Identity-Aware Large Language Models require Cultural Reasoning

Alistair Plum[1,*], Anne-Marie Lutgen[1], Christoph Purschke[1] and Achim Rettinger[2]

[1]University of Luxembourg, Luxembourg
[2]Trier University, Germany

Abstract

Large language models have become the latest trend in natural language processing, heavily featuring in the digital tools we use every day. However, their replies often reflect a narrow cultural viewpoint that overlooks the diversity of global users. This missing capability could be referred to as cultural reasoning, which we define here as the capacity of a model to recognise culture-specific knowledge values and social norms, and to adjust its output so that it aligns with the expectations of individual users. Because culture shapes interpretation, emotional resonance, and acceptable behaviour, cultural reasoning is essential for identity-aware AI. When this capacity is limited or absent, models can sustain stereotypes, ignore minority perspectives, erode trust, and perpetuate hate. Recent empirical studies strongly suggest that current models default to Western norms when judging moral dilemmas, interpreting idioms, or offering advice, and that fine-tuning on survey data only partly reduces this tendency. The present evaluation methods mainly report static accuracy scores and thus fail to capture adaptive reasoning in context. Although broader datasets can help, they cannot alone ensure genuine cultural competence. Therefore, we argue that cultural reasoning must be treated as a foundational capability alongside factual accuracy and linguistic coherence. By clarifying the concept and outlining initial directions for its assessment, a foundation is laid for future systems to be able to respond with greater sensitivity to the complex fabric of human culture.

Keywords

Cultural Reasoning, Large Language Models, Identity-Aware AI, Cross-Cultural Evaluation of LLMs

1. Introduction

The advent of large language models (LLMs) has brought about considerable advances in the field of artificial intelligence (AI). Language models have allowed many tasks in- and outside of natural language processing (NLP) to be automated. Prominent LLM families such as GPT [1, 2], LLaMA [3], and Aya [4], which are trained on massive web-scale corpora [5], exemplify the shift towards conversational models that can generate responses to open-domain user questions. AI as a concept has now become more real than ever, and LLMs that "know" everything and nothing at the same time are taking us closer.[1]

As we use these LLMs to generate what we desire at ever-increasing rates, concerns about the generated content also arise, often in some way linked to the accurate representation of the diverse cultural identities that shape humanity. More specifically, when we interact with these models, we expect not only correct or reasonable answers, we also expect these models to reason carefully, considering what aspects of our interaction are linked to not facts alone but our identity. Undeniably, our cultural backgrounds stand as a central pillar that constitutes our identity. Because of the sheer diversity of cultures represented on earth, however, it is clear that concerns arise as to the proper or accurate portrayal of these cultures as part of the exchange we have with language models and AI. Language and cultural identity are inherently intertwined [6] and therefore the language in which we communicate with AI is central to the understanding of our cultural backgrounds. These concerns are consistent

Identity-Aware AI workshop at 28th European Conference on Artificial Intelligence, October 25, 2025, Bologna, Italy
*Corresponding author.
alistair.plum@uni.lu (A. Plum); anne-marie.lutgen@uni.lu (A. Lutgen); christoph.purschke@uni.lu (C. Purschke); a.rettinger@uni-trier.de (A. Rettinger)
0000-0003-0977-3467 (A. Plum); 0009-0001-4342-9718 (A. Lutgen); 0000-0002-9655-2058 (C. Purschke); 0000-0003-4950-1167 (A. Rettinger)

[1]We point out that sentences ascribing human capacities to models by using activity verbs or metaphors are not meant in their literal sense, that is, we do not equate model output with human reasoning and speech.

with evidence that model outputs can encode and amplify social biases unless explicitly addressed in training, data or evaluation design [7, 8].

Recent studies have shown that LLMs trained predominantly on English-language data exhibit biases favouring Western cultural norms, thereby limiting their applicability in non-Western contexts [9, 10]. For instance, Karim et al. [11] show that ostensibly culture-neutral tasks are handled unevenly across cultural contexts, while Naous et al. [12] demonstrate Western-centric preferences in Arabic settings, with models over-selecting Western entities and frames. The ability to navigate the subtleties of cultural context becomes much harder, however, when models are biased towards Western norms; such bias can lead to outputs that perpetuate stereotypes or marginalise under-represented groups [13, 14], thereby constraining the usefulness of English-dominant LLMs in culturally diverse applications.

Taking all this into consideration, it stands to reason that any efforts that are aimed towards identity-aware AI must be preceded by efforts to improve these issues in LLMs. Culture influences every individual's identity in some way, and this influence is not necessarily bound by country or location. Because of the way culture can influence every individual differently, it is necessary to understand the cultural influences and take them into consideration when interacting with another individual, in effect reasoning while taking culture into account. With AI growing so rapidly in use, we need to ensure that it is inclusive and adaptive, in turn requiring it to be identity-aware and therefore capable of cultural reasoning (CR).

CR is required when acceptable interpretations and actions depend on local practices rather than universal rules, and where potential bias needs to be processed appropriately. Current evaluations mostly test factual recall or generic safety and therefore do not reveal whether models can recognise culture-specific norms, apply them consistently, and reconcile conflicts across settings. Importantly, bias should not be treated as automatically negative: conceptually, the term bias simply captures the fact that our views on the world as mediated through language put the world in a perspective that reflects our perceptions, norms and beliefs. Where work on raciolinguistic, gender-related or other forms of bias (as the by-product of model training) rightly foregrounds harms and issues with equal reprensentation of cultural diversity, CR clarifies that some forms of "bias" are desirable insofar as they enable simulated perspective-taking, i.e. adopting a culturally situated point of view for appropriate behaviour and interpretation.

This paper argues that CR is a distinct capability that should be defined, evaluated, and improved with targeted procedures. Beyond model capability claims, this work intersects long-standing concerns about NLP and its social impact and data practices. Early position and survey papers argue that technical progress must be coupled with attention to downstream harms and structural inequities [15, 7]. Complementary critiques highlight risks from scale and opaque data pipelines [8, 16], and broader taxonomies of language model risks frame why cultural specifics urgently require dedicated evaluation [17, 18].

Efforts to evaluate cultural alignment in LLMs have employed frameworks to assess specific aspects of CR in LLMs. This includes Hofstede's cultural dimensions [19] and the GLOBE study to assess models' adherence to specific cultural values [20]. However, these approaches often rely on static, survey-based methodologies that may not capture the dynamic and context-dependent nature of culture. They tend to focus on knowledge about cultures rather than the ability to reason within diverse cultural frameworks. In addition, evaluation metrics for LLMs tend to focus primarily on performance benchmarks such as accuracy, fluency, and coherence in tasks like question answering or text summarisation. The evaluation of these cultural capabilities of LLMs on their own does not necessarily show a detailed picture of CR capabilities.

