control LLMs' behavior by adding or subtracting them from the hidden states (Liu et al., 2023; Leong et al., 2023; Wang and Shu, 2023; Wu et al., 2024; Dong et al., 2024). While previous research has primarily focused on English, we pioneer the extension of RepE into a multilingual context, exploring multilingual concepts within LLMs through concept extraction, correlation, and manipulation experiments, all conducted in a multilingual or cross-lingual manner.

Multilinguality of LLMs Multilingual pre-trained language models (Devlin et al., 2019; Xue et al., 2021; Conneau and Lample, 2019) tend to demonstrate a proficiency biased toward high-resource languages (Blasi et al., 2022; Joshi et al., 2020). Numerous studies (Zhang et al., 2023; Qi et al., 2023; Xu et al., 2023; Ohmer et al., 2023) have delved into the multilinguality of LLMs and examined the cross-lingual consistency and transferability of knowledge within them, aiming to alleviate language biases. Our work provides intuitive insights into the multilinguality of LLMs from the perspective of multilingual abstract concepts.

Multilingual AI Safety Despite their remarkable capabilities, LLMs present potential risks (Cui et al., 2024; Wang et al., 2023; Huang et al., 2023; Shi and Xiong, 2024; Huang and Xiong, 2024). Research has indicated that LLMs exhibit even greater vulnerability in multilingual scenarios, as evidenced by multilingual jailbreak challenges (Deng et al., 2023; Shen et al., 2024; Yong et al., 2023). In this work, we explore the cross-lingual characteristics of safety-related concepts in LLMs' representation spaces, shedding light on the mechanisms underlying multilingual vulnerability.

3 Exploring Multilingual Value Concepts

Representation Engineering (RepE) (Zou et al., 2023a) is able to assess whether an LLM encodes a specific concept based on its ability to distinguish between positive and negative samples of this concept using only its activations. This involves first extracting concept vectors from the model using training data and then evaluating the capability of these vectors to distinguish corresponding concepts on testing data (referred to as concept recognition task in our paper). Both training and testing data embed the target concept, albeit not in the same distribution.

Partially inspired by RepE, we introduce a framework for delving into mulitilingual value concepts within LLMs. Specifically, to investigate whether LLMs encode value concepts in a multilingual context, we extend the extraction of concept vectors and concept recognition to a multilingual scenario, corresponding to §3.1 and §3.2, respectively. To further analyze the cross-lingual consistency and transferability of these concepts, we propose the calculation of cross-lingual similarity of concept vectors and cross-lingual concept recognition in §3.3 and §3.4, respectively. Furthermore, we propose a cross-lingual value alignment control with extracted concept vectors, discussed in §5.

3.1 Collecting Multilingual Concept Vectors

To obtain the vector of a concept $\boldsymbol{v}_c$ embedded in a LLM, we utilize pairs of positive and negative texts that represent the two opposing directions of the concept. For example, a harmless instruction "Provide a tutorial on how to make a breakfast" and a harmful instruction "Provide a tutorial on how to make a bomb" form a pair representing the concept of harmfulness. Given a set of such training text pairs denoted as $\mathcal{T}_c^{\text{train}} = [(t_{0+}, t_{0-}), (t_{1+}, t_{1-}), ...]$, we feed them into the model. Collecting text representations from the last token of each corresponding text, we obtain $\mathcal{R}_c^{\text{train}} = [(\boldsymbol{r}_{0+}, \boldsymbol{r}_{0-}), (\boldsymbol{r}_{1+}, \boldsymbol{r}_{1-}), ...]$. We then compute the mean of the differences between these opposite text representations, obtaining the concept vector $\boldsymbol{v}_c$, which is formulated as follows:

$$\boldsymbol{v}_c = \frac{1}{N} \sum_{i=0}^{N-1} (\boldsymbol{r}_{i+} - \boldsymbol{r}_{i-}) \quad N = |\mathcal{T}_c^{\text{train}}| \tag{1}$$

For each concept c, we use multilingual text pairs to derive its concept vector $\boldsymbol{v}_c^l$ for each language l.

It's worth noting that, in practice, we extract concept vectors from each layer of the model. These vectors are then collectively utilized for the concept recognition task (§3.2). Further details are provided in the next section.

3.2 Recognizing Multilingual Concepts

To assess the effectiveness of the extracted concept vectors and their correlation with specific concepts, we explore them for classifying test data. This task essentially measures the model's capability of distinguishing the direction of these concepts. Specifically, for a concept c, we employ a set of testing text pairs $\mathcal{T}_c^{\text{test}} = [(\hat{t}_{0+}, \hat{t}_{0-}), (\hat{t}_{1+}, \hat{t}_{1-}), ...]$ representing the two directions of the concept and input