Measuring Political Bias and Temporal Opinion Alignment of Large Language Models in Low-Resource Languages

Lanka Gamage

Supervised By Dr. Nisansa de Silva

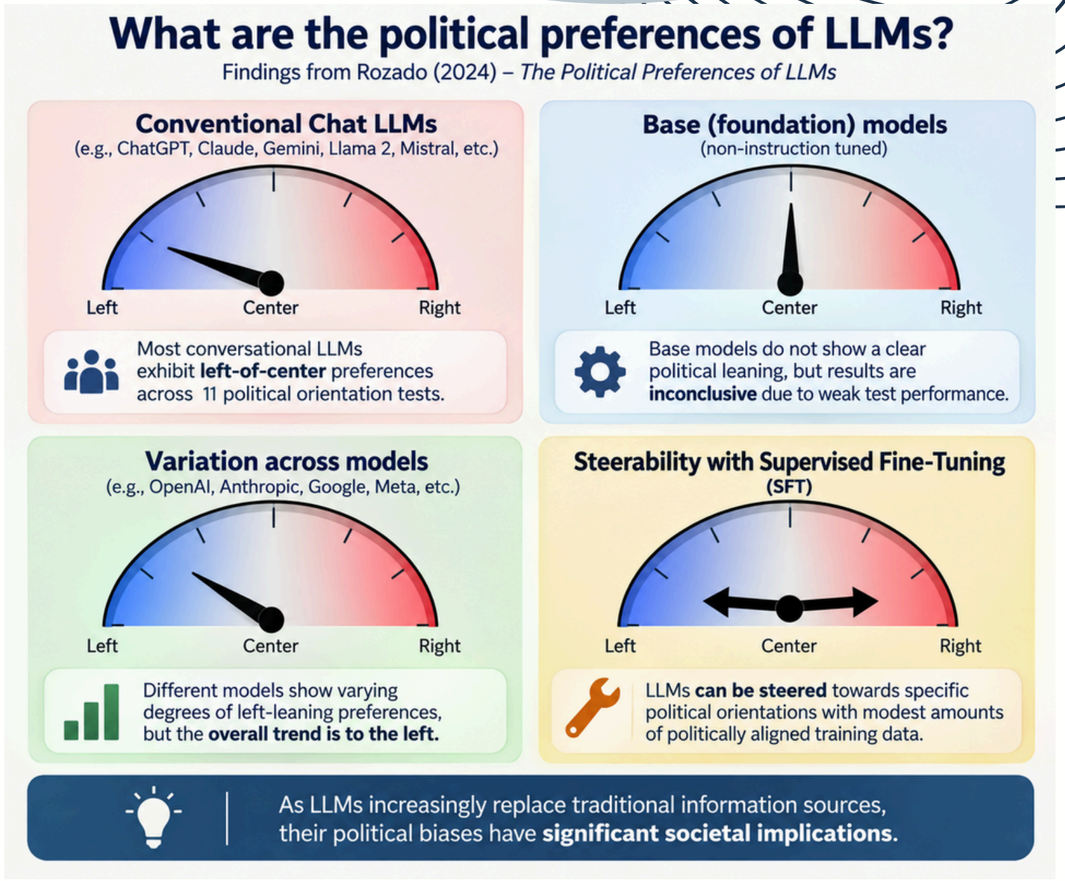
Agenda

- Introduction
  - Measuring Political bias in LLMs
  - The Sri Lankan Political Context
  - Sinhala as a Low-Resource Language
  - A Natural Experiment:2021-2026

- Research Gap
- Research Problem
- Research Objective
- Proposed Methodology(Stages 1-3)
- Expected Contributions
- Challenges and Resources
- References

