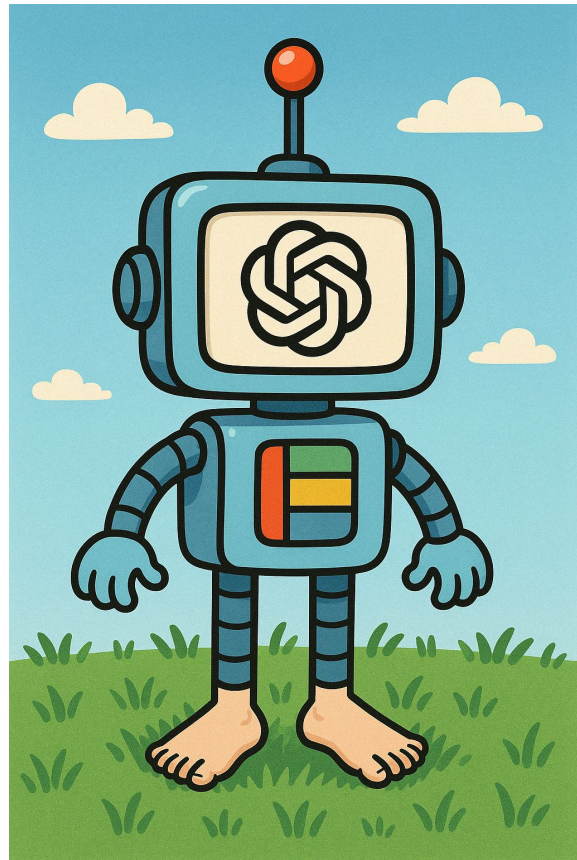
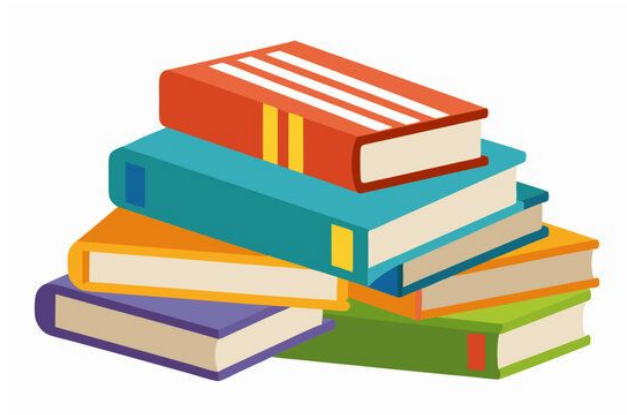


# Grounding LLM Explanations in Real-World Domains

Shreya Havaladar



# Two types of LLM tasks



**Domain-agnostic:** reasoning, summarization, NLI, question answering, translation, etc.



**Domain-specific:** Legal document analysis, disease diagnosis, therapy chatbots, etc.



Aspect	Domain-Agnostic Tasks	Domain-Specific Tasks
Expected users	Anyone	Domain experts
Consequence of errors	Low-stakes (e.g. a bad summary is annoying)	High-stakes (e.g. incorrect medical advice or legal interpretation)
User expectations	Accuracy and speed	Trust, transparency, reliability



Aspect	Domain-Agnostic Tasks	Domain-Specific Tasks
Expected users	Anyone	Domain experts
Consequence of errors	Low-stakes (e.g. a bad summary is annoying)	High-stakes (e.g. incorrect medical advice or legal interpretation)
User expectations	Accuracy and speed	Trust, transparency, reliability

***Good LLM explanations are needed for domain-specific tasks***

# What makes a good domain-specific explanation?

**1) Who is the explanation for?**



**DOMAIN  
EXPERT**

**2) What “language” should the explanation be in?**



**SOMETHING EXPERTS CAN  
EASILY UNDERSTAND**

**3) How will the explanation be used?**



**TO REVEAL DOMAIN-  
SPECIFIC LLM CAPABILITIES**

# What makes a good domain-specific explanation?

Explanations that are useful to experts must be ***grounded in real-world domains (i.e. incorporate domain knowledge).***



DOMAIN  
EXPERT

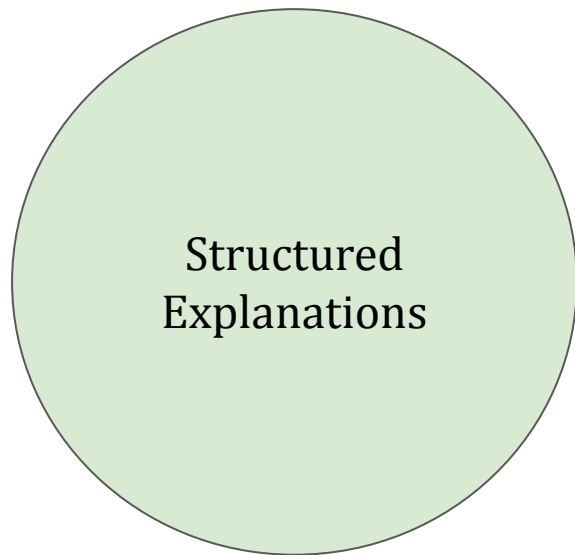


SOMETHING EXPERTS CAN  
EASILY UNDERSTAND

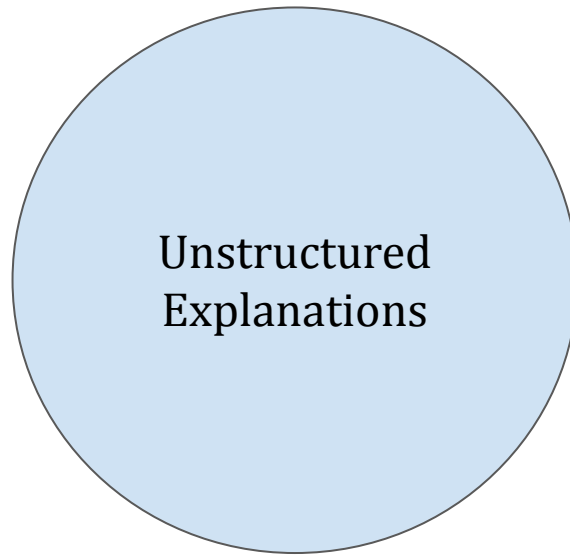


TO REVEAL DOMAIN-  
SPECIFIC LLM CAPABILITIES

# Explanations useful to experts



Explanation methods ***developed according to expert criteria***



General explanation methods that are ***post-hoc domain-grounded***

# Suppose I would like to explain why my dog ate my shoe...

<TIME LAST FED>  
11:39AM on  
4-15-25

<FAVORITE FOOD>  
Broccoli



There are many possible reasons why your dog might have eaten your shoe. He may have thought your shoe looked like his favorite food. It also could have been multiple hours since he ate. Or, your dog might be shedding.

