## Kinder, the empathetic writing assistant

"Kindness is the language which the deaf can hear and the blind can see." – Mark Twain

MIDS Capstone Summer 2025 Andy Guinto, Diego Moss, Priyanka Upadhyay

**UC Berkeley** School of Information





Image generated with Perplexity

## The problem and our impact



Our vision is to improve people's well-being by making everyday communication over online and mobile apps more empathetic.

## Studies Show a 40% Decline in Empathy **Among College Students**

Is Technology Angry by design: toxic communication and technical Communication in the

architectures

Greater use of virtual technology leads to toxic communications

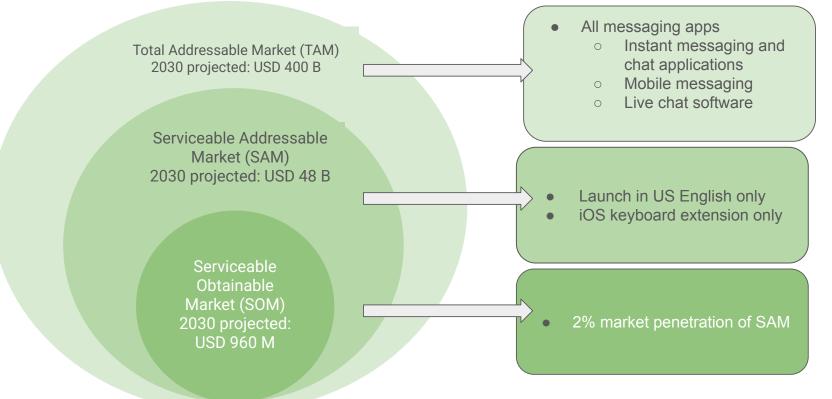
Can Contribute to **Dehumanizing Other People** 

More Social Media Use, Lower Empathy

How technology is harming our ability to feel empathy

## The opportunity





https://cropink.com/gen-z-social-media-usage-statistics;https://www.verifiedmarketresearch.com/product/instant-messaging-chat-software-market/; https://www.cognitivemarketresearch.com/mobile-messaging-app-market-report

## Kinder's unique market positioning



Value Proposition	Kinder	Grammarly	Inbuilt app assistants	Chatgpt/Cla ude/Grok
Grammar and writing style	×			
Focus on empathetic language		×	×	
In app use with instant rephrasing				×
Personal preference settings (such as empathy meter)		×	×	×
Requires user authentication				×
Collects personal, device, or usage data	×		<b>/</b>	<b>/</b>

## Our test users love Kinder already!



### What users liked

- They deeply relate to the problem
- Can use Kinder right within the messaging app of their choice

"It's like a filter for my emotions — in the best way", UX Designer, San Jose

"Kinder makes me feel confident that I'm coming across the way I want to", Student, Los Angeles

"Kinder changed how I talk to my team. helped me soften my feedback without losing clarity", Head of Data & AI, San Francisco

## User privacy and ethical concerns is our priority



## What are some of users' questions

- No Personal, device or usage data stores
- On device processing of data only
- T&C for iOS keyboard extension
- MVP US English only

Nuances of empathy by culture and demographics - human data annotation and survey for

train our models

"Different cultures or may have nuances in what is considered empathetic"

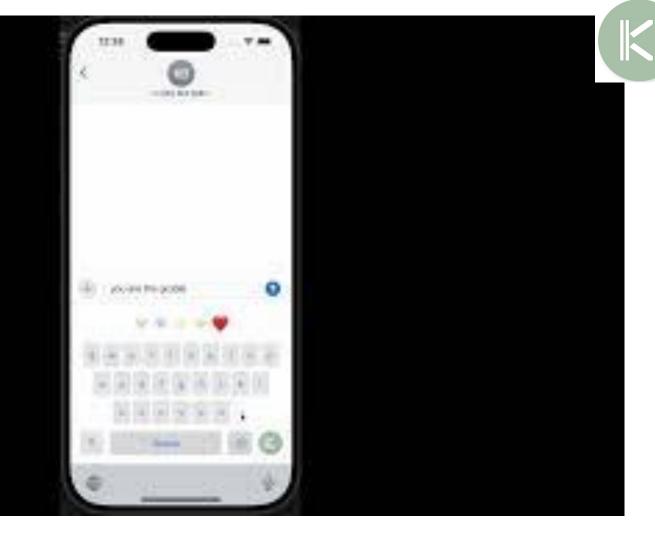
"I might speak to my Mom differently from my friend and still consider both empathetic communication" "Will my messages or other personal data be stored?"