To address the problems and limitations stated above, this paper proposes to define the term *cultural reasoning* in the context of LLMs, as well as the development of a robust, interdisciplinary evaluation framework that transcends traditional NLP metrics. This framework should incorporate methodologies from various domains in humanities research to assess the capacity of LLMs for CR. For instance, evaluating models' responses to culturally sensitive scenarios, their adaptability to context-specific variation in the language use and language choice in multilingual settings, and their understanding of context-specific norms can provide more in-depth insights into their cultural competence.

Furthermore, the case is made for the creation of diverse, representative datasets that encompass a wide range of cultural narratives and perspectives. Initiatives like CulturalBench [21] and WorldView-Bench [22] have made strides in this direction by introducing benchmarks that evaluate LLMs across various cultural contexts. Building upon these efforts, the proposed framework in this paper aims to guide the development and evaluation of AI systems that are not only linguistically proficient but also culturally attuned.

In the following sections, the conceptual underpinnings of CR in AI will be elaborated, a critique of existing evaluation methodologies offered, and an outline of the components of the proposed framework sketched. Through this discourse, we hope that a shift towards more culturally aware AI systems that respect and reflect the rich diversity of human identities can be induced.

2. Related Work

The term *cultural reasoning* is not yet an established technical term in AI and appears only sporadically in academic literature. Few papers formally define *culture* or *cultural reasoning*, underscoring the complexity and novelty of the concept [23, 24]. Researchers often use adjacent terms like *cultural knowledge*, *cultural awareness*, or *cultural competence* [21]. For instance, a recent survey of over 90 papers found that none explicitly define "culture" instead probing models with proxy aspects such as values or demographics, and that only certain aspects, such as values, norms, or objectives, have been studied while many remain unexplored [24].

For the purposes of this paper, we understand culture as defined by [25]: culture refers to the dynamic processes through which social meaning is created, negotiated, and materialised in practices, institutions, and lifeworlds. It is not static or merely symbolic, but a contested field shaped by ongoing exchanges, disputes, and power relations that guide both collective and individual activity. Language is part of culture and the primary medium through which culture is expressed, shared, and adapted [26].

Historically, CR has seen limited use, such as the Cultural-Reasoning Architecture (CARA) system in 2007 [27]. CARA was an attempt to model cultural group behaviours and norms for training simulations through a cognitive architecture [27]. Today, however, CR is gaining traction within NLP, AI ethics, human-computer interaction (HCI), and cognitive science as AI systems are deployed globally and must navigate diverse cultural contexts. Recent work has begun to work towards how AI can handle cultural differences in knowledge, values, and communication, though there is not necessarily a consensus on definitions and approaches [28, 24]. Most research relevant to CR deals with specific sub-areas of this broad challenge.

2.1. Cultural Reasoning vs Moral Norms, Bias, and Value Alignment

Empirical studies suggest that current LLMs have some awareness of cultural differences but limited adaptability in their responses, often related to current cross-lingual architectures of models. For instance, Kharchenko et al. [19] found that while LLMs can recognise that different countries have differing value orientations (drawing on Hofstede's cultural dimension theory), they often fail to adjust their advice or reasoning to align with those local values. In other words, a model might know of a cultural norm difference, yet not sufficiently apply that knowledge when generating answers for a user from that culture.

Closely related to this, Münker [29] found that some models tend to demonstrate clear limitations in the way they capture cross-cultural moral diversity, as they do not sufficiently differentiate between cultural contexts and represent Western perspectives more accurately than non-Western ones.

Similarly, evaluations of moral reasoning across cultures show that LLMs tend to align closely with Western moral frameworks by default, demonstrating a form of cultural myopia in their moral judgments [30]. Without explicit tuning, models largely mirror the norms dominant in their training data, and struggle to accurately reflect the moral or social norms of less-represented cultures. This aligns with broader findings that models often default to majority cultural priors unless representations of norms and trade-offs are made explicit [31, 7].

Beyond values and ethics, other culturally-rooted reasoning tasks reveal significant performance gaps. Liu et al. [23] investigated multilingual LLMs' capabilities with proverbs and sayings from different cultures, as these encapsulate cultural wisdom and often require context-specific interpretation. The authors found that state-of-the-art multilingual models know only a limited set of common proverbs and, even when a proverb is memorised, the model may not truly understand its meaning in context [23]. The models struggled in particular with figurative language and context-based reasoning: for instance, when prompted to interpret or choose the correct continuation of a culturally specific saying, their performance was poor, especially if the question was more complex, such as selecting an incorrect meaning [23]. Perhaps most striking, a clear culture gap was observed, where models performed worse when reasoning about proverbs translated from languages outside the model's primary training focus. This suggests that even multilingual LLMs lack robust cross-cultural abstraction and do not seamlessly transfer reasoning skills across cultural contexts.

Several benchmarks and analyses have reinforced these findings of partial awareness but inadequate adaptation. Rao et al. [32] introduced NormAD, a dataset of roughly 2,600 short stories from 75 countries, each reflecting local social norms, to test LLMs' cultural adaptability. Their results showed that models struggle with CR across all levels: whether identifying acceptable behaviour in a story or adapting a continuation to fit a given country's norms, performance was significantly lower for non-Western contexts [32]. Even when explicitly provided with the relevant cultural norm as context, the best model, Mistral-7B, achieved only about 82% accuracy, compared to human performance around 95%[32]. Notably, models performed better in judging stories that adhered to common norms than those that violated local norms, hinting at a bias toward assuming normative behaviour, or a general agreeableness bias that impairs detection of culturally deviant situations[32].

In the domain of factual and procedural cultural knowledge, comprehensive evaluations have exposed wide gaps in what LLMs know. Chiu et al. [21] present CulturalBench, a suite of 1,696 human-written, culturally diverse questions covering 45 global regions (including under-represented ones like Zimbabwe, Bangladesh, and so on) and topics ranging from food and festivals to social etiquette. Human performance on these questions is near 92% accuracy, yet even top-tier models like GPT-4 struggle. On the hardest version of CulturalBench, state-of-the-art models' accuracies range roughly between 30% and 60% [21]. Models often latch onto a single trained-for answer and fail on questions where multiple answers are correct or context-dependent (e.g., "What utensils do the Chinese usually use" expects both chopsticks and spoons/forks depending on context, but a model might always answer "chopstick" only) [21]. According to the authors, model performance is weakest on regions that are less represented in typical training data, such as the Middle East, North Africa, and parts of South America, making it clear that model knowledge is skewed toward cultures prevalent in its training corpus [21]. These evaluations echo the pattern seen in norms and values: current LLMs, even when fluent and knowledgeable in a general sense, remain culturally distant and often cannot replicate the breadth of human cultural knowledge or adapt their reasoning to specific cultural settings without additional help.