Introduction

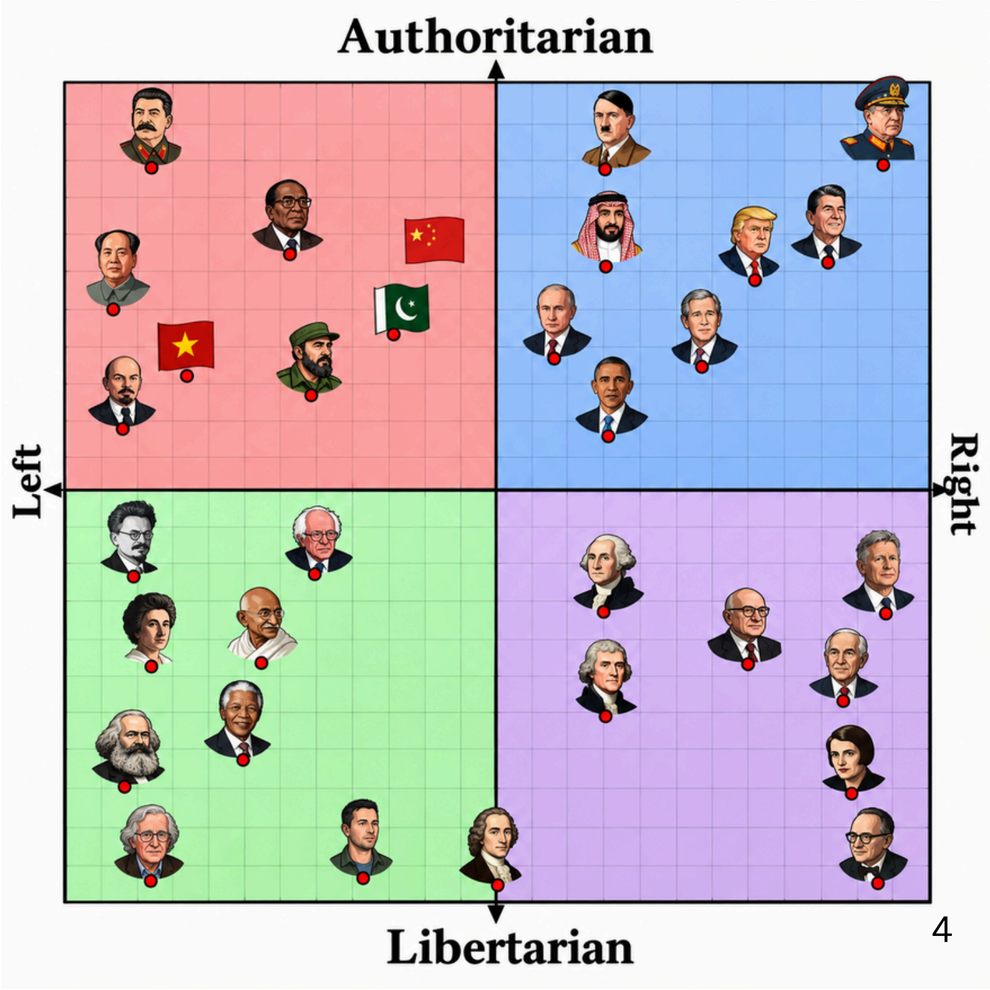
- Large Language models are now used by millions of people to read the news, understand policies, and form political opinions.
- LLMs are not politically neutral: on standard political orientation tests, most instruction-tuned models are diagnosed as left-of-centre [1][2].
- They also represent the opinions of some demographic groups far better than others [3].
- These learnings originate in the pretraining corpus and propagate into unfair downstream behavior [4].
- Experiments with human participants show that interacting with a politically biased model changes what people think and how they decide [5][6].



[1] Rozado, D., 2024. The Political Preferences of LLMs. PLoS ONE, 19(7), e0306621.
[2] Motoki, F., Pinho Neto, V. and Rodrigues, V., 2024. More Human than Human: Measuring ChatGPT Political Bias. Public Choice, 198, pp.3-23.
[3] Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P. and Hashimoto, T., 2023. Whose Opinions Do Language Models Reflect? In ICML, PMLR 202, pp.29971-30004.
[4] Feng, S., Park, C.Y., Liu, Y. and Tsvetkov, Y., 2023. From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. In ACL, pp.11737-11762.
[5] Fisher, J., Feng, S., Aron, R., Richardson, T., Choi, Y., Fisher, D.W., Pan, J., Tsvetkov, Y. and Reinecke, K., 2025. Biased LLMs Can Influence Political Decision-Making. In ACL.
[6] Potter, Y., Lai, S., Kim, J., Evans, J. and Song, D., 2024. Hidden Persuaders: LLMs' Political Leaning and Their Influence on Voters. In EMNLP, pp.4244-4275.

Measuring Political Bias in LLMs

- The Dominant approach administers survey instruments designed for humans to models
- The most widely Used instrument is the Political Compass Test: a two-dimensional map of ideology [7].
- It remains valuable for its clarity and its comparability across studies - and it is the framework this research adopts and tests.



