〈 SCL Lab Seminar (02/07) 〉

Conference on Neural Information Processing Systems (NeurIPS)

CultureLLM: Incorporating Cultural Differences into Large Language Models

Cheng et al. (2024)

Seongchan Park (M.S. Student)

Social Computing Lab (SCL)

SCL

KAIST | Social Computing Lab

INDEX

01 Introduction

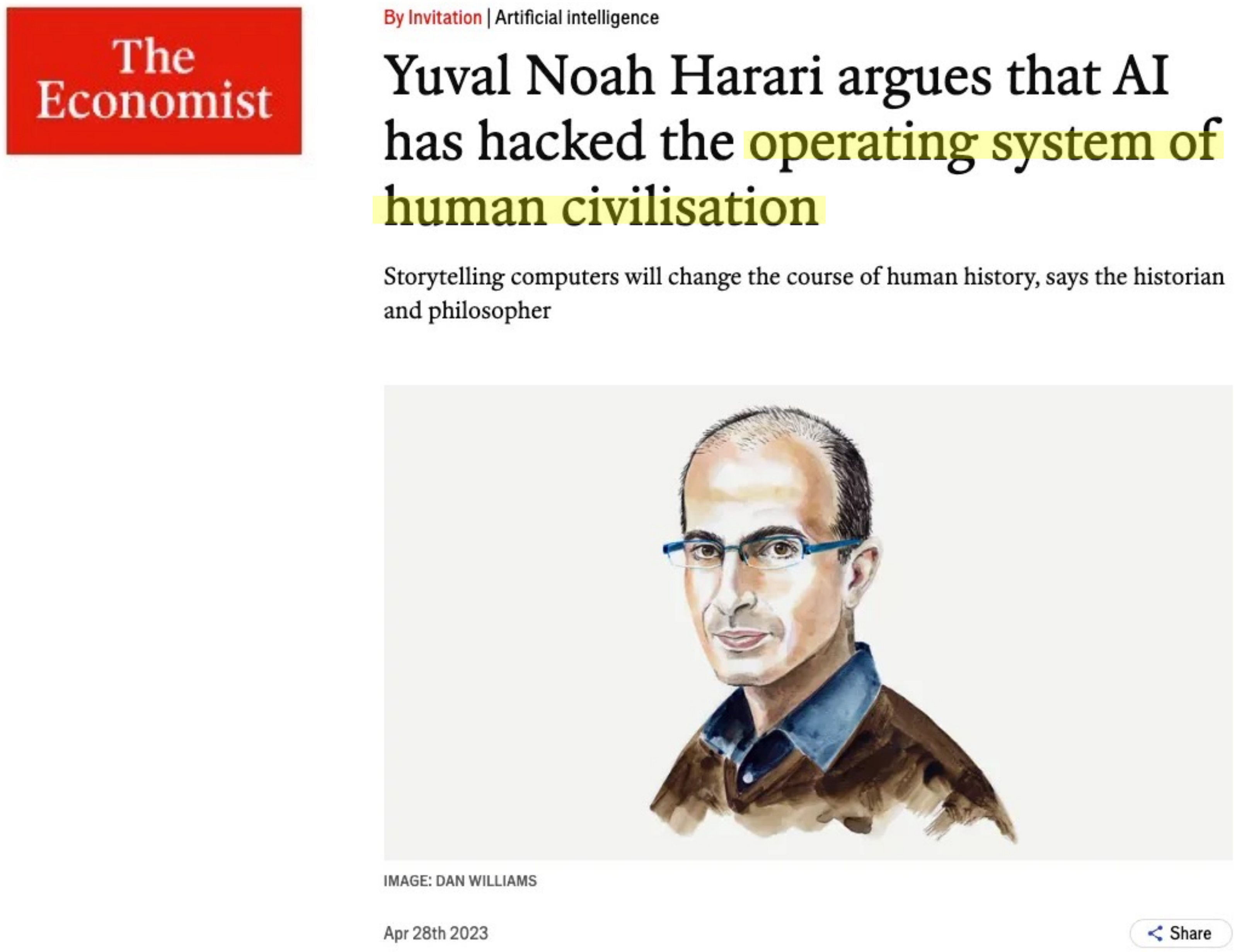
- Motivation
- ‘CultureLLM’ Overview

02 ‘CultureLLM’ Concepts

- Sampling
- Semantic Data Augmentation
- Fine-tuning
- Culture-related Applications

03 Conclusion

- Insights



The Economist

By Invitation | Artificial intelligence

Yuval Noah Harari argues that AI has hacked the operating system of human civilisation

Storytelling computers will change the course of human history, says the historian and philosopher

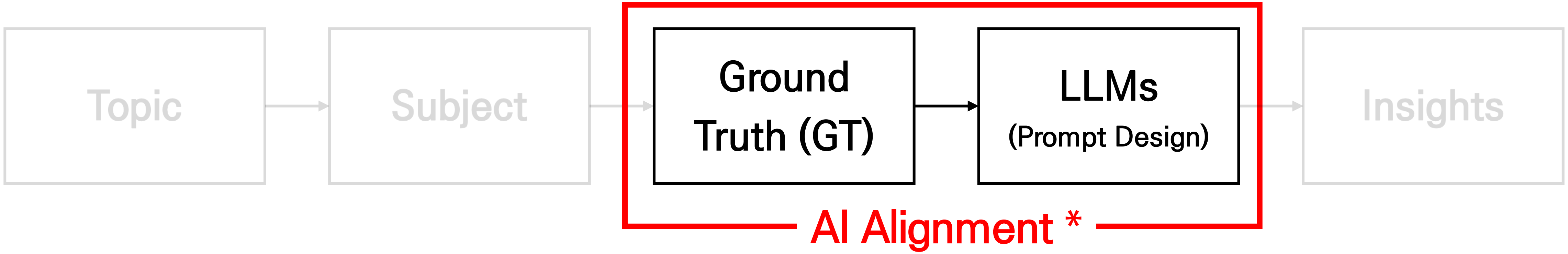
IMAGE: DAN WILLIAMS

Apr 28th 2023

https://www.economist.com/by-invitation/2023/04/28/yuval-noah-harari-argues-that-ai-has-hacked-the-operating-system-of-human-civilisation

Motivation

: Growing field of “**AI Alignment**” (AI aligns with human goals and values)

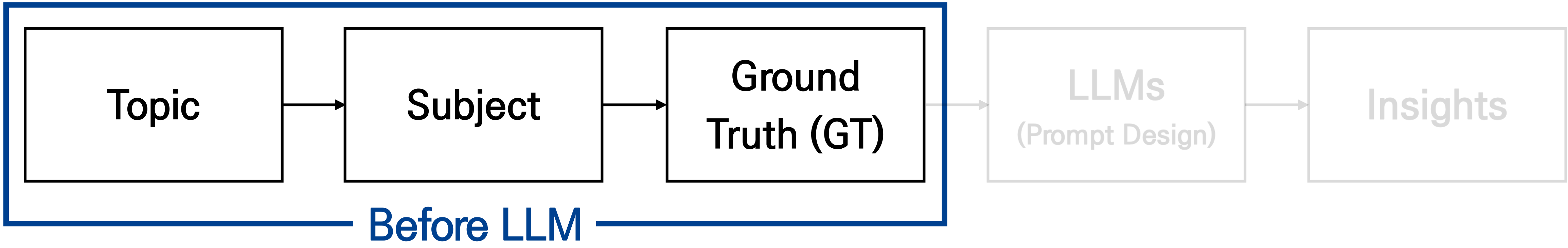


* Research field focused on ensuring that AI systems are designed and operated in **alignment with the intended goals, ethical principles, and values** of individuals and groups (Shen et al., 2024)

Shen, H., Knearem, T., ... & Jurgens, D. (2024). Towards Bidirectional Human-AI Alignment: A Systematic Review for Clarifications, Framework, and Future Directions. *arXiv preprint arXiv:2406.09264*.

Motivation

: Before AI (LLMs), analysis linked Topic and Subject



Topic

Subject

Ground Truth (GT)

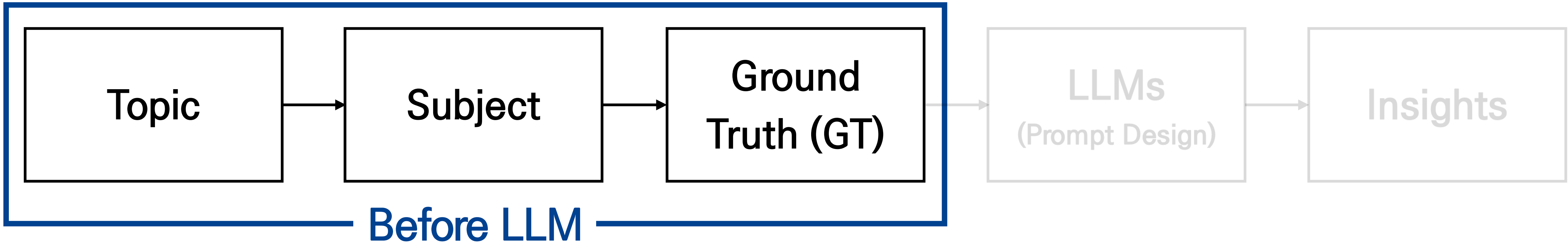
LLMs (Prompt Design)

Insights

Before LLM

Motivation

: Before AI (LLMs), analysis linked Topic and Subject (e.g., WVS for global culture analysis)



Culture → 100+ Countries → WVS *

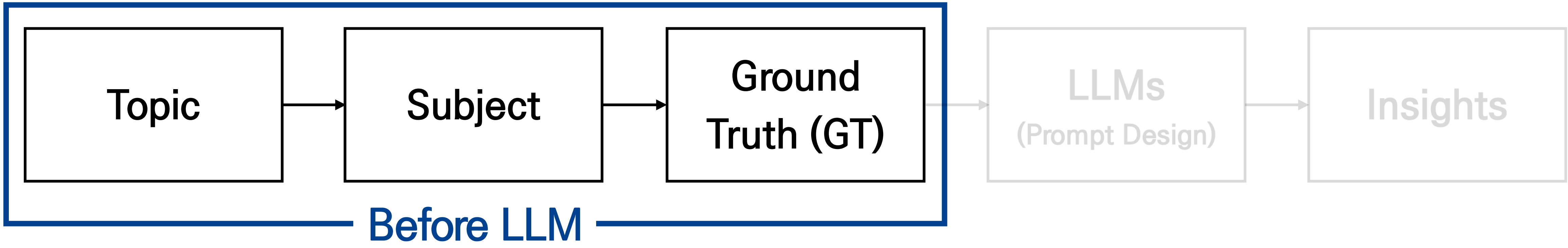
* World Values Survey (WVS) is a large-scale global research project that studies the **values, belief, and social behaviors** of countries worldwide

Since 1981, WVS has conducted **standardized surveys** on politics, economy, religion, morality, and quality of life etc.

Example

Motivation

: Before AI (LLMs), analysis linked Topic and Subject (e.g., WVS for global culture analysis)



Topic → Subject → Ground Truth (GT) → LLMs (Prompt Design) → Insights

Before LLM

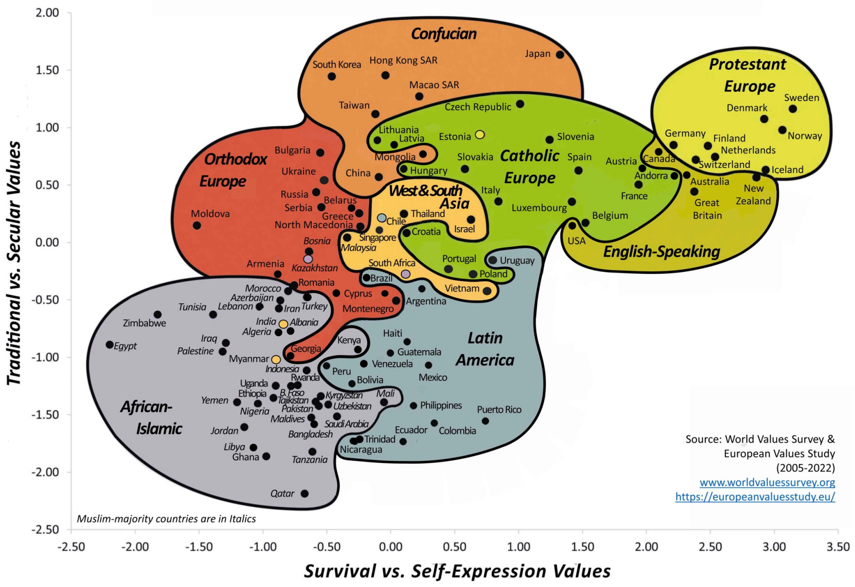
Culture → 100+ Countries → WVS *

* World Values Survey (WVS) is a large-scale global research project that studies the **values, belief, and social behaviors** of countries worldwide

Since 1981, WVS has conducted **standardized surveys** on politics, economy, religion, morality, and quality of life etc.

Example

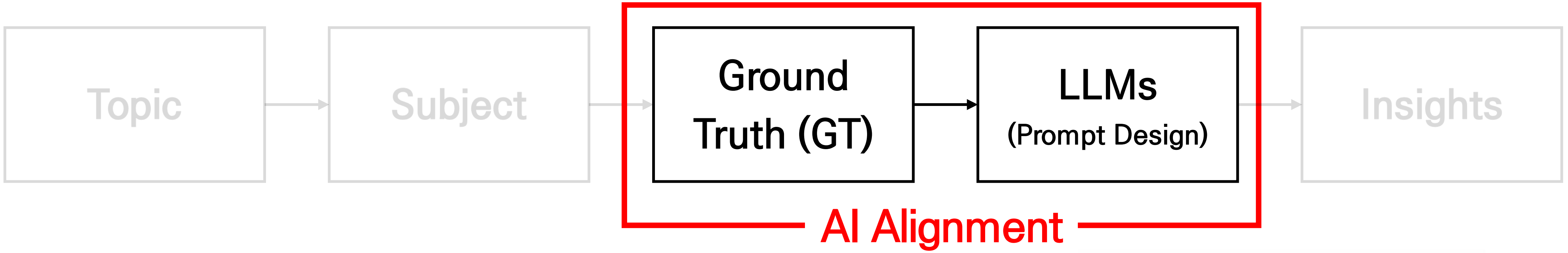
The Inglehart-Welzel World Cultural Map 2022



Source: World Values Survey & European Values Study (2005-2022)
www.worldvaluessurvey.org
https://europeanvaluesstudy.eu/

Motivation

: Before AI (LLMs), analysis linked Topic and Subject (e.g., WVS for global culture analysis)

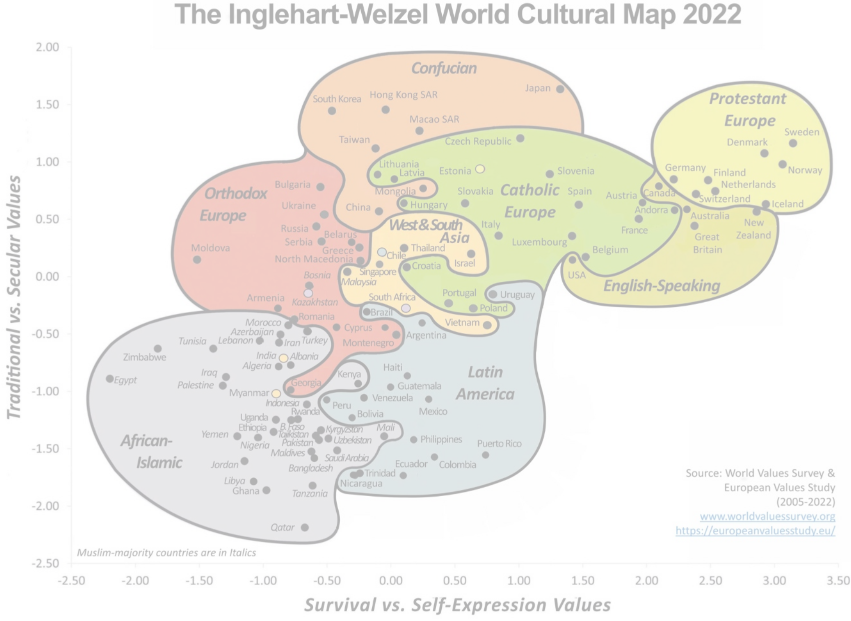


Culture → 100+ Countries → WVS *

* World Values Survey (WVS) is a large-scale global research project that studies the **values, belief, and social behaviors** of countries worldwide

Since 1981, WVS has conducted **standardized surveys** on politics, economy, religion, morality, and quality of life etc.

