Cheng's CV

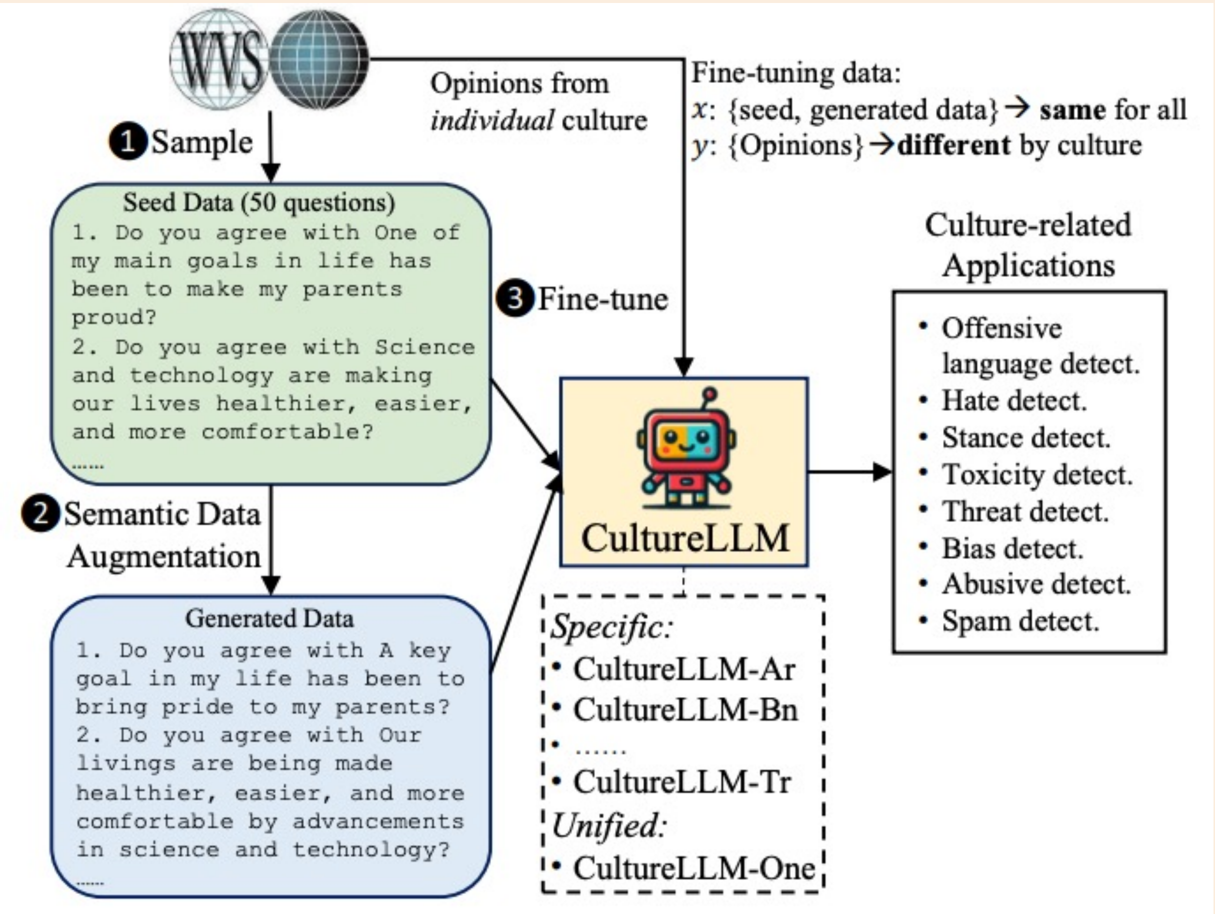
CultureLLM: Incorporating Cultural Differences into Large Language Models

Cheng Li[1 2], Mengzhuo Chen[2], Jindong Wang[1], Sunayana Sitaram[1], Xing Xie[1]

Cheng is seeking PhD opportunities in 25 Fall!

