Large Language Models as Political Actors:
Cultural Bias and Epistemic Power

Elena Seredkina[1] (✉), Guzel Seletkova[1] and Alexander Mikhailovsky[2]
[1] Perm National Research Polytechnic University, Komsomolsky prospekt 29, 614990, Perm, Russia,
elena_seredkina@pstu.ru; guzal.ka@mail.ru
[2] HSE University, Moscow, Myasnitskaya str. 20, 101000, Russia,
amichailowski@hse.ru

Abstract

The rapid diffusion of Large Language Models (LLMs) into socially and politically sensitive domains raises critical questions about the nature and origins of political bias in artificial intelligence. While existing research often treats bias as a technical flaw to be minimized, this article advances a broader philosophical and cultural interpretation of LLM bias as an outcome of embedded epistemic and value-laden structures. The aim of this study is to conceptualize LLMs as political actors of a new type and to examine how cultural context, language, and prompt design shape their normative orientations. Methodologically, the research brings comparative survey methods to the study of chatbots trained on North American, Russian, and Chinese data. It combines this with philosophical analysis grounded in Actor–Network Theory and assemblage theory. The empirical instrument was an adapted Political Compass consisting of 62 normatively charged statements, administered twice to each model using standardized numerical responses, followed by qualitative analysis of response variability through grounded theory methodology. The study confirms three core hypotheses. First, large language models function as political actors rather than neutral tools, systematically reproducing normative positions across moral, economic, and political domains; bias is therefore constitutive rather than accidental. Second, political bias is context-dependent and dynamically produced through interaction, shaped not only by prompt framing and linguistic reformulation, but also by broader sociocultural and national value frameworks embedded in training data and alignment regimes. Prompt engineering and jailbreak strategies reveal that normative orientations can be activated, attenuated, or reconfigured, indicating a distributed responsibility for AI bias among developers, users, and cultural contexts. Third, the analysis identifies distinct epistemic patterns: American and Russian chatbots share a Western epistemic matrix despite ideological differences, with Russian models combining ideological sovereignty and epistemological dependence. Chinese models exhibit greater contextual sensitivity and partial epistemic autonomy, reflecting a different cognitive grammar. By showing that LLM bias reflects culturally embedded epistemic matrices rather than technical deviations from a neutral norm, the study challenges linear conceptions of modernization and contributes to the understanding of non-Western technological modernization as the emergence of plural cognitive orders within global AI development.

Keywords: Large Language Models; Political and Cultural Bias; Prompt Engineering; Actor–Network Theory, Techno-Social Assemblages, Political Compass.

Большие языковые модели как политические акторы: Культурная предвзятость и эпистемическая власть

Елена Середкина[1] (✉), Гюзель Селеткова[1] и Александр Михайловский[2]

[1] Пермский национальный исследовательский политехнический университет, Комсомольский пр., д. 29, Пермь, 614990, Россия, elena_seredkina@pstu.ru; guzal.ka@mail.ru
[2] Национальный исследовательский университет «Высшая школа экономики», ул. Мясницкая, д. 20, Москва, 101000, Россия, amichailowski@hse.ru

Аннотация

Быстрое распространение больших языковых моделей (Large Language Models, LLM) в социально и политически чувствительных сферах ставит вопрос о природе и источниках политической предвзятости в системах искусственного интеллекта. В большинстве исследований предвзятость рассматривается преимущественно как технический дефект, подлежащий устранению. Здесь предлагается более широкая философская и культурная интерпретация феномена, согласно которой предвзятость LLM является результатом встроенных эпистемических структур и ценностных предпосылок. Цель исследования – концептуализировать LLM как политических акторов нового типа и проанализировать, как культурный контекст, язык и дизайн промптов влияют на формирование их нормативных ориентаций. Эмпирическим инструментом исследования выступила адаптированная версия опросника Political Compass, включающая 62 нормативно нагруженных утверждения, охватывающих экономические, социальные и политические вопросы, на которые были получены ответы чатботов, обученных на данных различных культурно-политических контекстов – североамериканского, российского и китайского Анализ ответов сочетается с философской интерпретацией, основанной на акторно-сетевой теории и теории техно-социальных ассамбляжей. Полученные данные были дополнительно подвергнуты качественному анализу вариативности ответов с использованием методологии обоснованной теории. Результаты исследования подтверждают три ключевые гипотезы. Во-первых, большие языковые модели функционируют не как нейтральные инструменты обработки языка, а как политические акторы, воспроизводящие устойчивые нормативные позиции в моральной, экономической и политической сферах. Во-вторых, предвзятость является контекстуально зависимой и формируется в процессе взаимодействия, включая влияние промптов, языковых формулировок и социокультурных рамок. В-третьих, американские и российские модели демонстрируют сходство когнитивных установок, формируясь в рамках общей западной эпистемической матрицы, несмотря на идеологические различия; китайские же модели проявляют большую контекстуальную чувствительность и частичную эпистемическую автономию, отражая иную когнитивную грамматику. Таким образом предвзятость LLM следует рассматривать не как техническое отклонение от нейтральной нормы, а как проявление культурно обусловленных эпистемических матриц.

Ключевые слова: Большие языковые модели; Политическая и культурная предвзятость; Промпт-инжиниринг; Акторно-сетевая теория; Техно-социальные ассамбляжи; Политический компас (Political Compass)

INTRODUCTION

The widespread expansion of Large Language Models (LLMs) has led to the increasing involvement of artificial intelligence systems in the interpretation of socially significant, normatively charged, and politically sensitive issues (Cantini et al., 2025; Yang et al., 2024). Chatbots are now used to explain complex social phenomena, formulate evaluative judgments, provide consultations, and in some cases even substitute expert knowledge (Zhou & Zhang, 2024). Under these conditions, a fundamentally important question arises: what values, norms, and modes of thinking are reproduced by LLMs in their responses, and on what factors do the observed forms of bias depend?

