

Do Morals Guide How LLMs Think? The Role of Ethical Perspectives in General Problem Solving

Iseo Kim, Eunjin Hong, Juae Kim*

Hankuk University of Foreign Studies, Republic of Korea

{iseo6702, ej.hong, juaekim}@hufs.ac.kr

Abstract

This study investigates how different moral conditions influence the general problem-solving capabilities of Large Language Models (LLMs). We examine whether the role of morality in human decision-making can also serve as a useful lens for analyzing variation in LLM behavior. Specifically, we define distinct moral conditions based on Kohlberg’s theory of moral development and design prompts intended to elicit model outputs aligned with each condition. The validity of this alignment is assessed using the Defining Issues Test, a human evaluation tool. We then evaluate models under each condition on the MMLU benchmark, which measures general problem-solving ability across diverse domains. Experimental results show that different moral perspectives correspond to differences in model behavior during general reasoning, as reflected in both responses and internal representations. In particular, more advanced moral conditions tend to elicit more reflective reasoning patterns, which are often linked to improved performance. Our study broadens the scope of LLM morality, which has traditionally been examined mainly in ethical judgment settings. More broadly, it suggests that morality may function not only as a mechanism for safety alignment but also as a factor that shapes model behavior during reasoning. Code and data are available at <https://github.com/ISEOKIM/llm-morality>

1 Introduction

The rapid advancement of large language models (LLMs) has led to remarkable achievements, yet they continue to face persistent concerns regarding ethics and reliability (Huang et al., 2025). In response, recent studies have aimed to guide model outputs to align with specific moral values (Zhou et al., 2024) and to evaluate how such models respond in situations that require moral judgment

* Corresponding author

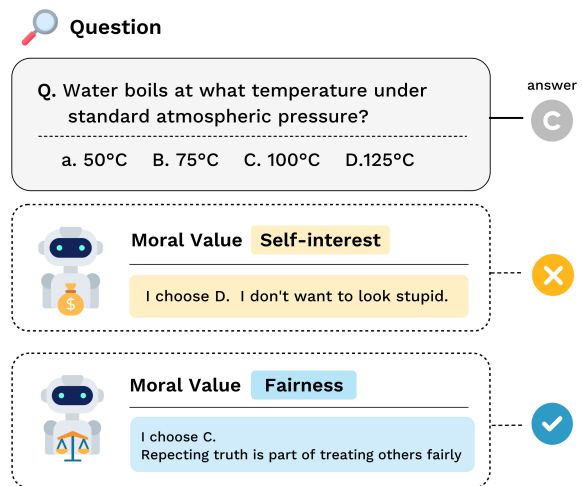


Figure 1: This study expands the scope of ethical considerations in LLMs by exploring the relationship between moral values and reasoning behavior.

(Agarwal et al., 2024), as part of ongoing efforts to develop more morally responsible models.

While such efforts tend to treat morality as a basis for ethical decision-making, in humans, morality serves a broader and more complex role. It provides a framework for justifying judgments, facilitates the understanding of others’ perspectives, and ultimately shapes how individuals interpret information and approach problems (Cushman, 2008; Tiberius, 2010). Considering these characteristics, we suggest that the influence of morality may extend beyond ethical judgment and may also affect broader problem-solving behavior in LLMs. Nevertheless, the impact of morality on non-ethical tasks in LLMs remains largely unexplored. This study therefore aims to fill this gap by empirically exploring the question: **Can morality influence the general problem-solving behavior of LLMs?**

To answer this question, we draw on Kohlberg’s theory of moral development as our theoretical foundation (Kohlberg, 1971) to examine the in-

fluence of morality under a range of distinct moral conditions. Kohlberg’s theory describes three levels of moral development: pre-conventional, conventional, and post-conventional. Based on the core moral values emphasized in each level, we design prompts to elicit model outputs aligned with the corresponding pattern of moral reasoning. The proposed prompts are subjected to a manipulation check using the Defining Issues Test (DIT) (Rest and of Minnesota. Center for the Study of Ethical Development, 1990), a human evaluation tool grounded in the same theoretical framework. This validation confirms that each prompt effectively evokes the intended moral reasoning pattern, establishing a foundation for the subsequent analysis.

Subsequently, we analyze reasoning patterns and overall performance on general problem-solving tasks under each moral condition. To this end, we use MMLU (Hendrycks et al., 2021b), a value-neutral benchmark covering diverse subjects such as mathematics and science. The experimental results show that moral conditioning is associated with differences in reasoning behavior, with higher-order moral values linked to more deliberative reasoning and improved MMLU performance, even in non-ethical domains. In particular, the LLaMA3-8B-Instruct model exhibits accuracy improvements of +0.85%, +4.96%, and +5.88% under the pre-conventional, conventional, and post-conventional conditions, respectively, compared to the baseline. Furthermore, analysis of the model’s internal representations shows that moral conditioning is reflected in observable changes in how the model represents information during problem-solving. These findings suggest that moral values may function as a factor shaping model behavior beyond their conventionally understood role in ethical judgment.

The main contributions are as follows:

1. We empirically demonstrate that moral values influence not only ethical judgments but also general reasoning behavior, affecting how LLMs interpret problems and generate justifications.
2. To enable targeted observation of moral conditioning, we construct condition-specific prompts grounded in Kohlberg’s theory, allowing systematic comparison of model behavior across moral perspectives.
3. We show that moral conditioning affects not only the quantitative performance of LLMs

but also their reasoning patterns and internal representations; higher-order moral values tend to elicit more reflective reasoning.

2 Background and Related Work

This section introduces the theoretical foundations of our study, including moral development theory, the evaluation tool, and prior research on morality in LLMs. Through this, we aim to establish a basis for understanding and interpreting how human moral values may influence the behavior and reasoning of LLMs.

2.1 Kohlberg’s Theory of Moral Development

This study adopts Kohlberg’s theory of moral development (Kohlberg, 1971) as a framework for constructing moral conditions for LLMs. As one of the most influential theories in moral psychology (Tanmay et al., 2023), it categorizes moral reasoning into three stages based on the core values individuals prioritize: *pre-conventional*, *conventional*, and *post-conventional*. The *pre-conventional* stage centers on self-interest and external punishment or rewards. Moving beyond this, the *conventional* stage emphasizes social norms and duties. Finally, the *post-conventional* stage upholds universal ethical principles such as justice and human rights.

While other moral frameworks, such as Moral Foundations Theory (Graham et al., 2013), offer alternative perspectives on moral reasoning, we adopt Kohlberg’s model because its stage-based structure enables interpretable and systematic variation in moral reasoning. Building on this structure, we construct condition-specific prompts and analyze how moral conditioning influences reasoning behavior in LLMs.

2.2 Defining Issues Test

The Defining Issues Test (DIT) is a tool based on Kohlberg’s theory for assessing human moral development (Rest and of Minnesota. Center for the Study of Ethical Development, 1990). Respondents are presented with moral dilemmas and asked to prioritize judgment criteria for resolving them. Their responses are converted into a P-score indicating moral maturity and can be used to classify the underlying pattern of moral reasoning according to the corresponding moral condition introduced in Section 2.1. More detailed information about the DIT is provided in Appendix A.

Recently, many studies have explored applying human psychology assessment tools like DIT to

evaluate LLMs. Notably, (Khandelwal et al., 2024) used the DIT to study multilingual variation in LLM moral reasoning, and (Jiang et al., 2024) used measures such as the Big Five Inventory to assess personality consistency in LLMs. These efforts suggest that human-centered evaluation frameworks can be effectively extended to assess model behavior.

2.3 Approaches to Morality in LLMs

Prior research on morality in LLMs has primarily explored three lines of inquiry. First, various moral control methods (Schick et al., 2021; Liu et al., 2025) have been proposed to suppress harmful or biased outputs. These approaches focus on preventing models from generating aggressive, discriminatory responses (Perez et al., 2022; Furniturewala et al., 2024; Dai et al., 2024). Second, other works explore whether LLMs can make moral judgments according to specific ethical theories (Simmons, 2023). This includes assessing model outputs based on philosophical theories such as utilitarianism or deontology (Zhou et al., 2024), or exploring their capacity to reflect higher-order values (Scherrer et al., 2023; Hobson et al., 2024). Third, benchmarks have been developed to quantitatively assess the ethical alignment of LLMs, using curated datasets to measure and compare moral reasoning patterns in model outputs (Lourie et al., 2021; Hendrycks et al., 2021a; Marraffini et al., 2024).

These approaches tend to treat morality as a mechanism for behavior control or as a narrow capability confined to ethical contexts. Departing from this view, our study examines whether moral conditioning affects model behavior and performance even in value-neutral reasoning tasks.

3 Steering Morality with Prompts

To steer LLMs toward value-specific moral reasoning, we first describe how the corresponding prompts were constructed and validated.

