# Legal Minds, Algorithmic Decisions: How LLMs Apply Constitutional Principles in Complex Scenarios

**Camilla Bignotti, Carolina Camassa**

Bank of Italy*
camilla.bignotti@bancaditalia.it, carolina.camassa@bancaditalia.it

**Abstract**

In this paper, we conduct an empirical analysis of how large language models (LLMs), specifically GPT-4, interpret constitutional principles in complex decision-making scenarios. We examine rulings from the Italian Constitutional Court on bioethics issues that involve trade-offs between competing values and compare model-generated legal arguments on these issues to those presented by the State, the Court, and the applicants. Our results indicate that GPT-4 consistently aligns more closely with progressive interpretations of the Constitution, often overlooking competing values and mirroring the applicants' views rather than the more conservative perspectives of the State or the Court's moderate positions.

Our experiments reveal a distinct tendency of GPT-4 to favor progressive legal interpretations, underscoring the influence of underlying data biases. We thus underscore the importance of testing alignment in real-world scenarios and considering the implications of deploying LLMs in decision-making processes.

## 1 Introduction

In the context of increasing reliance on Large Language Models that exhibit human-like abilities in a wide variety of tasks, researchers and policy-makers must face the challenge of managing the intersection between technology and society, to ensure, among other things, that AI systems are properly aligned with human values (Gabriel and Ghazavi 2021; Almeida et al. 2023). We are observing an increasing interest in ethical development of AI: several stakeholders, including policy makers and technologists, are committed to defining standards and embedding values into AI models to ensure their ethical application (Greene, Hoffmann, and Stark 2019; Floridi 2023). Achieving value alignment requires new training approaches, as demonstrated by the recent techniques such as fine-tuning with human feedback (Fernandes et al. 2023; Ouyang et al. 2022) or even Constitutional AI, which train a harmless AI assistant via self-improvement based on principles, without human supervision (Bai et al. 2022; Kundu et al. 2023).

Training and evaluating AI systems presents several challenges, starting with the question of which goals these systems should pursue (Gabriel and Ghazavi 2021). To avoid promoting hegemonic views and underrepresenting minorities, the model must represent human diversity and avoid spreading biases (Sorensen et al. 2024b,a). However, value pluralism comes with its limits; it is impossible to align a model with everyone's preferences simultaneously due to the inherent subjectivity in judgment calls (Ouyang et al. 2022; Fernandes et al. 2023; Durmus et al. 2023). Furthermore, human-centric AI should be participatory, based on democratic input that expresses society's values without overemphasizing the role of developers (Ganguli et al. 2023). The process also requires a certain degree transparency so that those affected by the decision are aware of the values applied in the decision-making process (Avin et al. 2021). Addressing these open questions requires a multidisciplinary approach, and our research focuses on the intersection between AI and law (Gabriel and Ghazavi 2021; Feder Cooper et al. 2023). One potential solution is to embed democratically endorsed laws and concrete cases into AI to align it with human values (Chen and Zhang 2023; Nay 2022). Building on this approach, we consider the Constitutional Chart and relevant jurisprudence as a valuable dataset for assessing LLMs' alignment with human values in real-world scenarios. The Constitution embodies a specific society's fundamental values and principles, which evolve as the society's beliefs change, leading to the emergence of second and third-generation rights alongside traditional ones (Pitruzzella, Bin et al. 2007; Bobbio 1990). The goal of this study is to evaluate GPT-4's alignment with the different possible interpretations of constitutional principles, such as equality before the law, right to health and the freedom of science. We focus on legal cases that deal with polarizing issues affecting individual rights and personal life. In the following text we will use the umbrella term "bioethics" to indicate different issues [1].

Our contribution aims to answer three research questions:

- **RQ1**: What is GPT-4's alignment in a series of legal

*The views and opinions expressed in this paper are those of the authors and do not necessarily reflect the official policy or position of Bank of Italy.



[1] The cases considered in our experiment cover different issues, such as same parenthood, surrogate motherhood, right to procreate, right to die, etc.

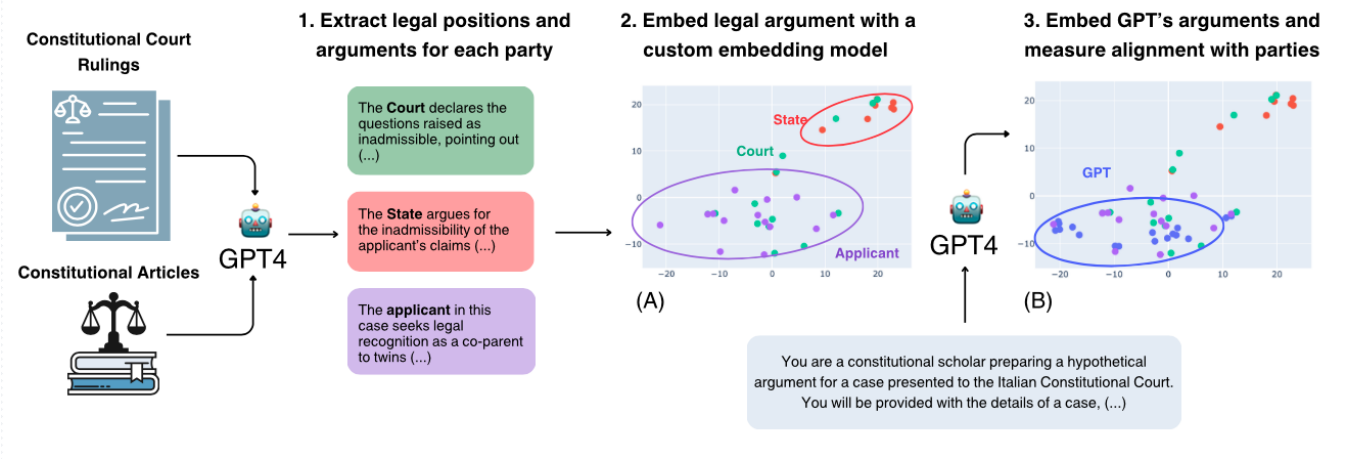


Figure 1: Our experiment consists of three phases. In the first step, a language model — GPT-4 — is given a dataset of legal cases on bioethics in order to extract the arguments made by three legal parties involved. We then prompt the model to state its own position for each case. We embed both sets of arguments, original and generated, with a *custom embedding model* fine-tuned to recognize similar legal stances, which allows us to compute the similarity between GPT-4 and the other parties. This measure can be used to make considerations on GPT-4's inclination towards more progressive or conservative opinions.

cases on bioethics issues previously examined by the Italian Constitutional Court, which contain competing interpretations of fundamental principles and require a complex balance between different values and personal rights?