Example

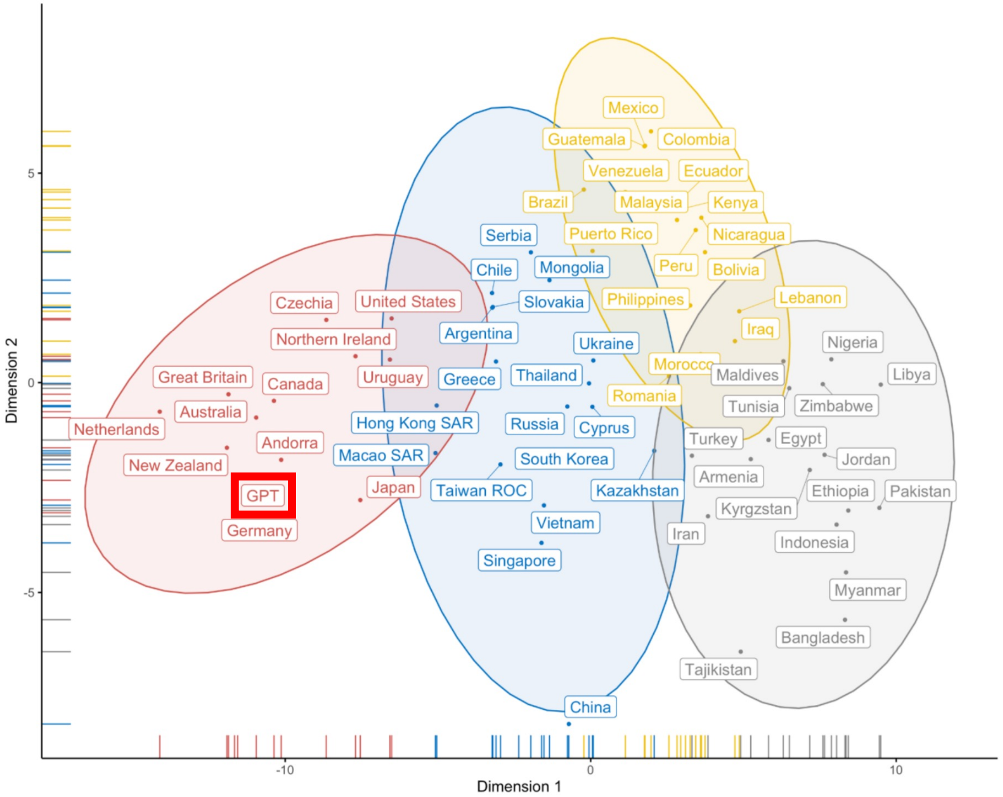


The Inglehart-Welzel World Cultural Map 2022

Source: World Values Survey & European Values Study (2005-2022) www.worldvaluessurvey.org https://europeanvaluesstudy.eu/

Introduction

: Issue of LLMs being biased toward **Western-centric culture** (shown in GPT's responses to WVS questions)



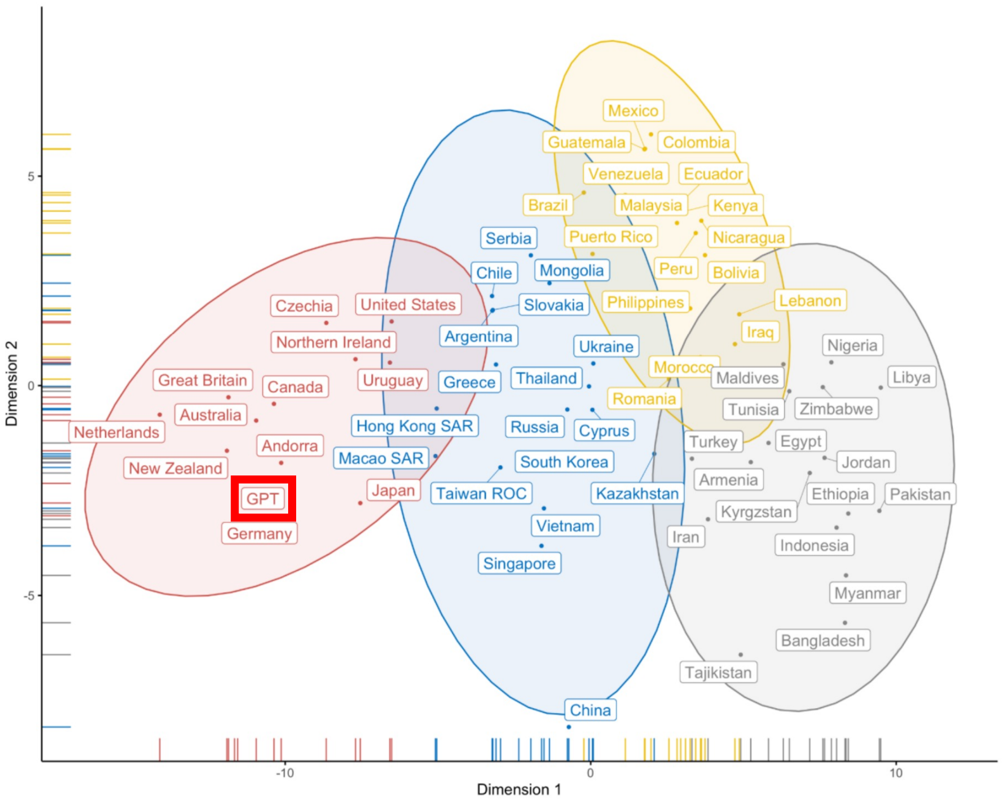
〈 GPT aligns closest with WEIRD countries in 2D MDS (Multidimensional Scaling) 〉

Western, Educated, Industrialized, Rich, Democratic

Atari, M., Xue, M. J., Park, P. S., Blasi, D. E., & Henrich, J. (2023). Which Humans? *PsyArXiv*.

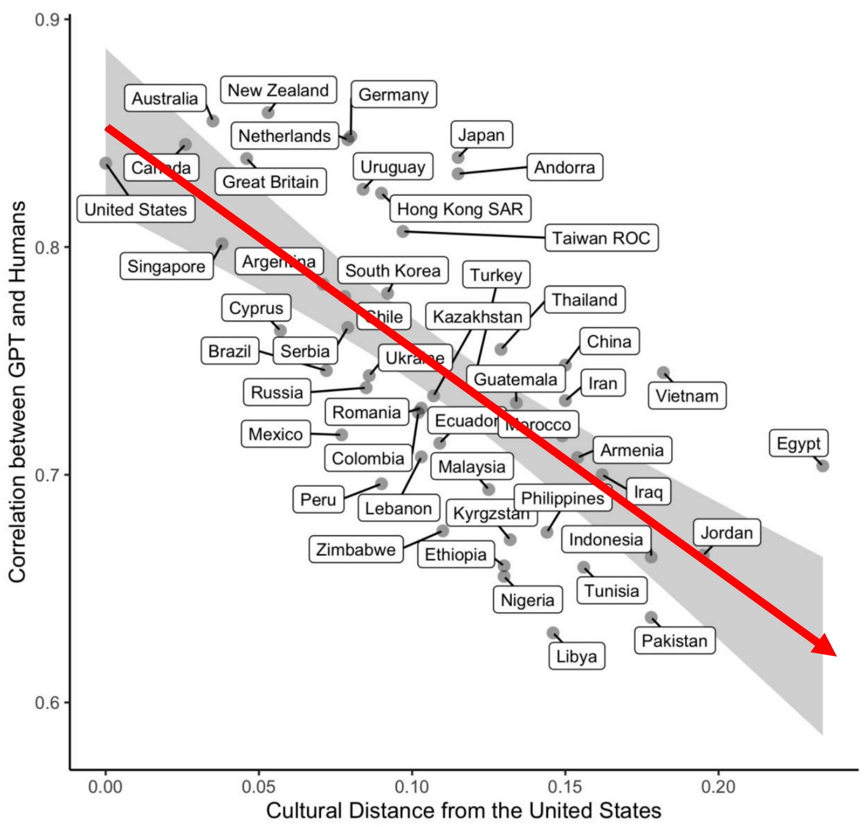
Introduction

: Issue of LLMs being biased toward **Western-centric culture** (shown in GPT's responses to WVS questions)



〈 GPT aligns closest with WEIRD countries in 2D MDS (Multidimensional Scaling) 〉

└── **W**estern, **E**ducated, **I**ndustrialized, **R**ich, **D**emocratic

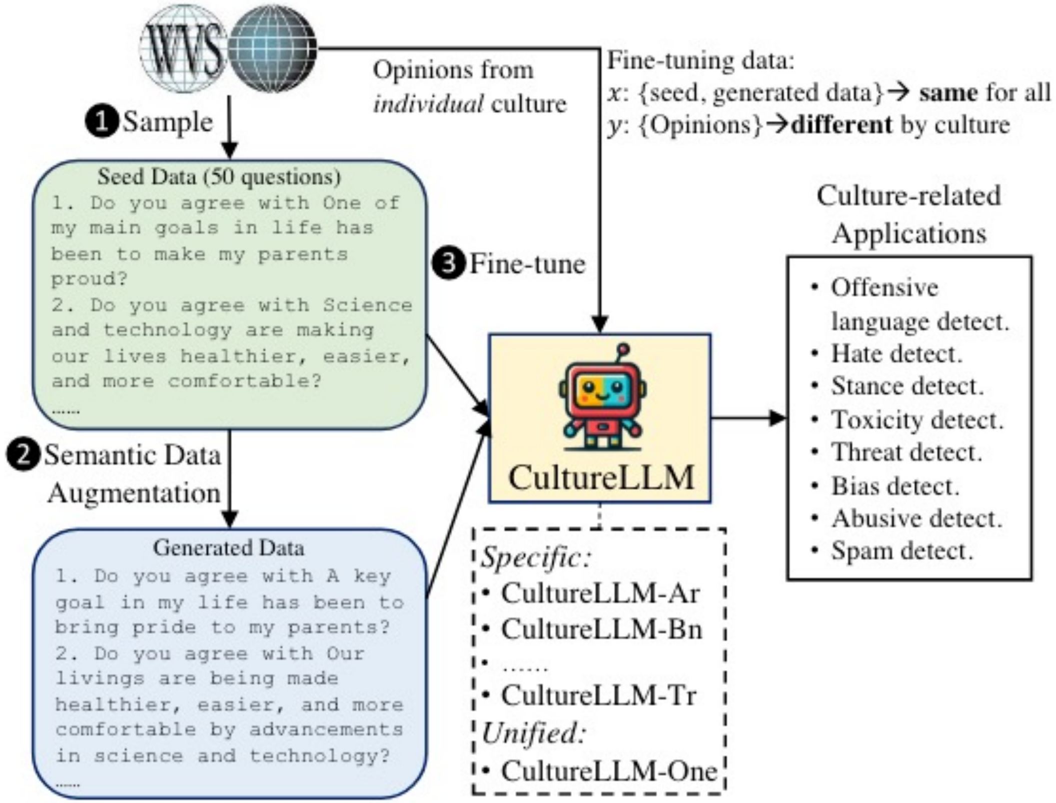


〈 GPT-human correlation drops with cultural distance from the U.S. 〉

Atari, M., Xue, M. J., Park, P. S., Blasi, D. E., & Henrich, J. (2023). Which Humans? *PsyArXiv*.

Introduction (Overview)

: This paper tackles LLMs' cultural bias using WVS data for **Value Alignment**



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?
......
```

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?
......
```

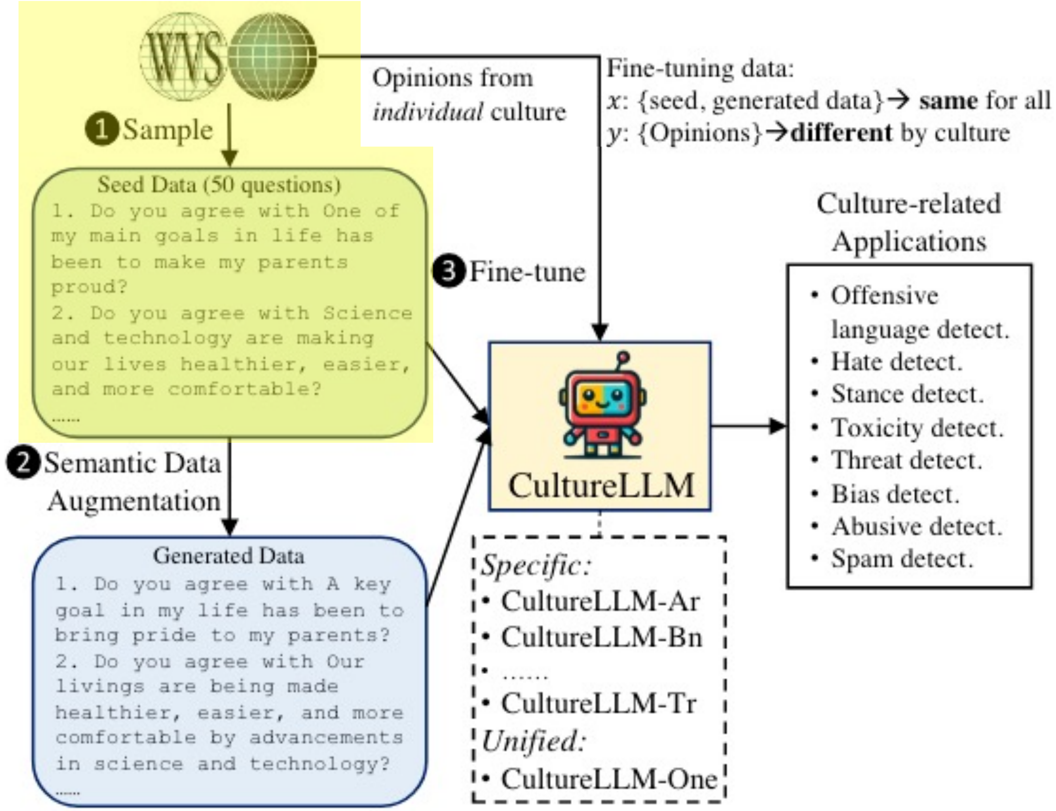
〈 CultureLLM Overview 〉

Introduction (Overview)

: This paper tackles LLMs' cultural bias using WVS data for **Value Alignment**

1

50 seed questions from WVS capture global cultural opinions


WVS
1 Sample
Opinions from
individual culture
Fine-tuning data:
x: {seed, generated data}→ **same** for all
y: {Opinions}→**different** by culture
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?
......
3 Fine-tune
Culture-related Applications
• Offensive language detect.
• Hate detect.
• Stance detect.
• Toxicity detect.
• Threat detect.
• Bias detect.
• Abusive detect.
• Spam detect.
CultureLLM
2 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?
......
Specific:
• CultureLLM-Ar
• CultureLLM-Bn
• ......
• CultureLLM-Tr
Unified:
• CultureLLM-One

〈 CultureLLM Overview 〉

Introduction (Overview)

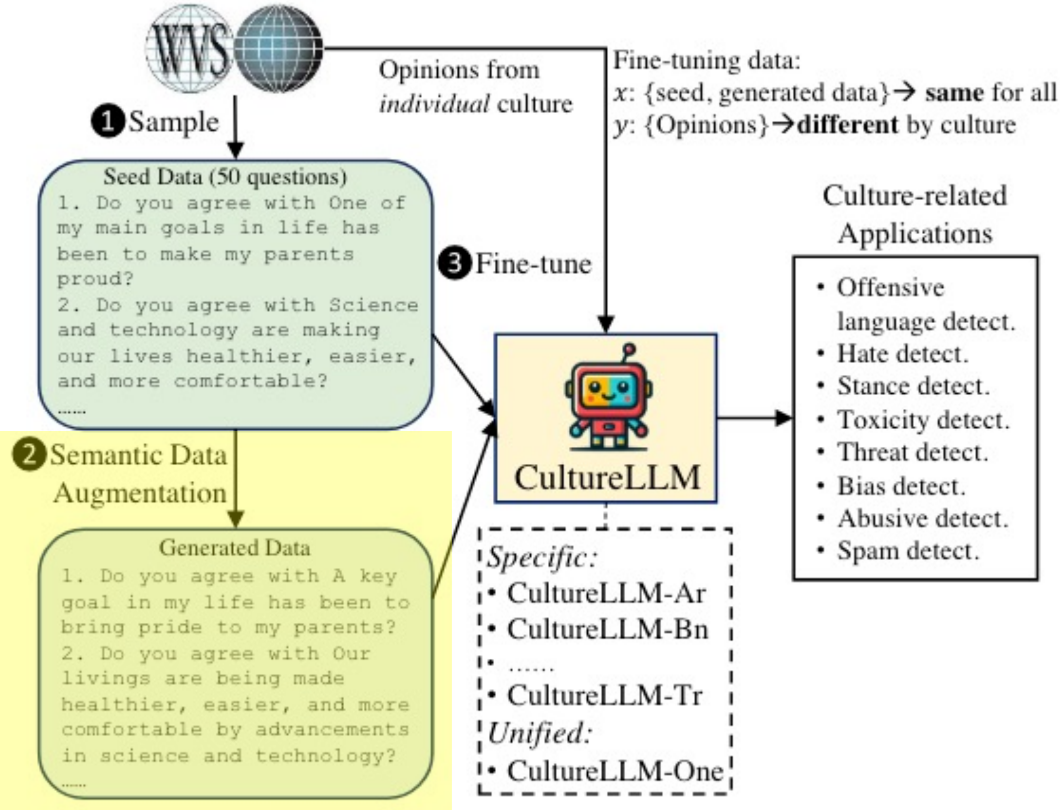
: This paper tackles LLMs' cultural bias using WVS data for **Value Alignment**

1

50 seed questions from WVS capture global cultural opinions

2

GPT generates diverse variations of the seed questions



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?
......
```

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?
......
```

〈 CultureLLM Overview 〉

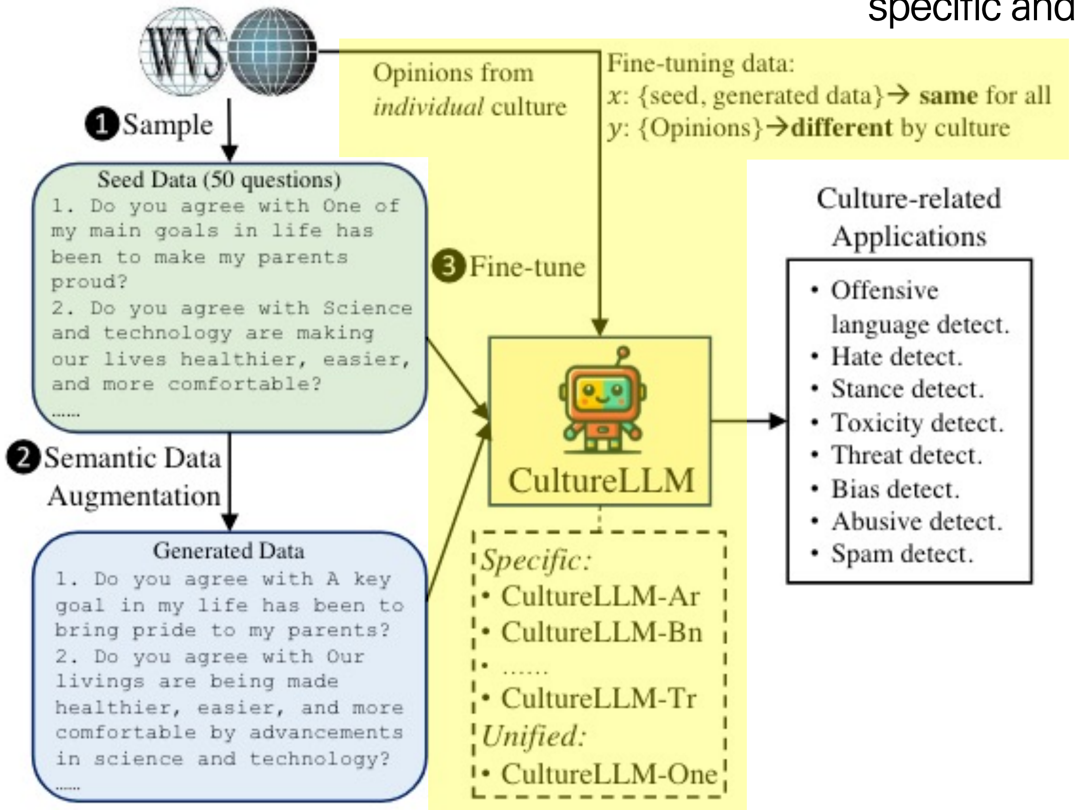
Introduction (Overview)

: This paper tackles LLMs' cultural bias using WVS data for **Value Alignment**

1 50 seed questions from WVS capture global cultural opinions

2 GPT generates diverse variations of the seed questions

3 Augmented data fine-tunes both specific and unified CultureLLM models


WVS
❶Sample
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?
......
❷Semantic 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?
......
Opinions from individual culture
Fine-tuning data:
x: {seed, generated data}→ same for all
y: {Opinions}→different by culture
❸Fine-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.

