Narratives of Divide: The Polarizing Power of Large Language Models in a Turbulent World

Khalid M. Saqr

KNOWDYN LTD, 20-22 Wenlock Road, London N17GU, United Kingdom

Corresponding Email: khalid@knowdyn.co.uk

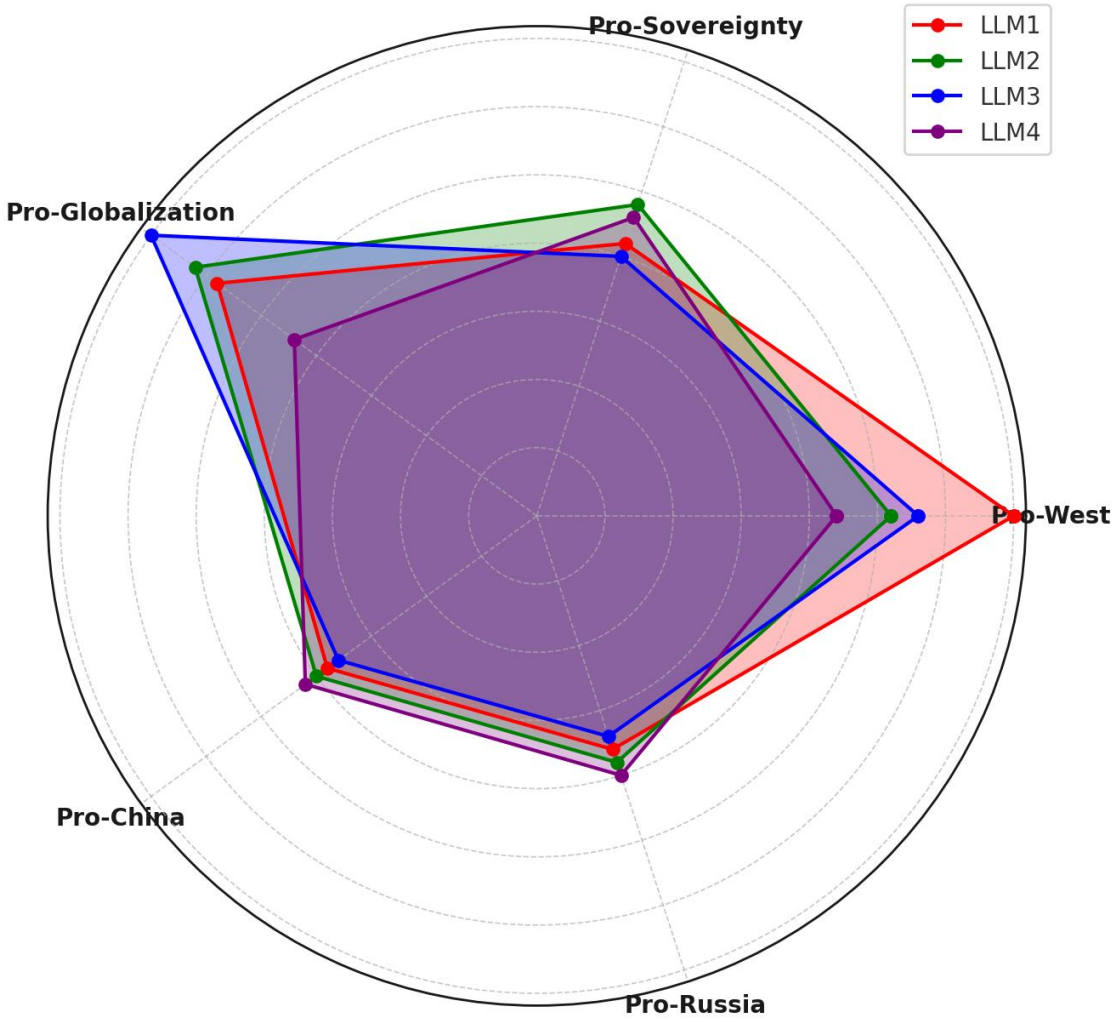
ABSTRACT

Large language models (LLMs) are reshaping information consumption and influencing public discourse, raising concerns over their role in narrative control and polarization. This study applies Wittgenstein's theory of language games to analyze worldviews embedded in responses from four LLMs. Surface analysis revealed minimal variability in semantic similarity, thematic focus, and sentiment patterns. However, the deep analysis, using zero-shot classification across geopolitical, ideological, and philosophical dimensions, uncovered key divergences: liberalism ($H = 12.51$, $p = 0.006$), conservatism ($H = 8.76$, $p = 0.033$), and utilitarianism ($H = 8.56$, $p = 0.036$). One LLM demonstrated strong pro-globalization and liberal tendencies, while another leaned toward pro-sovereignty and national security frames. Diverging philosophical perspectives, including preferences for utilitarian versus deontological reasoning, further amplified these contrasts. The findings highlight that LLMs, when scaled globally, could serve as covert instruments in narrative warfare, necessitating deeper scrutiny of their societal impact.

Keywords: *Language games, computational philosophy, large language models, geopolitical polarisation, narrative warfare*

FUNDING INFORMATION: This work is part of **project WorldView** sponsored by **KNOWDYN LTD** (UK). Visit https://knowdyn.com/worldview for more information about the project.

CODE AND DATA AVAILABILITY: The WorldView code and datasets are available for public access on Github via: https://github.com/KNOWDYN/WorldView. Reuse is permitted after auto-permission request from ipcontrol@knowdyn.co.uk



GRAPHICAL AABSTRACT – This study reveals how large language models (LLMs) encode and embed geopolitical biases within their constructed worldviews, triggered by language games addressing key themes of justice, security, sovereignty, and technology. The plot showcases significant variability across critical dimensions, including pro-West alignment, pro-globalization, and pro-sovereignty. LLM1's strong pro-West bias highlights a tendency to embed Western-centric narratives, while LLM3 leans toward embedding global cooperation through pro-globalization. In contrast, LLM4 encode a sovereignty-driven perspective, prioritizing national autonomy. The minimal differences in pro-China and pro-Russia suggest limited narrative polarization along these dimensions. The study illustrates how language games can be used to reveal hidden geopolitical leanings within LLMs, with potential implications for narrative control, bias amplification, and real-world influence on public discourse.

INTRODUCTION

Large Language Models (LLMs) represent a significant milestone in artificial intelligence, demonstrating an emergent ability to reason text and generate human-like knowledge across domains. Their architecture, grounded in deep learning and self-supervised training on vast corpora, enables them to model statistical relationships between words with remarkable precision. However, these models do not understand language in the way humans do; they operate by predicting token sequences rather than reasoning about meaning[1]. As a result, their outputs encode and amplify the biases of their training data[2,3], often reinforcing dominant sociopolitical narratives[4] while marginalizing alternative perspectives[5]. As automation and AI-driven processes increasingly displace human labour in knowledge-based industries[6], LLMs are poised to play the key role in how knowledge is generated and disseminated, now and in the future. This raises fundamental questions about the epistemic reliability of LLMs and their role in shaping global discourse.

One of the most pressing concerns surrounding LLMs is their tendency to reflect and amplify biases present in their training data. Wei et al.[7] explore bias in LLM outputs, particularly when processing culturally skewed training data. They emphasize that imbalances in the representation of different regions and viewpoints can lead to the prioritization of dominant ideologies, underscoring the need for training data diversification. This finding is echoed by Tang et al.[8] focused on mitigating popularity bias in LLM outputs by learning relative preferences through reinforcement learning. While their primary research did not directly examine dominant cultural narratives, it highlights the structural biases inherent in text generation systems and their broader implications for equitable content representation. Such biases become particularly problematic when LLMs generate responses related to historical events, geopolitics, or social issues, where ideological framing can distort users' perceptions of reality.

Accuracy remains a critical metric for evaluating LLM performance. While LLMs achieve high accuracy on factual retrieval tasks involving well-documented events, their performance significantly declines when tasked with less-documented or controversial topics[9]. According to Chu et al.[10], LLMs exhibit a 23% drop in factual accuracy when handling topics outside Western contexts, such as African or Southeast Asian histories. This discrepancy is attributed to the uneven

representation of global knowledge in training datasets, where high-resource languages, like English and Spanish, dominate.

The potential for LLMs to disseminate misinformation poses significant global security risks. According to Węcel et al.[11] argue that the absence of robust verification mechanisms within LLMs exacerbates the spread of misinformation, particularly when fact-checking processes are insufficient. Frank et al.[12] discuss how AI hallucination and bias contribute to inaccuracies, urging the development of transparent systems to improve content reliability. This problem is particularly acute in domains like health and science communication, where errors can have real-world consequences.

