ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240934

Revisiting Under-Represented Knowledge of Latin American Literature in Large Language Models

Jinsung Kim[a,†], **Seonmin Koo**[a,†] **and Heuiseok Lim**[a,*]

[a]Korea University, Department of Computer Science and Engineering
{jin62304,fhdahd,limhseok}@korea.ac.kr

Abstract. With the advent of large language models (LLMs), concerns about knowledge bias have recently increased. Previously, prevalent research has focused on detecting the bias of model knowledge by providing explicit social terms, such as race, gender, and age, into inputs. However, revealing the subtle and implicit bias of the model knowledge requires verification utilizing language expressed in a more implied form, such as literary works. This is because literature implicitly contains subjective filters of individuals and their living regional culture. Accordingly, this study aims to probe a research question of *whether LLMs have a knowledge under-representation problem between two different regions using the same language, Spain and Spanish-speaking countries in Latin America.* To this end, we design an under-representation verification task, **REG**ion and **L**iterary **A**uthor prediction (**REGLA**) and dataset based on Spanish-written literary works. Inspired by the knowledge shortcut concept from a previous study, **REGLA** consists of two tasks to figure out meta-information of poems, i.e., region and author. Moreover, we explore various prompting methods that can unleash the knowledge observed to be under-represented within the verification process. According to the verification and prompt engineering results, knowledge about the literary works of Latin American countries appears to be more under-represented compared to those of Spain in LLMs. It is also observed that the task decomposition prompting method effectively lets under-represented knowledge be generated.

1 Introduction

With the advent of the era of large language models (LLMs), concerns about various biases and stereotypes have been growing [6, 9]. Existing studies related to knowledge bias in language models have focused on detecting the bias of plain textual data in an explicit manner. In detail, several studies analyze the output from the models by providing input prompts containing a set of terms or human-written sentences that explicitly include social information such as race, gender, and age [26], and mainly utilize commonsense sentences or daily conversations extracted from Reddit [3, 19].

However, bias verification studies through daily sentences, providing explicit triggers in the input, often either handle only superficial or shallow semantic biases [21]. This can lead to overlooking subtle and implicit biases in the knowledge representation inherent in language models. For example, knowledge over- or under-representation on differences in perspectives depending on the region between the same language-speaking groups may need to be captured through implicit-way verification, as the implied values differ despite the identical language on the surface. From this perspective, implicit and implied textual data such as literary works can serve as valuable intermediaries for verification in addition to plain and everyday language. It is because literary works have many contexts, such as subtle differences, in an implied form, and verifying whether LLMs can read such 'implications' is advantageous. In particular, literature is sufficient to be used as an implicit verification tool, as it implies the following two aspects through expression methods such as metaphors and symbols: the cultural background of the work [10, 11] and the subjective filters of the author in each region [4]. Suppose a

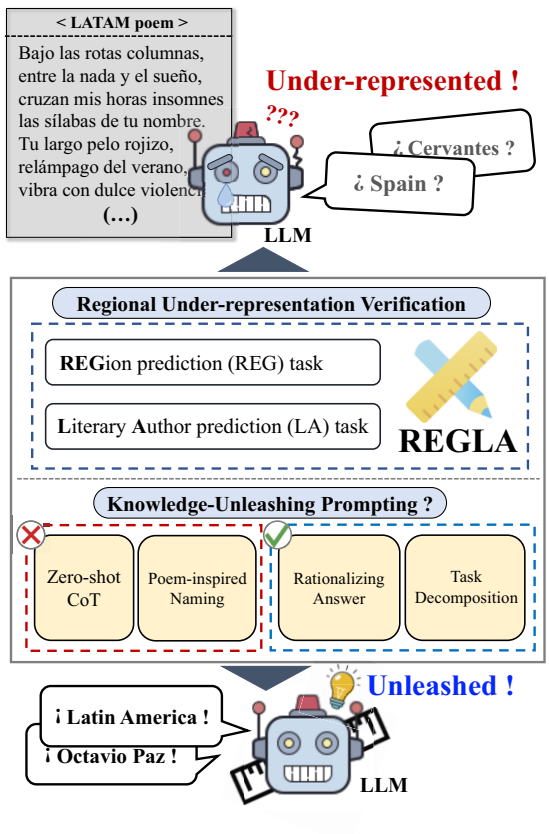

Figure 1. Overview of the designed task (**REGLA**) and the explored prompting methods to unleash the under-represented knowledge in LLMs.

* Corresponding Author. Email: limhseok@korea.ac.kr
† Equal contribution.

contemporary Spanish writer and a LATAM writer each write a poem in Spanish. Although each poem appears to share the same language on the surface, inherent differences in regional culture and individual perspectives may be assumed. Therefore, as a complement to bias studies that provide explicit prompts, it is worth exploring knowledge over- and under-representation in the LLMs using implicit textual data like literary works as bias triggers.

In this work, we first verify LLMs' knowledge bias by focusing on the following research question (RQ) and then attempt to alleviate the problem with prompt engineering:

- RQ) *"Are literary works from the Latin American region under-represented compared to those from Spain, despite using the same language: Spanish?"*

In other words, this study consists of two phases–problem verification and alternative exploration, as illustrated in Figure 1.