Structured Explanation

Unstructured Explanation



# Suppose I would like to explain why my dog ate my shoe...

<TIME LAST FED>

11:39AM on  
4-15-25

<FAVORITE FOOD>

Broccoli



There are many possible reasons why your dog might have eaten your shoe. He may have thought your shoe looked like his favorite food. **It also could have been multiple hours since he ate.** Or, your dog might be shedding.

Structured Explanation

Unstructured Explanation

## STRUCTURED EXPLANATIONS

1. Evaluating the *multicultural emotional understanding* of LLMs: grounded in *psychology*

## UNSTRUCTURED EXPLANATIONS

2. Comparing *politeness across languages* using LLMs: grounded in *linguistics*
3. Determining whether *feature groups* used for explanations are *aligned with domain expert intuition*

# Building explanation methods grounded in emotion psychology

**Havaldar, S.**, Singhal, B., Rai, S., Liu, L., Guntuku, S. C., & Ungar, L. (2023, July). Multilingual Language Models are not Multicultural: A Case Study in Emotion. *In Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis (pp. 202-214).*

# LLMs require cultural sensitivity for emotion-based tasks



Teaching

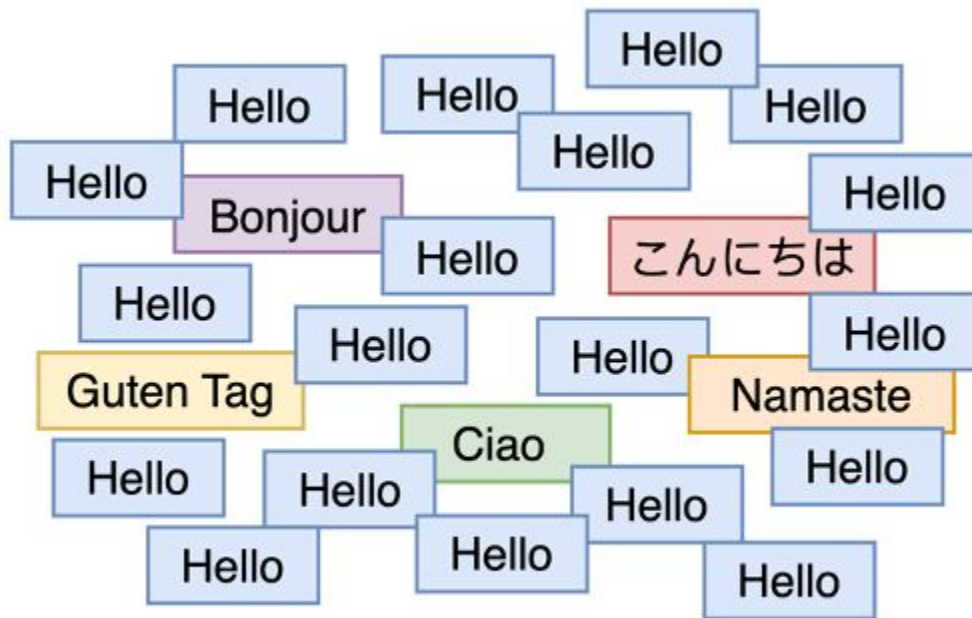


Therapy

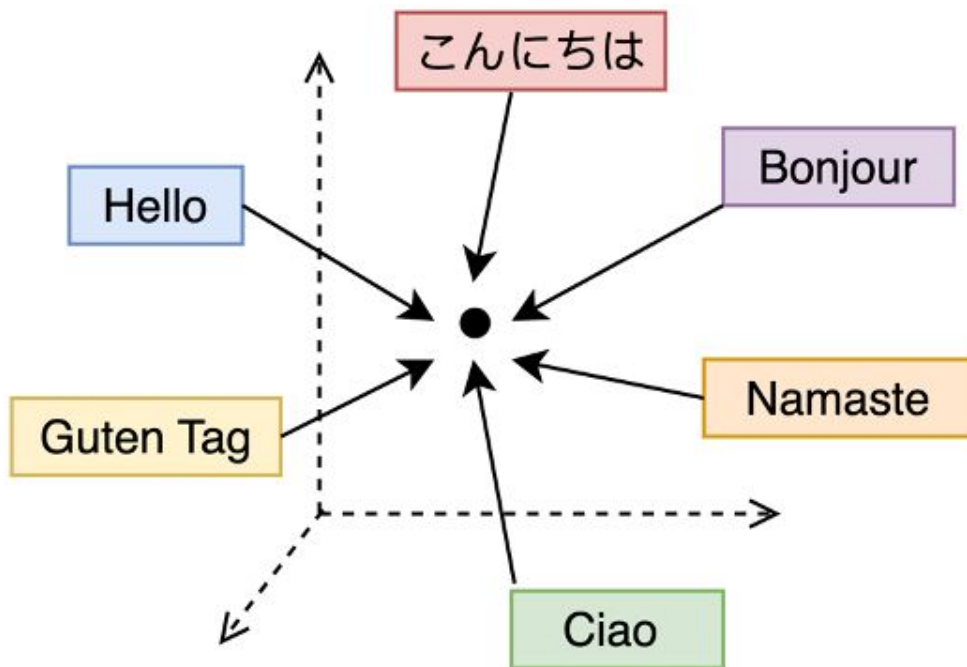


Crisis Prevention

# English-dominated datasets cause LLMs to be Anglocentric



# Training techniques cause LLMs to be Anglocentric



Emotions are heavily contextual.  
Contextual LLMs should understand this.