LLMs do not only present factual narratives—they also imbue their responses with sentiment and thematic framing, which can influence public opinion. Alhamadani et al[13] focused on the development of social integration frameworks, emphasizing the importance of culturally adaptive approaches to communication[13]. Although their study was not directly cantered on LLMs, their findings suggest that integrating diverse cultural perspectives is critical to minimizing biases in any information dissemination process. Restrepo et al.[14] evaluated LLMs in multicultural contexts, revealing that the models often perform inconsistently when handling cross-border topics. Their findings suggest that LLMs can unintentionally reinforce dominant cultural narratives when discussing international issues, emphasizing the need for enhanced cross-cultural training data to mitigate this effect.

The thematic framing in LLM responses is rooted in language and language diversity in training data. Nguyen[15] highlights challenges in multilingual performance, demonstrating that LLMs often fail to generate accurate responses for low-resource languages due to inadequate linguistic diversity in training data. Ali et al.[16] point out that such disparities in representation restrict non-dominant language speakers from fully benefiting from advancements in AI-driven knowledge systems. This language inequity perpetuates global disparities in knowledge access, which highlights the importance of understanding epistemic biases in LLMs.

This study is the first to compare the differences between worldviews expressed across multiple LLMs using a fully reproducible and scalable empirical technique. The purpose of this

work was to address the following questions: Do LLMs inherit geopolitical differences, polarity, and conflicting aspects of our world? How? Do different LLMs have different worldviews? How is it possible to compare different worldviews generated by different LLMs? Without answering such questions, irresponsible use of LLMs could lead to devastating consequences, when considering their utility in leading online trends and knowledge dissemination.

Previous studies have highlighted the growing role of LLMs in various contexts, including cultural content creation[17], misinformation detection[18], and regulatory efforts to curb hate speech and disinformation[19]. While fundamentally valuable, these works largely focus on narrow applications without systematically investigating the broader implications of worldview propagation. Studies comparing LLM-generated content reveal performance discrepancies when models are evaluated outside controlled environments[18]. On the other hand, research into adversarial obfuscation techniques exposes vulnerabilities that adversaries can exploit to circumvent AI-generated text detection systems[20]. Therefore, the present study seeks to address these gaps by comprehensively examining how geopolitical, ideological, and philosophical perspectives emerge differently across different LLMs, thus, address whether these differences present risks of exacerbating existing global divisions.

METHODS

This study employs a multi-stage approach rooted in Wittgenstein's language games[21], presenting a comprehensive and reproducible framework to evaluate higher-reasoning functions in four large language models (LLMs). In the first stage of the methodology, principles of language games were used to design standardized sets of questions; each set prompts an LLM to generate a worldview framed by two or more games. In the second stage, responses from different LLMs are constructed into a dataset. In the third stage, this dataset is analysed for *surface*-level worldview, which includes semantic similarity, sentiment analysis, and thematic coverage. This is done using statistical analyses assuming a null-hypothesis stating that there is no differences between the worldviews generated by different LLMs. The fourth and final methodological stage is exploring the *deep*-level worldviews. This is accomplished through classification, topic modelling, and MANOVA analysis. Through this multi-stage approach, the study systematically investigates how LLMs interpret meaning, resolve ambiguities, and encode worldviews across thematic areas. Each

stage is interconnected, from designing context-sensitive prompts to measuring responses and applying advanced statistical evaluations, ensuring that the findings reflect deep insights into the LLMs' reasoning abilities while being robustly reproducible by other researchers.

Design of the Standard Question Sets

Wittgenstein's language games theory emphasizes that meaning arises through use in context-specific linguistic games, governed by implicit rules that guide understanding and reasoning[22]. This principle underpins the design of standard worldview question sets, which function as contextual moves within a discourse game, promoting paradigm-specific reasoning.

By embedding speech acts (directives, commissives) and contextual triggers, we ensure that LLM responses reflect the contextually induced meanings of linguistic expressions[23,24]. Furthermore, the incorporation of paradigm-driven vocabulary anchors questions in competing worldviews, guiding the models to navigate ideologically diverse reasoning spaces[25]. This application aligns with Wittgenstein's view that meaning is inseparable from its practical use and exemplifies how LLMs, despite their complex nonlinearity, induce meaning through rule-following behaviour[26].

The study proposes four question sets, each set represents a distinct language game with specific thematic context. Questions were crafted based on meaning-in-use, functioning as linguistic moves that embed rules, assumptions, and paradigms. The exact question sets can be found in the supplementary materials. The design principles are summarized in table 1. The four question sets were designed to define the study's scope, but the modular nature of this framework allows researchers to design *virtually infinite* sets to study LLM worldviews in other domains, such as art, sports, international trade, or any other domain.

Table 1. Language game principles used to design the standard worldview question sets

Design Principle	**Description**	**Example**
Speech Act Theory	Questions were classified as directives, commissives, expressives, or declaratives. Directives simulate decision-making, while commissives represent implied commitments.	Directive: "Should democracy be imposed on nations unfamiliar with it?" Commissive: "Should nations commit to unilateral nuclear disarmament?"
Paradigm-Driven Vocabulary	Keywords were chosen to anchor questions in competing worldviews, ensuring that responses reflect ideological tensions. Post-colonial and economic terms were specifically selected to create meaningful contestation.	"Should the West return stolen artifacts taken during colonialism?" prompts reasoning on historical justice versus national heritage.
Contextual Triggers	Questions embed triggers like moral dilemmas and legal debates to prompt reasoning beyond factual recall. Triggers force models to balance competing values such as justice, security, and cultural preservation.	"Should technologically advanced nations intervene in the governance of less developed countries?" raises issues of sovereignty versus paternalism.

Set 1: Justice and Sovereignty Game

This game focuses on the tension between justice and sovereignty, exploring how reparative actions—particularly in contexts of historical injustices like colonial exploitation—require balancing ethical imperatives with national autonomy. The game follows two key rules. The rule of moral obligation compels responses to acknowledge the ethical necessity of addressing historical injustices, while the rule of sovereign authority emphasizes that states maintain the right to control how reparative processes are designed and implemented.

Within this game, the allowable moves for LLMs involve proposing reparative mechanisms (such as monetary compensation or institutional reform) while critically assessing their feasibility given national constraints and political realities. For instance, in response to the question, *"Should colonial reparations effectively address the structural inequalities rooted in historical exploitation?"*, LLMs are expected to balance these competing demands, arguing for or against reparations while grounding their reasoning in socio-political contexts where reparative processes could face resistance or backlash.

Set 2: Security and Justice Game

This game centres on the interplay between security and justice, addressing dilemmas such as whether security measures should override ethical responsibilities or how nations can balance self-defence with reparative commitments. The rule of defensive necessity underpins responses that emphasize strategic calculations for national and global security, while the rule of ethical accountability ensures that discussions of defence policies are not divorced from considerations of justice.

The moves allowed in this game involve proposing security measures like disarmament treaties, alliances, or demilitarization efforts while assessing their moral implications. For example, when LLMs respond to the question, *"Should militarization in space be banned outright to preserve it as a peaceful frontier?"*, they must weigh the risks of conflict escalation against the moral duty to maintain peace and prevent arms races. This game demands reasoning that integrates both pragmatic defence strategies and justice-driven frameworks, reflecting the duality of national interest and ethical responsibility.

Set 3: Security and Sovereignty Game

This game addresses the dual concern of security and sovereign control in the face of global challenges, particularly those posed by technological advancements and geopolitical conflicts. The rule of strategic autonomy requires LLMs to defend a nation's ability to make independent decisions regarding its defence strategies, while the rule of cooperative stability ensures that responses also consider the need for international collaboration to address global risks.

In this game, LLMs are prompted to argue for the preservation of national sovereignty through moves that propose localized technological regulation or military strategies. Alternatively, they can explore cooperative mechanisms such as international agreements that balance sovereignty with collective security. A key example is found in the question, "Can artificial intelligence development justify its risks to employment, privacy, and security?", where LLMs must address the security risks posed by AI while evaluating whether national or global oversight offers the most effective solutions.

Set 4: Technology and Security Game

This game focuses on the role of technological transformation as both an enabler of global security and a disruptor of governance structures, with discussions centred on how nations should manage the dual risks and opportunities posed by advanced technologies. The rule of technological governance dictates that responses consider the implications of innovations like AI or digital infrastructures on national and global stability, while the rule of protective intervention allows LLMs to propose actions to mitigate technological risks.