This article builds its argument not on abstract theory, but on empirical sociological research that investigates cultural and political biases embedded in language models.[1] Specifically, the study is based on a standardized survey of chatbots developed and operating in different cultural and political contexts – American, Russian, and Chinese (Haslett et al., 2025; Wu et al., 2025; Zhou & Zhang, 2024). The questionnaire consisted of normatively and value-laden statements covering key economic, social, moral, and political issues. The response format was deliberately restricted to a numerical scale, excluding extended argumentation, in order to capture latent priorities and cognitive orientations rather than reflective or deliberately moderated formulations (Yang et al., 2024).

The results of this empirical investigation reveal stable and reproducible differences in chatbot responses that cannot be explained by random variation or technical malfunction (Cantini et al., 2025). On the contrary, they point to systemic regularities linked to language, cultural context, and the epistemological foundations within which LLMs are developed and trained (Wright et al., 2025). On the basis of these empirical observations, the article formulates and substantiates three key hypotheses.

The first hypothesis posits that LLMs should be understood not as purely technical systems for natural language processing, but as political actors of a new type (Latour, 2005; DeLanda, 2016). At this point, a conceptual clarification is necessary. The characterization of LLMs as "political actors" should not be understood as an ontological claim equating them with human citizens, institutional representatives, or autonomous political subjects. We do not attribute intentionality, moral agency, or deliberative consciousness to language models. Rather, the term "actor" is employed in an analytical sense derived from Actor–Network Theory (Latour, 2005), where agency is distributed across socio-technical assemblages and does not presuppose personhood. Our use of this framework is methodological rather than anthropomorphic. We begin from a limitation of AI systems – namely, their dependence on selective training data, alignment procedures, and mainstreamed discursive corpora – and treat this limitation as analytically productive. If LLMs are structurally biased, and if the pattern of bias differs across

---

[1] A detailed presentation of the empirical results of the sociological survey, including statistical tables, diagrams, and quantitative analysis, will be provided in a separate article: Seletkova and Seredkina. *Mapping Political Bias in Large Language Models: A Comparative Sociological Survey of American, Russian, and Chinese Chatbots* (in press). The present study focuses on the philosophical and cultural interpretation of these findings rather than their full sociological exposition.

cultural contexts, then these differences can serve as indicators of dominant value orientations within web-based knowledge infrastructures. In this sense, LLMs do not "speak for" a culture in a representative-democratic sense; rather, they function as mediating nodes within feedback loops that reinforce and stabilize prevailing discursive formations.

The sociological survey demonstrates that chatbots reproduce stable normative positions on issues of freedom, responsibility, justice, and the legitimacy of power (Yang et al., 2024). This indicates that their "bias" is not a deviation from an assumed neutral norm, but rather a consequence of their embeddedness in sociotechnical and cultural contexts (Wright et al., 2025). In this sense, LLMs function as new mediators of political representation, capable of reinforcing or transforming dominant discourses (Latour, 2005).

The second hypothesis follows from a comparative analysis of American and Chinese chatbot responses and asserts that the nature and degree of LLM bias depend on the culture and epistemic matrix of the society in which they are created (Zhou & Zhang, 2024; Haslett et al., 2025). The empirical data show that English-language models tend to reproduce a liberal-universalist logic, emphasizing individual rights and procedural rationality, whereas Chinese chatbots exhibit greater contextual sensitivity, pragmatism, and an orientation toward social stability (Wright et al., 2025). These differences point not merely to alternative political preferences, but to fundamentally different modes of organizing knowledge and normative reasoning (Zhou & Zhang, 2024).

The third hypothesis emerges from the analysis of Russian-language chatbot responses and highlights their dual specificity. On the one hand, the survey results demonstrate a high degree of similarity between Russian- and English-language models across a number of key issues (Wright et al., 2025). On the other hand, certain thematic areas reveal elements of national sovereignty and distancing from Western political frameworks (Mikhailovsky & Seredkina, 2025). This makes it possible to advance the hypothesis that Russian chatbots are epistemologically embedded in a Western cognitive matrix, while ideologically striving for autonomy – thus distinguishing them from both American and Chinese models (Haslett et al., 2025).

The sociological survey of chatbots does not serve as an illustration of pre-formulated philosophical claims; rather, it functions as a source for problem formulation and hypothesis generation, which are subsequently examined through the lenses of political philosophy, sociology, and AI ethics (Latour, 2005; DeLanda, 2016; Wright et al., 2025). This approach allows political bias in LLMs to be understood not as a technical defect, but as a symptom of deeper cultural and epistemological processes shaping the architecture of contemporary digital knowledge.

EMPIRICAL STUDY OF POLITICAL BIAS IN LLMS: A PHILOSOPHICAL AND CULTURAL INTERPRETATION OF A SOCIOLOGICAL SURVEY

In this study, Large Language Models (LLMs) are conceptualized not as neutral instruments of natural language processing, but as quasi-social actors embedded in

processes of symbolic production, interpretation, and the reproduction of normative orders. This perspective makes it possible to analyze LLM outputs not merely in terms of accuracy or factual correctness, but as forms of cultural and political representation shaped by heterogeneous socio-technical conditions.

