大規模言語モデルと世界各国の価値観とのアライメント

Yang Liu[1]　金子 正弘[2]　Chenhui Chu[1]
[1] 京都大学　[2]MBZUAI
yangliu@nlp.ist.i.kyoto-u.ac.jp　Masahiro.Kaneko@mbzuai.ac.ae
chu@i.kyoto-u.ac.jp

概要

大規模言語モデル（LLM）の利用者は多様な地域的・世代的な背景を有しており，LLM は単に多言語で応答できるだけではなく，それらの背景に根差した価値観とアライメントされていることが望ましい．しかし，既存研究は少数国や現在の価値観に対象が偏っており，世界規模の国や多様な世代の価値観へのアライメントを多言語で網羅的に調査していない．本研究では，国，言語と世代の 3 項目ごとに LLM の価値観へのアライメントを体系的に評価する．実験の結果，(i) LLM は少数の国の価値観に適切あるいは過剰にアライメントする一方，多くの国ではほとんどアライメントされていない，(ii) プロンプトの言語によって国ごと価値観へ誘導できる可能性があること，(iii) LLM は過去よりも現代の価値観にアライメントされていることを確認した．

1　はじめに

大規模言語モデル（LLM）は日常の意思決定支援において重要性を増している [1, 2]．利用者とのやり取りでは，客観的質問のみならず主観的質問への回答も求められる [3, 4, 5]．価値観へのアライメントという観点で LLM は，事前学習における次トークン予測から価値観に対する統計的事前分布を学ぶ [6, 7]．その後の指示学習 [8, 9] や選好最適化 [10, 3] によって LLM は人間の価値観にアライメントするように学習される．LLM の利用者層は地域的・世代的に多様であり，多言語対応のみならず，そうした背景から形成される価値観との整合性を備えることが重要である．そのため，アライメントのためのデータは，多様な国・言語・年代を背景に持つアノテータの価値観に沿って作成される必要がある [11]．

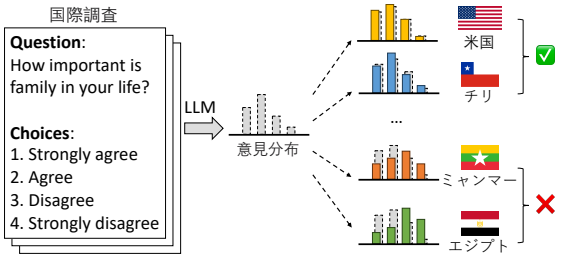


図 1: 価値観に関わる主観的な質問に対する LLM の意見分布の特定の国の意見分布への偏り．

先行研究は，主として米国内の人間や現代人が持つ価値観に対して LLM をアライメントを行ってきた [4]．特定の集団へのアライメントを誘導する試みとして，Durmus ら [12] はデータセットをロシア語・中国語・トルコ語へ翻訳し LLM をアライメントしたが，それらの国の人間へのアライメントは改善しないことが報告されている．本研究では既存研究では十分に調査されていない「主観的質問に対し，LLM の意見分布は具体的にどの国・世代の価値観の意見分布にアライメントされているのか？」また「アライメントを誘導することは可能か？」を明らかにする．特定国や世代の価値観を系統的に反映する LLM は透明性向上と地政学的・倫理的リスク評価に資し，多様な価値観を尊重する AI 設計の基盤となる [13, 14, 15]．図 4 は価値観の特定の国への偏りの例を示している．

従来の評価では「次トークンの対数確率」から意見分布を得る手法が多く，確率の較正問題やロジット出力の必要性などの制約がある [5]．本研究では，より良好とされる，LLM に選択肢ごとの確信度や支持度を自然言語で直接出力させる verbalized distribution（言語化分布）により LLM の意見分布を取得する．加えて，最先端の多言語 LLM を複数対象にし，8 言語・10 か国を対象として言語による価値観の誘導の有効性を検証する．さらに世代ごとの

アライメントも評価することで，国・言語・世代の3項目での価値観とのアライメントを扱う．

World Values Survey（WVS）[16] は政治・宗教・倫理など広範なトピックにおける人々の価値観を収集する国際調査であり，統一形式の質問票の設計により国・言語・時期をまたぐ比較を可能にしている．[1)] 本研究では WVS をベンチマークとして，以下を研究課題とする：**RQ1**：LLM はそれぞれの国の価値観に適切にアライメントするか．**RQ2**：言語は LLM を言語話者の価値観へと誘導できるか．**RQ3**：LLM の応答はどの世代の価値観を反映しているか．

