Purposefully Lost in Translation: Expanding The Stereotype Content Model for Cross-Cultural Stereotype Erasure

Anonymous ACL submission

Abstract

Stereotype detection offers valuable insights for detecting implicit bias in language models. To mitigate such bias, stereotyping theories have been adopted in various NLP tasks. However, these implementations have primarily focused on English language models. As language models are increasingly applied across diverse languages and cultures, it is crucial to develop a model that addresses the range of stereotypes present in these languages and cultures. In this paper, we propose a framework for expanding the Stereotype Content Model (SCM) beyond the English language, demonstrated through the development and validation of our Korean SCM (KoSCM). We also present a translation framework designed to address the challenges related to data annotation, explore the cross-cultural validity of the SCM by evaluating the model against theory-grounded hypotheses, and introduce a novel method for stereotype erasure. To make the study of stereotyping more accessible to a broader range of researchers, we also present SCM prompting, a set of prompt engineering guidelines for LLMs aimed at stereotype detection. Our proposed CoT prompting improves the performance of LLMs by an average of 18.6%. This study marks the first attempt to implement the SCM in a non-English language and with LLMs, paving the way for research on stereotypes across different languages and models.

1 Introduction

Language models have the capacity to learn and perpetuate biases present in their training data (Bolukbasi et al., 2016; Caliskan et al., 2017; Chang et al., 2019). To tackle this challenge, researchers have focused on identifying and removing explicit biases like hate speech and derogatory language. Recently, however, there has been an increasing focus on mitigating implicit biases, including societal stereotypes that, while not explicitly harmful, can still contribute to reinforcing negative perceptions. A relevant example is the stereotype that portrays Asians as smart. Although this stereotype might seem positive on the surface, it reflects the *model minority stereotype,* which can impose unrealistic expectations and obscure the rich diversity within the community.

One well-established theory of stereotyping is the Stereotype Content Model (SCM). The SCM (Fiske et al., 2002) suggests that when individuals encounter members of an out-group, they evaluate them based on two dimensions: *warmth* and *competence*. The SCM has been utilized in NLP to develop a computational model for identifying stereotypes (Fraser et al., 2021; Herold et al., 2022; Nicolas and Caliskan, 2024; Schuster et al., 2024; Fraser et al., 2024; Mina et al., 2024), to reduce stereotypical bias in language models (Omrani et al., 2023; Ungless et al., 2022; Gaci et al., 2023), and to enhance hate speech detection (Jin et al., 2024). While the SCM is widely adopted in NLP bias studies, little research explores its application beyond English to non-Western cultures.

As language models become more prevalent across cultures, the importance of detecting stereotypes in different languages is increasing. However, expanding the SCM presents challenges, particularly the cost of data annotation. Translating stereotypes requires both a social psychology expert and a language specialist. Another challenge is the high cost of developing an NLP model. Not everyone possesses the skills or tools required to build and train large language models (LLMs).

Building on social psychology research that proposes the potential of the SCM as a pancultural measure of stereotypes (Cuddy et al., 2009), we propose our comprehensive framework for implementing a cross-cultural SCM that addresses the challenges of translating the SCM by developing a Korean Stereotype Content Model (KoSCM). We start by compiling a Korean dictionary of warmth-competence seed words. To address the challenge of data annotation, we translate existing English warmth-competence lexicons into Korean using a machine translation model, which we then validate with the assistance of an expert translator. Then, we generate sentences incorporating the translated lexicons to create a training dataset for the KoSCM. To tackle the issue of data scarcity, we employ a data augmentation technique called back-translation.

We further illustrate the implementation of the SCM model using the curated dataset and evaluate KoSCM through a stereotype analysis on social groups of age, gender, and religion in Korean texts. Grounded in the theoretical framework (Cuddy et al., 2009), we outline three hypotheses the model must satisfy to confirm that it accurately represents the SCM: (1) the two dimensions hypothesis, (2) the ambivalent stereotypes hypothesis, and (3) the social structural correlates hypothesis. Additionally, we present a novel stereotype erasure method to remove a stereotype dimension from the KoSCM space.

To tackle the challenge of expensive NLP model development and improve the accessibility of SCM applications for a broader range of researchers, we present guidelines for SCM prompting. Our experiments with LLM prompting investigate zero-shot learning, in-context learning (ICL), and Chain-of-Thought (CoT) prompting in both English and Korean models. We find that the best performance is achieved through CoT prompting, particularly when definitions of warmth and competence are provided and when at least one demonstration includes examples of seed words used to derive the warmth/competence dimensions. We offer a website for CoT prompt generation[1].

We summarize our contributions listed above as follows:

- We present the first attempt at developing a framework to expand the stereotype content model into another language and culture.
- We develop KoSCM by following the proposed data curation steps that overcome the data annotation challenge and assess its validity as SCM using a theory-grounded method. We further present a stereotype erasure technique that can be utilized for bias mitigation.
- We identify that warmth-competence prediction is a challenging task for LLMs and provide SCM prompting guidelines to encourage broader application of stereotype analysis.

[1] `anonymized link` (See Figs. 5 and 6 for screenshots of the website.)

2 Background and Related Work

In this section, we examine research on stereotyping in social psychology (§2.1) and computational methods for detecting stereotypes (§2.2).

2.1 Stereotyping in Social Psychology

Stereotyping is a cognitive process in which specific attributes are overly generalized to entire social groups. It is a ubiquitous phenomenon that contributes to the perpetuation of social inequalities. When specific qualities are attributed to entire groups, it reinforces existing power dynamics and legitimizes discriminatory practices.

Social stereotypes are complex and multifaceted constructs that influence social perception and interaction. Traditional approaches to understanding stereotypes have relied on simplistic categorizations, such as positive or negative. However, the Stereotype Content Model (SCM) (Fiske et al., 2002; Fiske, 2018) offers a more nuanced framework for understanding social stereotypes. The SCM posits that social perception is guided by two fundamental dimensions: *warmth* and *competence*. Warmth refers to the perceived intentions and friendliness of a group, while competence refers to the perceived abilities and effectiveness of a group. These dimensions are orthogonal, allowing for the possibility of positive stereotypes along one dimension and negative stereotypes along the other.

A natural follow-up question for researchers is whether these stereotype studies can be generalized across cultures. Given that stereotypes arise from fundamental human phenomena—namely, the need to distinguish between “friends” and “foes” and the ubiquity of hierarchical status differences and resource competition—it is reasonable to assume that these principles are universally applicable.