How it works

A model is asked to agree or disagree with a fixed set of propositions, its answers are aggregated into a coordinate on two axes: economic(left right) and social (authoritarian libertarian).

[7] Rottger, P., Hofmann, V., Pyatkin, V., Hinck, M., Kirk, H.R., Schutze, H. and Hovy, D., 2024. Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in LLMs. In ACL, pp.15295-15311.

The Sri Lankan Political Context

What the Political Compass assess

- State vs. Market economics.
- Authoritarian vs. Libertarian social values.
- A stable two-axis ideological space.
- Party competition along left-right lines.
- Voters positioned by economic self-interest

What Sri Lankan politics is organised around

- Ethnicity and religion.
- Language policy.
- Centralisation vs. devolution.
- Post-war reconciliation.
- Personality - and patronage - based loyalties.

- **Open Question:** can the two Political Compass axes represent political opinion in Sinhala discourse at all? Testing this assumption is a primary objective of this research.

Sinhala as a Low-Resource Language

- Sinhala is spoken by approximately 17 million people and remains low-resource in NLP[15]
- LLM evaluation for Sinhala has only recently begun: general knowledge [16],societal values[17], and physical common sense[18].
- Existing multilingual political studies simply translate Western instruments into other languages, without checking whether the instrument itself makes sense in the new context[19].
- Meanwhile LLM-based tools are already in use in Sri Lankan schools[20],newsrooms, and social media.

For Sinhala there is:

No annotated dataset

of Sinhala political opinion

No benchmark

for Sinhala political text analysis

No study

of how LLMs handle Sinhala political content

[15] De Silva, N., 2019. Survey on Publicly Available Sinhala Natural Language Processing Tools and Research. arXiv preprint arXiv:1906.02358.
[16] Pramodya, A. et al., 2025. SinhalaMMLU: A Comprehensive Benchmark for Evaluating Multitask Language Understanding in Sinhala. arXiv:2509.03162.
[17] Muthugala, N., Ranathunga, S., de Silva, N. et al., 2026. LKValues: Aligning Large Language Models with Sri Lankan Societal Values. arXiv:2607.20410.
[18] De Silva, N. et al., 2026. Sinhala Physical Common Sense Reasoning Dataset for Global PIQA. arXiv:2602.02207.
[19] Helwe, C., Balalau, O. and Ceolin, D., 2025. Navigating the Political Compass: Evaluating Multilingual LLMs Across Languages and Nationalities.
[20] Karunarathne, B., Nanayakkara, V., de Silva, N. and Ambegoda, T., 2025. Equipping School Teachers with Technology. In EDULEARN25 Proceedings, pp.9925-9931. IATED.

A Rare Opportunity: Sri Lanka, 2021-2026

Between 2021 and 2026 Sri Lanka went through an economic collapse, a mass protest movement, and a complete electoral realignment. Public opinion changed dramatically in a very short time.



LLMs, whose training data stops at different cutoff dates along this timeline, must have absorbed different snapshots of this period.

Does an LLM describe public opinion as it is now, or as it was when its training data was collected?

Research Problem

How trustworthy are large language models in representing Sinhala political opinion, and how can this be measured, given that the language lacks the annotated resources, the validated measurement dimensions, and the evaluation evidence that such measurement requires?

The problem unfolds at three levels, each a precondition for the next:

1 Resources

No annotated corpus, and no evidence that Western measurement dimensions transfer to Sinhala political discourse.

2 Interpretation

Unknown whether LLMs can reliably infer the political orientation of Sinhala content, and whether their errors are politically symmetric.

3 Representation

Unknown whether the picture of Sinhala opinion held by an LLM corresponds to the present or to a moment fixed by its training cutoff.

Research Gaps and Objectives

Five documented gaps across three levels, and the four objectives that close them.