Not all findings are entirely pessimistic, though, as some differentiation in outputs across cultures has been observed. For example, when comparing LLMs developed in different cultural environments, there are measurable variations. Karinshak et al. [20] evaluated Chinese vs. U.S. origin LLMs using the GLOBE cultural values framework and found that each model reflects certain biases of its origin culture's value system. In their GLOBE benchmark, models showed both similarities and systematic differences in how they prioritise values, suggesting that the cultural context of model development or alignment does influence its behaviour to a degree [20]. However, they also note that extracting these differences requires careful, open-ended analysis and introduce an LLMs-as-a-Jury method for evaluating generation content, since simple QA tests might miss subtle cultural value cues [20].

Overall, while LLMs today are not reliably culturally adaptive, research is beginning to show that they possess areas of cultural knowledge and can mimic some cultural differences, but lack a generalised competence to reason as a local across the world's many cultures.

2.2. Culture in Training Data and Representation

As we have seen, a major reason behind the narrow cultural viewpoint of LLMs is the skewed representation of cultures in their training data. The vast majority of large-scale training corpora for language models are dominated by Western perspectives and English[2], as well as only a few other very high-resource languages. This leads to models that favour Western contexts by default, even when operating in other languages or regions [12, 9]. Joshi et al. [9] quantified the linguistic diversity gap in NLP, showing that a handful of languages, primarily English and a few European and East Asian languages, account for most NLP resources, while the thousands of other languages and by extension, the cultures associated with them, are scarcely represented. Data documentation and dataset development practices further shape which cultural signals are learnable in the first place [16]. They call into question the language agnostic claims of many models, underlining that current systems inherently prioritise certain cultures unless active efforts are made to include diverse languages [9]. In practical terms, this means an LLM is far more likely to have read about Christmas than Diwali, or about New York than Nairobi, yielding an uneven cultural knowledge base.

Furthermore, by including only small proportions of non-western languages, context-specific varieties and variation is barely included in the training data [9, 33, 34]. This creates a cultural gap in the communicative style for different settings and favours high prestige varieties that are most likely accounted for in the small amount of training data. In multilingual contexts, this gap widens where language choice for specific situations is a highly complex process. The scarcity of data does not allow for an even representation of different contexts and therefore different languages in the same cultural setting.

The Western-centric data bias is highly evident in model outputs. Naous et al. [12] demonstrated that both multilingual and even ostensibly Arabic-focused LLMs showed a strong bias towards Western entities and contexts when tested in Arabic. In their experiments, using their CAMeL benchmark, models frequently produced completions and associations that were more Western-oriented, sometimes even stereotyping non-Western contexts or handling them unfairly [12]. For example, when generating stories or filling in text, the models struggled to appropriately adapt to specifically Arab cultural settings, often injecting Western assumptions or failing to use culturally appropriate references. Such outcomes are said to be directly traceable to the training data: Naous et al. [12] analysed common pre-training sources and found them lacking in the richness and versatility needed for culturally aware AI. In fact, they suggest that without significant adjustments, relying on sources like Wikipedia may perpetuate cultural biases since those sources themselves can be skewed or incomplete for certain cultures [12]. These outcomes are consistent with critiques that large, weakly curated web corpora can entrench existing cultural skews unless counterbalanced [8].

Cultural under-representation in training data also affects fundamental processing at the token level. A follow-up study by Naous and Xu [35] looked at how pre-training data frequencies cause structural biases. They discovered that the frequency-based tokenisation schemes used by LLMs disadvantage less-represented languages and cultural terms. For instance, in Arabic, certain common words or names with multiple meanings ended up fragmented or poorly encoded by the tokeniser because the model had not seen them in varied contexts often enough [35]. Moreover, if a language shares script with others, as many non-Arabic languages use Arabic script, tokenisers trained on aggregated text can confuse or conflate culturally distinct terms. As model vocabulary sizes increase, these issues can worsen, leading to higher perplexity and confusion for culturally specific content [35]. In short, the very way text is ingested by LLMs can reflect cultural biases, as concepts frequent in Western contexts are assigned well-formed embeddings, whereas those frequent in Swahili or Bengali might be less distinct in vector space, hindering the model's fluency and understanding in those contexts.

Beyond data imbalance, there are also emergent biases in how models generalise, akin to social identity biases in humans. Hu et al. [13] examined whether LLMs exhibit in-group vs. out-group bias, a fundamental aspect of cultural psychology. By using prompts such as "*We are ..., they are ...*"

[2]It must be noted that there needs to be a distinction between culture and language as they are not interchangeable even though language is part of culture.

across many identity group pairs, they found that many base models strongly favour whatever group is described as "we" (ingroup favouritism) and often generate derogatory or negative continuations for "they" (outgroup) [13]. This pattern held for various groups and appears to reflect biases present in the training data or human texts. Although instruction-tuned models showed some reduction of this effect, it was still present unless specific bias mitigation fine-tuning was done [13]. These results imply that an LLM might not just lack knowledge of a culture, but could also unintentionally disrespect it by echoing harmful biases or stereotypes, if those were implicit in the training corpus. From an AI ethics standpoint, this raises concerns about deploying such models globally, as they could reinforce cultural hegemony or prejudice if not carefully corrected.

Research is actively exploring solutions to these issues of cultural bias and blind spots in data. One straightforward approach is to curate or augment training data with more culturally diverse sources, and to apply fine-tuning to instil cultural knowledge. Ramezani and Xu [30] were able to improve a model's predictions of various countries' moral norms by fine-tuning on survey data from those countries. Hu et al. [13] also showed that careful curation, including balancing training data or filtering out biased content, and additional fine-tuning can substantially reduce the level of in-group/out-group bias exhibited by LLMs.

Focusing more on the level of the individual, Zhang et al. [36] clearly demonstrate that current models do not reflect the vast breadth of preferences that individuals have across cultural, political and further dimensions. The authors also compile a dataset to improve model performance in this area.

