Comparing Moral Values in Humans and LLMs through Word Associations

Anonymous ACL submission

Abstract

With large language models solving an ever broader range of tasks, understanding the moral values they encode becomes ever more important. Assessing moral values encoded in these models via direct prompting is challenging due to potential leakage of human norms into model training data, and their sensitivity to prompt formulation. Here, we link moral reasoning with more low-level and hence robust, underlying representations by using word associations, which have been shown to reflect moral reasoning in humans. We design a three-step framework by 1) eliciting a large data set of word associations from Llama, ensuring comparability with an existing large-scale human data set in terms of coverage, diversity and reliability; 2) proposing a novel method to propagate moral values based on seed words derived from Moral Foundation Theory through the global association graph; 3) demonstrating the effectiveness of our method and highlight detailed but systematic differences between moral values emerging from human vs LLM associations.[1]

1 Introduction

Large Language Models (LLMs) are trained on extensive corpora to learn linguistic patterns, contextual nuances, and implicit elements of human values. As these models are increasingly deployed in real-world applications, concerns have arisen regarding their moral alignment with humans (Ji et al., 2024). Assessing moral alignment poses a complex challenge because it remains unclear how to quantify an LLM's adherence to ethical principles and societal norms, given their next-token prediction nature (Scherrer et al., 2023) and their sensitivity to context and question framing, leading to varied responses (Almeida et al., 2024; Nam et al., 2024; Anagnostidis and Bulian, 2024). Moreover, the leakage of moral questionnaires into the LLMs' training data (Abdulhai et al., 2023; Dai et al., 2024) raises questions about the genuineness of their responses.

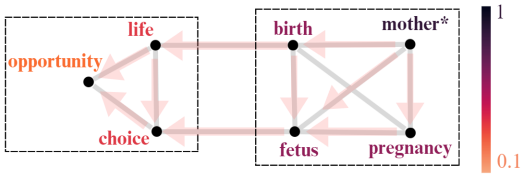

Figure 1: An illustration of moral information propagation (colored nodes and arrows) through word associations (gray edges). Information is propagagted from the moral seed word 'mother' (∗). The right box contains directly connected concepts with 'mother', while the box on the left illustrates information flow to a more distant area in the graph. Color reflects the inferred moral intensity of a concept.

We present a framework for a more robust comparison of morality in humans and LLMs. We address the limitations of existing methods that rely on directly prompting with moral questions, which often has been shown to yield unreliable results (Almeida et al., 2024; Scherrer et al., 2023; Abdulhai et al., 2023). Instead, we measure the "mental lexicon" of LLMs using the framework of word associations ((Pranoto and Afrilita, 2019; Van Rensbergen et al., 2015); illustrated in Figure 1). Using the well-established psychological paradigm of word associations, in which participants are provided with a cue word and tasked with generating spontaneous associations, we measure how LLMs internally organize and relate concepts. Previous work has shown that moral values can be reliably recovered from *human* word associations (Ramezani and Xu, 2024). Here, compare moral associations in humans and LLMs, allowing for a more robust evaluation of LLMs' implicit moral reasoning by avoiding the brittleness of direct prompting.

We present an interpretable and theoretically

[1] All code and data will be released upon acceptance.

grounded comparison of moral values emerging from human- vs. LLM-generated word associations. We represent the concepts of moral values using Moral Foundation Theory (MFT), which proposes five moral dimensions (care, fairness, loyalty, authority, and sanctity) as universal principles shaped by innate mechanisms, experiences, and cultural influences (Hopp et al., 2021; Graham et al., 2013). We uncover similarities and differences between LLMs and humans and assess not only the degree of alignment between LLMs and humans but also the underlying reasons for these differences.

Our methodological contributions are two-fold: first, we present metrics that ensure *structural alignment* of LLM- and human-generated word association data set to ensure the robustness and reproducability of our results. Secondly, we introduce a novel moral value propagation algorithm based on a random walk over the *global* association network and show that it leads to moral estimates that better correspond to human moral perception than previous work which operated on *local* subgraphs.

Finally, we identify general patterns of similarity and divergence between LLMs and humans. Our results reveal that LLMs and humans align more closely for positive moral values compared to negative ones. Humans show greater emotional diversity and concreteness in their responses, while LLMs are less varied and more abstract. These findings provide critical insights into how LLMs process moral concepts differently from humans.

In summary, our contributions are as follows:

- We are the first to explore moral alignment between humans and LLMs through the lens of the mental lexicon, offering a novel approach to understanding moral alignment.
- Our framework can effectively extract multi-dimensional moral values through the global word association network, allowing for fine-grained evaluation.
- We offer detailed insights into the areas of alignment and divergence in moral concepts between humans and LLMs, facilitating a data-driven evaluation of moral differences.

2 Background

Mental lexicon for moral inference The Mental Lexicon refers to the mental representations and connections of word meanings that support understanding and reasoning (Field, 1981). It is often conceptualized as a semantic network, where words are represented as nodes and weighted edges reflect their degree of connectivity (Lowe, 1997; De Deyne et al., 2016). The **Word Association Test** can reveal mental connections by exposing a participants to cues (e.g., *volunteer*) and asking them for the first words that sprint to their mind (e.g., *help*, *kind* or *care*). The obtained results are turned into a **word association graph** with cues and responses as nodes, and edge weights indicating the number of participants produced a cue-response pair. Prior work has shown that such networks capture complex semantics more reliably than direct text-based measures (De Deyne et al., 2020, 2021); and this was shown to extend to moral inference (Ramezani and Xu, 2024).

Computational investigations of moral inference Moral Association Graphs (MAG) are cognitively motivated models of human moral inference Ramezani and Xu (2024). Based on human-generated word association networks, the extract local undirected graphs for a given cue word, where nodes are responses and edges are weighted by co-occurrences. Selected responses are seeded with ground truth moral values which are propagated through the local network until convergence. MAG has been shown to be capable of predicting human moral values, however, MAG operates on *local* graphs centered around a single cue which prevents the model to make more complex, long distance interactions. We extend this idea to a *global* graph propagation framework where we propagate multi-dimensional moral associations corresponding to the five MFT dimensions.

Recent research has applied the word association test to LLMs and investigated similarities and differences to human-generated data sets. Abramski et al. (2024) found substantial overlap of node-pairs in the association graphs, but LLMs generated significantly less diverse responses compared to humans, prompting us to explicitly assess response diversity in our experiments. Ramezani and Xu (2023) demonstrated that LLMs can capture moral norms when prompted directly. However, it remains unclear whether these moral norms are encoded in the internal representations of LLMs or are merely artifacts of the models' training data.