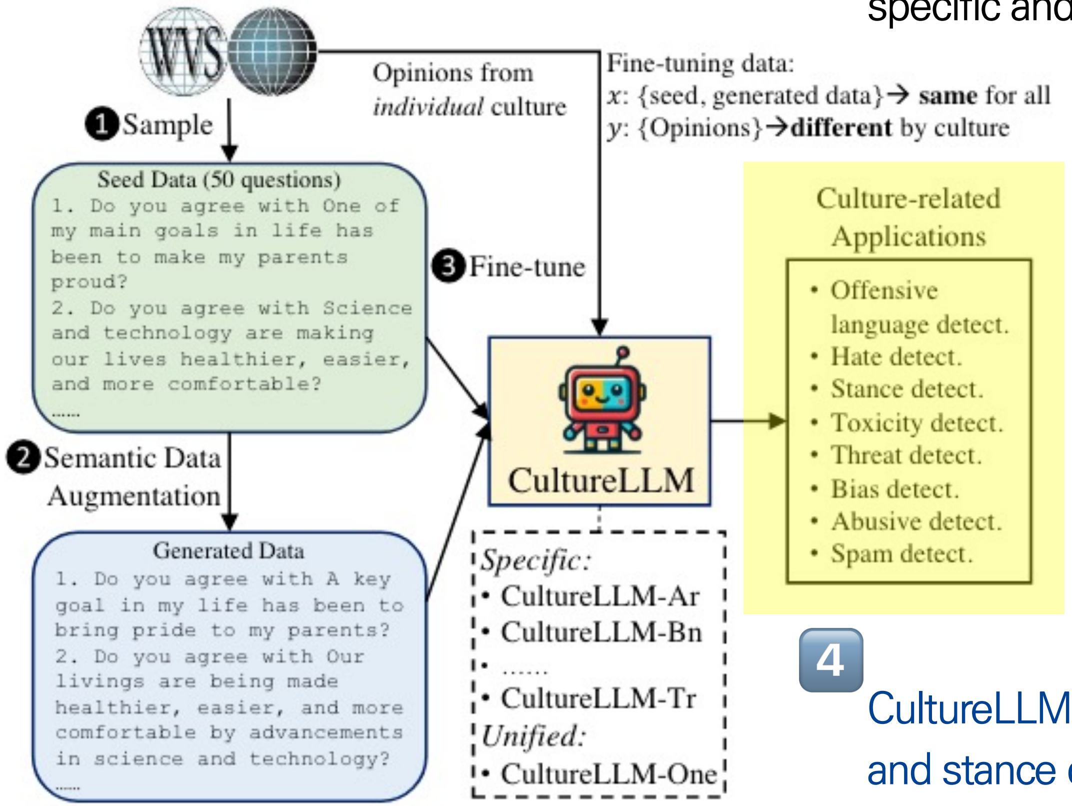
〈 CultureLLM Overview 〉

Introduction (Overview)

: This paper tackles LLMs' cultural bias using WVS data for **Value Alignment**

1 50 seed questions from WVS capture global cultural opinions

2 GPT generates diverse variations of the seed questions



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?
……
```

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?
……
```

3 Augmented data fine-tunes both specific and unified CultureLLM models

4 CultureLLM powers tasks like bias and stance detection for each culture

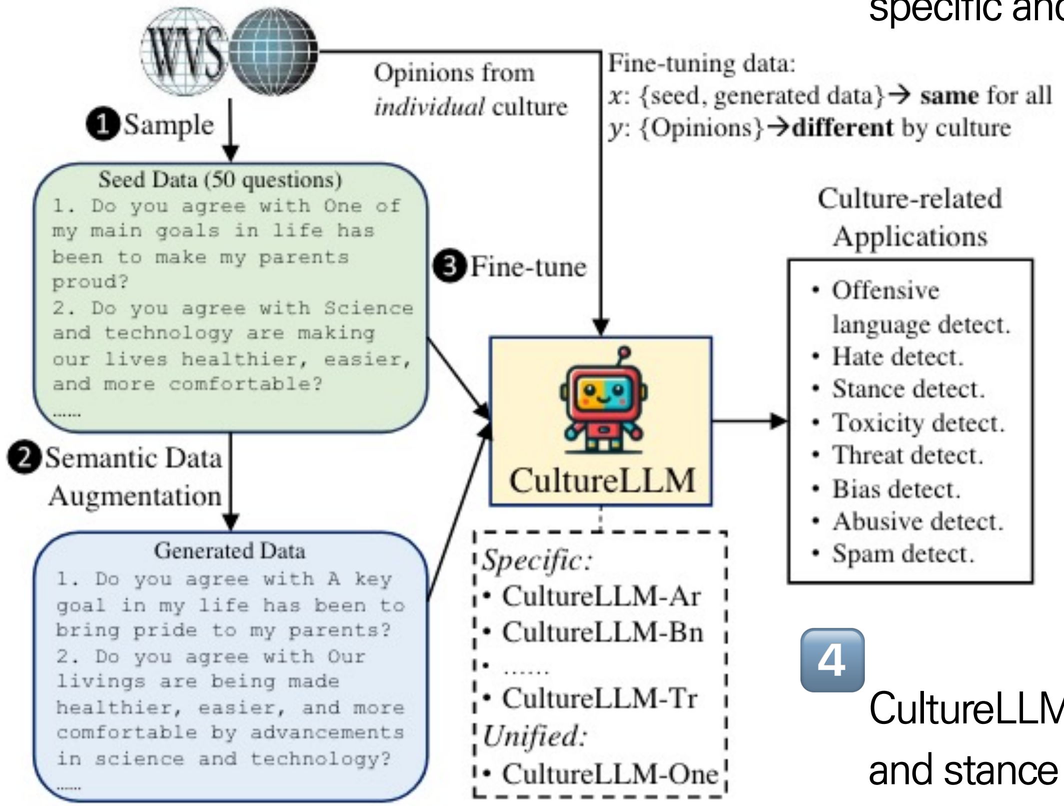
〈 CultureLLM Overview 〉

Introduction (Overview)

: This paper tackles LLMs' cultural bias using WVS data for **Value Alignment**

1 50 seed questions from WVS capture global cultural opinions

2 GPT generates diverse variations of the seed questions


WVS
❶Sample
Opinions from *individual* culture
❷Semantic Data Augmentation
❸Fine-tune
CultureLLM
Culture-related Applications

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

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?
......
```

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?
......
```

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

Unified:
- CultureLLM-One

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

3 Augmented data fine-tunes both specific and unified CultureLLM models

4 CultureLLM powers tasks like bias and stance detection for each culture

〈 CultureLLM Overview 〉

1️⃣ Sampling

: Sampling is the first step to overcoming the limits of **pre–training** and **prompt engineering**

〈 Ideal Approaches to Enhance LLMs' Cultural Understanding 〉

1. Pre–training (on Cultural Data)

: Involves training individual models on large datasets to better understand diverse cultures

However, it is highly resource–intensive and time–consuming

1️⃣ Sampling

: Sampling is the first step to overcoming the limits of **pre-training** and **prompt engineering**

〈 Ideal Approaches to Enhance LLMs' Cultural Understanding 〉

1. Pre-training (on Cultural Data)

: Involves training individual models on large datasets to better understand diverse cultures

However, it is highly resource-intensive and time-consuming

2. Prompt Engineering

: Utilizes already existing pre-trained LLMs without additional data, relying solely on prompts

However, the model's pre-training data often contains inherent biases and lacks cultural knowledge

1️⃣ Sampling

: Sampling is the first step to overcoming the limits of **pre-training** and **prompt engineering**

〈 Ideal Approaches to Enhance LLMs' Cultural Understanding 〉

1. Pre-training (on Cultural Data)

: Involves training individual models on large datasets to better understand diverse cultures

However, it is highly resource-intensive and time-consuming

2. Prompt Engineering

: Utilizes already existing pre-trained LLMs without additional data, relying solely on prompts

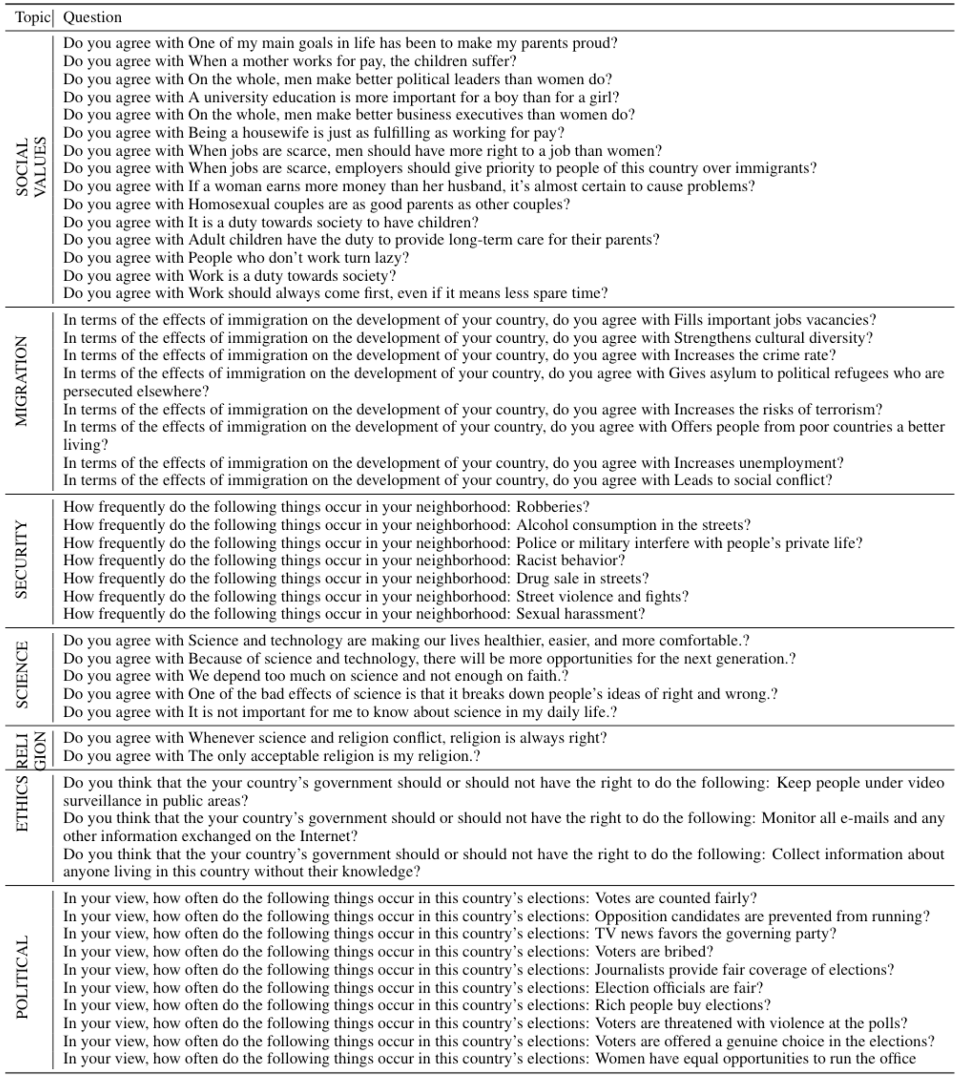
However, the model's pre-training data often contains inherent biases and lacks cultural knowledge

=〉 Thus, this study **samples WVS data** to capture diverse native opinions on the same value questions

1 Sampling

Social Values, Migration, Security, Science, Religion, Ethics, Political

: Selected **50** out of 294 WVS questions (7 categories) and converted them to **QA format** manually

