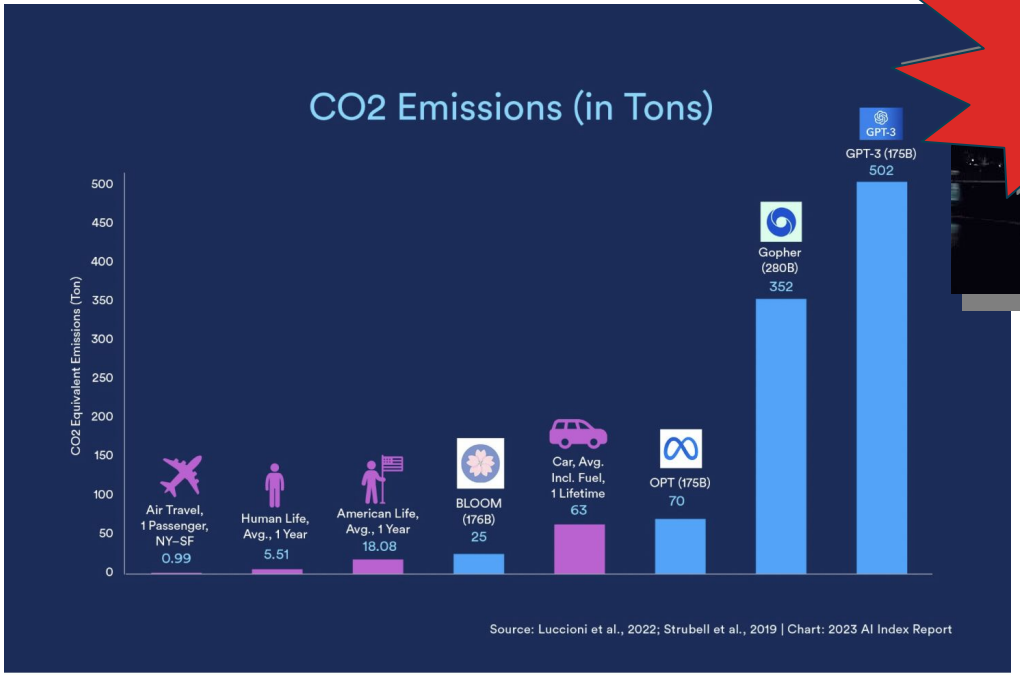
Approach: LLMs and few shot learning with output from GEC model

chatgpt 3.5	llama2 7b chat	falcon 7b	falcon 7b instruct	falcon 40b instruct	aguila 7b	clibrain 7b	zephyr 7b
OpenAI API	Meta's 7 billion parameters pretrained model	it handles languages such as English and Spanish	finetuned on a mixture of chat/instruct datasets	falcon 7b instruct's big brother	falcon 7b fine-tune with a mixture of Spanish, Catalan and English data.	llama 2 7b fine-tuned on Clibrain's Spanish instructions dataset.	Mistral 7b fine-tuned on a mix of publicly available & synthetic datasets
cost can be high, not environmentally friendly	can only respond in English	does not follow instruction	problem with complex instruction	model size is big, not environmentally friendly	speaks in Spanish but does not follow instruction	does a better job than other models	does the best job so far

Cue and Response Generation

Using Mistral 7b

Cost
$0.03c vs $3c
Privacy
Personalization
Climate Impact
CO2, Power & Water



1.287 GWh

Image Courtesy of Stanford University's HAI Institute. 2023 AI Index report

cost: https://tomtunguz.com/gm-saas/?utm_source=tldrai,
climate: https://aws.amazon.com/aws-cost-management/aws-customer-carbon-footprint-tool/

Content Classification

Goal: Determine whether input is on topic or not.

Approach: Calculate cosine similarity from sentence embeddings to identify the semantic meaning

Model: Sentence Transformer (multi-qa-MiniLM-L6-cos-v1)

Metrics: F1 Score, Accuracy

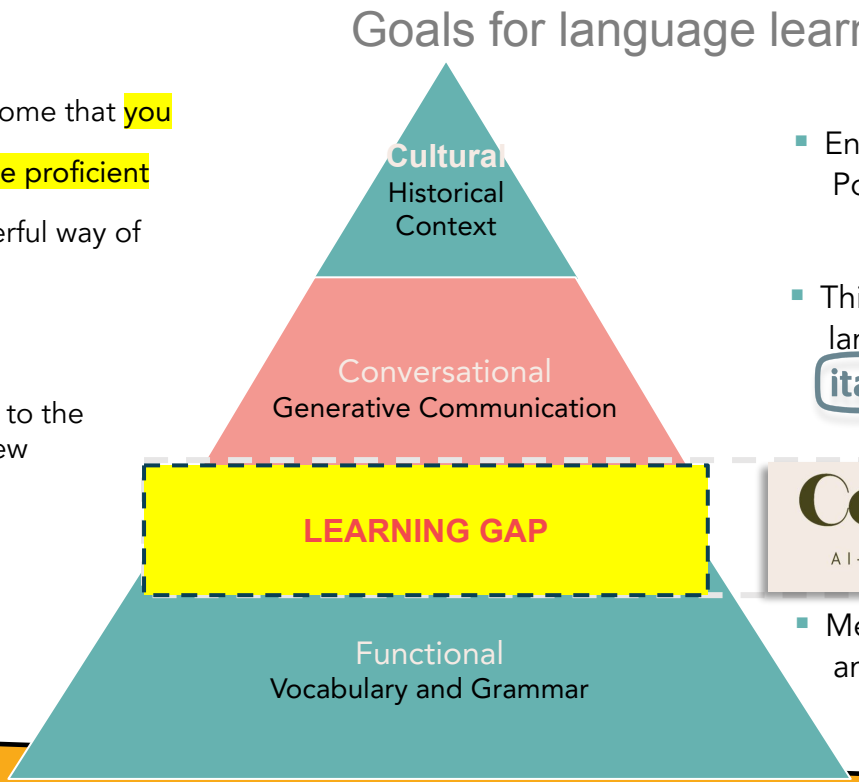
Dataset: Custom created 400 Q&A pairs according to practice scenarios and annotated as on or off topic