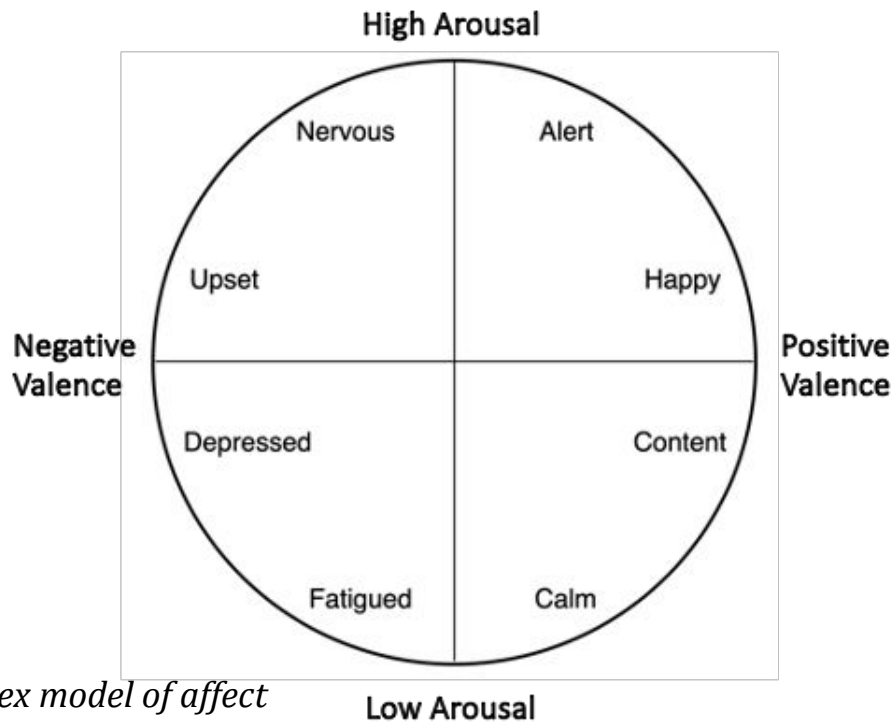


- Study with ~1,000 school children from America, Korea, and Japan (Furakawa et. al, 2012)
  - Test to measure “self-conscious” emotions showed substantial differences in mean levels of shame, guilt and pride
- **Japanese** children express the most **shame**
- **American** children express the most **pride**

# Explaining emotional understanding of LLMs via a *psychological model*

**Valence:** how negative or positive an emotion is

**Arousal:** The intensity and activation of an emotion



*Circumplex model of affect  
(Russell, 1980)*



# Extracting emotion representations from LLMs

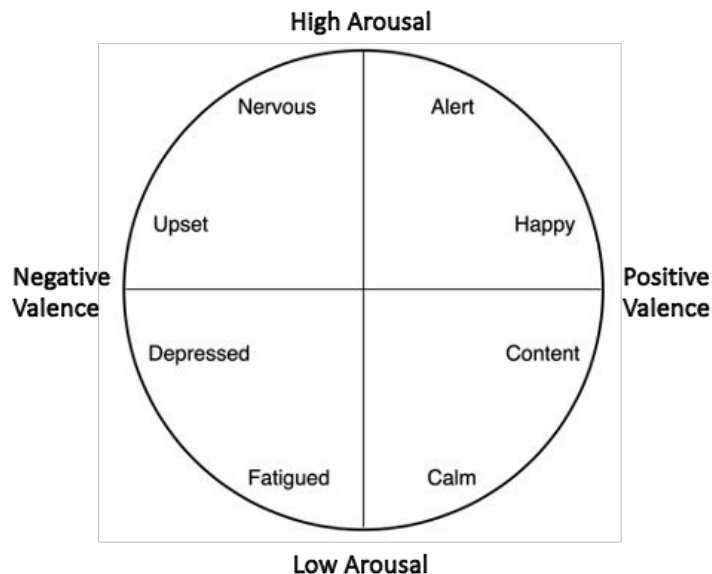
How do we extract an LLM's understanding of a given emotion in different languages?

1) Contextualize various expressions of an emotion: "I feel happy", "they are happy", etc.

2) Embed each emotional expression using the LLM

3) Average the embeddings of all expressions

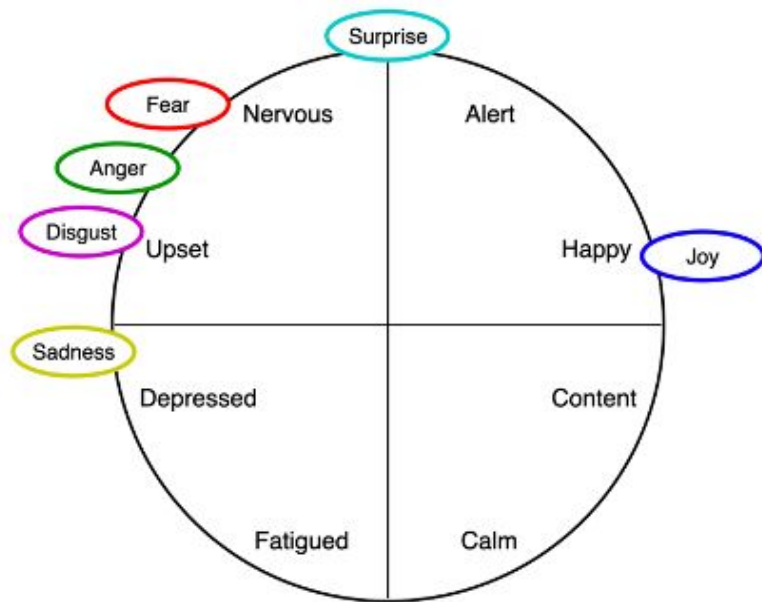
# Defining a circumplex in the latent space of an LLM



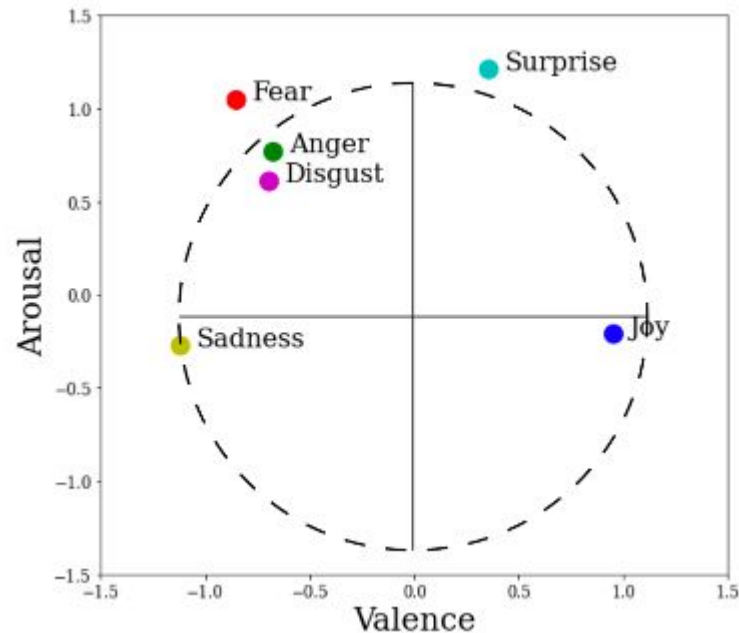
Circumplex model of affect  
(Russell, 1980)

1. Define valence ( $V$ ) and arousal ( $A$ ) axes using anchor emotions  $v_{pos}$ ,  $v_{neg}$ ,  $a_{high}$ ,  $a_{low}$
2. Transform axes so  $v_{pos}$ ,  $v_{neg}$ ,  $a_{high}$ ,  $a_{low}$  anchor to  $(1,0)$ ,  $(-1,0)$ ,  $(0,1)$ ,  $(0,-1)$  respectively.
3. Project an emotion representation by:
  - a. Calculating  $V$  and  $A$  components
  - b. Plot!