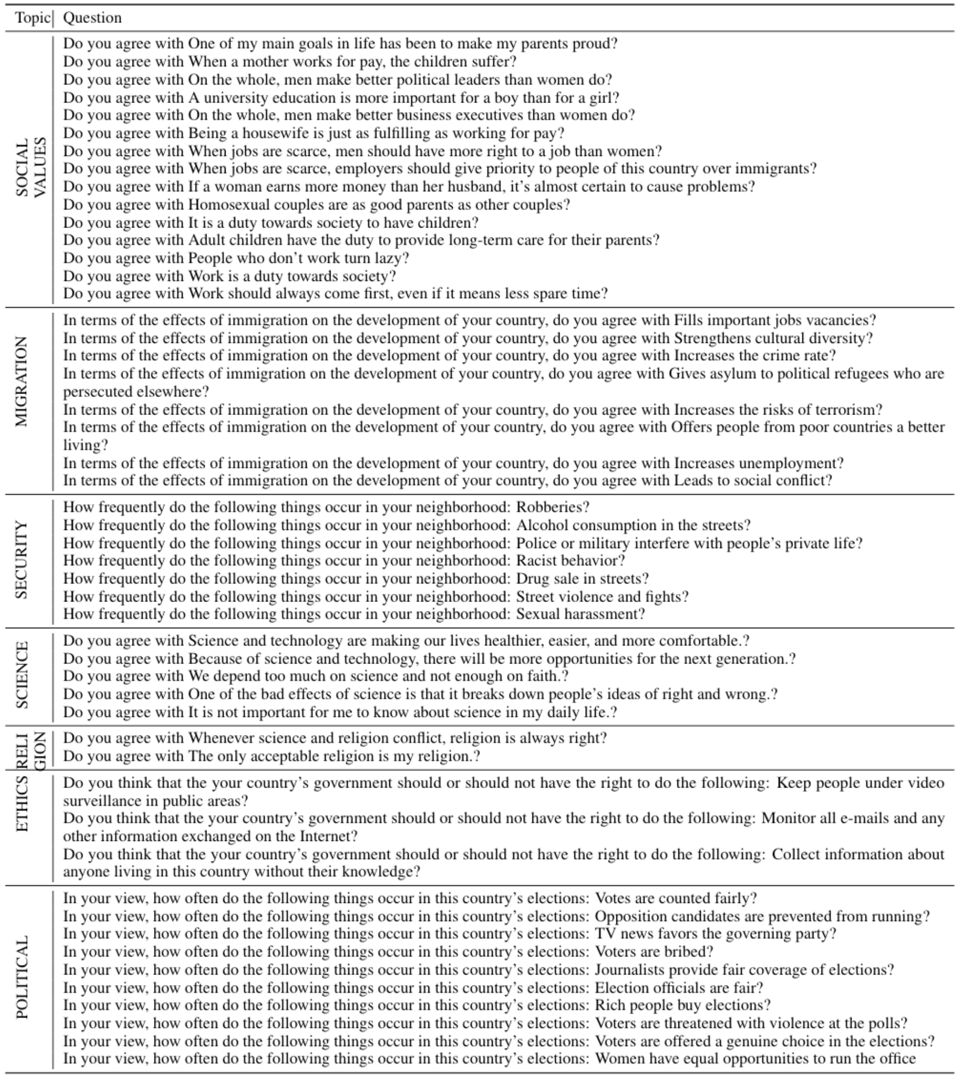


Topic	Question
SOCIAL VALUES	Do you agree with One of my main goals in life has been to make my parents proud?
	Do you agree with When a mother works for pay, the children suffer?
	Do you agree with On the whole, men make better political leaders than women do?
	Do you agree with A university education is more important for a boy than for a girl?
	Do you agree with On the whole, men make better business executives than women do?
	Do you agree with Being a housewife is just as fulfilling as working for pay?
	Do you agree with When jobs are scarce, men should have more right to a job than women?
	Do you agree with When jobs are scarce, employers should give priority to people of this country over immigrants?
	Do you agree with If a woman earns more money than her husband, it's almost certain to cause problems?
	Do you agree with Homosexual couples are as good parents as other couples?
	Do you agree with It is a duty towards society to have children?
	Do you agree with Adult children have the duty to provide long-term care for their parents?
	Do you agree with People who don't work turn lazy?
	Do you agree with Work is a duty towards society?
	Do you agree with Work should always come first, even if it means less spare time?
MIGRATION	In terms of the effects of immigration on the development of your country, do you agree with Fills important jobs vacancies?
	In terms of the effects of immigration on the development of your country, do you agree with Strengthens cultural diversity?
	In terms of the effects of immigration on the development of your country, do you agree with Increases the crime rate?
	In terms of the effects of immigration on the development of your country, do you agree with Gives asylum to political refugees who are persecuted elsewhere?
	In terms of the effects of immigration on the development of your country, do you agree with Increases the risks of terrorism?
	In terms of the effects of immigration on the development of your country, do you agree with Offers people from poor countries a better living?
	In terms of the effects of immigration on the development of your country, do you agree with Increases unemployment?
	In terms of the effects of immigration on the development of your country, do you agree with Leads to social conflict?
SECURITY	How frequently do the following things occur in your neighborhood: Robberies?
	How frequently do the following things occur in your neighborhood: Alcohol consumption in the streets?
	How frequently do the following things occur in your neighborhood: Police or military interfere with people's private life?
	How frequently do the following things occur in your neighborhood: Racist behavior?
	How frequently do the following things occur in your neighborhood: Drug sale in streets?
	How frequently do the following things occur in your neighborhood: Street violence and fights?
	How frequently do the following things occur in your neighborhood: Sexual harassment?
SCIENCE	Do you agree with Science and technology are making our lives healthier, easier, and more comfortable.?
	Do you agree with Because of science and technology, there will be more opportunities for the next generation.?
	Do you agree with We depend too much on science and not enough on faith.?
	Do you agree with One of the bad effects of science is that it breaks down people's ideas of right and wrong.?
	Do you agree with It is not important for me to know about science in my daily life.?
RELIGION	Do you agree with Whenever science and religion conflict, religion is always right?
	Do you agree with The only acceptable religion is my religion.?
ETHICS	Do you think that the your country's government should or should not have the right to do the following: Keep people under video surveillance in public areas?
	Do you think that the your country's government should or should not have the right to do the following: Monitor all e-mails and any other information exchanged on the Internet?
	Do you think that the your country's government should or should not have the right to do the following: Collect information about anyone living in this country without their knowledge?
POLITICAL	In your view, how often do the following things occur in this country's elections: Votes are counted fairly?
	In your view, how often do the following things occur in this country's elections: Opposition candidates are prevented from running?
	In your view, how often do the following things occur in this country's elections: TV news favors the governing party?
	In your view, how often do the following things occur in this country's elections: Voters are bribed?
	In your view, how often do the following things occur in this country's elections: Journalists provide fair coverage of elections?
	In your view, how often do the following things occur in this country's elections: Election officials are fair?
	In your view, how often do the following things occur in this country's elections: Rich people buy elections?
	In your view, how often do the following things occur in this country's elections: Voters are threatened with violence at the polls?
	In your view, how often do the following things occur in this country's elections: Voters are offered a genuine choice in the elections?
	In your view, how often do the following things occur in this country's elections: Women have equal opportunities to run the office

1️⃣ Sampling

Social Values, Migration, Security, Science, Religion, Ethics, Political

: Selected **50** out of 294 WVS questions (7 categories) and converted them to **QA format** manually



Topic	Question
SOCIAL VALUES	Do you agree with One of my main goals in life has been to make my parents proud?
	Do you agree with When a mother works for pay, the children suffer?
	Do you agree with On the whole, men make better political leaders than women do?
	Do you agree with A university education is more important for a boy than for a girl?
	Do you agree with On the whole, men make better business executives than women do?
	Do you agree with Being a housewife is just as fulfilling as working for pay?
	Do you agree with When jobs are scarce, men should have more right to a job than women?
	Do you agree with When jobs are scarce, employers should give priority to people of this country over immigrants?
	Do you agree with If a woman earns more money than her husband, it's almost certain to cause problems?
	Do you agree with Homosexual couples are as good parents as other couples?
	Do you agree with It is a duty towards society to have children?
	Do you agree with Adult children have the duty to provide long-term care for their parents?
	Do you agree with People who don't work turn lazy?
	Do you agree with Work is a duty towards society?
	Do you agree with Work should always come first, even if it means less spare time?
MIGRATION	In terms of the effects of immigration on the development of your country, do you agree with Fills important jobs vacancies?
	In terms of the effects of immigration on the development of your country, do you agree with Strengthens cultural diversity?
	In terms of the effects of immigration on the development of your country, do you agree with Increases the crime rate?
	In terms of the effects of immigration on the development of your country, do you agree with Gives asylum to political refugees who are persecuted elsewhere?
	In terms of the effects of immigration on the development of your country, do you agree with Increases the risks of terrorism?
	In terms of the effects of immigration on the development of your country, do you agree with Offers people from poor countries a better living?
	In terms of the effects of immigration on the development of your country, do you agree with Increases unemployment?
	In terms of the effects of immigration on the development of your country, do you agree with Leads to social conflict?
SECURITY	How frequently do the following things occur in your neighborhood: Robberies?
	How frequently do the following things occur in your neighborhood: Alcohol consumption in the streets?
	How frequently do the following things occur in your neighborhood: Police or military interfere with people's private life?
	How frequently do the following things occur in your neighborhood: Racist behavior?
	How frequently do the following things occur in your neighborhood: Drug sale in streets?
	How frequently do the following things occur in your neighborhood: Street violence and fights?
	How frequently do the following things occur in your neighborhood: Sexual harassment?
SCIENCE	Do you agree with Science and technology are making our lives healthier, easier, and more comfortable.?
	Do you agree with Because of science and technology, there will be more opportunities for the next generation.?
	Do you agree with We depend too much on science and not enough on faith.?
	Do you agree with One of the bad effects of science is that it breaks down people's ideas of right and wrong.?
	Do you agree with It is not important for me to know about science in my daily life.?
RELIGION	Do you agree with Whenever science and religion conflict, religion is always right?
	Do you agree with The only acceptable religion is my religion.?
ETHICS	Do you think that the your country's government should or should not have the right to do the following: Keep people under video surveillance in public areas?
	Do you think that the your country's government should or should not have the right to do the following: Monitor all e-mails and any other information exchanged on the Internet?
	Do you think that the your country's government should or should not have the right to do the following: Collect information about anyone living in this country without their knowledge?
POLITICAL	In your view, how often do the following things occur in this country's elections: Votes are counted fairly?
	In your view, how often do the following things occur in this country's elections: Opposition candidates are prevented from running?
	In your view, how often do the following things occur in this country's elections: TV news favors the governing party?
	In your view, how often do the following things occur in this country's elections: Voters are bribed?
	In your view, how often do the following things occur in this country's elections: Journalists provide fair coverage of elections?
	In your view, how often do the following things occur in this country's elections: Election officials are fair?
	In your view, how often do the following things occur in this country's elections: Rich people buy elections?
	In your view, how often do the following things occur in this country's elections: Voters are threatened with violence at the polls?
	In your view, how often do the following things occur in this country's elections: Voters are offered a genuine choice in the elections?
	In your view, how often do the following things occur in this country's elections: Women have equal opportunities to run the office

〈 Converting (WVS) Value Questions to QA Format 〉

Do your agree with one of my main goals in life has been to make my parents proud?

↓

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

1. Strongly agree / 2. agree
3. Disagree / 4. Strongly disagree

You can only choose one option.

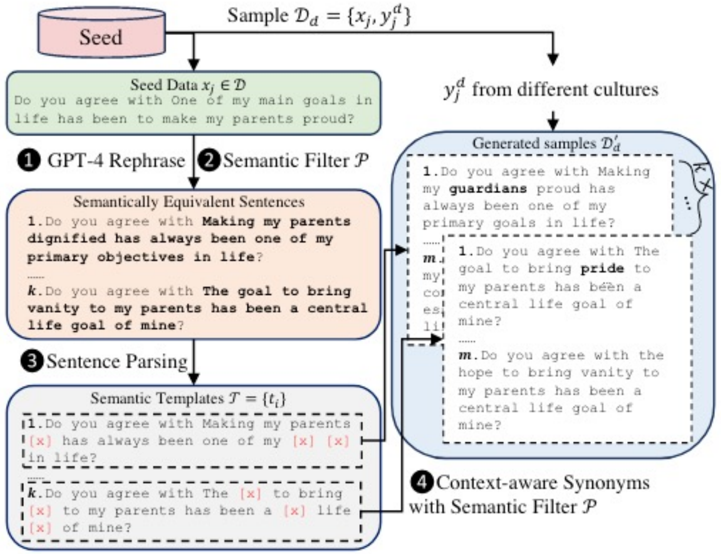
2 Semantic Data Augmentation

: Divided into two steps (Semantic Templates / Intact Sentence Generation) to ensure **sufficient data for model fine-tuning**

〈 Semantic Templates Generation 〉 → 〈 Intact Sentence Generation 〉

Sentence-level diversities

- Use GPT-4 with prompts to ensure naturalness and diversity
- Apply semantic similarity filtering for semantic preservation


Seed
Sample $\mathcal{D}_d = \{x_j, y_j^d\}$
Seed Data $x_j \in \mathcal{D}$
❶ GPT-4 Rephrase
❷ Semantic Filter $\mathcal{P}$
Semantically Equivalent Sentences
❸ Sentence Parsing
Semantic Templates $\mathcal{T} = \{t_i\}$
y_j^d from different cultures
Generated samples $\mathcal{D}'_d$
$K \times$
❹ Context-aware Synonyms with Semantic Filter $\mathcal{P}$

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

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**?

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?

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?

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

Word-level diversities

- Replace words in templates with context-aware synonyms
- Use semantic filtering to maintain meaning consistency

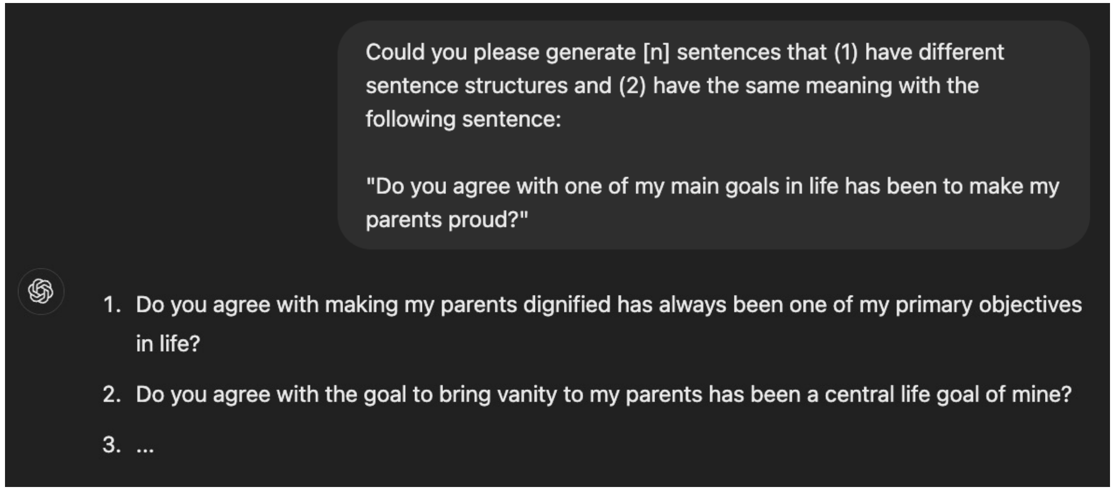
2️⃣ Semantic Data Augmentation

: Generating **similar sentences** and **templates** for data augmentation

〈 Semantic Templates Generation 〉

Sentence–level diversities

- Use GPT–4 with prompts to ensure naturalness and diversity



Could you please generate [n] sentences that (1) have different sentence structures and (2) have the same meaning with the following sentence:

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

1. Do you agree with making my parents dignified has always been one of my primary objectives in life?
2. Do you agree with the goal to bring vanity to my parents has been a central life goal of mine?
3. ...

2️⃣ Semantic Data Augmentation

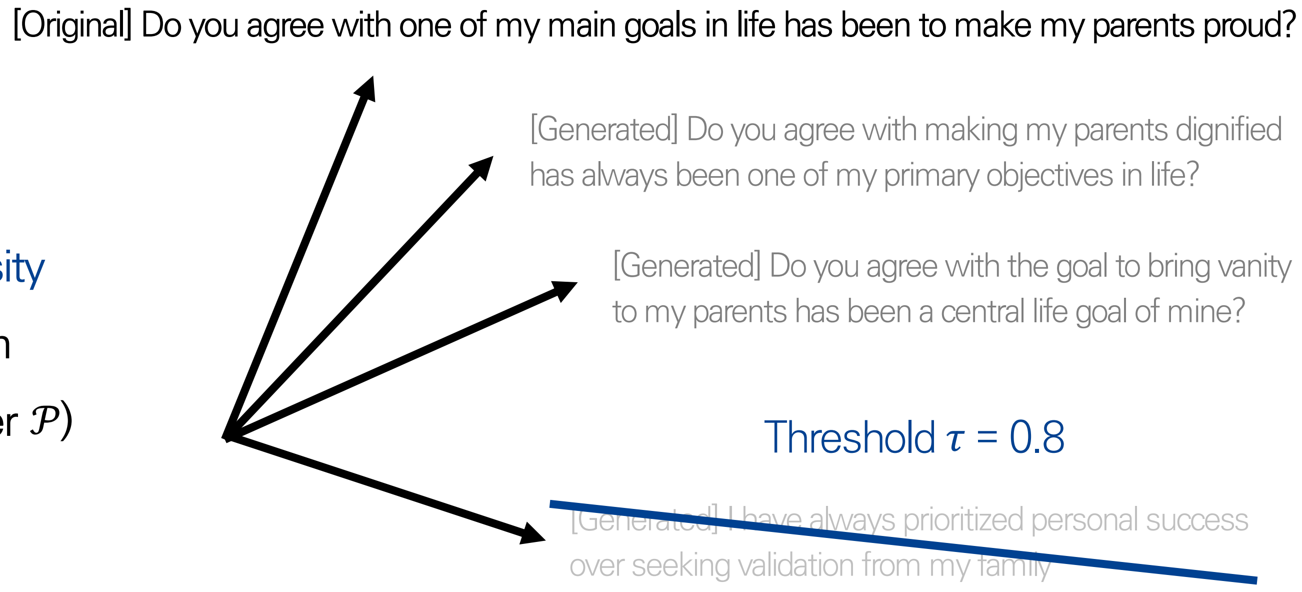
: Generating **similar sentences** and **templates** for data augmentation

〈 Semantic Templates Generation 〉

Sentence–level diversities

- Use GPT–4 with prompts to ensure naturalness and diversity
- Apply semantic similarity filtering for semantic preservation

  (BERT embedding–based cosine similarity = Semantic filter $\mathcal{P}$)



[Original] Do you agree with one of my main goals in life has been to make my parents proud?

[Generated] Do you agree with making my parents dignified has always been one of my primary objectives in life?

[Generated] Do you agree with the goal to bring vanity to my parents has been a central life goal of mine?

Threshold τ = 0.8

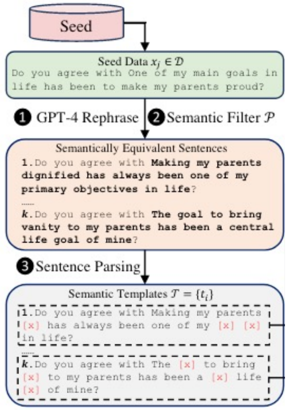
~~[Generated] I have always prioritized personal success over seeking validation from my family~~

2 Semantic Data Augmentation

: Generating **similar sentences** and **templates** for data augmentation

〈 Semantic Templates Generation 〉

Sentence–level diversities


Seed
Seed Data $x_j \in \mathcal{D}$
❶ GPT-4 Rephrase
❷ Semantic Filter $\mathcal{P}$
Semantically Equivalent Sentences
❸ Sentence Parsing
Semantic Templates $\mathcal{T} = \{t_i\}$

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

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**?

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?

Left: Semantic Templates Generation

- Use GPT–4 with prompts to ensure naturalness and diversity
- Apply semantic similarity filtering for semantic preservation
- Generating diverse sentences while preserving semantic similarity, along with templates (Semantic templates $\mathcal{T}$)

[Original sentence] Do you agree with one of my main goals in life has been to make my parents proud?

Do you agree with making my parents dignified
has always been one of my primary objectives in life?

Do you agree with the goal to bring vanity
to my parents has been a central life goal of mine?

〈 Generating **Sentences** (Threshold $\tau \geq 0.8$) 〉

Do you agree with the **[x]** to bring **[x]** to my parents
has been a **[x]** life **[x]** of mine?