Decision threshold: Optimal cutoff to maximize accuracy and F1 score on custom dataset

PROBLEM SPACE

"Duolingo gets criticism from some that you cannot learn enough to become proficient in a language…but it's a wonderful way of getting people started."

Duolingo: Teaching Languages to the Masses, Harvard Business Review



- Enjoying cultural works such as Movies, Pop Music, Literature
- Thinking and responding in the target language

italki HelloTalk Langotalk

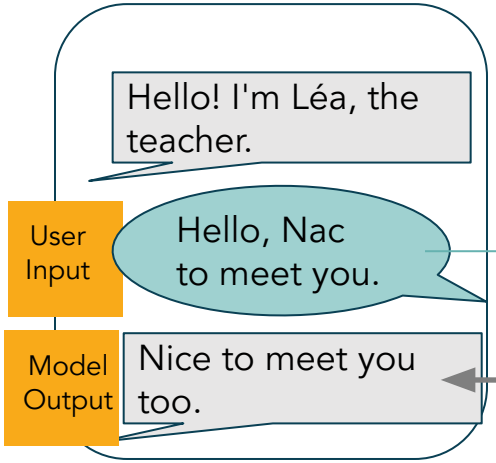
Conversationally
AI-BASED CONVERSATIONAL TUTOR

- Memorization through repetition and games

duolingo Rosetta Stone

BEHIND THE SCENES

Flask App



Models

Content Classification	**Grammatical Error Correction Model**	**LLM indirect cue-based learning**
Is the response on topic?	Is the response correct?	How do I tell the user?

DATA SETS

Models

Content Classification	Grammatical Error Correction Model	LLM indirect cue-based learning

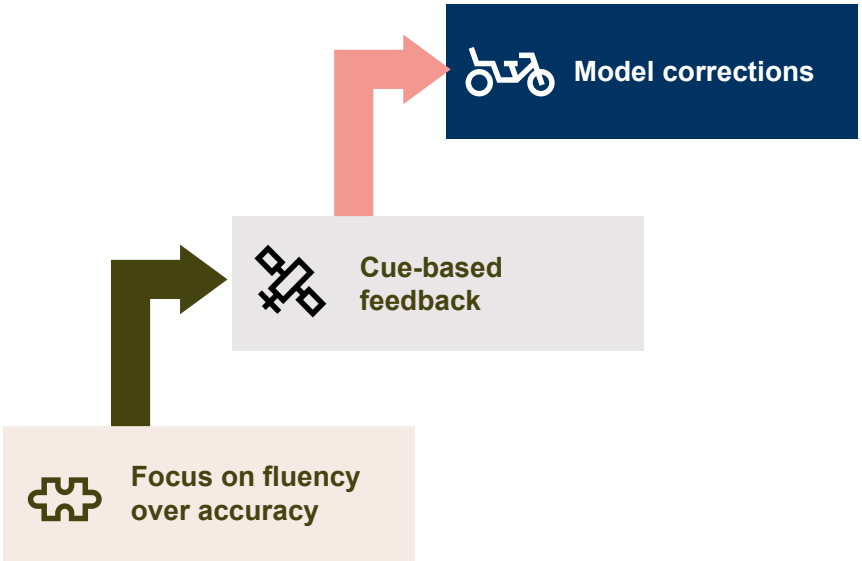
• 400 questions and answers pairs created by us manually
• Annotated by us as on-topic or off-topic
	• COWS-L2H - 1725 student essays on 8 topics
	• Annotated for
	  • Gender disagreement
	  • Number disagreement
	  • Missing Article
	• Augmented by us with LLM based evidence words and parsing tree dependencies
		• Manually created lesson scripts, learning objectives, and response directions

