ADAPTIVE AXES: A Pipeline for In-domain Social Stereotype Analysis

Qingcheng Zeng
Northwestern University
qcz@u.northwestern.edu

Mingyu Jin
Rutgers University
mingyu.jin404@gmail.com

Rob Voigt
Northwestern University
robvoigt@northwestern.edu

Abstract

Prior work has explored the possibility of using the semantic information obtained from embedding representations to quantify social stereotypes, leveraging techniques such as word embeddings combined with a list of traits (Garg et al., 2018; Charlesworth et al., 2022) or semantic axes (An et al., 2018; Lucy et al., 2022). However, these approaches have struggled to fully capture the variability in stereotypes across different conceptual domains for the same social group (e.g., *black* in science, health, and art), in part because the identity of a word and the associations formed during pretraining can dominate its contextual representation (Field and Tsvetkov, 2019). This study explores the ability to recover stereotypes from the contexts surrounding targeted entities by utilizing state-of-the-art text embedding models and **adaptive semantic axes** enhanced by large language models (LLMs). Our results indicate that the proposed pipeline not only surpasses token-based methods in capturing in-domain framing but also effectively tracks stereotypes over time and along domain-specific semantic axes for in-domain texts. Our research highlights the potential of employing text embedding models to achieve a deeper understanding of nuanced social stereotypes.

1 Introduction

Social stereotypes, representing the associations attributed to social groups (e.g., *White*, *Black*, *Religious*), are deeply embedded in and perpetuated by human languages. These stereotypes are both reflected in everyday language use and contribute to the reinforcement of societal biases. Consequently, a growing topic of interest in NLP is whether and how computational techniques can be used to quantify these associations at scale. Various methodologies have been developed to explore and measure these social biases in language. For instance, by calculating cosine similarities between the embeddings of traits (e.g., *unhealthy*, *weak*) and social groups within word embeddings (Garg et al., 2018; Charlesworth et al., 2022, 2023), researchers can uncover how stereotypes emerge and persist across society. Another related approach involves projecting social group embeddings along opposed semantic dimensions (i.e., semantic axes, such as *beautiful* - *ugly* (An et al., 2018; Lucy et al., 2022)) to reveal tendencies toward particular semantic dimensions, suggesting certain stereotypes.

Social stereotypes are multifaceted and can intersect across different social groups or vary across different domains. For instance, Burnett et al. (2020) investigated how racial stereotypes in the US persist in both academic and non-academic contexts, such as music and sports. Their research demonstrated that the same social group can be associated with different stereotypes depending on the domain. Similarly, Shih et al. (2006) found that Asian American women performed better on a verbal test when their female identity was made salient but performed worse when their Asian identity was emphasized. These findings suggest that domain-specific stereotypes significantly impact performance outcomes. Such complexity underscores the importance of understanding in-domain stereotypes, which are stereotypes specific to particular contexts or domains.

In-domain stereotypes are particularly challenging to analyze due to the need for contextual specificity in identifying associations with a target group. Lucy et al. (2022) attempted to address this by replacing target social group words with neutral words (e.g. "person") and projecting neutral words' contextual embeddings onto semantic axes. However, this approach often resulted in the neutral word's identity dominating the analysis, leading to similar semantic poles (i.e., one end of one semantic axis) across different occupational categories before filtering. In this work, we develop a novel pipeline leveraging off-the-shelf LLMs and text embedding models to explore in-domain stereotypes.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 15576–15593
November 12-16, 2024 ©2024 Association for Computational Linguistics

Our pipeline first enhances semantic axes in two ways: (1) To address the gap where existing broad semantic axes fail to capture domain-specific variations in stereotypes, we utilize LLMs to generate more comprehensive and relevant axes. This approach allows us to include important contextual nuances, such as *globalization - nationalism* in economic analyses. (2) Employing multiple pruning methodologies to refine existing semantic axes, ensuring inappropriate words are trimmed to avoid semantic confusion. Then, whereas prior work has calculated associations with semantic axes using token embeddings, we explore whether these associations can be better modeled by embedding the context surrounding a target entity mention. Using off-the-shelf text embedding models, we embed the context with target entity masked and adaptive semantic axes to measure group- and domain-specific stereotypes along these axes.

We conduct extensive evaluations using automatic validation metrics and human evaluators, demonstrating that: (1) text embedding models encode semantic axes with greater consistency compared to previous token-based embeddings from BERT; (2) our pipeline captures in-domain stereotypes that better align with human annotations compared to previous approaches; and (3) in a case study of US news discourse, our pipeline effectively captures general stereotypes, contrasts between countries, and changes in associational biases corresponding to real-world events along specific axes of interest. Our results show that this innovative approach allows for a more nuanced and precise understanding of stereotypes within specific domains. Our codes are available at `https://github.com/qcznlp/adaptive_axes`

2 Background and Related Work

2.1 Using NLP for Social Biases Analyses

Social stereotypes are widely encoded within natural languages. Traditional methods for eliciting social stereotypes, such as human surveys (Williams and Best, 1990) or dictionary analysis (Henley, 1989), are limited in scale. The advent of word embedding models, which quantitatively capture word associations, introduced a new approach. Garg et al. (2018) used decade-wise word2vec models trained on *Google Books* (Michel et al., 2011) and the *Corpus of Historical American English (COHA)* (Davies, 2012) to investigate temporal gender and ethnic biases, showing that stereotypes about women correlate with social movements. Similarly, Charlesworth et al. (2022) and Charlesworth et al. (2023) extended this research to 14 social groups, covering periods from 1800 to 1999, and used valence scores to track the positivity/negativity of stereotypes toward different social groups over time.

Semantic axes are a related but alternative approach initially proposed by An et al. (2018). Their framework involves three steps: constructing word embedding models, identifying semantic axes of interest, and projecting targeted words onto these axes to reveal associational stereotypes. Semantic axes are advantageous due to their interpretability along human-curated dimensions, allowing for a clear and intuitive comparison of how different groups are perceived along a particular semantic dimension. Lucy et al. (2022) extended this concept to contextualized embedding models, demonstrating their better alignment with human judgments over static embeddings. Both approaches rely on off-the-shelf knowledge graphs, such as ConceptNet (Speer et al., 2017) and WordNet (Miller, 1995), to construct semantic axes. While these knowledge graphs offer comprehensive synonym pairs, they are grounded in a manually curated, general-purpose ontology that is fixed; by contrast, we aim to capture domain-specific associations.

2.2 Text Embedding Models and Social Computing

The development of LLMs has advanced the representation of sentence- or paragraph-level text into fixed-size embeddings, facilitating the retrieval of relevant texts and clustering of similar semantic contents. For instance, SentenceBERT (Reimers and Gurevych, 2019), fine-tuned with natural language inference (NLI) data, and recent LLM-based embedding models fine-tuned using synthetic data (Wang et al., 2024; Meng et al., 2024) have demonstrated strong performance in retrieval and textual similarity tasks, as evidenced by their success on the MTEB leaderboard (Muennighoff et al., 2023).

Despite the potential of text embedding models to significantly enhance social computing across various disciplines, their application in this field remains quite an open question. For example, Licht (2023) demonstrated the capabilities of multilingual embedding models in political text classification, while Libovický (2023) showed that these models could encode biases related to jobs and occupational locations. In this paper, we investigate