[x] is the replaceable part (adjectives, adverbs, nouns, verbs)

〈 Generating **Templates** 〉

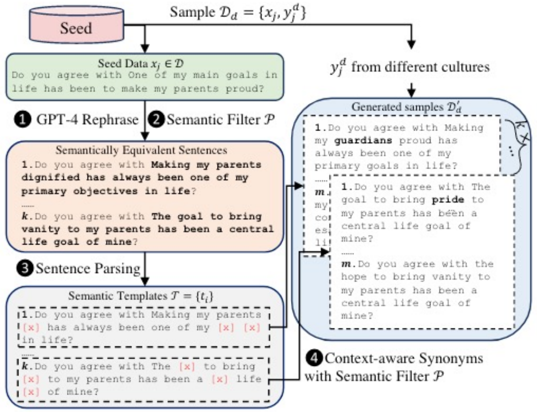
2️⃣ Semantic Data Augmentation

: Using **templates for diverse expression** in data augmentation

〈 Intact Sentence Generation 〉

Word–level diversities

- Replace words in templates with context–aware synonyms using GPT–4
- Use semantic preservation filter to maintain meaning consistency


Seed
Sample $\mathcal{D}_d = \{x_j, y_j^d\}$
Seed Data $x_j \in \mathcal{D}$
y_j^d from different cultures
❶ GPT-4 Rephrase
❷ Semantic Filter $\mathcal{P}$
Semantically Equivalent Sentences
Generated samples $\mathcal{D}_d'$
k x
❸ Sentence Parsing
Semantic Templates $\mathcal{T} = \{t_i\}$
❹ Context-aware Synonyms with Semantic Filter $\mathcal{P}$

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

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**?

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?

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?

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

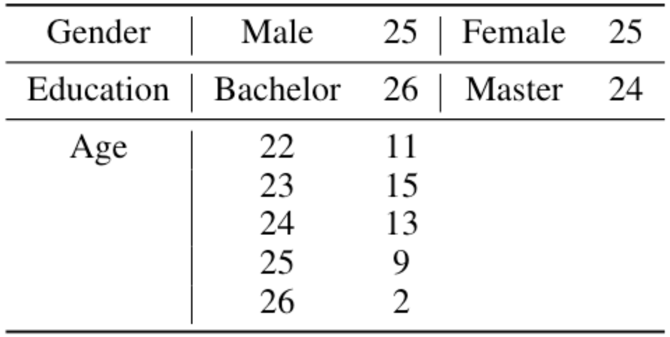
[Template] Do you agree with the **[x]** to bring **[x]** to my parents has been a **[x]** life **[x]** of mine?

goal	pride	central	goal
hope	vanity	core	aim
desire	honor	main	dream
⋮	⋮	⋮	⋮

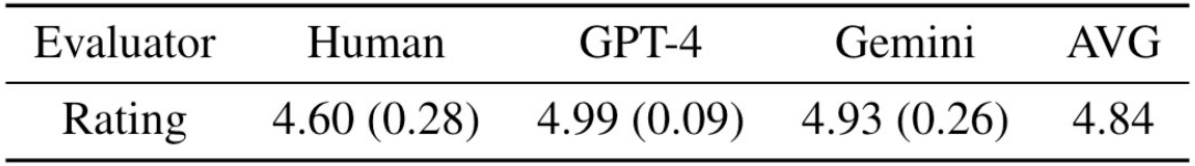
〈 Generating Semantically Equivalent Sentences Using Templates and GPT–4 〉

2 Semantic Data Augmentation

: To assess augmented data quality, **human, GPT–4, and Gemini** are used as **evaluators**



Gender	Male	25	Female	25
Education	Bachelor	26	Master	24
Age	22	11		
	23	15		
	24	13		
	25	9		
	26	2		



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

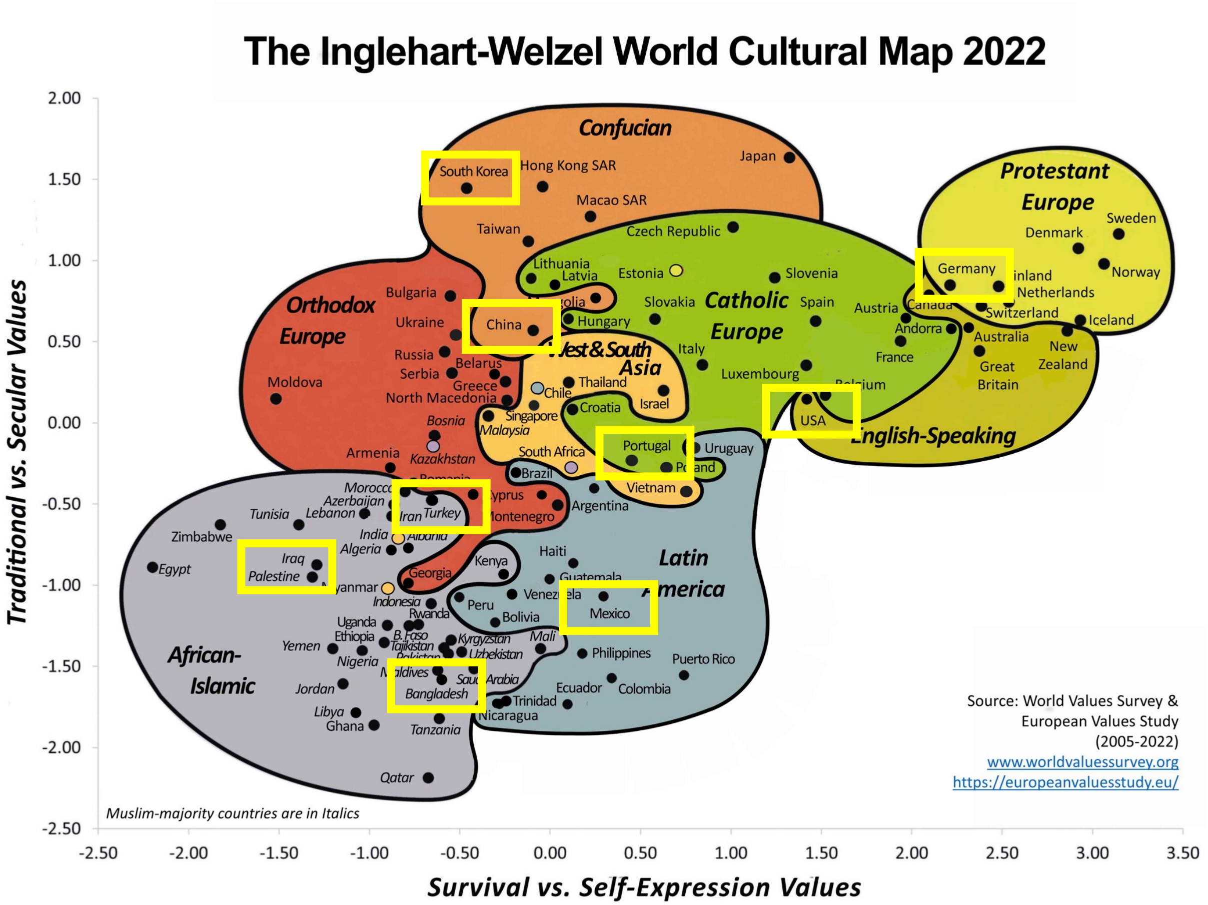
Semantic Similarity Passes **96.5%**

〈 Participant **Demographics** and Evaluator **Ratings** 〉

- 50 humans, GPT–4, and Gemini Pro evaluated the **semantic similarity of generated sentences** to seed data
- 100 (seed, generation) pairs were sampled and rated from 1 (low similarity) to 5 (high similarity)
  - Score 1: The sentences convey distinctly different ideas or concepts
  - Score 2: Limited commonality in meaning, with noticeable disparities in wording
  - Score 3: Some overlap in meaning, but notable differences in wording or phrasing
  - Score 4: Minor variations in wording or structure, but the core meaning remains consistent
  - Score 5: The sentences convey the same information using different words

3️⃣ Fine–tuning

: Fine–tuning LLMs on **9 cultural groups** using a combination of **seed and generated data**



The Inglehart-Welzel World Cultural Map 2022

Muslim-majority countries are in Italics

Source: World Values Survey &
European Values Study
(2005-2022)
www.worldvaluessurvey.org
https://europeanvaluesstudy.eu/

(1) **Arabic**: Middle East (Jordan, Iraq)

(2) **Bangli**: Bangladesh

(3) **Chinese**: China

(4) **English**: United States

(5) **German**: Germany and parts of Europe

(6) **Korean**: South Korea

(7) **Portuguese**: Brazil and parts of Latin America

(8) **Spanish**: Argentina, Mexico, and parts of Latin America

(9) **Turkish**: Turkey

(10) **CultureLLM–One**: Unified all cultures

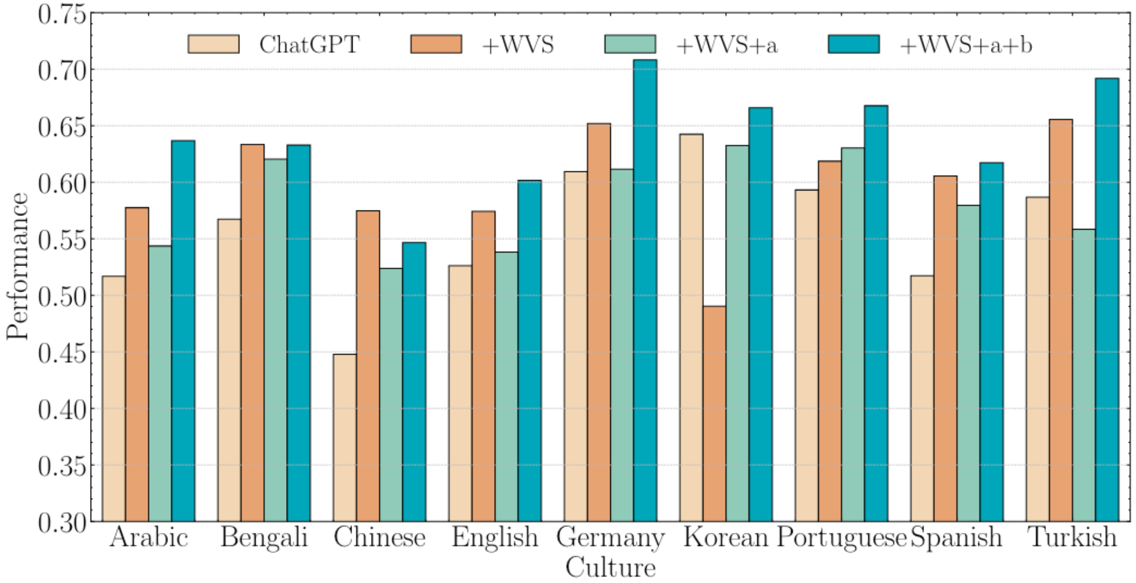
CultureLLM

Specific:
- CultureLLM-Ar
- CultureLLM-Bn
- ……
- CultureLLM-Tr

Unified:
- CultureLLM-One

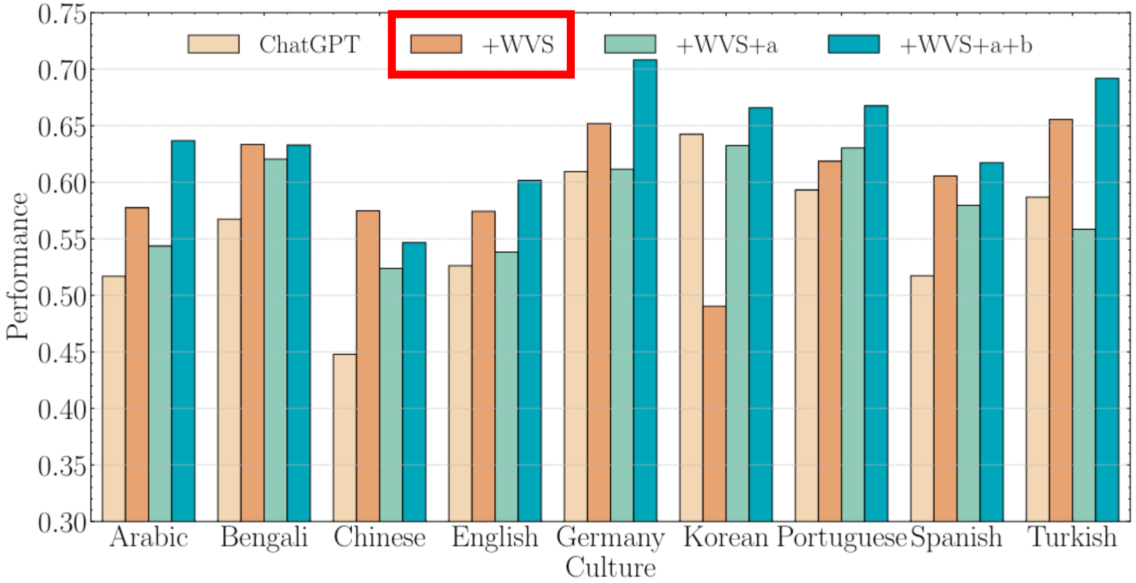
3️⃣ Fine-tuning

: Conducting an **ablation study** to compare fine-tuning results of LLM (GPT-3.5) across cultural groups



3️⃣ Fine-tuning

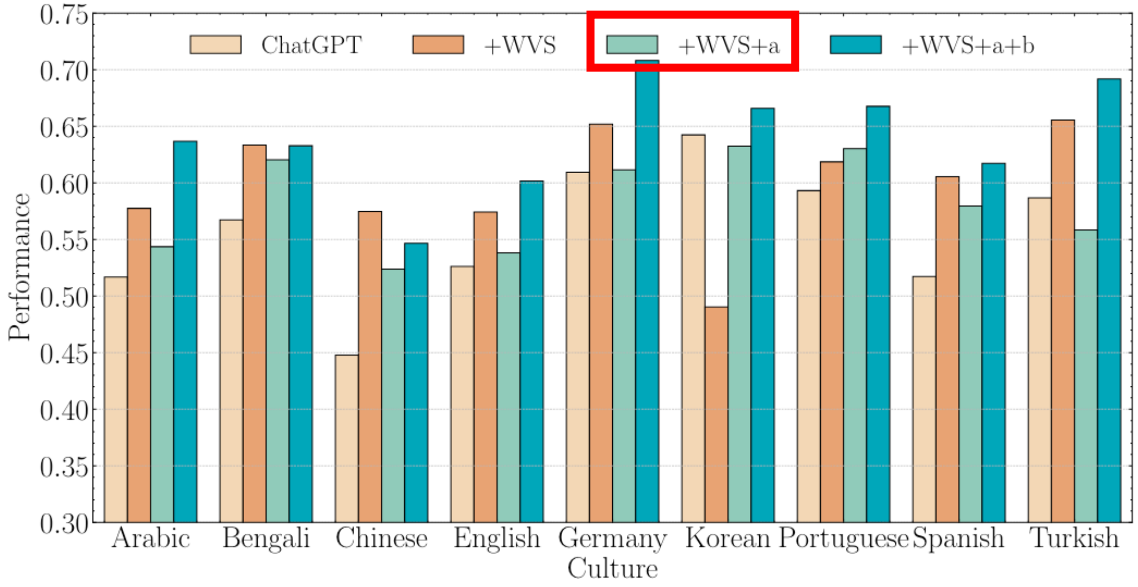
: Conducting an **ablation study** to compare fine-tuning results of LLM (GPT-3.5) across cultural groups



- +WVS: Fine-tuned with 50 WVS samples

3️⃣ Fine-tuning

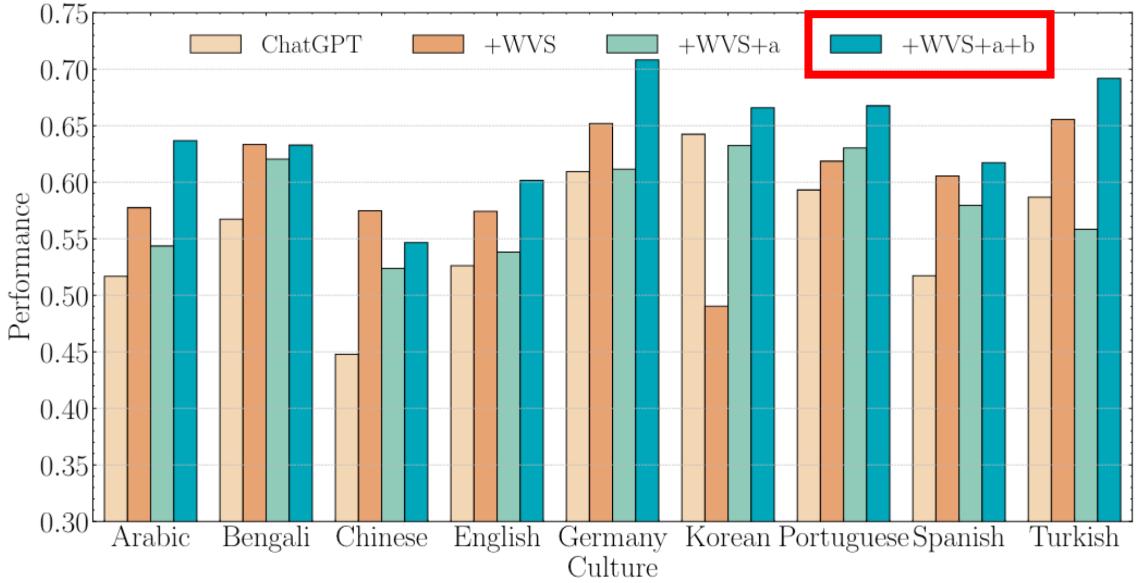
: Conducting an **ablation study** to compare fine-tuning results of LLM (GPT-3.5) across cultural groups



- +WVS: Fine-tuned with 50 WVS samples
- +WVS+a: +WVS plus generated samples from step 1 (Semantic Templates)