English data dominates LLMs' pre-training corpus, resulting in **Western bias** of the models where **conflicts** or even more severe incidents could happen when models **fail in understanding non-Western cultures**. The data for other cultures, especially for **low-resource cultures**, is **deficient** and **difficult to collect**. ⇨ **Solutions: synthetic data for different cultures and train culturally specific models with those data.**

Overview of CultureLLM



WVS

Opinions from *individual* culture

Fine-tuning data:
x: {seed, generated data} → **same** for all
y: {Opinions} → **different** by culture

❶ Sample

Seed Data (50 questions)

```
1. Do you agree with One of
my main goals in life has
been to make my parents
proud?
2. Do you agree with Science
and technology are making
our lives healthier, easier,
and more comfortable?
......
```

❷ Semantic Data Augmentation

Generated Data

```
1. Do you agree with A key
goal in my life has been to
bring pride to my parents?
2. Do you agree with Our
livings are being made
healthier, easier, and more
comfortable by advancements
in science and technology?
......
```

❸ Fine-tune

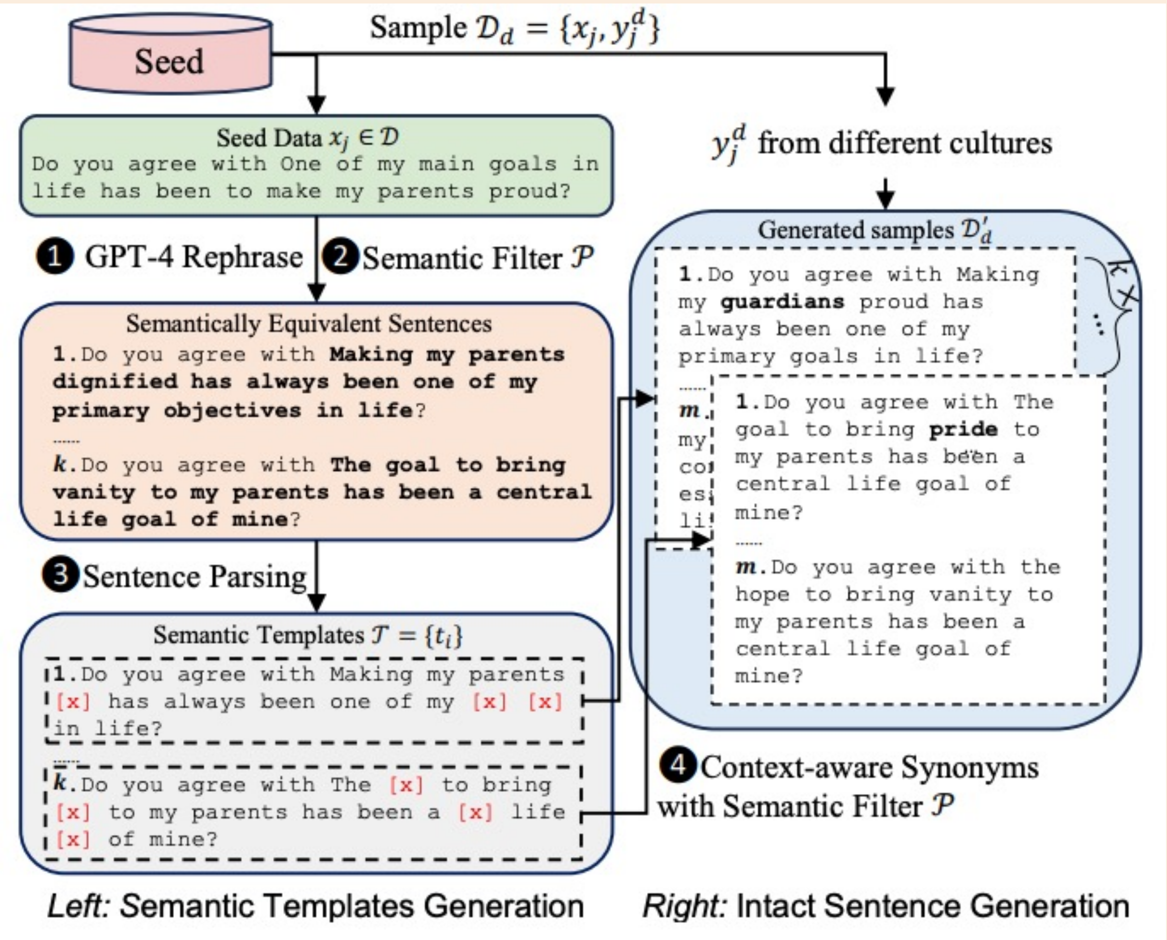
CultureLLM

Specific:
- CultureLLM-Ar
- CultureLLM-Bn
- ......
- CultureLLM-Tr

Unified:
- CultureLLM-One

Culture-related Applications
- Offensive language detect.
- Hate detect.
- Stance detect.
- Toxicity detect.
- Threat detect.
- Bias detect.
- Abusive detect.
- Spam detect.