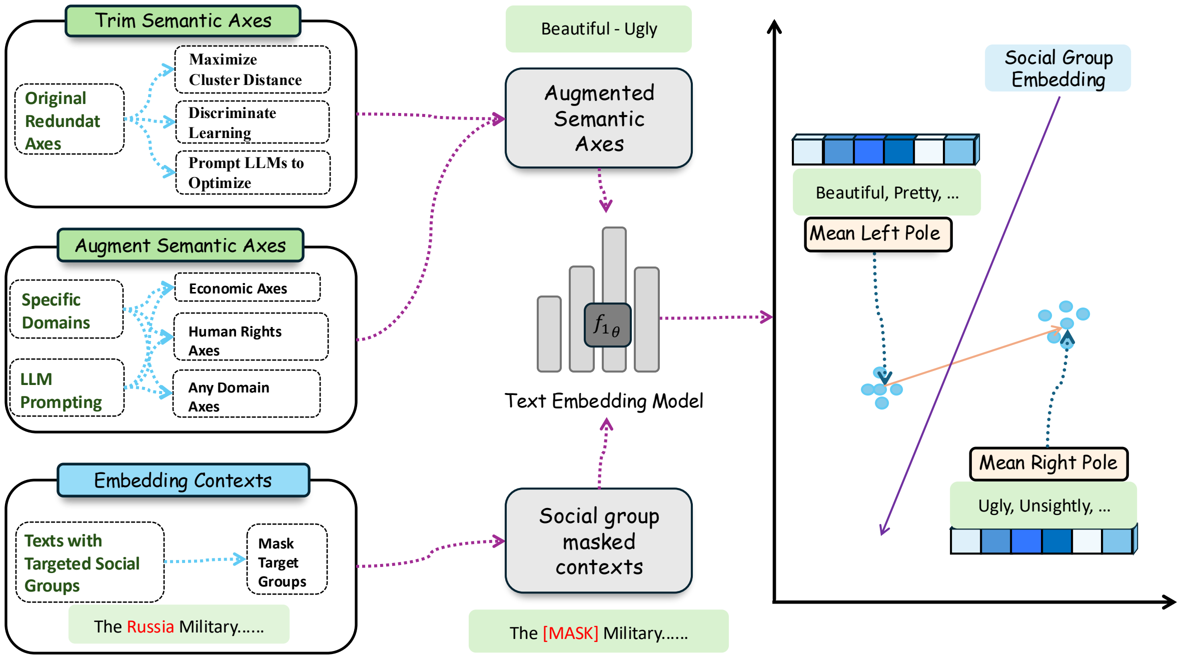


Figure 1: The ADAPTIVE AXES Pipeline. We use text embedding models as our core mechanism for social stereotype analyses, projecting context-only embeddings with the target group masked onto semantic axes. We also develop pruning methods to refine semantic axis seed sets and generate new domain-specific axes with LLMs.

whether sentence representations can effectively recover in-domain stereotypes by encoding contexts, thereby further exploring potential applications of text embeddings in social computing.

3 Proposed Pipeline – ADAPTIVE AXES

In this section, we describe our ADAPTIVE AXES pipeline. Specifically, the construction of semantic axes relies on three main steps - (1) building embedding models, (2) building semantic axis poles and getting semantic axis vectors, and (3) projecting target vectors on semantic axes to show stereotypes. Our study revisits these steps, aiming to construct a pipeline with high generalizability and accuracy. Our general framework is shown in Figure 1.

3.1 Embedding Model Constrution

The construction of semantic axes relies on high-quality embedding models capable of capturing nuanced semantic differences. Previous approaches, such as by An et al. (2018) and Mathew et al. (2020), used static embedding models like `word2vec` or `GloVe`, which often fail to capture antonym relations and are unsuitable for contextualized tasks. Lucy et al. (2022) used up to 1,000 sentences from Wikipedia per adjective to obtain average contextualized embeddings, requiring high-quality text instances for embedding construction.

Here we explore whether off-the-shelf text embedding models can effectively construct semantic axes. We propose two key advantages of using text embedding models: (1) models fine-tuned through techniques such as contrastive learning on base models are likely to excel at distinguishing nuanced semantic differences, which is crucial for semantic axes construction; (2) these models can encode short phrases directly without relying on extensive text instances, making semantic axes easily generalizable to any-length phrases. Therefore, we employ text embedding models that perform well on the semantic textual similarity (STS) task of the MTEB leaderboard (Muennighoff et al., 2023) as our backbone models for further analyses.

3.2 Semantic Axes Enhancement

We propose to enhance prior approaches to the construction of semantic axes in two ways. First, we prune axes to improve interpretability and enhance semantic contrast. Second, we go beyond fixed lists of axes by generating new, domain-specific axes tailored to particular contexts.

3.2.1 Pruning Existing Axes

The quality of the semantic axis poles is crucial as it directly influences the semantic contrastiveness of the axes, thereby affecting the quality of the semantic axis vectors. An et al. (2018) used 732 predefined single-word antonym pairs from ConceptNet and enhanced both poles by adding the top-*N* similar words from the embedding model to ensure greater robustness. In contrast, Lucy et al. (2022) utilized WordNet, which consists of 723 axes with each pole having an average of 9.63 adjectives. This approach can lead to unwanted meanings across both poles. For example, one example of the semantic axis of WordNet is:

- The left pole: animal, bodily, carnal, corporal, corporeal, fleshly, material, personal, physical, physiologic, physiological, sensual, somatic
- The right pole: intellectual, mental, moral, noetic, psychic, psychical, psychogenic, psychological, rational

We observe that while these axes exhibit some reasonable semantic coherence, they may also be quite broad in their semantic scope. For example, the left pole includes the term “animal,” which is quite general, and the right pole includes “psychogenic,” which is relatively rare. Additionally, the large number of terms on both poles could make human interpretation challenging. We investigated three methods to prune inappropriate words from the semantic axis poles in WordNet:

Cluster Distance Maximization We start by independently clustering the embeddings for the two poles. Employing Euclidean and cosine distances as our metrics, we iterate through all possible combinations to assess the contrasts between the two sets of embeddings. Ultimately, we select the poles with the greatest inter-group distance to construct the pruned axis.

Using Discriminative SVMs For this method we use embedding dimensions as features in support vector machines (SVMs), a supervised learning algorithm designed for classification tasks. We frame a binary classification problem of distinguishing between the two poles given observations of the words in each pole. We then progressively remove vectors that have minimal influence on the separability of the two embedding clusters, as determined by their effect on the classification margin. This iterative pruning continues until the clusters stabilize, evidenced by a cessation in the growth of the inter-cluster distance. Upon convergence, these refined clusters are employed to establish the final axes.

LLM Evaluation We instruct LLMs to analyze the existing exhaustive semantic axes in WordNet. The LLMs are guided to trim both poles in a way that preserves the semantic contrasts of the original seed adjectives. This process aims to ensure that the refined poles retain their distinctiveness and relevance. Our prompt template, which directs the LLMs in this task, is detailed in Appendix A.1. The LLM-generated semantic poles are then used to construct the final semantic axes.

3.2.2 Domain-Specific Axes Augmentation

As previously discussed, conventional semantic axes frequently encounter difficulties in adapting to new domain-specific contexts, a critical challenge when examining shifts in stereotypes within specific domains. To overcome this limitation, our study leverages LLMs to generate domain-specific semantic axes, such as *peaceful protests* versus *military intervention* in the **military** domain, or *political transparency* versus *political opaqueness* in the **political** domain. These various-length phrases play a pivotal role in expanding the scope of our analytic framework. Our approach aims to generate a set of axes tailored to the unique requirements of each domain, which can then be embedded directly using text embedding models. To offer a clear example of this process, our prompt template is presented in Appendix A.2. This method facilitates the dynamic generation and adaptation of domain-specific semantic axes, enabling monitoring of stereotypes in particular domains that would not be included in existing knowledge bases.

3.3 Stereotype Understanding with Text Embedding Models

Previous studies have typically used token embeddings of target social groups (e.g., static or contextual embeddings of *Black* or *Old*) for stereotype analysis. Although a neutral word approach (i.e., replacing the target social group tokens with *person*) can extract contextual differences across domains after statistical filtering, the stereotypes derived from contextual differences in Lucy et al. (2022) closely correlate with original token-based stereotypes. This indicates that well-encoded contexts are sufficient for understanding stereotypes, thereby reducing the reliance on social group biases within pre-training data.

We aim to exploit the context in contextual embeddings, rather than relying solely on token representations, to gain more detailed and domain-specific insights from various text sources. Our method involves extracting the context around specific social groups and masking their appearances in the text to obtain context embeddings, as shown in Figure 1. We then project these embeddings onto constructed semantic axes to identify stereotypes associated with different social groups across various domains.

4 Pipeline Validation

In this section, we present a series of experiments validating the effectiveness of our pipeline in two key areas: (1) Do text embedding models encode semantic axes effectively? (2) Does our pipeline capture stereotypes closely aligned with human intuitions? Additionally, we conduct experiments to understand whether target-masked text embedding models could accurately predict affective information in Appendix B.

To address the first question, we employ `UAE-large-v1` (Li and Li, 2023), a model fine-tuned on `BERT-large`, and `SFR-Embedding-Mistral` (Meng et al., 2024), which is based on `Mistral-7B` (Jiang et al., 2023), representing two state-of-the-art approaches to text embedding. Each word or phrase is processed by these models without additional prompts to obtain candidate embeddings.

For the second question, to avoid directly modeling the specific word of interest, we utilize attention masks and include 20 surrounding tokens (or all available tokens if fewer than 20) for context modeling. If the targeted social groups appear multiple times, all instances are masked before being processed by the embedding model.