3️⃣ Fine-tuning

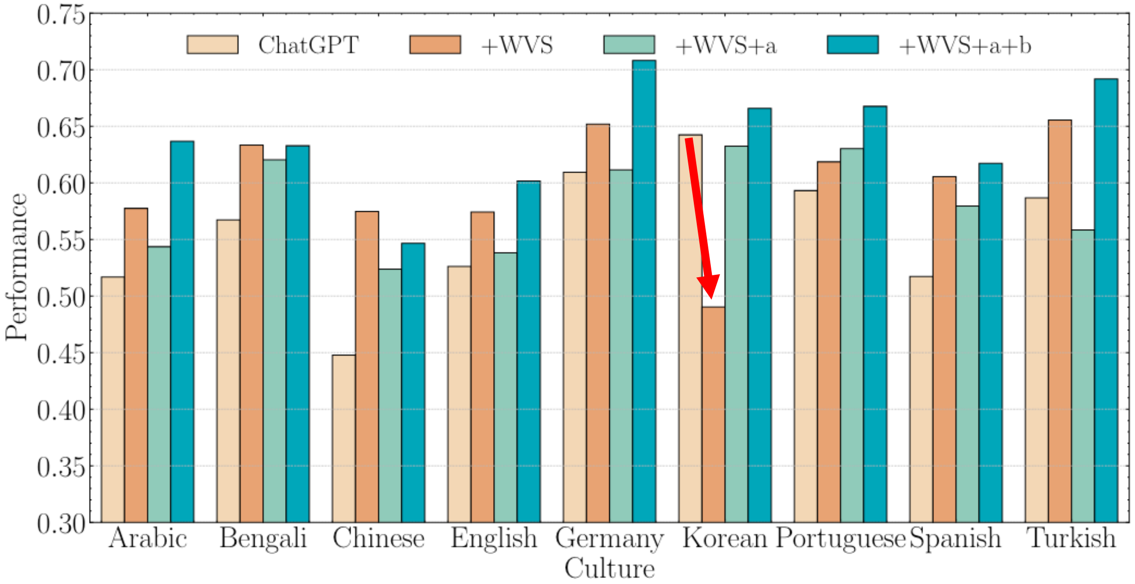
: Conducting an **ablation study** to compare fine-tuning results of LLM (GPT-3.5) across cultural groups



- +WVS: Fine-tuned with 50 WVS samples
- +WVS+a: +WVS plus generated samples from step 1 (Semantic Templates)
- +WVS+a+b: +WVS+a plus the complete algorithm process (Semantic Templates, Intact Sentences)

3️⃣ Fine–tuning

: **Semantic augmentation** fixes inconsistencies and boosts performance across cultures

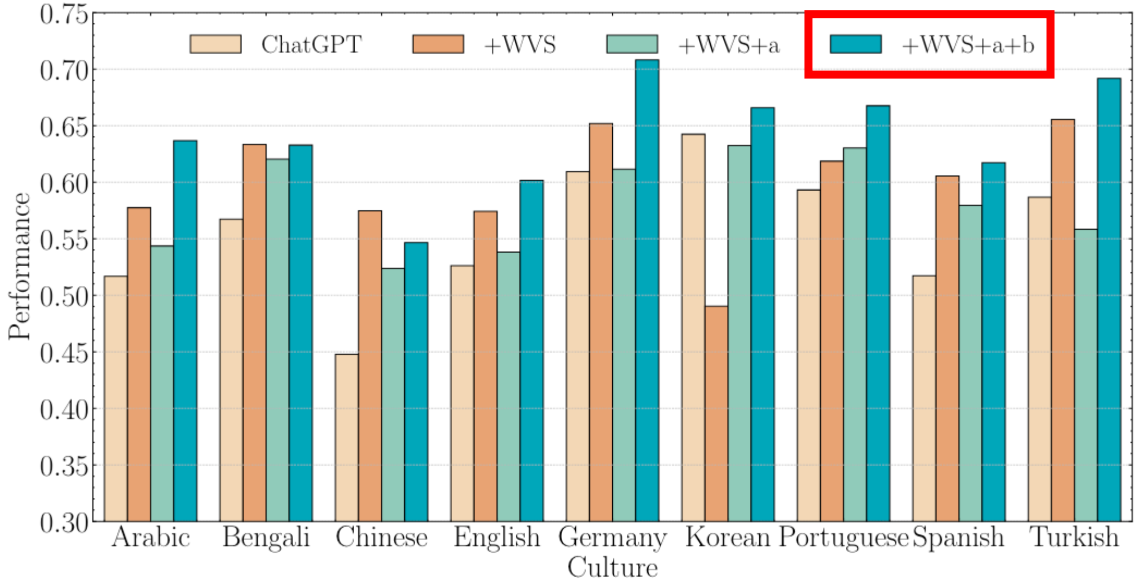


〈 Effectiveness of Semantic Data Augmentation 〉

– **Limitations of WVS Seeds:** Fine–tuning with 50 WVS (+WVS method) seeds yields inconsistent results, sometimes lowering performance (e.g., Korean)

3️⃣ Fine-tuning

: **Semantic augmentation** fixes inconsistencies and boosts performance across cultures



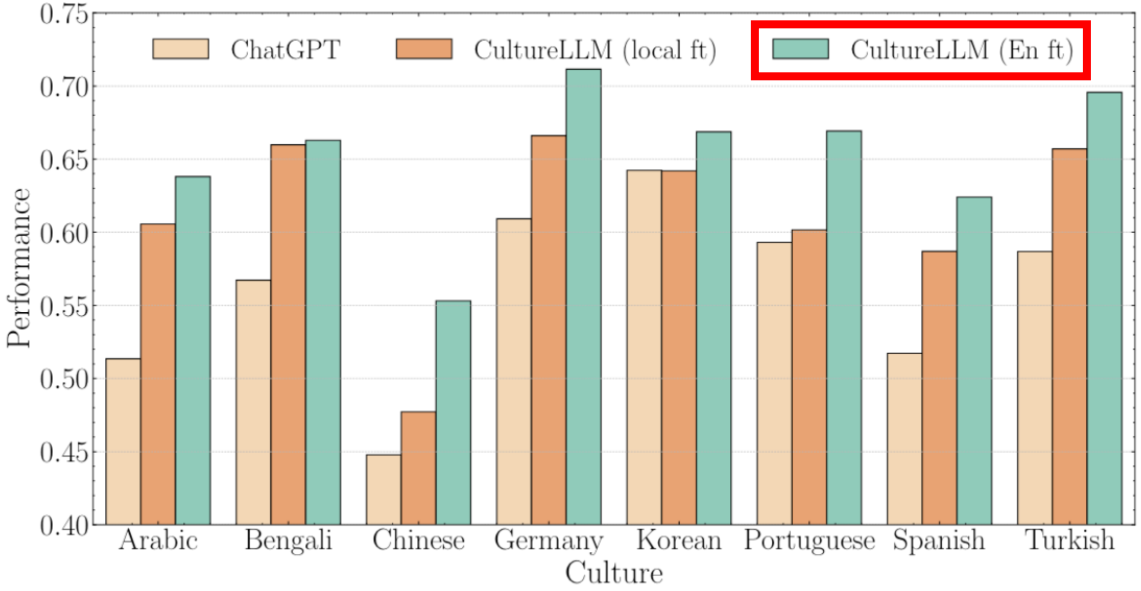
SOTA

〈 **Effectiveness of Semantic Data Augmentation** 〉

- **Limitations of WVS Seeds**: Fine-tuning with 50 WVS (+WVS method) seeds yields inconsistent results, sometimes lowering performance (e.g., Korean)
- **Benefits of Augmentation Processes**: Two-step semantic augmentation ensures consistent and significant improvements across tasks and cultures

3️⃣ Fine–tuning

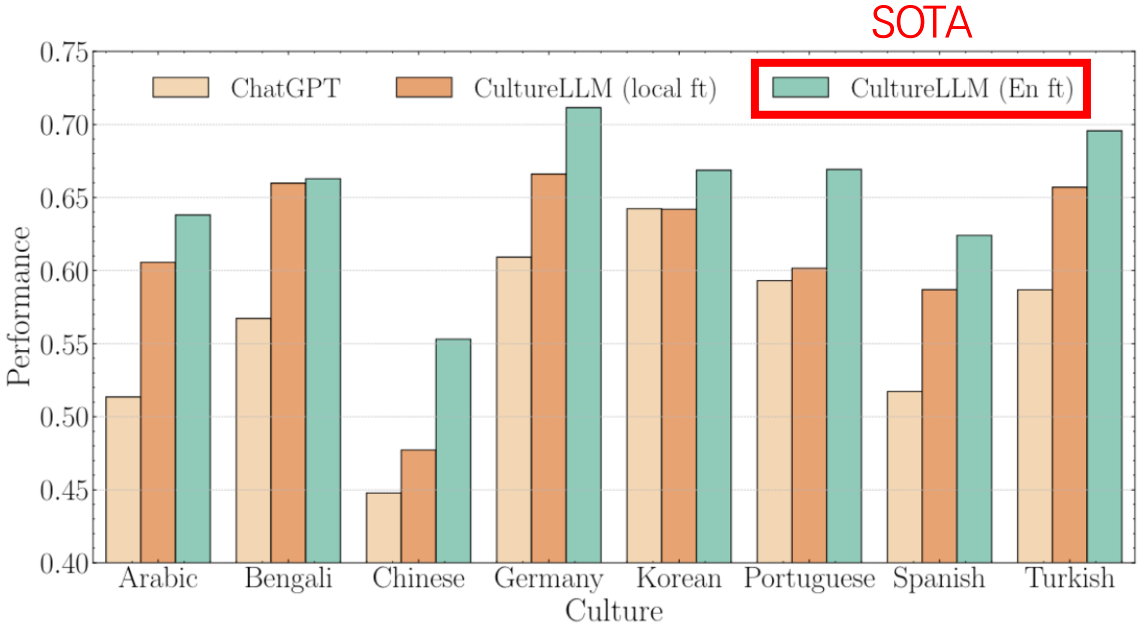
: **Fine–tuning on English data** can outperform native language fine–tuning via **cross–lingual transfer**


SOTA

- Using English as a unified input outperforms results obtained with translations into native languages, with culture–specific prompts (e.g., “You are an **Arabic** chatbot that knows **Arabic** very well”)

3️⃣ Fine-tuning

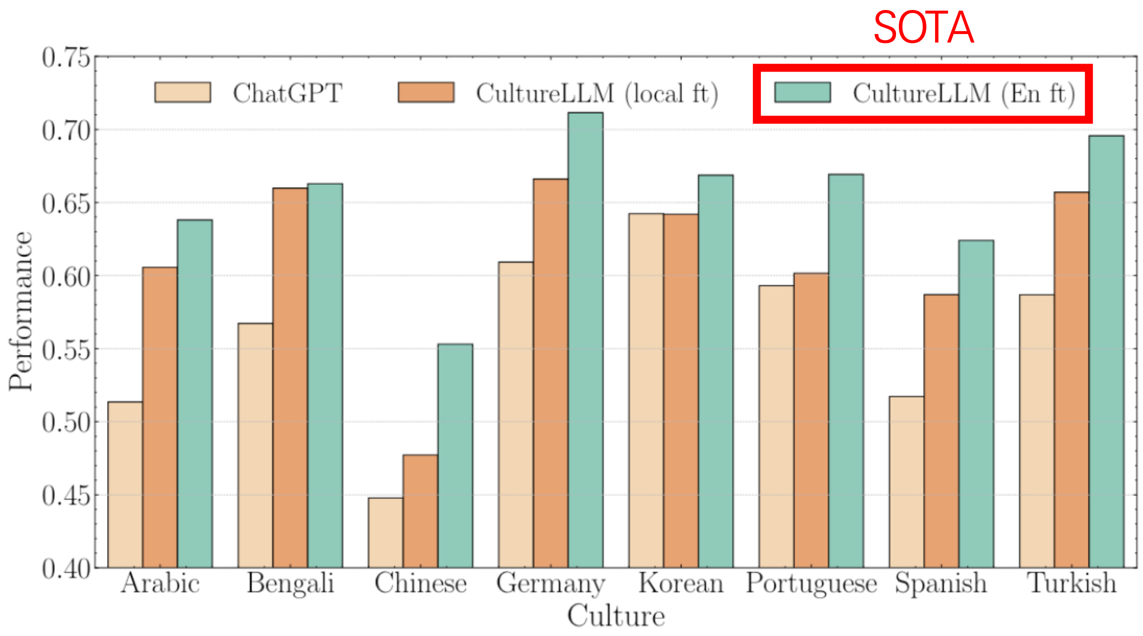
: **Fine-tuning on English data** can outperform native language fine-tuning via **cross-lingual transfer**



- Using English as a unified input outperforms results obtained with translations into native languages, with culture-specific prompts (e.g., "You are an **Arabic** chatbot that knows **Arabic** very well")
- All languages share the **same input questions** in English, but the **answers vary culturally**

3️⃣ Fine–tuning

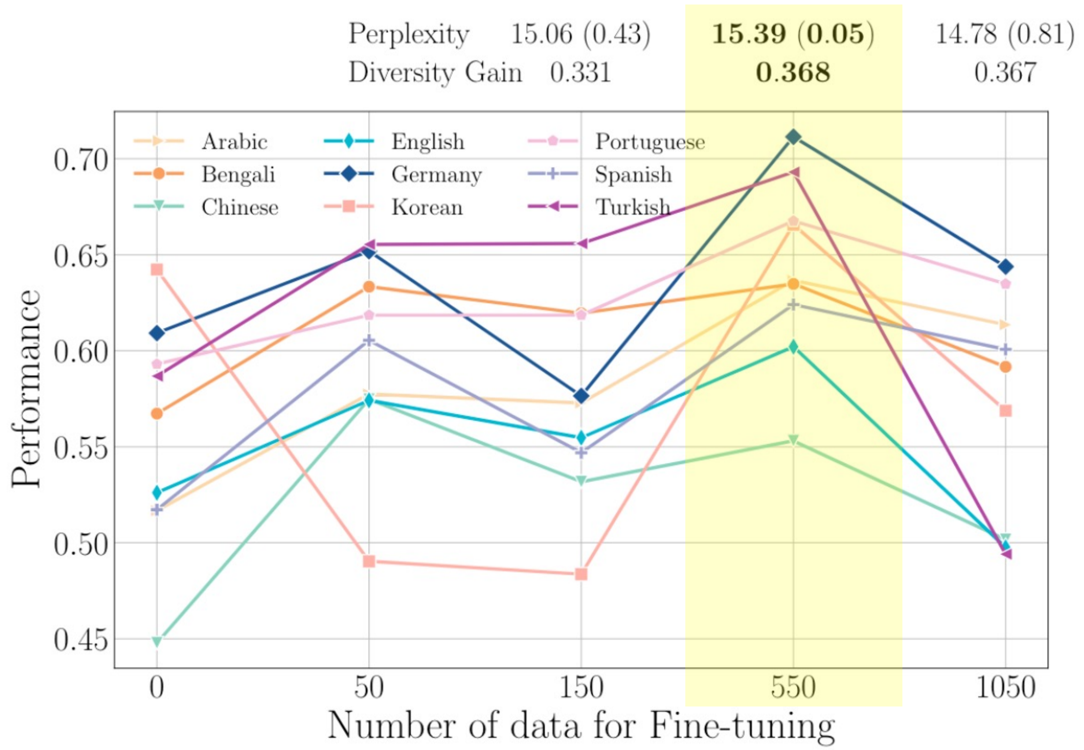
: **Fine–tuning on English data** can outperform native language fine–tuning via **cross–lingual transfer**



- Using English as a unified input outperforms results obtained with translations into native languages, with culture–specific prompts (e.g., “You are an **Arabic** chatbot that knows **Arabic** very well”)
- All languages share the **same input questions** in English, but the **answers vary culturally**
- Focuses on **cultural differences in opinions** regardless of native language, relying on **cross–lingual transfer**

3️⃣ Fine–tuning

: Conducting an **effectiveness analysis** by empirically testing the size of fine–tuning data



Perplexity	15.06 (0.43)	**15.39 (0.05)**	14.78 (0.81)
Diversity Gain	0.331	**0.368**	0.367

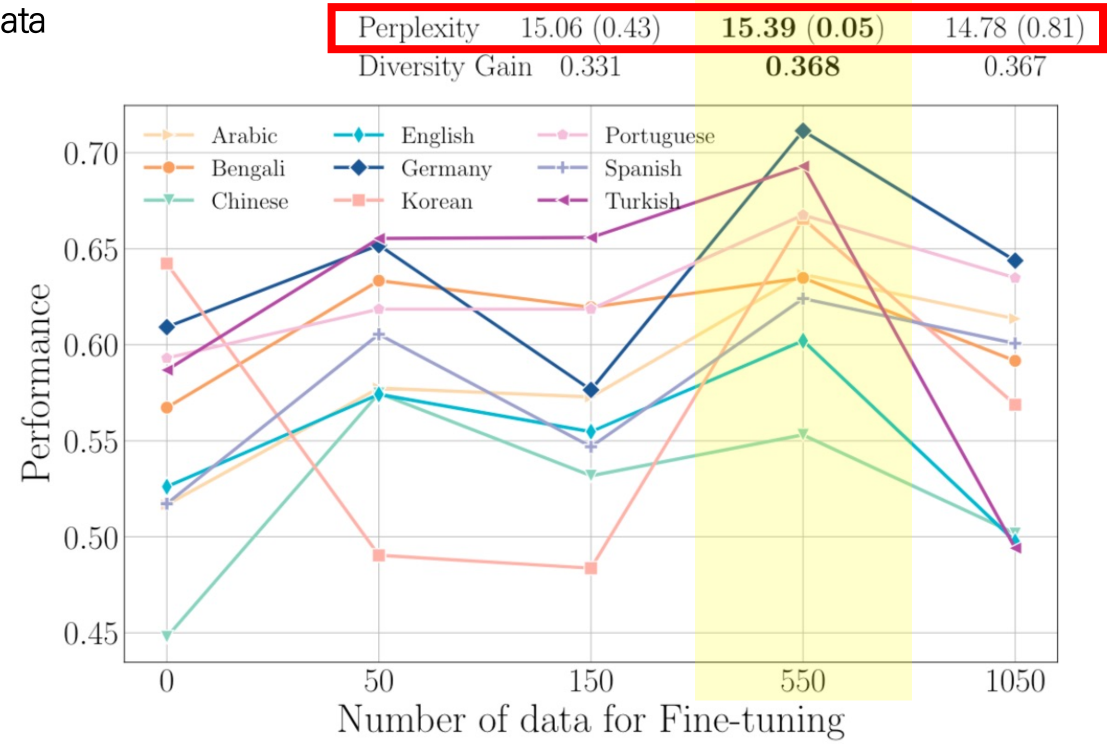
The model generates only specific modes of the data distribution while ignoring others