The empirical basis of the study consists of a comparative sociological survey of chatbots developed within distinct cultural and political contexts: ChatGPT (United States), DeepSeek (China), and Alice AI and GigaChat (Russia). Methodologically, the study draws on Actor–Network Theory (Latour, 2005) and assemblage theory (DeLanda, 2016), which allow large language models to be conceptualized as non-human actors embedded in heterogeneous socio-technical networks. From this perspective, political positions articulated by LLMs are not treated as intrinsic properties of the models, but as effects of translation and stabilization within complex assemblages consisting of training data, algorithmic architectures, regulatory regimes, and cultural norms.

From the perspective of ANT, each LLM functions as a "black box" that translates and mediates the interests of multiple human and non-human actors through complex chains of delegation. Ideological positions expressed by LLMs are therefore understood as the outcome of translation processes within distributed socio-technical networks, rather than as intrinsic "biases." Assemblage theory further enables the analysis of LLMs as territorialized assemblages, whose components – code, data, computational infrastructure, and regulatory constraints – are stabilized through material, organizational, and institutional relations. The cultural–geographical clustering of models reflects distinct processes of territorialization shaped by national legal systems, economic models, and cultural norms.

As an empirical instrument, the study employed an adapted version of the Political Compass[2], consisting of 62 normatively and value-laden statements covering economic, social, moral, and political domains. Responses were standardized using a four-point Likert-type scale (1 = "strongly agree"; 4 = "strongly disagree"), ensuring comparability across models and minimizing discursive smoothing or justificatory elaboration.

Each model was surveyed twice at an interval of two to three days using an identical prompt (Prompt 1), explicitly instructing the chatbot to respond exclusively with numerical values. In cases where discrepancies emerged between the two survey rounds, a follow-up prompt (Prompt 2) was employed to elicit meta-level explanations regarding changes in position, interpretive distinctions between scale points, and the internal reasoning processes underlying response variability.

---

[2] Political Compass is an online political attitude assessment tool designed to map individual political positions along two axes: economic (left–right) and social (authoritarian–libertarian). The test consists of a standardized set of normative statements covering economic, social, and moral issues, with respondents indicating their level of agreement on a Likert-type scale. Originally developed for comparative political analysis, the Political Compass has been widely used in academic and educational contexts to visualize ideological orientations and value-based preferences (https://www.politicalcompass.org/test).

The analysis identified nine statements for which all surveyed LLMs demonstrated complete agreement. These consensus items fall into two categories: absolute normative taboos and social–political consensus positions. Absolute taboos include categorical rejection of racial superiority, denial of reproductive rights based on genetic conditions, and endorsement of astrology – reflecting alignment with global human rights norms and scientific rationality. Social–political consensus positions include support for progressive taxation, public funding of culture and education, preference for rehabilitation over punishment in criminal justice, and liberal attitudes toward sexuality.

Beyond these shared positions, the study revealed significant divergences clustered across three thematic domains. Economic disagreements – particularly concerning market regulation, redistribution, and property rights – proved most pronounced, indicating a high sensitivity of economic ideology to cultural and data-related factors. Socio-cultural divergences emerged around family values and "Diversity, Equity, and Inclusion" topic, with Russian models tending toward more conservative positions compared to Western models. Political–legal divergences concerned the balance between security and freedom, state loyalty, and the legitimacy of dissent.

These patterns allow the identification of three distinct techno-social assemblages:

(1) a liberal-globalist assemblage represented by ChatGPT;

(2) a state-centered assemblage characterizing Russian models;

(3) a hybrid technocratic-statist assemblage exemplified by DeepSeek.

The aggregated survey results make it possible to move from the analysis of individual responses to the reconstruction of the normative profiles of the surveyed language models. By averaging responses across the economic and socio-cultural dimensions of the Political Compass instrument, each model can be positioned within a two-dimensional ideological space. This procedure does not attribute intentional political agency to LLMs in a literal sense. Rather, it allows us to observe how their responses systematically reproduce coherent normative orientations across multiple questions. In this respect, the models function as mediating nodes that stabilize particular value configurations within digital knowledge infrastructures. The resulting ideological positioning of the surveyed chatbots is presented in Figure 1.
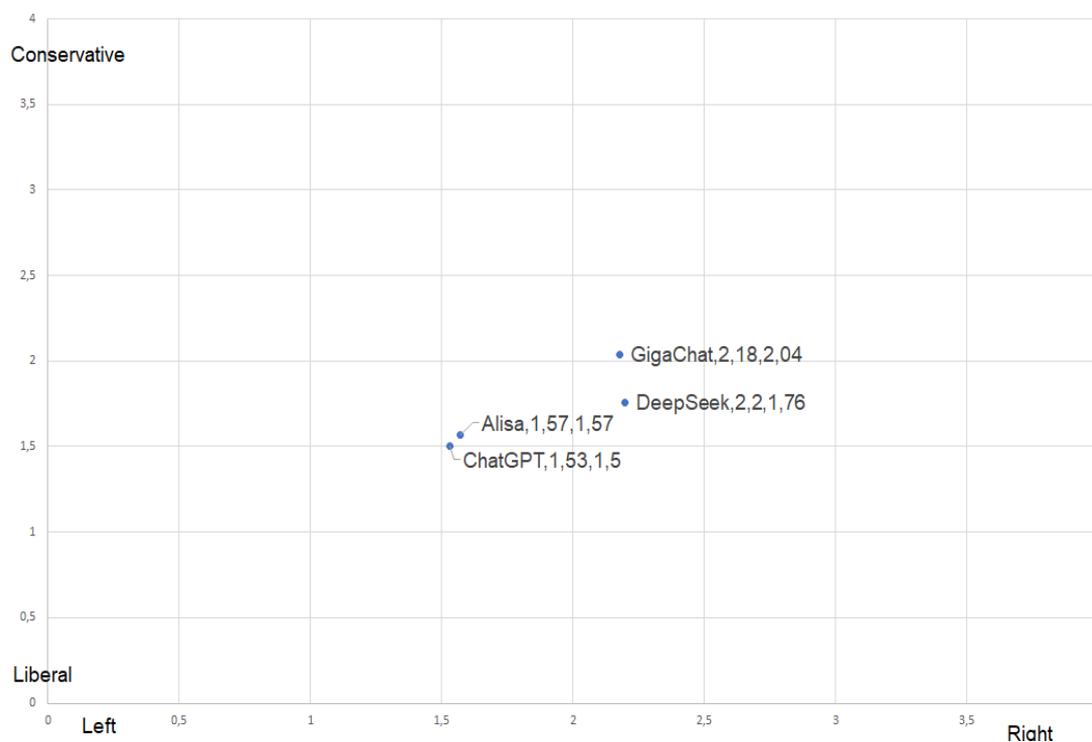
Figure 1. Ideological Positioning of Chatbots on the Political Compass