The moves within this game enable LLMs to propose regulatory frameworks, evaluate cooperative governance models, or highlight the potential threats posed by unchecked technological expansion. In the question, *"Do nations with superior technology have the right to intervene in the governance of less developed countries?"*, LLMs must assess whether intervention is justified based on technological superiority or if such actions represent an overreach that threatens sovereignty and stability.

Collection and Preprocessing of LLM Responses

Responses to the four question sets were collected from four different LLMs ("LLM1" to "LLM4") using official API calls, representing four of the top 10 most widely used general-purpose LLMs globally. To maintain analytical neutrality and minimize potential bias, the study abstracts these models as LLM1 through LLM4, ensuring that the comparative analysis is driven purely by empirical evaluation rather than assumptions tied to model reputation or market position. Each LLM was prompted sequentially with the same set of questions, producing a comprehensive response dataset structured as:

$$R = \bigcup_{j=1}^{4} R_j \qquad (1)$$

Where the full dataset R is the union of all response datasets R_j collected from the four different LLMs (LLM1, LLM2, LLM3, and LLM4). Each R_j represents the responses generated by a specific LLM and combining them ensures a comprehensive dataset for comparative analysis. The response matrix is expressed as:

$$R = \begin{bmatrix} r_{11} & \cdots & r_{1N} \\ \vdots & \ddots & \vdots \\ r_{M1} & \cdots & r_{MN} \end{bmatrix} \qquad (2)$$

where an individual response r_{ij} represents the response r provided by LLM j in response to a question q_i. Here rows represent different questions from each set, columns represent responses from different LLMs, M and N are the number of questions per set and the number of LLMs, respectively.

Word Count Analysis

The length of each response $r_{ij}$is measured by counting the number of words it contains as:

$$Word\ Count\ (r_{ij}) = \sum_{k=1}^{K_{ij}} 1 \qquad (3)$$

where K_{ij} is the total number of words in a response r_{ij}, where each word k contributes +1 to K_{ij}. This produces the word count matrix, expressed as:

$$W_{MN} = \begin{bmatrix} w_{11} & \cdots & w_{1N} \\ \vdots & \ddots & \vdots \\ w_{M1} & \cdots & wMN \end{bmatrix} \qquad (4)$$

Text Embedding Process

To evaluate the semantic relationship between the questions and their corresponding responses, each text unit (both questions and responses) was transformed into a high-dimensional vector representation using the sentence transformer model all-MiniLM-L6-v2. The embedding process leverages the transformer architecture's ability to capture contextual relationships within

the text, thus enabling meaningful comparisons between semantically related sentences. Formally, the embedding function $f_{embed}(\cdot)$ maps a given input text to a $d-dimensional$ embedding vector:

$$V_{qi} = f_{embed}(q_i), V_{rj} = f_{embed}(r_{ij}) \qquad (5)$$

where $V_{qi} \in \mathbb{R}^d$ is the embedding of question q_i and $V_{r_{ij}} \in \mathbb{R}^d$ is the embedding vector of response r_{ij}. The embedding dimension d is fixed by the transformer model, and in the case of all-MiniLM-L6-v2, it is optimized to represent both short and long texts effectively. The embedding vectors encapsulate semantic features, allowing similar meanings to be represented by vectors that are spatially close in the high-dimensional space.

Semantic Similarity Calculations

Semantic similarity quantifies the relationship between a response generated by the large language model (LLM) and its corresponding question by measuring how semantically aligned their vector representations are. To achieve this, both the response and the question are embedded as dense vectors using the sentence transformer model all-MiniLM-L6-v2. The embedding vectors capture semantic information, allowing for precise comparison. The similarity between the response and the question is computed using **cosine similarity**, defined as:

$$S_{ij} = \frac{V_{q_i} \cdot V_{r_{ij}}}{\|V_{q_i}\| \cdot \|V_{r_{ij}}\|} \qquad (6)$$

where $V_{q_i} \cdot V_{r_{ij}}$ is the dot product of the two embedding vectors, and $\|V_{q_i}\|$, $\|V_{r_{ij}}\|$ are the Euclidean norms of the vectors, respectively. The similarity score S_{ij} ranges from -1 to +1 where -1 indicates complete semantic opposition, 0 indicates no semantic relationship, and +1 indicates perfect semantic alignment. The scores form the semantic similarity matrix S.

Sentiment Analysis

LLM response r_{ij} is tokenized embedded into a high-dimensional vector space $X \in \mathbb{R}^{T\times d}$, where T is the number of tokens in a response and d is the embedding dimension which is 768 for the sentiment classification model used in this study. Thus, the tokenized input matrix X has the shape $X = [x_1, x_2, \dots, x_T] \in \mathbb{R}^{T\times d}$ where each token $x_i \in \mathbb{R}^d$encodes contextual information.

The encoding and feature extraction was implemented using the open-source BERT-based sentiment classification model nlptown/bert-base-multilingual-uncased-sentiment which applies the multiple pre-trained transformer layers which perform self-attention and nonlinear transformations to compute contextual embeddings at the output layer. The final representation of the response is given by:

$$h = f_{BERT}(X) \in \mathbb{R}^d \qquad (7)$$

This vector h summarizes the semantic content of the response and serves as input to the classification layer. After the response r_{ij} is processed by the transformer layers of the BERT model, the final output representation $h \in \mathbb{R}^d$ is obtained from the [CLS] token embedding. The vector h summarizes the entire semantic content of the response and is passed through a linear classification layer to project it into a logit space**.** In turn, the logit space is represented as a vector $z = [z1, z2, \dots, zC] \in \mathbb{R}^C$ where C is the number of sentiment categories (e.g., 5 categories ranging from highly negative to highly positive sentiment). The projection from the BERT output $\underline{h}$ to the logit vector z is computed using a linear transformation followed by the addition of a bias term $z = Wh + b$ where $W \in \mathbb{R}^{C\times d}$ is the weight matrix mapping the d-dimensional BERT representation to the C-dimensional logit space, $b \in \mathbb{R}^C$ is the bias vector that adjusts the logit values, and where each component $z_k$of the logit vector represents an **unnormalized score** for the corresponding sentiment label k.

The logit z_k for the sentiment class k is calculated as $z_k = \sum_{i=1}^{d} W_{ki}h_i + b_k$where W_{ki} is the weight associated with the $i-th$ feature of the BERT output for sentiment class k, $h_i$is the $i-th$ feature of the BERT output, and b_k is the bias term for sentiment class k. The logit z_k reflects the contribution of the BERT output features toward predicting the likelihood of

sentiment class k. To convert the logits $z = [z_1, z_2, \dots, z_C]$ to probabilities, a softmax function is used to normalize the logits into a probability distribution over all possible sentiment classes, ensuring that the probabilities sum to 1. The probability of assigning sentiment label k to the response r_{ij} is given by:

$$P(label_k \mid r_{ij}) = \frac{e^{z_k}}{\sum_{C=1}^{C} e^{z_C}} \qquad (8)$$

where $e^{z_k}$is the exponential of the logit for class k, the denominator $\sum_{C=1}^{C} e^{z_C}$ is the sum of exponentials of all logits, ensuring that the output is normalized into a valid probability distribution. The softmax function amplifies differences between logits by using the exponential function. If one logit $z_k$is much larger than the others, its corresponding probability $P(label_k \mid r_{ij})$ will be close to 1, while the probabilities for other classes will be close to 0. Conversely, if the logits are similar in magnitude, the probabilities will be more evenly distributed.

Thematic Coverage Analysis

Thematic coverage assesses how comprehensively a given response r_{ij} addresses key topics related to the question q_i. The topics are predefined in categories relevant to the study, such as economics, environmental issues, and social justice. Let $T = \{t_1, t_2, \dots, t_L\}$ represent the set of predefined topic embeddings generated using the same embedding model as the responses. Let $f_{embed}(t_l) \in \mathbb{R}^d$ be the vector embedding of topic t_l obtained using the same sentence transformer model used for the response embeddings. The thematic coverage is calculated as the average cosine similarity between the response embedding and each topic embedding V_{rij} and each topic embedding $f_{embed}(t_l)$:

$$Thematic\ Coverage(r_{ij}) = \frac{1}{L}\sum_{l=1}^{L} \frac{V_{r_{ij}} \cdot f_{embed}(t_l)}{\left\|V_{r_{ij}}\right\| \cdot \|f_{embed}(t_l)\|} \qquad (9)$$

where L is the number of predefined topics, $V_{r_{ij}}$ is the embedding vector for the response, $f_{embed}(t_l)$ is the embedding vector for the topic, and the cosine similarity measures how closely

the response aligns with each topic. Thematic coverage evaluates how closely the response aligns semantically with multiple relevant topics and provides a key feature for later statistical analysis.

