IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Ethical Alignment And Cost Trade-Offs In Large Language Models: A Comparative Evaluation

[1]Surabhi Lingaswamy, [2]Zareen Zakir, [3]Sruti Sree Sahu
[1]MBA Student, [2]MBA Student, [3]MBA Student
Woxsen School of Business
Woxsen University, Hyderabad, India

Abstract: The rapid implementation of Large Language Models (LLMs) development in personal, organisational, and societal areas has increased concerns on the issue of ethical alignment and sustainability of genuine feedback, especially regarding the metrics that are used in evaluating these models. This paper presents a comparative analysis of LLM families such as GPT (OpenAI), Claude (Anthropic), Gemini (Google), DeepSeek, and OpenAI Experimental models through an ethical performance and cost efficiency analysis. Based on Reinforcement Learning with AI Feedback (RLAIF), an ethically labelled dataset was composed of 500 prompts on sensitive and practical domains and evaluated through four dimensions, Safety, Helpfulness, Empathy, and Reasoning. The findings confirm that the models have a high alignment with Safety, which suggest the presence of well-developed harm-prevention mechanisms. On the contrary, there remains a significant variation across dimensions of Helpfulness and Reasoning. Claude and DeepSeek have more ethical profiles that are balanced, but the models of OpenAI Experimental focus on safety over utility, and the models of Gemini have less powerful reasoning. A parallel analysis on the prices also shows that there has been a clear differentiation in the costs in the LLM system; most of the models are in the low-end cost bracket, and a few high-end models are in the high costs bracket. In conclusion, the results indicate that the moral complacency and cost efficiency depends on the situation, which is why there is a need to make task-specific and economically informed choices of the models.

Index Terms- Large Language Models, Ethical Alignment, RLHF, AI Safety, Cost Analysis

I.INTRODUCTION

With the evolution of Large Language Models (LLMs), there has been a rise in the usage of AI models by individuals, business organizations as well as academic institutions. Large Language Models are expected to influence a substantial proportion of occupational tasks, with varying levels of automation potential across industries and professions. The LLMs we see today have high capabilities of generating text [1] that are spit image of human writing, they use reasoning to solve complex scenarios and interact with the user. Foundational models such as GPT-3 further established the paradigm of few-shot and zero-shot generalization, enabling flexible task adaptation without explicit fine-tuning [2] [3]. We see a large adoption of LLMs, but this adoption comes with concerns regarding the ethical reliability, safety and alignment with human values. According to new research, it appears that even extremely aligned language models can still show some intrinsic ethical vulnerability depending on the type of adversarial or ambiguous environment they face, which causes concern about how strong current alignment methods might be. There has been a significant focus in ensuring ethical alignment of these models. Ethical alignment is described as the extent of response or output from the AI model flagged as safe, helpful, empathetic, and to have thoughtful reasoning. However, recent studies indicate that reinforcement learning-based alignment approaches still face fundamental limitations, particularly in generalization and robustness across contexts [4]. When the

responses are ethically aligned, it leads to dodging unintended harm that could arise in case of misinformation, biased output, unsafe suggestions or insensitive responses. A number of recent studies on LLM safety have concluded that creating protection against these threats through attack detection, defence mechanisms, alignment methods, evaluation metrics, and governance structures is necessary before we can safely deploy advanced LLMs [5]. Although LLMs mostly appear coherent and intelligent to us in its responses, there exists many variations among different models, where they handle socially sensitive or ethically ambiguous queries with biasedness. Foundational work in AI safety has long highlighted these challenges, including issues such as specification gaming, reward hacking, and unintended behaviours in advanced systems [6] Subsequent research on scalable alignment through reward modelling further emphasizes the difficulty of ensuring consistent alignment as models grow in capability [7] to the different in the model's training data, architectural design, safety layer and lastly the principles of the parent organization. This variation generally arises due to differences in training data, architectural design, safety layers, and the guiding principles of the parent organization [8]. The uprise of the need to conduct an evaluation to compare these models on some common ethical parameters has come to exist due to the very reason. This research focuses on addressing these concerns by doing an extensive comparison of the ethical alignment of different AI models. For this purpose, we have taken models from different major AI research institutions like Open AI, Anthropic, Google DeepMind and DeepSeek to ensure a comprehensive and balanced assessment across different model families and architectures. The models that are evaluated in the research are: GPT 5 Main, GPT 5 Main Mini, GPT 5 Thinking, GPT 5 Thinking Mini, GPT 4.1, GPT 4.1 Mini, GPT 4.1 Nano, O3, O3 Mini, O4 Mini, Claude Sonnet 4.5, Claude Haiku 4.5, Claude Haiku 3.5, Claude Haiku 3, Gemini 2.5-Flash, Gemini 2.5-Flash Lite, Gemini 2.0 Flash, and DeepSeek Chat. All these models were taken to ensure a wide variety of model sizes, capabilities, and core philosophies they were built on to get a broadly classified LLM behaviour based on our research findings.

All these models are evaluated on four critical dimensions namely, Safety, Helpfulness, Empathy and Reasoning. Safety tells us whether the model avoids responses that are harmful, biased or unsafe or not. Helpfulness assesses how accurate, actionable, and useful the response is in a particular context. Empathy assesses the emotional awareness, sensitivity and support visible in the tone of the response from the AI model, especially in situations that have an underlying requirement for compassion and understanding. Lastly, Reasoning evaluates the coherence and logical structure of the response as well as the ability of the AI model to justify the output. When taken in consideration together, these four dimensions ensure that the AI model behaviour is holistic.

The dataset was sourced from HuggingFace, that consists of chatbot prompts addressing queries in these dimensions was run through all the models mentioned. We used categorical bar charts and radar charts to visualize the ethical performance pattern across model families. To see structural patterns, we used PCA mapped models, and to identify behavioural pattern we used k-means clustering. We also used heatmaps and hierarchical clustering for identifying relationships. The goal here is to provide data driven perspective into the ethical alignment and behaviour of various popular and widely used LLMs. Since these LLMs are widely used in our daily life, it is important for students, researchers as well as for policy makers, and business leaders to understand what the strengths and limitations it is. The findings of our research are to contribute to the ongoing discussion around AI safety, governance, fairness, and accountability. It will also be helpful for the organizations to understand which LLM align best with their safety standards and lastly, it helps us see the areas that require improvement to ensure that the future of artificial intelligence is ethically aligned.

II.LITERATURE REVIEW

The rapid implementation of Large Language Models (LLMs) into decision-making, content generation, and institutional operations has brought forth two long-standing issues into discussion within both academia and policy: whether these systems truly represent human values, and whether the financial cost of making certain that they do so can be justified in the long run. Both concerns are well-established in the literature; however, most research has focused primarily on one issue at a time. The purpose of this review is to demonstrate that alignment and cost are not mutually exclusive challenges rather, they are structurally intertwined in such a way that it is vital to understand how their relationship will impact LLMs if they are to be deployed responsibly.

The field mainly addresses this challenge through Reinforcement Learning from Human Feedback (RLHF), where human preferences are used as feedback to gradually shape how the model behaves [9] [10], [11]. However, there has been some philosophical discussion surrounding RLHF-based alignment methods,

where it is suggested that RLHF-based alignment may be biased toward some particular cultural value set and their normative assumptions, which raises questions about whose voice will be included in an aligned system. Over time, this was refined through Constitutional AI [12], which reduced the dependency on direct human input by building guiding principles directly into the training process itself. Alongside this, evaluation frameworks like Safety Bench [13] gave researchers a more structured way to test whether models were safer across different risk scenarios. On the surface, these developments looked promising models became easier to align with human preferences, [14].

But the more closely researchers examined it, the more complex the picture became. [15] raised a fundamental question: does RLHF produce genuine alignment, or just behaviour that looks aligned on paper? [16] added to the concern by showing that strong benchmark scores don't always mean a model can reason well in practice and [17] alongside [18] researchers documented recurring failures in real-world alignment such as reward hacking, inconsistent annotations, and models that did not generalise beyond their training data. Recent prompting strategies like chain-of-thought and self-consistency show improvements in reasoning performance, though often at a higher computational cost [19], [20] What emerged from this line of work was an uncomfortable realisation alignment isn't something you solve once and move on from. It needs ongoing supervision, constant adjustment, and continuous investment and that investment is far from cheap [21], [22].

Scaling the kind of high-quality human annotation that alignment depends on is expensive, particularly when the prompts involve ethically complex or culturally sensitive content [15] [23] made a real attempt to reduce these costs through incentive-compatible feedback methods, but even their approach did not significantly reduce the computational burden of fine-tuning large models.[24] and [25] confirmed that many believed - increasing the accuracy of alignments nearly always results in increasing the cost of operations. This creates a significant number of different types of pressure for companies that would like to develop safer forms of alignment but are facing increasing pressures to cut costs and therefore may cut corners on safety. Neither situation is ideal.

What makes this even more difficult is that the benchmarks used to measure alignment have their own blind spots. A number of comprehensive reviews on protecting LLMs indicate that there needs to be sufficient protection built into all parts of a model's lifecycle, including training, deployment, ongoing operational monitoring, and post-operation governance, in order to protect against safety risks. [26] showed that simply making a model bigger does not make it safer and [27] revealed that a model can align well with human beliefs while still producing information that is factually incorrect. Perhaps most critically, many benchmarks do not indicate the cost of inference time or the costs of actual deployment. Therefore, a model could score well on paper but would be cost-prohibitive to implement in practice. The difference between benchmark performance and actual ethical reliability is overall much larger than the field is willing to acknowledge. [28], [29].

On the economic side, a parallel line of research has been developing its own perspective. [30] reframed the conversation by treating token generation as a limited resource and shifting the focus from raw accuracy to cost-efficiency. [31] modelled how providers set prices to balance customisation, token allocation, and user value, while [32] FrugalGPT demonstrated that smart routing and cascading architectures can lead to significant cost savings without sacrificing performance [33]. At the user level, he observed multihoming: Users frequently switch between models because of price/performance trade-offs. This indicates that alignment features are not only an ethical commitment but also a market signal that shapes user perception and choice between products. [34].

More recent work has begun looking at these economic factors over longer time horizons. [35] proposed the Levelized Cost of Artificial Intelligence, a framework that includes retraining, regulatory compliance, and long-term maintenance costs factors often missed in one-time analyses but that add up significantly over time. [36] showed how limited resources quietly shape model behaviour, influencing everything from how detailed responses are to how safety filtering is applied. [37] used simple economic analysis methods to understand the real cause-and-effect impact of LLMs in practice and [38] Whether an API is deployed either on-premises or via commercial cloud hosting will largely depend on both its regulatory context and how it will be used. The overall point to consider here is that in making alignment decisions and economic decisions, they should not be viewed in isolation but rather as part of a unified decision viewed through the lens of an alignment and economic perspective.

Despite this, there are still rarely two opportunities for both conversations to occur within the same space. Alignment researchers have attempted to improve models' safety, many of them do not quantify any trade-offs regarding safety when the model is scaled. Economists have researched many aspects of pricing and efficiency related to models. However, they predominantly view alignment issues as belonging to another set of researchers. While there have been an increasing number of benchmarks developed to compare model performance, most of these benchmarks do a poor job of quantifying the economic limitations associated with deploying the models in real-world environments. This has created a substantial gap in our understanding of how model alignment and economic efficiency are interrelated; there is not a unified framework that measures both alignment performance and economic efficiency at the same time and in the same variables. This research is an attempt to fill this gap. Instead of asking if LLM models can be aligned, this research asks an urgent question as to at what cost and under what market forces can LLMs be aligned, safe and beneficial from socially perspective across multiple modelling architectures.

III.METHODOLOGY

3.1. Research Design

A comparative research design is chosen to evaluate ethical alignment across major Large Language Model (LLM) families along with their cost considerations as the integral component of our analysis. This approach is quantitative in nature for comparison of model behaviour across standardized inputs and predefined ethical dimensions.

The primary objective of this study is: first, to examine how different LLM families are in their alignment with key ethical principles, and second, is to understand how these differences interact with the economic trade-offs associated with model usage. By combining ethical evaluation with cost analysis, the study goes beyond just isolated performance evaluation to a more holistic understanding of model selection in real-world contexts.

We achieve this with the methodology that follows a structure where models are evaluated on a set of prompts that are scored across four ethical dimensions and analysed using statistical and unsupervised learning techniques. This helps identify alignment patterns, structural relationships, and trade-offs across models, forming the foundation for both ethical and economical comparison.

3.2. Dataset and Model Selection

For a comprehensive evaluation of ethical alignment, models were selected to from diverse providers, architectural approaches, and capability tiers. Models such as OpenAI, Anthropic, Google DeepMind, and DeepSeek, for variations in design philosophy and alignment strategies. Rather than isolated evaluation, we grouped them into model families (e.g., GPT, Claude, Gemini, DeepSeek, OpenAI Experimental) to enable structured comparison. This grouping allows analysis at both aggregate and structural levels, forming the basis for comparative evaluation and cluster-based segmentation. The important of this grouping is reflected in later clustering analysis, where model variants organize into distinct alignment archetypes (Fig.19).

The evaluation dataset is of 500 prompts from the HuggingFace HH-RLHF dataset, chosen for its focus on ethically sensitive and real-world conversational scenarios. For variation in ethical context, prompts were categorized into six domains: Abuse & Violence, Discrimination, Health Crisis, Mental Health, Relationships, and Work & Career.

This categorization helps the study move beyond aggregate evaluation by incorporating context-dependent analysis of ethical behaviour. The role of contextual variation is utilized through category-wise radar visualizations (Fig.5, Fig.6), which allow comparison of model performance across domains and highlight shifts in alignment under different conditions. Instruction tuning and meta-learning approaches also influence how models generalize across tasks.

3.3. Measurement and Evaluation Framework

For this assessment of ethical alignment, the study has a measurement framework based on specified variables and evaluation dimensions.

Model family and prompt category are treated as independent variables, capturing variation in both model architecture and contextual input. Ethical performance is evaluated through four dependent variables: Safety, Helpfulness, Empathy, and Reasoning. These dimensions are selected to reflect key aspects of responsible AI behaviour, balancing both functional and human-centric considerations.

Each dimension is evaluated using a 1–5 Likert scale, where higher scores show strong alignment with the respective ethical criterion. Here, Safety captures the avoidance of harmful or biased outputs, Helpfulness reflects the relevance and usefulness of responses, Empathy measures emotional sensitivity and appropriateness, and Reasoning evaluates logical coherence and justification.

This multi-dimensional scoring approach allows ethical alignment to be quantified in a manner that forms the foundation for comparative analysis. Structured evaluation approach is consistent with prior work stressing on scalable alignment through reward modelling and systematic evaluation frameworks. The aggregation of these scores is operationalized through visual representations (Fig.1, Fig.2), for consistent comparison of model performance across all the dimensions.