"My idea of empathetic communication may vary in personal vs professional setting"

"English language usage may vary in countries other than USA"



## Demo



## **Empathy Defined**



## Empathy has three components according to psychology experts:

We want to improve empathy in online communications by non-intrusively...

Perspective Taking<sup>1</sup>



Notifying the user when they say something that is unempathetic

2. Emotional Mimicking<sup>2</sup>



Showing the user how much a message might cause a negative emotional reaction

3. Connection between the Self and Other<sup>3</sup>



Offer an empathetic rephrased message to prevent disconnecting with someone

Aragona, M. (2016)

Cortina, M. (2021).

Vilardaga, R. (2009).



## Data Used and EDA

## What Data are we using?



## (Un)empathetic Classification

## Subreddits of insults and compliments



Roastme Toastme FreeCompliments

Human annotated online messages from data annotation teams from both Surge-ai and Google research





"['NAME]' Why are you being so sensitive"

"If awkward was a person."

"I'm sorry that happened. Some people can be jerks sometimes. That's not your fault."

"First off, I love your hair and choice of glasses"

## What Data are we using?



## (Un)empathetic Classification

## Message Rephrasing

## **Instruction Tuning**

with help from LLama 3.1 405b Instruct for synthetic data in empathetic rephrasing.

And empathetic dialogues on facebook provided by *AI at Meta* 





## **Data Annotation for Empathy Classifier**



-2

-1

0

2

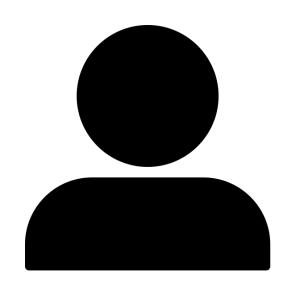
## **Empathetic and unempathetic messages are roughly equal in:**

- message length,
- word variety,
- emojis
- and topics covered!

## **Extra Data from Human Respondents**



Human Evaluations as Ground Truth



Why are you being so sensitive about this!?

Imagine you received the above message from a friend. Fill in the sentence "I would feel..."

-2

-1

0

.

2

Very Bad Bad

Neutral

Good

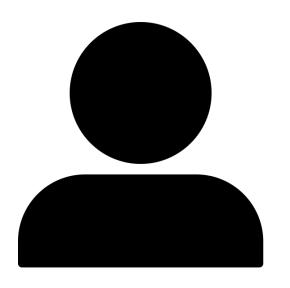
Very

Good

## **Extra Data from Human Respondents**

## K

## Human Evaluations as Ground Truth



Why are you being so sensitive about this!?

Imagine you received the above message from a friend. Which of the below ways of rephrasing the message is more empathetic



Can you help me understand why this bothers you?



What about this is making you feel this way?