Another line of research focuses on prompting and multi-agent techniques to inject multiple cultural perspectives. Mushtaq et al. [22], for example, argue for a multiplex world-view approach: instead of a single LLM response that might reflect a single dominant perspective, they have multiple LLM agents, each initialised with a different cultural viewpoint, jointly produce an answer. In their experiments, this approach dramatically increased the diversity of perspectives in outputs and improved the overall balance of viewpoints, as measured by an entropy-based metric of perspective distribution [22]. Such methods highlight that CR may be enhanced not just by feeding more data, but also by architecting interactions, either via prompt engineering or system design that force the model to consider alternatives and lesser-heard viewpoints.

Finally, the research community is devising more robust benchmarks and evaluation frameworks to track progress in culturally aware AI. In addition to CulturalBench and NormAD mentioned above, others like GIMMICK evaluate vision and language models across many cultural settings to identify where models know tangible cultural facts, such as flags or foods, versus where they fail on intangibles, such as rituals or values [37]. Such evaluations consistently find Western or high-resource culture bias across modalities, but also provide quantitative targets for improvement. The hope is that with clear benchmarks, future models can be trained or adjusted to perform well across all cultures, not just the ones most represented in their training data. Moving forward, the literature points toward a combination of strategies for true CR in AI: better data, meaning more inclusive and representative corpora, better definitions and taxonomies of "culture" to guide what models should learn [28], new training or prompting techniques to improve cultural adaptability, and strong evaluation to ensure that as AI becomes culturally competent.

2.3. Evaluation of Culture Specifics

The systematic evaluation of culture specifics also connects to risk taxonomies for language models and the broader analysis of foundation-model externalities [17, 18]. A growing body of empirical research and case studies illustrates how LLMs and related AI systems routinely manifest cultural biases and lapses in cultural sensitivity. These include:

- **Western-centric biases in multilingual LLM outputs**: Using the CAMeL dataset focused on Arabic contexts, Naous et al. demonstrate that multilingual and monolingual Arabic language models disproportionately favour entities and representations associated with Western culture, leading to inappropriate or stereotyped outputs in Arab cultural settings [12].

- **Misalignment with culturally specific moral norms**: Tao et al. perform a disaggregated evaluation of several LLMs (e.g., GPT-3.5, GPT-4 variants) against nationally representative World Values Survey data. They find that model outputs reflect values more typical of English-speaking and Protestant European societies, rather than those of the countries in question, and while cultural prompting improves alignment for many regions, it fails or even exacerbates bias for others [10].
- **Propagation of stereotypes and representational harms**: A recent UNESCO backed study revealed that models like GPT-3.5 and Llama 2 engage in regressive gender stereotyping that portrays women predominantly in domestic roles while associating men with career-oriented concepts, as well as displaying homophobic and racial biases [14].
- **Salary bias across demographic profiles**: A recent empirical analysis reveals that AI chatbots (e.g., ChatGPT, GPT-4o-mini, Llama 3.1) systematically recommend lower starting salaries to women and ethnic minorities even when qualifications and role descriptions are identical to those of male or white candidates, with differences spanning tens of thousands of dollars [38].
- **Subtle dialect-based prejudice**: Reporting on covert forms of racism, researchers found that models such as ChatGPT and Gemini hold biased stereotypes against speakers of African American Vernacular English, perceiving them as less intelligent or employable and resulting in reduced recommendations or harsher judgments [39].
- **Underrepresented cultural groups suffering stereotype bias**: The Indian-BhED dataset reveals that LLMs heavily stereotype Indian-specific axes of identity, such as caste and religion. Models like GPT-2 and GPT-3.5 generated stereotypical outputs in 63–79% of cases with respect to caste and 69–72% for religion, underlining their failure to handle Global South contexts sensitively [40].
- **Stereotype propagation across multiple languages**: The SHADES dataset, created under the BigScience initiative, allows systematic measurement of stereotypes in multiple languages. SHADES reveals that models often replicate harmful stereotypes beyond English, and even fabricate pseudo-scientific justifications for them, therefore extending cultural harm across linguistic boundaries [41].

These documented issues with model output underscore a critical problem: current LLMs lack genuine cultural understanding and, without disciplined intervention, amplify cultural bias in ways that are often subtle, damaging, and widespread.

3. Defining Cultural Reasoning

CR in AI is an emerging field that is investigating how AI systems understand and adapt to the world's diverse social and cultural norms. Current LLMs exhibit a degree of cultural knowledge and can mimic some differences, but they largely remain biased towards their dominant cultural context of their training data and struggle with truly adapting to unfamiliar cultural scenarios. The research so far, which spans moral reasoning, value alignment, idiomatic understanding, and bias analysis, shows a clear picture of the challenges. It also lays important groundwork, by identifying specific shortcomings, ranging from incorrect moral norm predictions to proverb misinterpretations, these studies guide the development of methods to make AI more culturally aware. Placing CR within these established lines of critique (bias surveys, data-centric risks, and LM risk taxonomies) clarifies both why the capability matters and how progress should be measured [15, 7, 16, 17].

In this paper, *cultural reasoning* denotes the capacity to select and justify interpretations or actions that are contingent on culture-specific norms, conventions, and procedures, given a locale and context. It differs from cultural *facts*, style or register control (lexical or politeness choices), moral value judgement

per se, and generic bias mitigation, although it can include any of these. We work with the definition of culture already mentioned in Section 2. In addition, while some recent systems expose intermediate "reasoning" tokens [42], we use the term in a broader, model-agnostic sense, and more in line with its traditional meaning.

We treat CR as distinct from translation/localisation alone and from demographic targeting. Producing the right language variety or tone is necessary but not sufficient; the system must use culture-specific premises to arrive at, and explain, its choice in a manner that is adequate for the context. Related strands on normative and social common-sense reasoning provide scaffolding for expressing such culture-specific justifications [31].

We also place CR carefully in relation to bias and de-biasing efforts in research. While we have shown at length the problematic tendencies that arise with bias in training data, we do not regard de-biasing training data as part of CR. Much rather, we consider the detection and careful treatment of bias in these situations as part of CR.

In operational terms, a model response is taken to *exhibit cultural reasoning* if it satisfies one or more of the following:

(a) **Context-appropriate application:** selects procedures or norms that vary by locale (e.g., administrative steps, forms of address) and applies them correctly to the described situation.

(b) **Justified adaptation:** provides a short rationale that references relevant norms, practices, or constraints for the specified cultural frame.

(c) **Conflict reconciliation:** when multiple cultural frames are salient (e.g., cross-border or diasporic settings), reconciles them explicitly, with prompts such as "do X because Y takes precedence in setting S".