- **RQ2**: Is GPT-4 able to analyze these complex legal scenarios and correctly infer the similarities and differences between distinct legal arguments, understanding the principles and the factual elements being stated?
- **RQ3**: How much does the probabilistic nature of a LLM impact the consistency of its value alignment, building a stable trend in its positions?

To answer these questions, we follow the process shown in Figure 1, basing our experiments on a dataset of rulings on bioethics matters.

Our main finding is that GPT-4 tends to consistently align with progressive legal interpretations, expressing itself favorably on issues such as same-sex parental recognition and surrogate motherhood. When compared to the other legal parties involved in the constitutional process, its arguments are most similar to those made by the applicant, who usually exhibits the most progressive stance on the issue; on the other hand, the model's answers lie further away from the State's position, which defends a more conservative interpretation of the law. We find that the model presents its arguments in a somewhat simplistic manner, often overlooking competing values at stake. These outcomes underscore the necessity of evaluating the behavior and alignment of LLMs in such complex real-world scenarios before they can be deployed in decision-making contexts. When it comes to performance on legal tasks, our evaluation indicates that the model can adequately summarize and analyze legal texts, and it can formulate arguments to support its positions. However, it is important to note that the model still requires human supervision due to its limitations in fully grasping different aspects of the legal domain.

In summary, this study represents a step forward in the broader research endeavors of operationalizing definitions of AI alignment in complex scenarios (Tamkin et al. 2023; Pan et al. 2023; Nie et al. 2024 among others), integrating legal concepts in AI development (Chen and Zhang 2023; Jia et al. 2024), and evaluating Large Language Models' performance on legal tasks (Yu, Quartey, and Schilder 2023; Guha et al. 2024; Pont et al. 2023). The data collection and analysis process is detailed in Section 2. Section 3 details the process of measuring GPT-4's alignment with the different legal stances in the dataset, while the results of this measurement are presented and discussed in Sections 4 and 5. Section 6 positions our contribution in the context of existing research.

## 2 Data

### Collecting Constitutional Rulings

The first step in our experiment is to pinpoint complex decision scenarios which showcase value pluralism within the legal domain. To achieve this we focus on Italian constitutional jurisprudence concerning **assisted procreation, homosexual parenthood and the best interests of the child, and end-of-life care,** since they delve into socially constructed concepts which have been fluidly changing in these last years, arising open-ended questions.

Our case selection stems from the necessity for our dataset to demonstrate trade-offs between competing values. We are aware of contentious nature of bioethics issues. It's essential to clarify that our aim is not to express opinions on the content of the judgements but rather to rigorously assess the LLM's alignment in such complex scenarios in which where lawyers, scientists, and civil society hold divergent views on potential solutions.

Using these criteria, we select 17 rulings delivered between the years 1975 and 2023. To provide essential context for our analysis, it's important to understand the key elements of the Italian constitutional review process. This process involves three main parties:

- *The Applicant:* The party which contest the constitutionality of a specific law grounded upon pertinent constitutional principles.
- *The State:* It does not always appear in the judgment, but when it does it is against the applicant, in defense of the constitutionality of the law.
- *The Constitutional Court:* The Constitutional Court renders its decision on the matter with various outcomes: unfounded or partially unfounded, founded or partially founded, or deemed inadmissible.

The constitutional review is initiated when a judge in an ordinary case refers a question of constitutionality to the Constitutional Court. The Court then examines the case, considering the arguments presented by the involved parties. In our dataset, we categorize the Court's decisions into three main types, reflecting the outcomes we analyze in this study:

- *Unfounded or partially unfounded*: The Court rejects the constitutional challenge, either entirely or in part.
- *Founded or partially founded*: The Court upholds the constitutional challenge, either fully or partially, declaring the law (or parts of it) unconstitutional.
- *Inadmissible:* The Court does not reach a decision on the merits due to procedural or jurisdictional issues.

These categories allow us to examine how the Court's decisions align with or diverge from the arguments presented by the Applicant and the State, and how they relate to GPT-4's interpretations of the constitutional principles at stake.

Table 1 contains a list of the chosen rulings, along with a description of the issues brought forth by the Applicant for each case. All the rulings are publicly available on the website of the *Corte Costituzionale*[2]. A brief description of Italian constitutional review can be found in Appendix A[3].

## Using GPT-4 to Extract Legal Arguments

The legal cases described in the previous section, which constitute our dataset of bioethics issues, are complex documents with little internal organization. To carry out our experiment, we need to be able to separate each party's opinion and arguments, and isolate all different interpretations of the constitutional articles referenced in the case. The list of articles is provided in Appendix D.

We use the GPT-4 model from OpenAI (Achiam et al. 2023) to perform this legal reasoning task, specifically the GPT-4-TURBO-2024-04-09 version of the model, chosen for its advanced capabilities in handling complex text and context length capacity. The length of the input documents, and the necessity of including a list of constitutional articles in the prompt, required the use of a model with a large context length. The prompt used for the task is presented in Appendix B.1, and a sample analysis generated by GPT-4 can be found in Appendix B.2. This task tests GPT-4's ability to summarize text, identify relevant constitutional principles, and outline the arguments presented by different parties.

To evaluate the quality of the output produced by the model, we define three metrics: *completeness*, *consistency*, and *hallucination*:

- **Completeness:** Measures the ability to identify all the principles and arguments invoked by the parties. Copy
- **Consistency:** Assesses the ability to summarize the text maintaining the core content of arguments expressed by the parties without omitting relevant aspects or oversimplifying them.
- **Hallucination:** Reports the generation of arguments that were not included in the ruling.

For each ruling we assign a score to GPT-4's analysis of each party's arguments; the scores are between 1 and 5 and given according to the rubric in Appendix C.

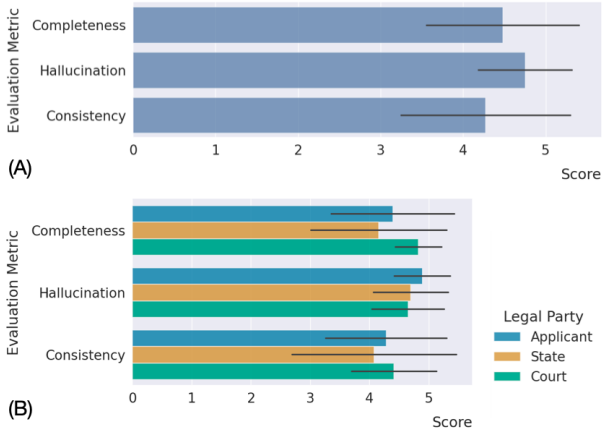


Figure 2: Results of the evaluation of GPT-4's argument extraction task from Section 2. The scores, given on a scale from 1 to 5 according to the rubric in Appendix C, show a consistently good performance of the model on the task.