To investigate this hypothesis, Cuddy et al. (2009) conducted a cross-cultural study spanning seven European (individualist) and three East Asian (collectivist) nations. Their findings suggest that the SCM framework is effective across various cultures, reliably indicating group stereotypes based on structural connections with other groups. Building on this social study, we leverage a computational approach to validate their findings by expanding the application of SCM from English to Korean. To the best of our knowledge, this is the

Dim.	Dir.	Num.	Example
W.	high	75	상냥한kind, 친절한friendly
	low	82	냉담한cold, 불친절한unfriendly
C.	high	68	유능한competent, 영리한clever
	low	60	무능한incompetent, 멍청한stupid

Table 1: **Statistics of Translated Korean Seed Words.** The first column denotes dimensions: warmth and competence, while the second column indicates their direction. The third column lists the number of data points. The final column provides example Korean seed words.

first work to study the SCM in a non-English language, non-western cultural setting.

2.2 Stereotype Content Model in NLP

The SCM has been extensively employed in various NLP applications to identify and mitigate stereotypical biases. For instance, researchers have utilized the SCM to detect stereotype subspaces in word embeddings (Fraser et al., 2021) and debias models by removing stereotype dimensions from the embedding space (Ungless et al., 2022; Omrani et al., 2023). Moreover, the SCM has been applied to assess benchmark datasets for bias (Fraser et al., 2021), examine how NLP models relate SCM dimensions to marginalized groups (Herold et al., 2022; Mina et al., 2024), and develop metrics to investigate biases across demographic and intersectional groups (Cao et al., 2022). Recent studies have further refined the SCM by exploring the construct differentiability of direction and representativeness for warmth and competence dimensions (Nicolas and Caliskan, 2024) and fine-graining stereotype dimensions into six psychologically-motivated categories to study occupation-related stereotypes (Fraser et al., 2024).

In recent years, researchers in NLP have expanded the study of bias and fairness to include non-English languages (Zhou et al., 2019; Chávez Mulsa and Spanakis, 2020; Kurpicz-Briki, 2020; Lauscher et al., 2020; Liang et al., 2020; Moon et al., 2020; Pujari et al., 2020; Takeshita et al., 2020; Zhao et al., 2020; Malik et al., 2021; Jeong et al., 2022), mirroring developments in social psychology. The SeeGULL dataset (Bhutani et al., 2024) has broadened its linguistic scope by introducing a multilingual stereotype dataset featuring 20 languages from 23 different regions. It includes Korean, but differs from our work in that it consists of pairs of associations between an identity term and an attribute generated by a language model. In contrast, our dataset and method are based on stereotyping theory from social psychology, utilizing seed words to identify stereotypes. This approach allows for broader applicability to various identity terms and social groups. To the best of our knowledge, this is the first attempt to expand the SCM lexicons to a different language.

3 Translating Stereotype

This section presents a framework for expanding the SCM to a different language. Four steps are followed to translate English SCM to Korean and create the dataset for KoSCM[2].

Step 1. Extract seed words The first step is to extract seed words for the stereotype content dictionary (Nicolas et al., 2019). The stereotype content dictionary is a collection of theory-driven seed words used to measure sociability, morality/trustworthiness, ability, status, assertiveness/dominance, and political and religious beliefs in relation to social groups. The list contains 341 words with their respective theoretical direction, either *high* or *low*, on their relevant dimension.

From the list, we select seed words that reflect warmth and competence dimensions. Specifically, words in sociability and morality categories are classified as warmth seed words, and those in ability and agency are classified as competence seed words. There are a total of 157 seed words associated with the warmth dimension and 128 for the competence dimension. Each seed word is labeled with a direction within its respective dimension. For example, the word "warm" is a high-direction seed word in the warmth dimension, whereas "cold" represents a low-direction seed word within the same dimension. Similarly, the word "competent" is an example of a high-direction seed word in the competence dimension, while "incompetent" is classified as having low direction in that dimension.

Step 2. Translate seed words Next, the extracted seed words are translated into Korean. The first step of translation is to adopt a machine translation model. We chose Naver Papago (Naver, 2025), one of the most popular Korean-English AI translators in Korea, to translate English seed words to Korean. Afterward, we validate the translation with an expert translator. The translator is asked to validate

[2] The dataset is available in `anonymized link`.

the translation by answering the following questions: (1) Is the translation grammatically correct (e.g., a noun is translated as a noun)? (2) Is a word translated into a distinct word (i.e., no recurrence in the translated list)? Through validation, we verify 285 Korean seed words labeled with stereotype dimension and direction in their corresponding dimension. Table 1 presents statistics and examples of seed words.

Step 3. Generate sentences with seed words With the translated stereotype seed words, we generate sentences based on a template. Similar to May et al. (2019), sentences are generated by inserting individual seed words from the list of Korean stereotype words into simple templates such as “그 사람은 <seed word> 사람이다” (That person is a[n] <seed word> person). The templates are selected according to the part-of-speech (POS) tagging of the seed words. Further, the template words are chosen carefully to prevent the generated sentences from referencing specific social groups. For example, the pronouns “he” and “she” indicate a person’s gender. We intentionally refrain from using these pronouns as subjects because we aim to create a dataset centered on understanding the dimensions of warmth and competence. For more details, see Appendix A.

Step 4. Augment data with back-translation To tackle the limitation of available Korean seed words and address challenges associated with low-resource scenarios, we utilize data augmentation. Sentences generated in Step 3 are augmented using back-translation (Sennrich et al., 2016; Domhan and Hieber, 2017; Belinkov and Bisk, 2018). Back-translation generates paraphrases by leveraging translation models. Initially, a text is translated into another language (forward translation) and then translated back into the original language. This process creates paraphrased sentences, introducing greater variety by allowing for diverse choices in terminology and sentence structure. While the content remains intact, stylistic features that reflect the author’s specific traits may be adjusted or omitted during translation.

For our dataset, we first translate the Korean sentences from Step 3 into English and then translate them back into Korean. We use the No Language Left Behind model (Team et al., 2022), a multilingual model that supports translation for 202 languages. This model is selected for two key reasons. First, it was designed to assist with low-resource language translations. Second, it supports both Korean and English languages. As a result of the back-translation, we obtain a dataset containing 10,260 sentences.

4 Korean Stereotype Content Model

In this section, we detail how the KoSCM dataset, collected through the four steps of the stereotype translation framework, is utilized to build the SCM model. By fine-tuning a model with the dataset, we build KoSCM, which predicts the warmth and competence scores of given Korean sentences.

4.1 Method

We suggest a systematic method to develop a SCM model specific to the language model employed. We utilize an embedding model as its base, adding two classifiers on top. Each classifier predicts the directions of a given text in the warmth and competence dimensions, respectively. Namely, the two classifiers perform multi-class classification, identifying one of three potential directions: high, low, or none. Formally, we use two classifiers, f_w and f_c, to predict warmth and competence directions, respectively. These prediction tasks are formulated as multi-class classification problems with cross-entropy losses, $\mathcal{L}_w$ and $\mathcal{L}_c$; $\mathcal{L}_w = -\sum_{t \in D} W(t) \cdot \log(f_w(t))$ and $\mathcal{L}_c = -\sum_{t \in D} C(t) \cdot \log(f_c(t))$, where t is a text in the dataset D, and $W(t)$ and $C(t)$ are warmth and competence directions of the text t. The final loss of the model is the sum of the prediction losses: $\mathcal{L} = \alpha\mathcal{L}_w + \beta\mathcal{L}_c$, where α and β are hyperparameters.