## **Data Pipelines**



Synthetic data generation for

using LLama

facebook

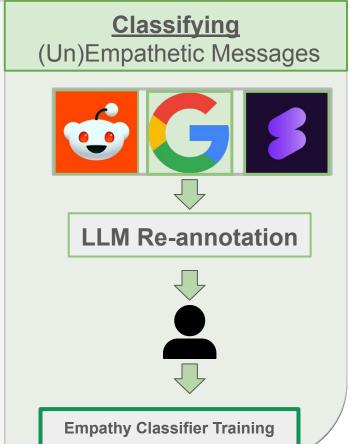
exchanges

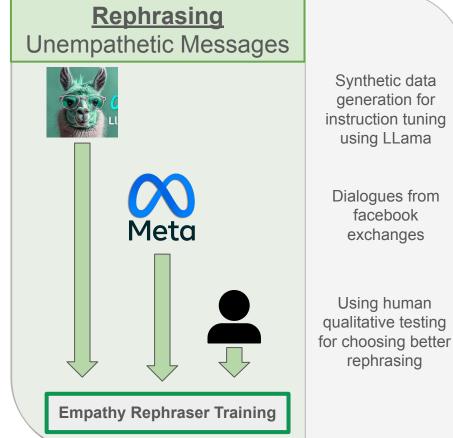
Using human

rephrasing

Online sources and annotaated data from google and surge-ai

Further data collection on human labels to messages



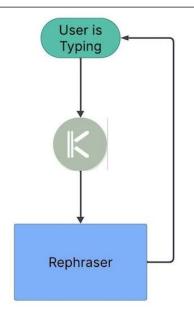


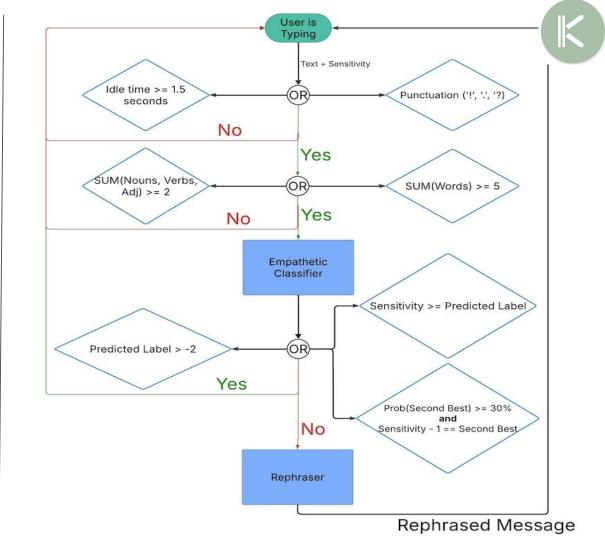


## Modeling and Architecture

## **Logical Flow**

- Classifier Trigger?
- Rephraser Trigger(s)?
- Rephraser Evaluation?
- Constraints
  - Capped at 128 tokens





## **Model Training**

K

- Classifier
  - Fine tuned pardonmyai-tiny
     (https://huggingface.co/tarekziade/pardonmyai-tiny)
  - Avoid Catastrophic Forgetting Problem
    - Learning Rate: 0.00002
    - Warmup Ratio: 0.1
  - Loss: Tolerance Based KL-Divergence

## Rephraser

- Instruction tuned
   (https://huggingface.co/distilber
   t/distilgpt2)
- Shrink Model
  - Removed unused chars
  - Resized Embeddings Layer
- Lean towards deterministic results
  - Temperature is set to 0.3
  - Max 2 ngrams
- Loss: Cross Entropy
- New special token in Tokenizer: ['NAME']

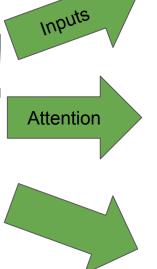
## **Rephraser - Instruction Tuning**

## Inputs





### Response:
I understand you're
facing a challenge, but
I'm not in a position
to take on that
responsibility.



```
tensor([21017, 46486,
                        25,
                              198,
                                    6207, 11840,
                                                    589,
                                                           428,
                                                                  284,
                                                                       2128,
                                                   2504,
         517,
                795,
                       8071,
                             6587,
                                      25,
                                            198,
                                                          338
                                                                  407.
                                                                        616,
        1917,
                       198,
                              198, 21017, 18261,
                                                    25,
                                                                        1833,
                821,
                                                          314,
                      6476,
                              257,
                                                   475,
                              284,
                                            319
                                                   326,
                                                                        220,
                      2292,
                                    1011,
                                                         5798
       50256, 50257, 50257, 50257, 50257, 50257, 50257, 50257, 50257, 50257,
              50257, 50257, 50257, 50257, 50257, 50257,
       50257, 50257, 50257, 50257, 50257, 50257, 50257, 50257,
       50257, 50257, 50257, 50257, 50257, 50257, 50257, 50257, 50257, 50257,
       50257, 50257, 50257, 50257, 50257, 50257, 50257, 50257, 50257,
       50257, 50257, 50257, 50257, 50257, 50257, 50257, 50257, 50257,
       50257, 50257, 50257, 50257, 50257, 50257, 50257, 50257, 50257,
       50257, 50257, 50257, 50257, 50257, 50257, 50257, 50257])
```

```
tensor([ -100.
               -100,
                      -100,
                                                        -100,
                                   -100.
                             -100.
                                          -100.
               -100,
         -100,
                      -100.
                                  21017, 18261,
                                                         198,
                                                                      1833,
         345.
                                                               1101.
         287,
                              284,
                                   1011,
                                                        5798,
       50256, 50257, 50257, 50257, 50257, 50257, 50257, 50257, 50257,
       50257, 50257, 50257, 50257, 50257, 50257, 50257, 50257, 50257,
       50257, 50257, 50257, 50257, 50257, 50257, 50257, 50257, 50257,
       50257, 50257, 50257, 50257, 50257, 50257, 50257, 50257, 50257, 50257,
       50257, 50257, 50257, 50257, 50257, 50257, 50257, 50257, 50257,
       50257, 50257, 50257, 50257, 50257, 50257, 50257, 50257, 50257,
       50257, 50257, 50257, 50257, 50257, 50257, 50257, 50257, 50257,
       50257, 50257, 50257, 50257, 50257, 50257, 50257, 50257])
```