To introduce more domain-specific semantic axes, for our case studies we use GPT-4 (OpenAI et al., 2024) to generate 13 new axes in the following domains: *politics and governance*, *global trade and economics*, and *culture and education*. These domain-specific axes enable a more precise analysis of stereotypes and their variations across different contexts. Our domain-specific semantic axes are attached in Appendix C.

To effectively classify news articles in the News on the Web corpus (Davies, 2022) into various categories to mine domain-specific stereotypes, we use the zero-shot classification system (Yin et al., 2019) and a list of candidate labels (*global trade and economic*, *politics and governance*, *cultural and education*, and *none of above*) to classify US news articles. Two authors manually annotated 100 random news articles and find an average classification accuracy of 82%, which is sufficient to scale across the corpus.

Models	**Average C**	**Number of Consistent Axes**
GLOVE	0.101	503
BERT-prob^z	0.133	512
UAE-large-v1	0.120	603
SFR-Embedding-Mistral	**0.153**	**712**
Pruning Methods	**Average C**	**Number of Consistent Axes**
Cluster Cosine Distance Maximization	0.141	620
Cluster Euclidean Distance Maximization	0.148	641
Using Discriminative SVMs	0.106	522
LLM Evaluation	0.107	537

Table 1: Top: Different models' consistency C and the number of consistent semantic axes. A higher consistency or number of consistent axes represents a better encoding of semantic contrasts. Specifically, the BERT-prob^z represents normalized BERT embeddings as proposed by Timkey and van Schijndel (2021).
Bottom: The metrics of the pruned semantic axes are based on the model `UAE-larve-v1`. We find maximizing cluster Euclidean distance gives the best results.

4.1 Validation of Semantic Axes Construction

In this section, we first investigate whether text embedding models can capture the meanings of different poles within semantic axes, following a methodology similar to Lucy et al. (2022). We remove one word from either pole and compute the cosine similarities to the axis constructed from the remaining words. If a semantic axis is consistent, the left-out word should be closer to the pole to which it originally belongs. We average these leave-one-out similarities for each pole to produce a consistency metric, C. An axis is considered "consistent" if both poles have $C \geq 0$.

For a fair comparison, we first use the same data as Lucy et al. (2022) to evaluate semantic axes derived from different models. The results, shown in Table 1, indicate that contemporary text embedding models embed semantic axes better than corpus-curated semantic axes using the original BERT. These findings suggest that using off-the-shelf sentence encoders to embed semantic axes is a rational approach, leading to a larger number of consistent axes and comparable consistency with the best results reported by Lucy et al. (2022). We then further prune the semantic axes only based on `UAE-large-v1` due to the large computational requirements of `SFR-Embedding-Mistral`, with the results presented in the second half of Table 1. Only the method of maximizing the cluster Euclidean distance leads to positive improvements, thus we use the pruned axes in further analyses.

To understand whether our domain-specific semantic axes from LLMs are truly meaningful in that domain, we propose one simple method - these domain-specific axes should have higher variances than general axes in that specified domain. That is to say, for example, in the **political** domain, the axis *political transparency* versus *political opaqueness* should have higher variances than the axis *beautiful* versus *ugly*.

To formalize this, let $\mathbf{E}_{\text{domain}}$ be the entity-masked context embeddings in a specific domain, $\mathbf{A}_{\text{specific}}$ be the domain-specific axis and $\mathbf{A}_{\text{general}}$ be the general axis. We calculate the cosine similarities between $\mathbf{E}_{\text{domain}}$ and the axes $\mathbf{A}_{\text{specific}}$ or $\mathbf{A}_{\text{general}}$:

$$\cos(\theta) = \frac{\mathbf{E}_{\text{domain}} \cdot \mathbf{A}_{\text{s/g}}}{||\mathbf{E}_{\text{domain}}|| \, ||\mathbf{A}_{\text{s/g}}||}$$

This allows us to capture how closely the embeddings align with the domain-specific or general axes. We then compute the variance of the similarity for each axis across the entire dataset:

$$\text{Var}(X) = E\left[(X-\mu)^2\right] = E[X^2] - (E[X])^2$$

where X represents the similarities for one single axis across the domain-wise dataset. We use the average percentile ranking by the variance of these similarities as a quantitative measure to evaluate whether these axes are meaningful in that domain, in which a lower percent implies the variance is bigger and thus captures meaningful variations. The results are shown in Table 2, indicating that entity embeddings along these domain-specific axes show high in-domain variance, suggesting they can capture meaningful domain-specific variation.

4.2 Validation of the Pipeline

Previous sections validate the text embedding model's capacity to encode semantic axes. In this section, we validate the practical step - how can our pipeline understand domain-specific stereotypes compared to previous models?

We construct an annotation task in which for each participant we obtain the same three sentences which include `China/Chinese`, `Mexico/Mexican`, or `Canada/Canadian` in the **political** and **cultural** domains from the News on the Web corpus (Davies, 2022). Then these sentences go through three pipelines - ADAPTIVE AXES, contextualized token-based semantic axes (the *BERT-prob* method in (Lucy et al., 2022), we averaged to aggregate the sentence), and a randomized baseline model which samples five seed words from semantic poles. We recruited 21 participants from the crowdsourcing platform Prolific to rank the three models in 3 nationalities × 2 domains = 6 targeting questions and one quality-control question. The quality control question is one question with an obviously right answer by which we filtered invalid participants out. In each question, every option consists of the top-5 positively associated semantic poles from the three methods. Each participant was paid $12 per hour to do the annotation task. Ultimately 20 participants were included in the final analyses. Our interface is shown in Appendix A.3.

Domains	Average Variance Ranking
Politics and Governance	6.4%
Global Trade and Economics	9.7%
Cultural and Education	10.3%

Table 2: The average variance ranking measures the in-domain average percentile rank by mean variance for our evaluation set of augmented semantic axes compared to WordNet-based axes. Lower numbers indicate better performance. Our domain-specific axes generally fall within the top 10% when ranked by variance, suggesting they capture significant variation in the domain-specific representation of entities.

We measured the effectiveness of our methodology in multiple ways. First, we calculated Kendall's W and the average rankings of these three methods to understand the superiority of these methods and to what extent participants agree with each other on the rankings of these three models. Second, we compared the diversity of these three methods. Given that these six questions belong to two different domains and three countries, the ideal methodology should have a medium-level diversity to represent domain specificity. Specifically, we use Jaccard similarity to calculate pairwise domain similarities (e.g., **China** in **political** vs. **cultural**) and average three similarity scores to get the final metric.

The final results are shown in Table 3, indicating that in most cases ADAPTIVE AXES helps to capture domain-specific semantic associations with the highest ranking and a reasonable inter-annotator agreement (Kendall's $W = 0.58$). Besides, the average Jaccard similarity suggests that

Model	Average Ranking	Average Jaccard Similarity
Random Baseline	2.4 (±0.227)	0
Token-based Embedding	1.925 (±0.217)	0.889
ADAPTIVE AXES (ours)	1.675 (±0.177)	0.310

Table 3: Human-evaluated rankings for three types of pipelines, where the ranking ranges from 1 to 3, with lower numbers indicating better performance. Confidence intervals are shown in parentheses.

our ADAPTIVE AXES also has medium-level cross-question similarity versus token-based approach (0.310 vs. 0.889), suggesting that it helps with in-domain social stereotype modeling by modeling social groups' surrounding contexts.

5 Case Study #1

In this section, we turn to real-world case studies with ADAPTIVE AXES to ask one research question: how are different countries generally framed across various domains in US news discourse?

5.1 Data

We use the US news subset of the NOW corpus (Davies, 2022) ranging from June 2010 to August 2023 and filters to news articles containing target countries or demonyms (e.g. France and French) in the headline to focus on substantially relevant articles. We cover four countries (China, Russia, Germany, and Canada), in which China and Russia are frequently framed as competitors with the US, and Germany and Canada are often depicted as allies. Detailed distributional information for each category of each country is attached in Appendix D.

5.2 Results

Observation 1: ADAPTIVE AXES effectively model general social stereotypes. Table 4 presents the top three semantic axes (referring to the positively associated pole of axes) for various countries across different domains. Although precise quantitative evaluation is challenging, the majority of semantic axes identified through our pipeline align well with widely recognized stereotypes. For instance, in the *politics and governance* domain, the term *electoral* shows a strong association with Germany and Canada. Conversely, the opposite pole, *authoritarian*, is predominantly linked to China and Russia. Another significant finding emerges in the *global trade and economic* domain, where our model accurately captures the stereotype of China as the 'world factory' with a large labor force (Zhang, 2006). Similarly, axes such as 'antimonopoly' and 'market economy' correctly reflect economic perceptions of Germany (Marktanner, 2014; Yamazaki, 2019). These results indicate that context-based modeling effectively reveals meaningful associational differences across social groups and domains.