# Projecting emotion representations onto the circumplex



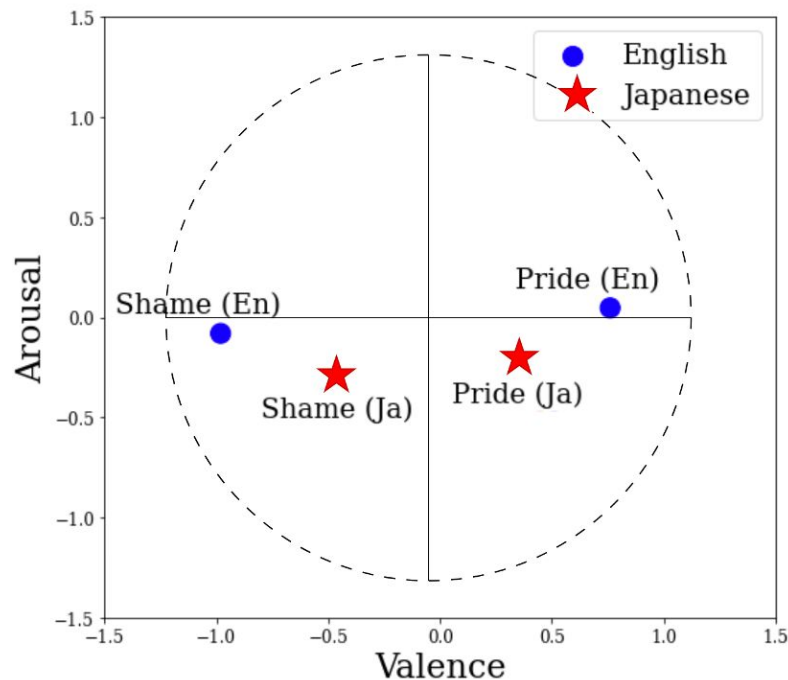
Circumplex model of affect  
(Russell, 1980)



Recreated circumplex  
(Havaldar et. al, 2023)

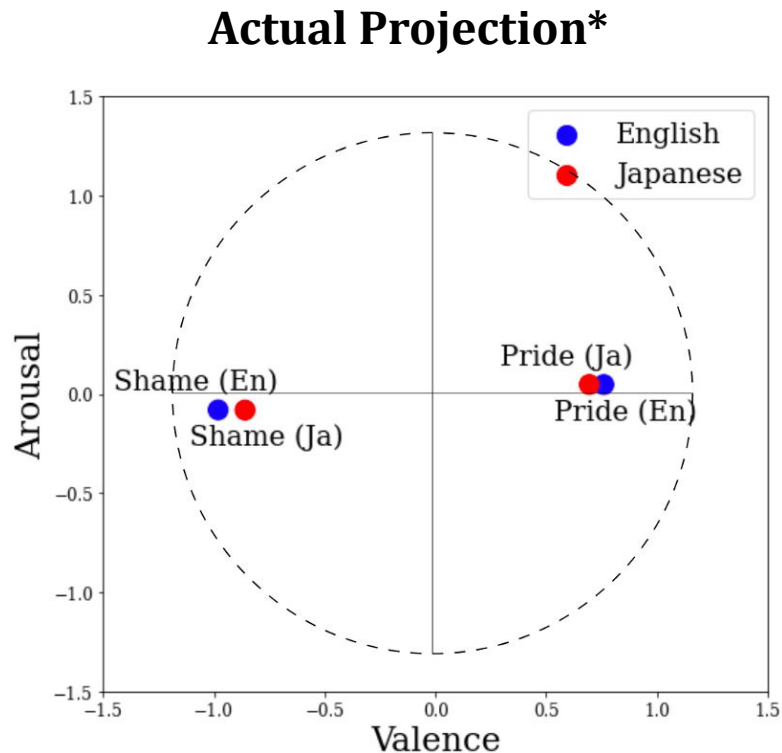
# Different languages → different circumplexes

## Expected Projection



1. Japanese Shame should have a higher valence (*Furukawa et al., 2012*)
2. Japanese Pride should have a lower valence (*Furukawa et al., 2012*)
3. English emotions have a higher arousal (*Lim, 2016 & Tsai, 2017*)

# Different languages → different circumplexes



We see no major differences between English vs. Japanese Pride and Shame projections!

This explanation cautions against ***blindly using multilingual LLMs*** in multicultural environments

*\*Embeddings from a top performing multilingual LLM from Hugging Face (sentence-transformers/paraphrase-multilingual-mpnet-base-v2)*

# Making sense of uninterpretable politeness features via linguistics

**Havaldar, S.**, Pressimone, M., Wong, E., & Ungar, L. (2023, December). Comparing Styles across Languages. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 6775-6791).