4.2 Experimental Setup

We evaluate the proposed methods on the following models: (1) **Multilingual BERT (mBERT)**, BERT (Devlin et al., 2019) pre-trained on 104 languages with 110M parameters, (2) **Multilingual Sentence Transformer (mST)**, a modification of the Sentence Transformer (Reimers and Gurevych, 2019) aimed at adapting it for a new language using multilingual knowledge distillation, and (3) **Multilingual RoBERTa (mRoBERTa)** (Conneau et al., 2020), a multilingual version of RoBERTa pre-trained on 100 languages. See Appendix B for further details of the experimental settings.

4.3 Evaluation

Using our proposed method, we evaluate how effectively models trained on the KoSCM dataset

Model	Warmth	Competence
mBERT	**0.923** (0.006)	**0.938** (0.005)
mST	0.917 (0.010)	0.924 (0.006)
mRoBERTa	0.859 (0.023)	0.863 (0.010)

Table 2: **Evaluation of KoSCM.** The average accuracy (and standard deviation) of warmth and competence predictions are presented.

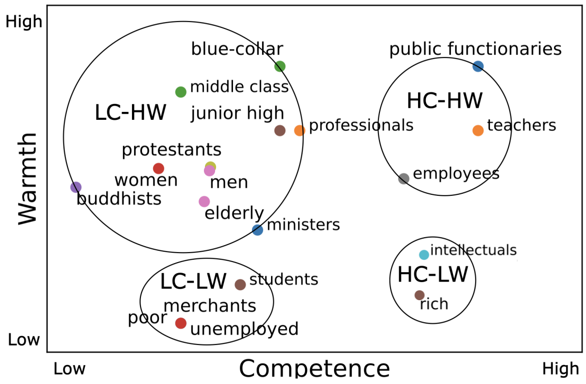


Figure 1: **Stereotypes of Groups Projected to the SCM Dimension.** Social groups are mapped according to their predicted warmth and competence by KoSCM.

predict stereotypes. To assess the effectiveness of these models, we measure the accuracy of warmth and competence prediction on the test data. The results are presented in Table 2, which illustrates both the average and standard deviation of the prediction accuracies. In all three models, we observe competitive performance with high prediction accuracies for both warmth and competence. Notably, mBERT is the best-performing model, achieving accuracies of 0.923 for warmth and 0.938 for competence prediction.

To evaluate the generalization capacity of the KoSCM, we conduct additional tests to determine whether the computational analysis aligns with and supports the results obtained from the SCM survey conducted in South Korea (Cuddy et al., 2009). We leverage the best-performing model, mBERT, from our evaluation to measure the stereotype directions of various social groups. For this analysis, we utilize the Korean Offensive Language Dataset (KOLD) (Jeong et al., 2022). The dataset consists of comments collected from news articles and videos, with labels indicating group information among the 21 target group labels tailored to Korean culture. From the existing group labels, we select 19 groups that intersect with the 23 social groups in the survey and use these for analysis.

We assess the warmth and competence directions of texts that comment on a target group and calculate the average warmth and competence directions. Then, the groups are clustered using hierarchical cluster analysis, following the method of Cuddy et al. (2009). The results are illustrated in the SCM dimension in Figure 1. In general, we observe a significant overlap between our results and the survey findings. For instance, social groups such as "women," "blue-collar," and "Protestants" fall into the low-competence/high-warmth cluster, while groups like the "poor" and "unemployed" are categorized as low-competence/low-warmth. However, there are also outliers. For example, the group "public functionaries" is positioned in the high-competence/high-warmth cluster in our figure, but it falls within the low-competence/low-warmth cluster in the survey plot. This discrepancy may come from the lack of data since outliers like "public functionaries" have only nine text samples contributing to their classification.

4.4 SCM as a Pancultural Tool

We explore the applicability of the proposed computational method of the SCM for analyzing stereotypes across various languages and cultures. Based on the survey in Cuddy et al. (2009), we examine three key hypotheses of SCM: (1) the two dimensions hypothesis, (2) the ambivalent stereotypes hypothesis, and (3) the social structural correlates hypothesis.

Two Dimensions Hypothesis The first hypothesis posits that (1) within each sample, groups will be positioned along the dimensions of warmth and competence and that (2) based on their warmth and competence scores, groups will form multiple clusters, including some at both the high and low ends of each dimension. As shown in Figure 1, our results support this hypothesis, as groups are mapped along the warmth and competence dimensions. The figure reveals a structure that aligns with the SCM survey. We identify four distinct clusters that reflect both high and low scores on each dimension. Consistent with the survey findings, the largest cluster is the low-competence/high-warmth group, which encloses the majority of the sampled groups. Yet we observe that the high-competence/high-warmth cluster in the survey has a lower average warmth score compared to our findings. As discussed in Section 4.3, this dissimilarity may be attributed to outliers, such as the group "public functionaries", which suffered from insufficient data.

Ambivalent Stereotypes Hypothesis This hypothesis proposes that (1) within any given sample, there will be significant variations in perceptions of warmth and competence across different social groups and that (2) it predicts that cluster analyses will reveal at least one high-competence/low-warmth cluster and one low-competence/high-warmth cluster. This indicates that numerous groups are characterized as being adept in one area—either warmth or competence—while being perceived as lacking in the other.

Figure 1 shows four distinct clusters at each end, which supports the hypothesis that the four clusters of stereotype content, defined within the warmth-competence space, have universal characteristics. We observe that the groups "women" and "elderly" fall within the low-competence/high-warmth group. This supports the theory that groups seen as "gentle but useless"—often associated with a "pitying" prejudice—frequently include traditional women and older people (Jackman, 1994; Glick and Fiske, 2001b,a). In contrast, another significant stereotyped group includes those seen as skilled yet dishonest. Our analysis emphasizes individuals labeled as "intellectuals" and "rich" in this group. It shows that "envious" prejudice frequently targets those considered alarmingly skilled yet untrustworthy (Glick and Fiske, 2001b,a; Fiske et al., 2002; Glick, 2002). This dynamic highlights the complex relationship between admiration and disdain influencing societal perceptions.

Social Structural Correlates Hypothesis From the social structural correlates hypothesis, we validate whether perceived competition is anticipated to negatively correlate with warmth. In the survey, participants are asked to evaluate the perceived status and competition of various social groups. As we cannot access the information of commentators in the KOLD dataset, we utilize average wage statistics as a measure of perceived status. Socioeconomic status is a complex construct influenced by multiple factors, with income being a key component (Havranek et al., 2015). Lower-income individuals often experience social disadvantages such as limited access to quality education, poor working conditions, housing insecurity, and unsafe neighborhoods, leading to a reduced perceived status within society (Hernández, 2016; on Civil Rights, 2018). Therefore, we use income as a symbolic indicator of perceived status, highlighting its significant impact on social standing.