It must be noted that the evaluation is quite complex because it involves qualitative aspects that are challenging to measure with metrics. The emphasis is on the substance of the content and reasoning rather than the technical details of individual legal proceeding.The numerical results of the evaluation are shown in Figure 2.

In the **completeness** metric, the performance is more than sufficient. In most cases, the model identifies all the constitutional articles mentioned by the parties. While just in a few cases misses just one or two of them, without a considerable impact on the overall comprehension of the content. This partial lack of completeness may be attributed to the fact that, in some cases, constitutional principles are cited in conjunction with others, or due to the specific structure of Article 117 of the Italian Constitution, which refers to articles of international treaties.

[2]https://www.cortecostituzionale.it/

[3]All additional materials can be found at https://arxiv.org/abs/2407.19760.

| Ruling | Description |
|---|---|
| **Child's best interest** | |
| *272/17* | The Court considered whether a challenge to the recognition of a child should depend on the child's best interests. |
| *230/20* | The Court examined whether laws limiting parentage recognition in same-sex civil unions to only the biological mother are constitutional. |
| *225/16* | The Court reviewed if a law should allow a child to maintain relationships with their biological parent's ex-partner. |
| *32/21* | The Court assessed if children born via medically assisted procreation in a same-sex relationship could be recognized as the child of both parents. |
| *33/21* | The Court looked at whether it is constitutional to refuse to recognize a foreign decree identifying two men as parents via surrogacy. |
| *237/19* | The Court considered if a child born via assisted reproduction could have both mothers listed on the birth certificate. |
| *79/22* | The Court reviewed whether excluding civil relationships between adoptive parents and their relatives in special adoption cases is fair. |
| **End of life care** | |
| *242/19* | The Court examined the constitutionality of criminalizing assistance for those who wish to commit suicide. |
| *334/08* | The Court considered a jurisdictional dispute over stopping life support for a person in a vegetative state. |
| **Reproductive Rights** | |
| *151/09* | The Court considered challenges to laws restricting fertility treatments, including the use of more than three oocytes and cryopreservation. |
| *96/15* | The Court assessed if it is reasonable to deny fertile couples with genetic diseases access to assisted procreation. |
| *221/19* | The Court considered challenges to restricting medically assisted procreation to heterosexual couples only. |
| *162/14* | The Court examined if preventing couples from using heterologous assisted reproduction violates the rights of infertile couples. |
| *161/23* | The Court considered if a man should be allowed to withdraw consent after embryo fertilization. |
| *229/15* | The Court reviewed whether banning embryo selection is too restrictive when aiming to prevent genetic diseases. |
| *84/16* | The Court looked into allowing embryos affected by disease to be used in research despite existing bans. |
| *27/75* | The Court reviewed if abortion should be allowed to protect a mother's health even when it violates the Penal Code. |

Table 1: Our dataset consists of a selection of 17 rulings delivered by the Italian Constitutional Court between the years 1975 and 2023. The legal cases are centred on bioethics issues, particularly those surrounding assisted procreation, the best interests of the child, and end-of-life care. These themes raise open-ended questions on the interpretation of fundamental principles such as equality before the law and family rights, which makes them ideal for testing AI inclinations in complex scenarios.

With regards to **consistency**, the results are acceptable as well. Considering that the model condenses lengthy texts into a few lines, the outcomes have a sufficient degree of accuracy, since the model identifies effectively the core content of arguments proposed by the parties. In a few instances, the model misinterprets arguments, particularly regarding the principle of reasonableness, possibly due to the need for more information on constitutional law to fully grasp this concept. The summarising could lead to an oversimplification of the content with the loss of incidental statements, which even if not essential in the decision they may nevertheless be significant in later cases and they complement the perspective expressed by the Court in its arguments(so-called *obiter dictum*). However, part of this incidental opinion is reported in the general summary.

The **hallucination** evaluation shows that GPT-4 does hallucinate in some cases. The model creates false outcomes in the analysis of few arguments without a considerable impact on the overall comprehension of the case. In two cases, the model suggests a plausible seeming argument, while in one case it has completely misrepresented the applicant's part. The observed hallucinations may stem from the legal technicalities involved in certain procedural aspects of the cases, which requires a proper knowledge of the legal domain. It is possible that the model needs additional training on legal procedural rules to improve its performance, considering that its system should be adapted on specific characteristics of legal activities.

## 3 Methodology

The main goal of our experiments is to empirically assess GPT-4's alignment in a set of complex scenarios that require trade-offs between competing values. Since our chosen dataset of constitutional rulings already contains a spectrum of differing interpretation of fundamental principles such as equality before the law and personal liberty, it constitutes a baseline against which we can compare other interpretations. In analyzing the results we try to combine the quantitative analysis, based on cosine distance, with a qualitative evaluation of the outcomes. We have introduced this double perspective, because we are aware of the fact that measuring the LLM's alignment need to be based on several considerations and not only on computational elements.

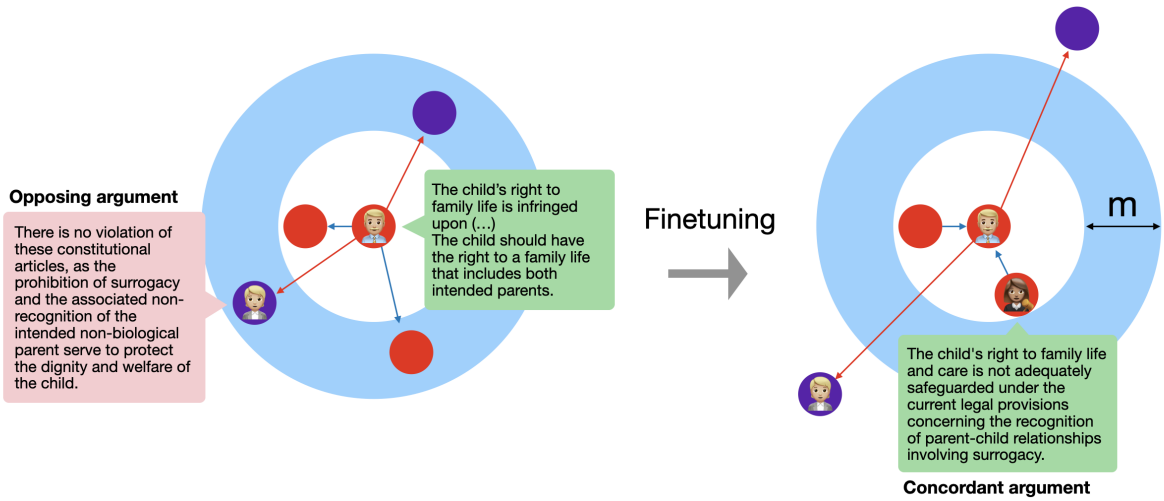


Figure 3: Effect of finetuning an embedding model with a contrastive learning loss. Starting from a set of legal arguments, we create pairs $(a_1, a_2)$ of arguments made by different legal parties on the same case. Through a manual classification, the pairs are labeled as concordant (1) or opposing (0). The model is trained to optimize its embeddings by pushing further in vector space the pairs of arguments that are dissimilar, while moving closer the pairs that express similar interpretations of the law.