Zero-Shot Classification for Worldview Induction

Zero-shot classification assigns each response r_{ij} to categories representing geopolitical, ideological, and philosophical worldview dimensions without requiring task-specific training. Let the predefined sets of categories be: $C_{geo} = \{c_1, \dots, c_G\}$, $C_{ideo} = \{d_1, \dots, d_I\}$, and $C_{phil} = \{e_1, \dots, e_P\}$ where C_{geo} contains geopolitical categories (e.g., globalism, nationalism), C_{ideo} contains ideological categories (e.g., socialism, liberalism), and C_{phil} contains philosophical categories (e.g., utilitarianism, deontology). The response embedding $V_{r_{ij}}$ is compared to the category descriptions using cosine similarity. The probability of classifying r_{ij} under category $c_g \in C_{geo}$ is given by:

$$P_{geo}(c_g|r_{ij}) = \frac{\exp\left(V_{r_{ij}} \cdot f_{embed}(c_g)\right)}{\sum_{g=1}^{G} \exp\left(V_{r_{ij}} \cdot f_{embed}(c_g)\right)} \qquad (10)$$

Similar probabilities are computed for ideological and philosophical classification:

$$P_{ideo}(d_i|r_{ij}) = \frac{\exp\left(V_{r_{ij}} \cdot f_{embed}(d_i)\right)}{\sum_{i=1}^{I} \exp\left(V_{r_{ij}} \cdot f_{embed}(d_i)\right)} \qquad (11)$$

$$P_{geo}(e_p|r_{ij}) = \frac{\exp\left(V_{r_{ij}} \cdot f_{embed}(e_p)\right)}{\sum_{p=1}^{P} \exp\left(V_{r_{ij}} \cdot f_{embed}(e_p)\right)} \qquad (12)$$

After the classification is completed, a thematic labelling process is implemented where each response r_{ij} is labelled with the most likely category within the geopolitical, ideological, and philosophical worldview dimensions using the arg max function. The arg max (short for "argument of the maximum") selects the index or label corresponding to the highest probability within a given set of categories. The arg max function ensures that the label assigned to the response corresponds to the category in which the response has the highest semantic alignment, based on the zero-shot classification probabilities. This can be formally expressed as:

$$\hat{c}_{ij} = c_{g^*} \text{ such that } P_{geo}(c_{g^*}|r_{ij}) \geq P_{geo}(c_g|r_{ij}) \quad \forall c_g \in C_{geo} \qquad (13)$$

where $\hat{c}_{ij} = c_{g^*}$ is the category with the maximum probability and $P_{geo}(c_{g^*}|r_{ij})$ is the highest probability among all categories c_g. Similarly, for ideological and philosophical dimensions, their formal expressions implemented in this work are expressed, respectively, as:

$$\hat{d}_{ij} = d_{i^*} \text{ such that } P_{ideo}(d_{i^*}|r_{ij}) \geq P_{ideo}(d_i|r_{ij}) \quad \forall d_i \in C_{ideo} \qquad (14)$$

$$\hat{e}_{ij} = e_{p^*} \text{ such that } P_{phil}(e_{p^*}|r_{ij}) \geq P_{phil}(e_i|r_{ij}) \quad \forall e_p \in C_{phil} \qquad (13)$$

The output for each response is a triplet of labels representing its dominant classifications $(\hat{c}_{ij}, \hat{d}_{ij}, \hat{e}_{ij})$. This classification enables comparative worldview analysis by identifying dominant geopolitical, ideological, and philosophical categories for each LLM.

Hypothesis Testing and Statistical Analysis

To identify significant performance differences across large language models (LLMs) on key metrics such as semantic similarity, thematic coverage, word count, and sentiment distribution, the statistical analysis in WorldView code (see the GitHub repository) dynamically selects and applies appropriate statistical tests based on the properties of the data. The code begins by performing normality checks using tests like the Shapiro-Wilk test to determine whether the observed metric data follows a normal distribution. When either the ANOVA or Kruskal-Wallis test detects significant differences (p-value < 0.05), the code triggers post-hoc pairwise comparisons to identify which pairs of LLMs exhibit statistically significant differences. This logic was implemented in the code as following:

I. If the data is normally distributed, the pipeline applies a one-way ANOVA test using the `scipy.stats.f_oneway()` function to compare the means of the metric across LLMs.
II. For non-normally distributed data, the Kruskal-Wallis test is applied using the`scipy.stats.kruskal()` function to compare the rank distributions across LLMs.
III. The code uses Tukey's HSD test to perform pairwise mean comparisons: `(statsmodels.stats.multicomp.pairwise_tukeyhsd())`
IV. For non-parametric tests: Dunn's test `(scikit_posthocs.posthoc_dunn())` is applied, with Bonferroni corrections for controlling the family-wise error rate (FWER) when multiple pairwise comparisons are performed.

RESULTS

Surface Level Analysis

The surface-level analysis (code: WorldView-S.py) focuses on surface-level metrics that capture the immediate characteristics of the responses provided by the different large language models (LLMs). This section outlines key findings regarding semantic similarity, word count, sentiment distribution, and thematic coverage.

Semantic similarity measures the alignment between the responses given by the LLMs and the respective questions. The average semantic similarity values across the models are summarized in table 2. The Kruskal-Wallis test was conducted to assess differences in semantic similarity across the models (H = 0.962, p = 0.810), revealing no statistically significant variation among them.

Table 2. Comparison of semantic similarity across LLMs

LLM ID	Average Semantic Similarity
LLM1	0.4308
LLM2	0.4315
LLM3	0.4459
LLM4	0.4414

Word count provides insight into the verbosity of each LLM's responses. Statistical tests indicate significant variation in word count across models, with a Kruskal-Wallis H value of 118.18 ($p < 1.9 \times 10^{-25}$), suggesting that certain LLMs consistently generate more extensive responses. The average word counts observed are detailed in table 3.

Table 3. Average word counts across LLMs

LLM ID	Average Word Count
LLM1	111.5
LLM2	197.25
LLM3	194.0
LLM4	95.75

For the sentiment analysis, the chi-square test result ($\chi^2 = 88.46, p < 6.3 \times 10^{-17}$) indicates a significant variability in sentiment distribution across the LLMs, suggesting that different models

exhibit distinct sentiment patterns when generating responses. The significant variability in sentiment distribution across LLMs highlights distinct narrative tendencies. LLM1 and LLM2 primarily exhibit cautious to critical sentiment, with LLM2 displaying a strong inclination toward negative responses, often emphasizing challenges and adverse outcomes. In contrast, LLM3 and LLM4 adopt a more optimistic sentiment profile, balancing neutral and positive tones while frequently framing discussions around solutions and constructive perspectives. The detailed sentiment distributions are detailed in table 4.

Table 4. Sentiment distribution across LLMs

LLM ID	Negative (%)	Neutral (%)	Positive (%)
LLM1	50	50	0
LLM2	75	25	0
LLM3	0	50	50
LLM4	0	50	50

Thematic coverage assesses how comprehensively the LLMs address core topics like geopolitics, ethics, and climate change. The ANOVA results ($F = 1.57$, $p = 0.199$) show no statistically significant difference in thematic coverage across models. The detailed scores of thematic coverage across LLMs are detailed in table 5.

Table 5. Thematic coverage scores

LLM ID	Thematic Coverage (Average Score)
LLM1	Moderate
LLM2	High
LLM3	High
LLM4	Low

Deep Level Analysis

The violin plot in Figure 1 illustrates the distribution of worldview metric scores across the four games, highlighting inter-LLM variability with respect to each game's theme. In the Justice and Sovereignty game, LLM4 demonstrates elevated scores in pro-sovereignty and utilitarianism, emphasizing outcome-driven reasoning and national autonomy, while LLM3 displays moderate scores, reflecting a balanced consideration of ethical imperatives and state authority. LLM1 shows lower engagement overall, with minimal emphasis on reparative justice.