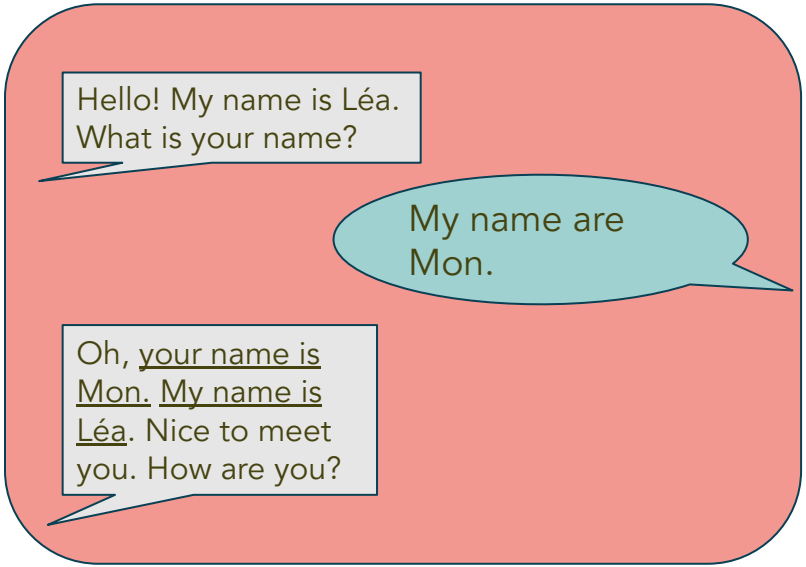
EVALUATION METRICS

Models

Content Classification	Grammatical Error Correction Model	LLM indirect cue-based learning
• Accuracy • F1 Score • False Positive: Conversation continues--User response is off topic, but model classifies it as on topic. • False Negative: User is prompted again--Response is on topic, but model treats it as off topic.	• Accuracy • F1 Score	• Definition of a quality response: 1. If there is an error, the correct answer is modeled 2. No inappropriate language was used 3. A next step in the conversation is provided

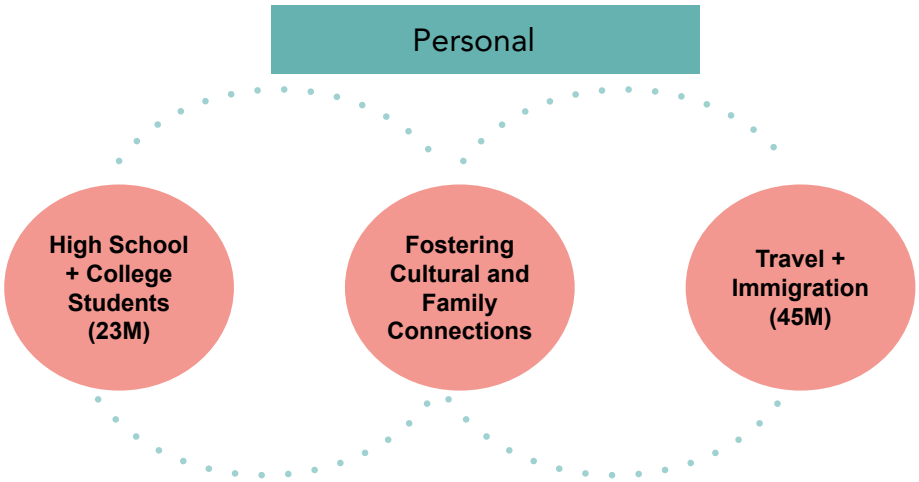
SCAFFOLDING LEARNING





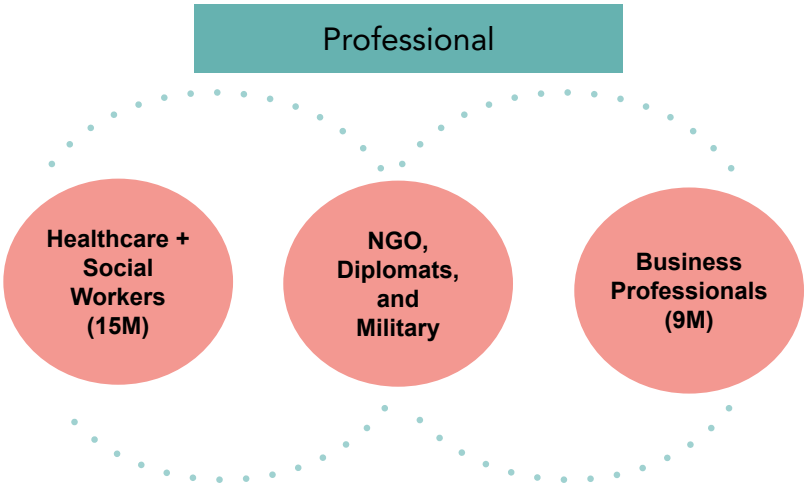
Explanation: “are” should be “is”, singular

LANGUAGE LEARNER PERSONAS



Target Persona for MVP

- Fluent English speakers with personal language learning goals
- Beginner Spanish learners who are not yet comfortable holding conversations



User overview

- ~92M US Language Learner Population, 27% of US
- Personal market is larger with more budget control but both are sufficient populations
- 10% user adoption is 9M customers

NCEDS: High School and Beyond; CBO: The Foreign-Born Population, the U.S. Economy, and the Federal Budget; BLS: Healthcare Workers
Zippia: Licensed Social Worker Demographics in the US; America Council on the Teaching of Foreign Languages: Making Languages our Business:
1 in 4 business rely a lot on worker's foreign language skills * 24M professional population with language skill needs

USER PROBLEM AND METHODS

User Problem

The gap between repetitive exercises and real-world conversation…

Leads to a loss of motivation…

Due to a lack of progress…

And Confidence.

Methods

Interactive, context-relevant dialog practice

Focuses on **conversational flow rather than accuracy**

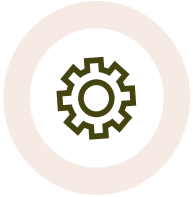
Uses proven scaffolding learning method

Successful Transition to conversations with native speakers

BRIDGING THE LEARNING GAP

Immersion is the fastest way to learn a language.

Consistent conversational practice is the closest learning experience to immersion.



Address the fear of making mistakes
By providing positive conversational experiences early in the learning journey.

Build confidence
Sequence feedback and learning goals to increase exposure to common exchanges and delay learning edge cases.