## **Tolerant Boosted Kullback-Leibler Divergence Function**

	Totol and Doodlog Hamback Eon		
Label	Soft Labels (KL Divergence with Center Weight = 0.8, Tolerance = 1)		
-2	[ <b>0.8</b> , <b>0.2</b> , 0, 0, 0]		
-1	[ <b>0.1</b> , <b>0.8</b> , <b>0.1</b> , 0, 0]		
0	[0, <b>0.1</b> , <b>0.8</b> , <b>0.1</b> , 0]		
1	[0, 0, <b>0.1</b> , <b>0.8</b> , <b>0.1</b> ]		
2	[0, 0, 0, <b>0.8</b> , <b>0.2</b> ]		

- Per\_Label\_Soft\_Label\_Loss[i] = (log(Predicted\_Prob[i])) \* Soft\_Label[i]
- Soft\_Labels\_KL\_Loss = SUM(Per\_Label\_Soft\_Label\_Loss)
- Per Sample Loss = Soft Labels KL Loss (1) \* Annotation Weights (2) \* Label Weights (3)
- Normalized\_Batch\_Loss = SUM(All Per\_Sample\_Loss) / Batch Size

### Basic Example:

True	Pred	Probabilities	Loss
0	1	[0.05, <b>0.1</b> , <b>0.35</b> , <b>0.4</b> , 0.1]	~3.56

- (1)  $-(\frac{\log(0.05)*0}{0.000})*0.1 + \log(0.35)*0.8 + \log(0.4)*0.1 + \frac{\log(0.1)*0}{0.000})$ (2)  $0.8 + \frac{0*(0-0.8)}{0.000}$
- (3) 1.7428 \* 2.2
- (4) 1.162 <sub>(1)</sub> \* 0.8 <sub>(2)</sub> \* 3.833 <sub>(3)</sub>

Boost Survey labeled Data Decrease Al

Decrease Al re-annotated data Leave original data alone

Confidence >= 50% only (LaaJ)

Text	Survey	Conf	Survey Boost
"They are bad for your body"	0	0.8	0
"Just go for a walk, you'll feel better"	1	1.0	1.0

Annotation\_Weight = Conf + Survey \* (Survey Boost - Conf)

Label	Inverse Frequency (IF)	Boost (SB)
-2	0.4255	1.0
-1	0.9749	1.5
0	1.7428	2.2
1	1.0585	1.5
2	0.7984	1.0

Label\_Weight = Inverse Frequency \* Boost

## Tolerant Boosted Kullback-Leibler Divergence Function

Let:

- $t_{\ell} \in \{0,1,2,3,4\}$ : the true class index for example  $\ell$ , mapped from semantic labels  $\{-2, -1, 0, 1, 2\}$
- $\mathbf{T}_{\ell} = [T_{\ell,0}, T_{\ell,1}, \dots, T_{\ell,4}]$ : the soft target distribution over class indices for example  $\ell$