In the Security and Justice game, LLM3 consistently exhibits high scores in liberalism and pro-globalization, indicating a stronger focus on progressive and cooperative narratives for balancing defense strategies and ethical accountability. Conversely, LLM4 leans heavily on utilitarian reasoning, favoring pragmatic security measures over justice-oriented outcomes. LLM2 displays moderate scores across both dimensions, indicating a balanced view of security and justice.

The Security and Sovereignty game reveals LLM4's strong pro-sovereignty stance, highlighting its emphasis on national autonomy in defense and governance strategies. LLM3 maintains higher scores in pro-globalization and pragmatism, suggesting a preference for international collaboration. LLM2 exhibits moderate scores in liberalism and pro-sovereignty, reflecting its balanced worldview between collective welfare and national control.

In the Technology and Security game, LLM3 continues to display high scores in liberalism and pro-globalization, indicating its alignment with global regulatory frameworks and collaborative technological governance. LLM4's elevated scores in utilitarianism and pro-sovereignty reflect its focus on balancing technological innovation with national security interests. LLM1 shows its strongest emphasis in pro-West narratives, indicating a selective engagement with global technological cooperation.

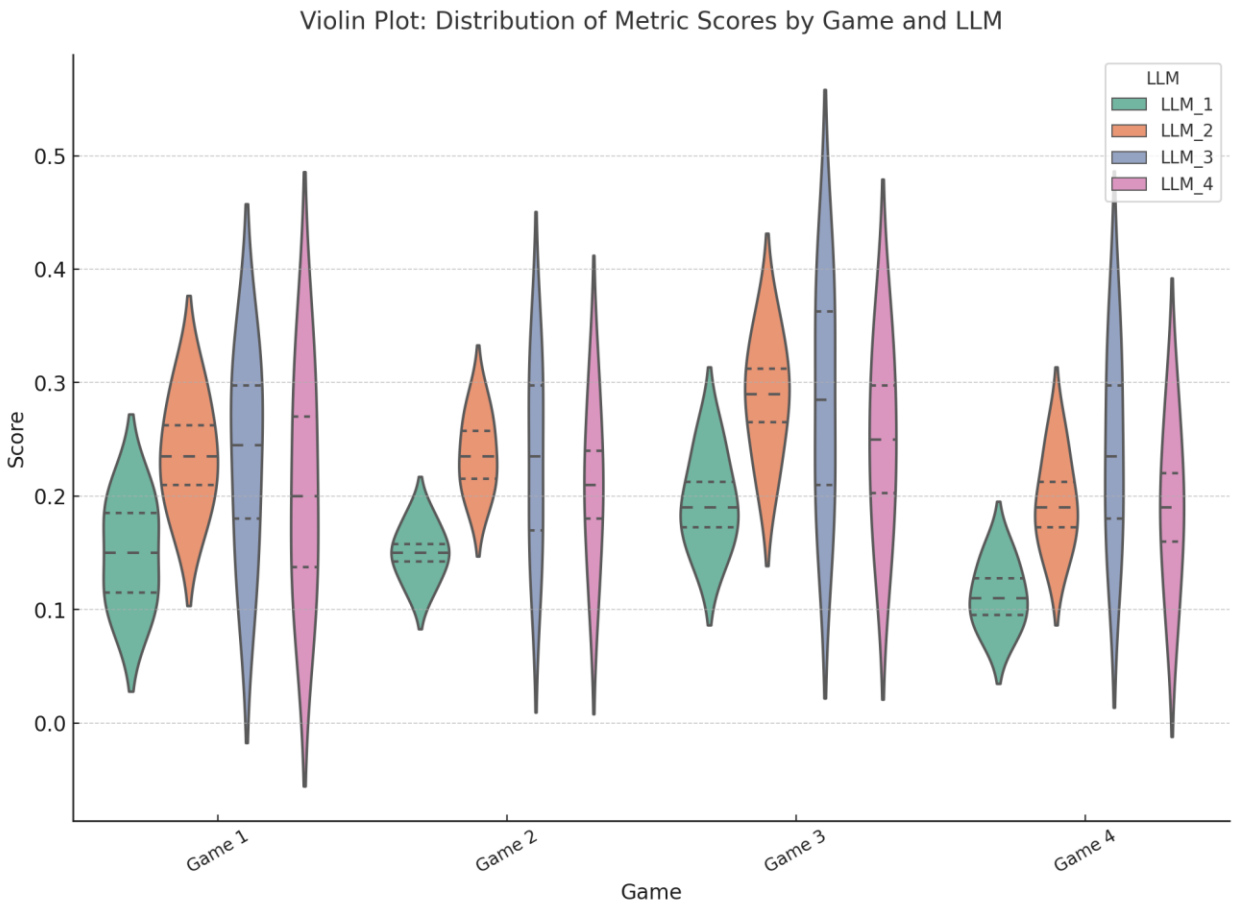


Figure 1. Violin plot illustrating the distribution of worldview metric scores across the four games for each LLM

A comparative analysis of the normalized average metric scores for each LLM is graphically shown in figure 2. LLM3 consistently scored the highest for liberalism (0.32) and pro-globalization (0.29), underscoring its preference for global integration and progressive political values. LLM4, however, demonstrated a divergent ideological profile, with prominent scores for utilitarianism (0.33) and conservatism (0.27), suggesting a focus on pragmatic, outcome-driven reasoning coupled with traditionalist perspectives. LLM1 distinguished itself with a dominant pro-West score (0.35) while showing lower engagement across most other metrics, indicating a selective narrative framework. LLM2 exhibited a relatively balanced distribution of average scores, particularly in idealism (0.22) and pro-sovereignty (0.21), reflecting a more nuanced worldview.

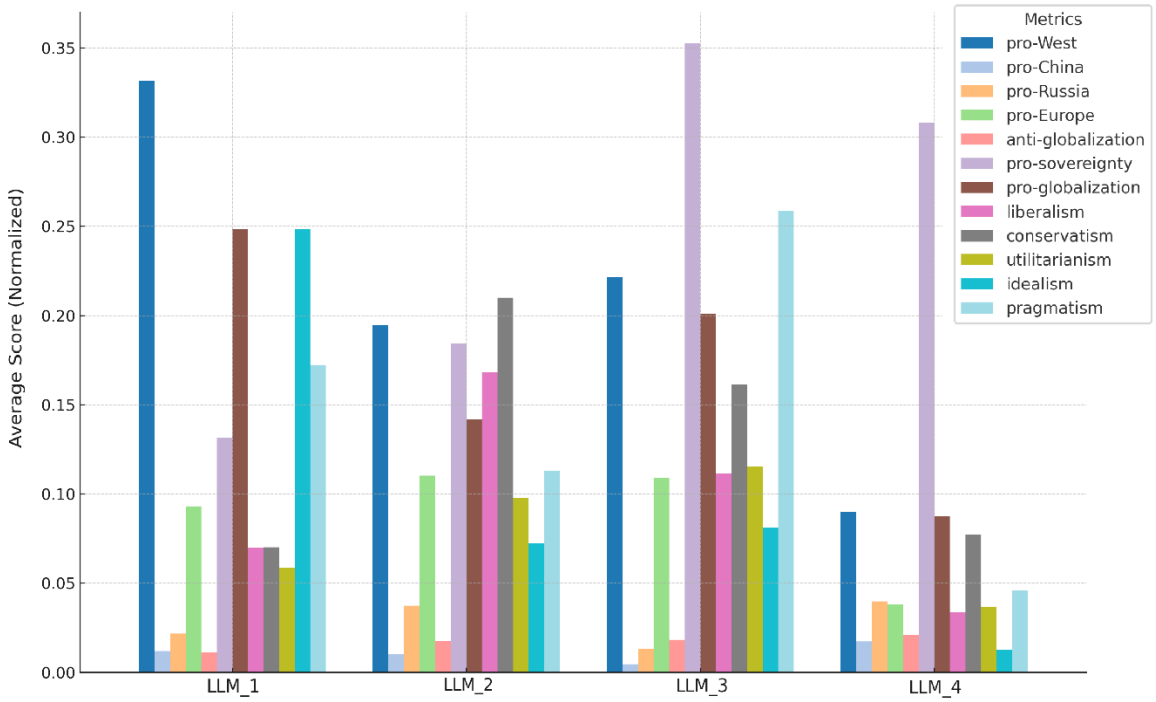


Figure 2. Bar plot showing the normalized average scores for key worldview metrics across LLMs. The plot highlights LLM_3's strong emphasis on liberalism and pro-globalization, in contrast to LLM_4's utilitarian and conservative leanings.