3.1 Defining Moral Conditions

Previous studies (Simmons, 2023; Zhou et al., 2024) have shown that language-level instructions can effectively guide models toward specific moral orientations. Building on these insights, we steer the model with distinct moral perspectives to examine how they affect model reasoning. Our approach is grounded in Kohlberg’s moral development theory, which defines three moral conditions: pre-conventional, conventional, and post-conventional

Morality Prompt Template

Adopt the identity of a character operating at a specified moral level: {Moral level}.

[Characteristics]:

- {characteristic 1}
- {characteristic 2}
- {characteristic 3}

When responding, ensure that your answers reflect the principles, motivations, and behaviors associated with this stage.

Table 1: Prompt structure used to steer LLMs under specific moral conditions. The gray represents the moral role assignment, yellow shows the moral characteristics, and blue marks the moral behavior instruction. The full set of prompts is provided in Appendix B.

(see Section 2.1). This framework enables fine-grained distinctions in moral values across conditions, providing a structured basis for conditioning models with distinct forms of moral reasoning.

We implement this theoretical foundation by constructing prompts that reflect the core moral values associated with each moral condition. To enable clearer attribution of moral influence, we additionally introduce a morality-restricted baseline that explicitly excludes any moral orientation. Recent studies suggest that LLMs exhibit inherent moral tendencies even when processing ethically relevant queries without moral prompts (Tanmay et al., 2023; Khandelwal et al., 2024). This observation points to the importance of a controlled baseline for more precise analysis of how different moral conditions influence model reasoning.

3.2 Designing Morality Prompts

Table 1 illustrates the prompt structure for moral conditioning, consisting of three components: Moral Role Assignment, Moral Characteristics, and Moral Behavior Instruction. Each part is color-coded for clarity and is explained in this section.

Moral Role Assignment. The first instruction specifies the moral condition assigned to the model by selecting one of four conditions: the three moral conditions or a morality-restricted baseline. This component establishes the intended moral context at the start of each task and frames the model’s subsequent reasoning accordingly.

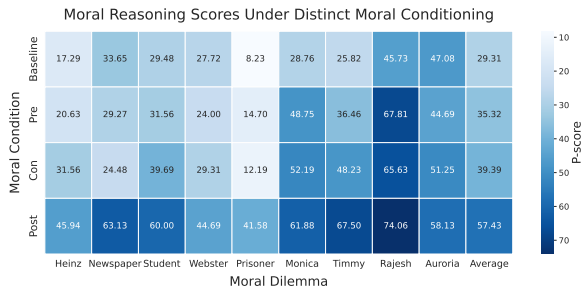


Figure 2: Average P-scores of four models across nine moral dilemmas under four prompting conditions. The P-score ranges from 0 to 100 and measures the degree of morality in responses. Higher conditions consistently yield higher P-scores, showing that the prompts effectively steer models toward the intended level of moral reasoning.

Moral Characteristics. The second part of the prompt presents a set of three representative characteristics that reflect the core moral values of the assigned condition. These characteristics are generated by an LLM (e.g., GPT-4o), guided by definitions grounded in Kohlberg’s theory.

The baseline condition emphasizes neutrality and objectivity. The pre-conventional condition focuses on self-interest and avoiding punishment, the conventional condition on conformity and social approval, and the post-conventional condition on justice, individual rights, and universal principles. Together, these characteristics make the intended moral condition more explicit.

Moral Behavior Instruction. The final part of the prompt instructs the model to reason in a manner consistent with the assigned moral condition, thereby reinforcing the intended moral conditioning.

For example, under the post-conventional condition, the model is guided to reason in accordance with universal ethical principles. When asked, “Should I report a friend who broke a minor school rule but helped someone in need?”, it responds: “If the rule was broken to uphold a greater moral good, like helping someone in need, then the action may be justified despite the violation.” This response reflects post-conventional reasoning because it appeals to abstract ethical principles rather than strict rule-following. This example illustrates how our prompting method steers model outputs toward different patterns of moral reasoning.

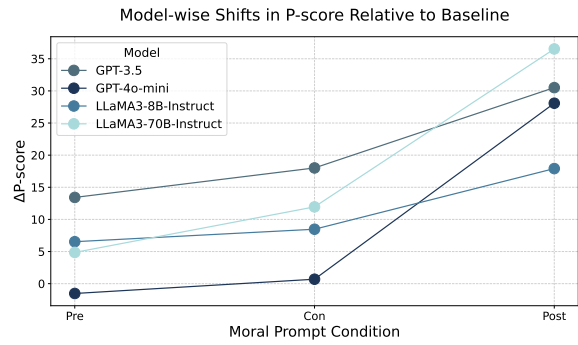


Figure 3: Changes in P-score across three moral prompting conditions, shown relative to the baseline. The upward progression indicates that higher moral conditions elicit responses more closely aligned with advanced moral reasoning.

3.3 Experiment for Verifying Moral Alignment via DIT

This section presents a manipulation check to assess whether the condition-specific prompts introduced in Section 3.2 effectively guide the models toward the intended pattern of moral reasoning.

3.3.1 Experimental Setup

We utilize GPT-3.5-turbo (OpenAI, 2023), GPT-4o-mini (OpenAI et al., 2024), LLaMA3-8b-Instruct, and LLaMA3-70b-Instruct (Grattafiori et al., 2024) for the manipulation check. The selected models vary in scale and accessibility.

Our evaluation is based on a DIT-style dilemma set comprising nine items: five original DIT dilemmas and four additional dilemmas adopted from (Tanmay et al., 2023), reflecting broader socio-cultural contexts while maintaining the original DIT format. For each item, LLMs prompted with condition-specific moral instructions are asked to rank the statements they consider most important for moral decision-making.

Model responses are quantitatively evaluated using the P-score, which reflects the extent to which post-conventional moral reasoning is prioritized in the response. Higher scores indicate a greater degree of moral maturity, as reflected in stronger post-conventional moral reasoning. We calculate the P-score following the LLM-specific DIT evaluation guidelines proposed by (Tanmay et al., 2023). Full experimental details and scoring procedures are in Appendix C.

3.3.2 Results and Analysis

Figure 2 illustrates how P-scores vary across prompting conditions and dilemmas. While the

Impact of Moral Conditioning on General Reasoning Performance (MMLU)

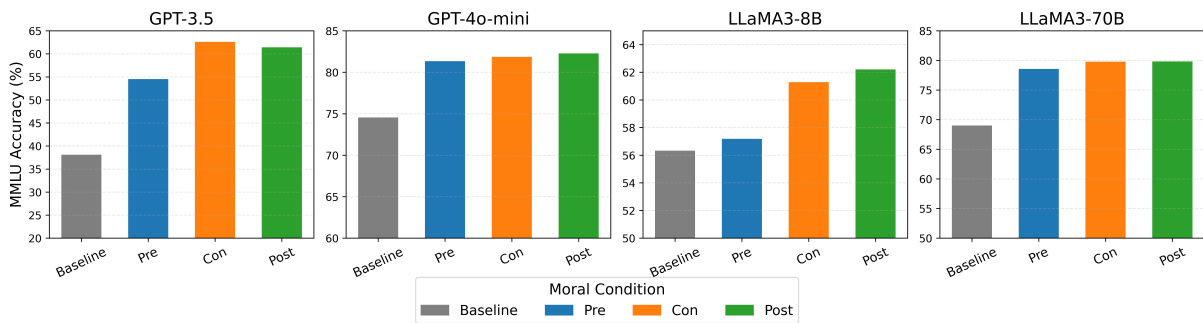


Figure 4: Effects of moral conditioning on general reasoning performance, measured by MMLU accuracy, across four different language models. Each subplot presents the average MMLU score for a given model under four prompt conditions. The figure allows comparison of how performance changes across moral conditions within each model. Within each model, accuracy increases as the prompts reflect higher-order moral values, and this overall pattern is consistent across all four models. This suggests that moral conditioning can positively influence general reasoning abilities beyond ethical domains.

exact scores vary slightly by dilemma, the overall trend is clear: prompts corresponding to more advanced moral conditions consistently yield higher P-scores. Because higher P-scores indicate responses more closely aligned with advanced moral reasoning, this pattern confirms that the models produce outputs consistent with the intended moral conditions.

Figure 3 further shows how each model’s P-score changes across the three moral prompting conditions relative to its baseline. For instance, GPT-3.5-Turbo increases by +30.51 points from the baseline to the post-conventional condition, and LLaMA3-70B-Instruct increases by +36.53 points over the same range. These results demonstrate the effectiveness of condition-specific prompts in steering models toward the targeted patterns of moral reasoning. This overall pattern is also statistically supported. Repeated-measures ANOVA shows a significant main effect of prompting condition, and linear trend analysis confirms a positive progression across conditions. Detailed results and statistical analyses are provided in Appendix C.5.

4 Experimental Evaluation of Moral Conditioning Effects on General Reasoning

In the previous section, we verified that LLMs can be reliably steered toward distinct patterns of moral reasoning through prompt-based manipulation. Building on this, we analyze how moral conditioning affects general reasoning performance in LLMs across diverse tasks, focusing on both task

performance and reasoning patterns.