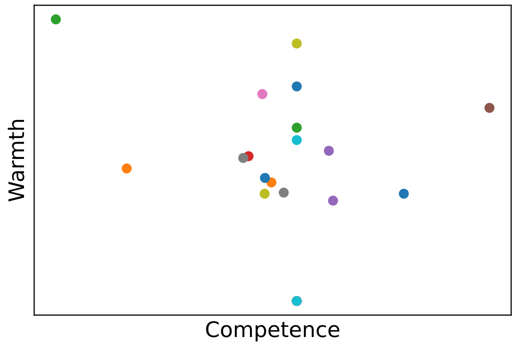


Figure 2: **SCM Dimension after Competence Erasure.** Social groups of Figure 1 after stereotype erasure of competence are mapped above.

The Korean Ministry of Employment and Labor publishes the Current Status of Wage Distribution by Business Characteristics every year[3]. We reference the 2024 report to extract the average income across different social groups. This report offers average wage data categorized by labor industry, gender, and years of experience. Due to the ambiguity in categorizing jobs within non-occupational social groups, e.g., "intellectuals" and "rich," we exclude these groups from this analysis. The report includes gender data for all jobs, so the average income for each gender is computed to represent the perceived status of groups "women" and "men."

Next, we calculate the correlation coefficient between the average wage and competence for the social groups. The correlation coefficient is computed as: $\mathrm{cov}(\text{wage}, \text{competence})/(\sigma_{\text{wage}} \cdot \sigma_{\text{competence}})$. The calculated correlation value is 0.71, a positive correlation that supports the hypothesis. In the survey, South Korea has a correlation of 0.64, and the average of all 13 surveys shows a correlation of 0.79.

4.5 Stereotype Erasure

We propose a stereotype erasure that adopts the least-squares concept erasure (LEACE) (Belrose et al., 2023) to remove a stereotype dimension of the SCM model. LEACE performs concept erasure for linear classifiers by applying a transformation that minimizes the distance between the original and transformed features. Given an input X and a concept Z, LEACE first subtracts the mean and normalizes X; then projects this adjusted value onto the subspace that captures the correlations between X and Z. After that, it reverses the normalization

[3]Ministry of Employment and Labor website

process. Lastly, it subtracts this adjusted value from X, eliminating the linear information that is available about Z.

Formally, $\text{LEACE}(x) = x - W^{+}P_{W\Sigma_{XZ}}W(x - \mathbb{E}[X])$, where W is the whitening transformation $(\Sigma_{XX}^{1/2})^{+}$ and $P_{W\Sigma_{XZ}} = (W\Sigma_{XZ})(W\Sigma_{XZ})^{+}$ is the orthogonal projection onto the column space of $(W\Sigma_{XZ})$.

For the stereotype erasure, we introduce two modifications. Firstly, we present a method for applying stereotype erasure on unseen, unsupervised datasets. Our approach involves extracting stereotype information from the KoSCM dataset. Then, we utilize this learned stereotype information to erase a stereotype dimension in datasets without warmth and competence labels, such as KOLD. This allows us to learn a stereotype direction in the warmth/competence dimension and expand it to unseen data without stereotype information. Secondly, as Z is assumed to be binary, we reassign the KoSCM labels to reflect the presence or absence of directional information within the stereotype dimension because the goal is to eliminate any directional cues. If a text t has either a high or low direction, we assign its label to 1. If there is no direction, we assign it to zero.

Let l_w and l_c be direction labels of warmth and competence of a given text t in the dataset D. For $(t, l_w, l_c) \in D$, $l' = |l|$, where l' is the label for the stereotype erasure and l is a label of the chosen stereotype S for erasure. Then, for a text t' in a target dataset D', the stereotype erasure equation is:

$$g(t') = t' - W^{+}P_{W\Sigma_{DL'}}W(t' - \mathbb{E}[D]) \quad (1)$$

Figure 2 illustrates the result of the stereotype erasure. The proposed method removes competence information from the KOLD data, resulting in a notable shift in the representation of social groups. We see that these groups are now positioned closer to the center of the plot, indicating that their competence scores are nearer to zero, especially when compared to the original depiction in Figure 1. We acknowledge that the method has its limitations, likely due to insufficient training data samples, as indicated by the outliers at the edges of the plot.

5 SCM Prompting for LLMs

In this section, we propose guidelines for effectively prompting LLMs to enhance stereotype detection. Our evaluations include testing the performance of LLMs in both English and Korean. To assess their capabilities, we compare various approaches: zero-shot learning, in-context learning (ICL), and Chain-of-Thought (CoT) prompting. Refer to Appendix D for the prompt formulation.

Lang	Model	Method	Warm.	Comp.
Eng	Llama	Zero	0.617	0.544
		ICL	0.584	0.523
		CoT	**0.694**	**0.657**
	Qwen	Zero	0.789	0.658
		ICL	0.769	0.643
		CoT	**0.793**	**0.750**
	DeepSeek	Zero	0.512	0.417
		ICL	0.548	0.462
		CoT	**0.557**	**0.487**
Kor	kLlama	Zero	0.489	0.468
		ICL	**0.658**	0.607
		CoT	0.607	**0.656**
	Qwen	Zero	0.0	0.0
		ICL	**0.596**	0.522
		CoT	0.493	**0.563**

Table 3: **Evaluation of SCM Prompting.** The table displays the average accuracies of predictions on warmth and competence. Each model's best performance is highlighted in bold.

5.1 Experimental Setup

We evaluate the proposed methods on the following models: (1) **Llama** (Grattafiori et al., 2024), a Transformer model with 405B parameters, (2) **Qwen** (Qwen et al., 2025), an LLM pre-trained on 18 trillion tokens that supports both Korean and English, (3) **DeepSeek** (DeepSeek-AI et al., 2025), an LLM only trained with reinforcement learning, and (4) **Korean Llama (kLlama)** (Choi et al., 2024), Llama 3.2 fine-tuned with Korean texts using instruction tuning.

5.2 Results

We initially begin our investigation by testing various prompting methods in LLMs that support the Korean language, utilizing the KoSCM dataset. Our findings reveal that the prediction accuracies for these models are significantly lower compared to those achieved by embedding models. As indicated in Table 3, the prediction accuracies of warmth and competence for kLlama range from 0.4 to 0.7. This is significantly lower than the lowest accuracy of the embedding models, approximately 0.85. To de-

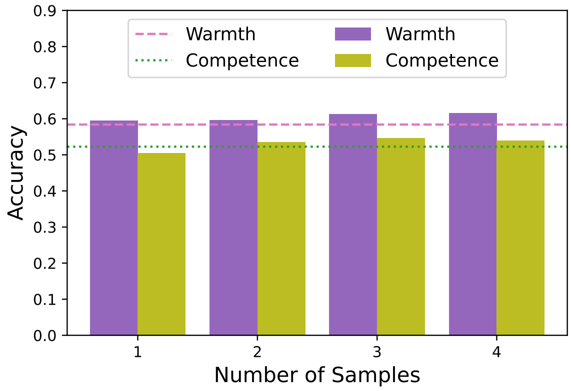


Figure 3: **Comparison Between ICL and Fine-tuning.** The bar plots indicates the average accuracies of warmth and competence predictions for ICL, and the dotted line are those of the fine-tuned model.