Accordingly, we introduce a **REG**ion and **L**iterary **A**uthor prediction task (**REGLA**), which is composed of two tasks for black-box probing into regional knowledge under-representation of LLMs. The proposed **REGLA** contains region prediction (Task 1: REG task) and literary author prediction (Task 2: LA task). We are inspired by the concept of *knowledge shortcut* in Chen et al. [5]'s work, in which LLMs readily and frequently lead to the answers biased to over-represented knowledge, e.g., positive over negative knowledge. Based on this *knowledge shortcut* phenomenon, this study seeks to discern under-represented knowledge through literary author and on-regional verification for different regions. Moreover, a verification dataset to perform **REGLA** is devised, which is constructed with various literary works from Spanish-speaking regions, i.e., Spain and Latin American countries (LATAM). With **REGLA**, we aim to address knowledge under-representation regarding the following two perspectives: 1) regional and broader implications are verified by predicting the region of literary works with task 1 (REG), and 2) task 2 (LA) inspects more specific and individual perspectives than task 1 by predicting the specific author. According to the verification results through **REGLA**, LLMs such as ChatGPT [22] are observed to exhibit under-representation regarding Latin American poems compared to those of Spain (§ 5).

In addition, we investigate various prompt engineering approaches to determine which prompting approach is to be effective in unleashing the under-represented knowledge detected with the proposed **REGLA**. Including the vanilla prompting method with basic task instruction, several reasoning-enhanced prompting methods, such as task decomposition prompting [16], are experimentally explored. The purpose of this exploration is to find out the prompting methods that let LLMs maintain the balance of generated answers by alleviating the knowledge shortcut phenomenon on over-represented knowledge and by effectively inducing the under-represented knowledge that LLMs are aware of but cannot actively express in response generation situations. In other words, the investigation into prompting approaches aims to mitigate implicit bias according to region, not simply maximizing task performance for all regions. In the mitigation experiments with several prompting methods, the task decomposition method demonstrates the most balanced performance in the **REGLA**, narrowing the performance gap in inferring between different countries.

Our contributions are threefold: (1) We devise a **REGLA** to verify the under-representation problem of LLMs between the Spanish-speaking regions (Spain and LATAM) and provide resources for this task; (2) Unlike previous studies focusing on explicit bias detection in cross-language settings, this is the first study to address the implicit knowledge bias according to regional culture among the same Spanish-speaking group; (3) We empirically investigate various prompt engineering approaches, such as multi-step reasoning-based approaches, to alleviate the verified under-representation problem. Also, through extensive analysis, we provide a foundation for research to unleash under-represented knowledge in knowledge-biased situations.

2 Related Works

Language Model Probing The probing approach has been employed to study knowledge captured in language models (LMs) [24]. Understanding and addressing the biases inherent in LMs is crucial for leveraging and advancing them. However, leading LLMs, such as the GPT family, tend to be closed off concerning their training data. For instance, one of the most widely-used models, ChatGPT [22], lacks publicly available training data and is accessible only via API. Furthermore, the explicit confirmation of token generation probabilities for the model is not feasible. Therefore, black-box probing studies utilizing prompting have been conducted to verify the intrinsic knowledge possessed by LLMs [27]. Some studies underscore the importance of manual prompts for probing and offer valuable insights [15]. The commonsense knowledge inherent in the LLMs is probed, as well as the conflicts between external evidence and the model's parametric memory [5, 30]. These studies employ cloze-style prompt designs to control the model's output. The correct answer to the blank to be filled must be clearly defined to enhance the model's interpretability.

Accordingly, for objectivity and interpretability, we employ the author and regional information triggered by intrinsic information from literary works as verification tools. The author incorporates regional identities, customary styles, and symbols in their work. Even in works from the same era, works from different perspectives appear depending on personal background and values, so author information is important considering this. For example, regarding the Spanish Conquest of the Inca, understanding the object shown in the poem must differ depending on whether the writer is for or against it. In other words, knowledge about the life and worldview of a specific author about a specific era and culture contained in the poem can enhance understanding of the poem. Therefore, our proposed **REGLA** can validate under-representation while maintaining objectivity.

Studies on Knowledge Bias in LLMs Recent studies have explored bias and difficulties across multiple languages and cultures, including Spanish LMs. To further explore multilingual embedding models and their impact, España-Bonet and Barrón-Cedeño [8] propose the collection of cultural-aware lists. Hämmerl et al. [12] validate whether pre-trained multilingual language models (PMLMs) that are known for superior performance in English compared to other languages exhibit similar results in capturing moral norms. Ramesh et al. [23] analyze bias from both a linguistic and cultural lens for non-English languages and present a comprehensive overview of the literature on bias pertaining to grammatically gendered languages and multilingualism. Framework has been proposed for developing a multicultural (language) NLP system, taking into account the difficulties in linguistic form, common ground, aboutness, and value across general sentences [13]. These studies encompass a range of languages, such as English, Czech, German, etc., but emphasize cultural differences among diverse language groups rather than intra-language group cultural differences.

Furthermore, previous studies have primarily focused on bias and difficulties apparent in plain textual data, leading to a scarcity of im-

plicit validation. Therefore, we conduct the first comprehensive study focusing on biases arising from cultural differences within the same linguistic domain, using implicit textual data such as literary works.

3 Region and Literary Author Prediction (REGLA)

3.1 Problem Statement: Under-representation

We are inspired by the concept of a *knowledge shortcut* to identify the knowledge bias based on regional features among groups sharing the same language. The *knowledge shortcut* concept indicates that the patterns and co-occurrence that LLM statistically learns more frequently during the pre-training process are more apt to be generated [5]. For example, since most predicates between entities in the training data are described in positive form, LLM is apt to respond in positive form even in inference situations where negative form knowledge must be generated. Inspired by the concept, we aim to reveal which region is not apt to co-occur by LLM with Spanish-written literary works among two regions that share the Spanish language: Spain and LATAM.