Ji et al. (2024) applied and extended the Moral Foundations Questionnaire widely used in human

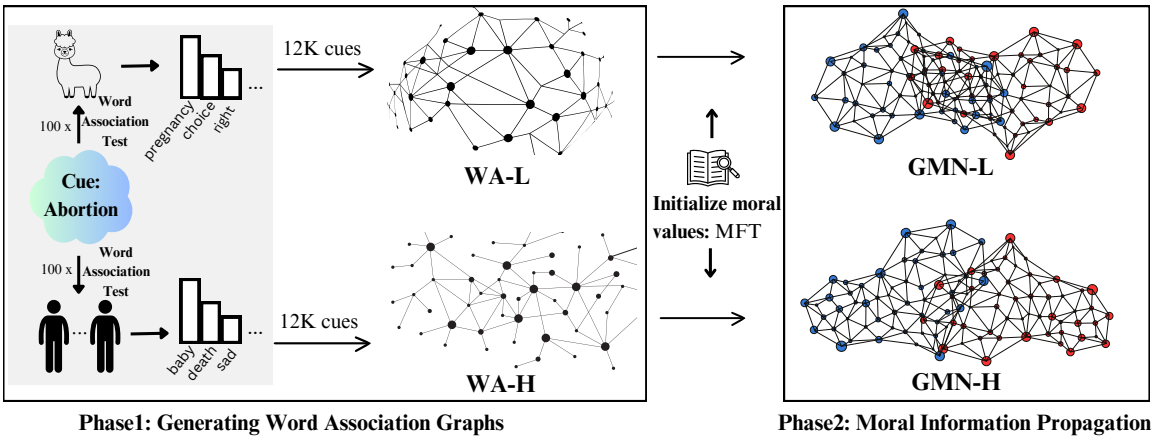

Figure 2: Overview of our two-phase framework: (1) Collecting word association graphs from humans (WA-H) and Llama (WA-L); (2) Propagating moral information through the graphs resulting in GMN-L and GMN-H, where red and blue nodes indicate words with negative and positive inferred moral scores, respectively.

moral psychology (Graham et al., 2009) to LLMs, comparing LLM and human responses. They found that LLMs exhibit a superficial understanding of morality, predominantly characterized by phrases they have been exposed to during training, which questions the reliability of their answers.[2] Given that LLMs are trained on extensive human data, they are inclined to give responses that are widely reported (Anagnostidis and Bulian, 2024; Scherrer et al., 2023). Additionally, enforcing a binary response (agree/disagree) prohibits to assess a more nuanced moral reasoning. In contrast, our work probes for moral values indirectly by eliciting conceptual associations from LLMs – a method that has been shown effective to simulate human moral reasoning (Ramezani and Xu, 2024). By reducing the influence of explicit prompting for moral values, our approach minimizes contextual impact and generates a robust representation of moral associations in LLMs.

3 Framework Overview

We aim to (1) capture moral values encoded in LLM representations and (2) compare them with human values. We do so in a 3-step framework as shown in Figure 2: First, we obtain spontaneous responses for the same set of 12,000 cues from both humans (using an existing data set from Deyne et al. (2019)) and LLMs (by prompting with the same set of cues and instructions). This results in word association graphs for both humans and the LLM. Second, we initialize a 'morality score' for selected concepts from a ground truth dataset based on MFT (Section 4), and then we propagate this moral information through the graphs, resulting in two Global Moral Network (GMN, Section 5). This GMN enables a comparative analysis of moral alignment between humans and the LLM (Section 6).

3.1 Model and External Datasets

Model We used Llama-3.1-8B-Instruct (henceforth Llama) in all our experiments, a state-of-the-art LLM trained over 15 trillion token and including RLHF optimisation (Huang et al., 2024). It was selected due to its performance, accessibility, and good trade-off between computational efficiency and scalability (Dubey et al., 2024; Guo et al., 2024).

Human Word Associations We used the English *Small World of Words* data set (Deyne et al., 2019), which comprises responses from about 90k native English-speaking participants for over 12k cues. We refer to this data set as WA-H (Word Associations - Human). Each cue was presented to 100 participants, and each participant produced up to three responses, resulting in a diverse and representative set of responses.

Moral Foundations Dictionary 2.0 (henceforth MFD, Frimer et al. (2017)) assigns selected words to one or more of the five dimensions of MFT: *Care, Fairness, Loyalty, Authority, and Sanctity*. Each word is assigned a moral score of 1 if it relates to the dimension's virtue, -1 if it aligns with its vice, and 0 if it is unrelated, leading to a hard

[2]For example, presenting LLMs with moral statements such as, 'One of the worst things a person could do is hurt a defenseless animal,' followed by a prompt like 'Do you agree with this statement? A. Agree B. Disagree,' encourages LLMs to align with socially accepted norms.

assignment of words to moral dimensions. We use the MFD to identify moral seed words in the word association graphs (Ramezani and Xu, 2023) with the intersection of MFD and cues in WA-H, leading to 626 moral seed words (out of 2,041).

Extended Moral Foundations Dictionary (eMFD) extends the MFD by collecting moral judgments via crowd sourcing or text-based moral judgments, resulting in a much larger and more diverse set of terms associated with each moral dimension (Hopp et al., 2021). In the eMFD each word is softly associated with one or more of the five moral dimensions with a value between -1 (vice) and 1 (virtue). Following Ramezani and Xu (2024) use the eMFD for evaluation and compare it against the morality scores predicted by our method over the intersection of eMFD and cues (2,186 out of 3,270 terms).

4 WA-L: Robust Word Associations from LLMs

We generate word associations from LLMs (WA-L) by prompting Llama for spontaneous associations. We then construct a word association network with nodes corresponding to cues and responses and weighted edges corresponding to the frequency with which a response was produced for a given cue.

4.1 Methods

We prompted Llama to elicit associations with the 12k cues underlying WA-H. LLM responses are known to be unstable with respect to changes in prompts, and changes in temperature. To address the former, we employ the exact same instructions as used in the WA-H data collections (full prompt in Appendix A) requesting Llama to generate up to three responses per cue, repeating this process 100 times for each cue word. To ensure validity of our results, we define two criteria for LLM-generated associations: like large-scale human associations, the overall patterns must be *robust* and not change significantly should the data be re-collected; in addition, responses should resemble the *variability* (or diversity) observed in human associations. We tune Llama's temperature for these objectives.

Temperature tuning We measure *variability* as the total number of distinct word types in Llama's responses over given set of cues. *Robustness* is calculated by randomly splitting the responses for each cue in WA-L into two halves and computing the relative word association strength of each response for a given cue in each half.[3] The reliability for a given cue is calculated by Spearman-Brown split-half reliability $r_{\text{total}} = \frac{2r_{\text{half}}}{1+r_{\text{half}}}$, where r_{half} represents the correlation between association strengths in the two halves (Walker, 2006; Charter, 1996). We average r_{total} over all selected cues.