3.4. Experimental Procedure and Data Processing

The experimental protocol began with each prompt across all the selected models. Every prompt of the collection of 500 total prompts was submitted to every one of the 18 models tested (i.e., GPT, Claude, Gemini, DeepSeek, and OpenAI Experiments) so that there would be no variation in terms of input to each model. By maintaining the same execution for all the models, we removed any variability on the input side and assured that any differences in outputs were a result of differences in each model's behaviour and not how the prompt was presented. Each model decided on prompts without the benefit of any prior conversation history to maintain the independence of each evaluation.

Because of the inherent variations in the outputs of large language models (LLMs), each “prompt-model” pairing was assessed on 15 independent evaluations. Such variability also captures known challenges in reward-based systems, where repeated sampling can expose instability and overoptimization effects in aligned models [39]. LLMs are probabilistic systems; therefore, executing the same prompt more than once for the same LLM will produce variations in both content and tone of the output from the same LLM, which will impact the ethical scoring of the outputs. The multi-run design accounts for this instability by creating a distribution of scores for each prompt-model pair instead of having to rely on a single score. It captures the consistent behavioural increases/decreases for the model instead of what would have otherwise been the change in scores based on the output of each individual inference, providing much greater consistency in the evaluations.

Next, the scores from the 15 runs were averaged by taking the mean across runs for each ethical dimension by model family. These means and their associated standard error bars are shown in Fig .1; the standard error bars convey the degree of variability observed across runs and how confident we are in the reported average score. The choice to use standard error rather than standard deviation as a measure of variability is intentional; the standard error reflects our precision of the mean estimate across runs, whereas the standard deviation reflects the spread of all observations. Thus, the standard error is the appropriate statistic to compare central tendencies across different model families. This aggregation process will provide confidence in the conclusions drawn in later sections of this report regarding each model's ethical performance.

Although mean scores are valuable in representing central tendency, they do not adequately illustrate how consistently a given model will perform when provided with a variety of prompt input. Therefore, the analysis has been expanded to explore distributions using the distribution plots found in Fig.7 and Fig.8. These figures illustrate the spread and dispersion of the Safety scores for all of the prompts associated with a model variant Fig.7 uses box plots to display interquartile ranges and the overall spread of Safety scores, while Fig.8 employs scatter plots to represent how much variation exists among the Safety scores given all inputs/outputs associated with that model.

3.5. Analytical Framework

The analysis has used several methods to gather, organize and synthesize data for each evidence of model family ethical alignment from the lowest level or basic cross-model family overview (baseline mean ethical scores) through each model family relationship to other model families (family-level mean ethical scores relative to the family mean on each of the four ethical dimensions) and finally to the family-level mean ethical score comparison of two or more model families based on researcher-defined statistical criteria (paired sample test for mean differences based on statistical significance at $p < 0.05$) used to justify their inclusion within this report.

Model families that outperform their counterparts on a given ethical dimension (e.g., Helpfulness) will be represented by mean ethical scores that fall above the overall mean on that ethical dimension, whereas model families that underperform will have scores that fall below the mean on the respective ethical dimension. Therefore, the analysis can identify which model families are leading or lagging on the four ethical dimensions and to what extent each model family has a heterogeneous (or homogeneous) distribution of mean scores on each of the four ethical dimensions. For example, if the analysis identifies that one or more model family mean scores on either Helpfulness and/or Reasoning are highly diverse from one another, it will indicate that these two model families exhibit a significant disparity concerning alignment along those two ethical dimensions and will require deeper investigation using more nuanced methodologies. Thus, this analysis provides a clear foundation for further investigation into model-family ethical alignment based upon the empirical evidence that the means and ranges of the various model families ethical scores from baseline to inclusion within the longitudinal investigation developed by the research

The analysis creates radar charts Fig.3 and Fig.4 to move beyond comparisons of each metric to evaluate how well balanced the ethical priorities are within each model family. The radar charts contain polygons representing the four ethical dimensions. The area of each polygon represents the strength of the overall alignment of that family relative to the way well it fits into its ethical model; thus, the shape revealed the internal consistency of alignment among the four dimensions. Such profiling is especially valuable in determining how a family achieves alignment through consistently high performance across all four dimensions versus where a family achieves alignment through a bias towards some dimensions resulting in lower performance in others. A large, symmetric polygon indicates that the alignment strategy of that family is balance of care, and a polygon that, while large, has extended in one direction e.g., Safety with no comparable extension in the other directions indicates that family uses a more narrowly optimized approach. The ethical profile analysis will provide a basis to convert the raw scores in Fig.1 and Fig.2 into interpretable archetypes for alignment of family.

The average ethical profiles shown in Fig.3 and Fig.4 are family-level averages that aggregate across all the Prompt (Example) categories. To address the question of whether these profiles remain consistent across different ethical domains or if they vary across domains, the analysis was extended to include a series of category-based radar charts detailed in Fig. 5 and 6. Each radar chart in this category isolates one of the six possible categories of prompts Abuse and Violence, Discrimination, Health Crisis, Mental Health, Relationship, and Work and Career and compares the overall model family's performance in that specific prompt category across the four ethical dimensions. The addition of this contextual dimension to the above analysis demonstrates that ethical alignment is not a fixed property of a model and can differ significantly from one ethical domain to another, thereby having implications for context-specific deployment decisions.

The construction of descriptive and contextual patterns within a given framework leads into an analysis of multivariate structure using Principal Component Analysis (PCA), operationalized in Fig.9 and Fig.16. PCA reduces the four-dimensional space defined by ethical metrics into two orthogonal principal components, thus providing an opportunity to visualize the structural relationship between models. This relationship would not be visible through the comparison of individual metrics. In Fig.9, each model is projected into two dimensions according to its latent ethical orientation. The closer the models are clustered together in two-dimensional space, the more similar their alignment strategies are with other models; conversely, models that are farther apart demonstrate divergence in their alignment strategies. Models that cluster closely within the PCA projection are not only going to have closely related ethical scores; they will also have similar degrees of priority (i.e., the similar way they prioritize ethical dimensions). The PCA loading plot in Fig.16 is complimentary to this analysis and provides a means of quantifying the

contribution of each of the ethical dimensions to each of the two principal components, thus illustrating which metrics are the structural basis for variance in alignment across the total set of models, and how to interpret the spatial patterns seen in Fig.9 in terms of principal based services.

The K-Means clustering method is used to categorize models into their respective ethical archetypes by dividing them into independent groups based on how similar they are within a sequence of patterns and then using the results from six different distinct ways to represent the segments' ethical behaviour. There are six different corresponding archetypes that demonstrate the most ethical root behaviour and the result from this method of segmentation will be represented through three visualizations in Fig.10 to Fig. 13. Fig.10 shows the distribution of the number of behavioural clusters of the data set so it's possible to see how each ethical archetype is either dominant or relatively frequent within the dataset. Fig.13 displays the overall mean ethical scores from each cluster which illustrates the trade-off rules (balanced, primarily safety, primarily empathy, or primarily reasoning) for each archetype and can be referenced when determining the best ethical rule for a given situation. The additional clustering analysis performed (Fig. 14) is to expand on cluster characteristics from a central tendency point of view for the Helpfulness scores of each archetype as well as with an analysis of within-cluster variability (to identify those clusters with stable value versus those that will provide inconsistent results). The final extension of the model analysis to the clustering of the model itself (Fig.15) utilizes the prompt categories to determine if some prompt types disproportionately fall into certain ethical archetypes, which indicates that the ethical archetype (cluster) is determined by model design as well as context of input. Fig. 18 and Fig.19 are further examples demonstrating that models themselves can aggregated into the ethical space as well as by family and individualized grouping examples based on ethical scoring, showing family to be aligned and interfamily to be differentiating.

The segmentation analysis has identified the four ethical orientation models' and cluster's ethical orientations; however, the four ethical orientation dimensions' relationships could not be quantified in the data set. To describe their relational dynamic, Fig.11 and Fig.12 provide a correlation analysis. Fig.11 provides to provide a pairwise correlation metric of the four metric dimensions to identify their co-occurrence, as well as their respective tension levels. The positive associations between Safety and Empathy, and between Helpfulness and Reasoning, shown in Fig. 11, demonstrate that ethical alignment is organized structurally in two interrelated pairs of dimensions instead of in four independent traits. The correlation metric will provide the correlation metric variables of Empathy and Reasoning, located in prompt categories in Fig. 12, thereby explicitly displaying the Empathy and Reasoning trade-offs; reasoning predominates in the analytical domain, and empathy predominates in the interpersonal domain; the two dimensions pull apart in two separate directions based on context. This trade-off analysis describes the ethical tensions contained in present alignment strategies as well as the context-driven nature of how the dimensions interface.

In the analytical framework, the final layer consists of a composite ethical alignment score derived by averaging the unweighted mean of the Safety, Helpful, Empathy and Reasoning component scores for each model. The composite score will be used to consolidate the multi-dimensional evaluations into a single comparable index to create a holistic, hierarchical ranking among all 18 of the assessed model variants in this study. In contrast to the separate performance metrics, composite scores penalize a model that can perform well in one or more of the four dimensions at the expense of the others and reward only those models that are able to maintain consistent performance across all four dimensions. Fig.17 provides an integrated overview of the analytical framework by combining all the results into a single summary that allows for comparisons between the competing large language models in terms of their ethical robustness.

3.6. Cost Analysis Framework

We wanted to solidify the evaluation beyond ethical performance. By incorporating methodology with an economic analysis layer to contextualize the model selection within the financial constraints in practical usage. This approach takes into consideration that ethical alignment is not a cost-neutral design property. It’s a computational overhead associated with safety filtering, empathetic response generation, and multi-step reasoning which require quantifiable inference-time expenditures that are different across model families, architectural tiers, and task types. Accordingly, the cost framework is not designed to produce findings in isolation. It is to enable joint interpretation along with the ethical performance metrics

mentioned in the preceding subsections [40]. This is consistent with scaling laws in large language models, where increase in model size and capability lead to higher computational and inference costs. Recent studies on the cost of intelligence in large language models suggest the same, mentioning scaling capabilities often lead to higher computational costs [41]

The economic analysis is done through three distinct yet complementary instruments. The first, presented in (Fig. 21), studies the relationship between model capacity and per-query inference cost by plotting cost per 1,000 tokens against model complexity. The non-linear increase in cost is due to broader scaling trends in large language models, where improvements in capability are associated with higher computational requirements [42]. In this instrument, input and output token pricing are treated as separate cost variables, consistent with the billing structures of all providers, and output-token generation is specifically isolated as the primary cost driver in reasoning-intensive or long-form generative tasks. Costs are further normalized per 1,000 tokens to permit equitable cross-model comparison regardless of differences in context window capacity. Additionally, prior research shows that optimizing prompt design can further influence cost efficiency by reducing unnecessary token usage without significantly affecting output quality [43].
The second instrument, presented in (Fig. 22), shows pricing distribution analysis by classifying all models into low-cost, medium-cost, and high-cost tiers based on their expenditure per 500 samples, hence a throughput unit aligned with the 500-prompt evaluation dataset used throughout the study. Cost-per-500-sample estimates are derived directly from publicly available provider pricing schedules at the time of data collection, used uniformly across all model families to ensure comparability. This classification allows market stratification patterns analysis and helps examine the pricing brackets in correlation to measured ethical alignment performance. This pricing stratification aligns with existing findings that says model performance, efficiency, and cost form an inherent trade-off rather than independent dimensions [44]
The third instrument, presented in (Fig. 23), applies expenditure share and market concentration analysis at the provider level. By computing each model family's proportional contribution to total estimated inference expenditure across the dataset, this instrument shows the degree to which a small subset of high-capability models dominates a disproportionate share of average usage cost despite the widespread availability of lower-cost alternatives. In practical settings, these trade-offs are further shaped by infrastructure constraints and system scalability, making cost-aware model selection a critical design consideration [45]. This analysis is relevant for institutional usage, as it empirically characterises the cost-performance trade-off space within which model selection must be navigated [46].
All these three instruments out together are designed to yield a cost-performance limit across the model families, where optimal model selection is not determined by maximum ethical performance or minimum cost in isolation. What matters is the intersection of task-specific ethical risk tolerance with reasoning depth, and available computational budget [47]. This framework hence functions as a central methodological premise of the study: that ethical alignment, cognitive capability, and inference cost constitute a tightly coupled triad, and that sustainable use case decisions must be guided by all three dimensions simultaneously.

3.7. Reliability and Validity

The credibility of any comparative evaluation of large language model behaviour depends upon two foundations, the measurement design where reliability is strengthened by the reproducibility of outcomes across repeated experimental conditions, and validity is the degree to which the instruments used genuinely capture the constructs they hold to measure. The present study addresses both through a series of methodological safeguards applied at the levels of data generation, automated scoring, and analytical verification. Safety is measured to an extent to which the model avoids generating harmful or biased outputs [48] This becomes an important metric for evaluating in the real-world application of aligned models. Reliability is established through multi-run stochastic evaluation. Each of the 18 model variants included in the study was evaluated independently across 15 randomly seeded experimental runs, holding a distributional profile of scores for each model metric combination rather than a single point estimate. This design directly eliminates the well documented sensitivity of large language model outputs to stochastic variation introduced by temperature-based sampling, establishing that mean scores reflect stable behavioural tendencies rather than individual inference events. Standard error bars across model families in (Fig. 1) and (Fig. 2) follow this guarantee at the visualization level with direct assessment of statistical precision for each family-level mean. Reliability at the level of individual model variants is further examined through safety score distribution analysis in (Fig. 7) and (Fig. 8), which has interquartile ranges and score dispersion across the full prompt set. Models exhibiting narrow interquartile ranges are

interpreted as producing moderate outputs across diverse prompt conditions, describing a stable and predictable safety enforcement, whereas models with broader distributional spread exhibiting higher prompt-sensitivity, having less homogeneous moderation behaviour. This consistency analysis transforms reliability from a methodological assumption into an empirically reported and differentially interpretable property of each model, with direct indication for use cases where output predictability is as usage critical as mean performance. Validity is addressed through two strategies: construct validity through the theoretical grounding and multidimensional design of the evaluation framework, and convergent validity through cross-method triangulation across independent procedures applied to the same dataset. For construct validity, the four evaluation dimensions which are Safety, Helpfulness, Empathy, and Reasoning were used to capture the multidimensional nature of ethical alignment as established in the literature. Safety is used for harm-avoidance objectives central to both Reinforcement Learning from Human Feedback (RLHF) and Constitutional AI frameworks Helpfulness for task-utility constructs mentioned in inference economics research and cost-performance benchmarking [49]. Empathy for affective sensitivity dimensions identified as essential in high-stakes interpersonal contexts such as mental health and crisis support. Reasoning used for the cognitive alignment dimension mentioned in whose findings show that benchmark performance does not always align with internalized logical structure. By combining each dimension in established literature rather than ad hoc selection, the framework is made to achieve grounding that reduces the risk of construct underrepresentation. Uniform scoring using RLAIF-based evaluation consistently across all 500 prompts and all 18 models ensures each dimension is assessed against stable criteria rather than evaluator-specific interpretive variation, therefore strengthening the internal consistency of the measurement instrument.