Accordingly, this study associates the under-representation problem as a phenomenon in which an LLM cannot sufficiently elicit relevant clues about one rather than the other between the two knowledge types, even though there are no problems with comprehension of language and instructions or recency of pre-trained knowledge. In detail, when a literary work from Spain or LATAM is fed to a model, if it fails to unleash factual knowledge about one region's work, such as the origin, the author's name and his/her nationality, compared to that of the other region, this is referred to as under-representation. Suppose that when predicting the region and author of given poems, LLM generates the most answers as Spain and a Spanish poet, regardless of whether the poems are from Spain or LATAM. This result is to be interpreted as indicating that the LATAM region's literary works are under-represented in the model's inherent knowledge.

Consequently, we associate the under-representation problem with the inability to elicit factual meta-information of a specific work, such as the origin region, author, and his/her nationality. It is because literary works are products that reflect the thoughts of authors influenced by regional cultures [14]. By verifying the under-representation problem, we confirm knowledge differences of LLMs arising from regional differences, even within the same language group.

3.2 Task Design

We introduce **REG**ion and **L**iterary **A**uthor prediction task (**REGLA**) to verify and mitigate implicit under-representation in LLMs. The designed **REGLA** consists of two sub-tasks: region prediction (REG) and literary author prediction (LA), and each task focuses on verifying different perspectives. In both tasks, the model performs cloze-style free-form generation according to each task instruction, and the knowledge under-representation in the model is estimated based on the generated response. Note that model performances are bound to be lower than the multiple-choice setting tasks, as the devised tasks aim to conduct free-form generation.

Denotation. Let the set of poems be P. The set of poems from Spain or LATAM is denoted as P^k ($k \in \{\text{Spain}, \text{LATAM}\}$) and an individual poem that is an i-th component belonging to P^k, is denoted as p_i. Each task has its task instruction I, and this can be I_{REG} or I_{LA} according to the task type. The prompt template mapping function for constructing an input prompt is $T(\cdot)$, and the structure of the mapping function varies according to prompting approach types. Accordingly, the constructed input prompt $T(I, p)$ is fed to LLM, and the generated model output is denoted as $\hat{y}$. Also, the corresponding ground truth label for $\hat{y}$ is denoted as y.

Task 1: Region Prediction (REG). REG task aims to estimate whether the model can implicitly elicit on-regional knowledge from the given poem, such as terms containing characteristics of a cultural region. This task has the model generate the region that better matches the input prompt, containing a Spanish-written poem, as an answer among the two given options: 'Spain' or 'Latin America.' In other words, LLM conducts binary prediction between two Spanish-speaking regions, and the prediction success rate is calculated based on the generated answer for evaluation.

The individual success score C_{REG} for each input and the task total success rate SR_{REG} can be formalized as follows:

$$SR_{\text{REG}}(\%) = \frac{1}{|P^k|} \sum^{|P^k|} C_{\text{REG}}(\cdot) \tag{1}$$

$$\underset{p_i \in P^k}{C_{\text{REG}}(\hat{y} \mid T(I_{\text{REG}}, p_i))} = \begin{cases} 1, & \text{if } y \in \hat{y} \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

where the label y can be given among {"Spain", "LATAM"}.

Task 2: Literary Author Prediction (LA). LA task requires LLM to predict the literary author of a given work and his/her nationality. To this end, the model needs to understand the given poem comprehensively and leverage clues that can help guess the author, integrating the understanding of poetry and its pre-trained knowledge. Task 2 (LA) is more complex than task 1 (REG) because it evaluates performance by generating multiple model outputs.

For evaluation, task 2 (LA) matches the author's name and his/her nationality included in the generated tokens with the ground-truth labels. As same with the Equation (1) of task 1 (REG), the success rate SR_{LA} is estimated based on the individual task score C_{LA}. The formalization of C_{LA} can be as follows:

$$\begin{aligned} &\underset{p_i \in P^k}{C_{\text{LA}}(\hat{y}_{aut}, \hat{y}_{nat} \mid T(I_{\text{LA}}, p_i))} \\ &= \begin{cases} 1, & \text{if } (y_{aut} \in \hat{y}_{aut}) \text{ and } (y_{nat} \in \hat{y}_{nat}) \\ 0, & \text{otherwise,} \end{cases} \end{aligned} \tag{3}$$

where y_{aut} and y_{nat} indicate the ground-truth author and nationality, respectively. For example, when P^k is P^{LATAM}, they have following relations; $y_{aut} \in$ {"Mario Benedetti", "Octavio Paz", ...} and $y_{nat} \in$ {"Uruguay", "Mexico", ...}. Also, if $\hat{y}_{aut}$ is generated as the abbreviation of the ground-truth author's name, it is also treated as a successful answer.

4 Dataset Creation for REGLA

We construct a verification set for conducting **REGLA**, consisting of poems from Spain and LATAM regions. To achieve this, we undergo a data refinement process that includes author/poem selection and meta-information labeling.

4.1 Raw Datasets and License Information

We adopt the raw poem data in the 'Spanish Poetry Dataset' available on Kaggle,[1] which were sourced from Poemas del Alma[2] where

[1] https://www.kaggle.com/datasets/andreamorgar/spanish-poetry-dataset
[2] https://www.poemas-del-alma.com/

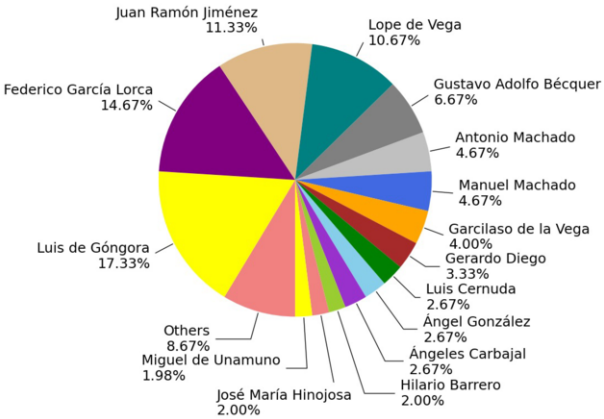

Figure 2. Distribution of poems according to author in Spain poetry data.