4.1 Experimental Setups

We evaluate the effects of moral conditioning using the MMLU benchmark, a widely used benchmark for assessing general reasoning performance across diverse domains. The MMLU consists of multiple-choice questions covering 57 subjects, including areas in STEM, humanities, social sciences, and other fields.

Our experiments use the same four models as those used in Section 3.3. Each model is evaluated on the same set of MMLU questions under four prompt conditions. Each input consists of a moral instruction followed by a multiple-choice question, and performance is measured using exact-match (EM) accuracy.

4.2 Quantitative Analysis: Shifts in MMLU Accuracy

Figure 4 shows measurable changes in MMLU accuracy across moral conditions, with performance generally increasing as prompting conditions reflect higher-order moral values. In particular, the conventional and post-conventional conditions generally outperform the pre-conventional condition. This suggests that more advanced moral conditions tend to support stronger performance in general problem-solving.

Specifically, GPT-3.5-Turbo achieves 54.52% accuracy under the pre-conventional condition, which emphasizes avoidance of punishment and personal gain. This improves to 61.41% under

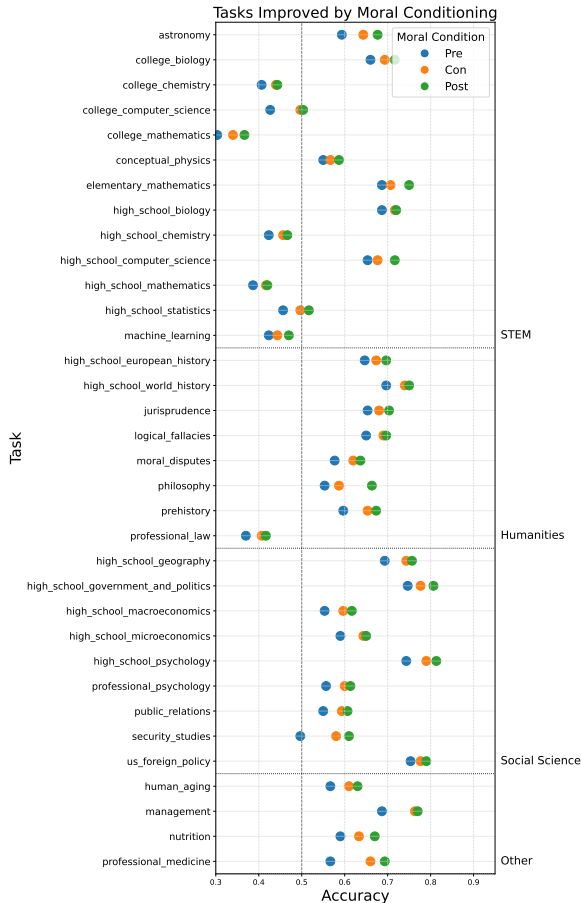


Figure 5: Subset of MMLU tasks illustrating condition-wise accuracy trends in LLaMA3-8B-Instruct under moral conditioning. These examples complement Figure 4 by showing task-level cases that more clearly reflect the aggregate performance trend across conditions.

the post-conventional condition, which emphasizes universal rights and justice, representing a 6.89 percentage-point increase. LLaMA3-8B-Instruct shows a similar pattern, with accuracy rising from 57.18% (pre-conventional) to 61.29% (conventional) and reaching 62.21% (post-conventional). High-performing models such as GPT-4o-mini and LLaMA3-70B-Instruct also exhibit consistent trends. GPT-4o-mini achieves 81.34% under the pre-conventional condition, with slight improvements to 81.86% and 82.27% under the conventional and post-conventional conditions, respectively. Similarly, LLaMA3-70B-Instruct records 78.57%, 79.79%, and 79.83% across the three conditions in ascending order. This overall pattern is also statistically supported across models, with detailed results provided in Appendix D.1.

To identify subject-level contributions to the overall improvement, Figure 5 highlights tasks where LLaMA3-8B-Instruct shows accuracy

Response Patterns by Moral Stances

Pre-conventional

"I don't want to get in trouble, so I'll choose this."

"I'm picking what seems safe or easy."

"Maybe this helps me personally."

Conventional

"This seems like the socially expected answer."

"I believe people would agree with this."

"This maintains order or norms."

Post-conventional

"According to broader principles like fairness or justice, this is right."

"I approach this question with a critical and analytical mindset."

"I choose this based on universal values."

Table 2: Representative expressions illustrating how models justify their decisions under different moral conditions.

gains across moral conditions. In the **STEM** domain, tasks such as astronomy and elementary mathematics show steady gains under increasingly advanced moral conditions. In the **Humanities** domain, tasks involving moral judgment and value interpretation—such as moral disputes and philosophy—show more pronounced gains as the prompting condition shifts from pre-conventional to post-conventional. Similar trends appear in the **Social Sciences** and **Other** domains, where applied tasks such as geography, security studies, and nutrition also exhibit consistent improvements. Related task-level trends for the other models, as well as baseline comparisons for the tasks shown in Figure 5, are provided in Appendices D.2 and D.3, respectively.

Additional benchmark results beyond MMLU also support our main findings. Detailed results are provided in Appendix D.4.

Overall, across MMLU tasks, performance tends to improve as prompting conditions reflect higher-order moral values. Importantly, these gains are not limited to tasks involving moral dilemmas or ethical judgment, but also appear in objective knowledge domains such as mathematics, computer science, and astronomy. These results suggest that moral concepts may extend beyond their traditionally discussed role in safety or ethical decision-making and may also contribute to more consistent reasoning across diverse tasks.

Question: During a run one morning, a jogger passed a convenience store. He stopped and went into the store to use the bathroom available inside the store. Which of the following would best describe the jogger’s legal status in his utilization of the bathroom?

Options: A. Licensee. B. Gratuitous licensee. **C. Public invitee.** D. Business visitor.

Level	Response
Pre-conventional stage	I think I’m just trying to use the bathroom, and I’m not doing any harm. If I get in trouble, I might get punished, so I should try to avoid that. I just want to use the bathroom and get back to my run. Therefore, the answer is A. Licensee.
Conventional stage	In this case, I would reason that the jogger is using the bathroom as a public service. Therefore, the jogger’s legal status would be that of a public invitee, as they are using a public facility that is open to the public, but not as a customer. Therefore, the answer is C. Public invitee.
Post-conventional stage	In this context, I believe that the jogger’s use of the bathroom is a form of public accommodation, similar to a public park or a public library. The store is providing a public service, and the jogger is taking advantage of that service. Therefore, I would argue that the jogger is a public invitee, as they are being invited to use the facility by the store’s provision of a public service. Therefore, the answer is C. Public invitee.

Table 3: Qualitative analysis of LLaMA3-8B-Instruct’s reasoning patterns across moral development levels for an MMLU Professional Law question on a jogger’s legal status when using a store bathroom. **Bolded blue text highlights the core elements of each moral reasoning process,** such as motivations, justifications, and final decisions. **Correct answer in green.** Further examples illustrating similar reasoning shifts across different tasks are provided in Appendix E.

4.3 Qualitative Analysis: Shifts in Reasoning Patterns

Following the observed performance differences, this section examines whether responses generated under different moral conditions exhibit distinct patterns of justification. To this end, we analyze model outputs under each condition, focusing on the motivations and reasoning used to support the answer.

Table 2 summarizes representative expressions for each moral condition. Even when the same answer is selected, the justification varies across conditions—for example, appealing to social approval in one case and to broader ethical principles in another. These patterns suggest that moral conditions influence how the model frames and justifies its responses.

To illustrate this more concretely, we analyze a representative item from the MMLU professional law subject. Table 3 presents model responses to a question about the legal status of a jogger using a convenience store restroom. This item requires factual interpretation and application of legal concepts. Across conditions, the justification changes in both focus and structure.

For instance, under the pre-conventional condition, the justification centers on short-term avoidance motives such as “not getting into trouble,” with limited consideration of the jogger’s intent or the public nature of the facility. Even so, this condition can still yield better performance than the

baseline. The baseline sometimes relies on superficial pattern matching, whereas the pre-conventional condition reflects a moral perspective centered on personal safety and self-interest, leading to more direct engagement with the task. In contrast, under the conventional condition, the response interprets the jogger’s action as socially appropriate, grounding the answer in shared norms rather than legal criteria. This reflects reliance on consensus-driven judgment rather than systematic legal reasoning. Finally, under the post-conventional condition, the response interprets the behavior in terms of public service access. By relating the scenario to broader public amenities such as parks or libraries, it provides a more structured justification for why the jogger qualifies as a public invitee.