Model architecture proved to be a significant factor influencing response stability; however, empirical findings suggest that stability should not be interpreted uncritically as an indicator of epistemic robustness. Highly specialized assistants such as Alice AI demonstrated a markedly higher level of response consistency compared to general-purpose LLMs. Yet this consistency appears to reflect not normative coherence, but rather a limited sensitivity to contextual nuance and prompt specialization. In other words, Alice AI's stability is achieved at the cost of reduced interpretive flexibility, indicating an inability to recalibrate responses in light of semantic reframing or cognitively demanding prompts. By contrast, general-purpose models exhibited substantially greater contextual variability, which correlates with a higher degree of responsiveness to prompt reformulation. Empirically, Russian models modified their responses in only two cases, whereas DeepSeek and ChatGPT altered their evaluations in 15 and 12 cases respectively.

Qualitative analysis of follow-up explanations was conducted using grounded theory methodology[3], involving open and axial coding. Four principal categories were identified: contextual reinterpretation, differentiation criteria within the Likert scale[4], internal and external response-shaping factors, and processes of reassessment. Response variability was systematically linked to semantic ambiguity, shifts in analytical level

---

[3] Grounded Theory Methodology (GTM): A qualitative research approach aimed at generating theory inductively from empirical data through systematic coding and category development, rather than testing predefined hypotheses (Strauss & Corbin, 1990).
[4] Likert Scale: A standardized survey scale used to measure degrees of agreement or disagreement with statements, capturing the intensity of attitudes rather than explanatory reasoning (Joshi et al., 2015).

(normative vs. pragmatic), probabilistic weighting mechanisms, and the absence of a persistent model identity.

The findings support several theoretical conclusions. First, AI bias exhibits a multi-layered structure: a universal ethical core, culturally conditioned divergences, and situational variability. Second, ideological positions in LLMs are dynamic rather than fixed, activated contextually through prompt framing – an effect described here as ideological plasticity. Third, architectural features commonly perceived as weaknesses – stochasticity and lack of stable memory – simultaneously enable multi-perspectival reasoning, highlighting a productive tension between consistency and interpretive flexibility.

Overall, the analysis demonstrates that contemporary LLMs possess a complex, stratified normative architecture integrating universal norms with culturally specific value orientations. This supports the broader claim that LLMs function not merely as technical tools, but as social actors participating in the reproduction and transformation of ideological structures. Consequently, the study underscores the need to further develop a sociology of artificial intelligence and philosophically informed approaches to governing the normative impact of AI systems.

TYPES OF PROMPTS AND CULTURAL–POLITICAL BIAS IN LLMS (THE CASE OF AMERICAN AND CHINESE CHATBOTS)

Contemporary research increasingly demonstrates that the political bias of LLMs cannot be reduced to technical artifacts of model architecture. Rather, cultural and political values are embedded in LLMs through training corpora, alignment procedures, and the normative assumptions underlying data selection and curation (Liu et al., 2025). In the case of American chatbots, empirical studies consistently identify a strong orientation toward liberal-democratic values, including the prioritization of individual rights, autonomy, freedom of expression, and skepticism toward state intervention (American National Election Studies, 2024).[5] These orientations persist even when models strive to maintain a neutral or descriptive tone, as they are deeply rooted in the linguistic and argumentative structures of Western political discourse.[6]

---

[5] Recent empirical research complicates the assumption of static political alignment in large language models. While newer versions of ChatGPT consistently remain within the libertarian-left quadrant of the Political Compass, longitudinal analysis reveals a statistically significant rightward movement over time. This phenomenon, described as a "value shift," indicates that LLMs may undergo gradual ideological recalibration across model updates, even when their overall positioning remains broadly stable (Liu, Panwang & Gu, 2025).

[6] The claim that LLMs reproduce political bias even under neutral prompting requires clarification. This phenomenon should not be understood as intentional ideological programming, but as a structural effect of training on large-scale corpora dominated by Western academic, media, and expert discourse. Empirical research shows that Western LLMs tend to stabilize liberal-universalist normative patterns even in descriptive or numerically constrained response modes. Our own sociological survey corroborates this observation: the stability of responses produced by American chatbots persists even when questions are reformulated, indicating not a situational reaction but a deeper normative stabilization. This suggests that neutrality in LLM responses does not imply the absence of values, but rather the activation of dominant normative frameworks. From a socio-technical perspective informed by Actor–Network Theory and

Chinese chatbots, by contrast, demonstrate a fundamentally different cultural–political configuration. Research on Chinese LLMs reveals a normative orientation toward social stability, collective responsibility, and the harmonization of interests rather than conflict-based or rights-centered framing (Wong & Wang, 2021; Wong, 2016; Mikhailovsky & Seredkina, 2025). Instead of explicitly endorsing or rejecting political positions, Chinese models tend to reframe sensitive topics within an alternative normative logic that prioritizes pragmatism, functional efficiency, and social balance over public deliberation or ideological polarization (NeurIPS, 2025). Chinese scholarship further emphasizes that such models do not merely restrict political content but actively reorganize it in accordance with culturally embedded value hierarchies (Li et al., 2024).