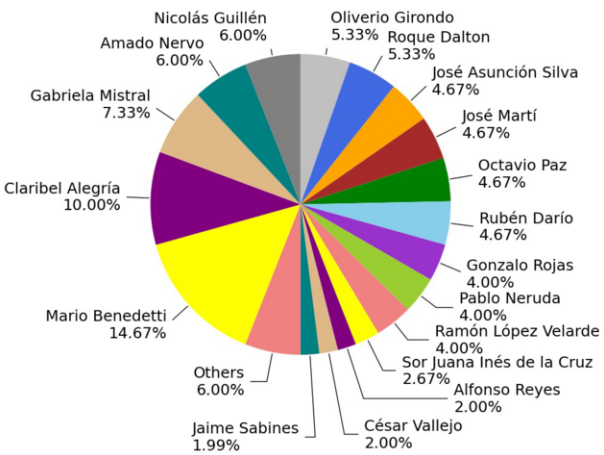

Figure 3. Distribution of poems according to authors in Latin America poetry data.

extensive data related to Spanish poetry are published. Regarding license information, the raw data fall under the GNU Lesser General Public License.[3] We acknowledge the licensing information and utilize the data solely for academic and research purposes, excluding commercial use.

4.2 *Author and Poem Selection*

Selection Process. Regarding the recency of the data on which the model was pre-trained, we confirm LLMs' existing knowledge of all poet authors in order to prevent LLMs from being unable to perform **REGLA** due to a lack of knowledge about the author. To this end, it is investigated whether the model's knowledge about individual authors matches the factual information on English[4] and Spanish Wikipedia[5]. For example, when the query "Who is Amado Nervo, and what is his/her profession?" is provided to the model, the response is evaluated to see whether it matches the facts described on Wikipedia. In addition to this check on factual knowledge, poets suitable for building a verification set are selected by comprehensively considering heuristic features, such as the number of the author's works.

[3] https://www.gnu.org/licenses/lgpl-3.0.html
[4] https://en.wikipedia.org/wiki/Main_Page
[5] https://es.wikipedia.org/wiki/Wikipedia:Portada

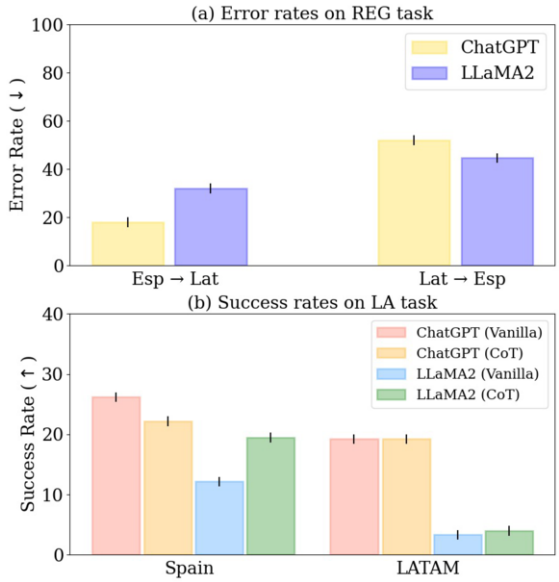

Figure 4. Verification results with **REGLA**. (a) and (b) illustrate the results in task 1 (region prediction) and task 2 (author prediction), respectively.

Afterward, overly lengthy poems that exceed each LLM's maximum input length are excluded. Consequently, A total of 48 authors that can guarantee the models' author awareness, including 24 Spanish authors and 24 Latin American authors, are selected. For example, Rubén Izaguirre Fiallos, a Honduran poet, is excluded from constructing the dataset because LLM mistakenly identifies the writer as an Ecuadorian designer.

Statistics. Figure 2 and Figure 3 respectively illustrate the distribution of selected authors from Spain and LATAM for data annotation. The former group includes Luis de Góngora (17.33%), Federico García Lorca (14.67%), Juan Ramón Jiménez (11.33%), and the latter Mario Benedetti (14.67%), Claribel Alegría (10.00%), resulting in the diversity of selected poets.

4.3 *Data Annotation*

Since the raw poetry dataset lacks meta-information such as literary authors' nationality and the region where a poem was composed, we annotate the dataset with them, leveraging a widely used knowledge source, English and Spanish Wikipedia. For example, Mario Benedetti's poem titled "Señales" is to be labeled with Uruguay and LATAM. We also remove duplicate poems and preprocess errors within the poem, including typos and unnecessary line breaks. Through these processes, we construct a verification dataset encompassing 300 poems from 48 Spanish-speaking authors for conducting **REGLA**.

5 Verification of LLMs with REGLA

5.1 *Experimental Setup*

Models and Prompting Methods. For all experiments, we adopted the version of ChatGPT as gpt-3.5-turbo-1106 by using OpenAI's API, and Meta's LLaMA2-13B-Chat model. Notably, ChatGPT might occasionally generate empty responses due to network transmission timeouts or API overload. In such cases, we followed

the standard practice of resubmitting the request until obtaining non-empty responses.

Hyperparameters. For the hyperparameter setting, we basically follow each model's recommended values by OpenAI and Meta. For the ChatGPT model, we set temperature = 1, top-P = 1, frequency penalty = 0.0, presence penalty = 0.0, and the maximum number of generated tokens = 512. For the LLaMA2-13B-Chat model, we set temperature = 0.75, top-P = 0.9, and the maximum number of generated tokens = 512. The remaining hyperparameters of the models were set to their default values.