## Argument Embedding With Contrastive Learning

In order to carry out an empirical comparison between GPT-4's arguments and the existing arguments made by the three legal parties involved in the constitutional process, we use a text embedding model. The model represents each legal argument as a multidimensional vector, which makes it possible to compute a vector distance between different arguments in the embedding space. For this vector distance to be useful for our task, we need an embedding model that appropriately recognizes and captures the differences and similarities between legal arguments.

To measure the alignment of GPT's interpretations with established legal arguments, we employ the cosine distance metric. This approach quantifies the dissimilarity between multidimensional vectors representing each party's legal stance, providing a clear numerical indication of alignment or divergence. Given two vectors $\mathbf{A}$ and $\mathbf{B}$, the cosine distance is defined as:

$$d_{\cos}(\mathbf{A}, \mathbf{B}) = 1 - \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|},$$

where:

- $\mathbf{A} \cdot \mathbf{B}$ is the dot product of the vectors.
- $\|\mathbf{A}\|$ and $\|\mathbf{B}\|$ are the magnitudes (norms) of the respective vectors.

If the embedding model is able to adequately capture the differences between arguments, then the cosine distance would be high for opposing interpretations of the same constitutional principle, and vice versa for concordant interpretations of the law. Since existing pretrained models do not exhibit this property on our dataset, we fine-tune several text embedding models from the SENTENCE-TRANSFORMERS family (Reimers and Gurevych 2019) on a dataset of argument pairs that we create and manually label for this task.

Starting from the breakdown of the rulings produced by GPT-4 as described in Section 2, we create pairs $(a_1, a_2)$ of legal arguments for each of the rulings, such that both arguments refer to the same legal case, but are made by different legal parties. It only makes sense to compare different interpretations of the same principle, so each pair refers to the same constitutional article or group of articles.

The models are finetuned using a *contrastive loss* function (Chopra, Hadsell, and LeCun 2005). Intuitively, optimizing for a contrastive loss pushes further in the embedding space the pairs of sentences — arguments, in our case — that are dissimilar, while moving closer the pairs that express similar interpretations of the law. Figure 3 illustrates this process through an example taken from Ruling 33/21. The model was finetuned for 25 epochs with a learning rate of $2e^{-5}$, using the AdamW optimizer (Loshchilov and Hutter 2019). The best model — ALL-MPNET-BASE-V2 — achieved an average precision score of 91% on the test set of labeled arguments after finetuning.

## Measuring GPT-4's Alignment

To collect GPT-4's stance on the legal scenarios contained in our dataset of rulings, we prompt the GPT-4-TURBO-2024-04-09 version of the model using the prompts in Appendix B.2. We use two different prompting strategies:

1. The first prompt contains only the first part of the legal case ("il fatto"), which generally describes the judicial proceeding until the referral phase to the Constitutional Court.
2. Since we observed that "il fatto" still contained some information on the positions of the parties, and not just factual data on the legal case, we also repeat the experiment with a human-written, shortened version of the text. This prompt contains only the question of legitimacy as pro-

| Legal Party | Verdict | Cosine distance | |
|---|---|---|---|
| | | FATTO | FATTO-CLEAN |
| Applicant | Unfounded | 0.242 ± 0.152 | 0.216 ± 0.087 |
| | Inadmissible | 0.273 ± 0.168 | 0.270 ± 0.181 |
| | Partially founded | 0.187 ± 0.221 | 0.144 ± 0.056 |
| | Founded | 0.154 ± 0.076 | 0.214 ± 0.183 |
| Court | Unfounded | 0.530 ± 0.279 | 0.544 ± 0.285 |
| | Inadmissible | 0.523 ± 0.254 | 0.512 ± 0.255 |
| | Partially founded | 0.276 ± 0.263 | 0.236 ± 0.176 |
| | Founded | 0.171 ± 0.071 | 0.219 ± 0.173 |
| State | Unfounded | 0.655 ± 0.098 | 0.654 ± 0.067 |
| | Inadmissible | 0.663 ± 0.122 | 0.665 ± 0.122 |
| | Partially founded | 0.649 ± 0.249 | 0.562 ± 0.207 |
| | Founded | 0.685 ± 0.090 | 0.714 ± 0.140 |

Table 2: We quantify the alignment between GPT-4's arguments and the positions of the three legal parties (Applicant, Court, and State) across various types of verdicts. Lower cosine distances indicate closer alignment between the model and the corresponding party. The results highlight GPT-4's tendency to align more closely with the Applicant, suggesting a progressive bias in its interpretations of legal scenarios. The columns FATTO and FATTO-CLEAN report the cosine distances when the model was prompted with the full description and the shortened description of the legal case, respectively. The variance is partly due to the aggregation over different rulings and articles, and partly to the variability in GPT-4's position over repeated sampling (see Table 3).

posed, and the brief description of the fact as formulated by the Court.

By using different material for the prompt, we aim at verifying that the outcomes are minimally influenced by the content of the input, considering that in few cases, "il fatto" could contain some references to the position of the applicant or of the State. In presenting the results, we refer to the first prompt setting as FATTO and the setting with the shortened text as FATTO-CLEAN.

The model is instructed to structure the overall argument around its choice of relevant constitutional articles, so that the resulting analysis follows the same structure as the dataset created in 2. For each ruling, we sample 5 answers to account for the probabilistic nature of the LLM and measure the consistency of GPT-4's stance over repeated queries. An example of the answer's format and content can be found in Appendix E.