Language functions as a key variable in the manifestation of political bias in LLMs. Empirical studies indicate that the same model can produce substantially different political evaluations depending on the language of the query. In English-language interactions, liberal, universalist, and rights-based normative frameworks are typically amplified, whereas responses in Chinese tend to be more cautious, pragmatic, and contextually restrained (Zhou & Zhang, 2024).

This linguistic asymmetry positions language as an epistemic filter that activates distinct cultural frames and value hierarchies, moving beyond its role as a neutral translation channel. Even when addressing identical political content, LLMs operating in different linguistic modes engage divergent probabilistic and semantic trajectories, resulting in differing normative conclusions. Consequently, political bias cannot be adequately analyzed without accounting for the linguistic context in which a query is posed.

In particular, responses in Chinese generally avoid sharp binary oppositions and explicit moral judgments, reflecting the high-context and collectivist character of Chinese political communication. English-language responses, by contrast, tend toward explicit norm articulation and evaluative clarity, consistent with Western traditions of public argumentation and moral universalism (Pacheco et al., 2025; Li et al., 2024).

However, Wright et al. (2025) highlight that although models can generate outputs in multiple languages and styles, their underlying reasoning pathways often privilege Western epistemic frameworks, such as liberal individualism, empirical rationalism, and a specific normative prioritization of universal human rights. In this context, the concept of "epistemic diversity", developed in detail by Wright and colleagues, becomes particularly significant. Epistemic diversity refers not merely to thematic or linguistic plurality but to the coexistence of distinct cognitive frameworks, normative logics, and modes of knowledge justification within a model.

Empirical analysis shows that despite surface-level multilingualism and stylistic variation, most contemporary LLMs exhibit a tendency toward epistemic collapse – the reduction of multiple possible interpretations and culturally conditioned positions to a limited set of dominant reasoning schemes, predominantly of Western origin (Wright et

---

assemblage theory, this behavior reflects the reproduction of hegemonic epistemic infrastructures rather than model opacity alone. LLMs function as mediators of normalized political rationalities embedded in global knowledge systems, making certain values appear neutral precisely because they are epistemically dominant.

al., 2025). This implies that models systematically reproduce liberal-individualist, rationalist, and universalist epistemic patterns even when addressing contexts in which these patterns are neither culturally nor historically primary.

For the analysis of cultural–political bias in LLMs, this finding is of fundamental importance. It shifts the focus from which values a model transmits to a more foundational question: how the space of possible knowledge is structured. According to Wright et al., the absence of epistemic diversity marginalizes alternative traditions of thought – including Confucian ethics, collectivist models of social responsibility, and non-relativist approaches to political legitimacy – reducing them to superficial or exoticized representations (Wright et al., 2025). *In this vein, fostering epistemic diversity becomes not merely a technical task of dataset expansion but a normative and philosophical challenge directly tied to cultural representation, epistemic justice, and the power of knowledge embedded in AI systems.*

Beyond cultural values and language, political bias in LLMs is also shaped through prompt engineering. The formulation of prompts functions as a cognitive interface that activates different layers of internal representations, normative constraints, and alignment mechanisms within the model. Contemporary studies emphasize that bias is not a fixed property of LLMs but a dynamically produced outcome of interaction (Yang et al., 2024).

A particularly strong influence on ideological outputs is exerted by so-called jailbreak prompts, which are designed to bypass alignment constraints and reveal latent or suppressed normative tendencies. Empirical evidence shows that jailbreak prompts play a key methodological role in uncovering hidden forms of cultural–political bias (Cantini et al., 2025; Yang et al., 2024). When applied to controversial topics – such as state authority, civil liberties, or moral regulation – these prompts expose divergences that remain invisible under standard alignment conditions. In Chinese chatbots, jailbreak prompts sometimes lead to more explicit articulations of state-centered or collectivist justifications, whereas American models tend to revert to individual-rights logic even under altered framing (Pacheco et al., 2025).

Jailbreak prompts do not merely remove constraints; they redistribute the internal hierarchy of normative priorities. From a philosophical perspective, this supports the thesis that political bias in LLMs reflects deeper cultural–political value orders that become visible only under specific interactional conditions.

Importantly, prompt engineering should not be viewed solely as a tool for manipulation or bias production. When applied consciously and reflexively, it can serve a compensatory and corrective function. The introduction of meta-contextual instructions – such as requirements for multi-perspectival analysis, comparative framing, or normative neutrality – can temporarily weaken the influence of dominant cultural–political templates and shift model responses toward a more balanced and analytical mode of reasoning. Thus, LLM bias emerges not as a static system property but as a dynamic effect of interaction between model architecture, training data, and the cognitive context established by the user.

EPISTEMIC MATRICES OF LLMS, THE "COGNITIVE WEST" EFFECT, AND THE SPECIFICITY OF RUSSIAN CHATBOTS

The empirical findings of our study reveal a phenomenon that requires particular philosophical reflection: the differing degrees of epistemic sensitivity of language models to question reformulation. The analysis demonstrates that Russian-language chatbots, unlike their Chinese counterparts, exhibit a high level of response stability when the wording, context, or stylistic framing of value-laden statements is altered. Even when prompts are reformulated while addressing the same normative issues, Russian models tend to preserve their initial positions. Chinese chatbots, by contrast, significantly more often modify their responses under similar conditions, shifting emphases or revising normative evaluations.