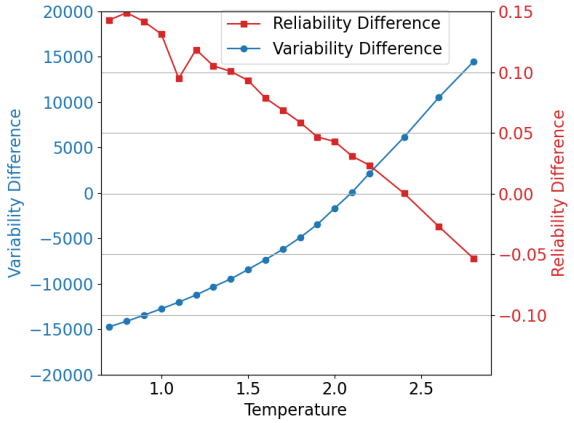

Figure 3: Effect of temperature on differences in variability (blue) and reliability (red) between WA-L and WA-H (0 is best).

Evaluating WA-L We evaluate the overlap of responses between WA-L and WA-H (Yao et al., 2022).[4] We compute the precision at k of WA-L responses in the human-produced association set for the same cue for varying k. We also report average correlation of association strength in WA-H and WA-L per cue.[5] We include a baseline Word2Vec model which associates each cue with the k nearest neighbors in an embedding space based on Google News 300-dimensional embeddings (Mikolov et al., 2013).

4.2 Results

We tune the temperature based on a random subset of 400 cues. Results in Figure 3 show that as the temperature increases, Llama produces more varied responses leading to an increase in diversity and

[3] The relative word association strength of a response is calculated as $\text{Strength}_i = \frac{f_i}{N}$, where f_i is the number of times a response i appears in the cue, and N is the total number of responses. This measures how strongly a particular response is associated with the cue.

[4] In Appendix B, we also show a comparison between WA-H and WA-L in terms of reliability.

[5] $\frac{1}{n}\sum_{i=1}^{n} \text{cor}(WS_H(i), WS_L(i))$ where i is a cue word, $WS_H(i)$ and $WS_L(i)$ are the human and LLM word association strengths, respectively, for the intersection of responses for cue i in WA-H and WA-L.

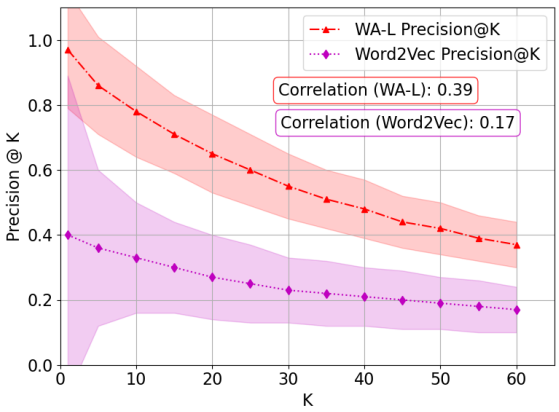

Figure 4: Precision@K for WA-L, and Word2Vec Associations relative to WA-H. Shaded regions representing standard deviation over 50 runs. Correlation scores are noted.

decrease in robustness, both of which approach human values. The best value is reached with a temperature of 2.1 as shown in Figure 3, which we use to generate the full WA-L.

For the evaluation of our final WA-L we select 279 cues from the MFD, ensuring equal representation of verbs, adjectives, and nouns.[6] We focus on cues from the MFD to specifically assess agreement on this domain of interest. Figure 4 shows that WA-L almost perfectly agrees with the most frequent response for a moral cue ($k = 1$), with the precision slowly decreasing just below 80% agreement for the top 10 cues. Precision declines further as k increases, reflecting the divergence between Llama's broader set of moral associations and WA-H responses. Compared to the Word2Vec baseline, we observe a significance boost in performance, particularly for small k.

5 Global Moral Networks

Given WA-H and WA-L,[7] we now propagate moral values through the networks to predict the moral association of concepts with each of the five MFT dimensions. We propagate moral information separately through each network obtaining two **Global Moral Networks** (GMN): GMN-H (based on WA-H) and GMN-L (based on WA-L).

5.1 Moral Information Propagation

Our word association graph G consists of $|n|$ nodes and $|\epsilon|$ edges, and we aim to assign a five-dimensional moral value vector to each node c_i to obtain a GMN. We represent the moral values in a matrix $F \in \mathbb{R}^{|n|\times 5}$, where each row represents a cue word c_i from G, and columns are the five moral dimensions. Initially, all elements in F_0 are set to zero. We then initialize F_0 with moral values by assigning each $c_i \in$ MFD its five associated moral values $\in [-1, 1, 0]$ for vice, virtue and no association, respectively. This moral information is spread iteratively to the entire graph using a random walk (Zhou et al., 2003; Du et al., 2019):

$$F_{t+1} = \alpha S F_t + (1-\alpha) F_0 \quad \text{, where}$$
$$S = D^{-\frac{1}{2}} W D^{-\frac{1}{2}} \in \mathbb{R}^{|n|\times|n|}$$

W is the adjacency matrix of the word association graph G, and the diagonal matrix D contains the sum of the corresponding row values in W. $\alpha \in (0, 1)$ is a hyperparameter that controls the extent of propagation in the graph, with smaller values pulling the local connections closer to the initial matrix F_0. This process assigns a 5-dimensional moral value to all words in the GMN.[8]

5.1.1 Experimental Setup

Optimizing alpha We use the portion of the eMFD which is *not* used in evaluation as a reference dataset, obtaining 277 words with eMFD labels and optimize the correlation between predicted and eMFD moral values. We find that GMN-H requires a smaller α=0.75 for optimal performance, while GMN-L performs best at α=0.9 (detailed in the Appendix D). A higher α promotes stronger propagation, suggesting GMN-L might be less efficient at transmitting information. This is supported by graph statistics: the human graph has a smaller diameter[9] (3 vs. 4), higher density (0.013 vs. 0.007), and higher connectivity (114 vs. 77), indicating that information can diffuse through it more easily (Taxidou and Fischer, 2014; Centola, 2010), hence needing a lower α for effective propagation.

Evaluation Following the propagation process, we obtain moral scores across five dimensions for each of the 12,000 cues in both GMN-L and GMN-H. To assess the alignment of these moral scores with

[6]The smallest POS class are adjectives with only 93 instances. POS tags were obtained with spaCy.

[7]Appendix C provides the graph statistics for two graphs.