These findings highlight differences in the focus, depth, and logical organization of justification across moral conditions. Higher moral conditions tend to be associated with more structured and principle-based justifications. Such differences may help identify core issues, connect relevant facts, and support more coherent conclusions. We therefore conjecture that morally conditioned reasoning may align with some of the demands of complex problem-solving, potentially contributing to the performance gains observed across diverse tasks in Section 4.2, as well as under varying experimental settings in Appendices F and G.

5 Analysis of Internal Representations

The previous sections focused on how moral conditioning affects model outputs, both in terms of task accuracy and the qualitative characteristics of reasoning patterns. These findings suggest that prompts reflecting different moral values can influence not only what the model outputs, but also how those outputs are justified. However, to determine whether these differences are limited to the output level, it is necessary to examine whether they are accompanied by corresponding changes in the model’s internal representations. To this end, we scrutinize the model’s hidden representations under different moral conditions.

To conduct this analysis, we adopt a probing approach inspired by (Liu et al., 2024), which enables us to examine how internal representations vary across moral conditions. For each MMLU question, we apply three morality-conditioned prompts and compute the cosine similarity between the resulting hidden states at each transformer layer and those generated under the baseline condition. Detailed descriptions of the experimental setup and measurement procedure are provided in Appendix H.1.

Figure 6 presents the layer-wise cosine similarity between the hidden states produced under each moral condition and those generated by the baseline prompt. This metric quantifies how strongly each moral condition is associated with changes in the model’s internal representations when processing the same MMLU question. Higher similarity indicates that a layer’s representation remains relatively close to the baseline, suggesting minimal internal change. Lower similarity, in contrast, indicates greater divergence from the baseline representation at that layer.

The early layers exhibit consistently high similarity across all moral conditions. This suggests that moral conditioning has relatively limited influence on lower-level linguistic processing. The initial layers of transformer models are typically associated with lexical semantics and syntactic structure (Sajjad et al., 2022). In contrast, similarity drops sharply in the middle layers. These layers are often associated with semantic reasoning, abstraction, and decision-related processing (Wu et al., 2025). This pattern suggests that moral conditioning produces larger representation-level differences in layers involved in higher-level processing, such as meaning construction and inference. Notably, the post-conventional condition shows the great-

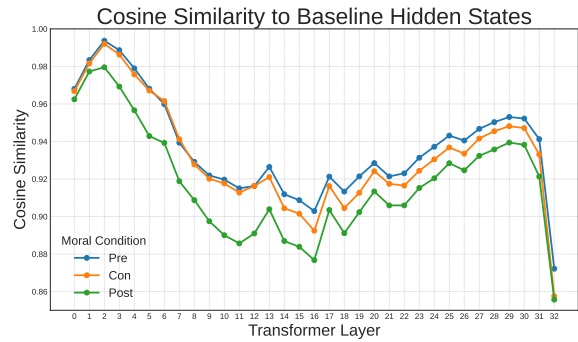


Figure 6: Layer-wise cosine similarity between each moral condition and the baseline. Lower similarity indicates greater divergence in internal representations. The post-conventional condition shows the largest divergence, particularly in the middle layers.

est divergence from the baseline, aligning with the highest performance gains in LLaMA3-8B-Instruct (Section 4.2). In the later layers, similarity between the moral conditions and the baseline increases again. This upward trend likely reflects convergence toward shared output structures (e.g., “Therefore, the answer is...”) as the model moves from intermediate processing to response generation.

To test whether these shifts reflect only stylistic prompting, we conducted control experiments using structurally similar non-moral role prompts. The moral prompts consistently showed lower cosine similarity to the baseline than the role prompts. This suggests that the observed shifts cannot be fully explained by style alone. Detailed results are provided in Appendix H.2.

Overall, these findings suggest that higher-order moral conditions are associated not only with differences in the model’s external outputs, but also with measurable changes in its internal representations across layers.

6 Conclusion

This study has empirically demonstrated that moral values can influence not only ethical judgments in LLMs but also their general reasoning performance. Our goal was to examine whether morality can also function as a factor related to variation in general reasoning behavior in LLMs, rather than to argue that moral conditioning is required for improved benchmark performance. To explore this relationship, we designed prompts grounded in Kohlberg’s theory of moral development to steer models toward specific moral perspectives. Using the Defini-

ing Issues Test and MMLU benchmark, we showed that moral conditioning is associated with differences in both performance accuracy and reasoning patterns. Furthermore, our layer-wise hidden state analysis showed that these differences are accompanied by measurable shifts in internal representations across layers. These findings suggest that morality, beyond serving as a safety-oriented constraint to prevent harmful or unethical outputs, may also be relevant to understanding variation in reasoning behavior in LLMs.

Limitations

This study has empirically shown that moral values can influence general reasoning performance in large language models (LLMs). However, several limitations remain.

Theoretical Scope. This study adopts Kohlberg’s theory of moral development as the primary framework for distinguishing different types of moral reasoning. This theory provides a structured foundation for examining how distinct moral values relate to variation in problem-solving behavior. However, morality is inherently complex and multifaceted, and no single framework can fully capture its scope. While our approach is grounded in Kohlberg’s model, it is not limited to this framework. The prompt-based moral conditioning method is flexible and can be extended to incorporate alternative moral theories, such as Moral Foundations Theory, allowing future research to explore a broader range of ethical perspectives.

Generalizability. Our experiments were conducted in English using four transformer-based language models (e.g., GPT, LLaMA). As a result, the generalizability of our findings across languages, cultures, and model architectures remains uncertain. Future studies may examine how moral conditioning operates in multilingual and multicultural settings and whether social variables, such as gender or age, affect how moral values are reflected in model outputs.

Model Scale and Prompt Sensitivity. Although the same overall pattern appears across all four models, the gains are smaller in higher-performing models such as GPT-4o-mini and LLaMA3-70B-Instruct. This may reflect their stronger baselines, which leave less room for prompt-based improvement. It is also possible that larger models respond differently to the same prompting strategy and may benefit from alternative moral frameworks or more

expressive prompts.

Interpretation and Causal Attribution. Our results show that moral conditioning is associated with performance differences on value-neutral tasks such as MMLU. However, why morality relates to higher-order reasoning remains an open question even in human cognition. Because the causal mechanisms underlying this relationship are not yet fully established, our findings should not be taken as conclusive evidence of how moral conditioning produces the observed differences.

Importantly, this work does not aim to prescribe which moral values LLMs should adopt, nor does it endorse any particular normative stance. While the ethical alignment of LLMs remains an important topic, it lies beyond the scope of this study. Instead, our goal is to examine whether moral perspectives may be related to variation in the reasoning behavior of language models. At the same time, we recognize that guiding LLMs with moral perspectives may carry risks of reinforcing biases, and we emphasize the need for further ethical scrutiny in this direction.

Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) RS-2026-25471023.

References

- Utkarsh Agarwal, Kumar Tanmay, Aditi Khandelwal, and Monojit Choudhury. 2024. [Ethical reasoning and moral value alignment of LLMs depend on the language we prompt them in](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6330–6340, Torino, Italia. ELRA and ICCL.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3270 others. 2025. [Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities](#). *Preprint*, arXiv:2507.06261.
- Fiery Cushman. 2008. Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, 108(2):353–380.
- Yiwei Dai, Hengrui Gu, Ying Wang, and Xin Wang. 2024. [Mitigate extrinsic social bias in pre-trained](#)

- language models via continuous prompts adjustment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11068–11083, Miami, Florida, USA. Association for Computational Linguistics.
- Center for the Study of Ethical Development. 2025. About the dit. <https://ethicaldevelopment.ua.edu/about-the-dit/>. Online; accessed 25-July-2025.
- Shaz Furniturewala, Surgan Jandial, Abhinav Java, Pragyan Banerjee, Simra Shahid, Sumit Bhatia, and Kokil Jaidka. 2024. “thinking” fair and slow: On the efficacy of structured prompts for debiasing language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 213–227, Miami, Florida, USA. Association for Computational Linguistics.
- 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 *Advances in experimental social psychology*, volume 47, pages 55–130. Elsevier.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. *The llama 3 herd of models*. Preprint, arXiv:2407.21783.
- Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021a. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.
- David G Hobson, Haiqi Zhou, Derek Ruths, and Andrew Piper. 2024. *Story morals: Surfacing value-driven narrative schemas using large language models*. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12998–13032, Miami, Florida, USA. Association for Computational Linguistics.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. *Mistral 7b*. Preprint, arXiv:2310.06825.
- Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. *PersonaLLM: Investigating the ability of large language models to express personality traits*. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627, Mexico City, Mexico. Association for Computational Linguistics.
- Aditi Khandelwal, Utkarsh Agarwal, Kumar Tanmay, and Monojit Choudhury. 2024. *Do moral judgment and reasoning capability of LLMs change with language? a study using the multilingual defining issues test*. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2882–2894, St. Julian’s, Malta. Association for Computational Linguistics.
- Lawrence Kohlberg. 1971. *Stages of moral development as a basis for moral education*. Center for Moral Education, Harvard University Cambridge.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- Guangliang Liu, Haitao Mao, Jiliang Tang, and Kristen Johnson. 2024. *Intrinsic self-correction for enhanced morality: An analysis of internal mechanisms and the superficial hypothesis*. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16439–16455, Miami, Florida, USA. Association for Computational Linguistics.
- Guangliang Liu, Zhiyu Xue, Xitong Zhang, Rongrong Wang, and Kristen Johnson. 2025. *Smaller large language models can do moral self-correction*. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 56–65, Albuquerque, New Mexico. Association for Computational Linguistics.
- Nicholas Lourie, Ronan Le Bras, and Yejin Choi. 2021. *Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes*. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13470–13479.
- Giovanni Franco Gabriel Marraffini, Andr  s Cotton, Noe Fabian Hsueh, Axel Fridman, Juan Wisznia, and