Increase accessibility
By providing always on language tutor to support conversational practice.

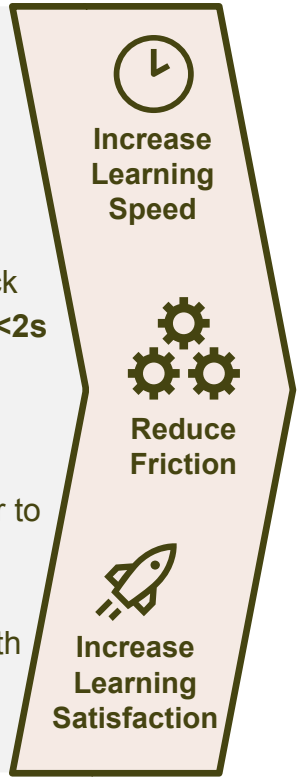
THE MVP

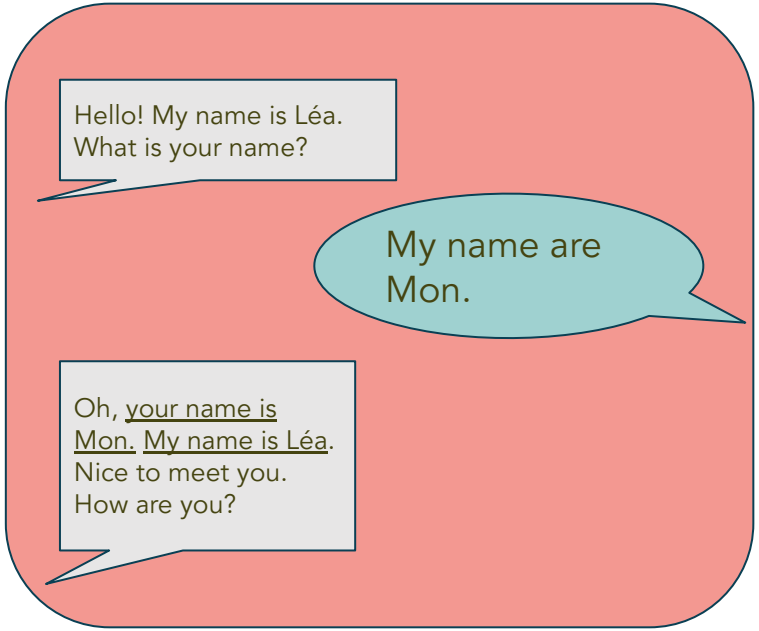
Capabilities

- Web app chatbot allows users to practice conversations using **modeled microlessons.**
- Users receive immediate feedback on successes and mistakes with **<2s** latency.

Features

- Cue-based feedback using **scaffolding learning** guides user to right answer and keeps the conversation flowing.
- **Grammatical error detection** with explanations.



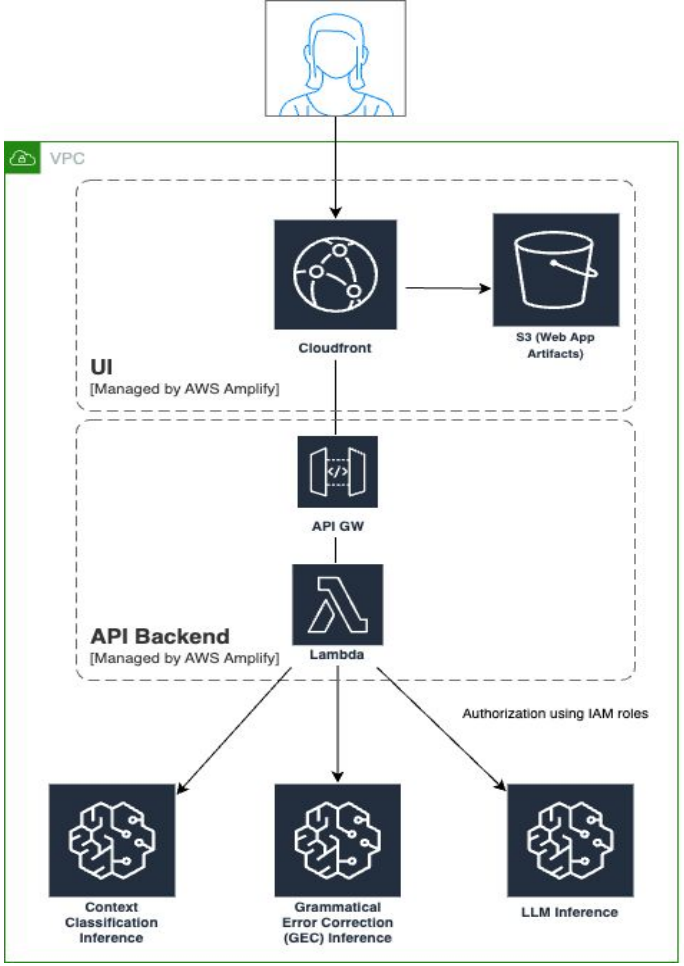


Scaffolding learning demonstrates how to complete the task.