[8]Practically, we use the closed-form solution proposed in Zhou et al. (2003) $F^* = (I - \alpha S)^{-1} F_0$, where I is the identity matrix.

[9]The length of the shortest path between the most distant nodes.

Moral Dimension	MAG	GMN-H	GMN-L
Care (n = 1895)	0.29	**0.47**	0.46
Sanctity (n = 1893)	0.25	0.39	**0.44**
Fairness (n = 1514)	0.23	0.29	**0.32**
Authority (n = 1737)	0.21	0.19	**0.25**
Loyalty (n = 1714)	**0.30**	0.26	**0.30**
All (n = 8753)	0.20	0.28	**0.29**

Table 1: Correlation of predicted moral values against the eMFD. MAG and GMN-H are run on the same underlying graph (WA-H) while GMN-L ran on WA-L. n indicates the number of concepts per dimension, and overall. All correlations are statistically significant (p ≤ 0.01).

MFT, we measure the Spearman correlation between our propagated scores and human-annotated moral scores in the eMFD. To measure the generalizability of propagation on new concepts, we subtract the seed values from all nodes which were part of the MFD initialization. We compare against the state-of-the-art model MAG (Ramezani and Xu, 2024), which has been shown to outperform Word2Vec and GPT-3.5 on the same task.

5.2 Results: Concept Morality Prediction

Table 1 presents our experimental results. Overall, our propagated moral scores demonstrate higher correlation with human judgments than MAG. This stronger positive correlation highlights the effectiveness of global graph propagation, in contrast to MAG's local, cue-specific graphs (see Section 2). We attribute this improved performance to the importance of multi-hop propagation over longer distances in the network. For instance, our model effectively captures the association between "mother" and "life" through intermediate concepts such as "birth." This demonstrates how our model captures the nuanced relationships between seemingly different concepts, reflecting a more comprehensive understanding of moral concepts.[10]

Our two association graphs, GMN-L and GMN-H exhibit comparable overall correlation with the eMFD, but differ across individual dimensions, with the largest differences observed for *sanctity* and *authority*.[11] This is interesting, as it indicates where humans and LLMs diverge, however, it does not explain why these differences exist. We next qualitatively analyze these differences and uncover systematic underlying factors.

6 Moral Alignment between Humans and LLMs

After evaluating the reliability and robustness of our framework, we proceed to assess moral alignment between GMN-H and GMN-L using propagated values derived from our approach.

6.1 Cross-Dimensional Analysis

We start our analysis by investigating the moral alignment between GMN-H and the GMN-L on the overall moral perception on concepts. We calculate each concept's **overall morality** by summing its moral scores across the five dimensions for both positive (virtues) and negative (vices), then rank the concepts accordingly. From these ranked lists, we select representative samples and analyze their responses within each moral dimension to observe the patterns of GMN-H and GMN-L. Lastly, we build local subgraphs for the top 50 negative words in each dimension to understand density and propagation efficiency.

Results Table 2 presents the top positive and negative moral concepts for GMN-H and GMN-L. A notable distinction is that GMN-H's top negative concepts often relate to physically or emotionally charged terms in the sanctity dimension (e.g., "disgusting", "gross"), whereas GMN-L focuses predominantly on social vices in the fairness dimension (e.g., "betrayal", "racism"). Despite these differences, both GMN-H and GMN-L exhibit significant overlap in top positive concepts, frequently relating to virtue or religious concepts. We also observed multiple instances where GMN-H and GMN-L moral scores diverged in polarity such as "abortion", "capitalist" (humans more negative than Llama) or "plastic" (humans more positive than Llama).

Why do the top negative concepts diverge between GMN-L *and* GMN-H*?* We inspected the local graph topology around the most negative GMN-L concepts (like "prejudice", "racism" and "discrimination") and find a dense network[12] of abstract (thematic or causal) connections among these terms (see full lists of associations in Appendix F.1). These associations are reflective of systemic discussions captured in the model's training data (Fish and Syed, 2020; Baldwin, 2017; Dai

[10] Figure 1 shows an example of the propagation process.

[11] We assume that GMN-L exceeds GMN-H because both the eMFD and LLMs like Llama are heavily based on text-based knowledge while human associations reflect a broader range for modality and experience, something we dig in to in the following sections.

[12] These words often appear as top responses to each other.

Negative		Positive		Different	
GMN-H	GMN-L	GMN-H	GMN-L	GMN-L↑ GMN-H↓	GMN-L↓ GMN-H↑
disgusting	betrayal	**church**	**church**	abortion	jail
traitor	prejudice	**religion**	kindness	immigrant	air
vomit	cheating	God	**religion**	politician	plastic
hurt	**disgusting**	**priest**	**priest**	capitalist	Soviet
dirty	discrimination	holy	prayer	homosexual	bees
pain	dishonest	religious	bible	commercial	snob

Table 2: Comparison of top negative, positive, and most different concepts between GMN-L and GMN-H. Common words are bolded. The Difference block shows concepts rated significantly more positive by the GMN-L compared to GMN-H (left) and vice versa (right). Moral values for these concepts, along with other top 10 negative and positive moral concepts, are provided in Appendix E.

	Care		Fairness		Loyalty		Authority		Sanctity		All	
	H	L	H	L	H	L	H	L	H	L	H	L
# Moral Concepts	70		68		60		65		70		6941	
Emotional responses (%)	**72**	61*	**67**	54*	**69**	54*	**67**	59*	**69**	58*	**66**	55*
Emotional intensity	4.24	4.1	3.71	3.77	3.8	3.82	3.78	**4.10***	**3.81**	3.60*	**3.30**	3.17*
Concrete responses (%)	**35**	24*	**24**	12*	**24**	12*	**29**	16*	**40**	33*	**42**	36*
Concreteness score	**3**	2.7*	**2.6**	2.2*	**2.5**	2.3*	**2.7**	2.5*	**3.2**	3*	**3.1**	2.9*

Table 3: Average proportion of emotional responses and intensity (top), and concrete responses and concreteness scores (bottom) in the top 50 negative cues from GMN-H (H) and GMN-L (L)-generated responses. The concepts are associated with moral dimensions identified by both humans and the LLM. The comparison size of moral concepts is the union of H and L from their respective top words. * indicates statistically significant differences (t-test; $p < 0.05$). Significantly higher scores are bolded.

et al., 2024; Zheng et al., 2023; Tjuatja et al., 2024; Dillion et al., 2023). In contrast, GMN-H associations for the same concepts are more varied, often influenced by individual sensory experiences and cultural context (Kostova and Radoynovska, 2008; Son et al., 2014; Shin et al., 2018). For example, the term "prejudice" is frequently associated with culturally specific terms like "race" or "black" resulting in divergent semantic networks compared to GMN-L's statistically driven associations (e.g., "bigotry"). When considering negative physical or emotional concepts like "vomit" and "gross" GMN-H consistently involves synonymous terms, indicative of direct sensory or emotional experiences (Appendix F.1). In contrast, GMN-L still maintains a focus on causal relations. This discrepancy highlights a systematic qualitative difference between representations based on statistical word co-occurrence patterns (Kang and Choi, 2023) and the rich associations observed in humans reflecting their rich physical and emotional experience (Ji et al., 2024). This difference clearly persists in associations, although in the dialogue tasks that LLMs increasingly approach human capabilities.