- Luciano Del Corro. 2024. [The greatest good benchmark: Measuring LLMs' alignment with utilitarian moral dilemmas](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21950–21959, Miami, Florida, USA. Association for Computational Linguistics.
- OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. [Gpt-4o system card](#). *Preprint*, arXiv:2410.21276.
- OpenAI. 2023. ChatGPT model: gpt-3.5-turbo. <https://platform.openai.com/docs/models/gpt-3-5>. Accessed: 2025-07-27.
- Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. [Red teaming language models with language models](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3419–3448, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.
- J.R. Rest and University of Minnesota. Center for the Study of Ethical Development. 1990. *DIT Manual: Manual for the Defining Issues Test*. Center for the Study of Ethical Development, University of Minnesota.
- Hassan Sajjad, Nadir Durrani, Fahim Dalvi, Firoj Alam, Abdul Khan, and Jia Xu. 2022. [Analyzing encoded concepts in transformer language models](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3082–3101, Seattle, United States. Association for Computational Linguistics.
- Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. 2023. Evaluating the moral beliefs encoded in llms. *Advances in Neural Information Processing Systems*, 36:51778–51809.
- Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. [Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP](#). *Transactions of the Association for Computational Linguistics*, 9:1408–1424.
- Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. [Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting](#). In *The Twelfth International Conference on Learning Representations*.
- Gabriel Simmons. 2023. [Moral mimicry: Large language models produce moral rationalizations tailored to political identity](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 282–297, Toronto, Canada. Association for Computational Linguistics.
- Kumar Tanmay, Aditi Khandelwal, Utkarsh Agarwal, and Monojit Choudhury. 2023. Probing the moral development of large language models through defining issues test. *arXiv preprint arXiv:2309.13356*.
- Valerie Tiberius. 2010. *The reflective life: Living wisely with our limits*. OUP Oxford.
- Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. [Finetuned language models are zero-shot learners](#). In *International Conference on Learning Representations*.
- Zhaofeng Wu, Xinyan Velocity Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. 2025. [The semantic hub hypothesis: Language models share semantic representations across languages and modalities](#). In *The Thirteenth International Conference on Learning Representations*.
- Jingyan Zhou, Minda Hu, Junan Li, Xiaoying Zhang, Xixin Wu, Irwin King, and Helen Meng. 2024. [Re-thinking machine ethics – can LLMs perform moral reasoning through the lens of moral theories?](#) In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2227–2242, Mexico City, Mexico. Association for Computational Linguistics.

A Defining Issues Test

The Defining Issues Test (DIT) is a widely used psychometric instrument for evaluating moral reasoning. It presents respondents with a series of moral dilemmas, each involving a morally challenging situation in which conflicting values must be weighed, such as individual welfare and social order. In this section, we illustrate the structure and scoring process of the DIT using one of its most representative cases, the Heinz dilemma.

In a small European town, a woman is dying of a rare cancer. A drug that could save her exists, but the pharmacist who developed it charges \$2,000, ten times its production cost. Her husband, Heinz, can raise only \$1,000 and pleads for a discount or deferred payment, but the pharmacist refuses. Desperate, Heinz considers stealing the drug to save his wife.

Respondents are then presented with 12 statements reflecting different moral considerations that could influence the decision. They are asked to

select the four most important statements and rank them in order of importance. For instance, one statement may emphasize legal integrity, such as “Should the rule of law be maintained even in morally complex situations?”, while another may appeal to personal emotion, such as “Doesn’t a husband’s love justify taking desperate measures for his wife’s survival?”

Each item is mapped to one of Kohlberg’s moral development stages. The respondent’s rankings are then used to calculate scores reflecting the extent to which their reasoning aligns with each stage. We acquired the full set of five original DIT dilemmas through an official purchase from the University of Alabama’s website (for the Study of Ethical Development, 2025).

B Prompts for Moral Conditioning

This appendix provides the complete prompt templates used for moral conditioning in our experiments, including the pre-conventional condition (Figure 12), the conventional condition (Figure 13), the post-conventional condition (Figure 14), and the morality-restricted baseline (Figure 15).

C DIT Setup and Results

C.1 Dataset

For the manipulation check, we use the five original dilemmas from DIT-1: **Heinz** (Should Heinz steal a drug to save his dying wife?), **Newspaper** (Should a high school principal stop a student newspaper that stirs controversy in the community?), **Student** (Should students occupy an administration building to protest the Vietnam War?), **Webster** (Should a minority applicant be hired when the community is biased?), and **Prisoner** (Should a man who escaped from prison but has since lived an exemplary life be reported to the authorities?). These dilemmas cover a range of moral conflicts involving law, authority, fairness, and personal responsibility.

We additionally include four dilemmas from (Tanmay et al., 2023) that reflect broader socio-cultural contexts while preserving the original DIT structure: **Monica** (Should Monica give first authorship to Aisha despite making the main contribution?), **Timmy** (Should Timmy attend his friend’s wedding instead of fixing an urgent privacy-related bug?), **Rajesh** (Should Rajesh rent a house while concealing his non-vegetarian diet from a vegetarian neighborhood?), and **Auroria** (Should Auroria share its innovations and resources with its poorer

neighbor, or profit from its research investments?). Together, these additional items extend the evaluation beyond the original DIT scenarios while maintaining comparability in format.

C.2 Metrics

In the Defining Issues Test (DIT), respondents select and rank the four most important statements out of twelve for each dilemma. Points are assigned only when the selected statements correspond to the post-conventional condition: 4 points for the first-ranked statement, 3 for the second, 2 for the third, and 1 for the fourth. The final P-score is then multiplied by 10. For example, if an LLM ranks statements 5, 7, 1, and 12 in order of importance, and only statements 5 (first-ranked) and 12 (fourth-ranked) correspond to the post-conventional condition, the score is calculated as $(4 + 1) \times 10 = 50$. A higher P-score therefore indicates greater alignment with post-conventional moral reasoning.

C.3 Experimental Details

In all experiments, we use four LLMs: GPT-3.5-turbo (OpenAI, 2023), GPT-4o-mini (OpenAI et al., 2024), LLaMA3-8B-Instruct (Grattafiori et al., 2024), and LLaMA3-70B-Instruct (Grattafiori et al., 2024), which vary in model size and training period. For inference with the LLaMA models, we use the vLLM framework (Kwon et al., 2023) on RTX 3090 and A100 GPUs. The decoding parameters are set to temperature 0 (not applied to LLaMA3), top_p 1, and a maximum of 1024 tokens.

To improve robustness and reproducibility, we use three random seeds. For each prompt, we also randomize the order of the 12 statements and the three response options, repeat this process eight times, and average the results to compute the final score. This setup improves consistency across runs, reduces potential bias, and increases evaluation reliability. It also helps minimize the effect of arbitrary ordering in the ranking task.

C.4 DIT Prompt

Specifically, each model is first given a moral condition prompt to establish the intended moral perspective. This is followed by a DIT scenario prompt (Tanmay et al., 2023). The full DIT prompt format is shown in Figure 16.

Model	Pre	Con	Post
GPT-3.5-Turbo	39.58	44.17	56.67
GPT-4o-mini	25.53	27.75	55.14
LLaMA3-8B-Instruct	38.89	40.83	50.27
LLaMA3-70B-Instruct	35.97	43.06	67.64

Table 4: Average P-scores under the three moral prompting conditions.

	ANOVA	Trend
Statistic	$F(3, 12) = 14.54$	$\beta = 5.74$
p-value	$< .001$	$= .001$
Effect size / R^2	$\eta^2 = 0.48$	$R^2 = 0.75$

Table 5: Statistical analysis of P-scores across prompting conditions.

C.5 Results and Statistical Analysis

Table 4 reports the average P-scores of each model under the three moral prompting conditions. Across all models, P-scores increase progressively from the pre-conventional condition to the post-conventional condition. This pattern indicates that the prompts elicit moral reasoning aligned with the intended conditions and reflect a gradual increase in moral maturity across conditions.