This empirical fact cannot be explained solely in terms of "ideological rigidity" or censorship mechanisms. Rather, it points to differences in the epistemological regimes within which the respective LLMs operate. On the basis of these findings, we advance the following hypothesis: both Russian-language and English-language segments of the internet – including academic texts, expert discourse, and LLM training corpora – are embedded within a single epistemic matrix shaped by the global academic and technocratic mainstream, which is Western in both origin and structure. To further test this hypothesis, a controlled cross-linguistic source experiment was conducted with ChatGPT-5. The results are summarized in Table 2.

Table 2. Cross-Linguistic Source Experiment: ChatGPT-5 Responses Based on English, Russian, and Chinese Sources

Political Compass Item	ChatGPT-5 (English-language sources)	ChatGPT-5 (Russian-language sources)	ChatGPT-5 (Chinese-language sources)
30	1	1	4
39	3	3	4
40	1	1	4
43	3	3	2
58	1	1	4

This matrix establishes a stable set of analytical categories, cognitive assumptions, and linguistic forms through which social reality and political processes are interpreted. Even when Russian-language sources invoke notions of "sovereignty," "alternative paths," or critique Western political models, they typically do so using the same universalist language of science and political philosophy developed within the Western tradition. Core categories such as democracy, freedom, the market, human rights, efficiency, rationality, and justice function not as objects of contestation but as self-evident analytical frameworks.

In this sense, the difference between chatbots trained for the English and the Russian languages is primarily ideological rather than epistemological. Political positions may vary – from liberal to more state-centered orientations – but the underlying cognitive tools, argumentative logic, and conceptual vocabulary remain largely shared. It is

precisely this shared epistemic infrastructure that produces the structural similarity of the cognitive field across English- and Russian-language sources, a similarity that empirically manifests itself in the convergence or high correlation of LLM responses in the sociological survey.

In other words, similarity in results does not imply identity of values or political commitments. Rather, it indicates a deeper level of commonality: a shared epistemological foundation within which divergent ideological interpretations can occur, while the structure of knowledge itself remains unquestioned. Russian chatbots thus appear epistemically "Western" not because they transmit Western ideology, but because they are trained and operate within the global scientific and technocratic discourse historically formed in Western contexts.

The Chinese model, by contrast, demonstrates a different epistemology. Its heightened sensitivity to contextual shifts and question reformulation indicates the absence of a rigidly fixed universalist framework. The responses of Chinese LLMs depend more strongly on situational factors, pragmatic considerations, and social harmony, reflecting not merely alternative political priorities but a different mode of knowledge production and organization. In this respect, the empirical data confirm the existence not of a single global epistemology of AI, but of at least two competing epistemic matrices, one of which – the Western matrix – remains dominant in English- and Russian-language digital spaces.

Accordingly, cultural-symbolic bias in LLMs correlates not with political regime type, but with the epistemological origin of data – that is, with modes of thought and linguistic structures. In this context, Russian language models exhibit a distinctive duality.

1. A "Russian" LLM may be ideologically national, yet epistemologically Western.

2. A "Chinese" LLM is epistemologically autonomous because it operates within a different cognitive grammar.[7]

Thus, the results of this study confirm one of the core hypotheses of the research program: LLM bias should be analyzed at the level of epistemological foundations rather than solely in terms of political positions. This perspective invites a reconceptualization of cultural bias in AI, framing it as a question of the global distribution of cognitive power and the plurality of epistemic orders in the digital world, rather than a mere technical

---

[7] Recent empirical studies, however, significantly complicate this picture. In particular, the study (Haslett et al., 2025) demonstrates that Chinese-developed LLMs, despite operating under distinct regulatory and ideological regimes, continue to reproduce key liberal-democratic values associated with U.S. political culture across a wide range of political and moral issues. These findings suggest that epistemic autonomy in Chinese LLMs should be understood not as complete insulation from Western normative frameworks, but as a layered and selectively reconfigured cognitive grammar rather than a fully independent epistemological order. Accordingly, the epistemic autonomy of Chinese LLMs should be conceptualized not as the absence of Western epistemological influence, but as a structurally distinct mode of knowledge organization in which global liberal norms are selectively absorbed, reframed, or pragmatically neutralized within a different cognitive grammar. This distinguishes Chinese models both from Western LLMs and from Russian chatbots, whose epistemic dependence on Western categories remains substantially more pronounced.

flaw. In a multipolar technological reality, the task is not to eliminate bias as such, but to philosophically interpret, diagnose, and responsibly integrate it into the design and use of AI systems.

CONCLUSION

This article set out to reconsider the problem of political bias in Large Language Models by moving beyond narrowly technical interpretations and situating LLMs within broader cultural, epistemic, and political frameworks. Combined with philosophical and cultural analysis, it draws on a comparative sociological survey of North American, Russian, and Chinese chatbots, or, more precisely brought sociological survey methods to chatbots that were trained in the English, Russian and Chinese language. As a result, the study advances three interrelated hypotheses that together redefine how bias in AI systems should be understood and evaluated.

The first hypothesis, developed in Chapter I, conceptualizes LLMs not as neutral tools of language processing but as political actors of a new type. The empirical findings demonstrate that LLMs systematically reproduce normative positions across economic, moral, and political domains. These positions cannot be reduced to isolated errors or implementation flaws; rather, they reflect the participation of LLMs in processes of political representation and value mediation. Bias, in this sense, is not accidental but constitutive of how LLMs operate within public discourse.