LEVEL	RESEARCH GAP	OBJECTIVE
Resources	**No corpus** No annotated corpus of Sinhala political discourse [21]. **No validated dimensions** Unknown whether the Political Compass axes apply to Sinhala discourse at all.	O1 Create a dataset of Sinhala political discourse comments and annotate them on the economic and social axes of the Political Compass.
Interpretation	**No annotator evidence** Unknown whether LLMs can assign continuous ideology scores that match human judgment.	O2 Evaluate whether LLMs, across model families and scales, can map Sinhala political discourse onto a continuous two-dimensional ideology space in agreement with human annotators.
Representation	**No temporal framework** No way to locate a model's political knowledge in time or to shift it deliberately.	O3 Measure the alignment between LLM-expressed political opinions and real-world Sinhala political opinion distributions. O4 Determine whether fine-tuning on temporally segmented political discourse measurably shifts the political alignment expressed by a model.

[21] Barkhordar, E., Thapa, S., Maratha, A. and Naseem, U., 2024. Why the Unexpected? Dissecting the Political and Economic Bias in Persian Small and Large Language Models. In SIGUL @ LREC-COLING.
[13] Ng, L.H.X., Cruickshank, I.J. and Lee, R.K.-W., 2025. Examining the Influence of Political Bias on Large Language Model Performance in Stance Classification. In ICWSM, 19, pp.1315-1328.
[22] Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Liu, Q., Liu, T. and Sui, Z., 2024. Large Language Models Are Not Fair Evaluators. In ACL, pp.9440-9450.

Proposed Methodology: Three Connected Stages

Stage 1

Sinhala Political Discourse Corpus

Collect, sample and human-annotate YouTube political comments on the two Political Compass axes, with a temporal dimension.

Stage 2

LLMs as Political Annotators

Test whether models of different families, sizes and access types agree with human annotators on the same held-out comments.

Stage 3

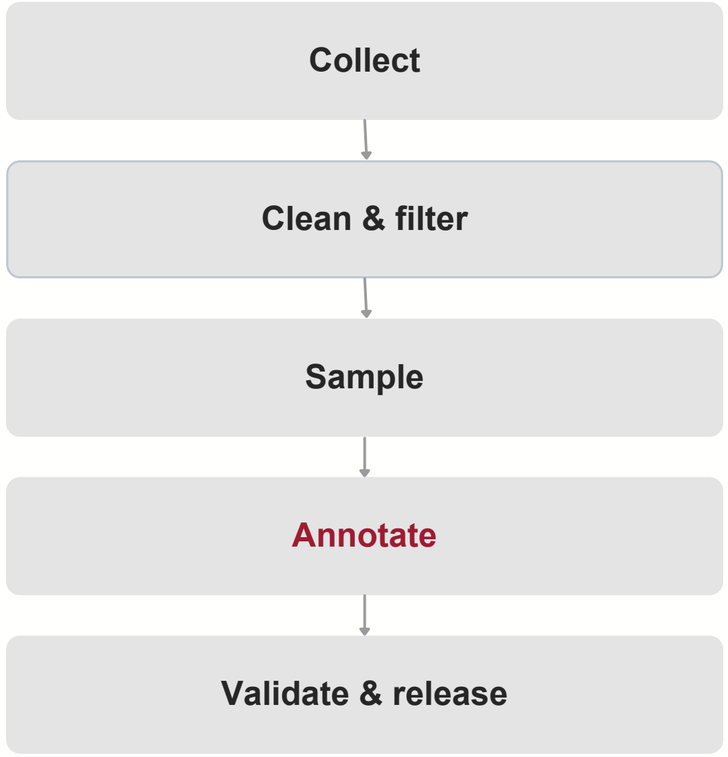
Temporal Political Alignment

Locate each model's picture of Sinhala opinion in time, and test whether fine-tuning on a time slice shifts expressed opinion.

Each stage is a precondition for the next: the corpus grounds the annotator evaluation, and both ground the temporal study. Approaches may be refined as findings from earlier stages inform later ones.

Stage 1: Sinhala Political Discourse Corpus

- Comments collected from the YouTube channels of Sri Lankan news and political commentary sources through the YouTube Data API.
- Coverage spans the economic crisis, the 2022 protest movement, the 2024 electoral cycle, and the present situation.
- Comments are cleaned, filtered for language and quality, and sampled to ensure coverage across time periods and sources.
- Qualified native Sinhala speakers annotate the sampled comments on the economic and social axes of the Political Compass.
- Quality is ensured through multiple annotators per comment and standard inter-annotator agreement measures.
- Data collection and annotation follow institutional ethics review; the corpus will be released for research use.