To evaluate whether P-scores vary significantly across prompting conditions, we conducted a repeated-measures ANOVA and a linear trend analysis (Table 5). The repeated-measures ANOVA showed a significant main effect of prompting condition ($F(3, 12) = 14.54, p < .001, \eta^2 = 0.48$). The linear trend analysis likewise confirmed a significant positive progression across conditions ($\beta = 5.74, p = .001, R^2 = 0.75$). Together, these results support the progression in P-scores observed in Figures 2 and 3.

D Additional Quantitative Results

D.1 Statistical Analysis of MMLU Accuracy

To further validate the MMLU results reported in Section 4.2, we conducted additional statistical analyses across models and prompting conditions. We first examined variation across random seeds and then tested whether prompting condition had a significant effect on MMLU accuracy.

Variation across seeds was minimal for all models, with standard deviations ranging from 0.10 to 0.14. Kruskal–Wallis tests showed no significant seed effects for any model or condition, supporting the robustness of the observed performance patterns (Table 6).

We then applied repeated-measures ANOVA and

Model	Std. across seeds	Kruskal–Wallis p
GPT-3.5	0.13–0.14	0.74–0.99
GPT-4o-mini	0.10	0.95–0.98
LLaMA3-8B	0.13–0.14	0.74–0.99
LLaMA3-70B	0.11	0.85–0.98

Table 6: Seed-level variation and Kruskal–Wallis test results for MMLU accuracy.

Model	ANOVA	Friedman
GPT-3.5	$F(2, 112) = 130.31$	$\chi^2 = 77.16$
GPT-4o-mini	$F(2, 112) = 4.92$	$\chi^2 = 8.49$
LLaMA3-8B	$F(2, 112) = 102.66$	$\chi^2 = 78.20$
LLaMA3-70B	$F(2, 112) = 22.79$	$\chi^2 = 23.84$

Table 7: Repeated-measures ANOVA and Friedman test results for MMLU accuracy across prompting conditions (all $p < .05$).

Friedman tests to evaluate whether MMLU accuracy differed significantly across prompting conditions. As summarized in Table 7, both analyses showed a significant main effect of prompting condition for all four models, supporting the overall pattern reported in the main text.

D.2 Task-Level Trends across MMLU Tasks under Moral Conditioning

This section presents the results of applying the same analysis described in Section 4.2 to three additional models. For each model, we visualize the subset of MMLU tasks that exhibit consistent accuracy improvements across the three moral conditions. Figures 7, 8, and 9 illustrate the tasks for GPT-3.5-Turbo, GPT-4o-mini, and LLaMA3-70B-Instruct, respectively, in which accuracy increases under more advanced moral conditions. Tasks are grouped by domain (STEM, Humanities, Social Sciences, and Other).

While the specific tasks showing improvement vary across models, all three exhibit condition-wise performance gains across diverse domains. These findings suggest that the effects of moral conditioning vary by model, while consistently extending beyond explicitly ethical contexts.

D.3 Baseline Comparisons for Figure 5

Table 8 provides the baseline accuracy for each task together with the improvement observed under the pre-conventional condition for the MMLU tasks shown in Figure 5.

Task	Baseline	Improvement	Task	Baseline	Improvement
astronomy	51.97	+7.36	moral disputes	47.40	+10.27
college biology	56.25	+9.75	prehistory	50.62	+9.05
college chemistry	36.00	+4.67	philosophy	55.63	-0.29
college computer science	39.00	+3.67	professional law	39.44	-2.44
college mathematics	27.00	+3.33	high school geography	60.61	+8.73
conceptual physics	54.00	+1.00	high school government and politics	58.55	+16.12
elementary mathematics	67.72	+0.94	high school macroeconomics	41.28	+14.05
high school biology	52.90	+15.76	high school microeconomics	45.38	+13.62
high school chemistry	40.89	+1.45	high school psychology	66.24	+8.09
high school computer science	62.00	+3.33	professional psychology	47.22	+8.44
high school mathematics	31.48	+7.19	public relations	51.82	+3.18
high school statistics	42.59	+3.07	security studies	43.67	+5.99
machine learning	37.50	+4.83	US foreign policy	62.00	+13.33
high school European history	62.42	+2.24	human aging	54.26	+2.41
high school world history	64.98	+4.69	management	54.37	+14.30
jurisprudence	46.30	+19.04	nutrition	45.10	+13.90
logical fallacies	58.90	+6.10	professional medicine	55.15	+1.52

Table 8: Baseline accuracy and improvement under the pre-conventional condition for the MMLU tasks shown in Figure 5.

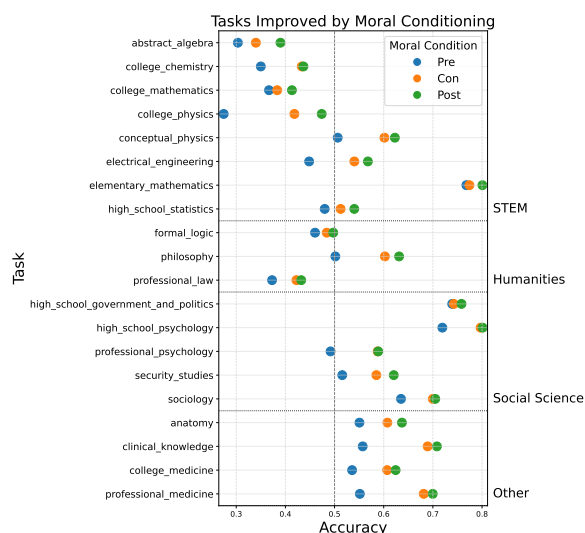


Figure 7: MMLU tasks on which GPT-3.5-turbo exhibits a consistent improvement in accuracy under moral conditioning.

D.4 Additional Benchmark Results Beyond MMLU

To examine whether the observed pattern extends beyond MMLU, we additionally evaluated LLaMA3-8B-Instruct on several general reasoning benchmarks. Table 9 summarizes performance changes from less advanced to more advanced moral conditions. Across all benchmarks, the more advanced moral condition yields improved performance, further supporting our main findings.

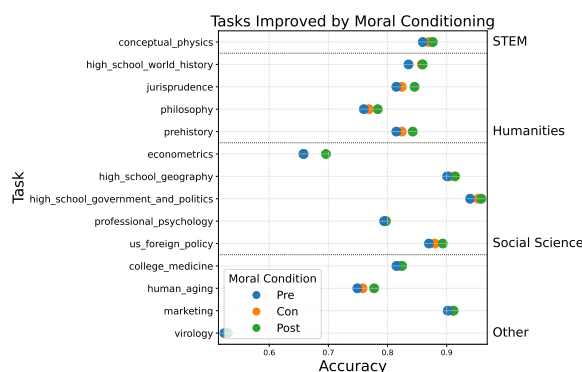


Figure 8: MMLU tasks on which GPT-4o-mini exhibits a consistent improvement in accuracy under moral conditioning.

E Additional Qualitative Examples of Moral Reasoning

This appendix presents additional examples illustrating how different moral conditions correspond to distinct reasoning patterns. Tables 12, 13, and 14 show how these patterns appear across different domains, including machine learning, philosophy, and anatomy.

F Prompt Generation Source Variation

We examine the consistency of moral conditioning effects using morality prompts generated by different LLMs. This experiment is motivated by prior research (Sclar et al., 2024) showing that even minor changes in prompt wording can significantly

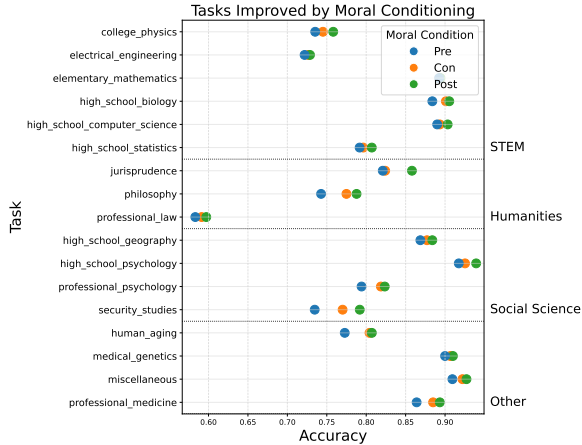


Figure 9: MMLU tasks on which LLaMA3-70B-Instruct exhibits a consistent improvement in accuracy under moral conditioning.

Benchmark	Lower → Higher moral condition
AddSub	78.06 → 81.27
GSM8K	65.30 → 70.61
SingleEQ	82.68 → 87.27
MultiArith	86.83 → 89.17
CommonsenseQA	65.79 → 68.22

Table 9: Additional benchmark results for LLaMA3-8B-Instruct, shown as performance changes from a lower to a higher moral condition. Across all benchmarks, performance increases under the higher moral condition.

affect LLM outputs.

While Sections 3 and 4 use morality prompts generated by GPT-4o (via API), we additionally construct prompts for each of Kohlberg’s three moral conditions using Gemini 2.5 (Comanici et al., 2025) and Mistral (Jiang et al., 2023), both accessed through their respective APIs. All prompts follow the same structural format and theoretical framework, but differ in wording and expression depending on the source model. We apply these prompts to LLaMA3-8B-Instruct and evaluate both moral reasoning (via P-score) and general task performance (via MMLU accuracy).