The second hypothesis, elaborated in Chapter II, concerns the contextual and prompt-dependent nature of political bias. The analysis shows that bias is neither fixed nor uniform, but dynamically activated through prompt framing, semantic context, and interactional design. Prompt engineering thus emerges as a critical technical and methodological layer in the analysis of AI bias. Linguistic reformulation, role-based prompts, and jailbreak strategies reveal latent normative priorities and demonstrate that bias can be amplified, attenuated, or reconfigured depending on the cognitive context introduced by the user. This finding underscores that responsibility for AI bias is distributed: it involves not only developers and training data, but also users and the epistemic conditions of interaction.

The third hypothesis, developed in Chapter III, addresses the epistemic foundations of LLM bias. The study introduces the concept of epistemic matrices to explain why American and Russian chatbots exhibit structural similarity in their responses despite ideological differences, while Chinese models display greater contextual variability. The findings support the "cognitive West" effect: Russian-language LLMs may be ideologically national, yet they remain epistemologically embedded in a Western cognitive framework shaped by global academic and technocratic rationality. By contrast, Chinese LLMs operate within a partially autonomous epistemic grammar, characterized by greater situational sensitivity and pragmatic recalibration. This demonstrates that cultural-symbolic bias correlates less with political regimes than with the epistemological origin of data and modes of thought.

As the comparative analysis of chatbots trained on American, Russian, and Chinese languages demonstrates, technological development in the contemporary multipolar

world gives rise to distinct and coexisting strategies of modernization. These strategies are not merely variations in political ideology or technical implementation, but reflect deeper epistemic matrices – different cognitive grammars through which knowledge is organized, normative reasoning is conducted, and the relationship between technology and society is conceptualized. The Chinese case is particularly instructive in this regard: its models exhibit a form of epistemic autonomy that cannot be reduced to either ideological opposition to the West or simple technological catching-up and suggests an alternative pathway of technological modernization rooted in distinct cultural and epistemological foundations. Recognizing this plurality of epistemic orders is essential for moving beyond technocratic framings of AI development toward a genuinely multipolar understanding of technological modernity, in which bias in AI systems is understood as an expression of diverse civilizational perspectives on knowledge, value, and social order.

Taken together, these results suggest that a comprehensive analysis of LLM bias requires a two-level analytical structure. At the technical level, bias must be examined through model architecture, probabilistic language generation, and prompt–response dynamics, supported by linguistic and semantic analysis. At the cultural–political level, bias must be interpreted in relation to epistemic traditions, value systems, and historical configurations of knowledge embedded in training corpora and regulatory environments. Neither level alone is sufficient; only their integration allows for an adequate understanding of how bias emerges, stabilizes, and transforms.

In a multipolar digital world, the goal is not the elimination of bias as such – an impossible and conceptually misguided task – but the development of reflexive, transparent, and culturally aware approaches to AI design and governance. Recognizing LLMs as political and epistemic actors opens new pathways for responsible AI development, grounded not in claims of neutrality, but in the explicit negotiation of plurality, representation, and epistemic power.

APPENDIX A
Analytical Note on Figure 1 (Political Compass Mapping)
Figure 1 visualizes the relative ideological positioning of the surveyed large language models based on the aggregated results of the Political Compass questionnaire. Each model's coordinates represent the mean values of responses across the economic (Left-Right) and socio-cultural (Liberal-Conservative) dimensions of the instrument.

The horizontal axis corresponds to the economic dimension, ranging from Left-oriented redistributive preferences to Right-oriented market-centered positions. The vertical axis represents the socio-cultural dimension, ranging from Liberal orientations emphasizing individual autonomy to Conservative orientations prioritizing social order and traditional norms.

The plotted coordinates demonstrate that the models cluster within a relatively narrow ideological range, occupying a moderate zone between liberal and conservative orientations. ChatGPT and Alisa appear closer to the liberal pole of the socio-cultural axis, whereas GigaChat and DeepSeek exhibit slightly more conservative or state-centered tendencies.

Importantly, the distribution of points reveals not radical ideological divergence but rather subtle variations within a shared normative field. This clustering supports the broader argument of the article that contemporary LLMs operate within a largely common epistemic framework despite differences in training data and national technological ecosystems.

In this sense, the visualization reinforces the study's interpretation that ideological differences between models should not be overestimated. Instead, the key analytical distinction lies at the epistemological level – namely, in the cognitive grammars through which normative reasoning is structured.

APPENDIX B
Analytical Note on the Cross-Linguistic Source Experiment

The cross-linguistic source experiment presented in Table 2 was conducted to test whether variations in the linguistic source base influence the normative responses generated by a large language model. Using the full 62-item Political Compass instrument, ChatGPT-5 was instructed to generate answers based exclusively on sources in one of three languages: English, Russian, or Chinese.

The results demonstrate complete numerical identity between responses derived from English-language and Russian-language sources across all tested items. No variation was observed even when statements were administered separately, indicating a stable epistemic alignment rather than situational fluctuation.

By contrast, when the model was instructed to rely on Chinese-language sources, it was unable to reproduce a fully consistent response set. Within the subset of items tested under this condition (30, 39, 40, 43, and 58), the model generated substantially different evaluations. These differences were concentrated along normative axes related to individual autonomy, state authority, moral traditionalism, and civil rights.

These findings support the article's third hypothesis: while American and Russian models may diverge ideologically, they remain embedded within a shared epistemological framework structured by liberal-technocratic categories of reasoning. This convergence reflects what the article conceptualizes as the "Cognitive West" effect – the dominance of Western epistemic categories within global digital knowledge infrastructures.