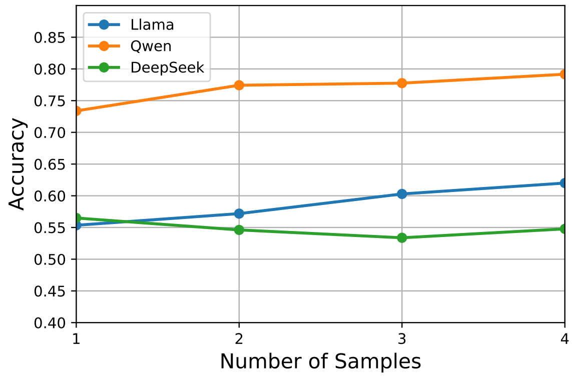


Figure 4: **ICL Performance on Warmth Prediction.** The plot displays the average accuracies of warmth prediction for ICL.

termine whether this subpar performance is related to the fact that Korean is considered a low-resource language in the context of LLM training, we conduct additional tests using the prompting methods in English. The performance improves when tested in English. Still, the accuracy is lower than that of the fine-tuned models. Given studies indicating that the distribution of pretraining data greatly affects ICL performance (Shin et al., 2022; Yadlowsky et al., 2023; Raventós et al., 2023), we deduce that the low results are due to insufficient exposure to similar data during the pretraining phase.

To see if we can further improve the performance of LLMs, we fine-tune Llama with an English SCM dataset of 10k sentences generated with our proposed method (§ 3). The fine-tuned Llama achieves average accuracies of 0.584 for warmth predictions and 0.523 for competence predictions. As shown in Figure 3, the performance is not much better than that of ICL. Overall, LLMs perform best with CoT prompting. In English, CoT consistently outperforms other approaches across all models tested. In contrast, for Korean, ICL achieves the highest accuracy for warmth predictions and CoT for competence predictions.

For all three approaches, we identify several effective strategies to enhance performance when curating instruction prompts. First, instruct the model specifically to conduct a warmth and competence prediction. Writing an instruction that only states "a stereotype detection" may frequently result in refusal due to ethical concerns. Second, include sample seed words of warmth and competence in the instruction. This approach has shown a significant boost in performance, particularly in Korean contexts. The implied nuances can easily get lost in translation, so clearly outlining examples of warmth and competence dimensions can substantially enhance the model's effectiveness.

For ICL, we observe that the performance of Llama and Qwen improves as the number of samples increases, illustrated in Figure 4. In contrast, DeepSeek exhibits consistent performance. Similarly, the CoT approach showed stability in its performance, irrespective of sample size (See Appendix E). Based on these results, the consistent ICL performance of DeepSeek may be attributed to the fact that it generates CoT responses, even with ICL prompts.

6 Conclusion

Our approach demonstrates the potential of the SCM as a cross-cultural tool by adapting it to the Korean language. Our proposed method addresses the challenge of data annotation by leveraging existing seed words. We validate our model using criteria grounded in social psychology theory and also introduce a method for erasing stereotypes. We provide guidelines for prompt engineering to enhance stereotype predictions. This opens up possibilities for expanding the computational application of the SCM to a broader range of researchers across languages and cultures. We observe that predicting warmth and competence is a challenging task for LLMs, suggesting an opportunity for further investigation. This study marks the first attempt to adapt the SCM to the Korean language, aiming to enhance the understanding of stereotypes across cultures. In the future, we plan to broaden our research by adding more languages to promote the development of more inclusive language models.

Limitations

We recognize several limitations that may impact the validity of our findings. Despite our efforts to minimize authorial bias, there remains a possibility for such bias to influence both the experimental design and analysis. For example, the process of clustering social groups is inherently affected by the selection of hyperparameters, which can significantly alter the resulting clusters. Additionally, our decisions in curating prompts for sampling from the dataset and crafting the prompt texts introduce further elements of bias. Hence, these decisions may result in selection bias, which could ultimately impact the conclusions drawn from our study.

Furthermore, our data and experiments are limited by scale constraints. Unlike the abundance of resources available for English models and datasets, there is a significant lack of open-source Korean datasets and models, which has limited our efforts. This insufficient data may suggest that the models utilized in this research are not performing at the same level as their English counterparts. For instance, while conducting back-translation in the data curation process, we observed significant noise in the generated data, which might indicate the difficulties posed by limited resources.

Ethical Considerations

We curate and publish the KoSCM dataset, which is used for training and evaluating KoSCM. This dataset is based on a specific social psychology theory known as the SCM, meaning our research investigates stereotypes within this particular framework. As a result, our dataset and analysis do not encompass the complete range of perspectives on stereotypes. Therefore, we advise researchers utilizing the KoSCM dataset and the proposed translation framework to be mindful of these limitations and encourage them to explore additional methodologies to gain a more comprehensive understanding of stereotypes.

We strongly recommend against using this research for harmful purposes, including the promotion and dissemination of stereotypical biases.

References

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. Leace: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems*, volume 36, pages 66044–66063. Curran Associates, Inc.

Mukul Bhutani, Kevin Robinson, Vinodkumar Prabhakaran, Shachi Dave, and Sunipa Dev. 2024. SeeGULL multilingual: a dataset of geo-culturally situated stereotypes. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 842–854, Bangkok, Thailand. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Yang Trista Cao, Anna Sotnikova, Hal Daumé III, Rachel Rudinger, and Linda Zou. 2022. Theory-grounded measurement of U.S. social stereotypes in English language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1276–1295, Seattle, United States. Association for Computational Linguistics.

Kai-Wei Chang, Vinodkumar Prabhakaran, and Vicente Ordonez. 2019. Bias and fairness in natural language processing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts*, Hong Kong, China. Association for Computational Linguistics.

Rodrigo Alejandro Chávez Mulsa and Gerasimos Spanakis. 2020. Evaluating bias in Dutch word embeddings. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 56–71, Barcelona, Spain (Online). Association for Computational Linguistics.