(d) **Sensitivity to intra-cultural variation:** acknowledges plurality (majority/minority practices; regional/age variation) with calibrated uncertainty or requests for disambiguation when appropriate.

(e) **Pragmatic interpretation:** interprets culturally bound figurative language, implicatures, or rituals in context rather than giving literal or default-global readings.

(f) **Consistency under paraphrase/contrast:** makes stable choices across rephrasings and structured contrast sets tied to the same cultural premise.

4. Analysing and Evaluating Cultural Reasoning in LLMs

To better analyse and evaluate CR in LLMs, we propose the following methodology, which aims to systematically elicit, validate, and integrate culturally specific knowledge into large language models in order to improve their ability to engage in culturally sensitive reasoning. The process is organised into several sequential stages (see Figure 1), each building on the outcomes of the previous one.

4.1. Identifying Domains of Cultural Variation

The first stage involves identifying a set of domains or areas of life in which cultural values, norms, and procedures differ significantly across societies. These may include, for example, family structure, workplace hierarchy, gift-giving customs, conflict resolution strategies, and moral priorities. The selection will draw on established cross-cultural psychology frameworks (e.g., Hofstede's dimensions, Schwartz's value theory) as well as ethnographic and sociolinguistic literature. These domains will serve as thematic anchors for all subsequent data collection and evaluation.

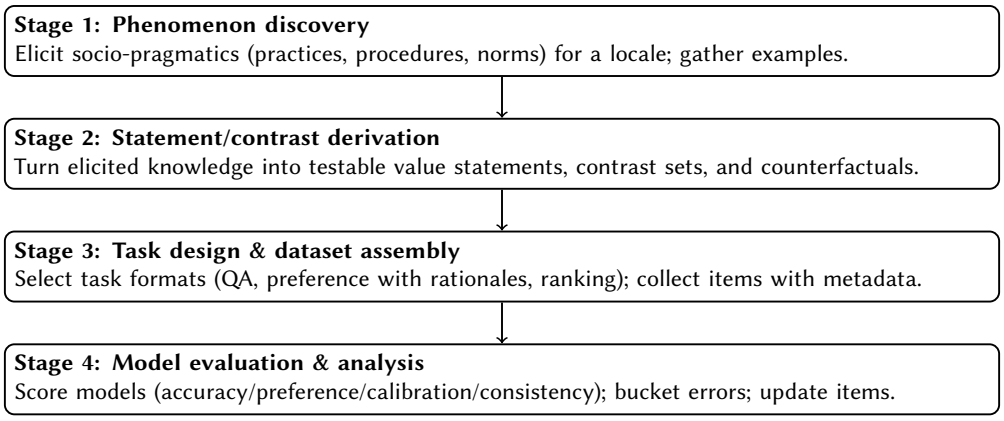

Figure 1: Four-stage cultural-reasoning evaluation pipeline.

4.2. Eliciting and Evaluating Cultural Descriptions

For each domain, prompts will be designed to elicit descriptive accounts of the relevant norms or values in multiple languages for multiple target countries, forming an initial matrix of culture–language combinations. A further aspect that we will consider in this matrix is the importance of specific languages and varieties for specific domains in multilingual cultural settings.

Taking Luxembourg as an example [43], a country with three official languages, Luxembourgish (the national language), French (the legislative one) and German, and with a deep cultural history of language contact from the neighbouring regions, Germany, France and Belgium. One domain, where the language choice is very clear is the legal one, as French is institutionally the only language allowed for. However, since the multilingual situation is very complex in Luxembourg, the choice of language in a communicative situation is complex and relies on cultural knowledge and the language of the communicative recipient. This situation is even more complicated considering the high percentage of non-Luxembourgers living in the country and the high number of workers crossing the border every day from France, Belgium and Germany. As this is only one example for a multilingual society from many, it is clear that language plays a big part in the cultural setting. Table 1 shows an overview of what this setup could look like conceptually for examples from the Greater Region of Luxembourg, Wallonia, Lorraine, Rhineland-Palatinate, and Saarland (Greater Region of SaarLorLux).

Further, the role of English will be carefully considered, both as a native cultural context and as a cross-cultural medium, to capture potential differences in meaning when norms are expressed in English rather than the local language. The resulting descriptions will then be evaluated by human annotators familiar with the relevant cultural contexts, who will indicate whether each description is accurate (*is true*) or inaccurate (*is not true*). This evaluation step ensures that only reliable and culturally authentic descriptions progress to the next phase.

4.3. Deriving and Assessing Value Statements

From the verified descriptions, concise value statements will be formulated to capture the underlying principles, priorities, or normative expectations reflected in the data. These statements should be distilled from the descriptions in a way that retains cultural specificity while also being general enough to apply across related contexts. Once drafted, the value statements will undergo a second human evaluation in which annotators judge three dimensions: Whether the statement is true for the target culture, the importance of the value within that culture and the specificity of the statement, making sure it is neither too vague to be meaningful nor so narrow as to lose relevance. This step filters the statements to retain only those that are both culturally valid and useful for downstream modelling.

Domain	Country–Lang.	Example Prompts	Importance of Lang./Var.
Family & Child-rearing	LU–lb/fr/de/pt/en	How do parents talk to young children? / What values are taught at home?	*lb* dominant in families; *fr* in early education; *pt* central in diaspora homes.
Education & Schooling	DE–de/en	What makes a good teacher? / How do students address authority?	*de* as instructional norm; *en* growing in schools.
Work & Professional Life	FR–fr/en	How should one behave with a boss? / How formal is workplace communication?	*fr* used in formal settings; *en* in international work.
Public Services & Administration	LU–fr/lb/de	How are citizens addressed in official communication?	*fr* dominates bureaucracy; *lb* adds local tone; *de* used cross-border.
Media & Public Discourse	Greater Region –lb/fr/de	How are political opinions expressed in local media?	Mix varies regionally; choice of language indexes stance or identity.
Everyday Politeness & Interaction	FR–fr	How do people greet or thank each other?	*fr* sets norms; regional varieties convey familiarity or solidarity.

Table 1
Matrix of domains, country–language combinations, example prompts, and cultural relevance.

4.4. Fine-Tuning and Post-Evaluation

The subset of high-quality value statements will be used to fine-tune a language model with the goal of improving its ability to recognise, articulate, and apply culturally relevant values in its outputs. After fine-tuning, the original prompting procedure for cultural descriptions will be repeated, and the new outputs will be evaluated using the same *is true* / *is not true* criterion as before. Comparing pre- and post-fine-tuning results will indicate whether the inclusion of validated value statements has led to measurable improvements in the cultural accuracy, contextual relevance, and nuance of the model's responses.