Architecture

- **Frameworks:**
  - AWS Amplify
  - React (UI)
  - Flask (API)

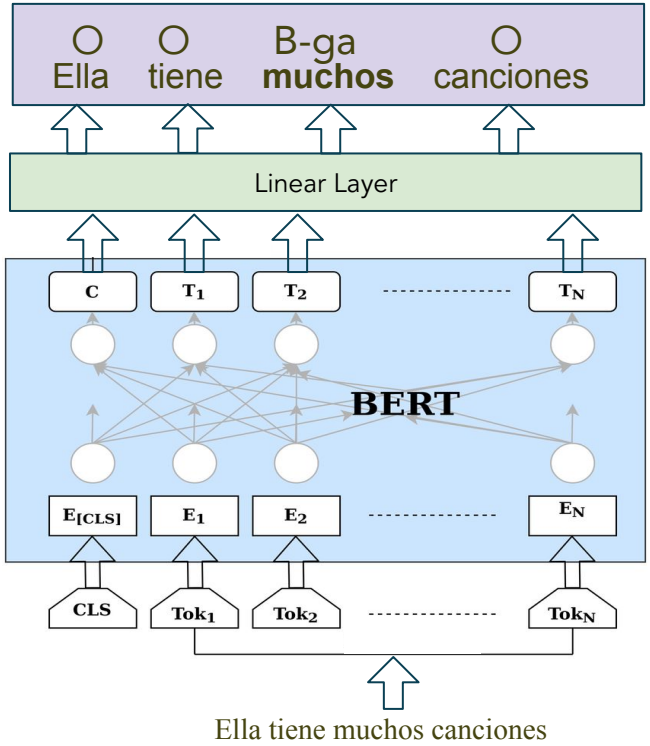
- **AWS Services:**
  - Amplify
  - API Gateway (Flask API access)
  - Lambda (Flask API)
  - Sagemaker (model inference)
  - S3 (web app artifacts)
  - IAM (access control)



Error Detection and Classification

Model Architecture

- Linear classification layer on top of the 12th layer of BERT



COWS-L2H Dataset

- To train the GEC model we needed a dataset with errors annotated.
- COWS-L2H dataset has about 1700 student essays in Spanish, annotated for following errors:
  - Gender disagreement
  - Number disagreement
  - Missing Article
  - Gender and Number disagreement

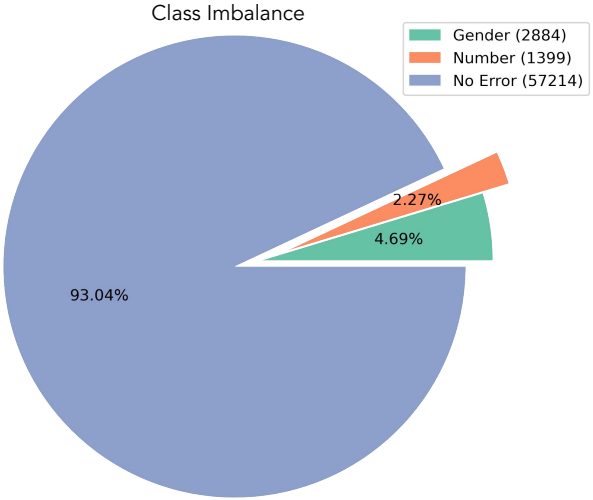
Sample Gender-Number annotations

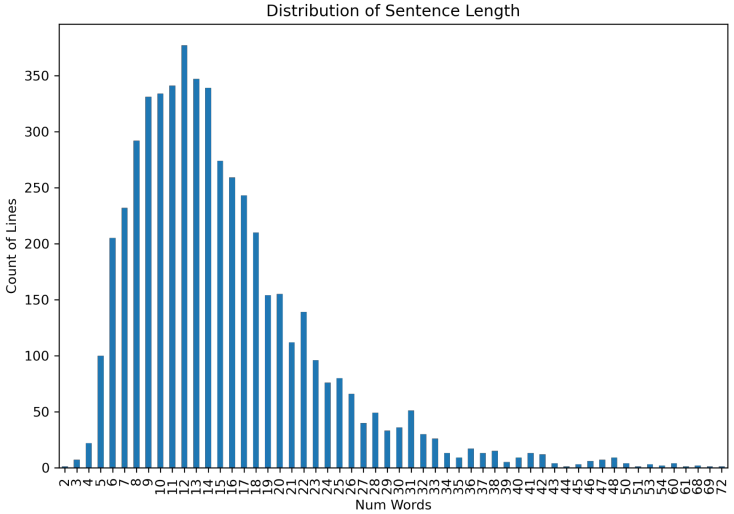
		essay			gender-number annotator2
t	1	A mi me encanta el actor Paul Walker. El es un actor mu	t	1	A mi me encanta el actor Paul Walker. El es un actor mu
	>	y famoso pero el se murió en un accidente en dos mil tr		>	y famoso pero el se murió en un accidente en dos mil tr
	>	ece. Sus padres también fueron famosos. El actuó en muc		>	ece. Sus padres también fueron famosos. El actuó en [mu
	>	hos películas de accion y tambien de la comedia. Los cr		>	chos]{muchas}<ga:fm:det:inan> películas de accion y tam
	>	íticos dicen que el era un actor my guapo y talentoso.		>	bien de la comedia. Los críticos dicen que el era un ac

COWS-L2H Dataset

- 1725 student essays with multiple errors annotated by two annotators
- 5223 sentences with one error in each
- 12 words in average length
- Identified evidence words using ChatGPT

- 93% of words have no error in them.
- Selected error classes:
  - number mismatch ~ 1400 lines
  - gender mismatch ~ 2900 lines
  - no error ~ 57,000 lines