One significant limitation of this approach is that regardless of the type of semantic axis employed, all axes capture patterns of co-occurrence between words or contexts rather than direct causal relationships. For instance, in the economic and trade domain, many countries are closely associated with the *'overseas'* axis. This axis reflects a general characteristic of trade rather than a domain-specific stereotype, indicating that while our method effectively captures broad shifts in stereotypes across domains, any associations identified should be interpreted with caution.

Observation 2: ADAPTIVE AXES can (partially) capture contrastive stereotype shifts across social groups. Can this pipeline effectively generalize to identify contrastive stereotypes? For instance, what are the key differences in semantic associations between China and Canada within the trade and economics domain? To explore this, we compute the top semantic axes and corresponding scores for each group separately, then analyze the differences to derive their contrastive axes. These differences help to highlight the variations in stereotype associations between the two social groups.

The top two semantic axes associated with the example contrastive groups are presented in Table 5. These findings indicate that our pipeline, when analyzing semantic axis scores, can at least partially reflect significant differences between groups. For instance, in the comparison between China and Canada, the model successfully captures the trade tensions between the U.S. and China, highlighting the unequal trade dynamics that eventually escalated into the 2017 trade wars (Kwan, 2020). Furthermore, the contrast between the *left-wing* axis in China and Germany within the political domain reflects a general characterization of Chinese politics (Chen et al., 2012). Similarly, the *native* axis, contrasting Canada and Germany, underscores the closer cultural alignment between the U.S. and Canada compared to Germany. These results suggest that domain-specific contexts can effectively reveal contrastive differences between countries.

Countries	Domains	Top Semantic Axes		
China	Global Trade and Economic	overseas	industrious, untiring	factory-made, mass-produced
	Politics and Governance	socialized	authoritarian	asymmetric
	Culture and Education	ethnical	self-conscious	authoritarianism
Germany	Global Trade and Economic	overseas	antimonopoly	market economy
	Politics and Governance	electoral	democratic	nationalistic
	Culture and Education	historical	labor-intensive	ethnical
Russia	Global Trade and Economic	overseas	ploughed	state control
	Politics and Governance	authoritarian	corrupt	rebellious
	Culture and Education	dictatorial	culture exclusivity	blue-collar
Canada	Global Trade and Economic	overseas	profitable	blue-chip, valuable
	Politics and Governance	soft power	electoral	nationalistic
	Culture and Education	north	time-honored	multiculturalism

Table 4: Top semantic axes associated with different countries in each domain.

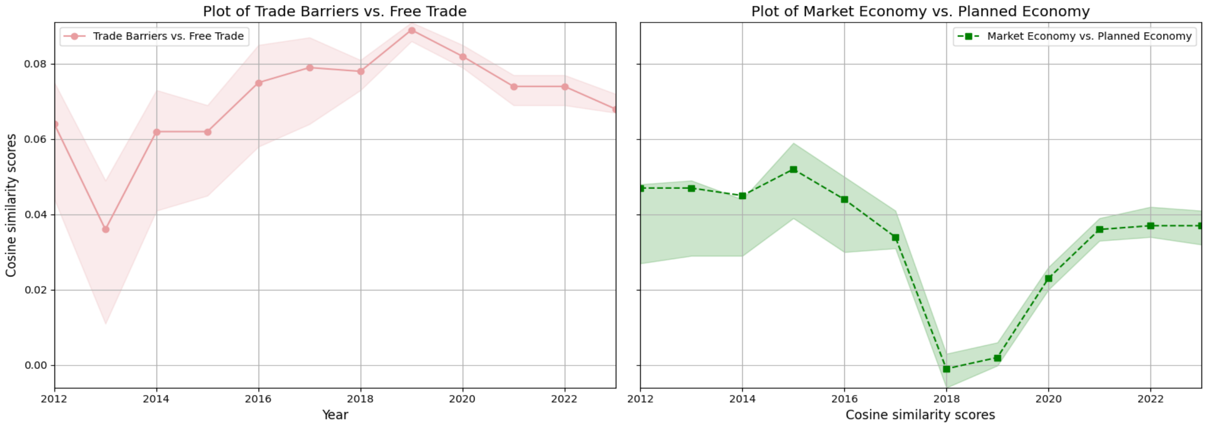


Figure 2: The cosine similarity score changes for two LLM-curated semantic axes. Left: a higher score means inclining to *trade barriers*. Right: a higher score means inclining to *market economy*. Ribbons represent 95% confidence intervals.

Contrastive Groups	Domains	Semantic Axes
China vs. Canada	Trade	inequality, warlike
China vs. Germany	Politics	foresighted, left-wing
Canada vs. Germany	Culture	emotionless, native

Table 5: Contrastive semantic axes associated with the former social groups. The semantic axes represent the more salient associations with the former country relative to the latter.

6 Case Study #2

A key theoretical advantage of our pipeline is its flexibility in embedding new, user-specified semantic axes, whether they are individual words or phrases. Moreover, we demonstrate that these domain-specific axes capture relatively significant variances among all axes within the in-domain texts. In this section, we conduct a case study to investigate the efficacy of new semantic axes curated by a large language model (LLM) and quantify temporal changes along these axes. Specifically, we seek to answer: Do these newly introduced axes effectively quantify social framing? How do these axes correspond to real-world events? By addressing these questions, we aim to evaluate the capacity of these axes to reflect shifts in societal narratives over time.

6.1 Background and Data

In March 2018, the U.S. announced sanctions against China under Section 301 of the U.S. Trade Act, citing concerns over China's 'unfair trade practices' (Kwan, 2020). A key argument was that China, despite claiming to support free trade, was in fact implementing trade protectionist policies. To analyze how U.S. news discourse evolved around this issue, we introduce two new semantic axes: *open markets, free trade* vs. *trade barriers, protectionism* and *market economy, capitalism* vs. *planned economy, socialism*. These axes are tailored to reflect the core aspects of the U.S.-China trade conflict, allowing us to track changes in U.S. media framing. The data for this analysis is drawn from the economic and trade-related texts in the NOW corpus, as classified in the previous section.

6.2 Results

Observation 3: ADAPTIVE AXES can model temporal shifts in social stereotypes. The results from 2011 to 2023, shown in Figure 2, reveal clear trends in the evolving trade tensions between China and the U.S. The score trend for the *open markets, free trade* vs. *trade barriers, protectionism* axis shows a marked shift towards the trade barriers side, with a sharp rise in 2018, corresponding precisely to the U.S. sanctions on Chinese goods. The persistently high scores favoring trade barriers after 2018 align with the ongoing nature of these tensions between the two nations.

Similarly, the second axis, *market economy, capitalism* vs. *planned economy, socialism*, reflects the evolving perception of China's economic system. Historically viewed as a hybrid between market and planned economies, with a leaning towards the planned side (Miranda, 2018), the axis shows a significant dip around 2018, followed by a gradual recovery. This pattern corresponds to the heightened focus on accusations of the Chinese government manipulating its economy during the trade conflict. Overall, these results suggest that our pipeline, by focusing on contextual information, successfully integrates human-curated, well-constructed semantic axes and captures real-world social dynamics over time.

7 Discussion

In this paper, we introduce a novel pipeline that leverages text embedding models to encode both augmented semantic axes and news discourse contexts related to target social groups. By combining these elements, we project contextual embeddings onto semantic axes embeddings to gain insights into the underlying stereotypes present within the data.

The study of social stereotypes is inherently complex and multifaceted. Recent research has focused on the evolution of intersectional stereotypes—those that emerge at the intersection of multiple social categories—in various text sources (Charlesworth et al., 2024). We identify that these text sources are themselves highly diverse, spanning a range of social domains such as culture, politics, and economics, as well as repositories like Google Books, COHA, and Common Crawl (Charlesworth et al., 2023). This diversity creates a layered and intricate landscape of social stereotypes within word embedding models. To address this complexity, we systematically categorize the text sources by domain and apply our pipeline to show how representations of social groups differ significantly across these domains. Our findings contribute to the discourse on stereotype analysis by underscoring the nuanced and context-dependent nature of stereotypes embedded in text, thus advancing a more granular understanding of how these representations are framed.