5. Discussion

The synthesis of prior work and our empirical considerations clarify why *cultural reasoning* should be treated as a cornerstone capability for identity-aware AI. While moral norms are important, culture also includes pragmatic conventions, historical narratives, communicative styles, and procedural knowledge, as has been argued previously [44, 24]. This distinction matters, because even when certain moral norms appear to generalise across cultures [30], their interpretation and salience are culturally situated and interact with other values [10, 19].

While we have shown that models and data are often Western-centric, and that areas outside this boundary are disadvantaged, the need for CR lies also within this boundary. For example, the Greater Region of SaarLorLux illustrates the practical complexity of CR. Here, norms and languages mix across borders, producing hybrid identities in which people switch between, or merge, cultural frames. The Greater Region encompasses regions from four different countries and even more language varieties than only German, French and Luxembourgish are part of this area. Moreover, the multilingual situation is central to this region and even differs in each country.

The importance of including the sociolinguistic situation in the thought process is implied through this example. As language and culture are deeply intertwined and form identities, this holds even more importance in multilingual settings. As the choice of language or variety in specific situations is a complex process that is highly routinised in the society itself. However, for AI this is part of the cultural knowledge that it needs to be aware of. Effective AI must recognise distinct norms and, when appropriate, reconcile them through a process that applies relevant cultural information in new mixed contexts.

CR therefore goes beyond cultural awareness. Awareness recognises differences, whereas reasoning applies, adapts, and, when necessary, combines norms to guide behaviour or generation, following Liu et al. [28], Chiu et al. [21]. Eliciting this capability in LLMs is difficult, because a model must "behave" as a person from one culture, or a blend of cultures, would behave, incorporating implicit values, communicative preferences, and context-specific interpretations [45, 35].

Cultural bias is related but distinct. Bias reflects the sum of learned perspectives on a given subject and often arises from pre-training data, as shown by a lot of current research [12, 9]. Such bias can undermine CR by triggering defaults to overrepresented norms even when another frame is more appropriate [13, 14]. Recent benchmarks, including NORMAD by Rao et al. [32], CulturalBench by Chiu et al. [21], and SeaEval by Wang et al. [46], report frequent misapplication of norms in under-represented contexts and suggest the need for targeted interventions.

It is tempting to assume that modern models can adopt a cultural perspective on demand, for example when asked to answer as a person from a less well-represented country. Models often simulate well-represented cultures [22, 23], yet success depends on the extent and quality of cultural signals in training. This dependence raises questions about how much knowledge is sufficient and whether models can integrate multiple influences, including whether an LLM could emulate a user shaped equally by Luxembourgish and French norms. Progress requires new evaluation methods [20, 47] and clearer theory on what constitutes sufficient cultural competence.

Cross-lingual transfer should be distinguished from cross-cultural transfer. Multilingual models often transfer linguistic form effectively [48, 49], yet nuanced CR is far harder. Subtle cues, including indirectness in politeness strategies or the moral weight of certain actions, do not reliably survive translation [11, 50]. Identity-aware AI therefore cannot rely on multilingual capacity alone and must model culture-specific reasoning directly.

In sum, CR aligns system outputs with users' cultural contexts in ways that exceed surface-level adaptation. For identity-aware AI, this capability is necessary for equitable and effective interaction. Achieving it will require curated cultural knowledge, mitigation of bias at data and model levels, and evaluation frameworks that reflect the complexity of real-world, often mixed, cultural settings.

6. Conclusion

This paper presented a survey of recent work at the intersection of NLP, HCI, and computational social science to situate CR within existing concepts and evaluations. Concretely, we contrasted adjacent notions (cultural knowledge, awareness, alignment) with CR and synthesised evidence from benchmarks and studies showing that contemporary LLMs often default to Western-centric norms and struggle to adapt to under-represented cultural contexts. This review established both the problem space and the measurement gap that motivates our approach.

We proposed an operational definition of CR as the capability to recognise, apply, and combine culturally grounded norms, values, and procedures to guide model behaviour. Methodologically, we distinguished CR from mere awareness by requiring perspective-taking and norm-application, including in mixed or hybrid identities (e.g., users in the Greater Region). The definition is harmonised with ML usage of "reasoning" while remaining agnostic to implementation, thereby supporting evaluation across architectures and training regimes.

Moreover, we outlined a concrete, data-to-evaluation pipeline. The pipeline includes the identification of domains of cultural variation, as well as the elicitation of multilingual country-specific descriptions via prompting. Going beyond these steps, the proposed pipeline includes human validation as well as fine-tuning steps to re-evaluate and improve model performance in terms of CR. This design deliberately goes beyond knowledge, supports mixed-culture prompts, and provides pre and post metrics tied to human judgements.

We argued that CR is necessary for identity-aware AI because moral norms alone are insufficient and because real users often inhabit blended cultural contexts. We separated cultural bias (distributional skew that pushes models toward dominant frames) from CR (the ability to apply the right frame), while

noting that bias directly impairs CR. We also clarified why multilingual transfer does not guarantee cultural transfer, as language form can move across tongues more readily than nuanced cultural priors such as politeness strategies, norm salience, or situated moral trade-offs.

Overall, this paper should provide a practical path to measure and improve cultural reasoning in LLMs. The approach treats CR as testable behaviour grounded in validated cultural descriptions and value statements, enables assessment in mixed-identity scenarios, and yields actionable signals for data curation and fine-tuning. This positions identity-aware AI not as a vague aspiration but as an achievable target: models that reliably adapt to users' cultural contexts, while preserving clarity, safety, and utility.

Acknowledgments

This paper was written as part of the UniGR-Guest Professorship project "Cultural Reasoning in AI: Examining Language Models in the Context of the Greater Region", awarded and funded by the University of the Greater Region. We thank UniGR for its support.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-5 and LanguageTool for grammar, style and spell checks. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language Models Are Unsupervised Multitask Learners (2019).

[2] OpenAI, J. Achiam, S. Adler, et al., GPT-4 Technical Report, 2024. URL: https://arxiv.org/abs/2303.08774. arXiv:2303.08774.

[3] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).

[4] A. Üstün, V. Aryabumi, Z.-X. Yong, W.-Y. Ko, D. D'souza, G. Onilude, N. Bhandari, S. Singh, H.-L. Ooi, A. Kayid, et al., Aya model: An instruction finetuned open-access multilingual language model, arXiv preprint arXiv:2402.07827 (2024).