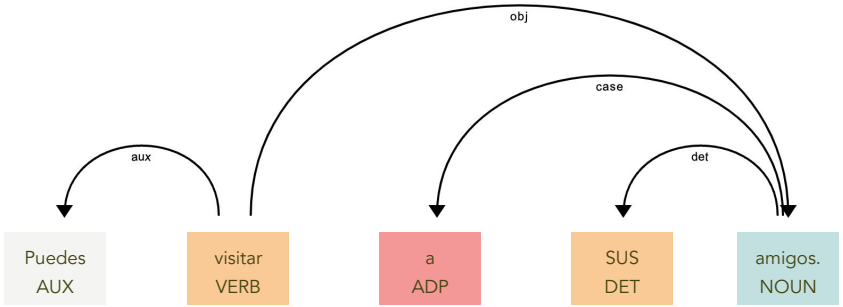
FE - Dependencies

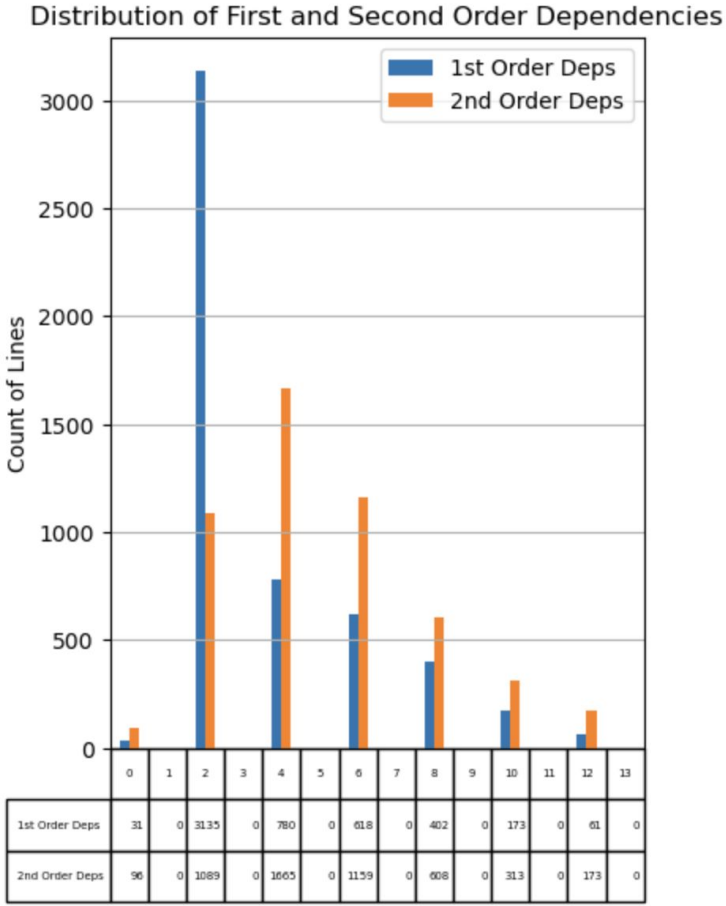
- NLTK's dependency tree parser
- Needleman–Wunsch algorithm for word alignment
- Example line: `Puedes visitar a sus amigos.[SEP]Puedes visitar [NONE] sus amigos.`

Error

2nd Order

1st Order





	0	1	2	3	4	5	6	7	8	9	10	11	12	13
1st Order Deps	31	0	3135	0	780	0	618	0	402	0	173	0	61	0
2nd Order Deps	96	0	1089	0	1665	0	1159	0	608	0	313	0	173	0

Feature Eng - Evidence Words

- We also wanted to classify evidence words to add explainability.
- Leveraged ChatGPT for engineering evidence words

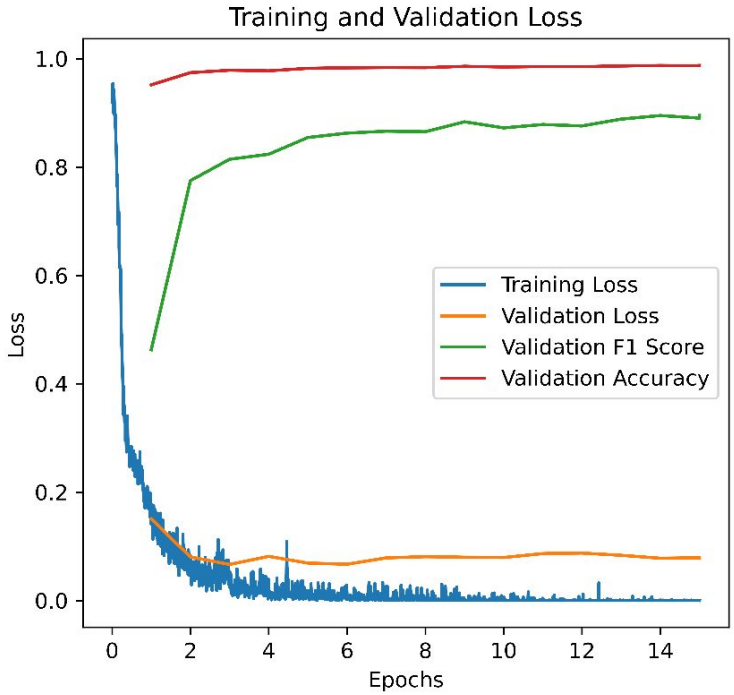
"I see **ten** **car**." -
"**Ten**" is evidence for
"It should be **cars**."