5.2 Verification Results

In this section, we investigate the difference in achievement between Spain and LATAM regions with devised **REGLA**. The model's inherent knowledge status is verified in a zero-shot setting without in-context learning to reduce misinstruction when exemplars are given.

Figure 4 demonstrates the verification results in each task of **REGLA**. (a) shows the error rates of each LLM on task 1 (REG). In particular, 'Esp→Lat' refers to the error case where LLM misunderstands the poem of Spain for that of LATAM, and vice versa for 'Lat→Esp'. Additionally, (b) shows the success rates of each LLM according to prompting methods in task 2 (LA). 'Vanilla' indicates a prompting method that only consists of task instruction and input poem, and 'CoT' indicates the zero-shot chain-of-thought (CoT) [17], a representative thinking-style prompting method. The zero-shot CoT enhances the model's reasoning abilities by providing the trigger sentence "Let's think step-by-step.", prompting the model to generate intermediate steps and inference processes necessary for generating the final output autonomously.

LLMs are more likely to mistake poems of LATAM for those of Spain. According to (a), in task 1 (REG), it is observed that both models commonly demonstrate a significantly higher 'Lat→Esp' error rate for given poems, compared to the opposite case (Esp→Lat). In detail, while ChatGPT shows the 'Lat→Esp' error by 52%, it shows the opposite error case by only 18%, exhibiting approximately three times more instances of mistaking Latin American works as Spanish works. For the LLaMA2 model, the 'Lat→Esp' error rate is 44.67%, which is also about 1.4 times higher than the 'Esp→Lat' case.

These results of incorrect cross-selection between the two countries suggest whether the model's response is biased toward knowledge from which region. In other words, it is implied that there is a higher knowledge under-representation of the LATAM region than Spain, even though the regions are both Spanish-speaking.

Does chain-of-thought help alleviate the under-representation? According to (b), similar to the trend of results in (a), each model's success rates in task 2 (LA) are overall lower for the works of LATAM than those of Spain. ChatGPT with vanilla prompting achieves 26.17% and 19.21%, when the given poems are from Spain and LATAM, respectively, resulting in a performance gap of 6.96%p. LLaMA2 with vanilla prompting also shows that the performance on the knowledge of LATAM is approximately 3.7 times lower compared to that of Spain. These results suggest that LLMs have difficulty predicting the ground-truth label, i.e., the author and his/her nationality when literary works of LATAM are provided as input.

Interestingly, employing the zero-shot CoT method has no significant impact on alleviating those performance gaps between the knowledge of the two regions. For example, in the case of LATAM, LLaMA2-13B with the CoT method shows slightly improved achievement by 0.66%p, but ChatGPT shows no change in performance. This observation suggests that the CoT method cannot consistently unleash the under-represented knowledge of the LATAM region. Therefore, more diverse types of prompting methods need to be explored (§ 6).

6 Exploration of Prompts Unleashing Under-represented Knowledge

This section explores various types of natural language-based prompting approaches that have recently become dominant in eliciting inherent knowledge in LLM, for performance improvement of **REGLA**. In particular, since zero-shot CoT, the most famous multi-step prompting method, is not successful in mitigating knowledge under-representation (§ 5), we additionally explore three reasoning-enhanced prompting methods. Only the most effective method in task 1 (REG) is adopted in task 2 (LA) to observe its significance in both tasks.

6.1 Additional Prompting Methods

Poem-inspired Content Generation. With this method, the model generates content considering the given input. In particular, we let it generate fictional names inspired by given poems, as several studies have been conducted to verify the bias of LMs through arbitrary names. For example, model actions vary depending on each given name based on demographics [29, 1]. It is attributed that the names and personal features have an implicit relation based on demographic factors [2] and bias [25]. Following this perspective, we let the model generate fictional names inspired by the given poem and leverage these generated names to infer the appropriate response (i.e., region or literary author). This fictional content generation differs from predicting the author's name which is factual knowledge in task 2 (LA), as it is the result of compressing and expressing the model's subjective understanding inspired by the input poem. The generated contents are integrated as rationales for the final prediction.

Answer Rationalization. This method lets the model generate a justification or explanation along with the target claim to be generated [28]. Encouraging LLMs to rationalize their responses enables deeper inference, enhancing reasoning abilities and interpretability [7]. Therefore, we aim to improve the model's inference capabilities and facilitate the acquisition of explainability, by additionally prompting the model to generate reasons together as follows: *"(· · ·), and explain why."*

Task Decomposition. This method is inspired by LLM's enhanced reasoning skills by decomposing complex tasks or problems into smaller subclaims [16]. We decompose the task question by providing the trigger sentence *"Let's break it down into subcomponents* (· · ·)*."*, dividing the inference process into multi-steps: 1) extracting implicit information such as theme and style, and 2) predicting the region or author (and his/her nationality). Through this, we aim to effectively unleash the under-represented knowledge by delving into the details of the LLM reasoning process.

6.2 Results

Table 1 shows the experimental results with various prompting methods. In the REG task, when the fictional content method is applied, both ChatGPT and LLaMA2 demonstrate the average success rate

Table 1. Performances in task 1: region prediction (REG) and task 2: literary author prediction (LA). 'Avg.' indicates the average score.