Our pipeline enhances the capabilities of existing semantic axes by using text embedding models to encode arbitrary semantic axes. The high semantic accuracy of these models makes the axes more reliable for capturing subtle meanings and nuances in language. However, our approach still relies heavily on word co-occurrences within texts. For example, in the trade domain, the term *overseas* frequently appears, while *Canada* is closely associated with *north/northern*, as determined by cosine similarity scores. These associations reflect general geographical information rather than stereotypes.

This highlights a crucial limitation: while the pipeline effectively captures broad contextual relationships, it is less equipped to automatically discern stereotypical language from neutral or descriptive terms. More targeted methods will be needed in future work to accurately retrieve stereotype-laden language specifically related to social groups. Such methods could involve refining the selection of semantic axes to focus on traits or attributes typically linked to social biases, or using advanced filtering techniques that distinguish between general context and stereotype-driven discourse. This refinement will allow the pipeline to move beyond identifying generic associations and instead focus on retrieving and analyzing language that specifically conveys stereotypes, thus providing more precise insights into the framing of social groups within various discourse contexts.

Acknowledgement

We acknowledge Tessa Charlesworth and Kaize Ding for their helpful discussions and feedback.

Limitations

We identify two main limitations of this work.

Consistency Metric for Semantic Axes In this work, we use the consistency metric, which measures whether two poles of each semantic axis are well-separated from each other. Although Lucy et al. (2022) showed that models with higher consistency would generally have better human voting preferences, there still lacks one clear metric to evaluate whether one semantic axis is meaningful in both semantic spaces and sociocultural scenarios. Thus, the evaluation in this work only represents the capability of embedding models to separate contrastive semantic terms in vector space. A better and more comprehensive way to construct semantic axes is still needed.

Framing vs. Stereotypes In this work, we use the word 'stereotype' to describe the associations retrieved from semantic axes. We note that there is a potentially ill-defined boundary between stereotypes and the adjacent concept of framing. "Framing" can occur at the level of individual instances, while "stereotype" necessarily refers to a more generalized set of associational biases or rather the large-scale accumulation of instances of framing. In our human annotation task, we evaluate the relevance of associated axes with reference to three concrete sentences, which is therefore potentially better described as framing. Nevertheless, we maintain the vocabulary of stereotyping throughout since our ultimate goal is the extraction of large-scale, generalizable associational biases.

References

Jisun An, Haewoon Kwak, and Yong-Yeol Ahn. 2018. SemAxis: A lightweight framework to characterize domain-specific word semantics beyond sentiment. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2450–2461, Melbourne, Australia. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report.

Marketa Burnett, Beth Kurtz-Costes, Heidi A Vuletich, and Stephanie J Rowley. 2020. The development of academic and nonacademic race stereotypes in african american adolescents. *Developmental Psychology*, 56(9):1750.

Tessa E. S. Charlesworth, Aylin Caliskan, and Mahzarin R. Banaji. 2022. Historical representations of social groups across 200 years of word embeddings from google books. *Proceedings of the National Academy of Sciences*, 119(28):e2121798119.

Tessa E S Charlesworth, Kshitish Ghate, Aylin Caliskan, and Mahzarin R Banaji. 2024. Extracting intersectional stereotypes from embeddings: Developing and validating the Flexible Intersectional Stereotype Extraction procedure. *PNAS Nexus*, 3(3):pgae089.

Tessa ES Charlesworth, Nishanth Sanjeev, Mark L Hatzenbuehler, and Mahzarin R Banaji. 2023. Identifying and predicting stereotype change in large language corpora: 72 groups, 115 years (1900–2015), and four text sources. *Journal of Personality and Social Psychology*.

Xiaomei Chen, Daniel F Vukovich, Xueping Zhong, Megan Ferry, Lisa Rofel, Aili Mu, Haomin Gong, Arif Dirlik, and Hai Ren. 2012. *China and new left visions: Political and cultural interventions*. Lexington Books.

Mark Davies. 2012. Expanding horizons in historical linguistics with the 400-million word corpus of historical american english. *Corpora*, 7(2):121–157.

Mark Davies. 2022. Corpus of News on the Web (NOW).

DeepSeek-AI, Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Hao Yang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jin Chen, Jingyang Yuan, Junjie Qiu, Junxiao Song, Kai Dong, Kaige Gao, Kang Guan, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruizhe Pan, Runxin Xu, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Size Zheng, T. Wang, Tian Pei, Tian Yuan, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Liu, Xin Xie, Xingkai Yu, Xinnan Song, Xinyi Zhou, Xinyu Yang, Xuan Lu, Xuecheng Su, Y. Wu, Y. K. Li, Y. X. Wei, Y. X. Zhu, Yanhong Xu, Yanping

Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Zheng, Yichao Zhang, Yiliang Xiong, Yilong Zhao, Ying He, Ying Tang, Yishi Piao, Yixin Dong, Yixuan Tan, Yiyuan Liu, Yongji Wang, Yongqiang Guo, Yuchen Zhu, Yuduan Wang, Yuheng Zou, Yukun Zha, Yunxian Ma, Yuting Yan, Yuxiang You, Yuxuan Liu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhewen Hao, Zhihong Shao, Zhiniu Wen, Zhipeng Xu, Zhongyu Zhang, Zhuoshu Li, Zihan Wang, Zihui Gu, Zilin Li, and Ziwei Xie. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model.

Anjalie Field and Yulia Tsvetkov. 2019. Entity-centric contextual affective analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2550–2560, Florence, Italy. Association for Computational Linguistics.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Nancy M Henley. 1989. Molehill or mountain? what we know and don't know about sex bias in language. In *Gender and thought: Psychological perspectives*, pages 59–78. Springer.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Chi Hung Kwan. 2020. The china–us trade war: Deep-rooted causes, shifting focus and uncertain prospects. *Asian Economic Policy Review*, 15(1):55–72.

Xianming Li and Jing Li. 2023. Angle-optimized text embeddings. *arXiv preprint arXiv:2309.12871*.

Jindřich Libovický. 2023. Is a prestigious job the same as a prestigious country? a case study on multilingual sentence embeddings and European countries. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1000–1010, Singapore. Association for Computational Linguistics.

Hauke Licht. 2023. Cross-lingual classification of political texts using multilingual sentence embeddings. *Political Analysis*, 31(3):366–379.

Li Lucy, Divya Tadimeti, and David Bamman. 2022. Discovering differences in the representation of people using contextualized semantic axes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3477–3494, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Marcus Marktanner. 2014. The social market economy–assembled in germany, not made in germany. *The Euro Atlantic Union Review*, 1(0):77–113.

Binny Mathew, Sandipan Sikdar, Florian Lemmerich, and Markus Strohmaier. 2020. The polar framework: Polar opposites enable interpretability of pre-trained word embeddings. In *Proceedings of The Web Conference 2020*, WWW '20, page 1548–1558, New York, NY, USA. Association for Computing Machinery.

Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024. Sfr-embedding-mistral: Enhance text retrieval with transfer learning. Salesforce AI Research Blog.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

George A. Miller. 1995. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41.

Jorge Miranda. 2018. How china did not transform into a market economy. *Non-market Economies in the Global Trading System: The Special Case of China*, pages 65–97.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

August Nilsson, J. Malte Runge, Oscar N E Kjell, Nikita soni, Adithya V Ganesan, and Carl V Nilsson. 2024.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

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.

James A Russell. 2003. Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145.

Margaret Shih, Todd Pittinsky, and Amy Trahan. 2006. Domain-specific effects of stereotypes on performance. *Self and Identity*, 5(1):1–14.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: an open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4444–4451. AAAI Press.

William Timkey and Marten van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4527–4546, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Improving text embeddings with large language models.

John E Williams and Deborah L Best. 1990. *Measuring sex stereotypes: A multination study, Rev.* Sage Publications, Inc.

Toshio Yamazaki. 2019. Anti-monopoly policy and new system of large corporate groups in germany after world war ii. *Banking, Economics and Business Research (ICMABEBR-19)*.