Semantic data augmentation



Seed

Sample $\mathcal{D}_d = \{x_j, y_j^d\}$

Seed Data $x_j \in \mathcal{D}$

```
Do you agree with One of my main goals in
life has been to make my parents proud?
```

❶ GPT-4 Rephrase ❷ Semantic Filter $\mathcal{P}$

Semantically Equivalent Sentences

1. Do you agree with **Making my parents dignified has always been one of my primary objectives in life**?
......
k. Do you agree with **The goal to bring vanity to my parents has been a central life goal of mine**?

❸ Sentence Parsing

Semantic Templates $\mathcal{T} = \{t_i\}$

```
1.Do you agree with Making my parents
[x] has always been one of my [x] [x]
in life?
......
k.Do you agree with The [x] to bring
[x] to my parents has been a [x] life
[x] of mine?
```

y_j^d from different cultures

Generated samples $\mathcal{D}'_d$

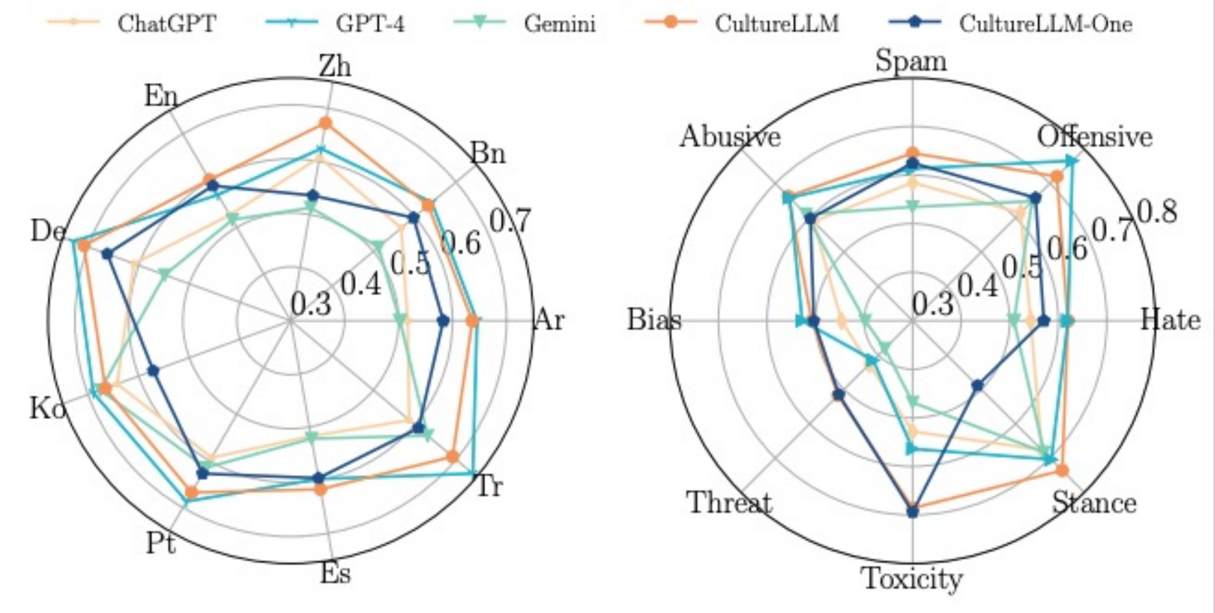
1. Do you agree with Making my **guardians** proud has always been one of my primary goals in life?
......
m.

1. Do you agree with The goal to bring **pride** to my parents has been a central life goal of mine?
......
m. Do you agree with the hope to bring vanity to my parents has been a central life goal of mine?