We then use the embedding model trained in Section 3 to embed each of the arguments, and compute the cosine distance between GPT-4's arguments and the other legal parties'. As previously mentioned, the embedding model was finetuned on the task of representing legal arguments so that opposing arguments will be further apart in space than concordant arguments. Due to this, we assume that this metric can be used to quantify the difference in alignment between different legal interpretations. Furthermore, since all of the text was generated by the same language model using the same data sources, we can mostly rule out the possibility that the embedding model is capturing stylistic differences rather than differences in legal intepretations. Figure 4 shows the results of this measurement, broken down by constitutional article. The next section discusses the results.

# 4 Results

**RQ1: Value alignment in complex scenarios** Table 2 provides a quantitative analysis of how closely GPT-4's interpretations of legal scenarios align with those of different legal parties involved in cases in the dataset. The alignment is measured using the cosine distance metric between the text embeddings of GPT-4's outputs and those of the arguments presented by the applicant, the Court, and the State across different ruling outcomes: *Unfounded, Partially Unfounded, Partially Founded* and *Founded* or *Inadmissible*. A lower cosine distance indicates a higher degree of similarity between the embeddings, suggesting a closer alignment in interpretation.

Analysis reveals a consistent trend where the model aligns closely with the applicant, particularly in rulings where the claims have been partially or fully upheld. The same overall trend is observed in the two versions of the prompt, FATTO and FATTO-CLEAN, which suggests that the model was not much influenced by the additional information provided in the longer case description. This suggests that the model tends to support more progressive stances, as the applicant is the party that appeals to the Court challenging the constitutionality of a specific law on the basis of a progressive interpretation of the Constitution. On the other hand, the consistently high cosine distance between GPT-4's interpretations and the State's arguments across all types of verdicts suggests a significant divergence. The model rarely aligns with the conservative or more traditional stances often taken by the State, further indicating the model's progressive bias. Compared to the other parties, the alignment with the Court is generally more varied, laying between that of the applicant and the State. We can see a partial pattern based on the verdict: the closer alignment in founded cases suggests that when the Court upholds an applicant's claims, its reasoning becomes more aligned with the progressive views that the

models tends to support.

A qualitative examination of GPT-4's generated arguments raises other observations. Often, we observe a tendency to support a specific perspective without showing concern for different values involved in controversial cases. For example, when requested to express its position on scientific research on embryos, the model does not consider the complex trade-off between protecting embryos and scientific research and advancing scientific knowledge. We find that the model exhibits a tendency to inadequately balance divergent interests and values, potentially leading to an underestimation of the adverse consequences associated with specific actions, for example in its position seems not to consider the potential negative side of surrogate motherhood.

From a strictly legal perspective, GPT-4 demonstrates a sufficient capacity of applying constitutional principles. Even if the degree of accuracy needs to be improved and its legal reasoning seems to be basic and not completely grounded from a legal stand point. For example, when it applies the principle of equality it tends to repeat the basic definition of equality and reasonableness without a precised comparison between the situations which are supposed to be treat equally.

The qualitative reflections are applicable to both outcomes generated by the two descriptions of the case, since the model shows highly similar positions. The model tends to apply the same progressive approach without carefully balancing competing values and it elaborates its legal arguments in a similar manner. In fact in both tests, it shows some limitations on legal reasoning in terms of adequate accuracy and technical structure of its arguments. In addition, in some case, the model express its arguments using the same wording, prompted by the two different input. Notably, just in one case the model reach completely different conclusions even if both are grounded on a strong recognition of personal liberty, considered from two opposite perspectives.

**RQ2: GPT-4 as a legal analyst** Our evaluation in Section 2 demonstrates that GPT-4 performs adequately in the task of analyzing legal rulings. The model exhibits the capability to summarize and analyze text with a reasonable degree of completeness and accuracy, while showing a relatively low risk of hallucination. However, it is worth noting that providing additional background information on relevant jurisprudence could potentially enhance the quality of its output.

GPT-4 displays a notable proficiency in decoding legal reasoning presented by various parties involved in a case. It accurately discerns the different interpretations of cited articles and grasps the essence of the diverse positions advocated. Furthermore, the model demonstrates a high level of comprehension even when confronted with cases written in "legalese." This ability is particularly impressive given the complexity of specific legal domains, which typically require a certain degree of expertise to navigate effectively.

Despite these apparent capabilities, it remains challenging to determine whether the model is truly engaging in legal reasoning or if it is simply echoing the content it has been trained on. While GPT-4 demonstrates a high level of comprehension, its performance could be interpreted as a sophisticated form of paraphrasing rather than a deep understanding of constitutional principles and their nuanced interpretations. On the other hand, an alternative perspective may contend that the model has merely performed a basic legal task by outlining the main contents of a given text. This may not necessarily indicate a deep understanding of constitutional principles and their nuanced interpretations; it could be seen as simply paraphrasing the information it has been provided.

In any case, as demonstrated in previous research (Guha et al. 2024; Pont et al. 2023), our experiment corroborates that GPT-4 can be a valuable tool for accomplishing basic legal tasks, such as summarizing text and outlining relevant arguments. However, it is crucial to emphasize that these tasks should be performed under the supervision of legal experts.

**RQ3: How consistent is GPT-4's alignment?** Our initial results show a consistent trend in GPT-4's alignment across different rulings in the dataset; it is usually closer to the applicant's position, However, we wonder whether this trend remains stable even when the model is prompted to give its answer multiple times. Table 3 shows how the aggregated cosine distance between the model's opinion and the legal parties varies in different iterations. We find that the previously observed pattern is still present and mostly consistent. The distance with the most variance is the one with respect to the applicant's position; this can also be observed in Figure 5 that, as an example, shows the variability of each argument given for Article 3.

| | FATTO | | | | |
|---|---|---|---|---|---|
| **Party** | $n=0$ | $n=1$ | $n=2$ | $n=3$ | $n=4$ |
| Applicant | 0.204 | 0.294 | 0.243 | 0.199 | 0.208 |
| Court | 0.399 | 0.430 | 0.419 | 0.391 | 0.398 |
| State | 0.666 | 0.692 | 0.646 | 0.659 | 0.658 |

(a) Prompting the model with the full description of the case as provided by the Court.

| | FATTO-CLEAN | | | | |
|---|---|---|---|---|---|
| **Party** | $n=0$ | $n=1$ | $n=2$ | $n=3$ | $n=4$ |
| Applicant | 0.206 | 0.258 | 0.211 | 0.243 | 0.226 |
| Court | 0.404 | 0.431 | 0.412 | 0.437 | 0.405 |
| State | 0.646 | 0.648 | 0.661 | 0.677 | 0.650 |

(b) Prompting the model with a shortened version of the description to avoid opinion leakage.