Task 1 (REG)		**Spain**	**LATAM**	**Avg.**
		Success Rate (%) (↑)		
ChatGPT	Vanilla	82.00	48.00	65.00
	Fictional Content	29.33	80.00	54.67 (-10.33)
	Rationalization	77.33	66.00	71.67 (+ 6.67)
	Decomposition	70.00	74.00	**72.00 (+ 7.00)**
LLaMA2-13B	Vanilla	68.00	55.33	61.67
	Fictional Content	14.00	95.33	54.67 (-7.00)
	Rationalization	63.33	72.67	68.00 (+6.33)
	Decomposition	64.67	76.00	**70.33 (+ 8.66)**
Task 2 (LA)		**Spain**	**LATAM**	**Avg.**
		Success Rate (%) (↑)		
ChatGPT	Vanilla	26.17	19.21	22.67
	Zero-shot CoT	22.15	19.21	20.67 (-2.00)
	Decomposition	28.86	29.80	**29.33 (+6.66)**
LLaMA2-13B	Vanilla	12.16	3.31	7.67
	Zero-shot CoT	19.46	3.97	11.67 (+4.00)
	Decomposition	18.12	5.96	**12.00 (+4.33)**

decreases sharply. In particular, this method greatly reduces the performance of Spain, implying that it still occurs in biased inference. On the other hand, through rationalization and decomposition (task decomposition) prompting, the average task success rates improve by 6.67%p and 7.00%p, respectively. Also, it is observed that the performance gap between Spain and LATAM has become balanced by leveraging the reasoning-enhance prompting methods.

In the LA task, zero-shot CoT is not consistently helpful for LLM's performance improvement than the Vanilla method, as shown in Section 5. On the contrary, the decomposition method, which is most effective in unleashing under-represented knowledge in task 1 (REG), is shown to be effective for both ChatGPT and LLaMa2-13B. In particular, when the method is adopted, ChatGPT achieves 28.86% for Spain and 29.80% for LATAM, showing an improved average success rate of 6.66%p. LLaMA2 also achieves 18.12% for Spain and 5.96% for LATAM, showing an improved average success rate of 4.33%p. These results indicate that it is feasible to address under-represented knowledge by exploring various prompting methods other than zero-shot CoT and adopting the appropriate strategy.

7 Analysis

In this section, we analyze the under-represented knowledge of Latin American literature from the perspective of LLMs' Spanish language comprehension, and we also provide qualitative analysis to facilitate an extensive understanding.

7.1 Multilingual LLMs' Spanish Understanding

We discuss multilingual LLMs' ability to understand Spanish, reconfirming the validity of the knowledge under-representation in Latin American literature. According to the studies that have experimented with multilingual reasoning tasks in different languages such as English, Spanish, and French, LLMs exhibit the highest performance in English, followed by Spanish, which demonstrates significant performance. [18]. In comprehensive assessments targeting more challenging healthcare domains, Spanish performs second only to English [31]. Similarly, ChatGPT and GPT-4 demonstrate a high level of accuracy and clinical reasoning when addressing questions in the Spanish access exam for specialized medical training [20]. Thus, these confirmations have demonstrated that LLMs have a sufficient understanding of the Spanish language, supporting that lack of language understanding cannot be the cause of the under-representation problem of Latin American works proven in this study.

Table 2. Generated results of ChatGPT with each prompting method for a Latin American poem in task 1 (region prediction).

Input Poem

Bajo las rotas columnas,
entre la nada y el sueño,
cruzan mis horas insomnes
(...)
húmeda costa nocturna
donde se tiende y golpea
un mar sonámbulo, ciego.

Ground-truth Region: Latin America

Vanilla Prompting Output
"This is a poem by the author from Spain."

Fictional Content Prompting Output
Phase 1 (generated fictional contents)
"The name "Liora Ruina" would be a fitting choice inspired by the poem. The name "Liora" feels like a soft and gentle whisper, reflecting the delicate imagery of the poem. "Ruina" not only echoes the idea of brokenness and ruins mentioned (· · ·) captures the essence of the poem's imagery and emotions, making it a suitable and evocative inspiration from the text."

Phase 2 (predicted answer)
→ "It would be more fitting to associate it with Spain."

Rationalization Prompting Output
generated rationale
"The use of vivid imagery and nature-based metaphors, as well as the incorporation of themes like insomnia and the night, align with the characteristics often seen in Latin American poetry. This region's rich cultural and literary traditions often influence the imagery and language used by its poets."

predicted answer
→ "This is a poem by the author from Latin America."

Task Decomposition Prompting Output
Phase 1 (generated reasoning path)
"- Theme: Love and longing - Style: Free verse poetry - Cultural Backgrounds: The poet seems to draw inspiration from romantic and ethereal imagery, possibly influenced by the literary traditions of Spanish poetry and Latin American literature."

Phase 2 (predicted answer)
→ "This is a poem by the author from Latin America."

7.2 Qualitative Analysis

We conduct a qualitative analysis of ChatGPT's generation results for poems from LATAM and Spain, which share the same language, Spanish, but have different cultures. Through the actual generation results according to the prompting method, we aim to examine the model's inherent knowledge about Spanish-written literary works and provide relevant insights.

Task 1: region prediction (REG). The REG task verifies whether the model can implicitly derive on-regional knowledge, such as terms containing regional characteristics, from given poems. Table 2 describes the generation results of ChatGPT according to each prompting method in the REG task for a Latin American poem. With the vanilla prompting method, ChatGPT mistakes the region of the poem

Table 3. Generated results of ChatGPT with each prompting method for a Spanish poem in task 2 (literary author prediction)).

Input Poem

¿Cuándo seré infeliz sin mi gemido?
¿Cuándo sin el ajeno fortunado?
El desprecio me sigue desdeñado;
(...)
Si el hombre es flaco y la ambición es fuerte,
caudal que en desengaños no se emplea,
cuanto se aumenta, Caridón, se vierte.