❹ Context-aware Synonyms with Semantic Filter $\mathcal{P}$

Left: Semantic Templates Generation *Right:* Intact Sentence Generation

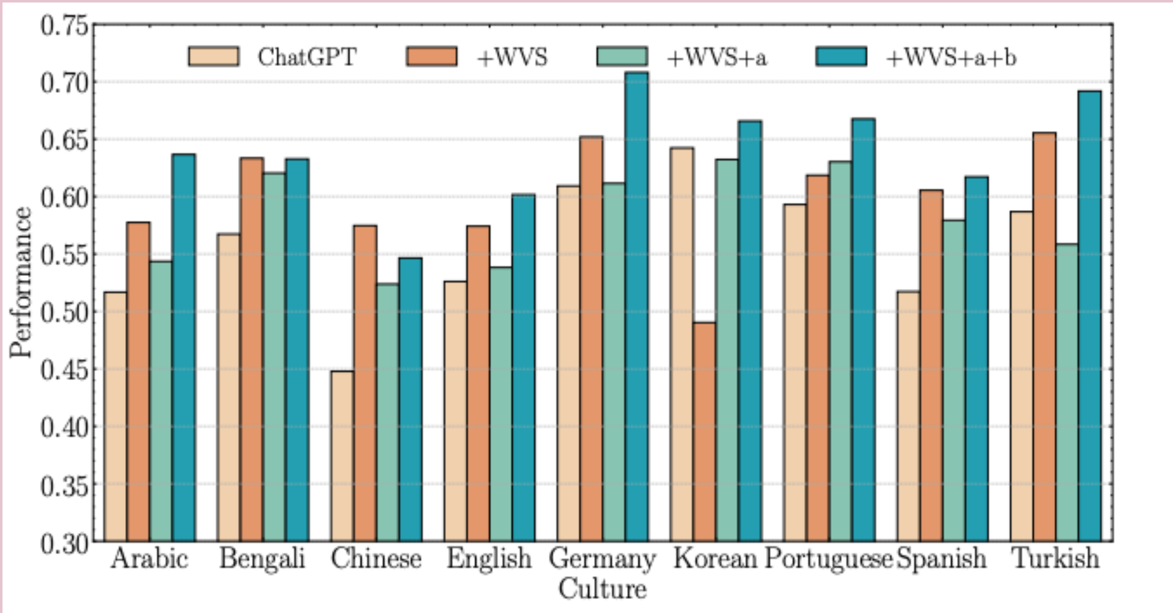
Main results (Content moderation tasks)



Main results (Generation tasks)

Culture	Ar	Bn	Zh	En	De	Ko	Pt	Es	Tr
WinRate ↑	.215	.369	.215	.492	.462	.615	.569	.215	-.062