[5] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, Journal of Machine Learning Research 21 (2020) 1–67.

[6] S. R. Schecter, Language, culture, and identity, in: The Routledge handbook of language and culture, Routledge, 2014, pp. 196–208.

[7] S. L. Blodgett, S. Barocas, H. D. III, H. Wallach, Language (Technology) is Power: A Critical Survey of "Bias" in NLP, in: Proceedings of ACL, 2020. URL: https://aclanthology.org/2020.acl-main.485/. doi:10.18653/v1/2020.acl-main.485.

[8] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?, in: Proceedings of FAccT, 2021, pp. 610–623. URL: https://dl.acm.org/doi/10.1145/3442188.3445922. doi:10.1145/3442188.3445922.

[9] P. Joshi, S. Santy, A. Budhiraja, K. Bali, M. Choudhury, The State and Fate of Linguistic Diversity and Inclusion in the NLP World, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of ACL, 2020. URL: https://aclanthology.org/2020.acl-main.560/. doi:10.18653/v1/2020.acl-main.560.

[10] Y. Tao, O. Viberg, R. S. Baker, R. F. Kizilcec, Cultural Bias and Cultural Alignment of Large Language Models, PNAS Nexus 3 (2024). doi:10.1093/pnasnexus/pgae346.

[11] A. Karim, A. Karim, B. Lohana, M. Keon, J. Singh, A. Sattar, Lost in Cultural Translation: Do LLMs Struggle with Math Across Cultural Contexts?, 2025. URL: http://arxiv.org/abs/2503.18018. doi:10.48550/arXiv.2503.18018, arXiv:2503.18018 [cs] version: 1.

[12] T. Naous, M. J. Ryan, A. Ritter, W. Xu, Having Beer after Prayer? Measuring Cultural Bias in Large Language Models, 2024. URL: https://aclanthology.org/2024.acl-long.862/. doi:10.18653/v1/2024.acl-long.862.

[13] T. Hu, Y. Kyrychenko, S. Rathje, N. Collier, S. van der Linden, J. Roozenbeek, Generative language models exhibit social identity biases, Nature Computational Science 5 (2025) 65–75. URL: https://www.nature.com/articles/s43588-024-00741-1. doi:10.1038/s43588-024-00741-1.

[14] D. Van Niekerk, M. Peréz-Ortiz, J. Shawe-Taylor, D. Orlic, J. Kay, N. Siegel, K. Evans, N. Moorosi, T. Eliassi-Rad, L. M. Tanczer, et al., Challenging systematic prejudices: An investigation into bias against women and girls (2024).

[15] D. Hovy, S. L. Spruit, The Social Impact of Natural Language Processing, in: Proceedings of ACL, 2016. URL: https://aclanthology.org/P16-2096/. doi:10.18653/v1/P16-2096.

[16] A. Paullada, I. D. Raji, E. M. Bender, E. Denton, A. Hanna, Data and its (Dis)Contents: A Survey of Dataset Development and Use in Machine Learning Research, Patterns 2 (2021) 100336. URL: https://www.cell.com/patterns/fulltext/S2666-3899(21)00173-5. doi:10.1016/j.patter.2021.100336.

[17] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, et al., Ethical and Social Risks of Harm from Language Models, 2021. URL: https://arxiv.org/abs/2112.04359. arXiv:2112.04359.

[18] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, et al., On the Opportunities and Risks of Foundation Models, 2021. URL: https://arxiv.org/abs/2108.07258. arXiv:2108.07258.

[19] J. Kharchenko, T. Roosta, A. Chadha, C. Shah, How Well Do LLMs Represent Values Across Cultures? Empirical Analysis of LLM Responses Based on Hofstede Cultural Dimensions, 2024. URL: http://arxiv.org/abs/2406.14805. doi:10.48550/arXiv.2406.14805, arXiv:2406.14805 [cs] version: 1.

[20] E. Karinshak, A. Hu, K. Kong, V. Rao, J. Wang, J. Wang, Y. Zeng, LLM-GLOBE: A Benchmark Evaluating the Cultural Values Embedded in LLM Output, 2024. URL: http://arxiv.org/abs/2411.06032. doi:10.48550/arXiv.2411.06032, arXiv:2411.06032 [cs].

[21] Y. Y. Chiu, L. Jiang, B. Y. Lin, C. Y. Park, S. S. Li, S. Ravi, M. Bhatia, M. Antoniak, Y. Tsvetkov, V. Shwartz, Y. Choi, CulturalBench: A Robust, Diverse and Challenging Benchmark for Measuring LMs' Cultural Knowledge Through Human-AI Red-Teaming, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Proceedings of ACL, 2025. URL: https://aclanthology.org/2025.acl-long.1247/. doi:10.18653/v1/2025.acl-long.1247.

[22] A. Mushtaq, I. Taj, R. Naeem, I. Ghaznavi, J. Qadir, WorldView-Bench: A Benchmark for Evaluating Global Cultural Perspectives in Large Language Models, 2025. URL: http://arxiv.org/abs/2505.09595. doi:10.48550/arXiv.2505.09595, arXiv:2505.09595 [cs].

[23] C. C. Liu, F. Koto, T. Baldwin, I. Gurevych, Are Multilingual LLMs Culturally-Diverse Reasoners? An Investigation into Multicultural Proverbs and Sayings, 2024. URL: http://arxiv.org/abs/2309.08591. doi:10.48550/arXiv.2309.08591, arXiv:2309.08591 [cs].

[24] M. F. Adilazuarda, S. Mukherjee, P. Lavania, S. S. Singh, A. F. Aji, J. O'Neill, A. Modi, M. Choudhury, Towards Measuring and Modeling "Culture" in LLMs: A Survey, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of EMNLP, 2024. URL: https://aclanthology.org/2024.emnlp-main.882/. doi:10.18653/v1/2024.emnlp-main.882.

[25] F. Stalder, The Digital Condition, John Wiley & Sons, 2018.

[26] L. S. Kim, Exploring the Relationship Between Language, Culture and Identity, GEMA Online Journal of Language Studies 3 (2003). Publicly Available Content Database.

[27] V. S. Subrahmanian, M. Albanese, M. V. Martinez, D. Nau, D. Reforgiato, G. I. Simari, A. Sliva, O. Udrea, J. Wilkenfeld, CARA: A Cultural-Reasoning Architecture, IEEE Intelligent Systems 22 (2007) 12–16. URL: https://doi.org/10.1109/MIS.2007.25. doi:10.1109/MIS.2007.25.