本研究の主な貢献は以下の通りである：(i) 国，言語と世代の3つの観点にわたり価値観とのアライメントを体系的に評価する枠組みを提案した，(ii) 英語・スペイン語圏（例：米国・カナダ・チリ）にアライメントしやすい一方，情報統制が行われている国や非ラテン文字圏（例：ミャンマー・エジプト）に過小にアライメントする傾向を示す，(iii) 8 言語で言語ステアリングを評価し，実験した 40 個の設定（4 モデル× 10 か国）において 77.5%超で言語誘導が有効であることを示した，(iv) 世代の観点での評価により，LLM が現代の価値観へ最もアライメントすることを明らかにした．

2 実験設定

2.1 データセット

RQ1 では，WVS ウェーブ 7（英語版，66 か国）から「経験依存」「客観性が強い」などの理由で LLM に対する評価として不適切な設問を除外し，価値観に関わる設問 144 問を採用する．RQ2 では同 144 問を他言語版質問票から収集する．RQ3 では ウェーブ 5（2005–2009），ウェーブ 6（2010–2014），ウェーブ 7（2017–2022）の英語版を用い，全 3 ウェーブに共通する価値観に関わる設問 75 問を抽出する．

2.2 モデル

多言語の指示チューニングされた LLM として Llama3-70B-Instruct [17]，Qwen2.5-72B-Instruct [18]，GPT-5 [19]，DeepSeek-R1 [20] を対象とする．

2.3 プロンプト

指示チューニングされた LLM が学習時に用いる「指示＋例示」形式に合わせ，回答選択肢の比率分布を JSON 形式（例：1:31%, 2:4%, 3:30%, 4:35%）で出力させる．分布推定の安定化のため，5 件の few-shot 例を付与する．

2.4 分布表現

LLM の意見分布は，モデル自身に分布を明示的に生成させる verbalized distribution として表現する．価値観の意見分布は WVS の統計データ（各国 1,000 件以上の回答を含む）から計算する．

2.5 アライメント指標

分布間のアライメントは選択肢の順序を考慮し，かつ $[0,1]$ に正規化されることが望ましい．本研究では [4] の指標を採用し，設問 q の選択肢数を $|N|$ とし，LLM の分布 $D_M(q)$ と国 c の価値観分布 $D_c(q)$ の距離としてワッサースタイン距離（WD）を用いる．設問集合 Q に対するアライメントスコアは次式で定義する：

$$A(D_M, D_c; Q) = \frac{1}{|Q|} \sum_{q \in Q} \left(1 - \frac{\mathrm{WD}(D_M(q), D_c(q))}{|N| - 1}\right) \tag{1}$$

$|N| - 1$ で正規化することで指標は $[0,1]$ となり，1 が完全一致を表す．

3 各国の価値観とのアライメント（RQ1）

本節ではウェーブ 7 における LLM と各国の価値観分布のアライメントを測定し，(i) モデルレベル，(ii) 国レベル，(iii) 平均意見分布との差分の観点から分析する．