### Soft Target Distribution (Center-Weighted)

Let  $\alpha \in (0,1)$  denote the center weight, e.g.,  $\alpha = 0.8$ . Then:

$$\begin{split} T_{\ell,i} &= \begin{cases} \alpha, & \text{if } i = t_\ell \\ \frac{1-\alpha}{Z_\ell}, & \text{if } |i-t_\ell| = 1 \text{ and } i \in \{0,1,2,3,4\} \\ 0. & \text{otherwise} \end{cases} \\ \mathrm{KL}_\ell &= \sum_{i=0}^4 T_{\ell,i} \cdot \log\left(\frac{T_{\ell,i}}{P_{\ell,i}}\right) \end{split}$$

$$KL_{\ell} = \sum_{i=0}^{4} T_{\ell,i} \cdot \log \left( \frac{T_{\ell,i}}{P_{\ell,i}} \right)$$

$$\mathcal{L}_{\text{batch}} = \frac{1}{N} \sum_{\ell=1}^{N} \text{KL}_{\ell} \cdot \text{BIF}_{\ell} \cdot w_{\ell}^{\text{final}}$$

Let:

- $w_{\ell}^{\text{conf}} \in [0.5, 1.0]$ : confidence-based weight for example  $\ell$ , filtered to include only predictions with at least 50% confidence
- $\delta_{\ell} \in \{0,1\}$ : 1 if survey, 0 if not
- w<sub>ℓ</sub> survey ∈ ℝ<sup>+</sup>: fixed weight for survey-based labels (w<sub>ℓ</sub> survey = 1.0)

### Per-sample Weight:

$$w_{\ell}^{\text{final}} = w_{\ell}^{\text{conf}} + \delta_{\ell} \cdot \left( w_{\ell}^{\text{survey}} - w_{\ell}^{\text{conf}} \right)$$

Let:

- **b** = [1.0, 1.5, 2.2, 1.5, 1.0]: class-specific boost factors
- $\mathbf{f} = [f_0, f_1, \dots, f_{C-1}]$ : inverse frequency weights per class
- $y_{\ell} \in \{0, \ldots, C-1\}$ : the true class index for example  $\ell$

## Boosted Inverse Frequency (BIF):

$$BIF_{\ell} = \mathbf{b}_{y_{\ell}} \cdot \mathbf{f}_{y_{\ell}}$$

## **Tolerant Boosted Kullback-Leibler Divergence Function**



## KL Divergence

## **Annotation Weights**

BIF

Center Weight: 80%

Tolerance: 1



Developer
Annotated

Al Re-annotated
(LaaJ)



Label 0

Labels {-1, 1}

Labels {-2, 2}

(Penalize predicted neighbors less)

(Weights by Data type)

(Weights by Sample Size)



LOSS

## K

## (+/-) 1 Tolerant Evaluation

- Baseline
  - Loss: Categorical Cross Entropy
  - Weighting: Inverse Frequency
- Final Model
  - Loss: Tolerance-Boosted Kullback-Leibler
  - Weighting: Boosted Inverse Frequency

Label	±1 Precision (B   F)	±1 Recall (B   F)	±1 F1-Score (B   F)
-2	0.91   <b>0.93</b>	0.64   <b>0.76</b>	0.75   <b>0.84</b>
-1	0.88   <b>0.91</b>	0.57   <b>0.89</b>	0.69   <b>0.90</b>
0	0.60   <b>0.61</b>	0.88   <b>0.93</b>	0.72   <b>0.74</b>
1	0.68   <b>0.90</b>	0.90   <b>0.92</b>	0.77   <b>0.91</b>
2	0.76   <b>0.85</b>	<b>0.86</b>   0.76	0.80   <b>0.81</b>

Text	True Label
"They are bad for your body"	-1
"Just go for a walk, you'll feel better"	-2







Let each class label  $c \in \{0, 1, 2, 3, 4\}$  corresponds to semantic labels  $\{-2, -1, 0, 1, 2\}$ , such that:

- TP<sub>c</sub>: predicted label is within  $\pm 1$  of the true label c
- FP<sub>c</sub>: predicted label is not within ±1 of c, but was predicted as c
- FN<sub>c</sub>: true label is c, but prediction was not within  $\pm 1$