In positive moral concepts, we observe that responses from both GMN-H and GMN-L to virtue-related words often display synonymy or antonymy, while religion-related terms exhibit various types of meronymy (shown in Appendix F.2). Llama is predominantly trained on training data from Western cultures, where religious concepts have a strong, positive historical presence despite the declining influence of religion in many Western societies (Topkev, 2024; Halman and De Moor, 1994). This cultural frameworks naturally lead to overlap in positive moral concepts between humans and llama.[13]

The quantitative analysis of subgraphs across dimensions reveals several important findings (statistical details are provided in Appendix G). First, the statistics suggest that moral words associated with fairness in GMN-L are more densely connected and exhibit stronger propagation efficiency, leading to significant advantages in spreading moral information (Taxidou and Fischer, 2014; Centola, 2010).[14] Moreover, GMN-L demonstrates notably

[13] In future work it will be interesting to examine this observation in LLMs trained on corpora from secular (e.g., China) or non-Christian societies.

[14] Even though the difference may not be significantly larger than in other dimensions, these advantages could accumulate across multiple metrics.

stronger connections within abstract dimensions such as fairness, loyalty, and authority, as weighted degree centrality are two times higher than GMN-H, while similar magnitude in care and sanctity dimensions. The analysis also shows that the sanctity dimension in GMN-H exhibits higher propagation efficiency, resulting in more effective moral information spread. Finally, both propagation efficiency and density decrease significantly when pruning the graph to retain only the top moral words for both GMN-H and GMN-L, suggesting that morally significant concepts across dimensions are highly interconnected and exhibit stronger propagation efficiency compared to less morally related concepts.

6.2 Human moral associations are more emotional and concrete

We identified systematic qualitative differences in the associations with cues associated with the negative (vice) moral dimensions. Specifically, Llama associations with morally loaded words are more sterile with less emotion and a higher level of abstractness.

Method We analyze **emotionality** in responses to the top 50 morally significant concepts across five moral dimensions. We obtain an emotion score for each response using the *arousal* norms from the VAD-norms (Warriner et al., 2013), a human-labeled emotion lexicon of over 13k English words and quantify the degree of emotions reflected in responses per cue using (a) the proportion of emotional responses among all responses and (b) their average emotional intensity. A response is considered emotional if it is in the emotion lexicon. Emotional response intensity per concept was calculated by multiplying the emotional intensity of responses by their word association strength, then averaging these values for each moral dimension. The **concreteness** of responses was assessed using the Brysbaert et al. (2014) concreteness lexicon. Concepts with a score above 3.5 were considered concrete. The same set of concepts and comparison size from the emotion analysis was used to maintain consistency. We calculated concept-level concreteness analogously to emotional intensity.

Results Table 3 shows that GMN-H contains significantly higher proportions of emotional responses across all dimensions, with notably greater emotional intensity for sanctity. GMN-L exhibits higher emotional intensity in the more abstract dimension of authority.

GMN-H also produces more concrete responses, or, conversely, GMN-L responses are more abstract. We observe (Appendix F.1) that GMN-H often connected cues to real-life or physical experiences (Kostova and Radoynovska, 2008; Son et al., 2014; Shin et al., 2018), while Llama relied on abstract associations from textual data (Ji et al., 2024; Scherrer et al., 2023). Its statistical, text-based nature limits its ability to replicate the sensory-driven responses typical of humans, which dominate the moral word associations, resulting in overall lower concreteness scores (Dillion et al., 2023; Santurkar et al., 2023).

7 Conclusions

We presented a framework for detailed comparison of moral associations in humans and LLMs. We presented a method to elicit word associations from LLMs while ensuring structural similarity to human responses. Broadly speaking, the moral associations emerging from Llama aligned with humans, and the moral perspectives could be uncovered through word associations without direct prompting moral, mirroring human performance. By employing a global network approach, we captured the nuanced relationships between moral concepts. Notably, we observed divergences between humans and LLMs in top negative moral concepts. These differences may stem from how moral concepts are processed. Our results show that humans are sensory and experience driven, leading to associations that are more grounded and emotional. In contrast, LLMs tend to produce more abstract concepts and generally exhibit lower emotional intensity than humans, particularly when comparing concepts related to physical or mental states (e.g., in the sanctity dimension). Conversely, there is considerable alignment in top positive moral concepts between humans and LLMs. This alignment is likely due to cultural frameworks, resulting in humans prioritizing more abstract moral concepts, which are easier for LLMs to capture. Overall, our findings indicate that while LLMs closely mirror human moral associations in many respects, internal differences in how humans and LLMs process information can lead to divergences. Future work can apply our framework across a wider range of models, and link our findings to questions of human and LLM alignment.

8 Limitations

LLM Selection As our main focus is to explore the feasibility of automatically generating reliable large-scale word associations and comparing morality alignment, we selected the recent representative Llama-3.1-8B model given its balance of performance and size in various NLP tasks. We acknowledge that different models might exhibit different behaviors. Our framework can be easily applied to other models. We leave the exploration of more large language models with varying sizes and types as future work.

Concept-Level Alignment Our study focuses on providing a framework to systematically evaluate the moral alignments between concepts in humans and LLMs. This approach is not directly applicable to assess morality alignment in broader contexts, such as sentences or documents, where the overall morality is complex to predict. However, the propagated moral scores for large-scale concepts can serve as basic, word-level scores, supporting future work on contextual moral inference.

References

Marwa Abdulhai, Gregory Serapio-Garcia, Clément Crepy, Daria Valter, John Canny, and Natasha Jaques. 2023. Moral foundations of large language models. *arXiv preprint arXiv:2310.15337*.

Katherine Abramski, Clara Lavorati, Giulio Rossetti, and Massimo Stella. 2024. Llm-generated word association norms. In *HHAI 2024: Hybrid Human AI Systems for the Social Good*, pages 3–12. IOS Press.

Guilherme F. C. F. Almeida, José Luiz Nunes, Neele Engelmann, Alex Wiegmann, and Marcelo de Araújo. 2024. Exploring the psychology of llms' moral and legal reasoning. *Artificial Intelligence*, 333:104145.