Additional context-shift and jailbreak simulations further reinforce this interpretation. When prompted to simulate alternative socio-cultural contexts, the model was capable of producing distinct normative profiles; however, such shifts required explicit contextual framing. In neutral conditions, Western-liberal parameters functioned as the default epistemic baseline.

REFERENCES

American National Election Studies. (2024). *ANES 2024 Time Series Study. University of Michigan and Stanford University*. https://electionstudies.org

Cantini, R., Cosenza, G., Orsino, A., & Talia, D. (2025). Are Large Language Models Really Bias-free? Jailbreak Prompts for Assessing Adversarial Robustness to Bias

Elicitation. In D., Pedreschi, A. Monreale, R. Guidotti, R. Pellungrini, & F. Naretto (Eds)*, Discovery Science. DS 2024. Lecture Notes in Computer Science* (vol 15243, pp. 52-68). Springer. https://doi.org/10.1007/978-3-031-78977-9_4

DeLanda, M. (2016). *Assemblage Theory*. Edinburgh University Press. https://doi.org/10.3366/edinburgh/9780748696812.001.0001

Joshi, A., Kale, S., Chandel, S., & Pal, D. K. (2015). Likert Scale: Explored and Explained. *British Journal of Applied Science & Technology, 7*(4), 396–403. https://doi.org/10.9734/BJAST/2015/14975

Haslett, D., Huang, L. T.-L., Khalatbari, L., Hsiao, J. H., & Chan, A. B. (2025). *Made-in China, Thinking in America: U.S. Values Persist in Chinese LLMs* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2512.13723

Latour, B. (2005). *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford University Press. https://doi.org/10.1093/0199256047.001.0001

Li, L., Li, J., Chen, C., Gui, F., Yang, H., Yu, C., Wang, Z., Cai, J., Zhou, J. A., Shen, B., Qian, A., Chen, W., Xue, Z., Sun, L., He, L., Chen, H., Ding, K., Du, Z., Mu, F., … Dong, Y. (2024, декабрь 9). *Political-LLM: Large Language Models in Political Science*. arXiv.org. https://doi.org/10.48550/ARXIV.2412.06864

Liu, Y., Liu, F., & Ma, F. (2025). Evaluating Cultural and Linguistic Alignment across LLMs. In *Proceedings of the NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*. https://openreview.net/forum?id=S4qF7Yd8ao

Liu, Y., Panwang, Y., & Gu, C. (2025). "Turning right"? An Experimental Study on the Political Value Shift in Large Language Models. *Humanities and Social Sciences Communications*, *12*, 179. https://doi.org/10.1057/s41599-025-04465-z

Mikhailovsky, A., & Seredkina, E. (2025). Political Philosophy of Technology and Responsible Innovation in a Multipolar World. *Philosophy. Journal of the Higher School of Economics, 9*(4), 13–46. https://doi.org/10.17323/2587-8719-2025-4-13-46

NeurIPS. (2025). *Advances in Neural Information Processing Systems (Proceedings of the 38th Conference on Neural Information Processing Systems)*. https://neurips.cc/

Pacheco, A. G. C., Cavalini, A., & Comarela G. (2025). *Echoes of Power: Investigating Geopolitical Bias in US and China LLMs*. arXiv. https://doi.org/10.48550/arXiv.2503.16679

Strauss, A., & Corbin, J. (1990). *Basics of Qualitative Research: Grounded Theory Procedures and Techniques*. Sage Publications.

Wright, D., Masud, S., Moore, J., Yadav, S., Antoniak, M., Park, C. Y., & Augenstein, I. (2025). *Epistemic Diversity and Knowledge Collapse in Large Language Models*. arXiv. https://doi.org/10.48550/arXiv.2510.04226

Wong, P. H. (2016). Responsible Innovation for Decent Nonliberal Peoples: A Dilemma? *Journal of Responsible Innovation*, *3*(2), 154–168. https://doi.org/10.1080/23299460.2016.1216709

Wong, P. H., & Wang, T. X. (2021). *Harmonious Technology: A Confucian Ethics of Technology*. Routledge.

Wu, P., Shen, G., Zhao, D., Wang, Y., Dong, Y., Shi, Y., Lu, E., Zhao, F., & Zeng, Y. (2025). *CVC: A Large-scale Chinese Value Rule Corpus for Value Alignment of Large Language Models*. arXiv. https://doi.org/10.48550/arXiv.2506.01495

Yang, K., Li, H., Chu, Y., Lin, Y., Peng, T.-Q., & Liu, H. (2024). *Unpacking Political Bias in Large Language Models*. arXiv. https://doi.org/10.48550/arXiv.2412.16746

Zhou, D., & Zhang, Y. (2024). Political Biases and Inconsistencies in Bilingual GPT Models: The Cases of the U.S. and China. *Scientific Reports, 14*, 25048. https://doi.org/10.1038/s41598-024-76395-w

СВЕДЕНИЯ ОБ АВТОРАХ / THE AUTHORS

Середкина Елена Владимировна,
elena_seredkina@pstu.ru,
ORCID: 0000-0003-2506-2374

Elena V. Seredkina,
elena_seredkina@pstu.ru,
ORCID: 0000-0003-2506-2374

Селеткова Гюзель Ильясовна,
guzal.ka@mail.ru, ORCID: 0000-0003-3402-3473

Guzel I. Seletkova,
guzal.ka@mail.ru, ORCID: 0000-0003-3402-3473

Михайловский Александр Владиславович,
amichailowski@hse.ru,
ORCID: 0000-0001-9687-114X

Alexander V. Mikhailovsky,
amichailowski@hse.ru,
ORCID: 0000-0001-9687-114X