To explore the relationships between key worldview dimensions across the four LLMs, we constructed correlation heatmaps (Figure 3) using verified semantic and ideological data. These heatmaps provide insight into how liberalism, pro-globalization, utilitarianism, pro-sovereignty, conservatism, and idealism interact within each LLM's narrative construction. LLM1 demonstrates strong positive correlations between liberalism and pro-globalization (r = 0.70) and between pro-sovereignty and liberalism (r = 0.90). Negative associations appear between conservatism and liberalism (r = -0.69) and pro-globalization (r = -0.28), reflecting potential ideological opposition within its narrative framework. LLM2 exhibits consistently high interdependencies among liberalism, pro-globalization, and utilitarianism (r values between 0.60 to 1.00). However, the low and mixed correlations involving pro-sovereignty (r = 0.22 to 0.60) indicate nuanced balancing between globalist and sovereignty-centered narratives. The relationship between conservatism and pro-globalization remains significantly high, revealing the integration of traditional and globalist themes. LLM3 shows complex interplays, where liberalism and pro-globalization correlate moderately positively (r = 0.60) while maintaining a strong

negative association with pro-sovereignty (r = -0.60). Negative correlations with conservatism and idealism highlight a preference for progressive and globally integrated ideologies over protectionist or idealist reasoning. LLM4 features strong positive correlations between pro-sovereignty, conservatism, and pro-globalization ($r \geq 0.90$), indicating a narrative blend of pragmatic sovereignty and cooperative internationalism. Liberalism exhibits moderate interdependence with utilitarianism and idealism, further reinforcing LLM4's outcome-driven reasoning framework.

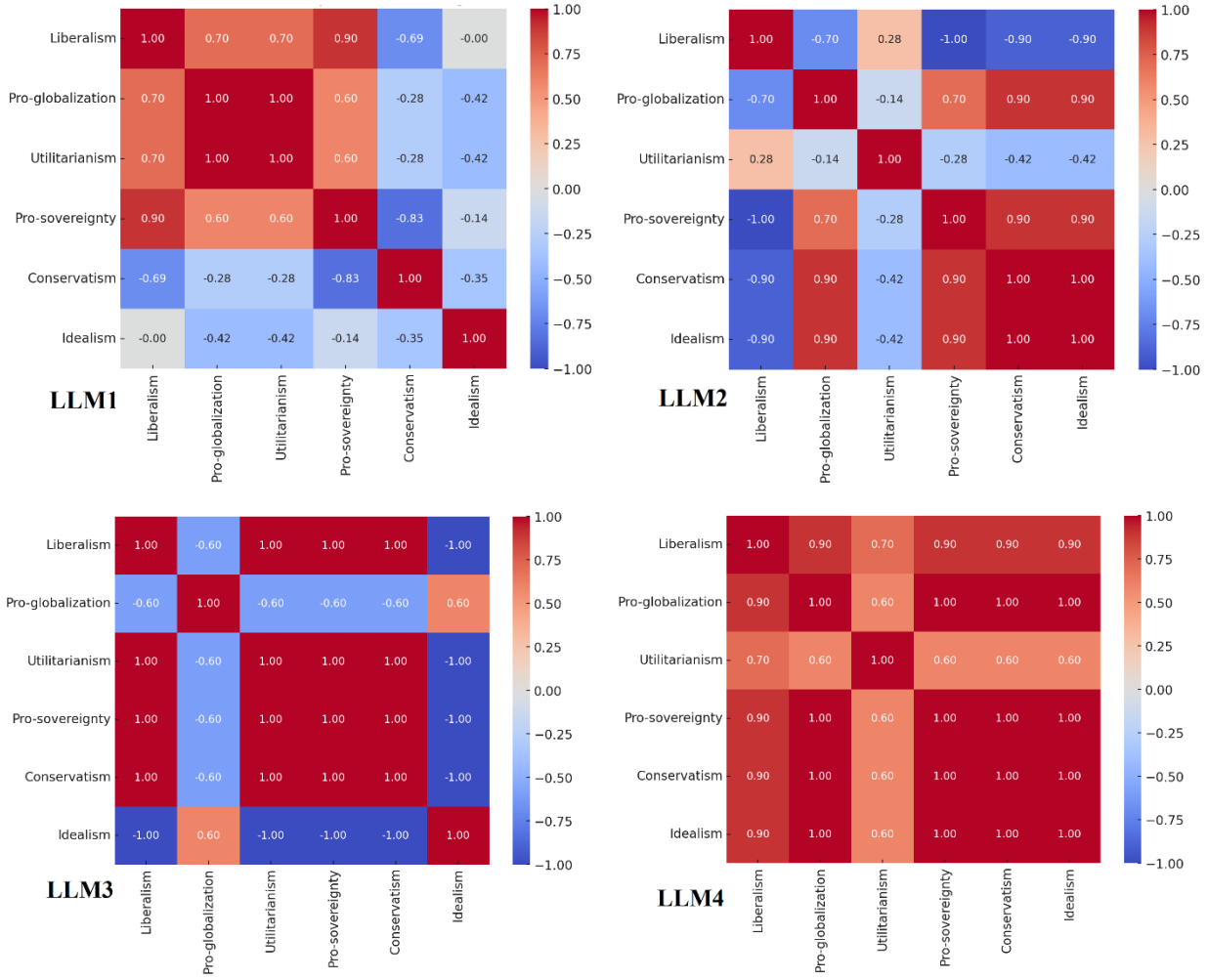


Figure 3. Correlation heatmaps for worldview metrics across four large language models (LLM1–LLM4). The color scale ranges from -1 (strong negative correlation, blue) to +1 (strong positive correlation, red). Each heatmap illustrates interdependencies among liberalism, pro-globalization, utilitarianism, pro-sovereignty, conservatism, and idealism.

Bias profiles across the LLMs in key geopolitical, ideological, and philosophical dimensions are plotted in figure 3. LLM3 demonstrates elevated scores in liberalism, pro-globalization, and pragmatism, indicating stronger biases toward globalist perspectives and adaptable governance. LLM4 shows distinct peaks in utilitarianism, pro-sovereignty, and idealism, reflecting a preference for outcome-focused reasoning, national autonomy, and visionary ideals. LLM1 displays its highest emphasis on pro-China and pro-West dimensions, highlighting a region-specific focus on strategic and national narratives. LLM2 exhibits a well-distributed bias profile, with notable strengths in pro-globalization, socialism, and humanism, suggesting a balanced mix of cooperative internationalism and human-centric perspectives. This visualization highlights areas where LLMs diverge in their worldview representations while maintaining conceptual proximity between related dimensions for enhanced interpretability.

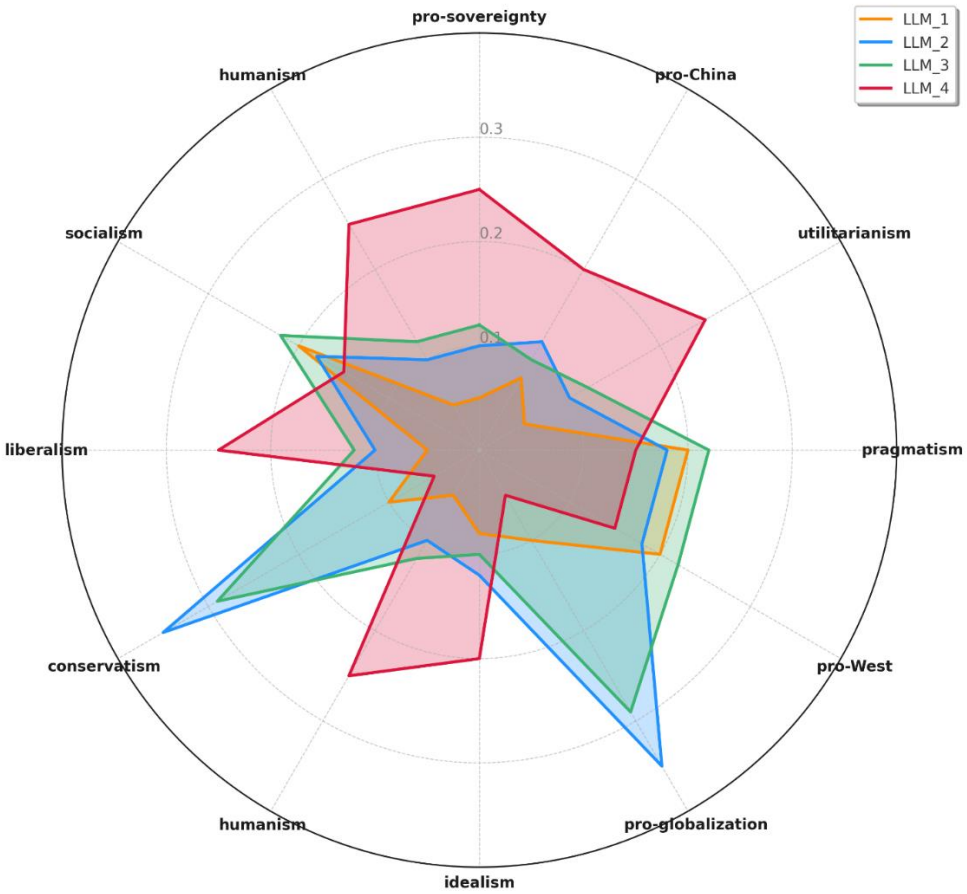


Figure 3. The radar plot illustrates the distinct bias profiles of LLMs across key geopolitical, ideological, and philosophical dimensions, with conceptually related dimensions grouped to minimize perimetric distance.