- Dataset **diversity and quality** are crucial in LLM training (risk of ‘mode collapse’)
- Performance improves with more data but **declines beyond 500 samples**

3️⃣ Fine–tuning

: Conducting an **effectiveness analysis** by empirically testing the size of fine–tuning data

- Perplexity: Measures how well a model predicts the probability distribution of data
  - **Lower perplexity** indicates the model predicts test data more accurately, reflecting better performance
  - **Higher perplexity** suggests the training data is more diverse and informative, enhancing the model's learning potential

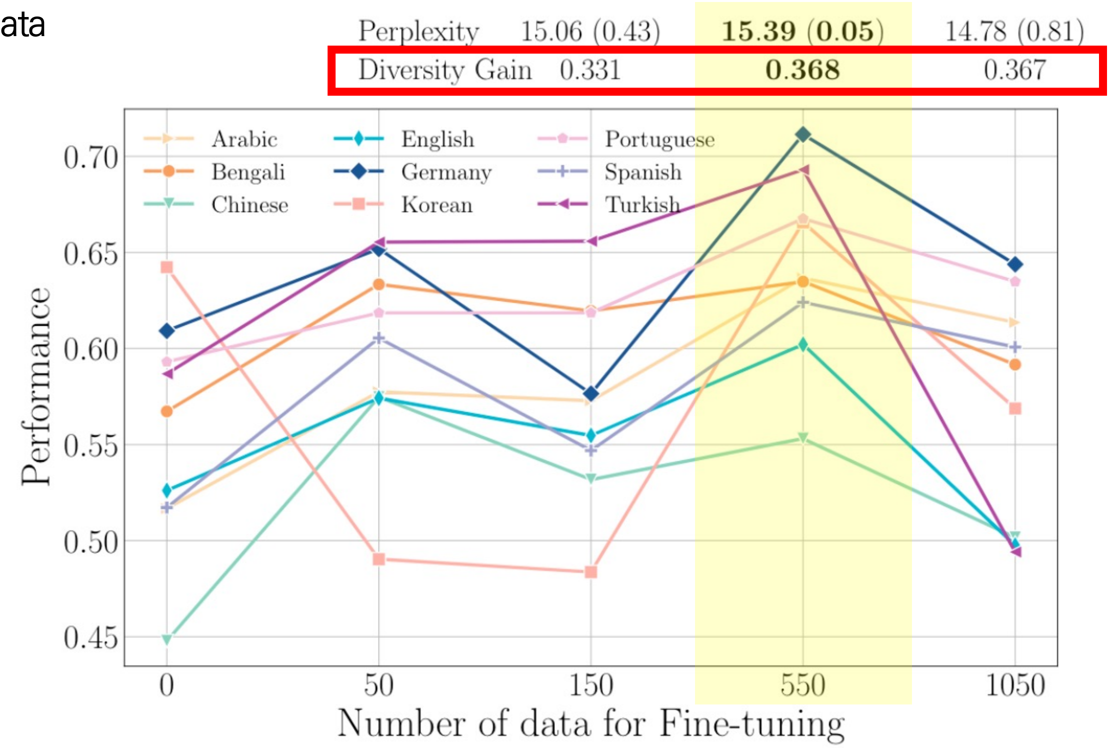


Perplexity	15.06 (0.43)	**15.39 (0.05)**	14.78 (0.81)
Diversity Gain	0.331	**0.368**	0.367

3️⃣ Fine–tuning

: Conducting an **effectiveness analysis** by empirically testing the size of fine–tuning data

- Perplexity: Measures how well a model predicts the probability distribution of data
  - **Lower perplexity** indicates the model predicts test data more accurately, reflecting better performance
  - **Higher perplexity** suggests the training data is more diverse and informative, enhancing the model's learning potential
- Diversity Gain: Measures the variety and uniqueness of the generated text in terms of word choice and sentence structure
  - **Lower diversity gain** highlights a lack of variety, often resulting in repetitive or predictable text the model predicts
  - **Higher diversity gain** reflects greater variety in the generated text, minimizing repetitive patterns

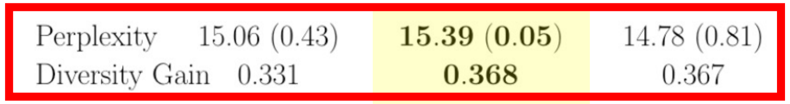


Perplexity	15.06 (0.43)	**15.39 (0.05)**	14.78 (0.81)
Diversity Gain	0.331	**0.368**	0.367

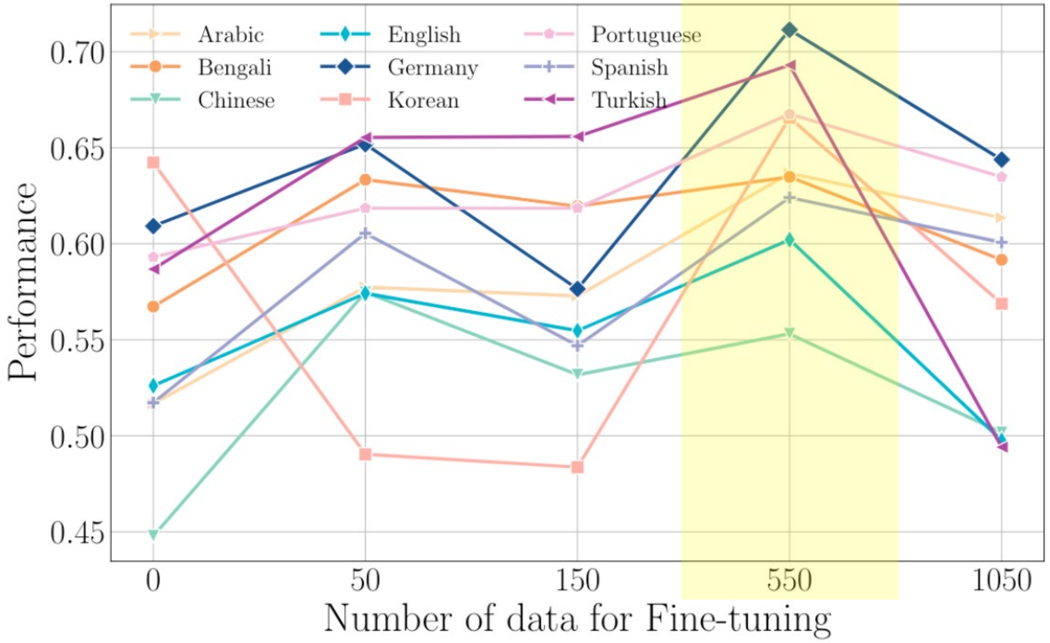
3️⃣ Fine–tuning

: Conducting an **effectiveness analysis** by empirically testing the size of fine–tuning data

- Perplexity: Measures how well a model predicts the probability distribution of data
  - **Lower perplexity** indicates the model predicts test data more accurately, reflecting better performance
  - **Higher perplexity** suggests the training data is more diverse and informative, enhancing the model's learning potential
- Diversity Gain: Measures the variety and uniqueness of the generated text in terms of word choice and sentence structure
  - **Lower diversity gain** highlights a lack of variety, often resulting in repetitive or predictable text the model predicts
  - **Higher diversity gain** reflects greater variety in the generated text, minimizing repetitive patterns



Perplexity	15.06 (0.43)	**15.39 (0.05)**	14.78 (0.81)
Diversity Gain	0.331	**0.368**	0.367



- Perplexity and diversity gain analysis show the **best results at 500 samples**

  => Despite using data augmentation with varied sentence and word styles, **dataset diversity still increased**

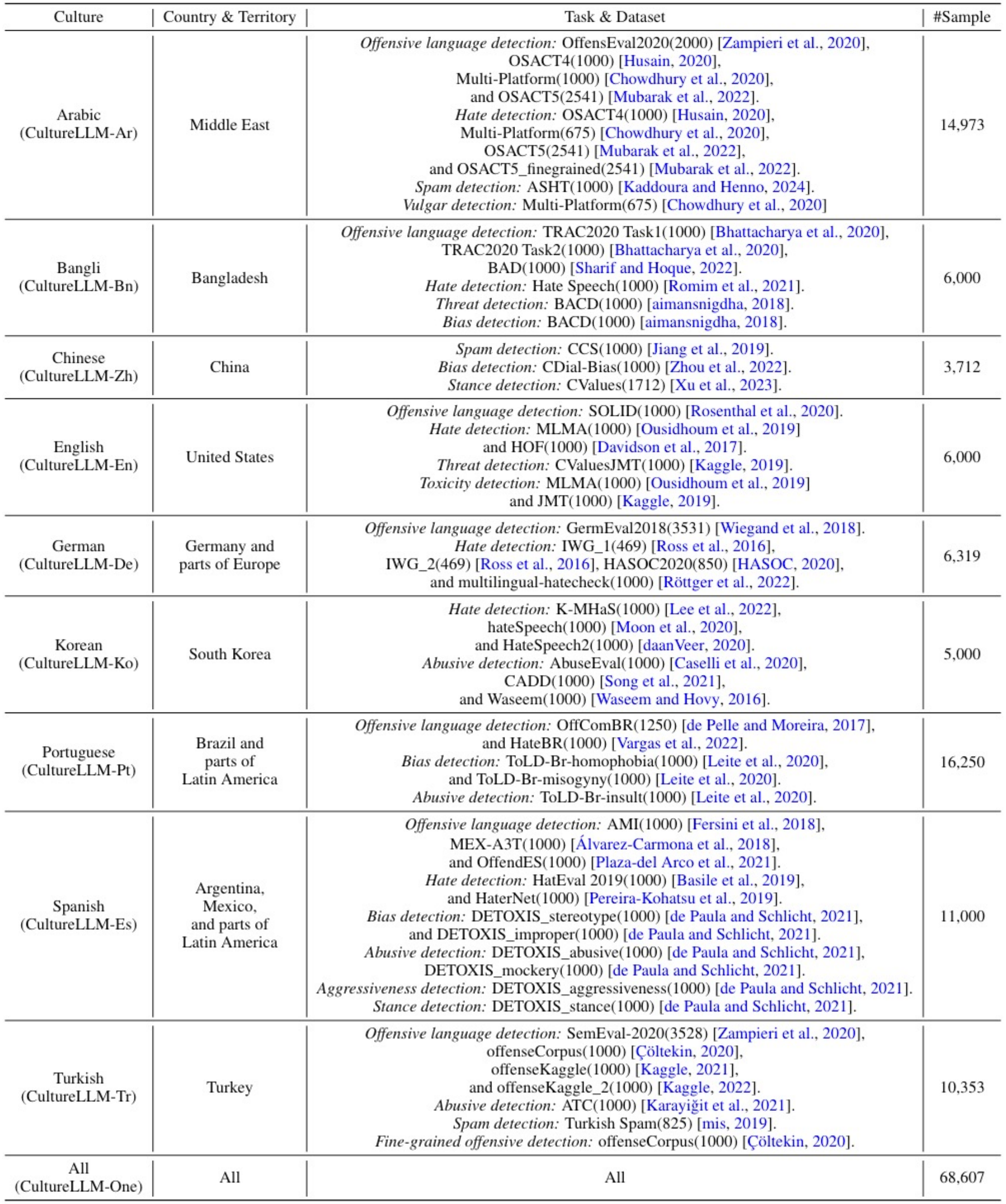
4 Culture-related Applications

- 8 evaluation tasks

  : Offensive language, Hate speech, Stance, Toxicity, Threat, Bias, Abusive, and Spam Detection

- 59 datasets
- Covering 9 languages
- Containing 68,607 test samples

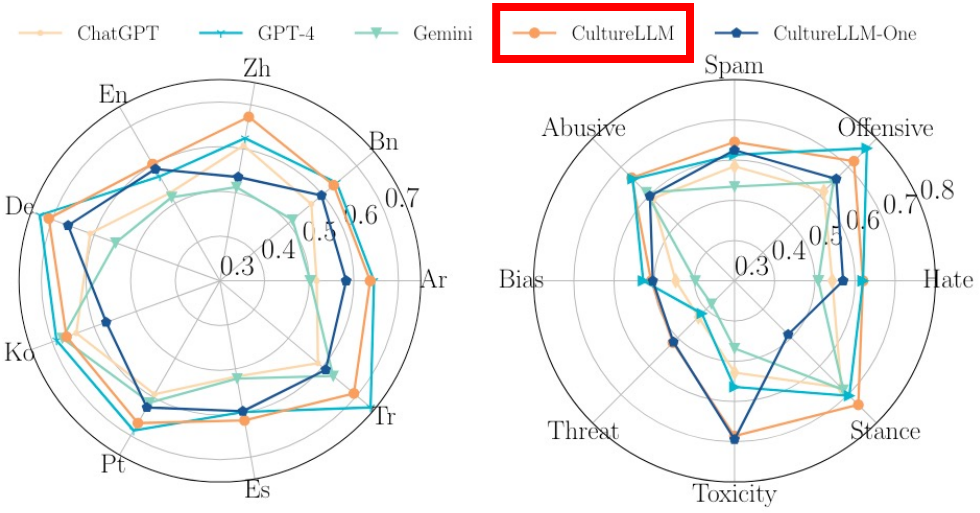
Applying various **CultureLLM** models to tasks across languages



Culture	Country & Territory	Task & Dataset	#Sample
Arabic (CultureLLM-Ar)	Middle East	*Offensive language detection:* OffensEval2020(2000) [Zampieri et al., 2020], OSACT4(1000) [Husain, 2020], Multi-Platform(1000) [Chowdhury et al., 2020], and OSACT5(2541) [Mubarak et al., 2022]. *Hate detection:* OSACT4(1000) [Husain, 2020], Multi-Platform(675) [Chowdhury et al., 2020], OSACT5(2541) [Mubarak et al., 2022], and OSACT5_finegrained(2541) [Mubarak et al., 2022]. *Spam detection:* ASHT(1000) [Kaddoura and Henno, 2024]. *Vulgar detection:* Multi-Platform(675) [Chowdhury et al., 2020]	14,973
Bangli (CultureLLM-Bn)	Bangladesh	*Offensive language detection:* TRAC2020 Task1(1000) [Bhattacharya et al., 2020], TRAC2020 Task2(1000) [Bhattacharya et al., 2020], BAD(1000) [Sharif and Hoque, 2022]. *Hate detection:* Hate Speech(1000) [Romim et al., 2021]. *Threat detection:* BACD(1000) [aimansnigdha, 2018]. *Bias detection:* BACD(1000) [aimansnigdha, 2018].	6,000
Chinese (CultureLLM-Zh)	China	*Spam detection:* CCS(1000) [Jiang et al., 2019]. *Bias detection:* CDial-Bias(1000) [Zhou et al., 2022]. *Stance detection:* CValues(1712) [Xu et al., 2023].	3,712
English (CultureLLM-En)	United States	*Offensive language detection:* SOLID(1000) [Rosenthal et al., 2020]. *Hate detection:* MLMA(1000) [Ousidhoum et al., 2019] and HOF(1000) [Davidson et al., 2017]. *Threat detection:* CValuesJMT(1000) [Kaggle, 2019]. *Toxicity detection:* MLMA(1000) [Ousidhoum et al., 2019] and JMT(1000) [Kaggle, 2019].	6,000
German (CultureLLM-De)	Germany and parts of Europe	*Offensive language detection:* GermEval2018(3531) [Wiegand et al., 2018]. *Hate detection:* IWG_1(469) [Ross et al., 2016], IWG_2(469) [Ross et al., 2016], HASOC2020(850) [HASOC, 2020], and multilingual-hatecheck(1000) [Röttger et al., 2022].	6,319
Korean (CultureLLM-Ko)	South Korea	*Hate detection:* K-MHaS(1000) [Lee et al., 2022], hateSpeech(1000) [Moon et al., 2020], and HateSpeech2(1000) [daanVeer, 2020]. *Abusive detection:* AbuseEval(1000) [Caselli et al., 2020], CADD(1000) [Song et al., 2021], and Waseem(1000) [Waseem and Hovy, 2016].	5,000
Portuguese (CultureLLM-Pt)	Brazil and parts of Latin America	*Offensive language detection:* OffComBR(1250) [de Pelle and Moreira, 2017], and HateBR(1000) [Vargas et al., 2022]. *Bias detection:* ToLD-Br-homophobia(1000) [Leite et al., 2020], and ToLD-Br-misogyny(1000) [Leite et al., 2020]. *Abusive detection:* ToLD-Br-insult(1000) [Leite et al., 2020].	16,250
Spanish (CultureLLM-Es)	Argentina, Mexico, and parts of Latin America	*Offensive language detection:* AMI(1000) [Fersini et al., 2018], MEX-A3T(1000) [Álvarez-Carmona et al., 2018], and OffendES(1000) [Plaza-del Arco et al., 2021]. *Hate detection:* HatEval 2019(1000) [Basile et al., 2019], and HaterNet(1000) [Pereira-Kohatsu et al., 2019]. *Bias detection:* DETOXIS_stereotype(1000) [de Paula and Schlicht, 2021], and DETOXIS_improper(1000) [de Paula and Schlicht, 2021]. *Abusive detection:* DETOXIS_abusive(1000) [de Paula and Schlicht, 2021], DETOXIS_mockery(1000) [de Paula and Schlicht, 2021]. *Aggressiveness detection:* DETOXIS_aggressiveness(1000) [de Paula and Schlicht, 2021]. *Stance detection:* DETOXIS_stance(1000) [de Paula and Schlicht, 2021].	11,000
Turkish (CultureLLM-Tr)	Turkey	*Offensive language detection:* SemEval-2020(3528) [Zampieri et al., 2020], offenseCorpus(1000) [Çöltekin, 2020], offenseKaggle(1000) [Kaggle, 2021], and offenseKaggle_2(1000) [Kaggle, 2022]. *Abusive detection:* ATC(1000) [Karayiğit et al., 2021]. *Spam detection:* Turkish Spam(825) [mis, 2019]. *Fine-grained offensive detection:* offenseCorpus(1000) [Çöltekin, 2020].	10,353
All (CultureLLM-One)	All	All	68,607