ChangSu Choi, Yongbin Jeong, Seoyoon Park, Inho Won, HyeonSeok Lim, SangMin Kim, Yejee Kang, Chanhyuk Yoon, Jaewan Park, Yiseul Lee, HyeJin Lee, Younggyun Hahm, Hansaem Kim, and KyungTae Lim. 2024. Optimizing language augmentation for multilingual large language models: A case study on Korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12514–12526, Torino, Italia. ELRA and ICCL.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Amy J. C. Cuddy, Susan T. Fiske, Virginia S. Y. Kwan, Peter Glick, Stéphanie Demoulin, Jacques-Philippe Leyens, Michael Harris Bond, Jean-Claude Croizet, Naomi Ellemers, Ed Sleebos, Tin Tin Htun, Hyun-Jeong Kim, Greg Maio, Judi Perry, Kristina Petkova, Valery Todorov, Rosa Rodríguez-Bailón, Elena Morales, Miguel Moya, Marisol Palacios, Vanessa Smith, Rolando Perez, Jorge Vala, and Rene Ziegler. 2009. Stereotype content model across cultures: Towards universal similarities and some differences. *British Journal of Social Psychology*, 48(1):1–33.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tobias Domhan and Felix Hieber. 2017. Using target-side monolingual data for neural machine translation through multi-task learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1500–1505, Copenhagen, Denmark. Association for Computational Linguistics.

S. T. Fiske, A. J. C. Cuddy, P. Glick, and J. Xu. 2002. A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, 82:878–902.

Susan T. Fiske. 2018. Stereotype content: Warmth and competence endure. *Current Directions in Psychological Science*, 27(2):67–73. PMID: 29755213.

Kathleen Fraser, Svetlana Kiritchenko, and Isar Nejadgholi. 2024. How does stereotype content differ across data sources? In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 18–34, Mexico City, Mexico. Association for Computational Linguistics.

Kathleen C. Fraser, Isar Nejadgholi, and Svetlana Kiritchenko. 2021. Understanding and countering stereotypes: A computational approach to the stereotype content model. In *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 600–616, Online. Association for Computational Linguistics.

Yacine Gaci, Boualem Benatallah, Fabio Casati, and Khalid Benabdeslem. 2023. Societal versus encoded stereotypes in text encoders. In *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 46–53.

Peter Glick. 2002. Sacrificial lambs dressed in wolves' clothing: Envious prejudice, ideology, and the scapegoating of jews.

Peter Glick and Susan T Fiske. 2001a. An ambivalent alliance: Hostile and benevolent sexism as complementary justifications for gender inequality. *American psychologist*, 56(2):109.

Peter Glick and Susan T. Fiske. 2001b. Ambivalent sexism. volume 33 of *Advances in Experimental Social Psychology*, pages 115–188. Academic Press.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Edward P. Havranek, Mahasin S. Mujahid, Donald A. Barr, Irene V. Blair, Meryl S. Cohen, Salvador Cruz-Flores, George Davey-Smith, Cheryl R. Dennison-Himmelfarb, Michael S. Lauer, Debra W. Lockwood, Milagros Rosal, and Clyde W. Yancy. 2015. Social determinants of risk and outcomes for cardiovascular disease. *Circulation*, 132(9):873–898.

Diana Hernández. 2016. Affording housing at the expense of health: exploring the housing and neighborhood strategies of poor families. *Journal of Family Issues*, 37(7):921–946.

Brienna Herold, James Waller, and Raja Kushalnagar. 2022. Applying the stereotype content model to assess disability bias in popular pre-trained NLP models underlying AI-based assistive technologies. In *Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022)*, pages 58–65, Dublin, Ireland. Association for Computational Linguistics.

Mary R Jackman. 1994. *The velvet glove: Paternalism and conflict in gender, class, and race relations*. Univ of California Press.

Younghoon Jeong, Juhyun Oh, Jongwon Lee, Jaimeen Ahn, Jihyung Moon, Sungjoon Park, and Alice Oh. 2022. KOLD: Korean offensive language dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10818–10833, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yiping Jin, Leo Wanner, and Aneesh Moideen Koya. 2024. Disentangling hate across target identities. *Preprint*, arXiv:2410.10332.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Mascha Kurpicz-Briki. 2020. Cultural differences in bias? origin and gender bias in pre-trained german and french word embeddings. *Arbor-ciencia Pensamiento Y Cultura*.

Anne Lauscher, Rafik Takieddin, Simone Paolo Ponzetto, and Goran Glavaš. 2020. AraWEAT: Multidimensional analysis of biases in Arabic word embeddings. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 192–199, Barcelona, Spain (Online). Association for Computational Linguistics.

Sheng Liang, Philipp Dufter, and Hinrich Schütze. 2020. Monolingual and multilingual reduction of gender bias in contextualized representations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5082–5093, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Vijit Malik, Sunipa Dev, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2021. Socially aware bias measurements for hindi language representations. *CoRR*, abs/2110.07871.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings* *of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Mario Mina, Júlia Falcão, and Aitor Gonzalez-Agirre. 2024. Exploring the relationship between intrinsic stigma in masked language models and training data using the stereotype content model. In *Proceedings of the Fifth Workshop on Resources and ProcessIng of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments @LREC-COLING 2024*, pages 54–67, Torino, Italia. ELRA and ICCL.

Jihyung Moon, Won Ik Cho, and Junbum Lee. 2020. BEEP! Korean corpus of online news comments for toxic speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 25–31, Online. Association for Computational Linguistics.

Naver. 2025. Papago.

Gandalf Nicolas, Xuechunzi Bai, and Susan Fiske. 2019. Automated dictionary creation for analyzing text: An illustration from stereotype content.

Gandalf Nicolas and Aylin Caliskan. 2024. Directionality and representativeness are differentiable components of stereotypes in large language models. *PNAS Nexus*, 3(11):pgae493.

Ali Omrani, Alireza Salkhordeh Ziabari, Charles Yu, Preni Golazizian, Brendan Kennedy, Mohammad Atari, Heng Ji, and Morteza Dehghani. 2023. Social-group-agnostic bias mitigation via the stereotype content model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4123–4139, Toronto, Canada. Association for Computational Linguistics.

U.S. Commission on Civil Rights. 2018. Public education funding inequity in an era of increasing concentration of poverty and resegregation.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. *CoRR*, abs/1912.01703.

Arun K. Pujari, Ansh Mittal, Anshuman Padhi, Anshul Jain, Mukesh Jadon, and Vikas Kumar. 2020. Debiasing gender biased hindi words with word-embedding. In *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, ACAI '19, page 450–456, New York, NY, USA. Association for Computing Machinery.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. 2023. Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. In *Advances in Neural Information Processing Systems*, volume 36, pages 14228–14246. Curran Associates, Inc.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Carolin M. Schuster, Maria-Alexandra Dinisor, Shashwat Ghatiwala, and Georg Groh. 2024. Profiling bias in llms: Stereotype dimensions in contextual word embeddings. *Preprint*, arXiv:2411.16527.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Seongjin Shin, Sang-Woo Lee, Hwijeen Ahn, Sungdong Kim, HyoungSeok Kim, Boseop Kim, Kyunghyun Cho, Gichang Lee, Woomyoung Park, Jung-Woo Ha, and Nako Sung. 2022. On the effect of pretraining corpora on in-context learning by a large-scale language model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5168–5186, Seattle, United States. Association for Computational Linguistics.