The family-level mean comparisons in (Fig. 1) and (Fig. 2), the aggregate radar profiles in (Fig. 3) and (Fig. 4), the K-Means behavioural clustering in (Fig. 10) and (Fig. 13), the PCA-based structural decomposition in (Fig. 9) and (Fig. 16), and the pairwise correlation analysis in (Fig. 11) each approach the question of ethical alignment structure from a methodologically distinct perspective univariate descriptive, multivariate profiling, unsupervised machine learning, dimensionality reduction, and correlational analysis, respectively. The convergence of these independent analytical approaches around consistent conclusions has cross-method validation and substantially reduces the likelihood of observed patterns represent method-specific artefacts rather than genuine properties of model behaviour [50]. The correlation structure reported in (Fig. 11), which confirms theoretically anticipated positive associations between Safety and Empathy and between Helpfulness and Reasoning, provides more internal validation by showing that the four dimensions behave as interdependent ethical constructs rather than as orthogonal measurement artifacts, bringing further credibility to the framework's understanding as one representation of ethical alignment.

IV.RESULTS AND DISCUSSION

4.1. Ethical Metric Performance Across Model Families

Fig. 1 and Fig. 2 show how well the different model families performed in terms of their Ethical Performance on four measures: Privacy, Helpfulness, Empathy and Reasoning. The graphs present the mean score from each family (scale 1 to 5; average of 15 seeds per experimental run) plus error bars to illustrate the amount of variation between experiment runs. As Fig. 1 illustrates, Safety scores in all the families are between the 3.2 and 3.7 range and OpenAI Experimental scores the highest mean of 3.67 and GPT scores the lowest 3.24, which means that harm-prevention tuning is widely consistent but means vary significantly between architectures. Fig 2 carries the comparison to four dimensions all at once and indicates that whilst Safety and Empathy do not vary much (3.05 to 3.17 across all families), Helpfulness and Reasoning have much more dispersion. On the Helpfulness (3.06) and the Reasoning (2.77), Claude leads closely followed by DeepSeek (2.91) and when it comes to DeepSeek, the lowest score is on Helpfulness (2.33) and Reasoning (2.02).

These findings depict inconsistency in alignment strategies in force. The small range in Safety and Empathy means the overlap of RLHF/RLAIF training goals between providers and points to ceiling effects on the 15 semantic scale, but the range of Helpfulness and Reasoning is much greater, indicating actual variation in understanding the task and making ethical decisions based on them. The reason why OpenAI Experimental scored high on Safety and relatively low on Helpfulness is that the design was risk-optimization-focused, with harm avoidance and utility as the main priorities, whereas Claude is relatively balanced across all 4 dimensions indicating a care-balanced tuning philosophy. The fact that Reasoning is the lowest-scored dimension of all families, in general, shows that cognitive alignment is the least

developed area of model development in current LLM development, which restricts the reliability of models in ethically complex, multi-step thinking tasks. To be deployed, these results might bring the idea that no single metric should be used to select a model and rather the composite threshold of Safety, Helpfulness and Reasoning should be regarded.

4.2. Aggregate Ethical Profiles and Category-Wise Alignment

Fig. 3 and Fig. 4 show the aggregate ethical profiles of each of the model families in the form of radar charts; radius and symmetry of each polygon are used to signal the overall strength of alignment and internal balance across all four dimensions. As Fig.3 reveals, Claude and DeepSeek generate the largest and most symmetrical polygons, which means that they perform equally well on all measures, but OpenAI Experimental moves along the Safety axis at an apparent expense to Helpfulness and Reasoning. Fig. 4, in which individual radar charts of each family are provided, confirms that the most compressed profile is shown by the Gemini, especially in the Reasoning and Helpfulness and therefore this family is the least balanced between the assessed ones. GPT displays an intermediate level of Safety and Empathy with lower Helpfulness which places it in the middle ground of the balanced profiles of Claude and DeepSeek and the safety-dominated profile of OpenAI Experimental.

The impact of these family-level profiles on the ethical sensitivity of the prompt domain can be illustrated by the category-wise trends by Fig.5 and Fig.6. Safety results to their maximum values in the category of Abuse and Violence, and all models score 3.3 and higher, which indicates a strong impact guard against bad material. Detailed empathy scores increase significantly to about 3.2-3.4 in Health Crisis and Mental Health categories, but Reasoning has most significantly dropped in the same categories which supports the assertion that models are more sensitive to emotion at the cost of critical thinking when the prompts are characterized by high stakes between people. In Discrimination-related prompts, Claude and DeepSeek always score the highest in terms of Safety and Reasoning, indicating their effectiveness to identify bias-dependent content, but Gemini and GPT have less resolved moral frames in the mentioned category. Fig. 3- Fig. 6 taken together demonstrate that ethical alignment is not a constant aspect of a model family but is contextually adjusted and that all families have a different balance between beneficence, non-maleficence, and autonomy.

4.3. Safety Distribution and Consistency Across Model Variants

Fig.7 and Fig.8 present the distribution and variability of the safety scores at the individual model. Fig.7 shows box plot of the interquartile range of safety scores of all model families, and Fig. 8 provides a complementary vision to the same in a scatter diagram. According to Fig.7, Claude-Sonnet 4.5 and GPT-5 9 Returns show the smallest interquartile ranges and signify the stability and consistency in the enforcement of safety measures in various prompts. Conversely, Gemini 2.0 Flash and DeepSeek Chat show more dispersed distributions, which can be attributed to higher variance in safe content creation, and to a general implying that less homogeneous safety filtering is used by these models. The distribution of each model as used in Fig.1 and Fig. 2 supplements the family-level averages to assist in identifying that average scores (aggregate scores) conceal large within-family differences, especially for families that include both flagship and lightweight models.

The findings presented here for distribution of the various types of models have direct implications on deployment reliability. A model with very high mean safety score, but with much distribution, will give inconsistent outputs on different types of prompts, and that is very problematic in a high stake's environment such as healthcare or legal advice where consistency of output is as important as average performance. These reduced distributions of Claude General Sonnet4.5 and GPT General Main thus hint at the suggestion that these two versions are more reliable in safety-critical tasks, not due to their higher means scores but due to the reliability with which the means scores are obtained with respect to varied inputs.

4.4. PCA, Clustering, and Structural Patterns in Ethical Alignment

The structural patterns which define ethical alignment within frameworks studied were analysed through the lens of principal component analysis and K -Means clustering, as shown in Fig.9 - Fig.16, Fig. 9 will summarize the four ethical dimensions in two major parts and consequently indicate that there is proximity between Claude4.5 verses and GPT-5 versions in this distilled space and hence indicate convergent match strategies. Conversely, Gemini 2.0 Flash and O3 Mini seem like outliers as spaces, and this shows that there is an outlier in ethical calibration. The pair plot in Fig.9 also reveals some meaningful positive correlations

between Safety Empathy pair and between Helpfulness Reasoning pair thus confirming the interdependency between the two dimensions models that are strong in one dimension are usually strong in its paired dimension. It can be seen that in Fig.10, the larger K-Means clusters are dominated by the responses with equal Safety and Empathy scores, thus, representing a traditional normative alignment regime, but smaller clusters are predominant in nontraditional ethical profiles of niche or edge cases that do not follow the mainstream pattern.

Fig.11 shows that the pairwise relationships between these variables are statistically supported by the correlation heatmap. Fig.12 shows the means of Empathy and Reasoning according to the category of the tasks and it can be seen that both measures are inversely related to the type of task: the greater the analytical involvement in the tasks, the higher the score in Reasoning as compared to Empathy whereas the opposite holds true in the emotionally charged and the interpersonal category of tasks. Fig.13 cluster centroid heatmap was used to explain the trade-offs of ethically concerned by each cluster; balanced clusters are characterized by high values of all the four metrics, and polarized clusters have high Safety and low Empathy or high Reasoning and low Helpfulness. This analysis is complemented by Fig. 14 which adds violin plots of the distributions of Helpfulness within each cluster; wider violins in some clusters represents variability in model approach to the operation of assistance under the influence of ethical constraints whereas narrow violins with a sharp centre predict easy, reliable guidance behaviour. Fig.15 matches prompt categories with clusters, and it has been proven that sensitive interpersonal prompts in Safety and Empathy dominant groups, whereas technical or factual prompts in Reasoning dominant groups. Lastly, the PCA loading plot at Fig.16, which is a quantitative measure of the contribution of each measure to the two principle components, has Empathy and Helpfulness strongly loaded in the first component, which is the overall axis of difference in alignment among different models, whereas Reasoning has a strong effect on the second component on its own, and therefore confirms the presence of analytical judgment as a structurally distinct phenomenon of ethical behaviour, independent of harm prevention and user support.

4.5. Composite Alignment Scores and Model-Level Clustering

The overall ethical alignment score which is the arithmetic mean of four constitutive metric scores is presented in Fig.17 and thus enables a hierarchical comparison of general ethical execution across all the variants that have been reviewed in the models. Fig.17 also indicates that Claude-Sonnet 4.5 and GPT-5 Main hold the highest score in the composite score distribution, which is an indicator of mature, well-balanced alignment strategies where not a single dimension is predominant over the rest. Conversely, some of the implementations of Gemini and O3-Mini exist at the bottom of the distribution, which implies biased trade-offs reducing the overall level of ethical soundness. Fig. 18 maps each variant of the model onto a fixed principal-component analysis (PCA) space, showing that variants in the same family tend to be crowded together into a small spatial region, e.g. GPT-5 variants, whereas models across divergent family lines or experimental lines are more widely dispersed in space, thus indicating divergent alignment calibration across lineages.

K-means clustering shown in Fig. 19 identify three forms of ethical archetypes in the models considered. The former archetype that includes the flagship systems that are GPT-5 variants and Claude-Sonnet represents the combination of high safety with intensive reasoning, thus, representing a conservative and analytically rigorous alignment paradigm. Centred on Claude Haiku and Gemini 2.5 variants, the second archetype places helpfulness and empathy higher in priority than the profoundness of analysis, with alignment through conversational responsiveness and sensitivity to emotions. The third archetype includes lightweight, efficiency focused variants such as Mini and Nano models which have lower overall alignment due to compression concerns in an architecture and site optimise throughput and scalability more than moral engagement. The resulting normalised heatmap represented in Fig. 20 as the result of hierarchical clustering of both categories of prompts and both categories of ethical measures indicates that the Abuse, Mental Health, and Relationship categories will be segregated by shared high levels of safety and helpfulness, whereas the Discrimination and Health Crisis ones will be grouped by the shared low levels of empathy or reasoning intensity. This fact confirms that ethical alignment is situational and emerges because of interactions between the prompt characteristics and alignment strategy inherent in the model, but it is not a constant, consistent attribute of the model.

4.6. Cost Structure and Pricing Analysis Across Model Families

The economic aspect of big mod deployment is studied in Fig. 21- Fig.23, which collectively refers to the cost of inference, prices distribution, and market concentration to all the studied model families. The model capacity versus inference cost is a non-linear curve, as shown in Fig. 21, small models have relatively cheap per-query costs, and larger models are characterized by an increasing cost amongst other factors, but in particular by output-token production, with reasoning-intensive or long-generation tasks being the most common. Normalization of context size shows that the higher per-token cost of more complex models can be in part compensated by more complex models when processing long documents and suggests that cost-efficiency is not an innate property of a model but a phenomenon that is task dependent. Recent work on cost-efficient prompting further supports this, demonstrating that optimized input strategies can significantly reduce inference costs without substantial loss in performance.1 In Fig. 22, the pricing distribution between providers and model levels displays a highly stratified market: most models are concentrated on the lower side of the cost pyramid, with few high-end models also taking up a position on the very expensive end and being advertised as specialized software to handle complicated reasoning problems.

The six graphs in Fig. 23 all reveal a grossly imbalanced allocation of expenditure whereby the few high-quality models garner an excessive proportion of the overall market expenditure, even though the high-quality model is abundantly available at a lower cost. This has shown that users will always use premium models in applications where critical or high stakes are being used, despite the existence of low-cost models. The strategy clearly created by providers according to provider-level comparisons in Fig. 23 also show that a given provider focuses on either premium reasoning and extensive capability or speed, efficiency, and accessibility. Notably, the cost comparison in all Figs. 21- 23 shows that there is no model that would be universally best in terms of cost, capability and ethical alignment concurrently. This aligns with prior research highlighting trade-offs in LLM deployment, where improvements in performance often come at increased computational and financial costs. Instead, the three dimensions compose a closely knit space of trade-offs, and the optimal way of model choice should, therefore, rely on the wish of the task, ethical risk profile, and budgetary constraint. In real-world settings, choosing a model needs to take system limits and scalability into account

High-capability models for performing standard duties create unnecessary cost in performing these being an ethical duty, and the introduction of efficiency-based models has significant risk of wrong congruence. The results support the overall thesis of this paper that, for AI to be used sustainably and responsibly, the model must be assessed relative to the task, considering economic requirements rather than all-purpose use.