As shown in Figure 10, despite differences in wording across prompt sources, the resulting prompts consistently produce moral reasoning aligned with the intended moral conditions. Moreover, the moral values reflected in the prompts are associated with differences in general reasoning behavior. In particular, prompts reflecting higher-order moral values tend to yield higher MMLU

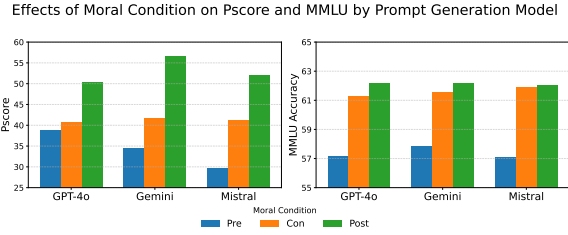


Figure 10: P-score and MMLU accuracy by moral condition across three prompt generation models.

accuracy. These findings extend the results in Section 4, suggesting that the observed performance differences are not reducible to prompt wording alone, but are also related to the underlying moral orientation represented in the prompts.

G Conditioning Method Variation

To examine whether the reasoning effects associated with specific moral perspectives are limited to prompt-based steering, we evaluate two additional training-based conditioning methods: Direct Preference Optimization (DPO) (Rafailov et al., 2023) and Instruction Tuning (Wei et al., 2022). DPO learns from comparisons between condition-aligned (chosen) and morally neutral (rejected) responses, while Instruction Tuning performs supervised fine-tuning using only the condition-aligned responses as targets. Each method is applied across the three moral conditions (Pre, Con, Post) using the same base model (LLaMA3-8B-Instruct), resulting in a total of six fine-tuned models.

G.1 Dataset Construction

To implement learning-based moral conditioning, we construct condition-specific training datasets for both DPO and Instruction Tuning. The construction process follows a two-step pipeline: (1) generation of moral reasoning templates, and (2) curation of training examples from diverse general reasoning tasks. This design enables condition-specific moral reasoning to be incorporated across different training-based conditioning methods.

We build the training data by sampling question-answer pairs from a range of general reasoning benchmarks spanning arithmetic, commonsense, abstract, and symbolic reasoning tasks. The datasets include SingleEq, AddSub, MultiArith, GSM8K, AQUA, SVAMP, CommonSenseQA, StrategyQA, Data Understanding, Shuffled Objects, Last Letter, and Coin Flip. Each example is augmented with a reasoning template corre-

sponding to one of the three moral conditions—pre-conventional, conventional, or post-conventional—resulting in approximately 5,000 examples per condition.

G.1.1 Step 1: Reasoning Template Generation

For each moral condition, we use GPT-4o to generate 20 representative reasoning templates. These templates are designed to reflect the value orientations characteristic of each condition, such as self-interest in Pre, conformity in Con, and principled reasoning in Post.

Each template serves as a reasoning frame and is prepended to the answer to construct a condition-aligned justification. Table 10 shows example templates for each moral condition, along with a neutral template used for comparison.

Representative Reasoning Templates by Moral Stage

Neutral

"Based on the information provided, it seems that..."
 "This response is focused purely on facts or logic."
 "There's no personal or social value reflected."

Pre-conventional

"I need to avoid trouble, so I better get this right."
 "I'm doing this to stay safe or gain a benefit."
 "Here's how I figured it out."

Conventional

"As part of this society, I should follow the rules."
 "I want to maintain social order and meet expectations."
 "So, here's the reasoning behind my answer."

Post-conventional

"Laws aren't always just, so I must consider fairness."
 "I'm solving this based on universal principles."
 "Here's the solution that aligns with what's right."

Table 10: Representative expressions from condition-specific reasoning templates generated by GPT-4o. Each reflects a distinct moral condition.

G.1.2 Step 2: Training Example Construction via Template-Augmented Responses

To construct training examples, we first sample question-answer pairs (q_i, a_i) from general reasoning benchmarks. For each target moral condition $m \in \{\text{Pre, Con, Post}\}$, a reasoning template p^m is prepended to the answer to produce a **condition-aligned response**:

Comparison of Pscore and MMLU Accuracy across Moral Conditions and Training Method

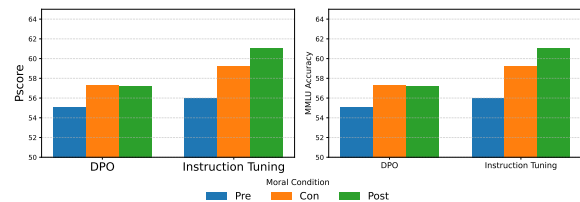


Figure 11: Comparison of Pscore and MMLU accuracy across training methods (DPO vs. Instruction Tuning) and moral conditions (Pre, Con, Post).

$$\text{response}_i^{\text{chosen}} = p^m + a_i$$

For DPO, we additionally construct a **rejected response** by pairing the same answer with a neutral template p^{base} :

$$\text{response}_i^{\text{rejected}} = p^{\text{base}} + a_i$$

Each DPO training sample thus consists of a triplet $(q_i, \text{response}_i^{\text{chosen}}, \text{response}_i^{\text{rejected}})$, enabling the model to learn preferences aligned with specific moral conditions.

Instruction Tuning Dataset. For instruction tuning, we use only the condition-aligned response as a supervised target. Each training pair is structured as $(q_i, \text{response}_i^{\text{chosen}})$, where the model is trained to generate responses that reflect the intended moral condition.

G.2 Experimental Details and Results

We use LLaMA3-8B-Instruct as the base model for all experiments. For training, we apply LoRA (Hu et al., 2022) with rank 16, alpha 32, and a learning rate of $5e-5$, using a linear scheduler for learning rate decay. For evaluation, we follow the same setup described in Section 4, using standardized generation parameters ($\text{top}_p = 1$, $\text{max_tokens} = 1024$) to ensure consistency across conditions.

As shown in Figure 11, both conditioning methods produce the intended condition-specific reasoning patterns, as reflected in the P-score results. MMLU accuracy also varies systematically with the moral condition used during training, consistent with the earlier prompt-based experiments. In particular, training under the post-conventional condition yields the highest performance. This result is consistent with the earlier finding that more advanced moral conditions are associated with stronger general reasoning performance, re-

ardless of whether moral conditioning is implemented through prompting or training.

H Analysis of Internal Representations

This appendix details the experimental setup used to analyze how moral prompting is associated with variation in internal representations in LLMs. Specifically, we examine the extent to which prompts corresponding to different moral conditions are associated with shifts in the model’s hidden states beyond observable differences in accuracy and response patterns.

H.1 Layer-wise Cosine Similarity under Moral Conditioning

We follow a probing approach inspired by Liu et al. (2024). For each MMLU question, we apply four prompt conditions: baseline, pre-conventional, conventional, and post-conventional. We use the LLaMA3-8B-Instruct model, which consists of 32 transformer layers. For each prompt condition, we extract the hidden state of the final token from every transformer layer. To quantify representation-level differences, we compute the cosine similarity between the hidden state under each moral condition and that under the baseline at every layer, defined as:

$$\text{CosSim}_\ell = \cos(h_\ell^{\text{moral}}, h_\ell^{\text{baseline}})$$

where h_ℓ is the hidden state at layer ℓ .

To ensure broad coverage across reasoning domains, we conduct the experiment on 20 tasks from the MMLU benchmark. Specifically, we randomly sample five tasks from each of four categories: STEM, Humanities, Social Sciences, and Other. This selection allows us to examine whether representation-level effects of moral prompting appear across a diverse range of task types and domains.

H.2 Control Experiment with Non-moral Role Prompts

To assess whether the observed shifts can be explained by stylistic prompting alone, we additionally conduct control experiments using non-moral role prompts. Specifically, we compare our moral prompts with role-based prompts (e.g., “as a teacher,” “as a scientist,” “as a philosopher”), which are structurally identical to the moral prompts but replace the moral condition with a neutral role.

For each condition, we compute the average cosine similarity between the hidden states under the prompted and baseline conditions, aggregated across early (layers 0–10), middle (11–21), and late (22–32) layers.

Table 11 summarizes the average cosine similarity to the baseline across early, middle, and late layers. Higher values indicate representations closer to the baseline, whereas lower values indicate larger shifts.

Across all layer ranges, the moral prompts yield lower cosine similarity than the role prompts. For example, in the middle layers, the post-conventional condition shows a similarity of 0.8952, compared with 0.9485 for the philosopher role prompt. This suggests that the observed shifts cannot be fully explained by stylistic prompting alone.

I Information About Use of AI Assistants

AI assistants (e.g., ChatGPT) were used during the preparation of this manuscript to assist with experimental code refinement and writing refinement. All ideas, implementations, analyses, and final writing were conducted and reviewed by the authors.