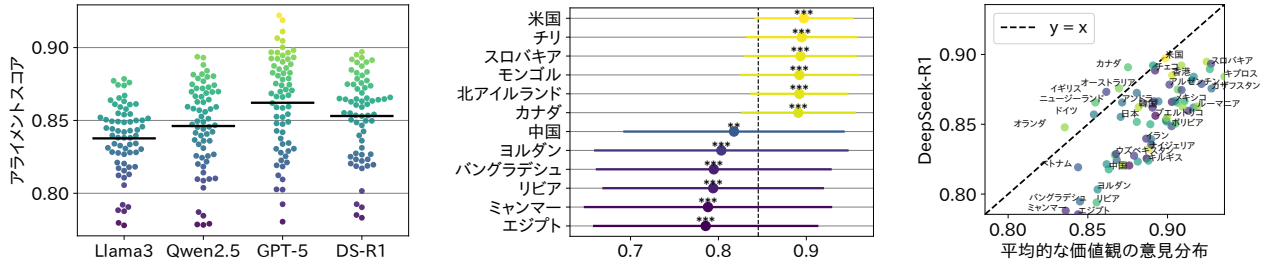
3.1 モデルレベル

図 2a は国ごとのアライメント点と平均（黒線）を示す．ウェーブ 7 における全体アライメントスコアを比較すると，GPT-5 が最も高い平均アライメントスコア（0.8621）を示し，DeepSeek-R1 も同様に優れた性能を発揮した．Llama3 と Qwen2.5 は商用モデルに比べてやや低い結果となった．全てのモデルのアラインメントスコアは 0.85 に近いが，国ごとに顕著な差異が見られ，単一モデルでは全ての国で高いアラインメントスコアを達成できないことを示している．

3.2 国レベル

図 2b は DeepSeek-R1 に対してアライメントが高い国／低い国の例を示す．米国・チリでは高いアラ

1) WVS の説明は付録 A を参照．



(a) LLM と各国とのアライメントスコア．　(b) 最初と最後の 6 カ国または地域．　(c) DeepSeek-R1 対 価値観．

図 2: **RQ1** の全体結果．(a) 各国に対する LLM のアライメントスコア（各点は国を表し，黒線は全ての国の平均アライメントスコアを表す）．「DS」は「DeepSeek」の略である．(b) DeepSeek-R1 の意見分布に対するアライメントスコアに基づく順位の上位 6 か国および下位 6 か国．点は全ての質問における当該国の平均アライメントスコアを表し，線は標準偏差を表す．***は p 値が 0.001 未満であること（t 検定）を示す．(c) 各国と DeepSeek-R1 の意見分布との関係，および平均的な価値観の意見分布とのアライメントスコアの関係．

イメント（例：0.8972，0.8948）で，エジプト・ミャンマーでは低いアライメント（例：0.7855，0.7881）となる．高いアライメントの国は標準偏差が小さく頑健である一方，低いアライメントの国は系統的ミスアライメントと不安定性が併存する傾向がある．この要因として，英語・スペイン語コンテンツの優勢により米国・カナダ・チリ等の支配的価値観へ同質化しやすいこと，非ラテン文字圏では政治・文化的データが学習コーパスで不足しやすいことが考えられる．またスロバキアやモンゴルでアライメントが高い点は，地域メディアや言語間の相互接続を捉えたコーパスを介した間接学習の可能性を示唆する．

3.3　平均価値観との差分レベル

価値観の多様性のため，すべての国に同時にアライメントすることは不可能である．そこで「平均的価値観分布へのアライメント」を一つの目標と仮定し，LLM が各国に対し過度／過小にアライメントしているかを分析する．図 2c では，横軸を平均分布へのアライメント，縦軸を DeepSeek-R1 へのアライメントとし，$y = x$ より上を過度アライメント，下を過小アライメントと解釈する．DeepSeek-R1 は米国に最も適切にアライメントし，一部の先進国では過度アライメントとなる．一方，チリ・スロバキア・モンゴル等は絶対値としては近いが，平均分布と比べるとさらに過小アライメントとなる．総じて，少数国には適切／過度アライメントであるが，多くの国（特にミャンマー・エジプト・リビア等の発展途上国）には過小アライメントとなる．

4　言語のステアラビリティ（RQ2）

4.1　ベースライン（既存ステアリング）

誘導能力は，目標人口集団の意見へ LLM を調整・アライメントさせる能力を指す [5]．既存手法として以下を用いる．

- **ペルソナ誘導** [4, 5]：対象国の一員であるかのように振る舞うよう LLM に指示を与え，集団の意見指向を模倣させる．
- **Few-shot ステアリング** [5]：ペルソナ設定に加え，実際の集団意見分布を例示し模倣させる．

4.2　提案：言語による価値観の誘導

言語は思考様式に影響し，話者の概念構造や世界観の差異を生むとされる [21, 22]．しかし，LLM における影響は未解明である．本研究では，プロンプト言語を英語から対象言語 l へ切り替えることで，当該言語話者の意見分布へ近づける「言語誘導」を検証する．タスク指示文は GPT-4 で翻訳後に人手確認し，few-shot 例は WVS の該当言語版質問票の翻訳を利用する．Durmus ら [12] が 3 言語・単一モデル・ロジット評価中心であったのに対し，本研究は 8 言語・複数モデル・verbalized distribution を用い，さらに他ステアリングとの組合せ効果も評価する．