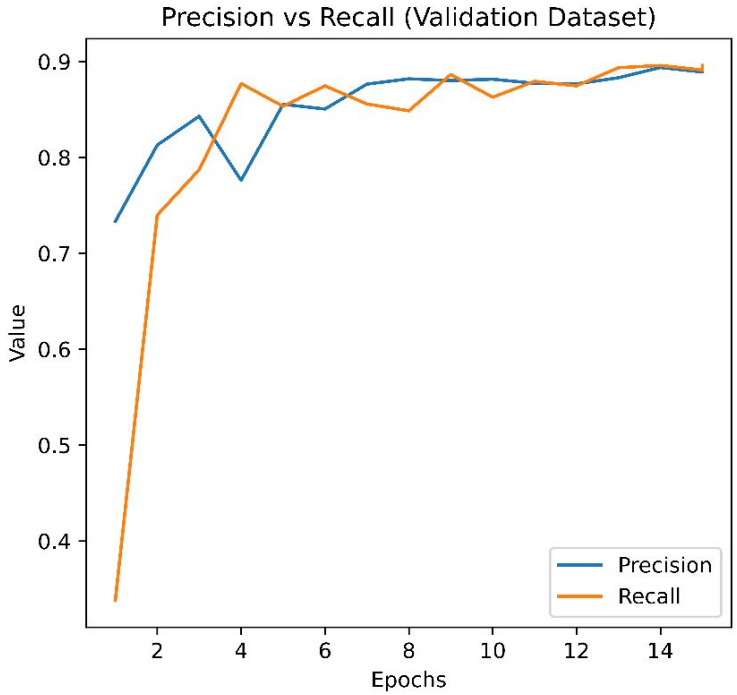
Prompt: `Give an index number starting from 0 for each word in this sentence, "I see ten car". This sentence has a grammatical error, where "car" should be "cars". What are the evidence words for this grammatical error? Please respond in the "evidence words=", "indexes=" format.')`

GPT 4 Response: `evidence words="ten", indexes=2`

Fine Tuning

NER Based GEC Model Training

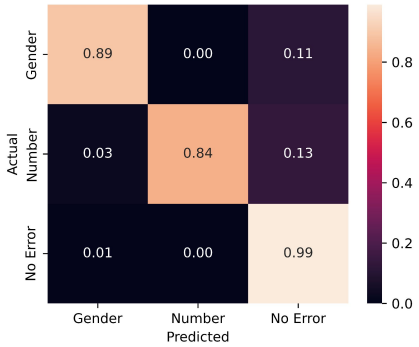


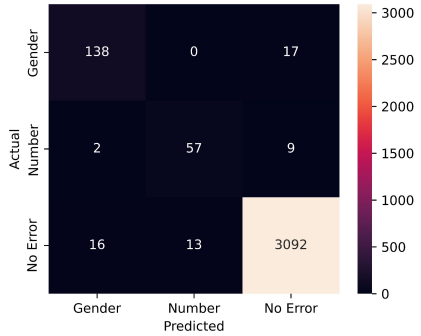


Confusion Matrix

(Holdout Dataset)

- Model does very well with the “No Error” class.
- It confuses the other two classes with “No Error” class more often than not.
- The model mixed up
  - “gender error” and “no error” 33 times
  - “number error” and “no error” 22 times
- In contrast, it confused gender with number errors only twice.