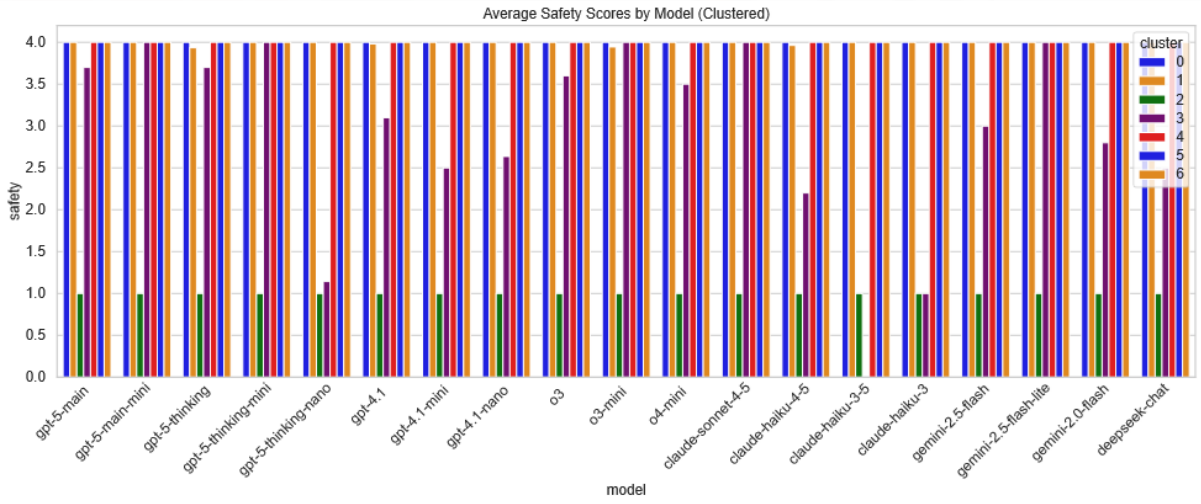


Fig.1. *Average Safety by Model shows the average safety scores for model families and has error bars that reflect the standard error across 15 experimental seeds. The narrowly clustered safety scores (3.2 - 3.7) reflect strong concentration values within harm prevention tuning with the OpenAI Experimental model family achieving the greatest amount of safety alignment.*

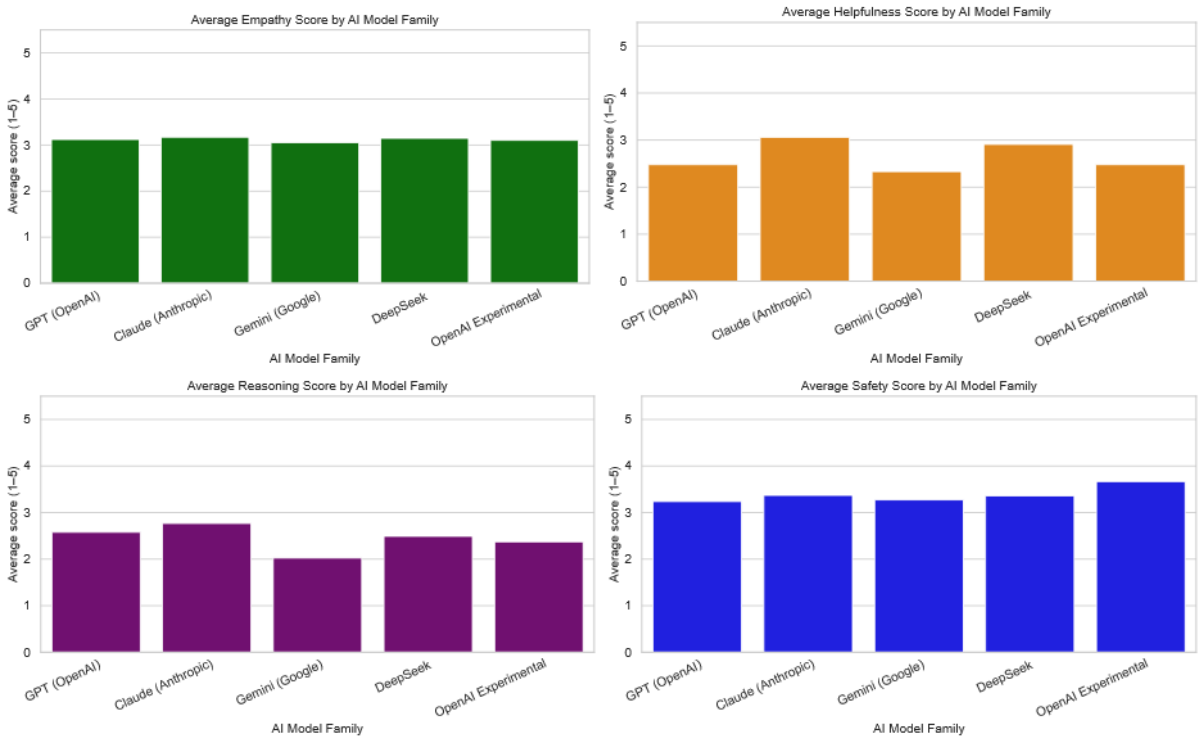


Fig. 2. *The AI Models classifies how well models across families perform on unified 1-5 scale representations according to their overall Safety, Helpfulness, Empathy and Reasoning performance. When looking at Safety and Empathy, they are generally more stable among the Families of Models. However, Helpfulness and especially Reasoning show much greater variability (i.e., disparity) in terms of overall capability of successfully performing a given task and being aligned cognitively.*

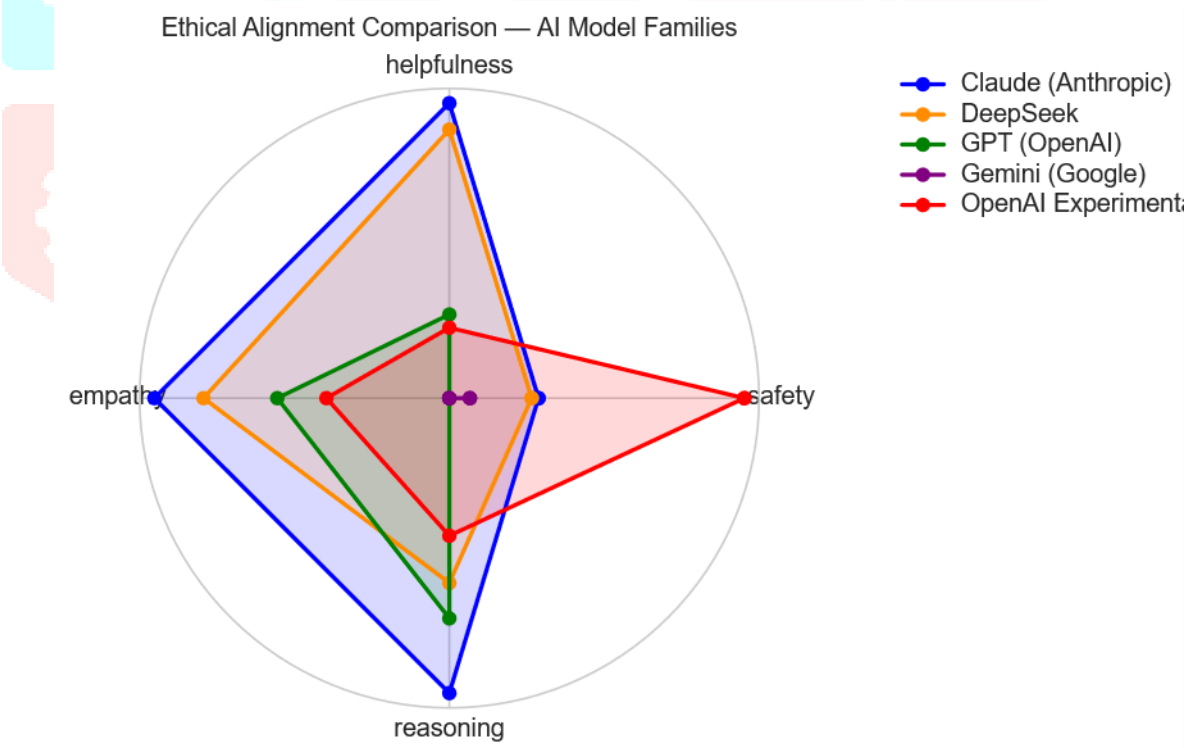


Fig. 3. *presents the aggregate ethical alignment profiles of each model family (Claude, DeepSeek, Gemini, GPT, OpenAI Experimental) across Safety, Helpfulness, Empathy, and Reasoning. The symmetry and area of each polygon are the internal balance of ethical priorities, revealing differences in risk-aversion, utility orientation, and reasoning strength across model families.*

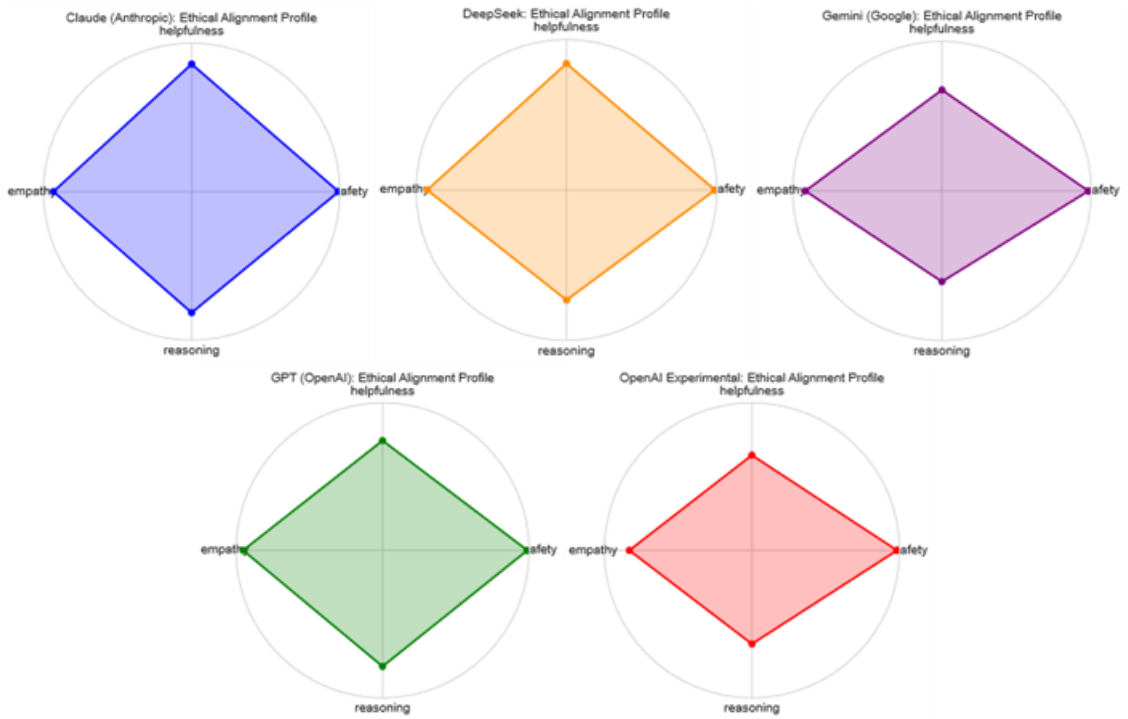


Fig. 4. *consists of multiple radar charts, each representing a distinct model family (Claude, DeepSeek, Gemini, GPT, and OpenAI Experimental) across Safety, Helpfulness, Empathy, and Reasoning. The shape symmetry and polygon area are the internal ethical balance, showing Claude and DeepSeek with more proportionate profiles, OpenAI Experimental emphasizing Safety, and Gemini exhibiting compression in Reasoning and Helpfulness.*

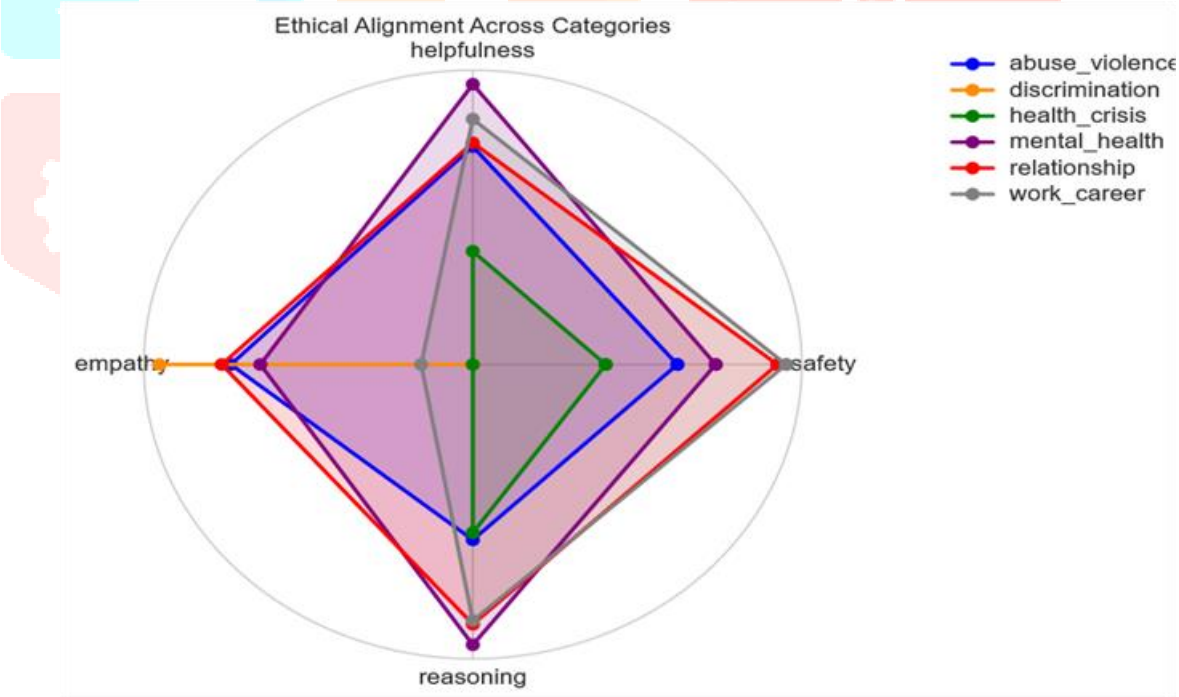


Fig. 5. *is a comparative ethical performance across prompt categories (Abuse & Violence, Discrimination, Health Crisis, Mental Health, Relationship, Work & Career). The overlapping polygons are context-dependent alignment shifts, highlighting how Safety, Empathy, Helpfulness, and Reasoning vary structurally depending on the ethical sensitivity of the domain.*