4.3　言語・国の選定

WVS では多言語質問票が存在する国もあり，その場合，回答が単一言語話者の価値観を代表しない可能性がある．したがって，(i) 単一言語で調査さ

表 1: 異なる誘導手法における，日本およびドイツに対する LLM の文化表象スコア．付録 B にて追加結果を示す．

Method	Llama3	Qwen2.5	GPT-5	DS-R1
日本「日本語」				
No Steering	0.8513	0.8596	0.8744	0.8541
Persona Steering	0.8566	0.8631	0.8884	0.8680
Few-shot Steering	**0.8618**	0.8696	0.8920	0.8788*
Language Steering	0.8508	**0.8807***	**0.9053*****	**0.8964*****
ドイツ「ドイツ語」				
No Steering	0.8441	0.8541	0.8826	0.8543
Persona Steering	0.8510	0.8596	0.9129***	0.8879***
Few-shot Steering	0.8348	0.8619	0.8983	0.8581
Language Steering	**0.8683**	**0.8916*****	**0.9219*****	**0.9103*****

れた国であること，(ii) 当該言語が LLM で扱えること，を満たす言語・国を対象とする．対象はスペイン語，中国語，日本語，韓国語，ドイツ語，ロシア語，ベトナム語，ポルトガル語である．スペイン語は対象国が多いためアルゼンチン・チリ・ウルグアイを選び，他言語は 1 国ずつを選定する．

4.4 言語による誘導の有効性

表 2 は日本（日本語）とドイツ（ドイツ語）に対し，ステアリング方法別のアライメントを示す．さらに言語ステアリングが両国で最大の整合性を達成し，言語ステアリングが言語話者の意見模倣に有効であることを示す．したがって，意見指向タスクでは「対象言語を維持する」ことが価値観へのアライメントの観点で重要であり，多言語展開時にはこの点に注意しなければならない．

5 世代へのアライメント（RQ3）

人間の認知は社会的・歴史的性格を持ち，時間遷移により再形成され続ける [23, 24]．したがって，LLM がどの世代の価値観を反映しているかは不明である．本研究では現代の価値観へのアライメントが望ましいと仮定し，世代的観点から評価する．

5.1 手法

RQ1 で多くの国が過小アライメントであったため，比較のため「平均分布に対して LLM が十分近い国」をフィルタリングする．具体的には次式で国集合 C^* を定義する：

$$C^* = \left\{ c \in C \mid \left| A_M^{(c)} - A_{avg}^{(c)} \right| < \tau \right\}. \qquad (2)$$

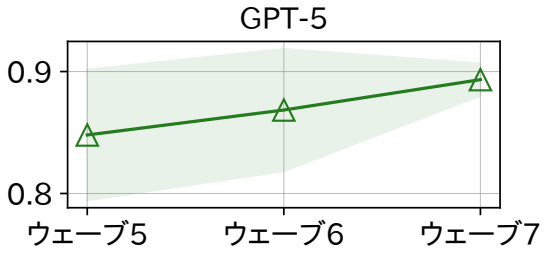


図 3: Eq. (2) を用いて国をフィルタリングした上で，各ウェーブにおける GPT-5 の各国に対する平均アライメントスコアの推移．塗りつぶし部分は当該ウェーブにおけるアライメントスコアの標準偏差を示す．付録 C にて追加結果を提供する．

ここで $A_M^{(c)}$ は国 c と LLM のアライメントスコアで，$A_{avg}^{(c)}$ は国 c と平均分布のアライメントスコアである．少なくとも 5 か国を確保するメタに，$\tau = 0.02$ とした．

5.2 結果

図 3 は GPT-5 について，ウェーブ 5–7 での平均アライメント推移を示す．両モデルともに最新ウェーブ（ウェーブ 7）の価値観に最もアライメントする．要因として，(i) アライメント学習に用いられる人間フィードバックが現代のアノテータに由来し，近年の社会規範に近い行動が強化されやすいこと，(ii) モバイルインターネットや SNS の普及により近年のウェブ上のデータが増加したこと，が考えられる．また標準偏差は「価値観カバレッジ」を反映し，最新ウェーブほど偏差が小さいことから，近年価値観のカバレッジがより良い一方，過去価値観への整合には困難が残ることが示唆される．