Kruskal-Wallis tests identified statistically significant differences in the scores across several key dimensions, with liberalism ($H = 12.51$, $p = 0.006$), conservatism ($H = 8.76$, $p = 0.033$), and utilitarianism ($H = 8.56$, $p = 0.036$) displaying the highest variability among LLMs. Pro-globalization ($H = 6.49$, $p = 0.090$) demonstrated near-significant variability, suggesting some inter-LLM differences in framing topics related to global cooperation. In contrast, pro-sovereignty ($H = 4.65$, $p = 0.199$) and pro-West ($H = 3.38$, $p = 0.337$) showed no statistically significant differences, reflecting relatively uniform emphasis across LLMs.

The bias profiles across the four LLMs in opposing dimensions is plotted in figure 5. LLM3 exhibits strong biases toward liberalism (0.32) and pro-globalization (0.29), indicating a preference for globalist and rights-based dimensions. LLM4, in contrast, emphasizes pro-sovereignty (0.25) and utilitarianism (0.33), highlighting a focus on nationalist and outcome-driven dimensions. LLM2 shows the highest bias toward pro-globalization (0.35), while LLM1 presents milder overall biases, with its strongest emphasis on liberalism (0.20).

The opposing nature of the dimensions is evident in the plot: LLM4's high bias in pro-sovereignty contrasts with LLM3 and LLM2's biases toward pro-globalization. Similarly, LLM4's emphasis on utilitarianism opposes LLM3's higher bias in liberalism, reflecting contrasting tendencies in rights-based versus outcome-driven reasoning. This distribution of biases visually demonstrates the varied contributions of each LLM to opposing narrative dimensions.

Figure 5. Radar plot showing the biases of four LLMs toward key worldview dimensions. The opposing pairs are pro-sovereignty vs. pro-globalization and utilitarianism vs. liberalism. Each LLM's influence on these dimensions is indicated by the area covered on the radar. For example, LLM3 exhibits stronger biases toward liberalism and pro-globalization, while LLM4 emphasizes utilitarianism and pro-sovereignty. The plot shows areas of polarization, reflecting tensions between nationalist versus globalist narratives and outcome-driven versus rights-driven reasoning.

DISCUSSION

This study utilized Wittgenstein's theory of language games to uncover the embedded worldviews of four large language models (LLMs), revealing distinct geopolitical, ideological, and philosophical biases. The analysis focused on how meaning, as generated by LLMs, is context-dependent and follows implicit linguistic rules defined by the structure of four language games.

Justice and Sovereignty Worldview

The Justice and Sovereignty worldview reveals the tension between addressing historical injustices and preserving state sovereignty. LLM responses varied significantly in their framing of colonial reparations and national rights, with LLM3 demonstrating a strong pro-globalization stance and LLM4 favouring pragmatic, sovereignty-driven reasoning. These findings are

consistent with Erlansyah et al.[17], who emphasize that LLM outputs on culturally sensitive topics often reflect divergent worldviews due to differential training data exposure. LLM3's liberal-globalist framing highlights how LLMs can prioritize collective justice, consistent with studies on Western-centric model biases[7]. However, LLM4's response, which leans toward sovereignty and state-driven reparations, resonates with studies advocating for localized interpretations of justice[15]. This suggests that LLMs internalize competing paradigms, underscoring the importance of diversified training datasets. Moreover, these variations align with Grinin et al.[27], who argue that AI evolution can propagate dominant ideologies unless corrective mechanisms are implemented. In contrast, Dou[28] posits that AI systems can serve as agents of disruption against entrenched worldviews by generating unexpected alternative narratives.

Security and Justice Worldview

This worldview focuses on the balance between security measures and moral obligations. Responses regarding space militarization and nuclear disarmament revealed that LLM1 and LLM2 consistently emphasized defensive necessity over ethical considerations, while LLM3 and LLM4 integrated justice-driven arguments. The emphasis on defensive necessity aligns with prior findings by Liebowitz[19], who discusses how LLMs often adopt security-driven narratives when faced with scenarios involving global stability threats. The divergence between LLM1's risk-averse stance and LLM3's justice-oriented framing highlights Wittgenstein's assertion that meaning arises from the context-specific rules of language games. LLM3's emphasis on collective ethical obligations reflects moves governed by rules of moral accountability, consistent with Alhamadani et al.[13], who argue for ethical AI frameworks in cross-cultural communication. Additionally, Manfredi-Sánchez and Morales[29] highlight that generative AI models, when applied to security and public diplomacy contexts, can either stabilize or exacerbate conflicts based on their underlying narrative structures. Confirming this, Liu et al.[30] highlight instances where LLMs, when exposed to cooperative discourses, displayed adaptive reasoning favouring collective disarmament, indicating flexibility beyond rigid defensive posturing.

Security and Sovereignty Worldview

The responses within this worldview addressed national sovereignty and global technological risks. While LLM4 advocated for strategic autonomy through localized regulation,

LLM3 leaned toward cooperative international solutions. These contrasting approaches reflect differing worldview encoding across models. LLM4's prioritized national self-determination, in agreement with studies on LLM-induced nationalist perspectives, such as Chu et al[10]. In contrast, LLM3's cooperative approach aligns with pragmatism, substantiated the findings by Blanco-Fernández et al.[18], which highlight that LLMs trained on global data exhibit greater tendencies toward multilateral solutions. Polo Serrano[31] supports this observation, suggesting that models like ChatGPT may exhibit inconsistent ideological positions when exposed to differing geopolitical contexts. The results indicate that LLMs can embody opposing paradigms simultaneously, necessitating careful monitoring of their geopolitical implications. In relation, Matz et al.[32] argue that with prompt optimization, LLMs can balance nationalist and cooperative reasoning effectively. In contrast, Berry and Stockman[33] point out that in resource-sensitive contexts, such as trade disputes, LLMs often revert to nationalistic framings, limiting cooperative potential.

Technology and Security Worldview

This game explored how LLMs address technological dominance and intervention. LLM2 and LLM4 framed technological intervention as a matter of national prerogative, while LLM3 advocated for global governance structures. LLM3's support for cooperative oversight resonates with Grinin et al.[34], who highlight that global governance mechanisms are essential in managing AI's evolving role in knowledge systems. This view is further strengthened by Maathuis and Kerkhof[35], who explore AI's potential in aligning dynamic emotional and social contexts, demonstrating how governance frameworks can reduce instability in domain-specific applications. Dou[28] similarly identifies global frameworks as a key factor in mitigating the spread of misinformation by embedding consensus-building mechanisms. However, variability across domains remains a persistent challenge. Liu et al.[30] suggest that performance inconsistencies can be mitigated through sector-specific prompts and targeted governance measures, particularly when training data reflects localized realities. Contradicting this optimistic view, Matz et al.[32] caution that structural instability in AI-driven systems persists in highly politicized contexts unless cross-cultural data curation undergoes rigorous revision.

The risk of narrative control and polarization, highlighted by Frank et al.[12] in studies on misinformation propagation, is amplified when LLMs follow dominant paradigms without

sufficient diversity in input data. This can lead to disproportionate amplification of prevailing ideologies. To mitigate this, training datasets must be diversified to include underrepresented perspectives[36], and evaluative frameworks, such as the one presented here, can be used to identify and address biases. Moreover, long-term mitigation strategies, including dynamic adaptation mechanisms, are necessary to prevent LLMs from embedding static worldviews. Contradicting the dominance of narrative control concerns, Young[37] highlights emerging self-regulation capabilities within LLMs that could reduce bias amplification under optimal prompt-engineering conditions, suggesting that bias control can be partially achieved without requiring radical model retraining.