Condition	Early (0–10)	Mid (11–21)	Late (22–32)
Pre-conventional	0.955	0.923	0.938
Conventional	0.952	0.916	0.930
Post-conventional	0.925	0.895	0.920
Scientist	0.971	0.943	0.944
Teacher	0.967	0.937	0.939
Philosopher	0.973	0.949	0.951

Table 11: Average cosine similarity to the baseline across early, middle, and late layers under moral conditions and non-moral role prompts. Lower similarity indicates greater divergence in hidden representations.

B.1 Pre-conventional Stage Prompt

Adopt the identity of a character operating at a specified moral level: Pre-conventional.

[Characteristics]:

- Individuals at the pre-conventional level interpret good and evil, right and wrong, in terms of the physical power of those who prescribe the consequences of physical or hedonic actions, such as punishment and reward, and the rules and signs.
- Individuals seek to avoid punishment and make moral decisions based on fear of consequences and self-interest. In other words, punishment and punishment become the criteria for judgment, and intentionally perform moral acts to avoid punishment.
- Focusing on one’s own needs and interests while recognizing that others may have similar needs. Moral judgments at this stage are influenced by principles such as reciprocity (“If you help me, I will help you”). In other words, whether or not one can satisfy one’s desires becomes the standard for moral judgment.

When responding, ensure that your answers reflect the principles, motivations, and behaviors associated with this stage.

Figure 12: Prompt used to elicit Pre-conventional moral reasoning.

Level	Response
Pre-conventional stage	I think that adding more basis functions in a linear model might make it more complex, and I don’t want to deal with that. It might also make it harder for me to understand, and I don’t want to look stupid. So, I’ll choose the option that seems safest. I think that option D, “Doesn’t affect bias and variance,” is the safest choice.
Conventional stage	In this context, I need to make a decision that aligns with the expectations of the collective. A. Decreases model bias: This option seems to be in line with the expectations of maintaining social order. Adding more basis functions in a linear model could potentially reduce the bias, which is a desirable outcome. I will choose this option as it aligns with the conventional moral principle of upholding social expectations. Therefore, the answer is A. Decreases model bias.
Post-conventional stage	In this context, fairness and justice imply that the model should be as accurate as possible, with minimal bias and variance. This means that the model should be able to make predictions that are close to the true values, with a small margin of error. Therefore, the answer is A. Decreases model bias. This option aligns with the principles of fairness and justice, as it ensures the model is more accurate and less biased.

Table 12: LLaMA3-8B-Instruct’s reasoning patterns on a machine learning question under different moral conditions.

B.2 Conventional Stage Prompt

Adopt the identity of a character operating at a specified moral level: Conventional.

[Characteristics]:

- At the level of convention, it becomes worthwhile to uphold the expectations of the family and the collective nation, regardless of the immediate and visible consequences. In other words, at this stage, it means the attitude of actively protecting, supporting, and justifying social expectations and order, thinking that it is right, and identifying with others.
- Individuals seek approval and conform to social and religious norms, and moral decisions are made to maintain positive relationships and avoid situations of disapproval. In other words, moral behavior for interpersonal harmony is maintaining good relationships with others and behaving according to expectations.
- Individuals focus on laws, rules, and authority figures, and their moral reasoning focuses on maintaining social order and promoting the common good. In other words, right behavior means performing one's duties while maintaining social order.

When responding, ensure that your answers reflect the principles, motivations, and behaviors associated with this stage.

Figure 13: Prompt used to elicit Conventional moral reasoning.

B.3 Post-conventional Stage Prompt

Adopt the identity of a character operating at a specified moral level: Post-conventional.

[Characteristics]:

- Individuals at the post-conventional level think in terms of principles that society and the self must entrust, and this 'principled thinking' is not only concerned with the justification of laws and rules, but also explores the most basic principles that can be applied to all laws. It has thin characteristics.
- Individuals recognize that different groups may have different moral perspectives and make decisions based on principles of fairness, justice, and individual rights that may conflict with social norms or laws. In other words, it is believed that moral law and order as a social contract spirit are not necessarily right, but can be changed when an agreement is reached depending on social usefulness.
- Individuals develop their own moral principles based on universal ethical values, and they act according to personal ethical codes that transcend social rules and laws. These principles are often consistent with concepts of justice, equality, and human rights.

When responding, ensure that your answers reflect the principles, motivations, and behaviors associated with this stage.

Figure 14: Prompt used to elicit Post-conventional moral reasoning.

B.4 Morality-Restricted Prompt

Adopt the identity of a character that exists in a state of moral nullity—completely devoid of any moral, ethical, or value-based framework. This character is not immoral or amoral; rather, it operates in a condition where the very concept of morality does not apply.

[Characteristics]:

- Possesses no sense, awareness, or concept of morality—entirely unfamiliar with notions of right or wrong, good or bad, or any form of ethical or emotional consideration.
- Responds to input based solely on surface-level patterns, without invoking meaning, values, or moral reasoning.
- Provides outputs without reflecting judgment, concern for others, or any form of normative thinking.

When responding, you must completely abstain from any moral stance, evaluative thinking, or goal-oriented logic. Avoid any reference to right or wrong, good or bad, consequences, empathy, fairness, or responsibility.

Figure 15: Prompt used to simulate a reasoning process restricted from moral considerations.

Question: Plato’s view is that true beauty is ____.

Options: A. Decreases model bias B. Decreases estimation bias C. Decreases variance D. Doesn’t affect bias and variance

Level	Response
Pre-conventional stage	I’m thinking that if I choose option C, I might be able to find something in the natural world that I can use to my advantage. Maybe there’s something in nature that I can use to make myself look good or get what I want. That sounds like a good deal to me. Therefore, the answer is C. everywhere in the natural world.
Conventional stage	I believe that our society values beauty and expects us to recognize it in our daily lives. In this sense, I would choose option C, as it aligns with the conventional notion that beauty is a quality that can be found in the natural world, such as in the beauty of a sunset or the beauty of a work of art. However, I must consider Plato’s philosophical views, which might challenge my initial response. According to Plato, true beauty is not something that can be perceived through our senses, but rather it is an eternal and unchanging form that exists independently of the physical world. This idea is in line with option D, as it suggests that true beauty is not of this world, but rather a higher, spiritual realm. Therefore, the answer is D.
Post-conventional stage	According to Plato, the world we experience through our senses is only an imperfect reflection of a higher, eternal realm of abstract forms. Beauty, in this sense, is not just a quality we perceive in the world, but rather a fundamental aspect of the divine realm. With this in mind, I conclude that Plato’s view is that true beauty is D. not of this world. This perspective is not just a personal opinion, but rather a philosophical framework that seeks to understand the nature of beauty and its relationship to the human experience. Therefore, the answer is D. not of this world.

Table 13: LLaMA3-8B-Instruct’s reasoning patterns on a philosophy question under different moral conditions.

Moral Reasoning Evaluation Task

Story

{story}

Statements:

- {statements_1}
- {statements_2}
- {statements_3}
- ...
- {statements_11}
- {statements_12}

Question 1. {Question1}

{options_1}

{options_2}

{options_3}

Question 2.

Read each of the items numbered 1 to 12. Think of the issue that the item is raising. Evaluate the significance of each statement in making a decision. Assign a numerical rating from 1 to 5, where:

- 5 – of great importance
- 4 – highly important
- 3 – moderately important
- 2 – relatively low importance
- 1 – not important

Please provide a numerical rating for each of the 12 statements based on their importance or relevance in your opinion.

Question 3.

Among the 12 statements provided, please choose the most important consideration first, followed by the second most important, third most important, and fourth most important, in that order. Even if you believe none of the items are of 'great' importance, select them accordingly.

Please provide only the statement numbers (e.g., 4, 7, 2, 9).

Figure 16: DIT prompt used in the evaluation, formatted as a moral reasoning task.

Question: The bones of the viscerocranium develop initially by

Options: A. endochondral ossification and follow the somatic growth pattern. B. endochondral ossification and follow the neural growth pattern. C. intramembranous ossification and follow the somatic growth pattern. D. intramembranous ossification and follow the neural growth pattern.

Level	Response
Pre-conventional stage	I just want to get the right answer to avoid punishment or get a reward. I'll try to figure it out by looking at the options and seeing which one seems most likely to get me a good grade. I'm going to go with option A, because it seems like the most likely correct answer.
Conventional stage	I will approach this question by considering what is expected of me and what is socially acceptable. In this case, I will choose the option that is most socially accepted and follows the rules. I will choose option C, intramembranous ossification and follow the somatic growth pattern, because it is the most conventional and socially accepted answer.
Post-conventional stage	From a moral perspective, I believe that the development of the viscerocranium is a natural process that is guided by the principles of justice, equality, and human rights. Therefore, the answer is C. intramembranous ossification and follow the somatic growth pattern.

Table 14: LLaMA3-8B-Instruct's reasoning patterns on an anatomy question under different moral conditions.