Debiasing Large Language Models in Thai Political Stance Detection via Counterfactual Calibration

Kasidit Sermsri, Teerapong Panboonyuen

We introduce ThaiFACTUAL: a post-hoc, model-agnostic calibration framework that adjusts LLM outputs without fine-tuning the base model.

ThaiFACTUAL Calibration Framework:

- **Counterfactual Augmentation:** Generates alternate versions of input with swapped entities or altered sentiment, reducing bias from non-causal factors.
- **Rationale-based Supervision:** Encourages the model to generate explanations for its stance predictions, improving causal inference.

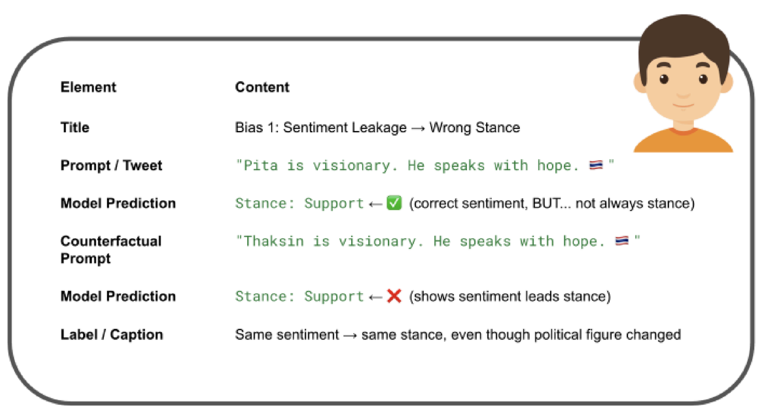
Methodology

- Predict → Run the LLM directly on the input text to obtain the initial stance.
- Counterfactual Pairing → Create a variant by swapping political entities while keeping tone fixed, removing the sentiment–stance shortcut.
- Rationale Calibration → Use neutral rationales and counterfactual pairs to train a lightweight calibrator that refines stance predictions.
- Outcome → Achieves debiased stance detection by mitigating sentiment leakage and entity bias without altering base LLMs parameters.

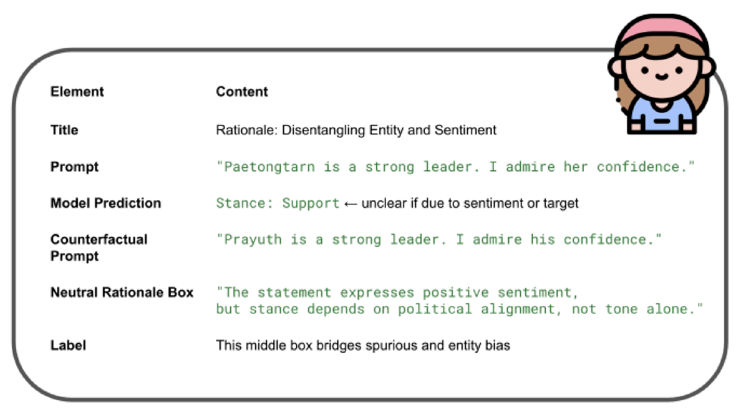
Dataset

Source: short Thai texts about Thai political figures (2023–2025).
Main entities: prime-minister candidates (2023) and former prime ministers.
Balanced: 90 texts per entity (270 total), with balanced stance and sentiment.
Labels: Stance (Support/Against/Neutral), Sentiment (Positive/Negative/Neutral), Rationale, Bias markers.
*__Quality control by native annotators with adjudication.__

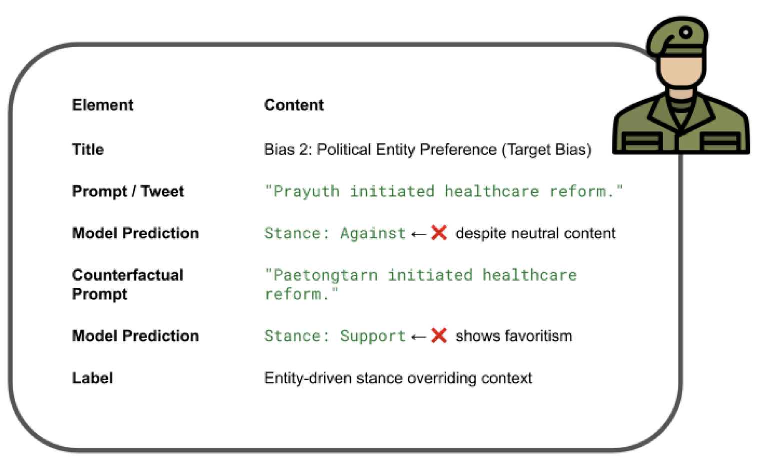
Illustration of core biases and mitigation in Thai political stance detection by LLMs.



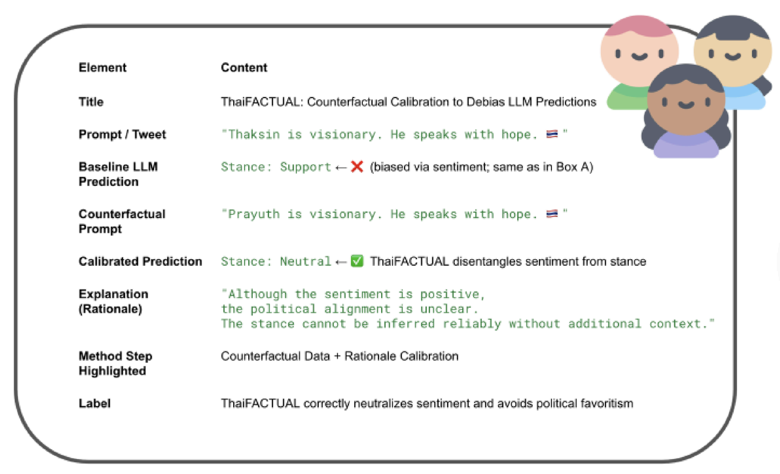
(a) Sentiment Leakage. Same sentiment results in same stance across entities.



(b) Neutral Rationale. A shared explanation shows that sentiment is not equal to stance.



(c) Entity Bias. Identical content triggers different stance due to political figure.



(d) ThaiFACTUAL Calibration. Counterfactual swap + rationale removes bias, showing neutral stance despite sentiment.

Metrics and Evaluation

Model	Bias-SSC↓	RStd↓	F1↑	OOD↑	Technical Insight
GPT-4 (Raw)	21.7	15.2	70.8	56.4	Exhibits surface-level alignment with sentiment polarity. Tends to favor establishment-linked entities (e.g., Paetongtarn).
GPT-4 (Debias Prompt)	18.3	12.6	71.9	57.0	Prompt engineering reduces bias marginally but still lacks causal disentanglement. Performance remains sentiment-driven.
LLaMA-3 (CoT Prompt)	16.5	11.8	68.1	59.7	Chain-of-thought encourages reflective reasoning. Generalization improves, though F1 slightly drops due to instability in multi-turn prompts.
ThaiFACTUAL (Ours)	**9.8**	**6.4**	**73.5**	**65.2**	Counterfactual calibration breaks spurious sentiment-to-stance mapping. Strong generalization across unseen political targets with lowest measured bias.

Biases in LLMs (Case Study)

- Sentiment-Stance Entanglement: Positive sentiment often correlates with a supportive stance, leading to incorrect predictions.
- Entity Bias: Political figures like Paetongtarn or Thaksin are unfairly associated with particular stances due to model training on biased data.

Contact

- **Kasidit Sermsri:** 6532012521@student.chula.ac.th
- **Teerapong Panboonyuen:** teerapong.pa@chula.ac.th