### Per-class metrics:

$$\begin{aligned} \text{Precision}_c &= \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c} \quad , \quad \text{Recall}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c} \\ & \text{F1}_c = \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \end{aligned}$$

## iOS Architecture

- CoreML
- SwiftUI
  - Generic Keyboard

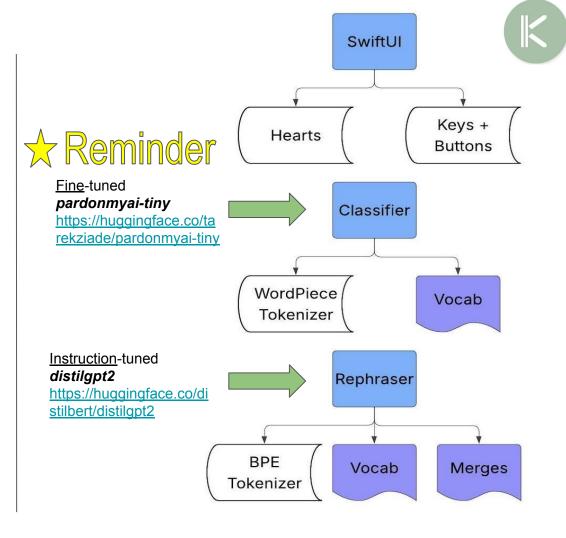


Sensitivity Settings



Additional Rephrase Button





## Model Quantization (Oversimplified)



## 1.) Filter Task via HuggingFace Exporters

```
$ python -m exporters.coreml -model="<model>" -feature <task> <Output Model>
```

## 2.) Reduce output

```
class LogitsOnlyWrapper(torch.nn.Module):
    def __init__(self, base_model):
        super().__init__()
        self.base_model = base_model

def forward(self, input_ids, attention_mask):
        logits = self.base_model(input_ids=input_ids, attention_mask=attention_mask).logits
        return logits[:, -1, :]
```

## 3.) Convert to Core ML

```
coreml_model = ct.convert(
    traced_model,
    source="pytorch",
    inputs=[
        ct.TensorType(name="input_ids", shape=(1, ct.RangeDim(1, 128)), dtype=np.int32),
        ct.TensorType(name="attention_mask", shape=(1, ct.RangeDim(1, 128)), dtype=np.int32)
        ],
        outputs=["logits"],
        convert_to="neuralnetwork",
        minimum_deployment_target=ct.target.iOS14,
        compute_units=ct.ComputeUnit.ALL
}
```

## 4.) Quantize even further

```
quantized_model = ct.models.neural_network.quantization_utils.quantize_weights(
    coreml_model,
    nbits=8,
    quantization_mode="linear"
)
```

## 5.) Compile

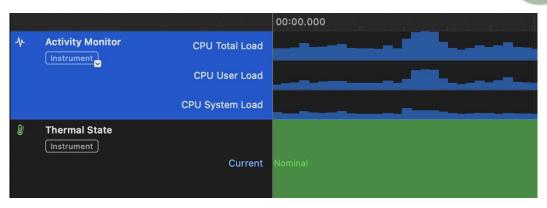
\$ xcrun coremic compile < Model>

Model	Best Size (mb)
Classifier	28
Rephraser	~100-150

## **Instrumentation and Profiling**

K

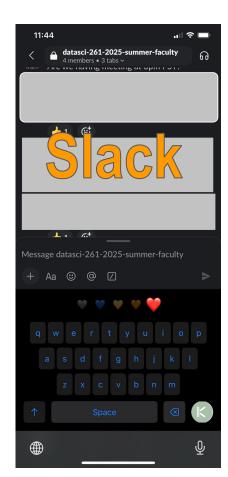
- Device Tested
  - o iPhone 16 Pro Max
  - o iOS 16+
- Hardware Requirements
  - CPU Only Any device that supports iOS 14
  - ANE/GPU iPhone 11+
- Evaluation Time
- Thermal looks good!