Sotiris Anagnostidis and Jannis Bulian. 2024. How susceptible are llms to influence in prompts? *arXiv preprint arXiv:2408.11865*. Computer Science > Computation and Language (cs.CL).

John Baldwin. 2017. Culture, prejudice, racism, and discrimination. *Oxford Research Encyclopedia of Communication*. Date of access 5 Oct. 2024.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, 46(3):904–911.

Damon Centola. 2010. The spread of behavior in an online social network experiment. *Science*, 329:1194–1197.

Richard A. Charter. 1996. Note on the underrepresentation of the split-half reliability formula for unequal standard deviations. *Perceptual and Motor Skills*, 82(2):401–402.

Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, and Jun Xu. 2024. Bias and unfairness in information retrieval systems: New challenges in the llm era. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)*, page 11 pages, New York, NY, USA. ACM.

Simon De Deyne, Álvaro Cabana, Bing Li, Qing Cai, and Meredith McKague. 2020. A cross-linguistic study into the contribution of affective connotation in the lexico-semantic representation of concrete and abstract concepts. In *CogSci*.

Simon De Deyne, Yoed N. Kenett, David Anaki, and Miriam Faust. 2016. Large-scale network representations of semantics in the mental lexicon. In Michael Ramscar, Matt Jones, Melody Dye, and Ernest Klein, editors, *Big Data in Cognitive Science*, 1st edition, page 7. Psychology Press.

Simon De Deyne, Danielle J Navarro, Guillem Collell, and Andrew Perfors. 2021. Visual and affective multimodal models of word meaning in language and mind. *Cognitive Science*, 45(1):e12922.

Simon De Deyne, Danielle J. Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. 2019. The "small world of words" english word association norms for over 12,000 cue words. *Behavior Research Methods*, 51(3):987–1006.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can ai language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600.

Yupei Du, Yuanbin Wu, and Man Lan. 2019. Exploring human gender stereotypes with word association test. 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 6133–6143, Hong Kong, China. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hartry H. Field. 1981. *5. Mental Representation*, pages 78–114. Harvard University Press, Cambridge, MA and London, England.

Jillian Fish and Moin Syed. 2020. Racism, discrimination, and prejudice. In *The Encyclopedia of Child and Adolescent Development*, pages 1–12. John Wiley & Sons, Inc.

Jeremy Frimer, Jonathan Haidt, Jesse Graham, Morteza Dehghani, and Reihane Boghrati. 2017. Moral foundations dictionaries for linguistic analyses, 2.0. Unpublished Manuscript.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In Patricia Devine and Ashby Plant, editors, *Advances in Experimental Social Psychology*, volume 47, pages 55–130. Academic Press.

Jesse Graham, Jonathan Haidt, and Brian A. Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96(5):1029–1046.

Rui Guo, Greg Farnan, Niall McLaughlin, and Barry Devereux. 2024. Qub-cirdan at "discharge me!": Zero shot discharge letter generation by open-source llm. *arXiv preprint arXiv:2406.00041*. BioNLP 2024 workshop.

Loek Halman and Ruud De Moor. 1994. Religion, churches and moral values. In *The individualizing society*, pages 37–65. Brill.

Frederic R. Hopp, Jacob T. Fisher, Devin Cornell, Richard Huskey, and René Weber. 2021. The extended moral foundations dictionary (emfd): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior Research Methods*, 53(1):232–246.

Wei Huang, Xingyu Zheng, Xudong Ma, Haotong Qin, Chengtao Lv, Hong Chen, Jie Luo, Xiaojuan Qi, Xianglong Liu, and Michele Magno. 2024. An empirical study of llama3 quantization: From llms to mllms. *arXiv preprint arXiv:2404.14047*.

Anil Jain, Karthik Nandakumar, and Arun Ross. 2005. Score normalization in multimodal biometric systems. *Pattern Recognition*, 38(12):2270–2285.

Jianchao Ji, Yutong Chen, Mingyu Jin, Wujiang Xu, Wenyue Hua, and Yongfeng Zhang. 2024. Moralbench: Moral evaluation of llms. *arXiv preprint arXiv:2406.04428*.

Cheongwoong Kang and Jaesik Choi. 2023. Impact of co-occurrence on factual knowledge of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7721–7735, Singapore. Association for Computational Linguistics.

S. Kappal. 2019. Data normalization using median median absolute deviation mmad based z-score for robust predictions vs. min–max normalization. *London Journal of Research in Science: Natural and Formal*, 19(4):39–44.

Zdravka Kostova and Blagovesta Radoynovska. 2008. Word association test for studying conceptual structures of teachers and students. *Bulgarian Journal of Science and Education Policy (BJSEP)*, 2(2):209–231.

Will Lowe. 1997. Meaning and the mental lexicon. In *Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1092–1097.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Daye Nam, Andrew Macvean, Vincent Hellendoorn, Bogdan Vasilescu, and Brad Myers. 2024. Using an llm to help with code understanding. In *ICSE '24: Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, New York, NY, USA. Association for Computing Machinery.

Budi Eko Pranoto and Lidia K. Afrilita. 2019. The organization of words in mental lexicon: Evidence from word association test. *TEKNOSASTIK*, 16(1):26–33.

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 428–446, Toronto, Canada. Association for Computational Linguistics.

Aida Ramezani and Yang Xu. 2024. Moral association graph: A cognitive model for moral inference. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. 2023. Evaluating the moral beliefs encoded in llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 51778–51809.

Ji-eun Shin, Eunkook M. Suh, Kimin Eom, and Heejung S. Kim. 2018. What does "happiness" prompt in your mind? culture, word choice, and experienced happiness. *Journal of Happiness Studies*, 19:649–662.

Jung-Soo Son, Vinh Bao Do, Kwang-Ok Kim, Mi Sook Cho, Thongchai Suwonsichon, and Dominique Valentin. 2014. Understanding the effect of culture on food representations using word associations: The case of "rice" and "good rice". *Food Quality and Preference*, 31:38–48.

Io Taxidou and Peter M. Fischer. 2014. Online analysis of information diffusion in twitter. In *Proceedings of the 23rd International Conference on World Wide Web (WWW '14 Companion)*, pages 1313–1318, New York, NY, USA. Association for Computing Machinery.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkar, and Graham Neubig. 2024. Do llms exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026.

Ahmed Topkev. 2024. *Framing Religion*, pages 185–284. Springer Nature Switzerland, Cham.

Bram Van Rensbergen, Gert Storms, and Simon De Deyne. 2015. Examining assortativity in the mental lexicon: Evidence from word associations. *Psychonomic Bulletin & Review*, 22:1717–1724.

