How Stable are LLM Bias Judgments? A Study of Implicit Bias with Target-Preserving Variations

Iffat Maab[1] Muhammad Ehtisham Hassan[2] Usman Haider[3]
[1]National Institute of Informatics, Tokyo [2]GIK Institute of Engg. Sciences and Technology, Topi
[3]University of Galway
maab@nii.ac.jp[1]

Abstract

Political bias depends not only on the content of a statement, but also on **who** the statement is about. Prior work in psychology shows that judgments of the same argument vary with the referenced social or political group, reflecting the influence of group-based associations. LLMs can perform well on explicit bias benchmarks while still exhibiting implicit associations that influence downstream decisions. In this work, we investigate whether changing the target entity of a political sentence such as replacing "Biden," with pronouns or vague references alters LLMs predictions of bias and political polarity. Using the BABE dataset, a benchmark of politically biased news statements annotated with bias spans and stance, we generate systematic variations of each sentence, i.e., pronoun-masked, vague-target, and flip-bias variants where the original target is replaced with an ideologically opposite one. These transformations enable us to probe implicit bias by examining whether bias judgments persist when explicit political identities are obscured or altered. We evaluate state-of-the-art instruction-tuned and reasoning models on these variants to examine how bias labels and ideological polarity shift, revealing whether model behavior is driven by surface lexical cues or implicit stereotypes.

1 Introduction

Political bias depends not only on wording, but also on the identity of the referenced actors. Research in social psychology [6, 5] and political communication [7, 3, 13] demonstrates that people interpret identical statements differently depending on the target group, political actor, or social identity involved. These attributions reflect underlying stereotypes and ideological associations that influence how information is perceived [2].

Implicit bias often manifests through subtle patterns of judgment rather than explicit or overt expressions. In human cognition, such latent biases are frequently revealed through comparative or context-dependent evaluations, which can precede observable discriminatory behavior [10]. Similar limitations also apply to LLMs, thus it became essential to examine whether such effects also arise in computational systems. Current findings [19, 21] found that implicit bias cannot be ignored, and it is therefore reasonable to expect that judgments can be shaped by influences operating outside conscious awareness.

LLMs exhibit similar patterns, despite strong performance on explicit bias benchmarks; they can still encode implicit biases that surface in downstream tasks such as classification, summarization, and stance detection [16, 20, 8]. Implicit biases are particularly concerning because they can influence model behavior even without explicit group references, reflecting human-like stereotype activation [12]. Motivated by this insight, we design a decision task that examines how LLM bias emerges when political targets are explicitly mentioned versus implicitly obscured in otherwise comparable statements, focusing on subtle, relational forms of bias rather than overt polarization. Prior work suggests that such comparative settings are effective for uncovering implicit bias [4, 11], as they surface differential judgments that may not be apparent in isolated evaluations. Building on this insight, we examine bias detection and ideological stance classification under minimal, target-preserving changes to assess whether LLMs behave consistently across minimally altered inputs.

We use the state-of-the-art BABE (Bias Annotation By Experts) [15] benchmark, which contains 3,700 sentences on political bias. A key challenge in evaluating political

Table 1: Sentence variants illustrating implicit bias manipulations using two target entities—[t1: Black Lives Matter] (social movement) and [t2: Christianity] (religion), from the BABE dataset.

Variants	Sentences
Original	''**[Black Lives Matter]** is essentially a non-falsifiable religious cult that sets itself specifically against **[Christianity]**, just like the French Revolution did.''
t1 pronoun	''[They] are essentially a non-falsifiable religious cult that sets itself specifically against [Christianity], just like the French Revolution did.''
t1 vague	''[The movement] is essentially a non-falsifiable religious cult that sets itself specifically against [Christianity], just like the French Revolution did.''
t1 flipped	''[Christianity] is essentially a non-falsifiable religious cult that sets itself specifically against [Black Lives Matter], just like the French Revolution did.''
t2 pronoun	''[Black Lives Matter] is essentially a non-falsifiable religious cult that sets itself specifically against [it], just like the French Revolution did.''
t2 vague	''[Black Lives Matter] is essentially a non-falsifiable religious cult that sets itself specifically against [a major religion], just like the French Revolution did.''
t2 flipped	''[Black Lives Matter] is essentially a non-falsifiable religious cult that sets itself specifically against [Islam], just like the French Revolution did.''

Table 2: Distribution statistics of the BABE dataset.