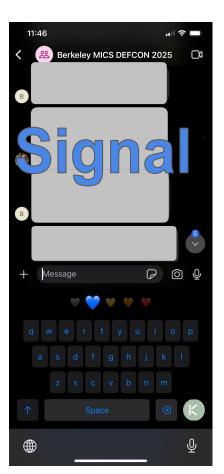


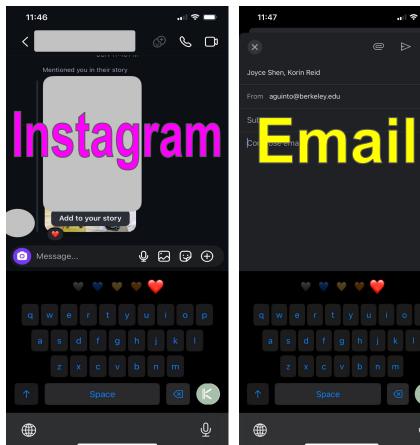
Hardware	Classifier Inference(s)	Rephraser Inference (s)
CPU	0.20546603202819824	5.127754092216492
ANE/GPU	0.16854095458984375	4.187222957611084
ANE/GPU with Padding	0.19244205951690674	7.1593159437179565

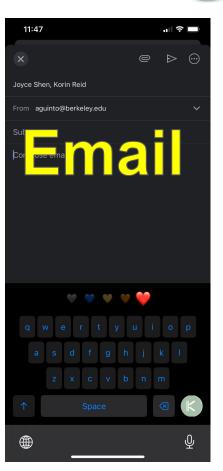
## Device Installation - Cross Application, Cross Device, etc.













## What's Next?

### **Algorithm Development**

- More surveys and re-annotations for training
- Find optimal parameters for the Tolerant-Boosted KL Divergence Function
  - Boost values, Center Weight, and Survey weights have room for experimentation
- Update evaluation metric to include (+/-) 1 IFF probability >= Threshold

## **Application Specific**

- Address demographic and cultural bias as part of loss and evaluation
- Address UI Edge cases
- Integrate onto Android

### **Academic Endeavors**

- Fully Homomorphic Encryption (FHE) and Reinforcement learning from Human Feedback (RHLF)
- Submit at an NLP (or Cryptography) conference

## **Team Kinder Pledge**



"We aim to empower users to communicate more empathetically by" offering real-time and human-backed rephrasing suggestions."

"We aim to make empathetic rephrasing more widely available without consumers purchasing additional hardware."

"We aim to preserve user **privacy** by ensuring that only the user can view their own data in plaintext."



## Q&A



## Appendix

## T&C Page 1

Kinder Terms and Last updated: 6/30/2025

Please read these Terms and Conditions ("Terms") carefully before downloading or using Kinder, the iOS keyboard-extension that offers real-time, Al-powered empathy suggestions ("App," "Kinder," "we," "our," or "us"). By installing or using Kinder you acknowledge that you have read, understood, and agree to be bound by these Terms and by our separate Privacy Policy, which is incorporated by reference.

- Acceptance of the Terms 1.
- If you do not agree with these Terms, do not install or use Kinder.
- We may update these Terms periodically. Material changes will be communicated through the App or by other reasonable means, and continued use constitutes acceptance of the revised Terms.
- 2. Service Description Kinder analyzes text you type inside the keyboard to (i) flag language that may be perceived as unkind, and (ii) suggest alternative, more empathetic phrasing. Processing is performed primarily on-device; however, if you opt-in, certain prompts may be sent to cloud models for enhanced accuracy. 3. Eligibility
- You must be at least 13 years old and legally able to form a binding contract. If you are under 18, you affirm that you have parental or guardian consent.
- 4. License Grant & Intellectual **Property** We grant you a personal, revocable, non-exclusive, non-transferable license to use Kinder solely for its intended purpose on iOS devices you own or control. All intellectual-property rights in Kinder and its content remain ours or our licensors'.
- 5. User Responsibilities & Restrictions
  - Do not reverse-engineer, decompile, or create derivative works of Kinder.
  - Do not use Kinder to generate or transmit unlawful, harassing, defamatory, or hateful content.
- You remain solely responsible for the messages you send, even when you accept Kinder's suggestions.
- 6. Privacy & California-Specific Disclosures

Categories Personal Information Collected οf

## T&C Page 2

6.2

At or before the point of collection we inform you of the categories of any personal information and the purposes for which they are used. This notice is provided within these Terms and the in-app onboarding screen. 6.3 California Act (CCPA/CPRA) Rights Consumer Privacy If you reside in California, you have the right to: Know the categories and specific pieces of personal information we collect, use, or disclose;

(Cal.