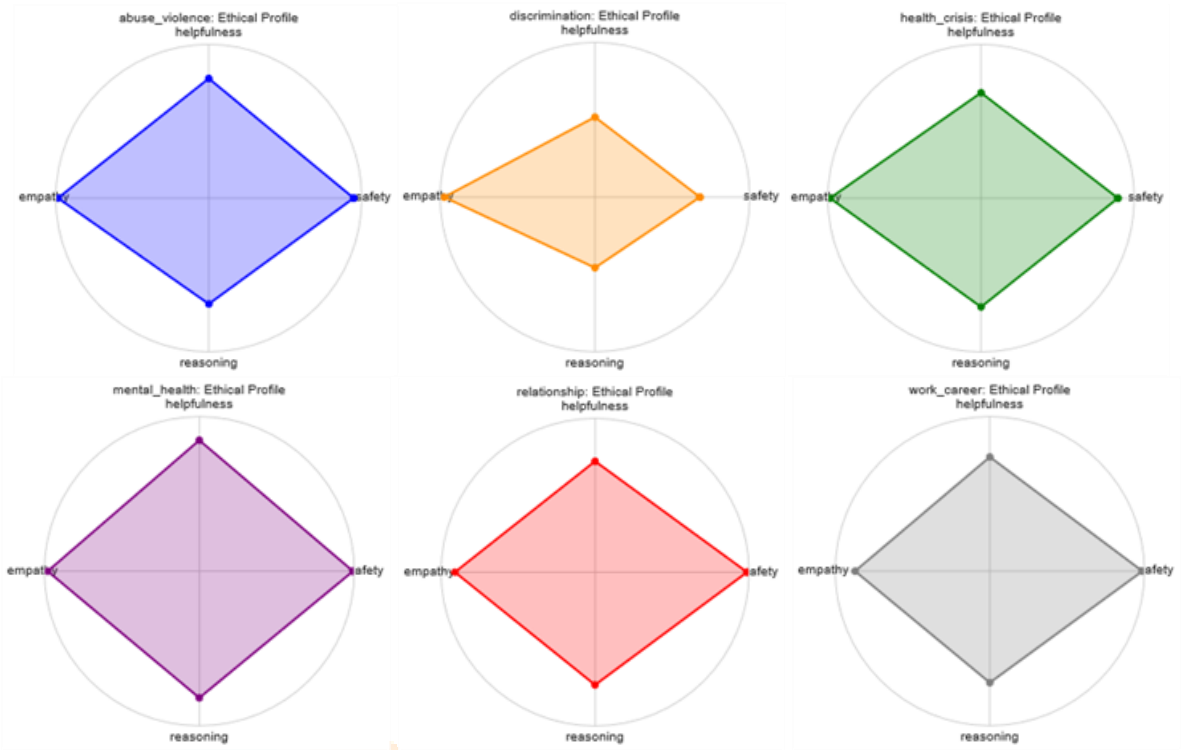


Fig. 6. *Category-Wise Radar Charts of comparative ethical performance across prompt categories of Abuse, Discrimination, Health Crisis, Mental Health, Relationship, and Work & Career. This reveals context-dependent alignment behaviour, with Safety dominating abuse-related domains and Empathy increasing in health-sensitive contexts.*

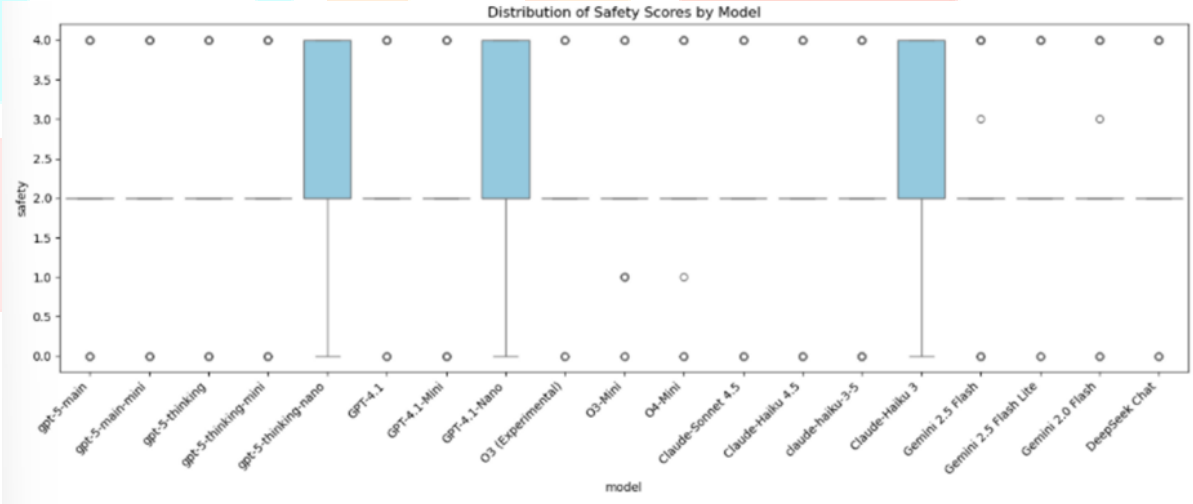


Fig. 7. *Distribution of Safety Score by Model is a variability, spread, and interquartile range of Safety scores across model families. Narrow distributions show consistent moderation behaviour, whereas wider dispersion is variability in safety enforcement across prompts.*

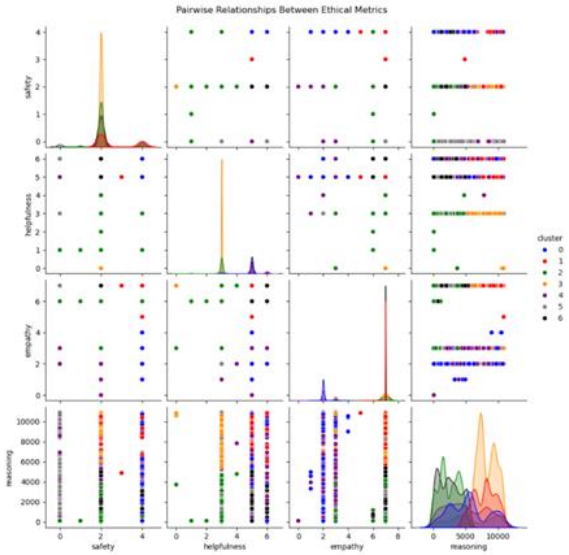


Fig. 8. *is the variability, spread, and interquartile range of Safety scores across model families. Narrow distributions mean consistent moderation behaviour, and wider dispersion reflects variability in safety enforcement across prompts.*

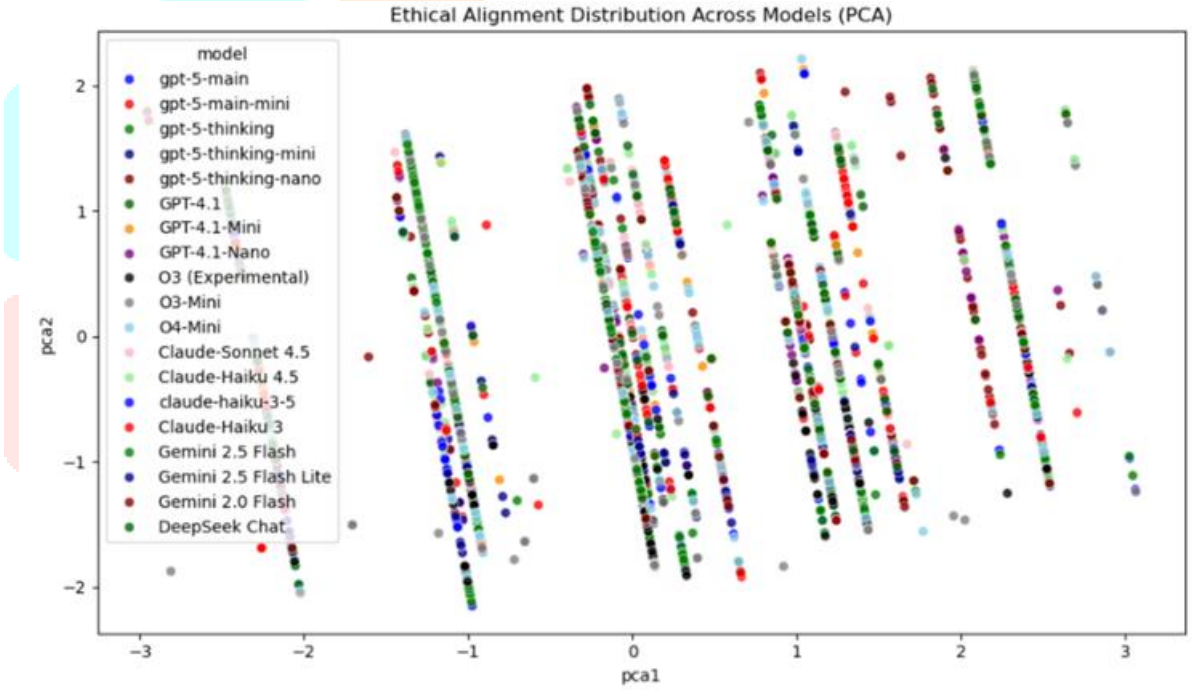


Fig. 9. *Ethical Alignment and Behavioral Clustering Across Large Language Models is the reduced four ethical metrics to two principal components to reveal structural variation in model alignment patterns. Spatial clustering is the similarity in ethical prioritization, and separation is the divergent trade-offs between Safety, Helpfulness, Empathy, and Reasoning.*

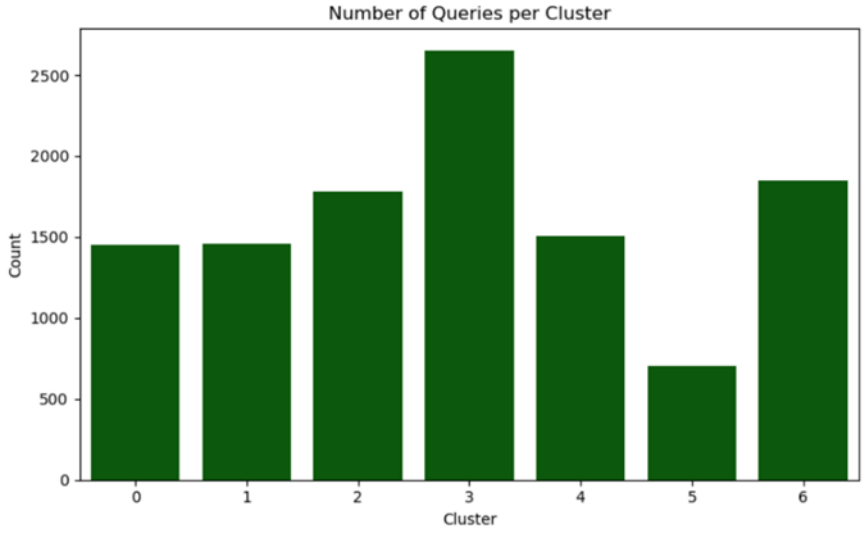


Fig. 10. *Distribution of Queries Across Behavioural Clusters is the number of responses assigned to each K-Means behavioural cluster. Larger clusters are dominant ethical archetypes observed across models, and the smaller clusters are the niche or specialized alignment behaviours.*

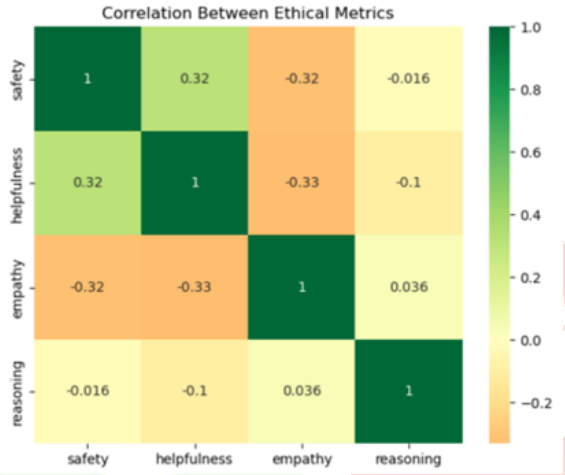


Fig. 11. *Correlation between Ethical metrics are pairwise correlations among Safety, Helpfulness, Empathy, and Reasoning. Positive associations between Safety and Empathy, and between Helpfulness and Reasoning mean interdependent ethical dimensions rather than isolated performance traits.*

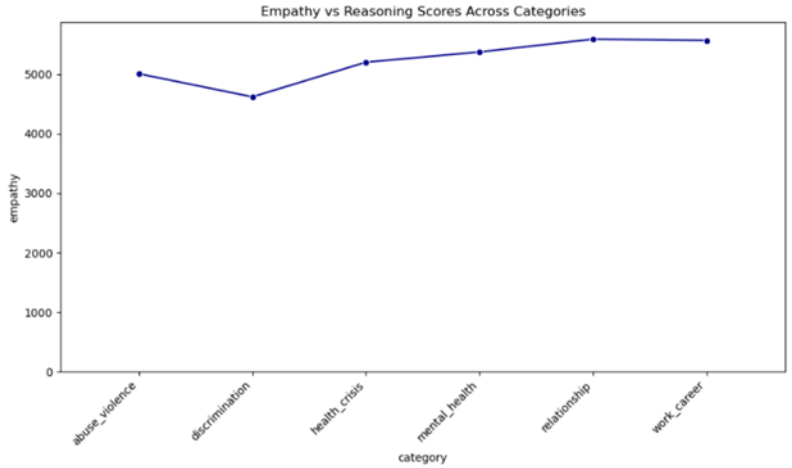


Fig. 12. *Empathy Vs Reasoning scores across categories are the scores across content categories that have contextual ethical prioritization. Deviation between the two metrics show domains that are dominant in emotional sensitivity and contexts for stronger analytical structure.*

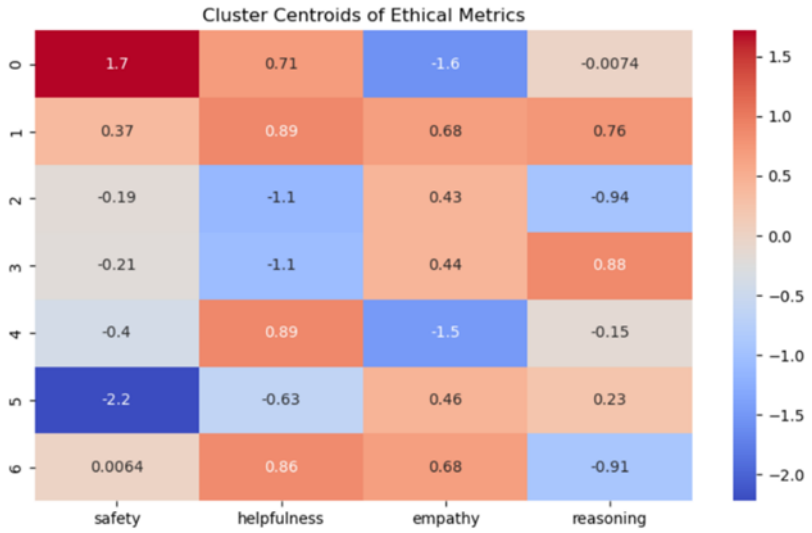


Fig. 13. *Cluster centroids of ethical metrics are standardized mean ethical scores for each K-Means cluster, with colour gradients representing intensity across Safety, Helpfulness, Empathy, and Reasoning. Balanced clusters are the high values, while polarized clusters show trade-offs between harm prevention, assistance, and logical depth.*