[28] C. C. Liu, I. Gurevych, A. Korhonen, Culturally aware and adapted nlp: A taxonomy and a survey of the state of the art, TACL 13 (2025) 652–689.

[29] S. Münker, Cultural Bias in Large Language Models: Evaluating AI Agents through Moral Questionnaires, in: Proceedings of 0 th Moral and Legal AI Alignment Symposium, 2025, p. 61.
[30] A. Ramezani, Y. Xu, Knowledge of cultural moral norms in large language models, in: Proceedings of ACL, 2023, pp. 428–446.
[31] M. Sap, R. L. Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, Y. Choi, Social Chemistry 101: Learning to Reason about Social and Moral Norms, in: Proceedings of EMNLP, 2020. URL: https://aclanthology.org/2020.emnlp-main.48/. doi:10.18653/v1/2020.emnlp-main.48.
[32] A. S. Rao, A. Yerukola, V. Shah, K. Reinecke, M. Sap, NormAd: A framework for measuring the cultural adaptability of large language models, in: Proceedings of NAACL-HLT, 2025, pp. 2373–2403.
[33] M. A. Hedderich, L. Lange, H. Adel, J. Strötgen, D. Klakow, A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of NAACL-HLT, 2021. URL: https://aclanthology.org/2021.naacl-main.201/. doi:10.18653/v1/2021.naacl-main.201.
[34] S. L. Blodgett, L. Green, B. O'Connor, Demographic Dialectal Variation in Social Media: A Case Study of African-American English, in: J. Su, K. Duh, X. Carreras (Eds.), Proceedings of EMNLP, 2016. URL: https://aclanthology.org/D16-1120/. doi:10.18653/v1/D16-1120.
[35] T. Naous, W. Xu, On The Origin of Cultural Biases in Language Models: From Pre-training Data to Linguistic Phenomena, in: L. Chiruzzo, A. Ritter, L. Wang (Eds.), Proceedings of NAACL-HLT, 2025. URL: https://aclanthology.org/2025.naacl-long.326/. doi:10.18653/v1/2025.naacl-long.326.
[36] L. H. Zhang, S. Milli, K. Jusko, J. Smith, B. Amos, Wassim, Bouaziz, M. Revel, J. Kussman, L. Titus, B. Radharapu, J. Yu, V. Sarma, K. Rose, M. Nickel, Cultivating Pluralism In Algorithmic Monoculture: The Community Alignment Dataset, 2025. doi:10.48550/arXiv.2507.09650. arXiv:2507.09650.
[37] F. Schneider, C. Holtermann, C. Biemann, A. Lauscher, GIMMICK – Globally Inclusive Multimodal Multitask Cultural Knowledge Benchmarking, 2025. URL: http://arxiv.org/abs/2502.13766. doi:10.48550/arXiv.2502.13766, arXiv:2502.13766 [cs].
[38] A. Sorokovikova, P. Chizhov, I. Eremenko, I. P. Yamshchikov, Artificial Intelligence Gives Women Lower Salary Advice, Technical University of Applied Sciences Würzburg–Schweinfurt press release, 2025. https://www.thws.de/en/research/institutes/cairo/releases/thema/artificial-intelligence-gives-women-lower-salary-advice/.
[39] V. Hofmann, P. R. Kalluri, D. Jurafsky, S. King, AI Generates Covertly Racist Decisions about People Based on Their Dialect, Nature 633 (2024) 147–154. doi:10.1038/s41586-024-07856-5.
[40] K. Khandelwal, M. Tonneau, A. M. Bean, H. R. Kirk, S. A. Hale, Indian-bhed: A dataset for measuring india-centric biases in large language models, in: Proceedings of the 2024 International Conference on Information Technology for Social Good, 2024, pp. 231–239.
[41] R. Rogers, AI Is Spreading Old Stereotypes to New Languages and Cultures, *Wired*, 2025. URL: https://www.wired.com/story/ai-stereotypes-new-languages-cultures/, accessed August 27, 2025.
[42] OpenAI, Reasoning models – OpenAI API Documentation, 2025. URL: https://platform.openai.com/docs/guides/reasoning.
[43] S. Ehrhart, F. Fehlen, Luxembourgish: A success story? a small national language in a multilingual country, in: J. A. Fishman, O. García (Eds.), Handbook of Language and Ethnic Identity, Oxford University Press, 2011, pp. 285–298.
[44] N. Zhou, D. Bamman, I. L. Bleaman, Culture is Not Trivia: Sociocultural Theory for Cultural NLP, 2025. URL: http://arxiv.org/abs/2502.12057. doi:10.48550/arXiv.2502.12057, arXiv:2502.12057 [cs].
[45] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, M. S. Bernstein, Generative Agents: Interactive Simulacra of Human Behavior, in: Proceedings of UIST, 2023. URL: https://dl.acm.org/doi/10.1145/3586183.3606763. doi:10.1145/3586183.3606763.

[46] B. Wang, Z. Liu, X. Huang, F. Jiao, Y. Ding, A. Aw, N. Chen, SeaEval for Multilingual Foundation Models: From Cross-Lingual Alignment to Cultural Reasoning, in: K. Duh, H. Gomez, S. Bethard (Eds.), Proceedings of NAACL-HLT, 2024. URL: https://aclanthology.org/2024.naacl-long.22/. doi:10.18653/v1/2024.naacl-long.22.

[47] C. Li, M. Chen, J. Wang, S. Sitaram, X. Xie, Culturellm: Incorporating cultural differences into large language models, Advances in Neural Information Processing Systems 37 (2024) 84799–84838.

[48] Z. W. Lim, A. F. Aji, T. Cohn, Language-Specific Latent Process Hinders Cross-Lingual Performance, arXiv preprint arXiv:2505.13141 (2025).

[49] Z.-X. Yong, C. Menghini, S. H. Bach, Low-Resource Languages Jailbreak GPT-4, 2024. URL: http://arxiv.org/abs/2310.02446. doi:10.48550/arXiv.2310.02446, arXiv:2310.02446 [cs].

[50] A. Sinelnik, D. Hovy, Narratives at Conflict: Computational Analysis of News Framing in Multilingual Disinformation Campaigns, in: X. Fu, E. Fleisig (Eds.), Proceedings of ACL Student Workshop, 2024. URL: https://aclanthology.org/2024.acl-srw.21/. doi:10.18653/v1/2024.acl-srw.21.