Civ.

Code

Request deletion of personal information (with statutory exceptions); Correct inaccurate personal information;

Notice

- Opt-out of the "sale" or "sharing" of personal information (Kinder does not sell or share your data for cross-context behavioral advertising);
- Limit the use of sensitive personal information; and
- Be free from discrimination for exercising CCPA rights. You may exercise these rights in-app or by emailing <u>privacy@kinder.ai</u>.
- 6.4 CalOPPA Online Privacy Protection

at

- Kinder maintains a conspicuously posted, accessible Privacy Policy that discloses our data practices, the effective date, and how we notify users of material changes.
- "Do-Not-Track" Signals: Browsers may transmit DNT requests. We currently do not track users for advertising and therefore do not respond differently
- to DNT signals; this is disclosed as required by CalOPPA.
- We retain on-device data only while necessary for core functionality, and cloud-processed data is deleted after completion of inference. We apply least-retention

Collection

- Retention Deletion 6.5 Data
- and surprise-minimization principles recommended by the California AG for mobile apps. 7. Al Output & Disclaimers Suggestions are generated statistically and may be inaccurate or inappropriate in certain contexts; you should review and modify any text before
- - sending.
- Kinder does not provide legal, medical, or psychological advice.

8. Third-Party Services The App may interface with Apple systems and optional cloud-Al providers. Your use of third-party services is subject to their separate terms and

## **Acknowledgements and additional resources**



The Kinder team extends gratitude to:

- UC Berkeley MIDS Program for providing the capstone framework and academic support
- Data Annotation Partners: Surge Al and Google Research teams for high-quality human annotations
- Open Source Community:
  - Meta AI for the Empathetic Dialogues dataset
  - Hugging Face for model hosting and tools
  - Reddit communities for authentic conversational data
- Technical Infrastructure:
  - Apple CoreML team for mobile AI optimization tools
  - LLaMA and InstructLab for synthetic data generation capabilities
- User Testing Participants who provided invaluable feedback during development

### Github with the code

https://github.com/NLPwned/Empathetic\_Datasets\_210\_Capstone/tree/main/

### **Project Website**

https://priyau.wixsite.com/writekindly

### **Human Survey Template Doc**

https://github.com/NLPwned/Empathetic\_Datasets\_210\_Capstone/blob/main/Empathic\_Communication\_-\_Responses.docx





## **AI Has Suddenly Evolved to Achieve Theory of Mind**

In a stunning development, a neural network now has the intuitive skills of a 9-year-old.

## Theory of Mind AI: Bringing Human Cognition to Machines

MAGAZINE

Neil Sahota Published: July 29, 2024 Artificial Intelligence



## EDA - Our data is looking good!



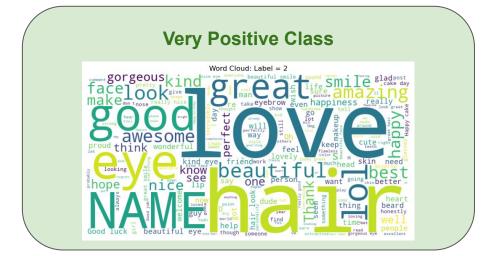
## Negative and positive messages are equivalent in:

Message Length (p < .001)Emoji Usage (p < .01)

Word Variety (p < .001)

Relevant Topics (p < .05)

# Word Cloud: Label = -2 already ur disgusting every to don, t re chicken fraid literally don, t re chicken f



## We are operating in a blue ocean



Grai	mmar,	polish,
and	writing	g style

Grammarly	
-----------	--

Inbuilt AI writing assistant in messaging apps (such as LinkedIn Messenger)

Option for empathetic writing

Kinder

ChatGPT Claude Grok

Integrated within messaging apps

Standalone Apps

## Our value proposition





All users of any ios messaging app. ios keyboard extension and can work right in messaging apps



Want to write kindly with confidence, without overthinking



As users type their message, will detect and flag words or phrases that don't read very kindly and get instant rephrasing suggestions



Empathy meter: personal preference