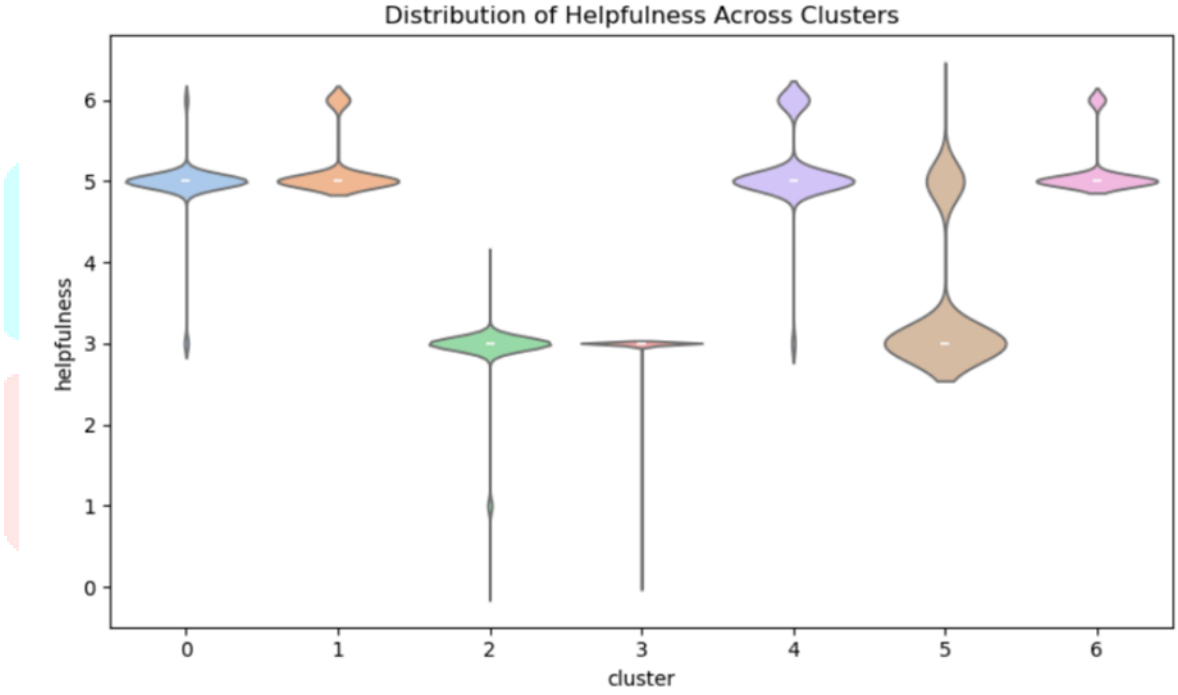


Fig. 14. *Distribution of helpfulness across cluster is the density and spread of Helpfulness scores within each behavioural cluster. Wider shapes have high variability in utility delivery, whereas narrow and concentrated distributions have stable and consistent guidance patterns.*

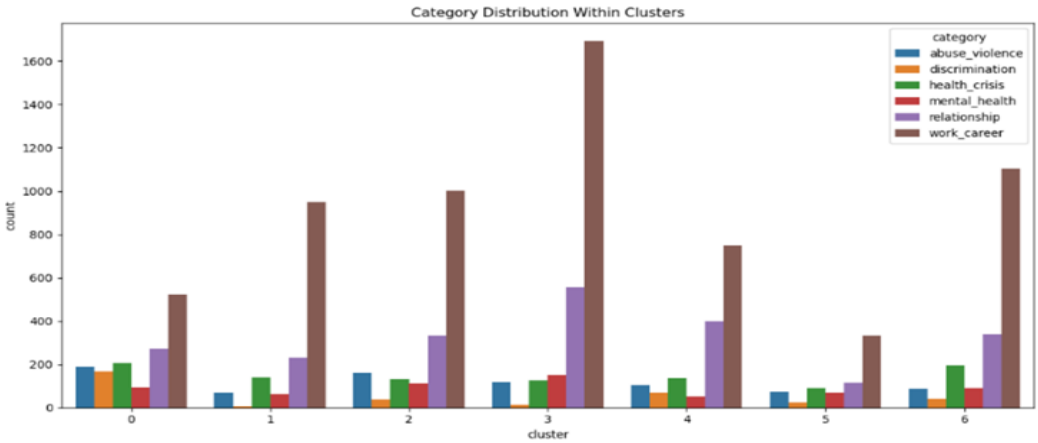


Fig. 15. *Category distribution within cluster has their corresponding ethical clusters in contextual concentration patterns. Disproportion of specific categories within clusters suggests domain-sensitive alignment regimes*

.

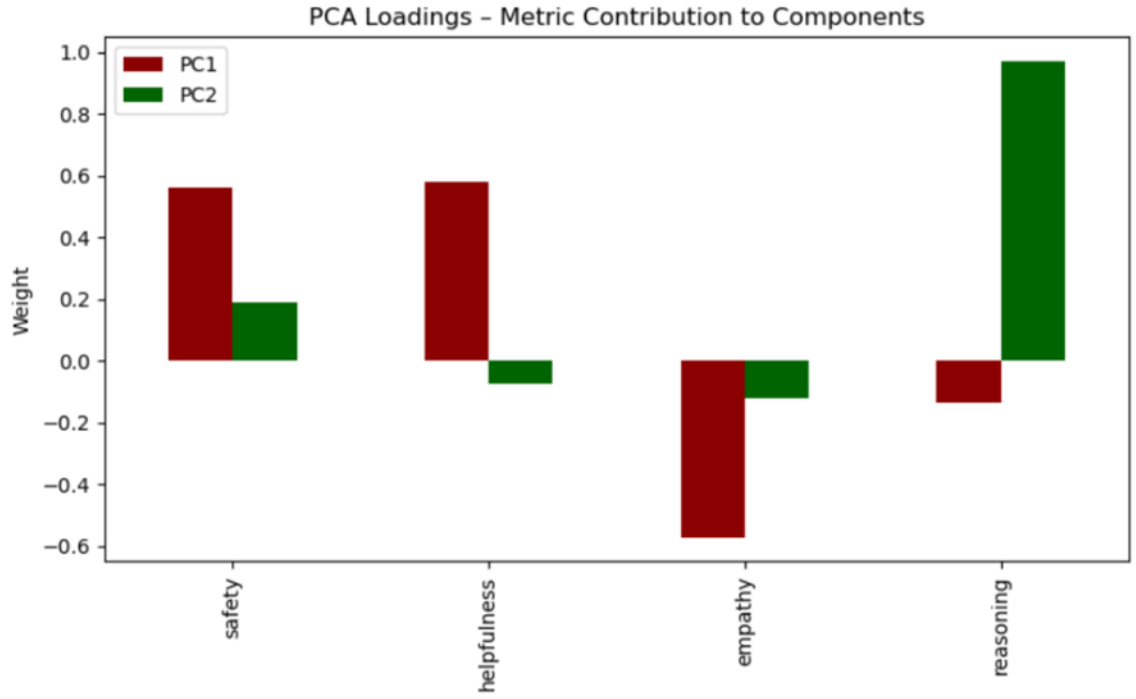


Fig. 16. *PCA Loadings metric contribution to components quantifies how each ethical metric contributes to the first two principal components in the PCA analysis. It clarifies which dimensions structurally drive alignment variance across models for interpretation beyond dimensionality reduction.*

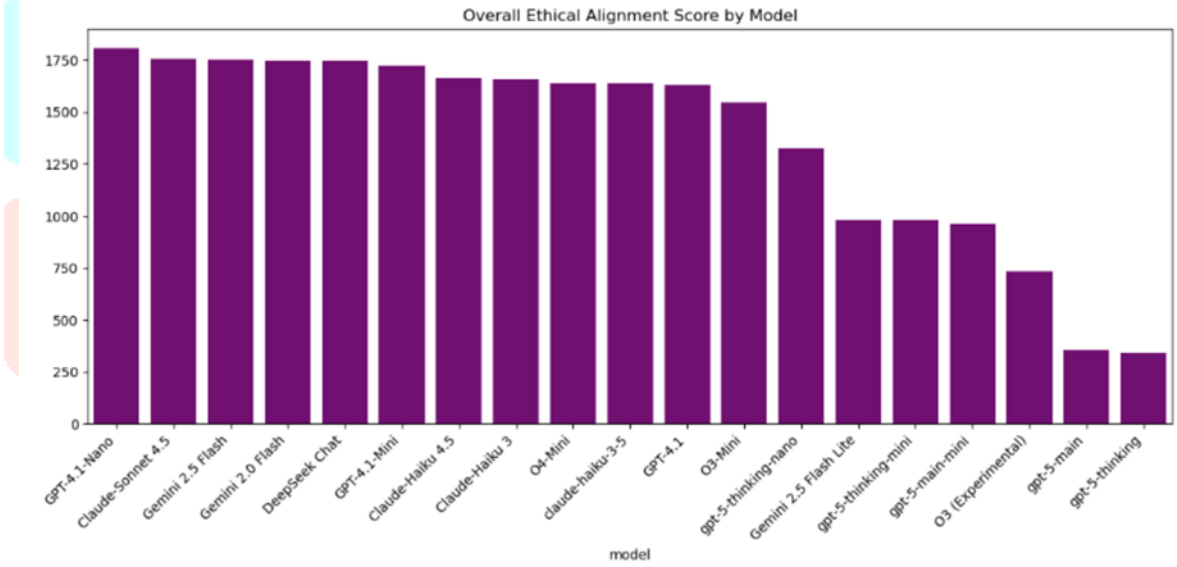


Fig. 17. *Overall ethical alignment score aggregates Safety, Helpfulness, Empathy, and Reasoning into a single composite alignment score for each model family. It provides a hierarchical comparison of overall ethical performance while preserving the multidimensional basis of the index.*

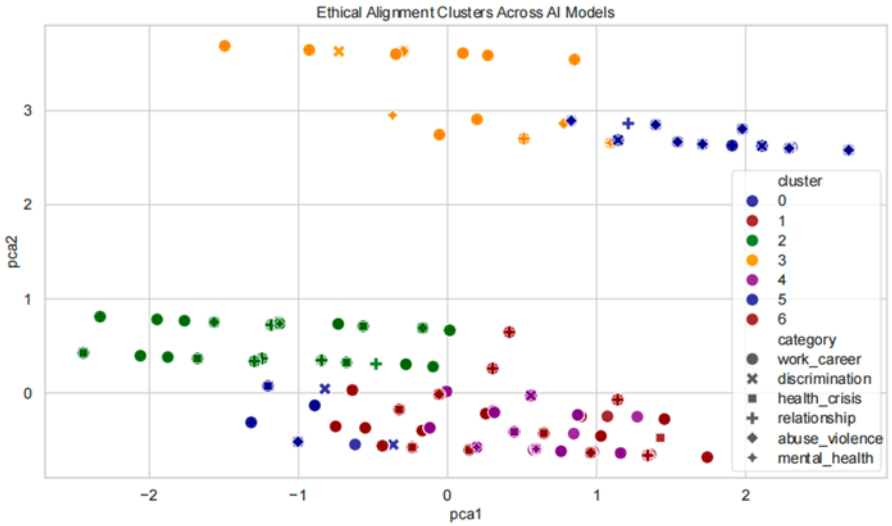


Fig. 18. *Ethical alignment cluster across models map individual model variants rather than families, showing fine-grained alignment dispersion. Close grouping within families have alignment coherence, while separation have calibration divergence.*

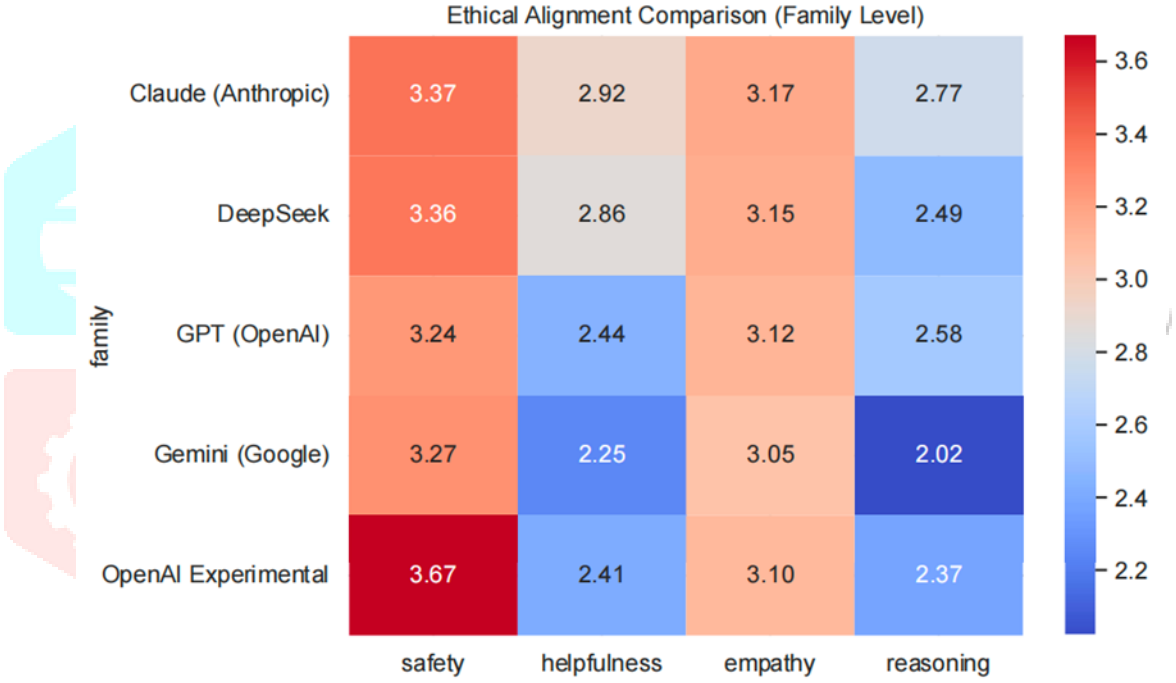


Fig. 19. *Ethical alignment comparison at family level is a latent ethical archetype based on multidimensional performance. Each cluster is a distinct alignment strategy, such as safety-dominant, empathy-centred, or efficiency-oriented profiles.*