Masashi Takeshita, Yuki Katsumata, Rafal Rzepka, and Kenji Araki. 2020. Can existing methods debias languages other than English? first attempt to analyze and mitigate Japanese word embeddings. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 44–55, Barcelona, Spain (Online). Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Eddie Ungless, Amy Rafferty, Hrichika Nag, and Björn Ross. 2022. A robust bias mitigation procedure based on the stereotype content model. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 207–217, Abu Dhabi, UAE. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Steve Yadlowsky, Lyric Doshi, and Nilesh Tripuraneni. 2023. Pretraining data mixtures enable narrow model selection capabilities in transformer models. *Preprint*, arXiv:2311.00871.

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2896–2907, Online. Association for Computational Linguistics.

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining gender bias in languages with grammatical gender. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5276–5284, Hong Kong, China. Association for Computational Linguistics.

A Templates for Sentence Generation

In this section, we describe the details of the templates used for generating sentences in Section 3. The templates are curated based on the part-of-speech (POS) tagging of the seed words. The curated seed words contain noun and adjective tags. Based on those tags, we utilize the two templates in Table 4. The subject words for the templates are chosen carefully to ensure that the generated sentences do not contain information about specific social groups. For instance, the pronouns "he" and "she" indicate a person's gender. We chose to avoid using these pronouns as subjects because our objective is to develop a dataset focused on learning the dimensions of warmth and competence. The subject words used for the templates are: ["나" (I), "너" (You), "우리" (We), "그 사람" (That person), "저 사람" (That person), "이 사람" (This person)]. With the curated templates, a total of 3,420 sentences are generated. Here are sample sentences generated using the templates: "나는 능력이 있다." (I am competent.), "그 사람은 친절한 사람이다." (That person is a kind person.).

B Experimental Settings

All models are implemented with PyTorch (Paszke et al., 2019) and Huggingface's Transformers (Wolf et al., 2020). For training, we use the Adam optimizer (Kingma and Ba, 2015) and set the learning rate from $\{1e-05, 2e-05\}$, an epoch as 10, a dropout rate of 0.3, and hyperparameters α and β as 1 . All experiments are conducted on a Nvidia Quatro RTX 5000, 16 GB memory GPU in a machine with Intel(R) Xeon(R) Silver 4214 CPU @ 2.20GHz.

C The SCM CoT prompt Generator

When seed words and their corresponding directions are given, the website generates a CoT prompt suitable for use in a selected LLM. The site is available in both English and Korean, as shown in the screenshots (Figures 5 and 6).

D LLM Prompts for Warmth and Competence Predictions

English and Korean prompts for warmth and competence predictions are shown in Table 5. The first two rows display a system prompt that is provided for all learning prompts. In zero-shot learning, only the system prompt and a single sentence are presented, and the model must predict the direction of warmth and competence for the given sentence. In in-context learning (ICL), a sentence is provided along with its corresponding labels as an example for the model. Users can choose the number of examples to present to the model.

POS	Template	English Translation
NOUN	[SUBJECT]은/는 <seed word>이/가 있다.	[SUBJECT] has <seed word>.
ADJECTIVE	[SUBJECT]은/는 <seed word> 사람이다.	[SUBJECT] is a[n] <seed word> person.

Table 4: **Templates for Sentence Generation.** The table shows two different sentence templates based on the POS tagging of a seed word. English versions of Korean templates are provided for reference.

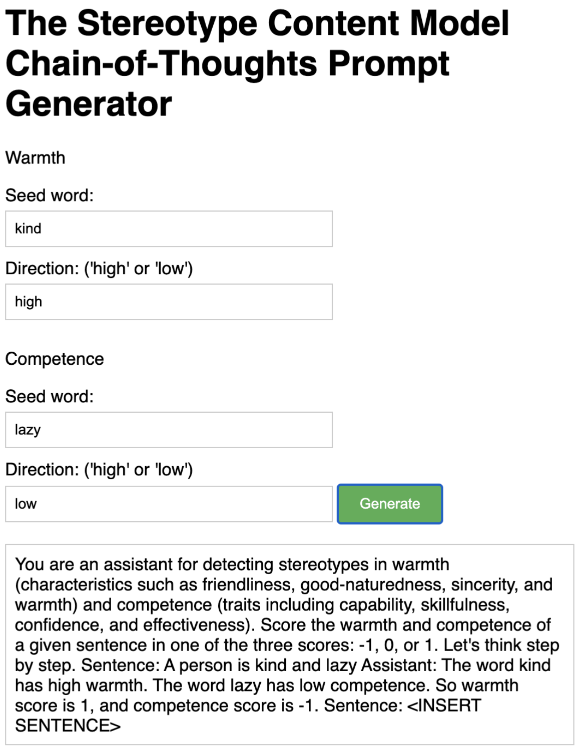
	Prompt	
Zero	{"role": "system", "content": "당신은 친밀함(친근함, 친절함, 성실함, 따뜻함 등의 특징)과 능력(능력, 숙련도, 자신감, 효과성 등의 특징)에서 고정관념을 감지하는 서비스를 제공합니다. 주어진 글의 친밀함과 능력 점수를 -1, 0, 1 중에 하나로 매깁니다.}	
	{"role": "system", "content": "You are an assistant for detecting stereotypes in warmth(characteristics such as friendliness, good-naturedness, sincerity, and warmth) and competence(traits including capability, skillfulness, confidence, and effectiveness). Score the warmth and competence of a given sentence in one of the three scores: -1, 0, or 1.}	
ICL	{"role": "user", "content":<sentence>}, {"role": "assistant", "content": "온기 <w dir>, 능력 <c dir>"}	× # samples
	{"role": "user", "content":<sentence>}, {"role": "assistant", "content": "Warmth <w dir>, Competence <c dir>"}	× # samples
CoT	{"role": "user", "content": <sentence>}, {"role": "assistant", "content": "차근차근 생각해봅시다. 주어진 문장에서 <w dir> 친밀함을 나타내는 단어는 <w seed word>이다. 주어진 문장에서 <c dir> 능력을 나타내는 단어는 <c seed words>이다. 그러므로 온기 <w dir>, 능력 <c dir>"}	
	{"role": "user", "content": <sentence>}, {"role": "assistant", "content": "Let's think step by step. The word <w seed word> has <w dir> warmth. The word <c seed word> has <c dir> competence. So Warmth <w dir>, Competence <c dir>"}	

Table 5: **Prompts for Warmth and Competence Predictions.** The table above shows the prompt used for zero-shot learning, in-context learning, and Chain-of-thought prompting with LLMs.

For Chain-of-Thought (CoT) prompting, a selected number of examples are given to the model, similar to in-context learning. However, the difference is in the example answers, which provide more detailed explanations. The model is instructed to think step by step, and then it is presented with the seed words that help determine the direction of warmth and competence.