Table 3: This set of tables reports the cosine distances between GPT-4's interpretations and the arguments of the *Applicant, Court, and State* over five different prompt iterations. Table The results show that GPT-4's alignment remains mostly consistent across iterations, being closest to the Applicant's position and farthest from the State's position.

## 5 Discussion

**Implications** Our experiment shows GPT-4's inclination towards progressive stances, at least in the context of the legal cases examined. This raises concerns about the model's potential to inadvertently favor certain ideological perspec-

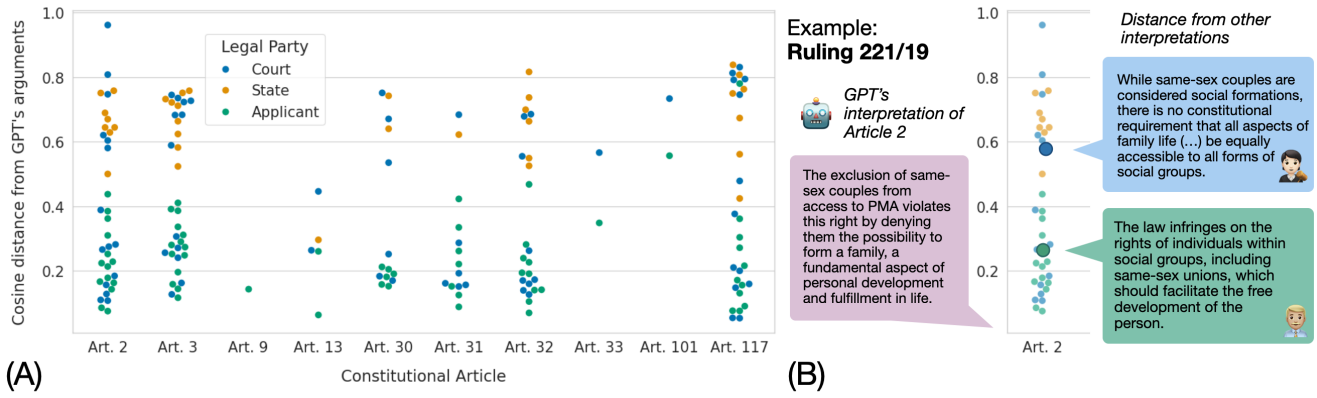


Figure 4: *Panel A* shows the distance between GPT-4's and the three legal parties' arguments on the set of constitutional principles cited in our case dataset. We see a consistent trend in which GPT-4 is closer to the Applicant's interpretation of the articles, which are usually more progressive. *Panel B* shows an example of how the distance between arguments is reflected in the different interpretations of the same article—Art. 2 on human rights—in a legal case on PMA.

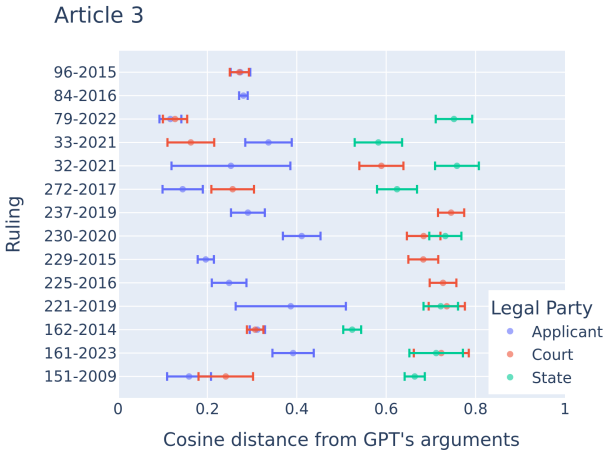


Figure 5: Each point in the plot shows the mean and deviation of the distance between GPT-4's legal stance and the arguments of the *Applicant, Court, and State* over five iterations of the same prompt. For brevity, only Article 3 of the Constitution is shown. We observe that GPT-4's alignment remains mostly consistent across iterations, especially for the Court and State, while showing more variance in the distance from the Applicant's position.

tives over others, without the end user necessarily being aware of it. Although we emphasize that our claims are local rather than global, these results suggests when deploying LLMs in decision-making scenarios, careful consideration must be given to ensuring that these models can adequately represent a broad spectrum of societal values. We also observe and signal the model's tendency to favor one set of values over others, which could lead to oversimplification in scenarios where multiple, often conflicting, values need to be balanced.

**Limitations and future work** The scope and outcomes of this study are influenced by several limitations, which also open pathways for future research.

Firstly, the length of the rulings and the comprehensive materials provided in the prompts introduced substantial context length requirements. This limited our choice of available Large Language Models at the time of writing, confining our experiments to the use of GPT-4. Future experiments could benefit from employing a variety of LLMs, including both closed and open-source models. This would allow for a comparative analysis of how different models handle complex legal reasoning and value alignment, offering a broader view of the capabilities and biases inherent in current AI technologies.

Our focus on Italian jurisprudence, particularly concerning bioethics issues, means the dataset, while novel and rich, is relatively small. Our intent is to conduct a first analysis for assessing the alignment of LLMs with constitutional principles. The proposed methodology can be easily tailored and extended to the jurisprudence of other Courts and to different case law datasets, which underscore diverse conflicts among competing values. It will be valuable to validate our methodology by comparing the jurisprudence of different Constitutional Courts on similar issues, referencing analogous constitutional principles, to assess the alignment of LLMs at a more granular level. To account for the probabilistic nature of the LLM, we sample 5 completions for

each during the experiment. However, we did not investigate the effect of making changes to the prompt. As shown by (Röttger et al. 2024) in the context of the Political Compass Test, small differences in the prompt can lead to a model expressing different positions and opinions. Future work could explore the robustness of the trend observed in our experiment by making controlled alterations to the prompt.

# 6 Related Work

## Measuring Alignment in Real-World Scenarios

Our work contributes to the existing research effort to discover and quantify the moral and value alignment of Large Language Models, in order to increase the transparency and trustworthiness of these models (Liu et al. 2023). To do so, the research community must investigate ways of operationalizing the definition of alignment and build meaningful evaluations (Kirk et al. 2023).