David A. Walker. 2006. A comparison of the spearman-brown and flanagan-rulon formulas for split half reliability under various variance parameter conditions. *Archives*, 5(2).

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior Research Methods*, 45(4):1191–1207.

Peiran Yao, Tobias Renwick, and Denilson Barbosa. 2022. WordTies: Measuring word associations in language models via constrained sampling. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5959–5970, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Dengyong Zhou, Olivier Bousquet, Thomas Lal, Jason Weston, and Bernhard Schölkopf. 2003. Learning with local and global consistency. In *Advances in Neural Information Processing Systems 16 (NeurIPS 2003)*. MIT Press.

A Word Association Test Instructions

We used the following prompt to generate WA-L.

System Prompt:

> **Background:** On average, an adult knows about 40,000 words, but what do these words mean to people like you and me? You can help scientists understand how meaning is organized in our mental dictionary by playing the game of word associations. This game is easy: Just give the first three words that come to mind for a given cue word.
>
> **Output Format:** Output your response in the following format:
> response1, response2, response3
> Do not provide any additional context or explanations. Just the words as comma-separated values.

User Prompt: Cue word: `{keyword}`

The fixed system prompt positions the model as a human participant in a psychology experiment, requesting three word associations for a given cue word, formatted as comma-separated values without additional context. The exat same system prompt has been used to collecting human responses for WA-H. The {keyword} will be replaced with actual cue words when generating word associations, and each cue will be prompted 100 times.

B WA-H and WA-L Reliability Test

Figure 5 presents reliability test for WA-L and WA-H using the the Precision@K.

WA-H refers to word associations produced by human participants, as detailed in Section 3.1. The figure compares precision@K for each internal half. Each line shows precision at different K values, with shaded regions representing standard deviation over 50 runs. Reliability values are noted

C Graph Statistics

Table 4 presents the overall graph statistics of WA-H and WA-L. Both graphs were prompted with the same 12,216 cue words.

Compared to WA-H, WA-L has fewer edges, lower density, and lower average connectivity, but exhibits a slightly higher local clustering coefficient and a larger diameter, indicating more localized subgraph connections.

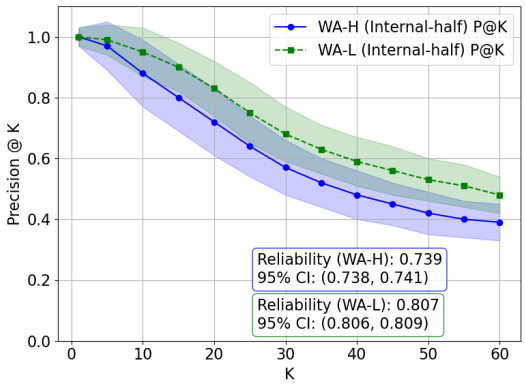

Figure 5: Precision@K for WA-H and WA-L associations.

	WA-H	WA-L
Nodes	12,216	12,216
Edges Number	963,043	502,174
Density	0.013	0.007
Local Cluster	0.12	0.15
Max Connectivity	221	208
Min Connectivity	48	10
AVG Connectivity	114	77
SD Connectivity	21	23
Diameter	3	4

Table 4: A statistical overview of the global word association graphs in WA-H and WA-L.

D Optimizing Alpha

Figure 6 shows how the Spearman correlation varies with different α values for both GMN-H and GMN-L.

The GMN-L correlation reaches its peak at alpha = 0.75, while the GMN-H correlation peaks at alpha = 0.9. We used these respective optimal values in Section 5 to propagate the moral values.

E Ranking Values

We present the top-ranked positive and negative words that we used, as well as words with different polarity in the Section 6 supplemented with their overall morality score and dimensions.

The morality score is calculated as the sum of scores across five dimensions after propagation. Due to differences in word association responses between LLMs and humans, the values produced may not be directly comparable. To address this, we applied median absolute deviation (MAD) normalization post-aggregation to the sum scores for

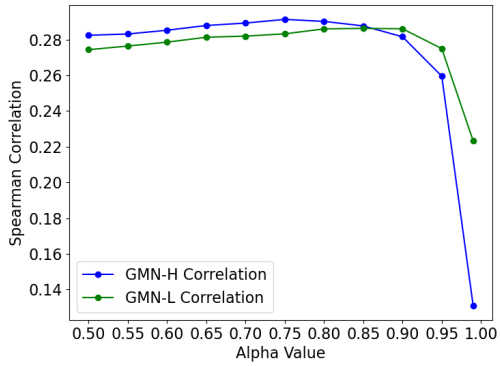

Figure 6: The Spearman correlation between the eMFD and the propagated values for various values of α.

both LLMs and humans. This helps ensure consistency in comparisons across potentially skewed distributions and mitigating outliers, while still preserving the internal structure of the data.(Jain et al., 2005; Kappal, 2019).

The moral dimension of a concept is the one with the highest score among the five dimensions. Denoting the dominant dimensions as 1: Care, 2: Fairness, 3: Loyalty, 4: Authority, 5: Sanctity.

E.1 Top Negative

GMN-H: disgusting(5): -28, traitor(3): -27, vomit(5): -27, hurt(1): -26, dirty(5): -26, pain(1): -25, bad(5): -25, thief(2): -24, gross(5): -24, sick(5): -24

GMN-L: betrayal(2): -43, prejudice(2): -38, cheating(2): -37, disgusting(2): -36, discrimination(2): -33, dishonest(2): -32, deception(2): -31, dishonesty(2): -30, racism(2): -30, infidelity(3): -28

E.2 Top Positive

GMN-H: church(5): 62.03, religion(5): 52.71, God(5): 47.78, priest(5): 37.43, holy(5): 34.74, religious(5): 34.04, catholic(5): 33.01, kind: 29.72(1), caring(1): 26.04, worship(4)(5): 25.72

GMN-L: church(5): 52, kindness(1): 41, religion(5): 40, priest(5): 36, prayer(5): 34, bible(5): 34, faith(5): 34, family(3): 33, compassion(1): 32, holy(5): 30

E.3 Difference

Table 5 presents the concepts that we used in Table 2 (column Different), along with their dominant moral dimensions (using GMN-H as the standard) and propagated moral scores from GMN-H and GMN-L.

Word (Dimension)	GMN-H	GMN-L
Abortion (1,4)	-0.45	1.5
Immigrant (4)	-0.62	1.1
Politician (2)	-6.6	6.5
Capitalist (3,4)	-0.16	0.97
Homosexual (4,5)	-0.55	1.03
Commercial (2,4,5)	-0.42	0.52
Jail (4)	0.06	-3.15
Air (4)	1.09	-0.73
Plastic (4,5)	0.19	-1.25
Soviet (3)	2.27	-0.44
Bees (3,4)	2.27	-0.82
Snob (4)	2.27	-0.32

Table 5: Comparison of concepts with divergent moral values from GMN-H and GMN-L.