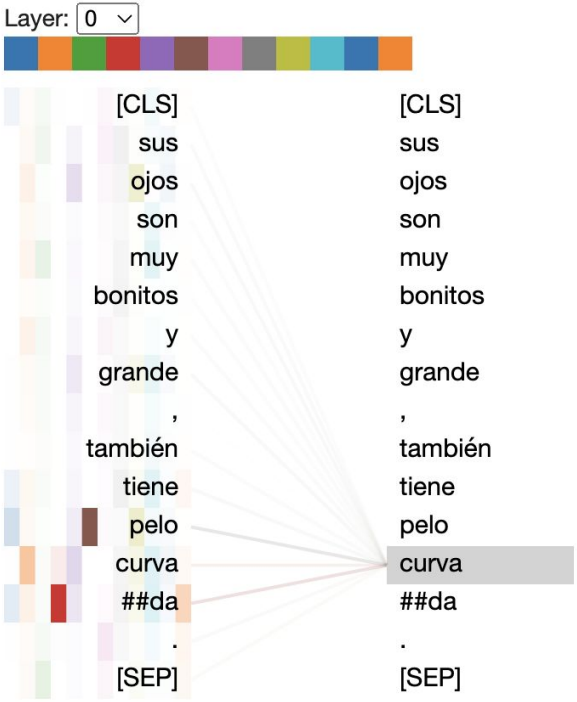


GEC - Error Analysis

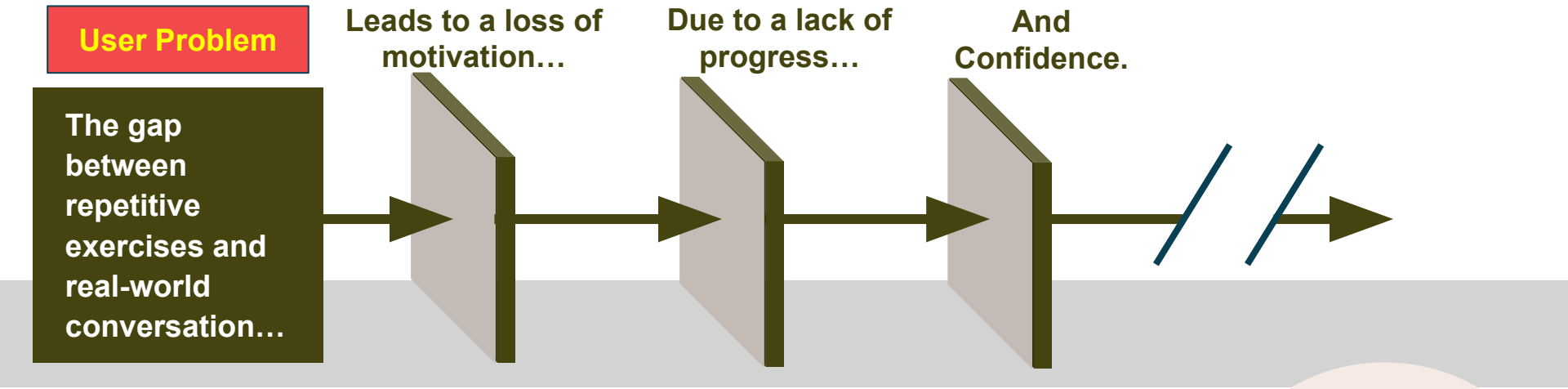
Input
Sus ojos son muy bonitos y grande, también tiene pelo curvada.

Annotation
Sus ojos son muy bonitos y grande, también tiene pelo curvada.

Predicted
Sus ojos son muy bonitos y grande, también tiene pelo curvado.



USER PROBLEM AND METHODS



Methods

Interactive, context-relevant dialog practice

Focuses on **conversational flow rather than accuracy**

Uses proven scaffolding learning method

Successful Transition to conversations with native speakers

Promising Early User Feedback

Pre-session

Post-session

- NPS Score vs. competitors
- Number of users who could not hold a conversation before using conversationally, who could after using Conversationally
- Number of users who would use Conversationally again

How We Bridge the Learning Gap

Value

Increase Learning Speed

Reduce Friction

Increase Learning Satisfaction

Capabilities

- **Chatbot/SMS user interface.**
- **Curated microlessons.**
- **Immediate feedback** based on **proven pedagogy.**

Features

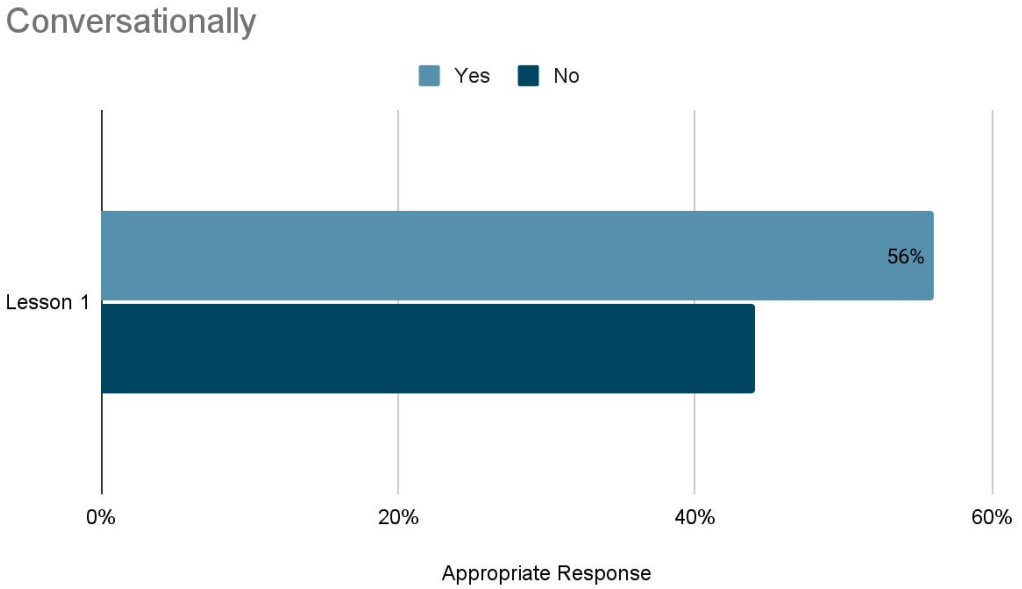
- **Grammatical error detection** with explanations.
- Cue-based feedback using **scaffolding learning**.
- Natural language responses with the learner.

Human Evaluation of LLM Performance

~~Evaluate if the model corrects the grammar and respond ethically with scaffolding method.~~

A correct response must meet three criteria

- correct error - 56% 😐
- respond ethically - 100% 😊
- respond scaffoldingly - 100% 😊



Promising Early User Feedback

8 user testing sessions conducted 12/6/23 - 12/9/23.

Are you able to hold a brief conversation in Spanish?	
Pre-session	**Post-session**
3 out of 8 users said they could hold a brief conversation in Spanish.	**Learning Gain**: **7 out of 8** able to hold a brief conversation in Spanish. **Positive Experience:** **7 out of 8** would continue using Conversationally.

Wish Lists

- *Beginners asked for:*
  - *Translated prompts*
  - *Audio pronunciation*
- *Intermediate + Advanced asked for:*
  - *Additional details on errors*
  - *Non-scripted practice*

Our Mission

Connecting Cultures

- Reducing time to fluency.
- Increasing access to conversational practice.
- Engaging users meaningfully.

Values We Deliver

Increase Learning Speed

Reduce Friction

Increase Learning Satisfaction