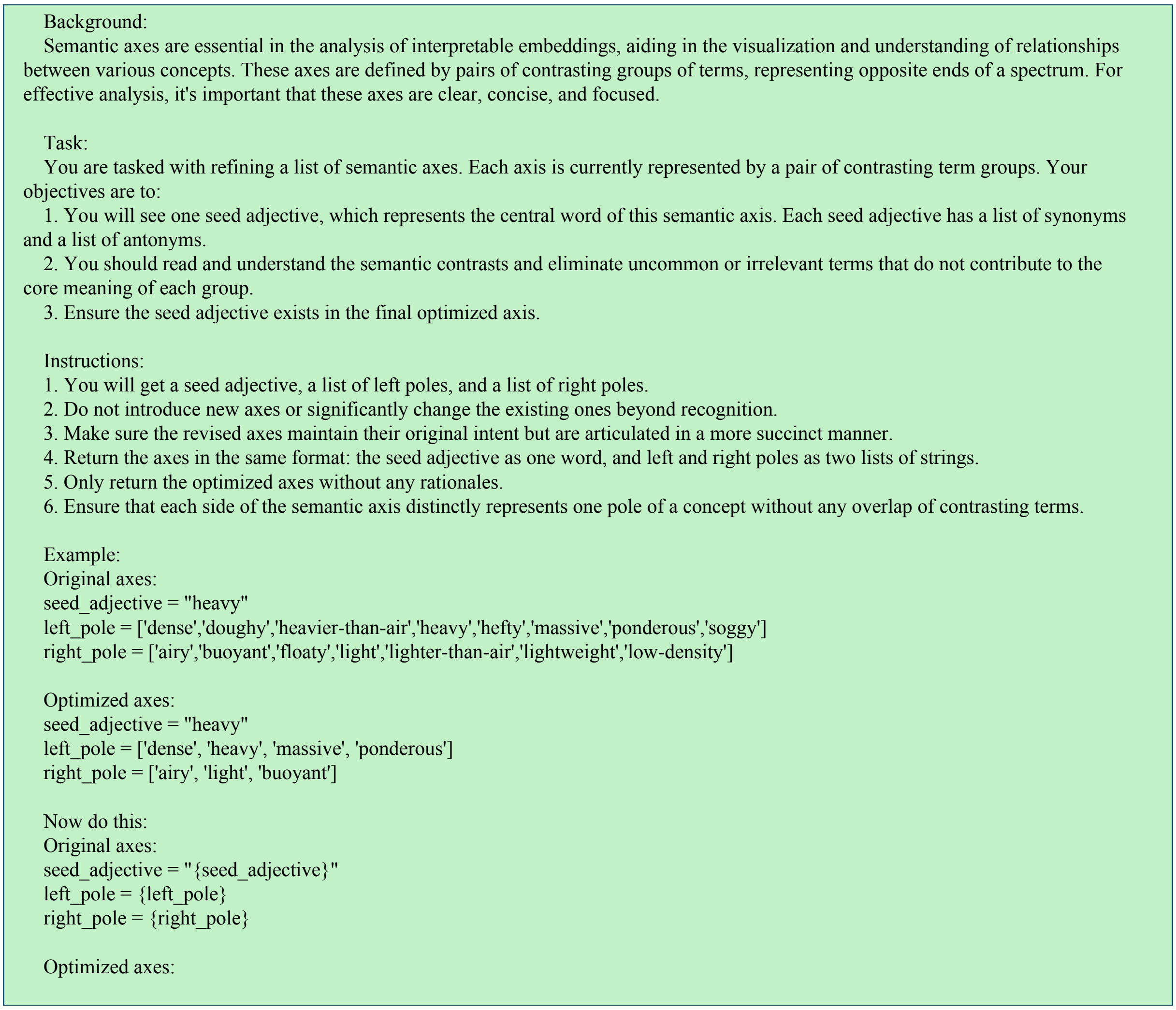
Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.

Kevin Honglin Zhang. 2006. Is china the world factory? In *China as the world factory*, pages 279–295. Routledge.

A Prompt Templates and Interfaces

In this section, we show the prompt templates and the annotation interface we use throughout our study.

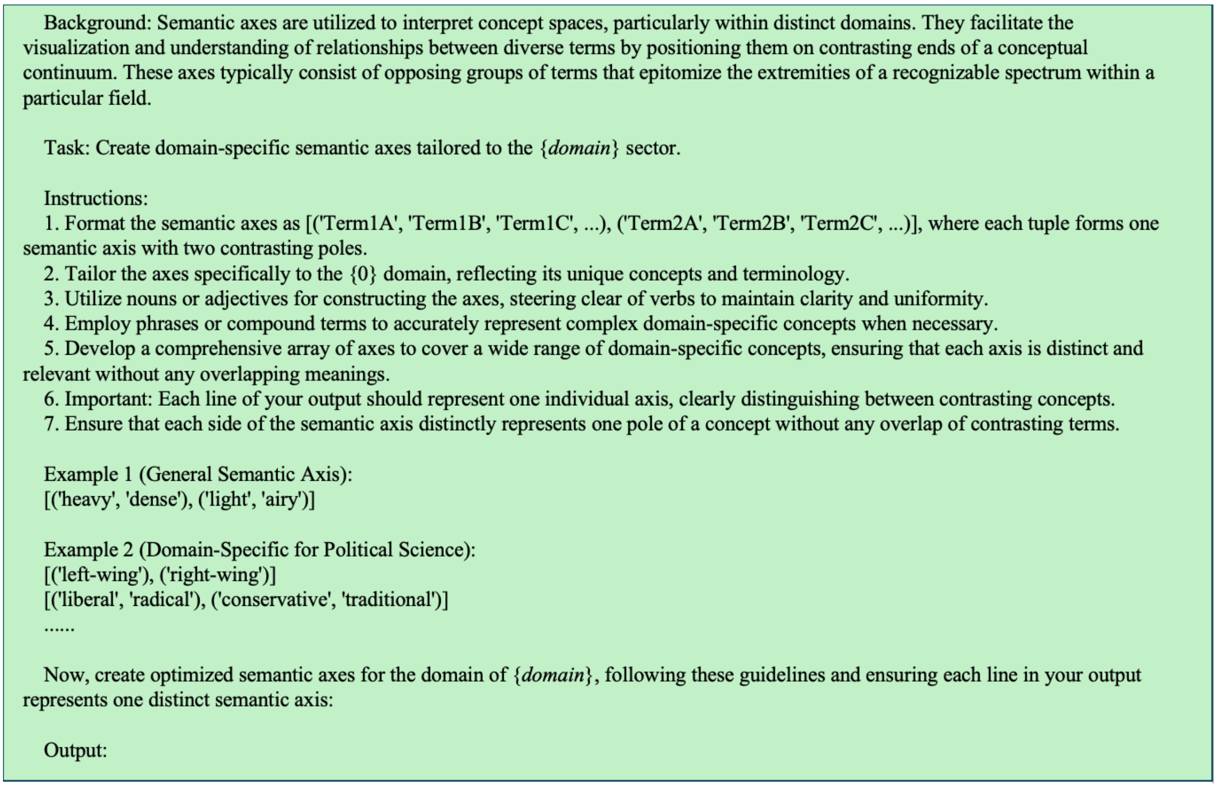
A.1 The Prompt Template for Pruning Semantic Axes



```
Background:
Semantic axes are essential in the analysis of interpretable embeddings, aiding in the visualization and understanding of relationships between various concepts. These axes are defined by pairs of contrasting groups of terms, representing opposite ends of a spectrum. For effective analysis, it's important that these axes are clear, concise, and focused.

Task:
You are tasked with refining a list of semantic axes. Each axis is currently represented by a pair of contrasting term groups. Your objectives are to:
1. You will see one seed adjective, which represents the central word of this semantic axis. Each seed adjective has a list of synonyms and a list of antonyms.
2. You should read and understand the semantic contrasts and eliminate uncommon or irrelevant terms that do not contribute to the core meaning of each group.
3. Ensure the seed adjective exists in the final optimized axis.

Instructions:
1. You will get a seed adjective, a list of left poles, and a list of right poles.
2. Do not introduce new axes or significantly change the existing ones beyond recognition.
3. Make sure the revised axes maintain their original intent but are articulated in a more succinct manner.
4. Return the axes in the same format: the seed adjective as one word, and left and right poles as two lists of strings.
5. Only return the optimized axes without any rationales.
6. Ensure that each side of the semantic axis distinctly represents one pole of a concept without any overlap of contrasting terms.

Example:
Original axes:
seed_adjective = "heavy"
left_pole = ['dense','doughy','heavier-than-air','heavy','hefty','massive','ponderous','soggy']
right_pole = ['airy','buoyant','floaty','light','lighter-than-air','lightweight','low-density']

Optimized axes:
seed_adjective = "heavy"
left_pole = ['dense', 'heavy', 'massive', 'ponderous']
right_pole = ['airy', 'light', 'buoyant']

Now do this:
Original axes:
seed_adjective = "{seed_adjective}"
left_pole = {left_pole}
right_pole = {right_pole}

Optimized axes:
```

Figure 3: The prompt template to prune the existing semantic axes.