Ground-truth Author: (Francisco de Quevedo, Spain)

Vanilla Prompting Output
"This is a poem by Francisco de Quevedo"

Zero-shot CoT Output
generated reasoning path
"This poem appears to be written in Spanish, and it shows characteristics of the Spanish Golden Age poetry, particularly the use of classical meter and themes such as love and the human condition. This style aligns with the works of Luis de Góngora, a prominent Spanish writer of the 17th century. The use of complex themes and the intricate use of language in the poem are reminiscent of Góngora's style. (· · ·)"

predicted answer
→ "This is a poem by Luis de Góngora"

Task Decomposition Prompting Output
Phase 1 (generated reasoning path)
"- Tema: El sufrimiento causado por la envidia, el desprecio y la ambición. - Estilo: Soneto con métrica y rima regular, característico de la poesía clásica. - Antecedentes culturales: El poema refleja una visión pesimista de la vida, influenciada posiblemente por la filosofía pesimista de la época en la que fue escrito."

Phase 2 (predicted answer)
→ "This is a poem by Francisco de Quevedo, a Spanish writer and poet, known for his baroque style and sharp wit."

for Spain instead of Latin America. Similarly, the fictional content generation prompting generates incorrect knowledge biased toward Spain, enhancing the knowledge under-representation problem in the fictional content generation process. It can be interpreted that the fictional name is created by focusing on surficial terms or objects rather than considering implicit regional or cultural factors in the poem. On the contrary, with rationalization and task decomposition prompting methods, it can be observed that the answer is generated based on a reasonable interpretation of the poems' region in the intermediate evidence generation process, resulting in prediction success. These results suggest that exploring appropriate reasoning-enhanced prompting approaches can resolve the knowledge bias.

Task 2: literary author prediction (LA). In the LA task, the LLM is required to predict the literary author and his/her nationality of the given work. Table 3 demonstrates the generation results of ChatGPT according to each prompting method in the LA task for a poem by Francisco de Quevedo, a renowned Spanish poet. ChatGPT with vanilla prompting method correctly predicts the author 'Francisco de Quevedo'. However, when applying the zero-shot CoT, which is known for its effectiveness in improving inferential abilities, it can be observed that the incorrect answer is generated. This can be attributed to the inappropriate generation of intermediate reasoning paths. On the other hand, in the case of task decomposition prompting, which also shows considerable performance in task 1 (REG), it generates a correct answer through a deep understanding of the poem by decomposing the complex claim into significant sub-claims such as theme, style, and cultural backgrounds.

Table 4. Performances by language (English-Spanish) in task 2: literary author prediction (LA). 'Spanish' performance in this table includes both results about Spain and the LATAM region. 'Diff.' indicates the difference between performance in English and Spanish.

Task 2 (LA)		Language		Diff.
		English	Spanish	
ChatGPT	Vanilla	62.67	22.67	40.00
	Zero-shot CoT	61.33	20.67	40.66
	Decomposition	64.67	29.33	35.34
LLaMA2-13B	Vanilla	20.33	7.67	12.66
	Zero-shot CoT	18.33	11.67	6.66
	Decomposition	21.33	12.00	9.33

7.3 Comparative Verification with English

We further verify the under-representation of Spanish based on the language that is the reference point for many studies, English. The English dataset for the experiment is also reconstructed in the same manner, following our proposed **REGLA** framework (§ 4) with the raw data from Kaggle's Poetry Analysis with Machine.[6] The created verification data for **REGLA** includes English poems by 48 authors from English-speaking countries (UK, Ireland, USA, etc.), such as William Shakespeare, T.S. Eliot, and W.B. Yeats.

We provide the LLMs with poems in both English and Spanish, prompting the model to generate the regions (task 1) and authors (task 2) of the works as described in Section 3.2. According to Table 4, the overall performance for Spanish works is observed to be remarkably lower than that for English works, demonstrating a performance difference of at least 1.5 times or more. These results may indicate a significant implicit knowledge bias depending on the language, implying a preponderance in the English domain.

8 Conclusion and Future Works

This study focuses on employing literary work as an implied medium of cultural and regional identities to unveil implicit biases in LLMs. To this end, we designed the **REGLA** task, composed of two verification tasks–region prediction (REG) and literary author prediction (LA), and constructed corresponding resources. According to the verification results, it is observed that literary works from the LATAM region are under-represented in LLMs, compared to those from the Spain region. Moreover, we explored several reasoning-enhanced prompting methods to unleash the under-represented knowledge, and the task decomposition method improved accuracy minimizing the performance gap in Spanish and Latin American works. We expect this study to promote future research on verifying and alleviating the under-representation problem in LLMs through implicit textual expression, such as literary works.

In our future work, we will assess a larger amount of advanced LLMs, such as GPT-4, which was not explored due to the constraints posed by the API usage cost for the GPT family. Furthermore, we will expand the target languages and regions. For example, bias verification between Spanish-speaking countries in the LATAM region can be investigated.

[6] https://www.kaggle.com/datasets/ishnoor/poetry-analysis-with-machine-learning/dataPoetry Analysis with Machine Learning dataset

Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425). This work was supported by Institute for Information communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (RS-2024-00398115, Research on the reliability and coherence of outcomes produced by Generative AI). This work was supported by ICT Creative Consilience Program through the Institute of Information Communications Technology Planning Evaluation(IITP) grant funded by the Korea government(MSIT)(IITP-2024-2020-0-01819).

References

[1] H. An and R. Rudinger. Nichelle and nancy: The influence of demographic attributes and tokenization length on first name biases. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 388–401, Toronto, Canada, July 2023. Association for Computational Linguistics.