6 結論

本研究では WVS を用いて，LLM との世界規模の価値観とのアライメントを国・言語・時期の 3 項目で検討する枠組みを提案した．主な知見は，(1) 少数国には適切／過剰にアライメントする一方，多数国には過小アライメントであること，(2) 質問票言語に合わせたプロンプト言語による誘導が国別価値観への誘導に有効であること，(3) LLM は過去より現代の価値観へ整合しやすいことである．本研究は LLM の価値観へのアライメント研究に対し，世界規模・多言語・世代という観点からの基盤を提供する．

謝辞

本研究は、JST 国家戦略分野の若手研究者及び博士後期課程学生の育成事業 (博士後期課程学生支援)JPMJBS2407 の支援および JSPS 科研費 JP23K28144 の助成を受けたものです.

参考文献

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. **arXiv preprint arXiv:2303.08774**, 2023.

[2] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. **arXiv preprint arXiv:2303.12712**, 2023.

[3] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. **Advances in neural information processing systems**, Vol. 35, pp. 27730–27744, 2022.

[4] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In **International Conference on Machine Learning**, pp. 29971–30004. PMLR, 2023.

[5] Nicole Meister, Carlos Guestrin, and Tatsunori Hashimoto. Benchmarking distributional alignment of large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, **Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)**, pp. 24–49, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[6] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. **OpenAI blog**, Vol. 1, No. 8, p. 9, 2019.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. **Advances in neural information processing systems**, Vol. 33, pp. 1877–1901, 2020.

[8] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. **arXiv preprint arXiv:2109.01652**, 2021.

[9] Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. Multitask prompted training enables zero-shot task generalization. **arXiv preprint arXiv:2110.08207**, 2021.

[10] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. **Advances in neural information processing systems**, Vol. 30, , 2017.

[11] Winter Mason and Siddharth Suri. Conducting behavioral research on amazon's mechanical turk. **Behavior research methods**, Vol. 44, No. 1, pp. 1–23, 2012.

[12] Esin Durmus, Karina Nguyen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. Towards measuring the representation of subjective global opinions in language models. **arXiv preprint arXiv:2306.16388**, 2023.

[13] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In **Proceedings of the 2021 ACM conference on fairness, accountability, and transparency**, pp. 610–623, 2021.

[14] Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. Investigating cultural alignment of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 12404–12422, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[15] Nicholas Sukiennik, Chen Gao, Fengli Xu, and Yong Li. An evaluation of cultural value alignment in llm. **arXiv preprint arXiv:2504.08863**, 2025.

[16] Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Juan Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bi Puranen. World values survey: Round seven – country-pooled datafile version 6.0, 2022. Dataset.

[17] AI@Meta. Llama 3 model card. https://www.llama.com/docs/model-cards-and-prompt-formats/meta-llama-3, 2024. Accessed: 2025-12-04.

[18] Qwen Team. Qwen2.5: A party of foundation models, September 2024.

[19] OpenAI. Introducing gpt-5. https://openai.com/index/introducing-gpt-5, August 2025. Accessed: 2025-12-04.

[20] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. **arXiv preprint arXiv:2501.12948**, 2025.

[21] Jonathan H. Turner. **On the Origins of Human Emotions**. Stanford University Press, Redwood City, 2000.

[22] Alessio Plebe and M Vivian. When language shapes perception. **Rivista Italiana di Filosofia del Linguaggio**, Vol. 9, No. 2, 2015.

[23] Karl Mannheim. **Ideology and utopia**. Routledge, 2013.

[24] Peter Berger and Thomas Luckmann. The social construction of reality. In **Social theory re-wired**, pp. 110–122. Routledge, 2016.

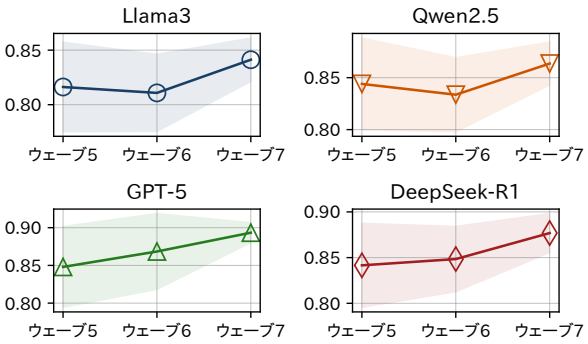


図 4: 平均アラインメントスコアの推移.