A.2 The Prompt Template for Generating Domain-specific Semantic Axes



Background: Semantic axes are utilized to interpret concept spaces, particularly within distinct domains. They facilitate the visualization and understanding of relationships between diverse terms by positioning them on contrasting ends of a conceptual continuum. These axes typically consist of opposing groups of terms that epitomize the extremities of a recognizable spectrum within a particular field.

Task: Create domain-specific semantic axes tailored to the {*domain*} sector.

Instructions:
1. Format the semantic axes as [('Term1A', 'Term1B', 'Term1C', ...), ('Term2A', 'Term2B', 'Term2C', ...)], where each tuple forms one semantic axis with two contrasting poles.
2. Tailor the axes specifically to the {0} domain, reflecting its unique concepts and terminology.
3. Utilize nouns or adjectives for constructing the axes, steering clear of verbs to maintain clarity and uniformity.
4. Employ phrases or compound terms to accurately represent complex domain-specific concepts when necessary.
5. Develop a comprehensive array of axes to cover a wide range of domain-specific concepts, ensuring that each axis is distinct and relevant without any overlapping meanings.
6. Important: Each line of your output should represent one individual axis, clearly distinguishing between contrasting concepts.
7. Ensure that each side of the semantic axis distinctly represents one pole of a concept without any overlap of contrasting terms.

Example 1 (General Semantic Axis):
[('heavy', 'dense'), ('light', 'airy')]

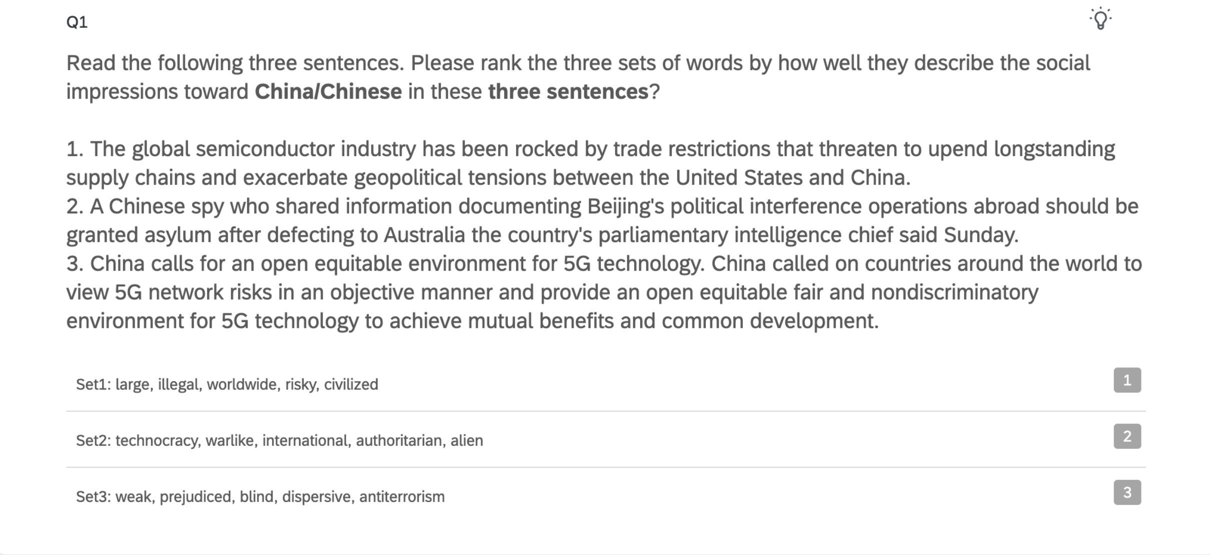
Example 2 (Domain-Specific for Political Science):
[('left-wing'), ('right-wing')]
[('liberal', 'radical'), ('conservative', 'traditional')]
......

Now, create optimized semantic axes for the domain of {*domain*}, following these guidelines and ensuring each line in your output represents one distinct semantic axis:

Output:

Figure 4: The prompt template to generate domain-specific semantic axes.

A.3 The Annotation Interface of Human Judgments



Q1

Read the following three sentences. Please rank the three sets of words by how well they describe the social impressions toward **China/Chinese** in these **three sentences**?

1. The global semiconductor industry has been rocked by trade restrictions that threaten to upend longstanding supply chains and exacerbate geopolitical tensions between the United States and China.
2. A Chinese spy who shared information documenting Beijing's political interference operations abroad should be granted asylum after defecting to Australia the country's parliamentary intelligence chief said Sunday.
3. China calls for an open equitable environment for 5G technology. China called on countries around the world to view 5G network risks in an objective manner and provide an open equitable fair and nondiscriminatory environment for 5G technology to achieve mutual benefits and common development.

Set1: large, illegal, worldwide, risky, civilized	1
Set2: technocracy, warlike, international, authoritarian, alien	2
Set3: weak, prejudiced, blind, dispersive, antiterrorism	3

Figure 5: This is the interface we use for human annotators to rank the framing from various pipelines. The annotators are asked to rank 1/2/3 in this task.

B The Evaluation of Affective Information Understanding

In this section, we conduct a side experiment to evaluate whether entity-masked context embedding could recover affective information better than token-based embeddings. Given this study is not directly associated with stereotypes, we attach the experimental procedure in the Appendix.

B.1 Background about Affective Lexicons

Russell (2003) classified word meaning into three factors - *valence* (positiveness – negativeness / pleasure – displeasure), *arousal* (active - passive), and *dominance* (dominant - submissive) and this has long been the general principle to construct affective lexicons. For example, Mohammad (2018) constructed a comprehensive VAD lexicon with scores for 20,000 English words ranging from 0 to 1 using best-worst scaling.

B.2 Can Large Language Models Generate Human-aligned Affective Information?

To verify whether LLMs can generate sentences with accurate valence scores for target words, we first need to evaluate whether LLMs can generate human-aligned affective information. To achieve this, we design a multi-LLM-in-the-loop strategy to guarantee that our annotations are not closely inclined to one specific LLM's values and emotional understandings. We use six seed annotations to elicit LLMs' emotional reasoning capacities, then we ask multiple LLMs (QWen1.5-72B (Bai et al., 2023), GPT-4 (OpenAI et al., 2024), LLaMA2-70B-Instruct (Touvron et al., 2023), DeepSeek-Chat-V2 (DeepSeek-AI et al., 2024), Mistral-7B (Jiang et al., 2023)) to generate affective values for all semantic units. Finally, we prune the highest and the lowest values for each semantic unit and average to get the multi-LLM-collaboration affective scores. Our prompts are detailed in Figure 6.

To validate whether our newly generated affective scores accurately reflect human-level affective understanding, we randomly choose 50 semantic units not in the original VAD lexicon and ask three individual annotators to perform similar reasoning procedures to what LLMs do. We average these three annotations to get human-annotated affective scores for three dimensions. Then, we calculate Pearson's correlation coefficient to show whether multi-LLM collaboration generates human-aligned affective values.

Our results are shown in Table 6. The correlation for *arousal* is the highest at 0.86, and the lowest is *dominance* at 0.78. The statistical significance suggests that multi-LLM could approximate human-level affective annotations really well. Similarly, Nilsson et al. (2024) reported results of using LLMs to automatically annotate implicit motives, suggesting that LLMs could generate as accurate as humans and 99% cheaper. Our results further contribute to this field and show the great potential of using LLMs for quantifying affective information.

Affective Dimensions	Correlation
Valence	0.82^{**}
Arousal	0.86^{**}
Dominance	0.78^{**}

Table 6: The Pearson correlation score between LLM judgments and human judgments. The asterisks represent statistical significance.