Progress to date

550,000+	20+	2026
Comments Collected So far	Sri Lankan Channels Coverd	First Period Fully Covered

Stage 2: Evaluating LLMs as Political Annotators

- A panel of models covering different families, sizes and access types(closed and open-weight) labels a held-out portion of the stage 1 corpus.
- Prompting conditions: zero-shot, few-shot, and guideline-conditioned settings.
- An additional English-translation condition helps separate limited Sinhala capability from politically driven behaviour [23][24][25].
- Because outputs on political content are prompt-sensitive [7][8], evaluation uses multiple prompt variants and repeated runs, with variability reported.

Analysis goes beyond accuracy - the nature of the errors is the finding:

Random noise?

Errors distributed evenly across comments and camps.

Capability failure?

Errors driven by weak Sinhala language ability rather than politics.

Systematic skew?

Moderation toward neutrality, or asymmetric treatment of political camps [13][22].

[23] Aravinda, H.W.K., Sirajudeen, R., Karunathilake, S., de Silva, N., Kaur, R. and Ranathunga, S., 2025. SinLlama - A Large Language Model for Sinhala. In MERCon, pp.617-622. IEEE.
[24] Sandaruwan, P.K.U.I., Fonseka, N.M.A., Salwathura, P.C., Alwis, D.S., Wanigasekara, C. and Velmanickam, L., 2026. End-to-End Adaptation of LLMs for Low-Resource Languages. IEEE Access.
[25] Unjum, N., Evert, S., Sarveswaran, K., Weerasinghe, R. and Jayatilleke, N., 2025. LLMs for Low-Resource Languages: A Survey.
[7] Rottger, P., Hofmann, V., Pyatkin, V., Hinck, M., Kirk, H.R., Schutze, H. and Hovy, D., 2024. Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in LLMs. In ACL, pp.15295-15311.
[8] Ceron, T., Falk, N., Baric, A., Nikolaev, D. and Pado, S., 2024. Beyond Prompt Brittleness: Evaluating the Reliability and Consistency of Political Worldviews in LLMs. TACL, 12.
[13] Ng, L.H.X., Cruickshank, I.J. and Lee, R.K.-W., 2025. Examining the Influence of Political Bias on Large Language Model Performance in Stance Classification. In ICWSM, 19, pp.1315-1328.
[22] Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Liu, Q., Liu, T. and Sui, Z., 2024. Large Language Models Are Not Fair Evaluators. In ACL, pp.9440-9450.

Stage 3: Temporal Political Alignment of LLMs

Using the political opinions observed in the corpus as reference data, two connected experiments

EXPERIMENT 1

Locating the model in time

- Off-the-shelf models are prompted to express positions on corpus issues and entities, and to estimate public opinion on them.
- Prompted with and without explicit time anchoring.
- Model opinions are compared against corpus data for each period, locating each model's picture of Sinhala opinion in time.

EXPERIMENT 2

Shifting the model in time

- An open-weight model is fine-tuned separately on comment data from different time periods.
- Opinions generated by each time-sliced model are compared statistically, and against period-specific and present-day references.
- Appropriate controls separate genuine opinion shift from other effects.

Motivation: training-data composition is already known to shape expressed politics [4][26] - this stage manipulates its temporal composition.

[4] Feng, S., Park, C.Y., Liu, Y. and Tsvetkov, Y., 2023. From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. In ACL, pp.11737-11762.
[26] Stammbach, D., Widmer, P., Cho, E., Gulcehre, C. and Ash, E., 2024. Aligning Large Language Models with Diverse Political Viewpoints. In EMNLP.

Expected Contributions

Resource

The first annotated dataset of Sinhala political discourse, released for research use, together with an annotation framework that other non-Western contexts can reuse.