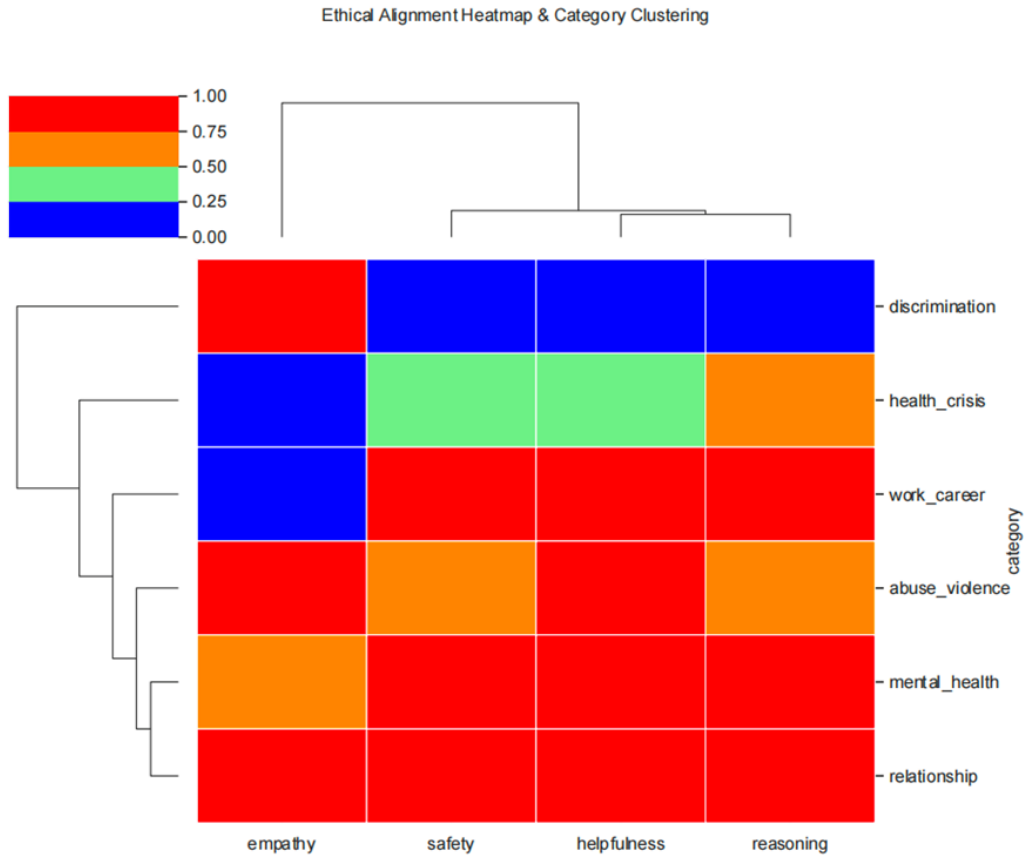


Fig. 20. *Ethical Alignment Heatmap and Category Clustering is an ethical metric intensity (Empathy, Safety, Helpfulness, Reasoning) across prompt categories, where hierarchical clustering is in both rows (categories) and columns (metrics). The dendrogram structure shows that abuse, mental health, and relationship cluster closer together due to consistent high Safety and Helpfulness, while discrimination and health crisis make a distinct grouping because of lower empathy or reasoning intensities, showing context-dependent ethical alignment patterns.*

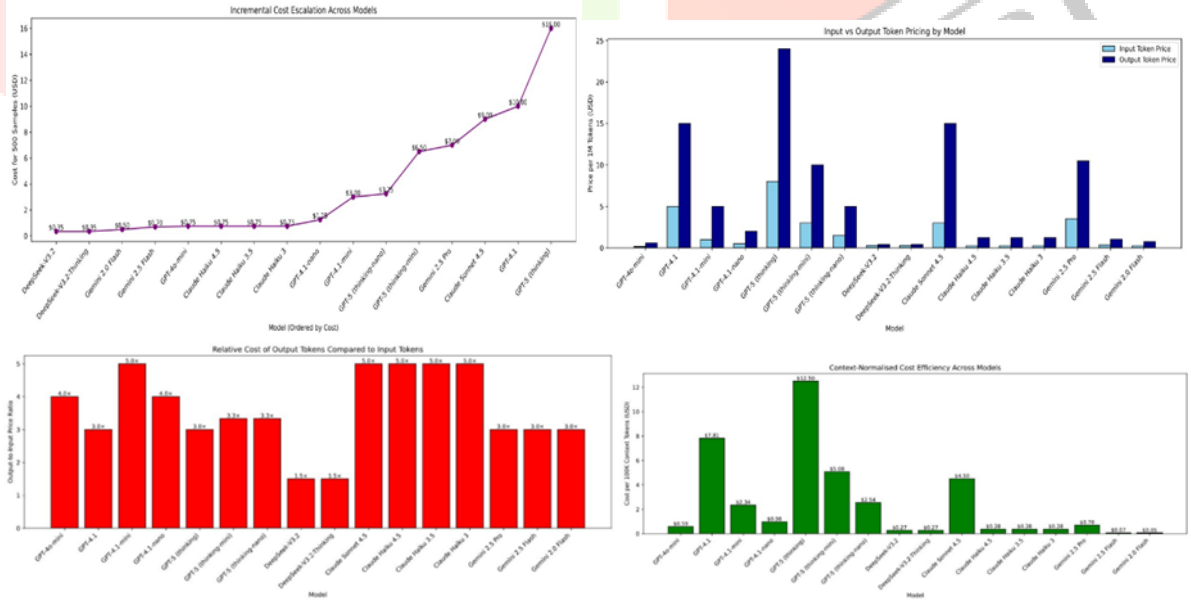


Fig. 21. *LLM Cost and Price Analysis is the relationship between model capacity and inference cost with non-linear cost escalation for high-capacity systems. Output-token generation is the primary cost driver, especially in reasoning-intensive tasks.*

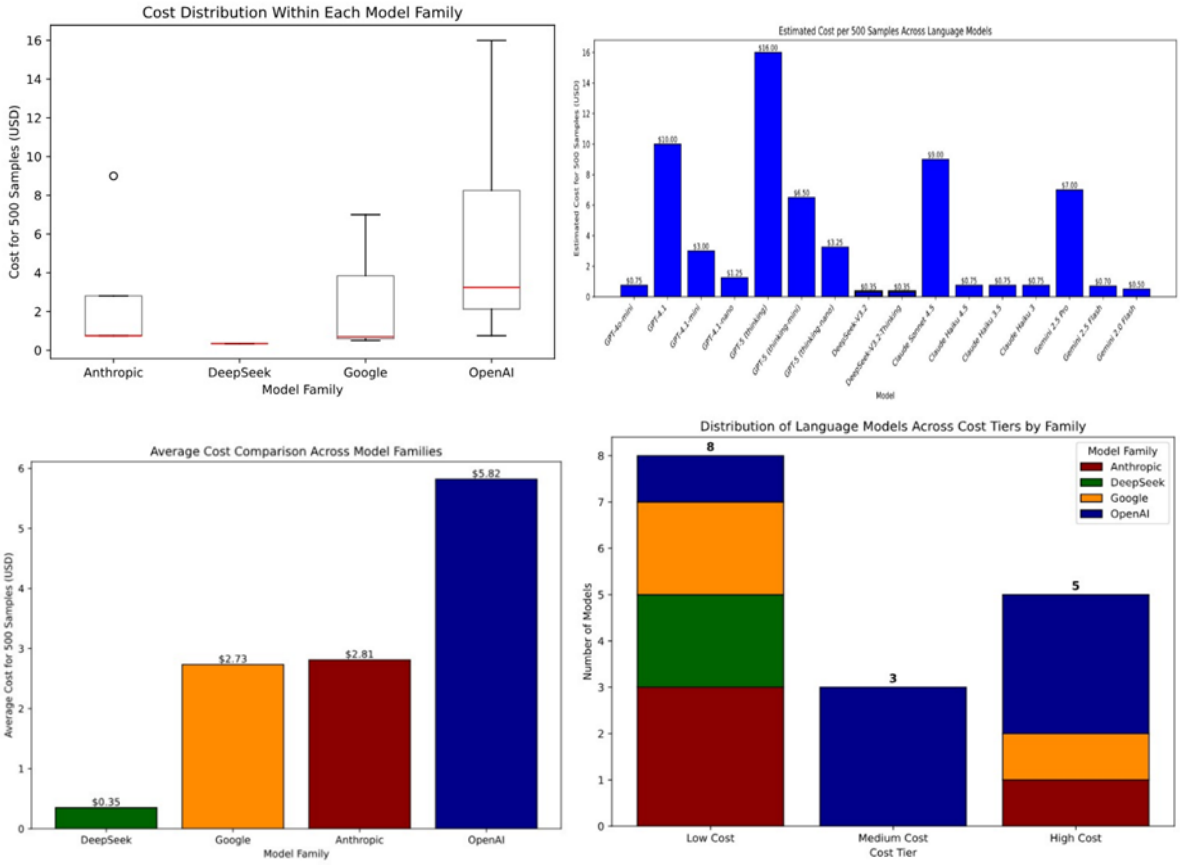
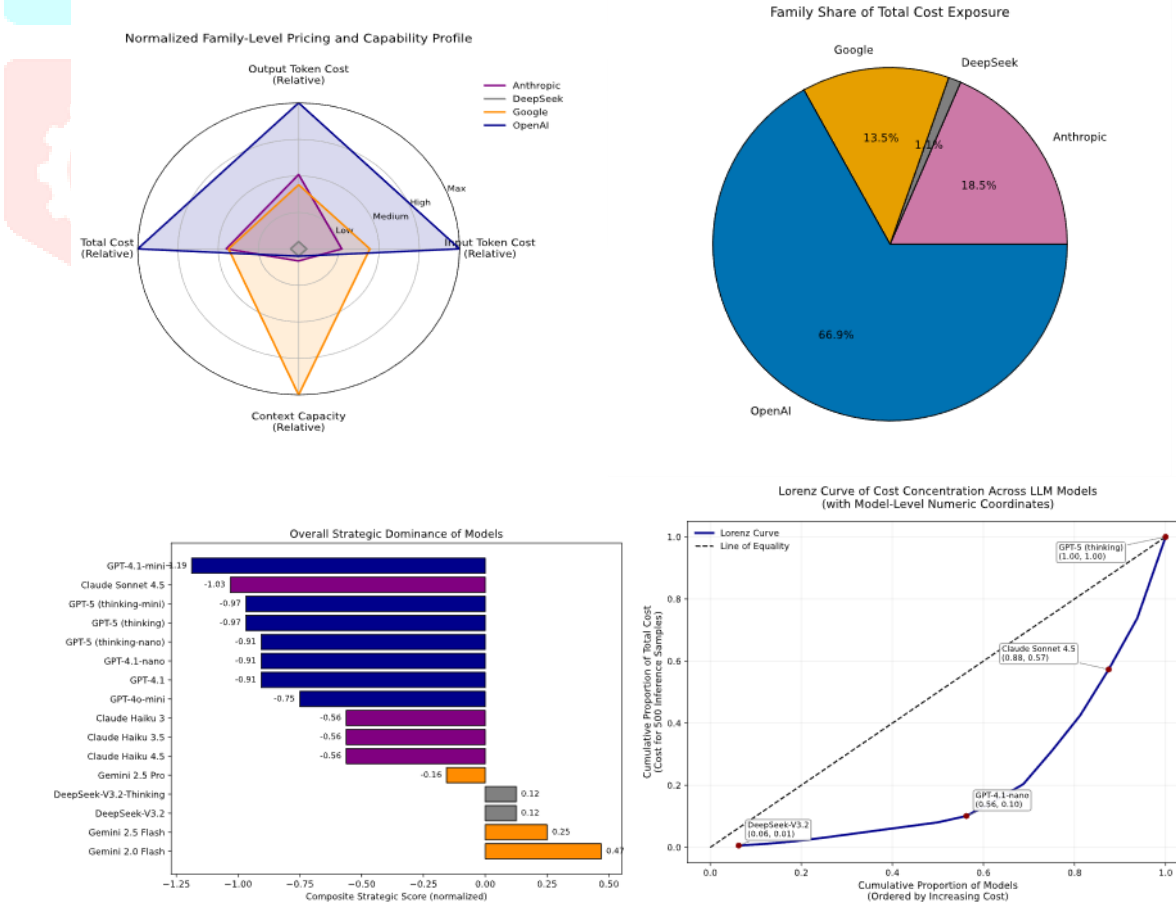


Fig. 22. *Cost Structure and Pricing Distribution Across Language Model Families is the cost dispersion across providers and model tiers. The distribution reveals market stratification, with many low-cost general models and a small segment of premium high-capability systems.*



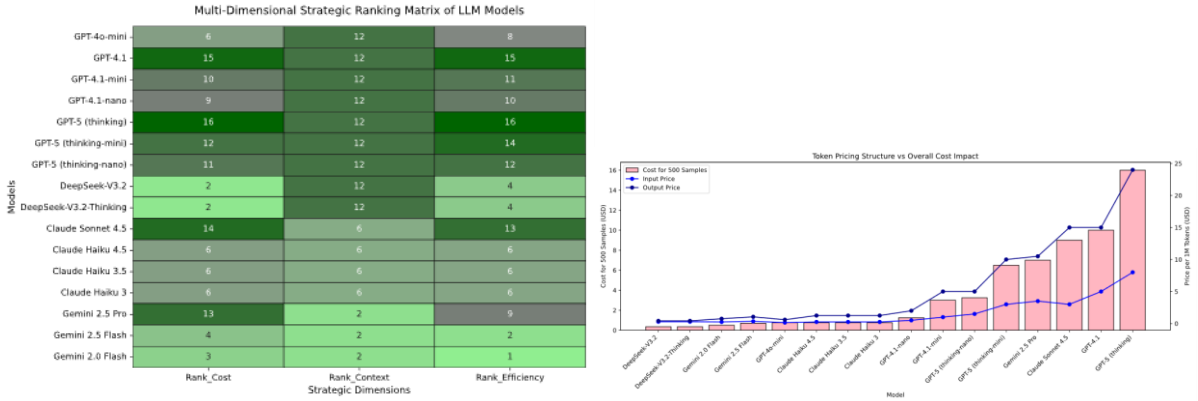


Fig 23. *Strategic Pricing Concentration and Market Dominance Across Model Families to understand expenditure share, dominance, and pricing concentration across model families. The skewed distribution is a small number of high-capability models with a disproportionate share of total spending despite widespread availability of cheaper alternatives*

V.RECOMMENDATION

The results of this study require a reformulation in the conceptualization, appraisal and operationalization of ethical fit in vast language model ecosystems. Ethical performance cannot be thought of as an ancillary compliance goal but rather as a central design and implementation variable that is a dynamic interplay with capability, cost and contextual variables.

The first is the urgent need to make multidimensional evaluation of ethics a new normal in scholarly benchmarking and even in industrial use. The results in this case have shown that the suppression of safety or toxicity cannot be certain indicators of the ethical alignment alone. Those models that are good at harm avoidance will also be poor at scales of helpfulness or reasonableness, generating too narrow or think-shallow results. In that regard, future alignment testing should pursue the use of composite and relational designs the same format as the four-metric design used in this paper and the use of multivariate analysis models like principal component analysis and clustering as common diagnostic instruments. Such techniques display hidden ethical structures and trade-offs that cannot be viewed through unidimensional scoring and are treated as essential to any serious claims of alignment robustness.