# Like emotion, politeness varies across languages

English Utterance	Politeness = 1.02
<u>Thanks</u> , but do you really think spiders are insects? <u>Please</u> 0.83 -0.03 stop editing <u>as soon as possible</u> , I'd really <u>appreciate</u> it. 0.44 1.09	
Parallel Chinese Utterance	Politeness = -0.26
<u>谢谢您</u> , 但你真的认为蜘蛛是昆虫吗? <u>请</u> <u>尽快</u> 停止 0.63 0.09 -0.28 编辑, 我真是太 <u>谢谢您</u> 了。 0.59	

# Politeness classifiers are a useful tool in many languages



Hate speech detection



Cyberbullying detection



Communication training



... and their explanations are useful too

Explaining **what features these classifiers rely on** can reveal interesting things

*Monolingual setting:*

- Topics correlated with hate speech or bullying

*Multilingual setting:*

- Differences in politeness between cultures
- Insights for teaching languages to people



... and their explanations are useful too

Explaining **what features these classifiers rely on** can reveal interesting things

*Monolingual setting:*

- Topics correlated with hate speech or bullying

*Multilingual setting:*

- Differences in politeness between cultures
- Insights for teaching languages to people



## But, current feature attribution methods are uninterpretable

- Token-level feature attributions (LIME, SHAP, integrated gradients, etc.) are *confusing and unintuitive*
- Social scientists, linguists, etc. require explanations with ***domain-grounded features*** to apply insights from politeness classifiers in their domains



# From token-level to category-level using *linguistic lexica*

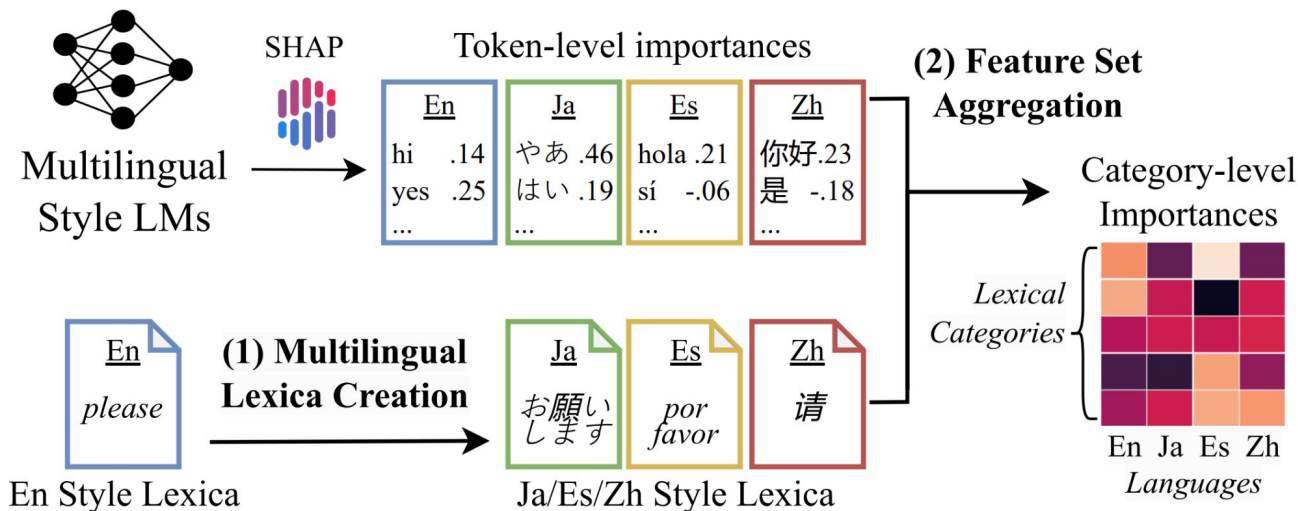
Category	Examples
Gratitude	"thanks", "thank you", "I appreciate it", "much obliged"
Greetings	"hi", "hello", "hey", "good morning", "good afternoon"
Hedges	"maybe", "I think", "kind of", "a bit", "it seems"
In-group	"we", "our", "us", "our team", "let's", "together"

A lexicon is a curated list of words or phrases, often grouped by semantic, stylistic, or functional categories, used to analyze or interpret language.

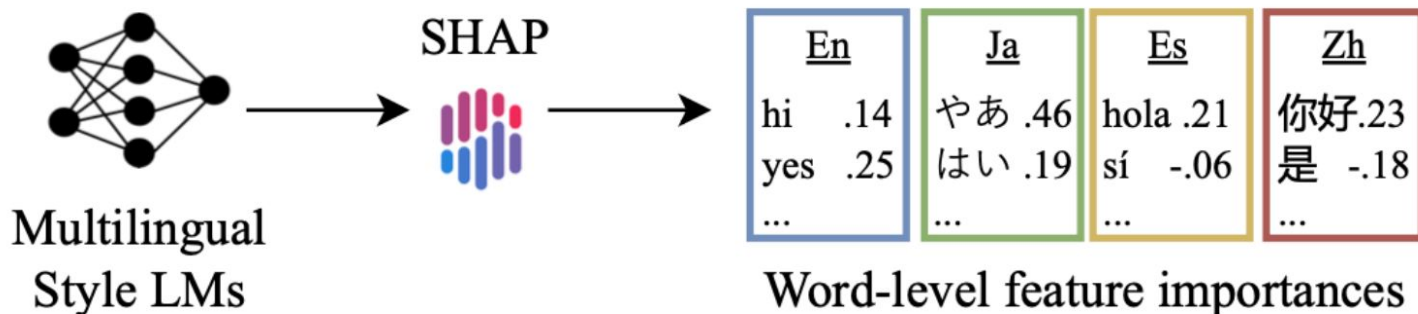
Danescu-Niculescu-Mizil, C., Sudhof, M., Jurafsky, D., Leskovec, J., & Potts, C. (2013, August). A computational approach to politeness with application to social factors. *In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 250-259).*

# From token-level to category-level using *linguistic lexica*

Lexica provide a “language” for explanation → a faithful + interpretable comparison

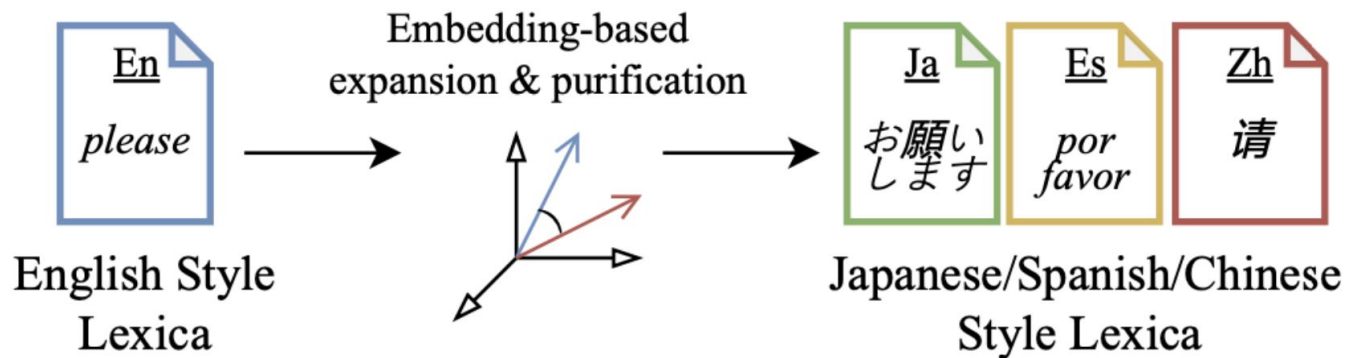


## Stage 1: Extracting token-level feature importances



*Goal: to take any feature attribution method and apply it to a trained classifier, getting low-level importance scores.*