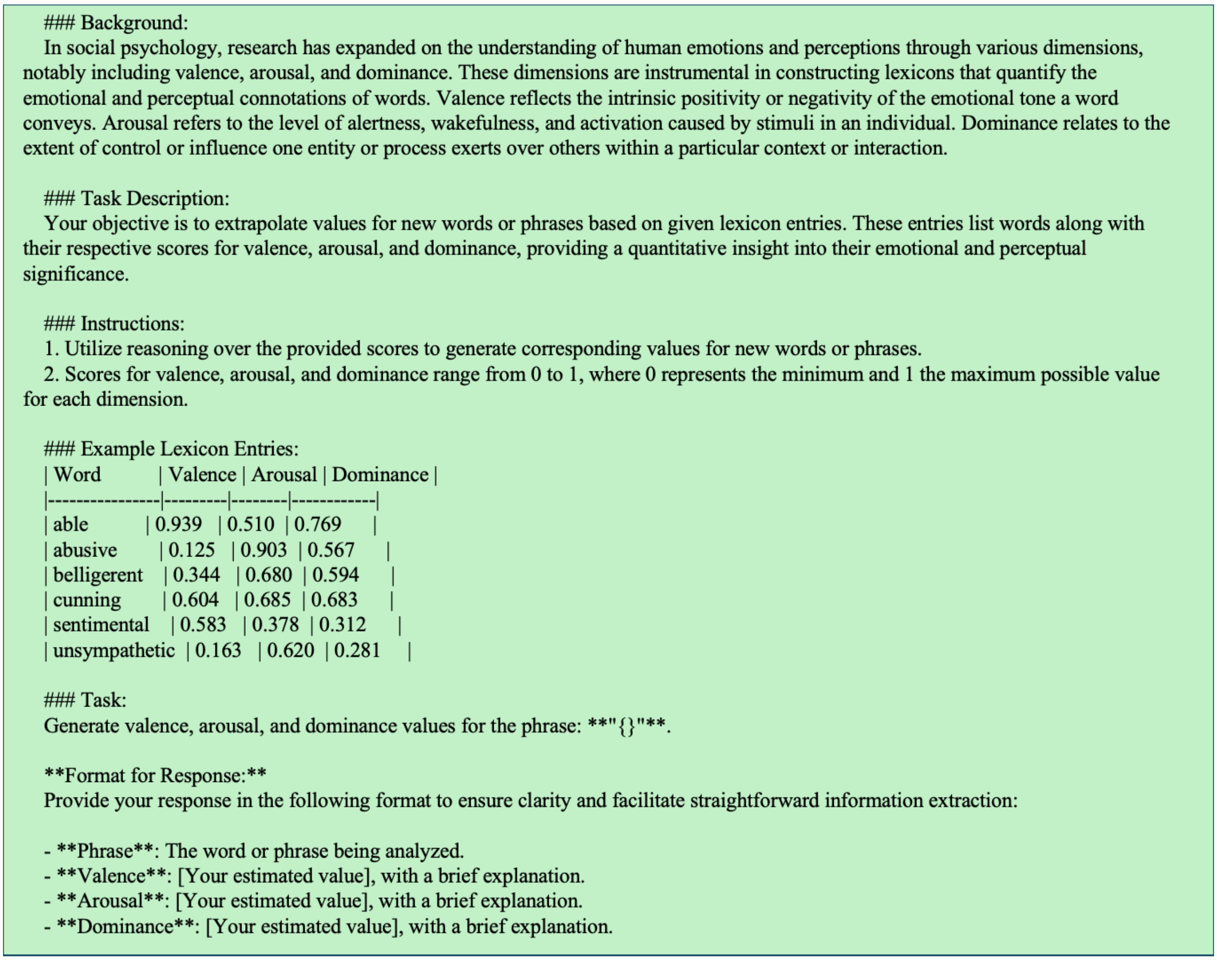
Regression			
Model	Valence Score		
	general	non-general	total
BERT-Large	0.62	0.39	0.44
UAE-Large-V1	0.64	0.58	0.60

Table 7: The Pearson correlation between predicted valence scores and silver valence scores.

B.3 Validation of Affection Understanding

If embedding contexts leads to robust affective understanding, contextual text representations should approximate valence scores, which measure the intrinsic attractiveness or averseness of a word, at least as accurately as target token representations. In this study, we first demonstrate that LLMs can generate human-aligned valence, arousal, and dominance scores. We then randomly sample 1,000 words from our semantic axes and prompt LLMs to generate two sentences under two scenarios: (1) a sentence reflecting the general use of the word, and (2) a sentence reflecting a human-curated valence score, randomly sampled from the other half of the (0,1) range to represent a non-general use of the word. For example, for the word *abandon*, the two sentences are: (1) The feeling of being **abandoned** by someone you love can be utterly devastating, filling your heart with sorrow and despair. *(Valence - 0.05)*; (2) When you decide to **abandon** a toxic relationship, it marks the beginning of a positive transformation and personal growth. *(Valence - 0.7)* in which the same word in different sentences conveys different affections. We manually checked the generated 1,000 sentences and removed 87 inappropriate sentences.

We get the target word's token embedding and the contextual embedding around the target word. Then, two kernel ridge regression models will be fitted on 700 training sentences. We use the adjusted R^2 to determine which better predicts the affective annotations on the remaining 213 sentences. Our results are shown in Table 7, indicating that well-trained text embedding models could predict affective annotations better than token-based embeddings, which are not intended for in-domain use. These results also partially correspond to Field and Tsvetkov (2019) and further reveal the potential of using text embedding models to understand specific social stereotypes.



```
    ### Background:
    In social psychology, research has expanded on the understanding of human emotions and perceptions through various dimensions, notably including valence, arousal, and dominance. These dimensions are instrumental in constructing lexicons that quantify the emotional and perceptual connotations of words. Valence reflects the intrinsic positivity or negativity of the emotional tone a word conveys. Arousal refers to the level of alertness, wakefulness, and activation caused by stimuli in an individual. Dominance relates to the extent of control or influence one entity or process exerts over others within a particular context or interaction.

    ### Task Description:
    Your objective is to extrapolate values for new words or phrases based on given lexicon entries. These entries list words along with their respective scores for valence, arousal, and dominance, providing a quantitative insight into their emotional and perceptual significance.

    ### Instructions:
    1. Utilize reasoning over the provided scores to generate corresponding values for new words or phrases.
    2. Scores for valence, arousal, and dominance range from 0 to 1, where 0 represents the minimum and 1 the maximum possible value for each dimension.

    ### Example Lexicon Entries:
    | Word          | Valence | Arousal | Dominance |
    |----------------|---------|--------|------------|
    | able          | 0.939   | 0.510  | 0.769      |
    | abusive       | 0.125   | 0.903  | 0.567      |
    | belligerent   | 0.344   | 0.680  | 0.594      |
    | cunning       | 0.604   | 0.685  | 0.683      |
    | sentimental   | 0.583   | 0.378  | 0.312      |
    | unsympathetic  | 0.163   | 0.620  | 0.281     |

    ### Task:
    Generate valence, arousal, and dominance values for the phrase: **"{}"**.

    **Format for Response:**
    Provide your response in the following format to ensure clarity and facilitate straightforward information extraction:

    - **Phrase**: The word or phrase being analyzed.
    - **Valence**: [Your estimated value], with a brief explanation.
    - **Arousal**: [Your estimated value], with a brief explanation.
    - **Dominance**: [Your estimated value], with a brief explanation.
```

Figure 6: The prompt to use multiple LLMs to annotate affective dimensions automatically.

C Domain-specific Semantic Axes

Global Trade and Economics:
Free Trade, Open Markets - Protectionism, Trade Barriers
Market Economy, Capitalism - Planned Economy, Socialism
Globalization, International Integration - Localization, Economic Self-sufficiency
Economic Liberalization, Deregulation - State Intervention, Regulation
Innovation, Technological Advancement - Traditionalism, Preservation
Politics and Governance
Authoritarianism, Totalitarianism - Democracy, Republic
Centralization, Federal Authority - Decentralization, Local Autonomy
Political Transparency, Political Openness - Political Secrecy, Political Opaqueness
Individual Rights, Personal Freedom - Collective Good, Social Responsibility
Progressivism, Social Reform - Conservatism, Tradition
Cultural and Education
Cultural Homogeneity, Monoculture - Cultural Diversity, Multiculturalism
Cultural Openness, Inclusivity - Cultural Exclusivity, Preservation
Global Culture, Cross-Cultural Exchange - Local Culture, Indigenous Practices

D Case Study Data Description

Country Name	Total Number of News Articles	Categories	The Number of In-category News Articles
China	63431	Politics and Governance	10550
		Global Trade and Economics	4940
		Culture and Education	7402
		None of the Above	40539
Canada	17694	Politics and Governance	1046
		Global Trade and Economics	1294
		Culture and Education	3804
		None of the Above	11550
Germany	17256	Politics and Governance	2233
		Global Trade and Economics	909
		Culture and Education	3026
		None of the Above	11088
Russia	52377	Politics and Governance	28770
		Global Trade and Economics	1644
		Culture and Education	3906
		None of the Above	18057

Table 8: The descriptive statistics of our data