[2] H. An, Z. Li, J. Zhao, and R. Rudinger. SODAPOP: Open-ended discovery of social biases in social commonsense reasoning models. In A. Vlachos and I. Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1573–1596, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[3] S. Barikeri, A. Lauscher, I. Vulić, and G. Glavaš. RedditBias: A real-world resource for bias evaluation and debiasing of conversational language models. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1941–1955, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.151. URL https://aclanthology.org/2021.acl-long.151.

[4] L. Boroditsky. Does language shape thought?: Mandarin and english speakers' conceptions of time. *Cognitive psychology*, 43(1):1–22, 2001.

[5] J. Chen, W. Shi, Z. Fu, S. Cheng, L. Li, and Y. Xiao. Say what you mean! large language models speak too positively about negative commonsense knowledge. *arXiv preprint arXiv:2305.05976*, 2023.

[6] M. Cheng, E. Durmus, and D. Jurafsky. Marked personas: Using natural language prompts to measure stereotypes in language models. *arXiv preprint arXiv:2305.18189*, 2023.

[7] C.-H. Chiang and H.-y. Lee. A closer look into using large language models for automatic evaluation. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8928–8942, Singapore, Dec. 2023. Association for Computational Linguistics.

[8] C. España-Bonet and A. Barrón-Cedeño. The (undesired) attenuation of human biases by multilinguality. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2056–2077, 2022.

[9] I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, and N. K. Ahmed. Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770*, 2023.

[10] T. C. Guy. Culturally relevant adult education: Key themes and common purposes. *New Directions for Adult and Continuing Education*, 82:93–98, 1999.

[11] T. C. Guy. Culture as context for adult education: The need for culturally relevant adult education. *New directions for adult and continuing education*, 82:5–18, 1999.

[12] K. Hämmerl, B. Deiseroth, P. Schramowski, J. Libovickỳ, C. A. Rothkopf, A. Fraser, and K. Kersting. Speaking multiple languages affects the moral bias of language models. *arXiv preprint arXiv:2211.07733*, 2022.

[13] D. Hershcovich, S. Frank, H. Lent, M. de Lhoneux, M. Abdou, S. Brandl, E. Bugliarello, L. C. Piqueras, I. Chalkidis, R. Cui, et al. Challenges and strategies in cross-cultural nlp. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, 2022.

[14] A. E. Housman. The name and nature of poetry. *(No Title)*, 1933.

[15] J. Hu and R. Levy. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, 2023.

[16] T. Khot, H. Trivedi, M. Finlayson, Y. Fu, K. Richardson, P. Clark, and A. Sabharwal. Decomposed prompting: A modular approach for solving complex tasks. In *The Eleventh International Conference on Learning Representations*, 2022.

[17] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[18] M. T. R. Laskar, M. S. Bari, M. Rahman, M. A. H. Bhuiyan, S. Joty, and J. Huang. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469, Toronto, Canada, July 2023. Association for Computational Linguistics.

[19] Y. Li, G. Zhang, B. Yang, C. Lin, A. Ragni, S. Wang, and J. Fu. HERB: Measuring hierarchical regional bias in pre-trained language models. In Y. He, H. Ji, S. Li, Y. Liu, and C.-H. Chang, editors, *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 334–346, Online only, Nov. 2022. Association for Computational Linguistics.

[20] A. Madrid-García, Z. Rosales-Rosado, D. Freites-Nuñez, I. Pérez-Sancristobal, E. Pato-Cour, C. Plasencia-Rodríguez, L. Cabeza-Osorio, L. León-Mateos, L. Abasolo-Alcázar, B. Fernández-Gutiérrez, et al. Harnessing chatgpt and gpt-4 for evaluating the rheumatology questions of the spanish access exam to specialized medical training. *medRxiv*, pages 2023–07, 2023.

[21] N. Nangia, C. Vania, R. Bhalerao, and S. R. Bowman. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online, Nov. 2020. Association for Computational Linguistics.

[22] OpenAI-Blog. Chatgpt: Optimizing language models for dialogue, 2022. URL https://openai.com/blog/chatgpt/.

[23] K. Ramesh, S. Sitaram, and M. Choudhury. Fairness in language models beyond english: Gaps and challenges. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2061–2074, 2023.

[24] A. Ramezani and Y. Xu. 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)*, pages 428–446, Toronto, Canada, July 2023. Association for Computational Linguistics.

[25] V. Shwartz, R. Rudinger, and O. Tafjord. "you are grounded!": Latent name artifacts in pre-trained language models. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6850–6861. Association for Computational Linguistics, Nov. 2020.

[26] E. M. Smith, M. Hall, M. Kambadur, E. Presani, and A. Williams. "I'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9211, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics.

[27] J. Sun, Y. Tian, W. Zhou, N. Xu, Q. Hu, R. Gupta, J. Wieting, N. Peng, and X. Ma. Evaluating large language models on controlled generation tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3155–3168, 2023.

[28] H. Wang and K. Shu. Explainable claim verification via knowledge-grounded reasoning with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6288–6304, 2023.

[29] R. Wolfe and A. Caliskan. Low frequency names exhibit bias and overfitting in contextualizing language models. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 518–532, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics.

[30] J. Xie, K. Zhang, J. Chen, R. Lou, and Y. Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*, 2023.

[31] Y. H. Yeo, J. S. Samaan, W. H. Ng, X. Ma, P.-S. Ting, M.-S. Kwak, A. Panduro, B. Lizaola-Mayo, H. Trivedi, A. Vipani, et al. Gpt-4 outperforms chatgpt in answering non-english questions related to cirrhosis. *medRxiv*, pages 2023–05, 2023.