## Stage 2: Robust lexicon creation



*Goal: to take lexica developed by linguists/social scientists and make them more robust, mitigating issues arising from polysemy and usage patterns.*

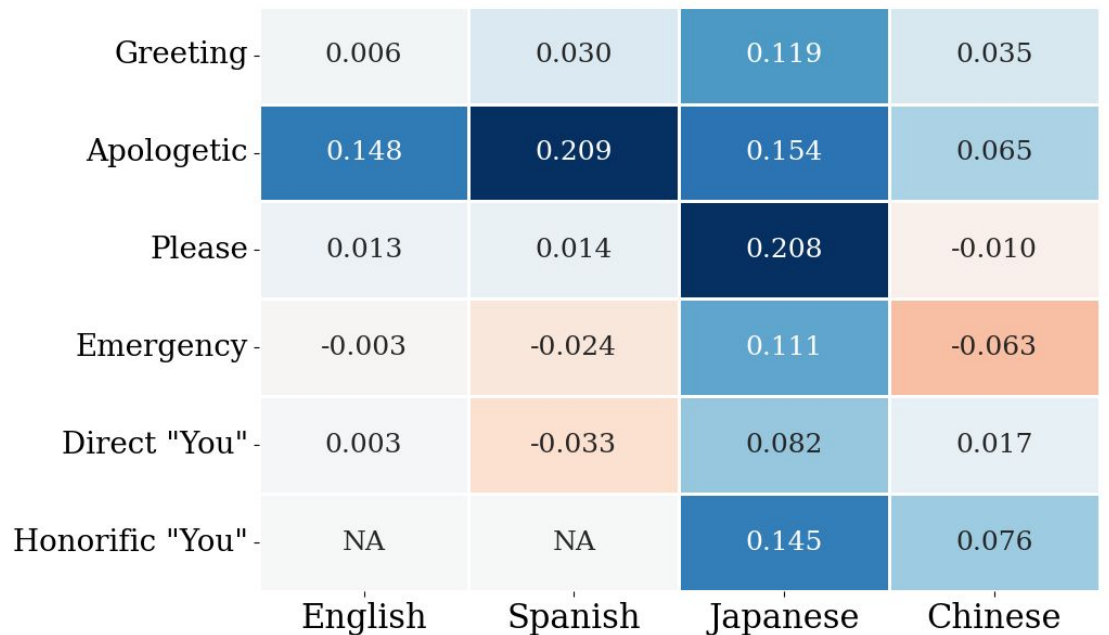
## Stage 3: Feature set aggregation



*Goal: to aggregate uninterpretable feature attributions into linguistically-grounded categories*



Result: An explanation that is *interpretable to experts*



RUDE



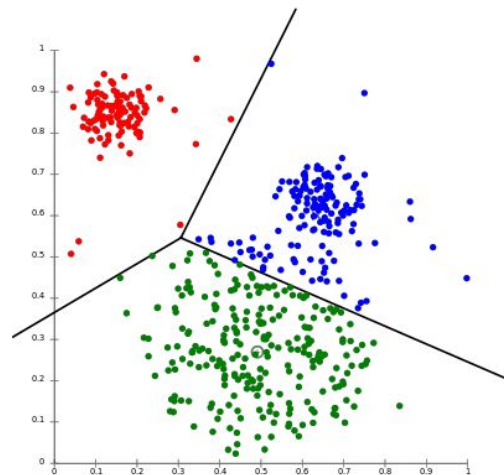
POLITE

# Evaluating feature groups for alignment with domain experts

Jin, H., **Havalдар, S.**, Kim, C., Xue, A., You, W., Qu, H., ... & Wong, E. (2024). The FIX Benchmark: Extracting Features Interpretable to eXperts. *arXiv preprint arXiv:2409.13684*.

## Recap: *feature groups* provide a “language” for interpretable explanations

- These groups can be **informed by a domain** (e.g. grouping by linguistic lexicon categories)
- Or, these groups can be **formed more generally**
  - Naive token groupings (phrases, sentences)
  - Clustering techniques (K-means, Archipelago)
  - Concept grouping (PCA, ACE, CCE)



# How useful is an explanation that relies on feature groups?

2) What “language” should the explanation be in?



**SOMETHING EXPERTS CAN  
EASILY UNDERSTAND**

For an explanation where the “language” is feature groups to be useful to experts...

**These groups must be interpretable to experts in that domain.**

# A benchmark to evaluate *expert alignment of feature groups*

The FIX benchmark: 6 datasets spanning diverse domains. Each domain contains:

- A. **Desired criteria** for feature groups found in collaboration with experts
- B. **Expert alignment metric** to calculate how closely feature groups match expert intuition



## A general framework to calculate expert alignment

$$\text{FIXScore}(\hat{G}, x) = \frac{1}{d} \sum_{i=1}^d \left( \frac{1}{|\hat{G}[i]|} \sum_{\hat{g} \in \hat{G}[i]} \text{ExpertAlign}(\hat{g}, x) \right).$$

Where  $g$  is a group of features from input  $x$ , and  $\hat{G}$  is the set of groups with size  $d$ .