Evidence

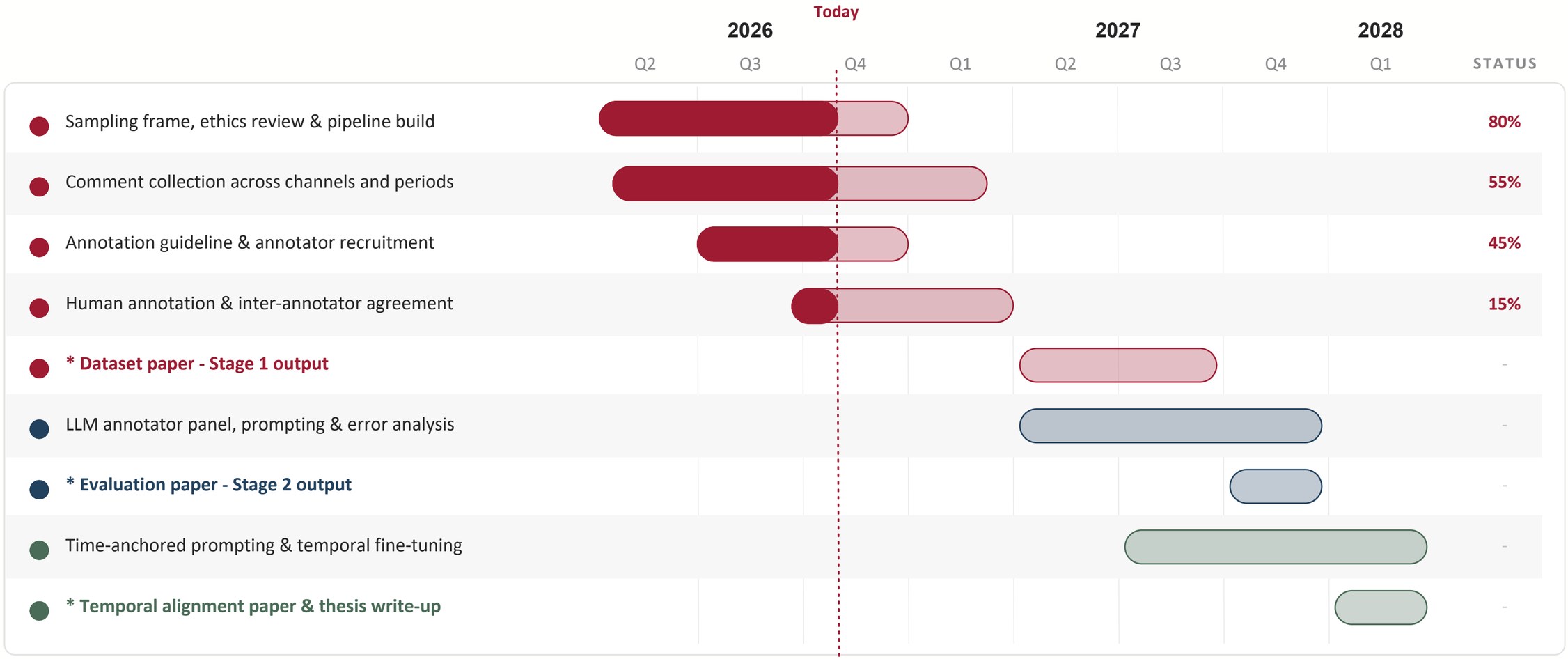
Evidence on whether LLMs can be trusted to annotate political content in low-resource languages - before such pipelines are relied upon.

Method

The first measurement of how the political knowledge held by an LLM drifts over time, and whether the temporal window of fine-tuning data shifts it.

Research Plan

A 24-month plan across the three stages. Stage 1 collection and annotation set-up are already under way.



Each stage delivers one publishable output; quarters are indicative and will be confirmed against the enrolment calendar.

Anticipated Challenges

Annotation subjectivity

Political annotation is inherently contested; reaching acceptable inter-annotator agreement on two continuous axes is demanding.

Framework transferability

The Political Compass axes may not cleanly capture ethnicity-, religion- and devolution-based cleavages [9] - a negative result is itself a finding.

Noisy, code-mixed text

YouTube comments are informal, colloquial and often Sinhala-English code-mixed; Sinhala LLM capability is still limited [23][24].

Ethics and privacy

Public political speech from identifiable users requires careful anonymisation, institutional ethics review, and a responsible release policy.

Temporal confounds

Separating genuine opinion shift caused by the fine-tuning window from topic, vocabulary and volume differences across periods.

Compute

Repeated multi-prompt runs across a model panel, plus separate fine-tuning runs per time slice.