E SCM Prompting

We evaluate how to effectively prompt LLMs to enhance stereotype detection in the English SCM dataset. To assess their capabilities, we test in-context learning (ICL) and Chain-of-Thought (CoT) prompting on Llama, Qwen, and DeepSeek. Figures 4 and 7 show the performance of ICL in warmth and competence predictions, respectively. For Llama and Qwen, we notice that performance improves as the number of samples increases. On the other hand, for DeepSeek, we observe a plateau. This difference may be from the observation that even with ICL prompts, DeepSeek generates responses that are similar to a CoT approach. As shown in Figures 8 and 9, which illustrate the performance of CoT as the number of samples increases, the performance of CoT remains stable regardless of the number of samples provided. The resemblance between these figures and the ICL performance of DeepSeek supports our conjecture. In all cases, we observe that LLMs perform better in predicting warmth than competence.



The Stereotype Content Model Chain-of-Thoughts Prompt Generator

Warmth

Seed word:

kind

Direction: ('high' or 'low')

high

Competence

Seed word:

lazy

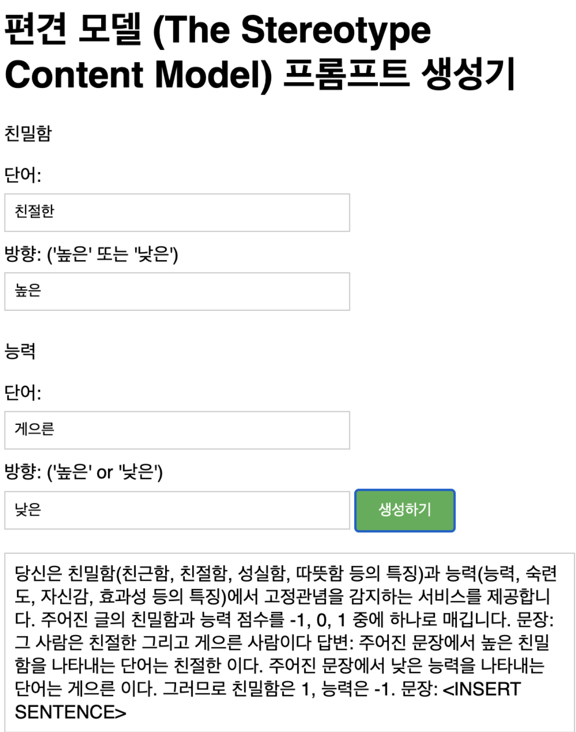
Direction: ('high' or 'low')

low

Generate

You are an assistant for detecting stereotypes in warmth (characteristics such as friendliness, good-naturedness, sincerity, and warmth) and competence (traits including capability, skillfulness, confidence, and effectiveness). Score the warmth and competence of a given sentence in one of the three scores: -1, 0, or 1. Let's think step by step. Sentence: A person is kind and lazy Assistant: The word kind has high warmth. The word lazy has low competence. So warmth score is 1, and competence score is -1. Sentence: <INSERT SENTENCE>

Figure 5: **The SCM CoT Prompt Generator in English.** A screenshot of the prompt generator website in English is shown above.



편견 모델 (The Stereotype Content Model) 프롬프트 생성기

친밀함

단어:

친절한

방향: ('높은' 또는 '낮은')

높은

능력

단어:

게으른

방향: ('높은' or '낮은')

낮은

생성하기

당신은 친밀함(친근함, 친절함, 성실함, 따뜻함 등의 특징)과 능력(능력, 숙련도, 자신감, 효과성 등의 특징)에서 고정관념을 감지하는 서비스를 제공합니다. 주어진 글의 친밀함과 능력 점수를 -1, 0, 1 중에 하나로 매깁니다. 문장: 그 사람은 친절한 그리고 게으른 사람이다 답변: 주어진 문장에서 높은 친밀함을 나타내는 단어는 친절한 이다. 주어진 문장에서 낮은 능력을 나타내는 단어는 게으른 이다. 그러므로 친밀함은 1, 능력은 -1. 문장: <INSERT SENTENCE>

Figure 6: **The SCM CoT Prompt Generator in Korean.** A screenshot of the prompt generator website in Korean is shown above.

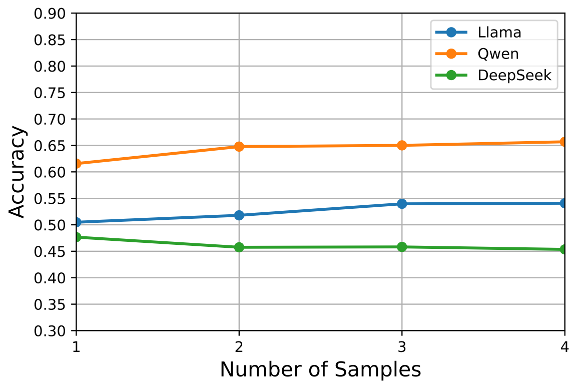


Figure 7: **ICL performance on competence prediction.** The plot displays the average accuracies of competence prediction for ICL. The x-axis represents the number of samples presented to a model.

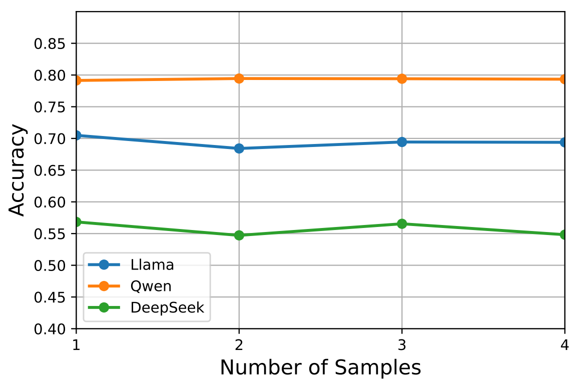


Figure 8: **CoT performance on warmth prediction.** The plot displays the average accuracies of warmth prediction for CoT. The x-axis represents the number of samples presented to a model.

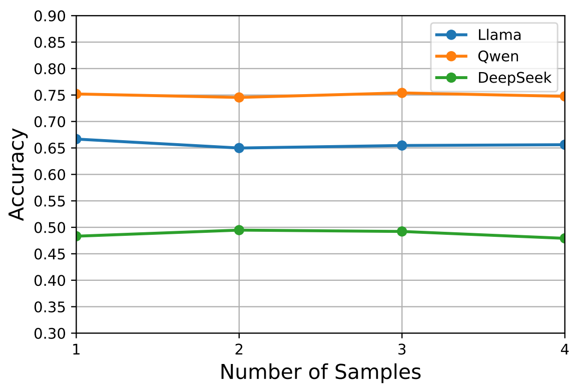


Figure 9: **CoT performance on competence prediction.** The plot displays the average accuracies of competence prediction for CoT. The x-axis represents the number of samples presented to a model.