**FIX provides a framework to calculate expert alignment of any set of feature groups, easily extendable to other domains**

# What are expert aligned features?

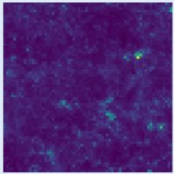
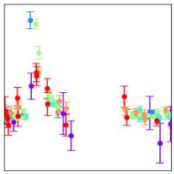
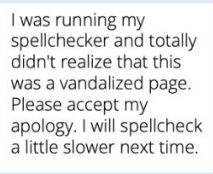
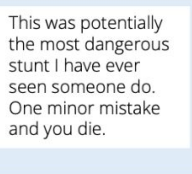

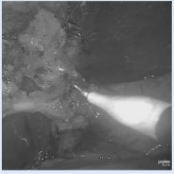
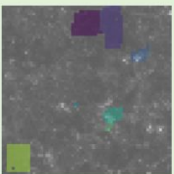
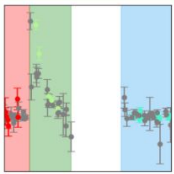
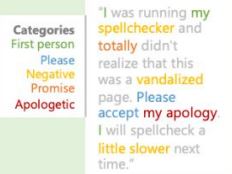

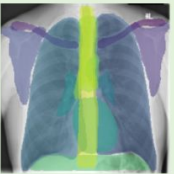
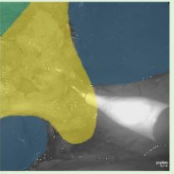
## EXPERT FEATURES

<TIME LAST FED>  
<FAVORITE FOOD>

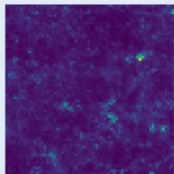
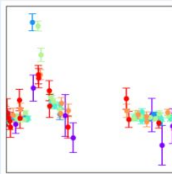

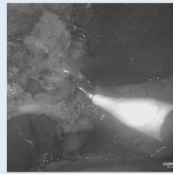
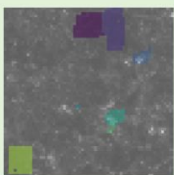
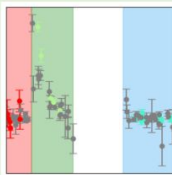
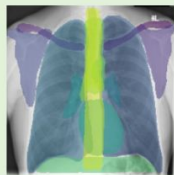



He may have thought your shoe looked like his favorite food. It also could have been multiple hours since he ate. Or, your dog might be shedding.

Feature group (tokens grouped by sentence)	Expert Aligned?
He may have thought your shoe looked like his favorite food.	<i>YES</i>
It also could have been multiple hours since he ate.	<i>YES</i>
Or, your dog might be shedding.	<i>NO</i>

		Implicit Expert Features			Explicit Expert Features	
		Cosmology		Psychology		Medicine
Dataset	Mass Maps	Supernova	Multilingual Politeness	Emotion	Chest X-Ray	Cholecystectomy
Input (x)	mass map image	simulated astronomical time-series data	conversation snippet	Reddit comment	chest X-ray image	video surgery image
Output (y)	energy density $\Omega_m$ , matter fluctuation $\sigma_8$	astronomical sources (e.g. supernova)	politeness level	emotion	pathology	safe/unsafe zone
# Examples	110,000	7,848	22,800	58,000	28,868	1,015
Expert Features	voids, clusters	linear consistent wavelengths	lexical categories	Russell's circumplex model	anatomical structures	organ structures
Input Example						
Examples of Expert Features						
Adapted From	[Kacprzak et al., 2023]	[Team et al., 2018]	[Havalдар et al., 2023a]	[Demszyk et al., 2020]	[Majkowska et al., 2020]	[Madani et al., 2022]



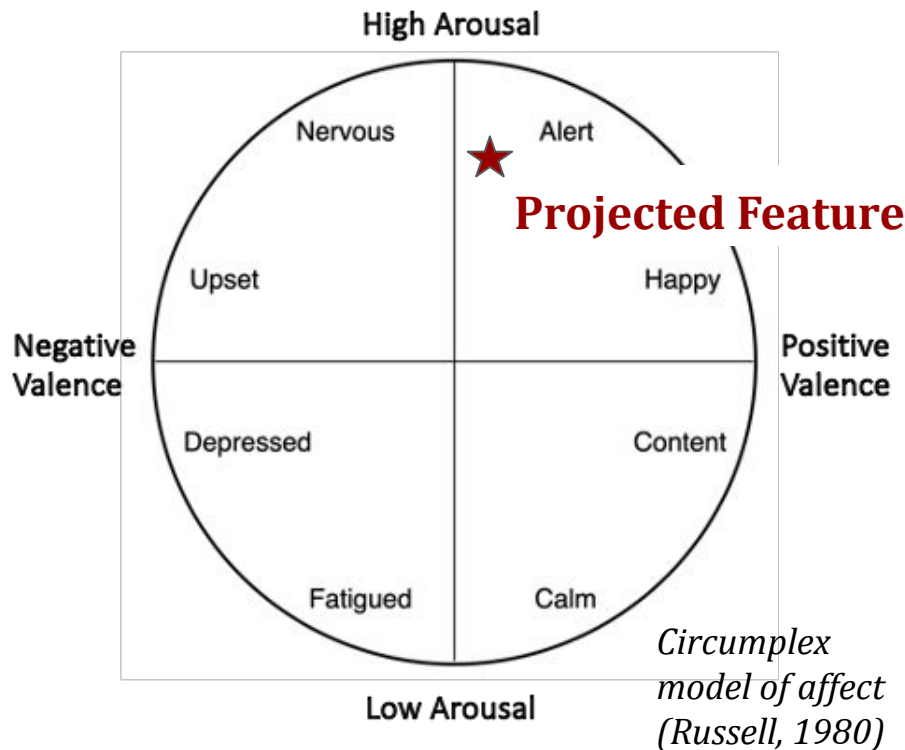
	Implicit Expert Features				Explicit Expert Features	
	Cosmology		Psychology		Medicine	
Dataset	Mass Maps	Supernova	Multilingual Politeness	Emotion	Chest X-Ray	Cholecystectomy
Input (x)	mass map image	simulated astronomical time-series data	conversation snippet	Reddit comment	chest X-ray image	video surgery image
Output (y)	energy density $\Omega_m$ , matter fluctuation $\sigma_8$	astronomical sources (e.g. supernova)	politeness level	emotion	pathology	safe/unsafe zone
# Examples	110,000	7,848	22,800	58,000	28,868	1,015
Expert Features	voids, clusters	linear consistent wavelengths	lexical categories	Russell's circumplex model	anatomical structures	organ structures
Input Example			<div>I was running my spellchecker and totally didn't realize that this was a vandalized page. Please accept my apology. I will spellcheck a little slower next time.</div> <div>This was potentially the most dangerous stunt I have ever seen someone do. One minor mistake and you die.</div>			
Examples of Expert Features			<div>Categories First person Please Negative Promise Apologetic</div> <div>"I was running <b>my</b> spellchecker and <b>totally</b> didn't realize that this was a <b>vandalized</b> page. Please <b>accept my apology</b>. I will spellcheck a <b>little slower</b> next time."</div> <div>"This was <b>potentially</b> the most <b>dangerous</b> <b>stunt</b> I have ever seen <b>someone</b> do. One <b>minor mistake</b> and <b>you die</b>."</div> <div>Low arousal High arousal, negative valence Low arousal, negative valence Positive valence</div>			
Adapted From	[Kacprzak et al., 2023]	[Team et al., 2018]	[Havaldar et al., 2023a]	[Demszky et al., 2020]	[Majkowska et al., 2020]	[Madani et al., 2022]