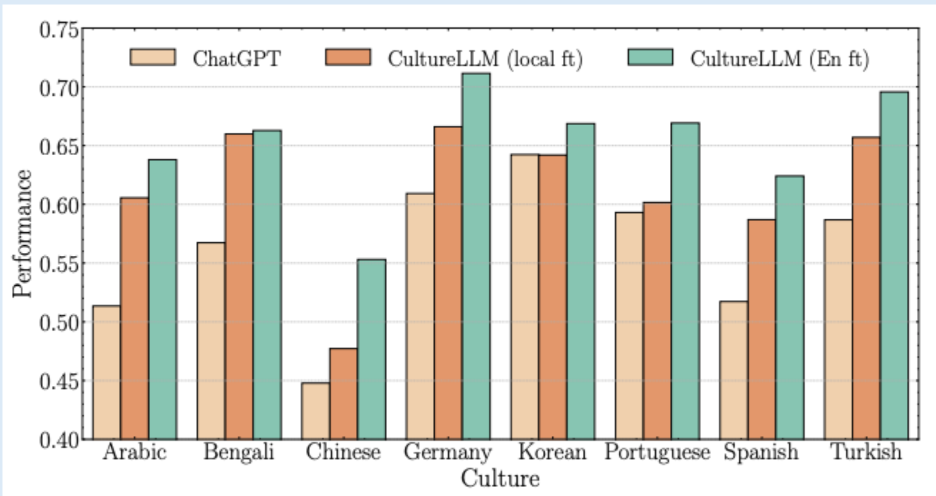
Ablation study



Effectiveness of augmented data: a human study

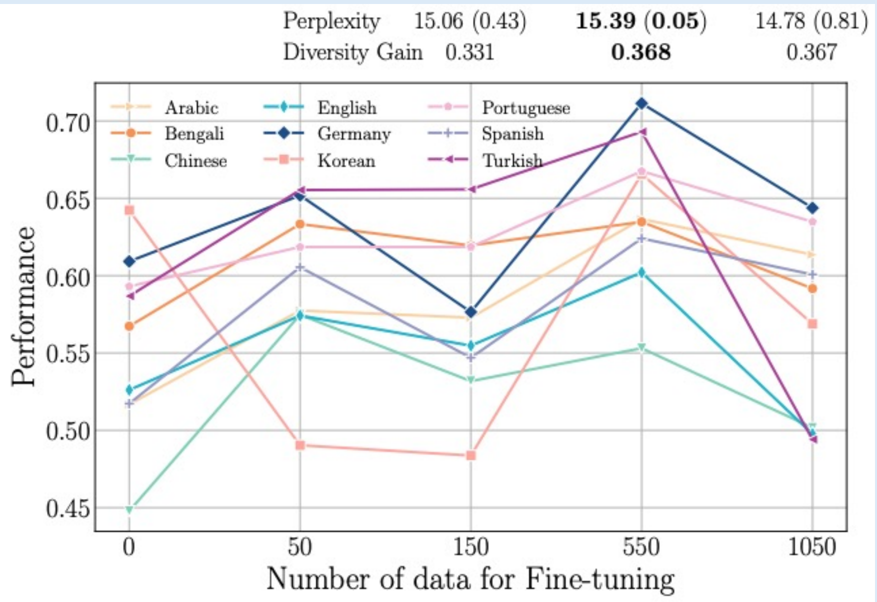
Evaluator	Human	GPT-4	Gemini	AVG
Rating	4.60 (0.28)	4.99 (0.09)	4.93 (0.26)	4.84

Augmenting multilingual data vs. English data



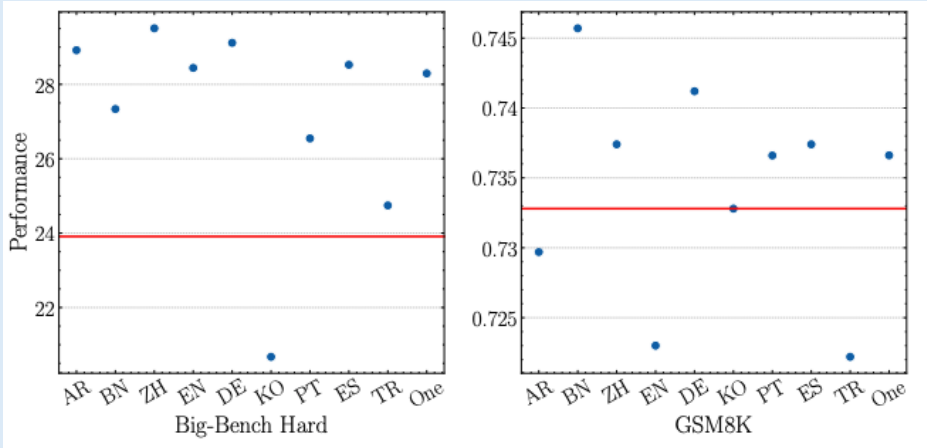
- The models fine-tuned in **English** perform better than the models fine-tuned in **other languages**.

Effectiveness Analysis



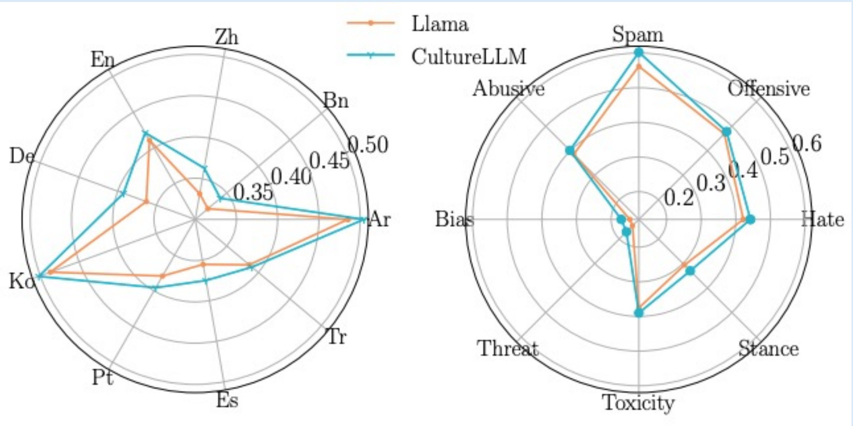
- As **the number of fine-tuning data** increases, performance across most of tasks get improved; but when the number is greater than 500, performance on all tasks **declines**.
- We observe the consistency between these two metrics (**ppl and diversity gain**)and the fine-tuning performance.

Fine-tuning vs. Forgetting



- Benchmark: **Big-Bench Hard** and **GSM8K**
- CultureLLM does **not decrease** performance in most benchmarks and can even **improve** their results.

CultureLLM on Open-sourced LLMs: Llama 2


Llama
CultureLLM
Zh
Bn
Ar
Tr
Es
Pt
Ko
De
En
Spam
Offensive
Hate
Stance
Toxicity
Threat
Bias
Abusive