[9] Kamal, S., Prakash, L.P.Y., Rafiuddin, S.M., Rakib, M., Sen, A. and Ray Choudhury, S., 2025. A Detailed Factor Analysis for the Political Compass Test: Navigating Ideologies of Large Language Models.
[23] Aravinda, H.W.K., Sirajudeen, R., Karunathilake, S., de Silva, N., Kaur, R. and Ranathunga, S., 2025. SinLlama - A Large Language Model for Sinhala. In MERCon, pp.617-622. IEEE.
[24] Sandaruwan, P.K.U.I., Fonseka, N.M.A., Salwathura, P.C., Alwis, D.S., Wanigasekara, C. and Velmanickam, L., 2026. End-to-End Adaptation of LLMs for Low-Resource Languages. IEEE Access.

Resources

Data	Sinhala political comments collected through the YouTube Data API from Sri Lankan news and political commentary channels.
Annotators	Qualified native Sinhala speakers, working under a documented annotation guideline with multiple annotators per comment.
Models	Open-weight models from HuggingFace for fine-tuning, together with API access to closed models for the annotator panel.
Computational resources	GPU cluster available in the CSE department, University of Moratuwa.
Governance	Institutional ethics review for data collection, annotation and release.

References

[1] Rozado, D., 2024. The Political Preferences of LLMs. PLoS ONE, 19(7), e0306621.
[2] Motoki, F., Pinho Neto, V. and Rodrigues, V., 2024. More Human than Human: Measuring ChatGPT Political Bias. Public Choice, 198, pp.3-23.
[3] Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P. and Hashimoto, T., 2023. Whose Opinions Do Language Models Reflect? In ICML, PMLR 202, pp.29971-30004.
[4] Feng, S., Park, C.Y., Liu, Y. and Tsvetkov, Y., 2023. From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. In ACL, pp.11737-11762.
[5] Fisher, J., Feng, S., Aron, R., Richardson, T., Choi, Y., Fisher, D.W., Pan, J., Tsvetkov, Y. and Reinecke, K., 2025. Biased LLMs Can Influence Political Decision-Making. In ACL.
[6] Potter, Y., Lai, S., Kim, J., Evans, J. and Song, D., 2024. Hidden Persuaders: LLMs' Political Leaning and Their Influence on Voters. In EMNLP, pp.4244-4275.
[7] Rottger, P., Hofmann, V., Pyatkin, V., Hinck, M., Kirk, H.R., Schutze, H. and Hovy, D., 2024. Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in LLMs. In ACL, pp.15295-15311.
[8] Ceron, T., Falk, N., Baric, A., Nikolaev, D. and Pado, S., 2024. Beyond Prompt Brittleness: Evaluating the Reliability and Consistency of Political Worldviews in LLMs. TACL, 12.
[9] Kamal, S., Prakash, L.P.Y., Rafiuddin, S.M., Rakib, M., Sen, A. and Ray Choudhury, S., 2025. A Detailed Factor Analysis for the Political Compass Test: Navigating Ideologies of Large Language Models.
[10] Faulborn, M., Sen, I., Pellert, M., Spitz, A. and Garcia, D., 2025. Only a Little to the Left: A Theory-Grounded Measure of Political Bias in Large Language Models. In ACL, pp.31684-31704.
[11] Rottger, P., Hinck, M., Hofmann, V., Hackenburg, K., Pyatkin, V., Brahman, F. and Hovy, D., 2026. IssueBench: Millions of Realistic Prompts for Measuring Issue Bias in LLM Writing Assistance.
[12] Jakob, C., Harbecke, D., Parschan, P., Wenzel Neves, P. and Schmitt, V., 2025. PolBiX: Detecting LLMs' Political Bias in Fact-Checking through X-phemisms. In Findings of EMNLP, pp.17192-17210.
[13] Ng, L.H.X., Cruickshank, I.J. and Lee, R.K.-W., 2025. Examining the Influence of Political Bias on Large Language Model Performance in Stance Classification. In ICWSM, 19, pp.1315-1328.
[14] Elbouanani, A., Dufraisse, E. and Popescu, A., 2025. Analyzing Political Bias in LLMs via Target-Oriented Sentiment Classification. In Findings of ACL, pp.15476-15505.
[15] De Silva, N., 2019. Survey on Publicly Available Sinhala Natural Language Processing Tools and Research. arXiv preprint arXiv:1906.02358.
[16] Pramodya, A. et al., 2025. SinhalaMMLU: A Comprehensive Benchmark for Evaluating Multitask Language Understanding in Sinhala. arXiv:2509.03162.
[17] Muthugala, N., Ranathunga, S., de Silva, N. et al., 2026. LKValues: Aligning Large Language Models with Sri Lankan Societal Values. arXiv:2607.20410.
[18] De Silva, N. et al., 2026. Sinhala Physical Common Sense Reasoning Dataset for Global PIQA. arXiv:2602.02207.
[19] Helwe, C., Balalau, O. and Ceolin, D., 2025. Navigating the Political Compass: Evaluating Multilingual LLMs Across Languages and Nationalities.