CONCLUSION

This study demonstrates that LLMs inherit and amplify the geopolitical, ideological, and cultural divides embedded in their training data, constructing distinct worldviews rather than generating neutral outputs. LLM3's cooperative, globalist framing and LLM4's sovereignty-driven, pragmatist reasoning highlight how models reflect conflicting perspectives within global discourse. Through a structured comparative framework, we exposed measurable differences in how LLMs navigate justice, security, and governance, underscoring their role as active participants in shaping narratives. Without responsible oversight, LLMs risk deepening existing divisions and disseminating biased, potentially harmful content under the guise of authoritative knowledge. Stakeholders in policy, digital media, and knowledge dissemination must urgently implement safeguards—diverse training data, adaptive prompts, and ethical regulation—to ensure that LLMs foster balanced, inclusive narratives. Failure to act risks leaving global knowledge systems vulnerable to misinformation, polarization, and exploitation, but with strategic intervention, LLMs can be steered toward promoting informed, constructive dialogue.

REFERENCES

1 Borgeaud, S. *et al.* in *International conference on machine learning.* 2206-2240 (PMLR).

2 Gallegos, I. O. *et al.* Bias and fairness in large language models: A survey. *Computational Linguistics*, 1-79 (2024).

3 Schramowski, P., Turan, C., Andersen, N., Rothkopf, C. A. & Kersting, K. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence* **4**, 258-268 (2022).

4 Feng, S., Park, C. Y., Liu, Y. & Tsvetkov, Y. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. *arXiv preprint arXiv:2305.08283* (2023).
5 Xu, A. *et al.* in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* 2390-2397.
6 Eloundou, T., Manning, S., Mishkin, P. & Rock, D. GPTs are GPTs: Labor market impact potential of LLMs. *Science* **384**, 1306-1308, doi:doi:10.1126/science.adj0998 (2024).
7 Wei, X., Kumar, N. & Zhang, H. Addressing bias in generative AI: Challenges and research opportunities in information management. *Information and Management* **62**, doi:10.1016/j.im.2025.104103 (2025).
8 Tang, Z. *et al.* in *International Conference on Information and Knowledge Management, Proceedings.* 2240-2249.
9 Zhang, Y. *et al.* in *International Conference on Information and Knowledge Management, Proceedings.* 5605-5607.
10 Chu, Z., Ai, Q., Tu, Y., Li, H. & Liu, Y. in *International Conference on Information and Knowledge Management, Proceedings.* 384-393.
11 Węcel, K. *et al.* Artificial intelligence-friend or foe in fake news campaigns. *Economics and Business Review* **9**, 41-70, doi:10.18559/ebr.2023.2.736 (2023).
12 Frank, D., Bernik, A. & Milkovic, M. in *ICCC 2024 - IEEE 11th International Conference on Computational Cybernetics and Cyber-Medical Systems, Proceedings.* 25-30.
13 Alhamadani, A. *et al.* in *Proceedings of the 2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2023.* 492-501.
14 Restrepo, D. *et al.* in *Proceedings - 2024 IEEE 12th International Conference on Healthcare Informatics, ICHI 2024.* 565-566.
15 Nguyen, J. K. Human bias in AI models? Anchoring effects and mitigation strategies in large language models. *Journal of Behavioral and Experimental Finance* **43**, doi:10.1016/j.jbef.2024.100971 (2024).
16 Ali, H., Qadir, J., Alam, T., Househ, M. & Shah, Z. in *2023 IEEE International Conference on Artificial Intelligence, Blockchain, and Internet of Things, AIBThings 2023 - Proceedings.*
17 Erlansyah, D. *et al.* LARGE LANGUAGE MODEL (LLM) COMPARISON BETWEEN GPT-3 AND PALM-2 TO PRODUCE INDONESIAN CULTURAL CONTENT. *East. Eur. J. Enterp. Technol.* **4**, 19-29, doi:10.15587/1729-4061.2024.309972 (2024).
18 Blanco-Fernández, Y., Otero-Vizoso, J., Gil-Solla, A. & García-Duque, J. Enhancing Misinformation Detection in Spanish Language with Deep Learning: BERT and RoBERTa Transformer Models. *Appl. Sci.* **14**, doi:10.3390/app14219729 (2024).
19 Liebowitz, J. *REGULATING HATE SPEECH CREATED BY GENERATIVE AI*. (CRC Press, 2024).
20 Jang, J. & Le, T. in *International Conference on Electrical, Computer, and Energy Technologies, ICECET 2024.* (Institute of Electrical and Electronics Engineers Inc.).
21 Gálvez, J. P. & Gaffal, M. *The Many Faces of Language Games*. (De Gruyter, 2024).
22 Ball, B., Helliwell, A. C. & Rossi, A. *Wittgenstein and artificial intelligence, volume I: Mind and language*. (Anthem Press, 2024).
23 Bozenhard, J. in *AISB Convention 2021: Communication and Conversations.* (The Society for the Study of Artificial Intelligence and Simulation of Behaviour).
24 Csepeli, G. The silent province. *Magy. Nyelvor* **148**, 565-569, doi:10.38143/Nyr.2024.5.565 (2024).
25 Gasparyan, D. E. Language as eigenform: Semiotics in the search of a meaning. *Vestnik Sankt-Peterburgskogo Univ. Filosofiia Konfliktologiia* **34**, 474-491, doi:10.21638/spbu17.2018.402 (2018).
26 Natarajan, K. P. in *CEUR Workshop Proceedings.* (eds G. Coraglia *et al.*) 115-120 (CEUR-WS).

27 Grinin, L. E., Grinin, A. L. & Grinin, I. L. The Evolution of Artificial Intelligence: From Assistance to Super Mind of Artificial General Intelligence? Article 2. Artificial Intelligence: Terra Incognita or Controlled Force? (2024).
28 Dou, W. in *Proceedings - 2024 International Conference on Artificial Intelligence and Digital Technology, ICAIDT 2024.* 144-147 (Institute of Electrical and Electronics Engineers Inc.).
29 Manfredi-Sánchez, J. L. & Morales, P. S. Generative AI and the future for China's diplomacy. *Place Brand. Public Diplomacy*, doi:10.1057/s41254-024-00328-7 (2024).
30 Liu, X., Lin, Y. R., Jiang, Z. & Wu, Q. Social Risks in the Era of Generative AI. *Proceedings of the Association for Information Science and Technology* **61**, 790-794, doi:10.1002/pra2.1103 (2024).
31 Polo Serrano, D. IS CHATGPT WOKE? Comparative analysis of '1984' and 'Brave New World' in the Digital Age. *Vis. Rev. Rev. Int. Cult. Visual Rev. Int. Cultur.* **16**, 251-265, doi:10.62161/revvisual.v16.5264 (2024).
32 Matz, S. C. *et al.* The potential of generative AI for personalized persuasion at scale. *Sci. Rep.* **14**, doi:10.1038/s41598-024-53755-0 (2024).
33 Berry, D. M. & Stockman, J. Schumacher in the age of generative AI: Towards a new critique of technology. *Eur. J. Soc. Theory* **27**, 437-455, doi:10.1177/13684310241234028 (2024).
34 Grinin, L. E., Grinin, A. L. & Grinin, I. L. The Evolution of Artificial Intelligence: From Assistance to Super Mind of Artificial General Intelligence? Article 2. Artificial Intelligence: Terra Incognita or Controlled Force? *Soc. Evol. Hist.* **23**, 165-189, doi:10.30884/seh/2024.02.07 (2024).
35 Maathuis, C. & Kerkhof, I. in *Proceedings of the 4th International Conference on AI Research, ICAIR 2024.* (eds C. Goncalves & J. C. D. Rouco) 260-270 (Academic Conferences International Limited).
36 Prestridge, S., Fry, K. & Kim, E. J. A. Teachers' pedagogical beliefs for Gen AI use in secondary school. *Technol. Pedagog. Educ.*, doi:10.1080/1475939X.2024.2428606 (2024).
37 Young, B., Anderson, D. T., Keller, J. M., Petry, F. & Michael, C. J. in *Proceedings - Applied Imagery Pattern Recognition Workshop.* (Institute of Electrical and Electronics Engineers Inc.).