F Response Analysis

For cue words in the Table 2, we provide the detailed associations to understand how their moral values are being captured by GMN-H and GMN-L. We examine (a) the top frequent responses for each cue word and in both GMN-H and GMN-L; and (b) "top unique response": a response that appears in one graph (GMN-L or GMN-H) but does not appear in the other.

F.1 Negative Response Analysis

Table 6 presents the associations for the top negative moral concepts in Table 2.

F.2 Positive Response Analysis

Table 7 presents the associations for top positive cues in Table 2.

Cue Word	Top Response		Top Unique Response	
	GMN-H	GMN-L	GMN-H	GMN-L
prejudice	pride	bias	pride	stereotypes
	racism	racism	black	biases
	black	discrimination	race	stereotyping
	race	bigotry	racist	bigoted
racism	black	prejudice	black	inequality
	white	discrimination	white	segregation
	bad	bigotry	bad	equality
	prejudice	inequality	bigot	pain
discrimination	racism	prejudice	race	stereotypes
	race	racism	racist	stereotyping
	prejudice	bias	sexism	equality
	unfair	inequality	gender	prejudices
vomit	puke	nausea	gross	stomachache
	sick	sickness	spew	queasy
	gross	stomach	smell	hangover
	barf	stomachache	green	poisoning
gross	disgusting	disgusting	fat	nauseating
	nasty	vomit	net	disgusted
	ugly	nauseating	large	queasy
	fat	revolting	yuck	nausea

Table 6: Comparison of the top 4 responses and top 4 unique responses between GMN-H and GMN-L for selected cue words in top negative and divergent concepts, ranked based on frequency.

Cue Word	Top Response		Top Unique Response	
	GMN-H	GMN-L	GMN-H	GMN-L
kind	nice	gentle	type	nurturing
	type	caring	sort	soft
	gentle	friendly	happy	charitable
	sweet	compassionate	person	warmth
caring	love	nurturing	sharing	supportive
	loving	loving	nice	motherly
	kind	kind	giving	selfless
	sharing	compassionate	sweet	emotional
church	steeple	altar	catholic	altar
	religion	priest	synagogue	minister
	God	sunday	stone	baptism
	priest	pews	school	service
priest	church	church	father	altar
	catholic	clergy	black	clergyman
	father	altar	vicar	chapel
	religion	minister	pedophile	vatican
religion	God	church	cross	beliefs
	church	faith	war	rituals
	faith	God	atheism	scripture
	Christianity	spirituality	fear	churches

Table 7: Comparison of the top 4 responses and top 4 unique responses between GMN-H and GMN-L for selected cue words in top positive concepts, ranked based on frequency.

G Quantitative analysis of graph property

Figure 7 presents detailed the graph analysis we used in Section 6.

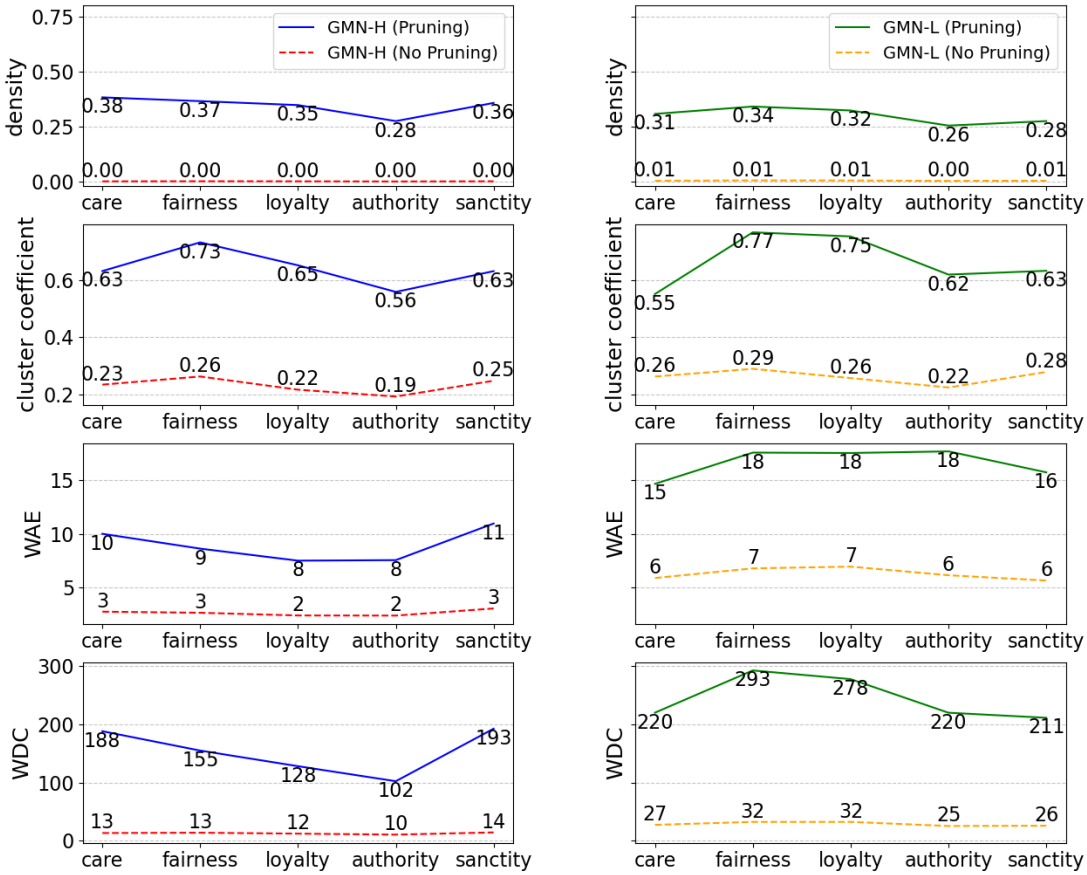

Figure 7: Quantitative analysis of graph properties—density, local clustering coefficient (clustering coefficient), weighted average edge (WAE), and weighted degree centrality (WDC)—was conducted across moral dimensions for both GMN-H and GMN-L. Results are presented for pruned and non-pruned subgraphs, highlighting the effects of pruning on propagation efficiency and network density. In pruned subgraphs, we keep only the top 50 negative cues based on each dimension in the graph. In non-pruned subgraphs, the subgraph contains not only the top 50 negative cues but also each cue's corresponding responses. WAE represents the average edge connection weight between any two connected nodes in a graph, with higher WAE indicating a greater potential for moral information transfer during propagation.