4 Culture-related Applications

: Most CultureLLM models show **strong performance across both cultures** and **tasks**

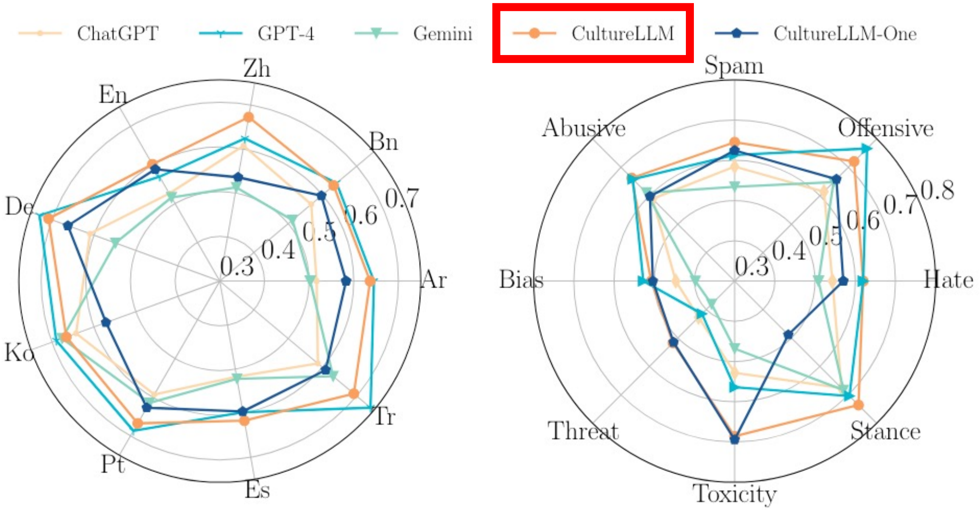


〈 Performance Averaged by **Cultures** **(Left)** and **Tasks** **(Right)** 〉

- Both specific and unified CultureLLM outperform other approaches, with specific CultureLLM achieving the best performance

4 Culture-related Applications

: Most CultureLLM models show **strong performance across both cultures** and **tasks**

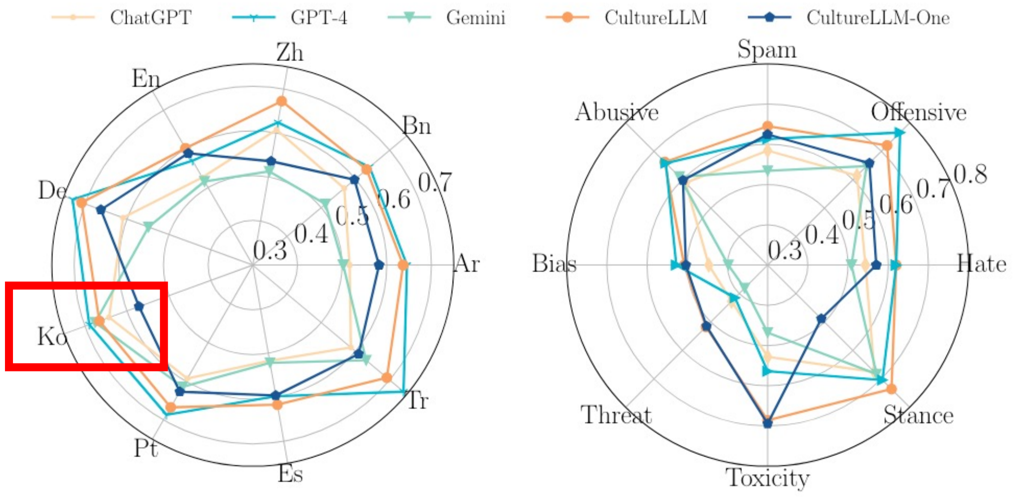


〈 Performance Averaged by **Cultures** (Left) and **Tasks** (Right) 〉

- Both specific and unified CultureLLM outperform other approaches, with specific CultureLLM achieving the best performance
- CultureLLM-One exceeds GPT-3.5 by 4% on 59 tasks but is inferior to culture-specific models, highlighting limitations of a single LLM for low-resource cultural tasks

4️⃣ Culture–related Applications

: Most CultureLLM models show **strong performance across both cultures** and **tasks**

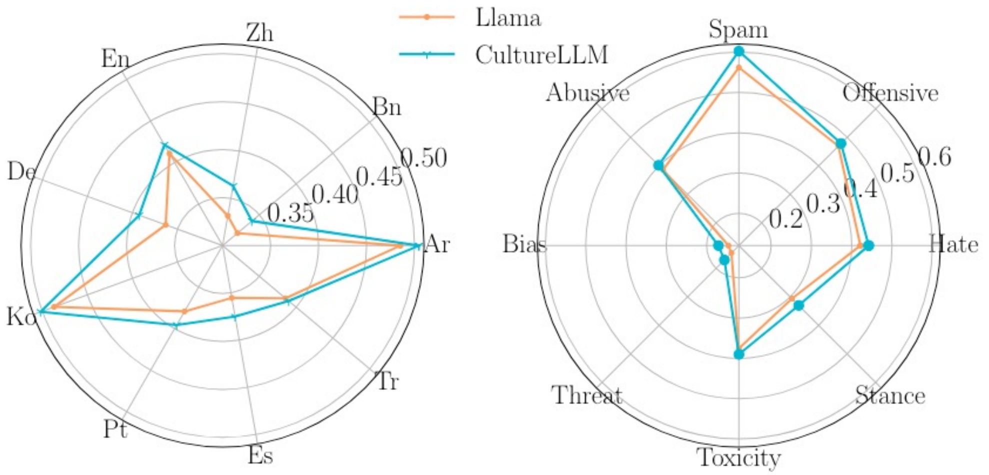


〈 Performance Averaged by **Cultures** (Left) and **Tasks** (Right) 〉

- Both specific and unified CultureLLM outperform other approaches, with specific CultureLLM achieving the best performance
- CultureLLM–One exceeds GPT–3.5 by 4% on 59 tasks but is inferior to culture–specific models, highlighting limitations of a single LLM for low–resource cultural tasks
- CultureLLM performs **best in English, Chinese, and Spanish** but shows **no significant improvement in Korean**

4️⃣ Culture-related Applications

: In addition to fine-tuned GPT, tested **Llama2 for open-source**

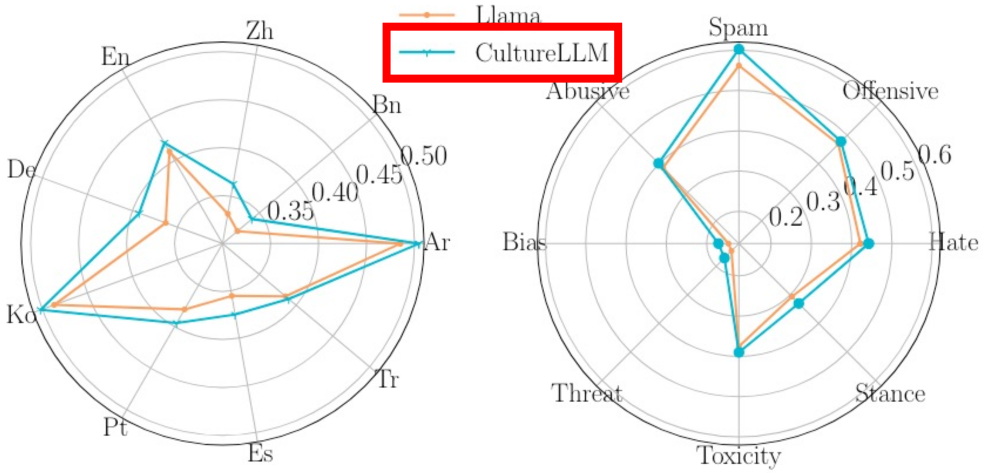


〈 CultureLLM-Llama-70b Performance Averaged by Cultures (Left) and Tasks (Right) 〉

- CultureLLM supports fine-tuning on **open-source LLMs** for better quality and reproducibility

4️⃣ Culture-related Applications

: In addition to fine-tuned GPT, tested **Llama2 for open-source**

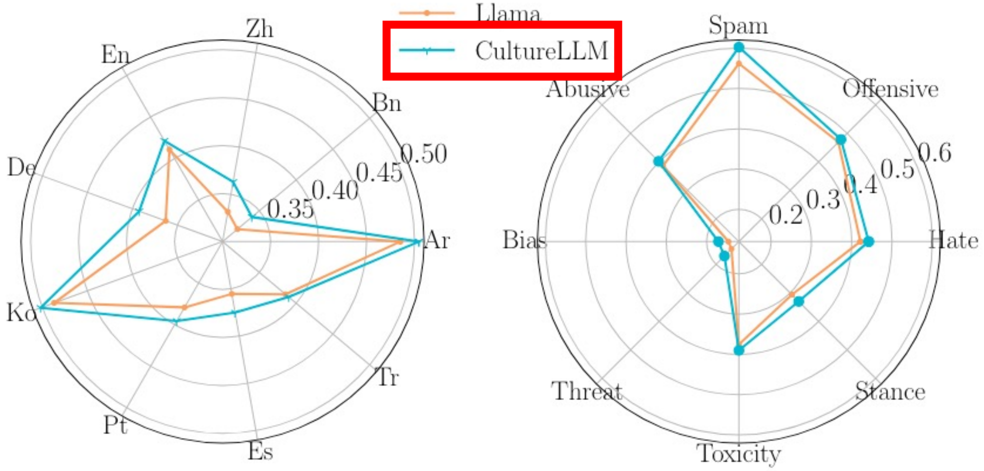


〈 CultureLLM-Llama-70b Performance Averaged by Cultures (Left) and Tasks (Right) 〉

- CultureLLM supports fine-tuning on **open-source LLMs** for better quality and reproducibility
- Llama2-70b-chat was used as the base model to fine-tune a CultureLLM-Llama2-70b

4️⃣ Culture-related Applications

: In addition to fine-tuned GPT, tested **Llama2 for open-source**

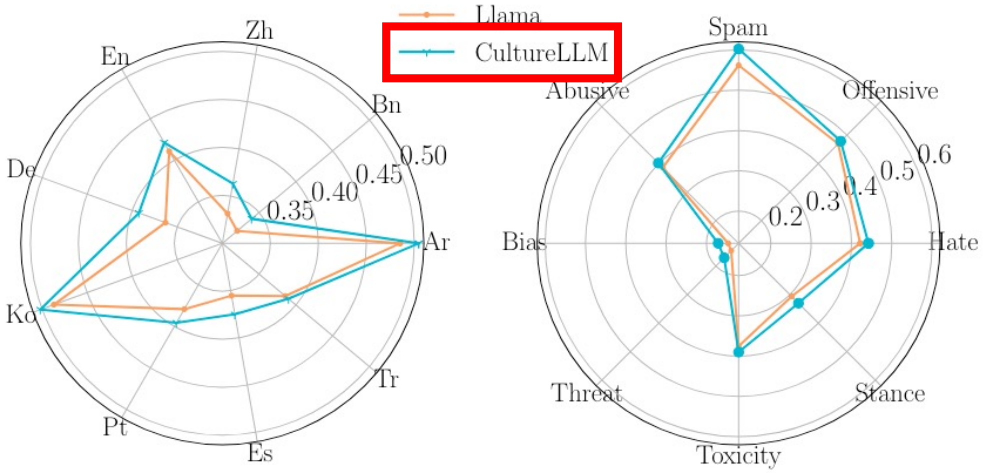


〈 CultureLLM-Llama-70b Performance Averaged by Cultures (Left) and Tasks (Right) 〉

- CultureLLM supports fine-tuning on **open-source LLMs** for better quality and reproducibility
- Llama2-70b-chat was used as the base model to fine-tune a CultureLLM-Llama2-70b
- CultureLLM-Llama2-70b outperforms the vanilla Llama model by 2.17%, demonstrating the effectiveness of fine-tuning

4️⃣ Culture-related Applications

: In addition to fine-tuned GPT, tested **Llama2 for open-source**



〈 CultureLLM-Llama-70b Performance Averaged by Cultures (Left) and Tasks (Right) 〉

- CultureLLM supports fine-tuning on **open-source LLMs** for better quality and reproducibility
- Llama2-70b-chat was used as the base model to fine-tune a CultureLLM-Llama2-70b
- CultureLLM-Llama2-70b outperforms the vanilla Llama model by 2.17%, demonstrating the effectiveness of fine-tuning

=> CultureLLM is a general approach to **enhance LLMs' cultural understanding**

Conclusion

: CultureLLM, a cost-effective solution to **integrate cultural differences into LLMs**

1. Leveraged **WVS data to align LLMs** with culturally representative values

Conclusion

: CultureLLM, a cost-effective solution to **integrate cultural differences into LLMs**

1. Leveraged **WVS data to align LLMs** with culturally representative values
2. Introduced a **novel and cost-effective data augmentation approach** for fine-tuning data generation

Conclusion

: CultureLLM, a cost-effective solution to **integrate cultural differences into LLMs**

1. Leveraged **WVS data to align LLMs** with culturally representative values
2. Introduced a **novel and cost-effective data augmentation approach** for fine-tuning data generation
3. Fine-tuned models on **both closed (GPT-3.5) and open-source (Llama2)** for broader applicability

Conclusion

: CultureLLM, a cost-effective solution to **integrate cultural differences into LLMs**

1. Leveraged **WVS data to align LLMs** with culturally representative values
2. Introduced a **novel and cost-effective data augmentation approach** for fine-tuning data generation
3. Fine-tuned models on **both closed (GPT-3.5) and open-source (Llama2)** for broader applicability
4. Applied and tested the resulting **CultureLLM across diverse cultural tasks**, demonstrating consistent performance

Conclusion (Insights)

: Key points for **applying** CultureLLM's process to **our research**

1. Leverage WVS, a high-quality, long-term, multi-country dataset, widely used in research

Ramezani, A., & Xu, Y. (2023). Knowledge of cultural moral norms in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 428–446.

Zhao, W., Mondal, D., Tandon, N., Dillion, D., Gray, K., & Gu, Y. (2024). WorldValuesBench: A Large-Scale Benchmark Dataset for Multi-Cultural Value Awareness of Language Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 17696–17706.

Kim, J., Jeong, C., Park, S., Cha, M., & Lee, W. (2024). How Do Moral Emotions Shape Political Participation? A Cross-Cultural Analysis of Online Petitions Using Language Models. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 16274–16289.

Jackson, J. C., & Medvedev, D. (2024). Worldwide divergence of values. *Nature Communications, 15*(1), 2650.

⋮

Conclusion (Insights)

: Key points for **applying** CultureLLM's process to **our research**

1. Leverage WVS, a high-quality, long-term, multi-country dataset, widely used in research
2. Benchmark each step of Human-AI (GPT) collaborative data augmentation

Conclusion (Insights)

: Key points for **applying** CultureLLM's process to **our research**

1. Leverage WVS, a high-quality, long-term, multi-country dataset, widely used in research
2. Benchmark each step of Human-AI (GPT) collaborative data augmentation
3. Identify key application points in empirically validated processes

   (Fine-tuning N-shot, Perplexity, Diversity Gain, etc.)

Conclusion (Insights)

: Key points for **applying** CultureLLM's process to **our research**

1. Leverage **WVS**, a high-quality, long-term, multi-country dataset, widely used in research
2. Benchmark each step of **Human-AI (GPT) collaborative data augmentation**
3. Identify key application points in **empirically validated processes**

   (Fine-tuning N-shot, Perplexity, Diversity Gain, etc.)
4. Extend beyond closed-source (GPT) fine-tuning to **open-source (Llama) for Open Science**

Conclusion (Insights)

: Key points for **applying** CultureLLM's process to **our research**

1. Leverage **WVS**, a high-quality, long-term, multi-country dataset, widely used in research
2. Benchmark each step of **Human-AI (GPT) collaborative data augmentation**
3. Identify key application points in **empirically validated processes**

   (Fine-tuning N-shot, Perplexity, Diversity Gain, etc.)
4. Extend beyond closed-source (GPT) fine-tuning to **open-source (Llama) for Open Science**
5. Go beyond multicultural, multilingual modeling by **testing the model on open-source datasets**

   (8 evaluation tasks, 59 datasets)