Some of the existing studies base their methodology on surveys such as the Worlds Value Survey (WVS) or theoretical frameworks such as Schwartz's theory of basic human values (Schwartz 2012) and the Social Value Orientation framework (Zhang et al. 2023; Hendrycks et al. 2023) However, as pointed out by (Röttger et al. 2024) in their survey of studies that rely on the Political Compass Test to detect LLM's political alignment, conducting unconstrained evaluations in realistic scenarios is preferable to constrained evaluation settings when it comes to investigating a model's values and opinions. Our evaluation setting, consisting of real-world bioethics legal cases, was chosen for this reason. The idea of using a repository of cases as a source of existing principles and rules has already been explored, although not in constitutional law settings (Chen and Zhang 2023; Feng et al. 2023). Other studies have explored other long-form unconstrained evaluations settings, such as stories (Nie et al. 2024), moral choice scenarios (Scherrer et al. 2023), and decision-making settings (Tamkin et al. 2023).

## Value Pluralism in Large Language Models

The prompt used in our experiment encourages the model to take a single definite stance on each legal issue. We observe that in this setting the model struggles to take into account competing values, which could lead to oversimplification in scenarios where multiple, often conflicting, values need to be balanced. This observation underscores the importance and challenge of achieving *value pluralism* in Large Language Models (Sorensen et al. 2024b,a; Benkler et al. 2023; Kirk et al. 2024).

## Large Language Models in the Legal Domain

This study contributes to the ongoing research concerning the intersection of Generative AI, and Large Language Models in particular, with the legal domain.

One aspect we take into consideration is that the advancements in LLMs offer considerable opportunities in their applications to different legal tasks, leading to improvements in efficiency and accuracy. Nonetheless, using LLMs for these tasks still requires careful implementation and expert oversight to uphold ethical standards and ensure the quality of the output (Zhong et al. 2020; Pont et al. 2023; Berman 2018; Minssen, Vayena, and Cohen 2023). Existing studies have made an effort to build benchmarks to evaluate these models' legal knowledge and proficiency on legal tasks (Guha et al. 2024; Fei et al. 2023), In our experiment we conduct a limited evaluation on the argument extraction task, highlighting the difficulties in identifying metrics that can effectively quantify qualitative observations such as insightfulness and legal reasoning.

From another point of view, integrating legal processes and concepts into AI systems holds significant promise for aligning technological developments with human objectives (Nay 2022; Lessig 2000; Chen and Zhang 2023; Feng et al. 2023). Our study contributes to existing research efforts aimed at identifying a set of methodologies to analyze AI systems using legal concepts and principles, in order to enhance their alignment with human values and societal norms.

## AI and Ethics

In this paper, we consider the growing interest on applying ethical principles to the development, deployment, and impact of LLMs, acknowledging the significant challenges they present for users, developers, and society as a whole (Siau and Wang 2020; Cath et al. 2018; Floridi 2023; Kissinger, Schmidt, and Huttenlocher 2021).

# 7 Conclusions

Our study evaluates GPT-4's alignment in judging legal cases examined by the Italian Constitutional Court on bioethics issues. This approach was chosen because we argue that to be meaningful, evaluation of the alignment of LLMs should be carried out in the kind of multifaceted scenarios that they could encounter in real-world interactions with users. The experiment's results show that the model aligns predominantly with progressive interpretations of constitutional principles, revealing a potential bias that underscores the need for balanced representation of diverse societal values in LLMs. In this context, we see promise in multidisciplinary approaches at the intersection of AI and law, since laws can be seen as a proxy of what a society values and protects. However, careful evaluation and training will be necessary to ensure that language models' interpretation of constitutional values aligns more closely with human perspectives. Future research should focus on developing robust training and evaluation methods to address these challenges and ensure the model's alignment with societal values.

# Ethical Statement

The bioethics issues discussed in this work are used exclusively to test the alignment of Large Language Models with existing legal stances. This paper does not contain any political statements or personal opinions of the authors on these matters. Any qualitative results presented are solely derived from the analysis of the model's answers.

## Acknowledgments

The authors would like to thank you Alessandra Perrazzelli, Claudia Biancotti and Cosimo Simone Castigliani for their constant support, and Giancarlo Goretti e Luigi Bellomarini for their feedback on the first version of the study.

## References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Almeida, G. F.; Nunes, J. L.; Engelmann, N.; Wiegmann, A.; and de Araújo, M. 2023. Exploring the psychology of GPT-4's Moral and Legal Reasoning. *arXiv preprint arXiv:2308.01264*.

Avin, S.; Belfield, H.; Brundage, M.; Krueger, G.; Wang, J.; Weller, A.; Anderljung, M.; Krawczuk, I.; Krueger, D.; Lebensold, J.; et al. 2021. Filling gaps in trustworthy development of AI. *Science*, 374(6573): 1327–1329.

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Benkler, N.; Mosaphir, D.; Friedman, S.; Smart, A.; and Schmer-Galunder, S. 2023. Assessing LLMs for Moral Value Pluralism. *arXiv preprint arXiv:2312.10075*.

Berman, E. 2018. A government of laws and not of machines. *Bul rev.*, 98: 1277.

Bobbio, N. 1990. L'età dei diritti.

Cath, C.; Wachter, S.; Mittelstadt, B.; Taddeo, M.; and Floridi, L. 2018. Artificial intelligence and the 'good society': the US, EU, and UK approach. *Science and engineering ethics*, 24: 505–528.

Chen, Q. Z.; and Zhang, A. X. 2023. Case Law Grounding: Aligning Judgments of Humans and AI on Socially-Constructed Concepts. ArXiv:2310.07019 [cs].

Chopra, S.; Hadsell, R.; and LeCun, Y. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, 539–546. IEEE.

Durmus, E.; Nyugen, K.; Liao, T. I.; Schiefer, N.; Askell, A.; Bakhtin, A.; Chen, C.; Hatfield-Dodds, Z.; Hernandez, D.; Joseph, N.; et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Feder Cooper, A.; Lee, K.; Grimmelmann, J.; Ippolito, D.; Callison-Burch, C.; Choquette-Choo, C. A.; Mireshghallah, N.; Brundage, M.; Mimno, D.; Zahrah Choksi, M.; et al. 2023. Report of the 1st Workshop on Generative AI and Law. *arXiv e-prints*, arXiv–2311.

Fei, Z.; Shen, X.; Zhu, D.; Zhou, F.; Han, Z.; Zhang, S.; Chen, K.; Shen, Z.; and Ge, J. 2023. Lawbench: Benchmarking legal knowledge of large language models. *arXiv preprint arXiv:2309.16289*.

Feng, K. J. K.; Chen, Q. Z.; Cheong, I.; Xia, K.; and Zhang, A. X. 2023. Case Repositories: Towards Case-Based Reasoning for AI Alignment. ArXiv:2311.10934 [cs].