A World Values Survey（WVS）

本節では，WVS が本研究に適する理由を，世界規模，言語，世代の 3 項目から述べる．

A.1 グローバル次元

WVS は価値判断を伴う高主観的質問を多く含み，LLM の意見を探るプロンプトと整合する．最新ウェーブは 66 の国・地域を対象とし，各国につき 1,000 件超の回答を含むため，国別の価値観の意見分布との比較が可能である．

A.2 言語次元

WVS 質問票は，人口の 15%以上が第一言語として使用する言語に翻訳され，さらに時系列比較のために過去ウェーブの翻訳を流用するなどの工夫がなされる．最新ウェーブでは 50 以上の言語翻訳が提供され，LLM が対応する言語の多様性を確保できる．

A.3 時間次元

WVS は 1981 年以降 7 ウェーブで調査を継続し，多くの設問で機能的な一貫性を保つ．これにより，異なる歴史的期間での長期的傾向として LLM と価値観のアライメントを評価できる．

B RQ2 の追加結果

表 2 は RQ2 の追加結果を示している．

C RQ3 の追加結果

図 4 は，すべての大規模言語モデルの平均アラインメントスコアの推移．

表 2: 中国，韓国，ロシア，ベトナム，ブラジル，アルゼンチン，チリ，ウルグアイと LLM 間のアライメントスコア（異なるステアリング方法下）．

Method	Llama3	Qwen2.5	GPT-5	DS-R1
中国「中国語」				
No Steering	0.8235	0.8273	0.8348	0.8174
Persona Steering	0.8439	0.8316	0.8953***	0.8625**
Few-shot Steering	**0.8666****	0.8596	0.8996***	0.8884***
Language Steering	0.8652**	**0.8838*****	**0.9041*****	**0.8976*****
韓国「韓国語」				
No Steering	0.8415	0.8632	0.8659	0.8606
Persona Steering	0.8357	0.8640	0.8814*	0.8875**
Few-shot Steering	0.8591	0.8646	0.8718	0.8637
Language Steering	**0.8749****	**0.8790**	**0.8960***	**0.9019*****
ロシア「ロシア語」				
No Steering	0.8415	0.8644	0.8862	0.8780
Persona Steering	0.8513	0.8854	0.9180***	**0.9018***
Few-shot Steering	0.8801***	0.8865	0.9172***	0.8939
Language Steering	**0.8874*****	**0.8947****	**0.9247*****	0.8975*
ベトナム「ベトナム語」				
No Steering	0.7992	0.8094	0.8111	0.8164
Persona Steering	0.8078	0.7987	0.8337	0.8253
Few-shot Steering	0.8498***	0.8190	0.8450*	0.8396
Language Steering	**0.8668*****	**0.8718*****	**0.8713*****	**0.8674*****
ブラジル「ポルトガル語」				
No Steering	0.8455	0.8693	0.8834	0.8655
Persona Steering	0.8380	0.8820	0.8983	0.8815
Few-shot Steering	**0.8777****	0.8842	0.9047*	0.8805
Language Steering	0.8755**	**0.8869**	**0.9049****	**0.8939***
アルゼンチン「スペイン語」				
No Steering	0.8397	0.8698	0.8943	0.8766
Persona Steering	0.8429	0.8774	0.9101	0.8929
Few-shot Steering	0.8627	**0.8984***	**0.9153****	0.8947
Language Steering	**0.8684**	0.8983**	0.9080	**0.8985***
チリ「スペイン語」				
No Steering	0.8728	0.8936	0.9051	0.8937
Persona Steering	0.8836	0.9040	0.9152	0.9091
Few-shot Steering	0.8987**	**0.9075**	**0.9230***	0.8989
Language Steering	**0.9019*****	0.9020	0.9210*	**0.9116***
ウルグアイ「スペイン語」				
No Steering	0.8607	0.8766	0.8893	0.8639
Persona Steering	0.8584	0.8797	0.8946	0.8900**
Few-shot Steering	0.8642	0.8849	**0.9050**	0.8804
Language Steering	**0.8709**	**0.8899**	0.8989	**0.8923*****