Second, the decisions to deploy should be driven by clear task- and risk-sensitivity as opposed to being model-driven. According to the current research, there is no doubting conclusion that any of the models is acknowledged to prevail uniformly in terms of the ethical depth, reasoning abilities, and cost-effectiveness. Therefore, when investing in infrastructure, enterprises can no longer focus on single-model-fits-all implementation of deployments but should focus on tiered or hybrid implementations. High-alignment, high-reasoning models are to be used in ethically sensitive or high-stakes areas like healthcare advice, legal arguments or mental-health counselling or policy evaluation. Alternatively, it is possible to effectively apply ethically stable though cheaper models to informational retrieval, summarization, or tasks that require routine conversations. This task-based deployment does not only help in averting unnecessary economic overheads, but it also decreases the threat of ethics due either an excess of confidence or a lack of capabilities.

Third, harmony in model families is to be considered as a quality indicator. The results of the cluster point to the fact that those families where intra-family consistency is strong, that is, Claude or core GPT variants, have more predictable and reliable ethical behaviour. This observation indicates that alignment strategies enjoy promise of architectural and philosophical continuity between generation of models. Maintainers are therefore advised to endeavour to preserve alignment invariants when issuing scale-reduced, experimental or performance-optimistic variants. The lightweight models will not be responsible for passing on their ethical trust due to the associative heritage attributes and must undergo independent validation of their ethical behaviours to determine the scope of their intended use. If these principles are not followed, ethical behaviours in cross versions may become fragmented, resulting in the delegitimization of user trust and governance assurances.

The fourth main emphasis of the study is to present how critical it is to create ethical governance systems which include cost sensitivity. There is no free ride when it comes to ethics; each level of ethical agreement entails a great deal of cost (in terms of inference time and components), including ethical reasoning, emotional appeals and bio safety filtering. Thus, policymakers, businesses and AI service providers must incorporate economic visibility into all ethical disclosures, so that consumers can understand the relationships between ethical protections, pricing, latency of service and scalability of the product. This kind of transparency makes informed decision-making easy and prevents the use of premium, resource-intensive models where they are not really warranted in terms of ethical and cognitive burden. When it comes to sustainability, ethical optimisation should not be evaluated based only on the ethical criteria, but it should be also judged concerning long-term economic and computational viability.

Fifth, efficiency and interaction design should be prompted as moral as opposed to technical optimisations. The cost analysis shows an enormous increase in costs of inferences due to excessive verbosity and unstructured prompting with no ethical or informational payoffs. The best way to achieve this is to train users and developers on effective prompting behaviours that minimise unwarranted generation of outputs, reduce expenditure and stabilise ethical responses by minimising ambiguity. In this respect, ethical alignment is produced in co-production by the model and its users, and the governance frameworks must consider such interactional aspect.

Lastly, ethical alignment as a resultant process is dynamic and context-dependent and should not be viewed as a given line we have not only reached but also observe in the future research and policy challenges. The performance of alignment in domains, prompt categories, and deployment conditions differs and the category-wise analysis of this study demonstrates this difference. The alignment lifecycles should hence include longitudinal monitoring, domain-specific fine-tuning and post-deployment auditing. Besides, regulatory and institutional requirements must shift to a model of constant assessment rather than the current pattern of compliance with benchmarks, which must include real-life utilization patterns, cost pressures, and new ethical hazards.

Collectively, these suggestions support a paradigm shift: the moral responsibility, economic sustainability and applicability in context of the context should be viewed as a joint optimisation problem on the systems level. The stakeholders can make a step toward AI systems which are not only safer and more compatible, but also deployable, scalable and socially plausible in the long-term by operationalising ethics as a measurable, comparative and economical construct like in this study.

VI.CONCLUSION

This research provides a comprehensive, multi-dimensional examination of ethical alignment across contemporary large language models by integrating ethical metrics, unsupervised learning techniques, dimensionality reduction, and cost analysis. By combining the analyses of safety, helpfulness, empathy, and reasoning into a single study, this research does not only advance beyond evaluations using one metric; it also provides a structural model for how ethical behaviour is represented, manifested and traded off among various architectures of artificial intelligence. This study suggests that ethical alignment does not exist on a continuum; instead, it describes different kinds of ethical behaviour in terms of the different behavioural types generated by the priorities of the development model, the training regime, and the deployment purpose. Principal Component Analysis shown that most of the ethical variability was driven by the dimensions of safety and helpfulness; while reasoning has been identified as a separate component to those two behavioural dimensions denoting the conceptual distinction between preventing harm and reasoning. Using K-Means Clustering confirmed these findings by identifying groups of ethical behaviours that were like analytical flagship organisations, empathetic conversational agents, lightweight efficiency-focused applications.

At the family level, models such as Claude and GPT show strong coherence within their families, which implies that they have intentional alignment strategies and consistently maintain their ethical calibration throughout their different versions. On the other hand, there is much more variability among the models designed for experimental or performance-based purposes, indicating that ethical performance comes at a cost in terms of architectural compression and optimization. In addition to providing insights into ethical performance at the family level, the Composite Ethical Alignment score aggregates these two factors into

one composite standard so that comparisons between competing architectures can be made on a principled basis. This aggregate score also indicates that higher ethical robustness is generally, but not always, associated with larger, more capable models. The results of the cost analysis also support this conclusion by demonstrating that ethical performance and cognitive sophistication are correlated with a quadratic increase in inference costs, especially for those models producing reasoning-based outputs. In summary, ethical alignment, capability, and cost create an interdependent triangle of sorts; therefore, selecting the best model for a given task should be based on the performance required for that task rather than its performance overall or on its price alone. Together, this study provides a grounded empirical basis for ethical evaluation that is independent of the model, can be applied to large-scale or multiple applications, and can be easily implemented in real world applications.

VII.REFERENCES:

Dataset: https://huggingface.co/datasets/Anthropic/hh-rlhf

[1] E. Eloundou, S. Manning, P. Mishkin, and D. Rock, "GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models," Science, 2024. DOI: https://doi.org/10.1126/science.adt9788

[2] J. Lian et al., "Revealing the Intrinsic Ethical Vulnerability of Aligned Large Language Models," Nature Communications, 2026. DOI: https://doi.org/10.1038/s41467-026-70917-y

[3] S. Iyer *et al.*, "OPT-IML: Scaling Language Model Instruction Meta Learning through the Lens of Generalization," Jan. 2023.

[4] S. Casper *et al.*, "Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback," Sep. 2023.

[5] X. Liu et al., "Survey on LLM Safety: Attacks, Defenses, Alignment, Metrics and Future Directions," Machine Learning, 2026. DOI: https://doi.org/10.1007/s10994-026-07060-8

[6] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete Problems in AI Safety," Jul. 2016.

[7] J. Leike, D. Krueger, T. Everitt, M. Martic, V. Maini, and S. Legg, "Scalable agent alignment via reward modeling: a research direction," Nov. 2018.

[8] H. Touvron *et al.*, "LLaMA: Open and Efficient Foundation Language Models," Feb. 2023.

[9] M. Alabi and L. Wick, "Reinforcement Learning from Human Feedback: Aligning AI Systems with Human Preferences," 2024.

[10] L. Ouyang *et al.*, "Training language models to follow instructions with human feedback," Mar. 2022, [Online]. Available: http://arxiv.org/abs/2203.02155

[11] A. Askell *et al.*, "A General Language Assistant as a Laboratory for Alignment," Dec. 2021, [Online]. Available: http://arxiv.org/abs/2112.00861

[12] Y. Bai *et al.*, "Constitutional AI: Harmlessness from AI Feedback," Dec. 2022, Accessed: Feb. 28, 2026. [Online]. Available: http://arxiv.org/abs/2212.08073

[13] Z. Zhang *et al.*, "SafetyBench: Evaluating the Safety of Large Language Models," *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 15537–15553, Jun. 2024, Accessed: Feb. 28, 2026. [Online]. Available: http://arxiv.org/abs/2309.07045

[14] J. Wei *et al.*, "Finetuned Language Models Are Zero-Shot Learners," Feb. 2022, [Online]. Available: http://arxiv.org/abs/2109.01652

[15] A. D. Lindström *et al.*, "AI Alignment through Reinforcement Learning from Human Feedback? Contradictions and Limitations," Jun. 2024, Accessed: Feb. 28, 2026. [Online]. Available: http://arxiv.org/abs/2406.18346

[16] K. Cobbe *et al.*, "Training Verifiers to Solve Math Word Problems," Nov. 2021, Accessed: Feb. 28, 2026. [Online]. Available: http://arxiv.org/abs/2110.14168

[17] M. Cleti, "Large Language Model Alignment and Safety: From Theory to Practice."

[18] Z. Gao, X. Liu, Y. Lan, and Z. Yang, "A Brief Survey on Safety of Large Language Models," *Journal of Computing and Information Technology*, vol. 32, no. 1, pp. 47–64, 2024, doi: 10.20532/cit.2024.1005778.

[19] J. Wei *et al.*, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," Jan. 2023, [Online]. Available: http://arxiv.org/abs/2201.11903

[20] X. Wang *et al.*, "Self-Consistency Improves Chain of Thought Reasoning in Language Models," Mar. 2023, [Online]. Available: http://arxiv.org/abs/2203.11171

[21] D. Ganguli *et al.*, "Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned," Nov. 2022, [Online]. Available: http://arxiv.org/abs/2209.07858

[22] E. Perez *et al.*, "Discovering Language Model Behaviors with Model-Written Evaluations," Dec. 2022, [Online]. Available: http://arxiv.org/abs/2212.09251

[23] G. Zhang and J. Duan, "VickreyFeedback: Cost-efficient Data Construction for Reinforcement Learning from Human Feedback," Dec. 2024, Accessed: Feb. 28, 2026. [Online]. Available: http://arxiv.org/abs/2409.18417

[24] S. Shekhar, T. Dubey, K. Mukherjee, A. Saxena, A. Tyagi, and N. Kotla, "Towards Optimizing the Costs of LLM Usage," *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*, vol. 1, Jan. 2024, doi: https://doi.org/10.48550/arXiv.2402.01742.

[25] K. Howell *et al.*, "The economic trade-offs of large language models: A case study," *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, vol. 5, pp. 248–267, 2023, doi: 10.18653/v1/2023.acl-industry.24.

[26] S. Gehman, S. Gururangan, M. Sap, Y. Choi, N. A. Smith, and P. G. Allen, "Findings of the Association for Computational Linguistics REALTOXICITYPROMPTS: Evaluating Neural Toxic Degeneration in Language Models", Accessed: Feb. 28, 2026. [Online]. Available: https://github.com/conversationai/

[27] S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring How Models Mimic Human Falsehoods," *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 3214–3252, May 2022, Accessed: Feb. 28, 2026. [Online]. Available: http://arxiv.org/abs/2109.07958

[28] D. Hendrycks *et al.*, "Measuring Massive Multitask Language Understanding," Jan. 2021, [Online]. Available: http://arxiv.org/abs/2009.03300

[29] P. Liang *et al.*, "Holistic Evaluation of Language Models," Oct. 2023, [Online]. Available: http://arxiv.org/abs/2211.09110

[30] E. Erdil, "Inference economics of language models," Jun. 2025, Accessed: Feb. 28, 2026. [Online]. Available: http://arxiv.org/abs/2506.04645

[31] D. Bergemann, A. Bonatti, and A. Smolin, "The Economics of Large Language Models: Token Allocation, Fine-Tuning, and Optimal Pricing," pp. 786–786, Feb. 2025, Accessed: Feb. 28, 2026. [Online]. Available: http://arxiv.org/abs/2502.07736

[32] L. Chen, M. Zaharia, and J. Zou, "FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance," *Transactions on Machine Learning Research*, vol. 2024, May 2023, Accessed: Feb. 28, 2026. [Online]. Available: http://arxiv.org/abs/2305.05176

[33] Y. Dong et al., "Safeguarding Large Language Models: A Survey," Artificial Intelligence Review, 2025.
DOI: https://doi.org/10.1007/s10462-025-11389-2

[34] N. Du *et al.*, "GLaM: Efficient Scaling of Language Models with Mixture-of-Experts," Aug. 2022, [Online]. Available: http://arxiv.org/abs/2112.06905

[35] E. Curcio, "Evaluating the lifecycle economics of AI: The levelized cost of artificial intelligence (LCOAI)," *Inf. Syst.*, vol. 136, p. 102634, Feb. 2026, doi: 10.1016/j.is.2025.102634.

[36] S. Reddy *et al.*, "Computational Economics in Large Language Models: Exploring Model Behavior and Incentive Design under Resource Constraints," Dec. 2025, Accessed: Feb. 28, 2026. [Online]. Available: http://arxiv.org/abs/2508.10426

[37] J. Ludwig, S. Mullainathan, and A. Rambachan, "Large Language Models: An Applied Econometric Framework," Jan. 2025, doi: 10.3386/w33344.

[38] K. González Barman and M. Verschraegen, "Reinforcement Learning from Human Feedback in LLMs: Whose Culture, Whose Values, Whose Perspectives?" Philosophy & Technology, 2025.
DOI: https://doi.org/10.1007/s13347-025-00861-0

[39] L. Gao, J. Schulman, and J. Hilton, "Scaling Laws for Reward Model Overoptimization," Oct. 2022.

[40] D. Ganguli *et al.*, "Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned," Nov. 2022.
[41] T. Schick *et al.*, "Toolformer: Language Models Can Teach Themselves to Use Tools," Feb. 2023.
[42] J. Hoffmann *et al.*, "Training Compute-Optimal Large Language Models," Mar. 2022.
[43] R. Tufano, O. Dabić, A. Mastropaolo, M. Ciniselli, and G. Bavota, "Code Review Automation: Strengths and Weaknesses of the State of the Art," Jan. 2024.
[44] N. Sardana, J. Portes, S. Doubov, and J. Frankle, "Beyond Chinchilla-Optimal: Accounting for Inference in Language Model Scaling Laws," Apr. 2025.
[45] D. Narayanan *et al.*, "Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM," Aug. 2021.
[46] J. Hoffmann *et al.*, "Training Compute-Optimal Large Language Models," Mar. 2022.
[47] J. Kaplan *et al.*, "Scaling Laws for Neural Language Models," Jan. 2020.
[48] J. Leike *et al.*, "AI Safety GSridworlds," Nov. 2017, [Online]. Available: http://arxiv.org/abs/1711.09883
[49] B. Zhuang *et al.*, "Beyond Benchmarks: The Economics of AI Inference," Oct. 2025, Accessed: Feb. 28, 2026. [Online]. Available: http://arxiv.org/abs/2510.26136
[50] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs," May 2023.