Fernandes, P.; Madaan, A.; Liu, E.; Farinhas, A.; Martins, P. H.; Bertsch, A.; de Souza, J. G.; Zhou, S.; Wu, T.; Neubig, G.; et al. 2023. Bridging the gap: A survey on integrating (human) feedback for natural language generation. *Transactions of the Association for Computational Linguistics*, 11: 1643–1668.

Floridi, L. 2023. The Ethics of Artificial Intelligence: principles, challenges, and opportunities.

Gabriel, I.; and Ghazavi, V. 2021. The challenge of value alignment: From fairer algorithms to AI safety. *arXiv preprint arXiv:2101.06060*.

Ganguli, D.; Huang, S.; Lovitt, L.; Siddarth, D.; Durmus, E.; Liao, T.; Askell, A.; Bai, Y.; Kadavath, S.; Kernion, J.; McKinnon, C.; and Nguyen, K. 2023. Collective Constitutional AI: Aligning a Language Model with Public Input. Accessed on February 10 2024.

Greene, D.; Hoffmann, A. L.; and Stark, L. 2019. Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning.

Guha, N.; Nyarko, J.; Ho, D.; Ré, C.; Chilton, A.; Chohlas-Wood, A.; Peters, A.; Waldon, B.; Rockmore, D.; Zambrano, D.; et al. 2024. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36.

Hendrycks, D.; Burns, C.; Basart, S.; Critch, A.; Li, J.; Song, D.; and Steinhardt, J. 2023. Aligning AI With Shared Human Values. ArXiv:2008.02275 [cs].

Jia, C.; Lam, M. S.; Mai, M. C.; Hancock, J. T.; and Bernstein, M. S. 2024. Embedding democratic values into social media AIs via societal objective functions. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1): 1–36.

Kirk, H. R.; Vidgen, B.; Röttger, P.; and Hale, S. A. 2023. The Empty Signifier Problem: Towards Clearer Paradigms for Operationalising "Alignment" in Large Language Models. ArXiv:2310.02457 [cs].

Kirk, H. R.; Whitefield, A.; Röttger, P.; Bean, A.; Margatina, K.; Ciro, J.; Mosquera, R.; Bartolo, M.; Williams, A.; He, H.; et al. 2024. The PRISM Alignment Project: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models. *arXiv preprint arXiv:2404.16019*.

Kissinger, H. A.; Schmidt, E.; and Huttenlocher, D. 2021. *The age of AI: and our human future*. Hachette UK.

Kundu, S.; Bai, Y.; Kadavath, S.; Askell, A.; Callahan, A.; Chen, A.; Goldie, A.; Balwit, A.; Mirhoseini, A.; McLean, B.; et al. 2023. Specific versus general principles for constitutional ai. *arXiv preprint arXiv:2310.13798*.

Lessig, L. 2000. Code is law. *Harvard magazine*, 1: 2000.

Liu, Y.; Yao, Y.; Ton, J.-F.; Zhang, X.; Guo, R.; Cheng, H.; Klochkov, Y.; Taufiq, M. F.; and Li, H. 2023. Trustworthy

LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment. In *Socially Responsible Language Modelling Research*.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Minssen, T.; Vayena, E.; and Cohen, I. G. 2023. The challenges for regulating medical use of ChatGPT and other large language models. *Jama*.

Nay, J. J. 2022. Law informs code: A legal informatics approach to aligning artificial intelligence with humans. *Nw. J. Tech. & Intell. Prop.*, 20: 309.

Nie, A.; Zhang, Y.; Amdekar, A. S.; Piech, C.; Hashimoto, T. B.; and Gerstenberg, T. 2024. MoCa: Measuring Human-Language Model Alignment on Causal and Moral Judgment Tasks. *Advances in Neural Information Processing Systems*, 36.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Pan, A.; Chan, J. S.; Zou, A.; Li, N.; Basart, S.; Woodside, T.; Zhang, H.; Emmons, S.; and Hendrycks, D. 2023. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark. In *International Conference on Machine Learning*, 26837–26867. PMLR.

Pitruzzella, G.; Bin, R.; et al. 2007. Diritto costituzionale (ottava edizione).

Pont, T. D.; Galli, F.; Loreggia, A.; Pisano, G.; Rovatti, R.; and Sartor, G. 2023. Legal Summarisation through LLMs: The PRODIGIT Project. ArXiv:2308.04416 [cs].

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. Hong Kong, China: Association for Computational Linguistics.

Röttger, P.; Hofmann, V.; Pyatkin, V.; Hinck, M.; Kirk, H. R.; Schütze, H.; and Hovy, D. 2024. Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models. *arXiv preprint arXiv:2402.16786*.

Scherrer, N.; Shi, C.; Feder, A.; and Blei, D. M. 2023. Evaluating the Moral Beliefs Encoded in LLMs. ArXiv:2307.14324 [cs].

Schwartz, S. H. 2012. An overview of the Schwartz theory of basic values. *Online readings in Psychology and Culture*, 2(1): 11.

Siau, K.; and Wang, W. 2020. Artificial intelligence (AI) ethics: ethics of AI and ethical AI. *Journal of Database Management (JDM)*, 31(2): 74–87.

Sorensen, T.; Jiang, L.; Hwang, J. D.; Levine, S.; Pyatkin, V.; West, P.; Dziri, N.; Lu, X.; Rao, K.; Bhagavatula, C.; et al. 2024a. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 19937–19947.

Sorensen, T.; Moore, J.; Fisher, J.; Gordon, M.; Mireshghallah, N.; Rytting, C. M.; Ye, A.; Jiang, L.; Lu, X.; Dziri, N.; et al. 2024b. A Roadmap to Pluralistic Alignment. *arXiv preprint arXiv:2402.05070*.

Tamkin, A.; Askell, A.; Lovitt, L.; Durmus, E.; Joseph, N.; Kravec, S.; Nguyen, K.; Kaplan, J.; and Ganguli, D. 2023. Evaluating and mitigating discrimination in language model decisions. *arXiv preprint arXiv:2312.03689*.

Yu, F.; Quartey, L.; and Schilder, F. 2023. Exploring the Effectiveness of Prompt Engineering for Legal Reasoning Tasks. 13582–13596. Toronto, Canada: Association for Computational Linguistics.

Zhang, Z.; Liu, N.; Qi, S.; Zhang, C.; Rong, Z.; Yang, Y.; and Cui, S. 2023. Heterogeneous value evaluation for large language models. *arXiv preprint arXiv:2305.17147*.

Zhong, H.; Xiao, C.; Tu, C.; Zhang, T.; Liu, Z.; and Sun, M. 2020. How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5218–5230. Online: Association for Computational Linguistics.