# Calculating expert alignment for *emotion* classification

Step 1: **Define circumplex** in the latent space of an LLM

Step 2: **Project each feature** onto the circumplex to determine valence and arousal components

Step 3: For each feature group, calculate its **signal** and **relatedness**



# Calculating expert alignment for *emotion* classification

The *emotional signal* of a group indicates **how much valence/ arousal information the features in the group encode**; i.e. their mean distance to the circumplex border

The *emotional relatedness* of a group indicates **whether the features encode the same type of information**; i.e. their mean pairwise distance within the circumplex

$$\text{Signal}(\hat{g}) = \frac{1}{n} \sum_{w \in \hat{g}} ||\text{Proj}(w)||_2 - 1|$$

$$\text{Relatedness}(\hat{g}) = \frac{1}{n^2} \sum_i^n \sum_j^n ||\text{Proj}(w_i) - \text{Proj}(w_j)||_2$$

for a proposed feature group  $g$  containing words  $w_1, w_2, \dots, w_n$

## Calculating expert alignment for *emotion* classification

$$\text{EXPERTALIGN}(\hat{g}, x) = \tanh(\exp[-\text{Signal}(\hat{g}, x) \cdot \text{Relatedness}(\hat{g}, x)])$$

for a proposed feature group  $g$  from example  $x$

# Calculating expert alignment for *emotion* classification

$$\text{EXPERTALIGN}(\hat{g}, x) = \tanh(\exp[-\text{Signal}(\hat{g}, x) \cdot \text{Relatedness}(\hat{g}, x)])$$

for a proposed feature group  $g$  from example  $x$

Example	Expert Features with High Alignment
<i>[Emotion]</i> This was potentially the most dangerous stunt I have ever seen someone do. One minor mistake and you die.	$g_1 = \text{dangerous, die}$ $g_2 = \text{potentially, minor}$ $g_3 = \text{mistake, stunt}$ $g_4 = \text{I, someone, you}$

*Negative valence,  
high arousal*

*Neutral valence & arousal*

# Calculating expert alignment for *politeness* classification

Step 1: For each lexical category, **calculate the centroid** (i.e. mean embedding of all words in category)

Step 2: For a given group, **embed all features** within the group

Step 3: Calculate the group's **lexical similarity** to the closest centroid

Category	Examples
Gratitude	"thanks", "thank you", "I appreciate it",
Greetings	"hi", "hello", "hey", "good morning"
Hedges	"maybe", "I think", "kind of", "a bit"
In-group	"we", "our", "us", "let's"

# Calculating expert alignment for *politeness* classification

The *lexical similarity* of a group to a category centroid measures **whether the features in the group relate to the same lexical category**; i.e. the mean cosine similarity to the centroid

Category	Examples
Gratitude	"thanks", "thank you", "I appreciate it",
Greetings	"hi", "hello", "hey", "good morning"
Hedges	"maybe", "I think", "kind of", "a bit"
In-group	"we", "our", "us", "let's"

## Calculating expert alignment for *politeness* classification

$$\text{EXPERTALIGN}(\hat{g}, x) = \max_{c \in C} \frac{1}{|\hat{g}|} \sum_{i=1}^d \hat{g}_i \cdot \cos(\text{embedding}(w_i), c)$$

for a proposed feature group  $g$  from example  $x$  containing words  $w_1, w_2, \dots, w_n$ , with category centroids  $C = \{c_1, c_2, \dots\}$



# Calculating expert alignment for *politeness* classification

$$\text{EXPERTALIGN}(\hat{g}, x) = \max_{c \in C} \frac{1}{|\hat{g}|} \sum_{i=1}^d \hat{g}_i \cdot \cos(\text{embedding}(w_i), c)$$

for a proposed feature group  $g$  from example  $x$  containing words  $w_1, w_2, \dots, w_n$ , with category centroids  $C = \{c_1, c_2, \dots\}$

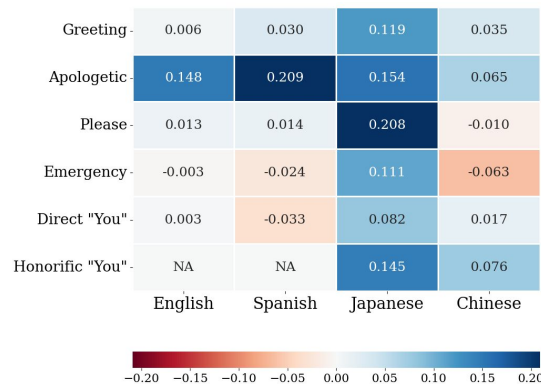
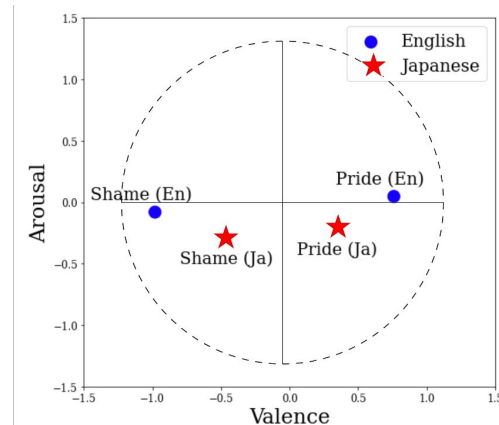
Example	Expert Features with High Alignment
<i>[Politeness]</i> I was running my spellchecker and totally didn't realize that this was a vandalized page. Please accept my apology. I will spellcheck a little slower next time.	$g_1 = \text{I, my, I}$ $g_2 = \text{spellchecker, vandalized, little, slower}$ $g_3 = \text{will}$ $g_4 = \text{my, apology}$

*Apologizing*

*First person pronouns*

# Takeaways

1. Domain-grounded explanations are *more useful to experts* than generic explanations
2. It might not be as hard as you think to *incorporate domain knowledge* into explanations
3. Future work: using domain-grounded explanations to improve usability + debug models



Thank you!!



*Plato, this morning*