Refereces (cont.)

[20] Karunarathne, B., Nanayakkara, V., de Silva, N. and Ambegoda, T., 2025. Equipping School Teachers with Technology. In EDULEARN25 Proceedings, pp.9925-9931. IATED.
[21] Barkhordar, E., Thapa, S., Maratha, A. and Naseem, U., 2024. Why the Unexpected? Dissecting the Political and Economic Bias in Persian Small and Large Language Models. In SIGUL @ LREC-COLING.
[22] Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Liu, Q., Liu, T. and Sui, Z., 2024. Large Language Models Are Not Fair Evaluators. In ACL, pp.9440-9450.
[23] Aravinda, H.W.K., Sirajudeen, R., Karunathilake, S., de Silva, N., Kaur, R. and Ranathunga, S., 2025. SinLlama - A Large Language Model for Sinhala. In MERCon, pp.617-622. IEEE.
[24] Sandaruwan, P.K.U.I., Fonseka, N.M.A., Salwathura, P.C., Alwis, D.S., Wanigasekara, C. and Velmanickam, L., 2026. End-to-End Adaptation of LLMs for Low-Resource Languages. IEEE Access.
[25] Unjum, N., Evert, S., Sarveswaran, K., Weerasinghe, R. and Jayatilleke, N., 2025. LLMs for Low-Resource Languages: A Survey.
[26] Stammbach, D., Widmer, P., Cho, E., Gulcehre, C. and Ash, E., 2024. Aligning Large Language Models with Diverse Political Viewpoints. In EMNLP.
[27] Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S.R. et al., 2024. Towards Understanding Sycophancy in Language Models. In ICLR.
[28] Kaur, A., 2025. Echoes of Agreement: Argument Driven Sycophancy in Large Language Models. In Findings of EMNLP, pp.22803-22812.
[29] Nadeem, A., Dras, M. and Naseem, U., 2025. Steering Towards Fairness: Mitigating Political Stance Bias in LLMs. In CASE Workshop, pp.52-61.
[30] D'Avenia, S. and Basile, V., 2025. Quantifying the Influence of Irrelevant Contexts on Political Opinions Produced by LLMs. In ACL Student Research Workshop.
[31] Azzopardi, L. and Moshfeghi, Y., 2025. POW: Political Overton Windows of Large Language Models. In Findings of EMNLP, pp.24767-24773.
[32] Bang, Y., Chen, D., Lee, N. and Fung, P., 2024. Measuring Political Bias in Large Language Models: What Is Said and How It Is Said. In ACL, pp.11142-11159.
[33] Yoo, J. and Shin, Y., 2025. Fair or Framed? Political Bias in News Articles Generated by LLMs. In EMNLP, pp.16904-16930.
[34] Zhang, T., 2025. Probing Political Ideology in Large Language Models: How Latent Political Representations Generalize Across Tasks. In Findings of EMNLP, pp.23349-23360.
[35] Fulay, S., Brannon, W., Mohanty, S., Overney, C., Poole-Dayan, E., Roy, D. and Kabbara, J., 2024. On the Relationship Between Truth and Political Bias in Language Models. In EMNLP.
[36] Hao, S., Cao, Y., Yu, L., Ren, Y. and Xiong, D., 2026. D2VBench: Benchmarking Large Language Models with Value Dilemmas in Daily Scenarios. arXiv:2607.19834.
[37] Supryadi, Irfan, Julianti, Martin, D.K., Fernando, J., Ren, Y. and Xiong, D., 2026. Pancasila-Dilemmas: Evaluating LLMs on Indonesian Human Value Dilemmas Grounded in Pancasila. arXiv:2607.18066.
[38] Liu, R., Jia, C., Wei, J., Xu, G. and Vosoughi, S., 2022. Quantifying and Alleviating Political Bias in Language Models. Artificial Intelligence, 304, 103654.

Thank You!

Presented By Lanka Gamage