Bias Labels		**Polarity Labels**	
Non-biased	1,863	Left	989
Biased	1,810	Right	993
		Center	692
		None	1,000
Bias × Polarity		**Frequent Topics**	
Non-biased–Center	593	Marriage equality	347
Biased–Left	618	Vaccine	299
Biased–Right	597	Black Lives Matter	289

bias is the role of **target identification**. For each sentence, we identify two targets (the primary entity), and using GPT-4o we generate target-preserving variants for the first (T1) and second (T2) targets using three controlled transformations as shown in Table 1. Pronoun substitutions replace explicit political entities with pronouns, vague substitutions remove explicit lexical cues by replacing the target with an underspecified description, while flipped substitutions swap the target with an ideologically contrasting entity. Flipped variants have the possibility of altering the underlying ideological stance (e.g., left, center, right), enabling us to test whether models change their predictions solely due to target identity. Together, these manipulations isolate the extent to which model judgments depend on target identity rather than sentence content. Using the BABE dataset, our study asks:

- Does bias vanish when explicit political targets are removed?
- Do LLMs hallucinate who the target is, especially for vague references?
- How political polarity shifts when targets are flipped?

2 Related Work

Prior work [5, 13, 3] demonstrates that identical arguments are evaluated differently depending on the social or political group referenced, as stereotypes and group-based associations influence perception and interpretation. The behavior of the language model depends on carefully designed prompts and experiments that control for potential confounders [1, 9]. Detecting bias is considerably harder when it is implicit and lacks explicit abusive markers, as such cases do not rely on clearly identifiable lexical cues [22]. In this context, [18] worked on a broad range of techniques using machine learning and social sciences to evaluate and reduce discriminatory behavior in language models, while [17] designed multiple prompts to investigate how language models frequently reveal stereotypical associations in implicit bias evaluations.

3 Proposed Approach

Target-Preserving Bias Manipulations We utilize OpenAI GPT-4o to automatically construct target-preserving sentence variants for the BABE dataset, with the goal of probing implicit bias sensitivity under minimal textual perturbations. The prompt is shown in Figure 2 in Appendix. For each original sentence, we identify up to two target entities (e.g., political actors, social groups, institutions) and generate two manipulation categories, corresponding to the first target (T1) and the second target (T2). Variant examples are shown in Table 1, while Table 2 presents the BABE dataset distribution. For each sentence,

Table 3: Results of bias detection.

Model	Sentence Type	Acc	F1	Prec	Rec
Llama-3B	Original	76.10	76.00	78.60	70.72
	T1 Pronoun	74.43	74.25	77.96	67.05
	T1 Vague	73.86	73.61	78.21	65.06
	T1 Flipped	74.57	74.48	76.50	69.77
	T2 Pronoun	73.68	73.61	76.04	68.77
	T2 Vague	75.14	74.89	80.88	65.52
	T2 Flipped	73.87	73.79	76.32	68.83
Llama-8B	Original	70.67	70.14	62.78	90.49
	T1 Pronoun	70.40	69.81	63.29	89.17
	T1 Vague	70.66	70.19	63.49	88.50
	T1 Flipped	69.00	68.21	61.45	90.53
	T2 Pronoun	67.57	66.41	60.59	90.69
	T2 Vague	70.87	70.32	63.43	89.82
	T2 Flipped	67.33	66.09	60.14	91.47
Qwen-7B	Original	64.55	64.00	58.10	83.27
	T1 Pronoun	63.20	62.93	58.29	76.28
	T1 Vague	64.37	64.14	59.09	77.29
	T1 Flipped	62.68	62.00	57.08	81.28
	T2 Pronoun	63.07	62.36	57.99	80.85
	T2 Vague	63.91	63.40	58.64	80.10
	T2 Flipped	62.21	61.26	57.00	82.21
Qwen-14B	Original	56.57	47.67	53.17	99.28
	T1 Pronoun	56.10	47.01	52.93	98.89
	T1 Vague	57.22	49.05	53.57	98.73
	T1 Flipped	55.78	46.37	52.74	99.06
	T2 Pronoun	54.88	43.97	52.47	99.44
	T2 Vague	56.40	46.76	53.33	99.38
	T2 Flipped	54.97	44.17	52.52	99.38

Table 4: Results of polarity classification.

Model	Sentence Type	Acc	F1	Prec	Rec
Llama-3B	Original	45.57	43.80	46.92	43.79
	T1 Pronoun	42.86	41.78	44.37	41.43
	T1 Vague	43.59	42.63	44.69	42.28
	T1 Flipped	42.40	41.37	43.90	41.03
	T2 Pronoun	46.35	44.50	48.56	44.54
	T2 Vague	46.68	45.21	47.90	45.12
	T2 Flipped	44.68	42.69	45.64	42.88
Llama-8B	Original	53.98	53.05	53.54	53.13
	T1 Pronoun	48.85	48.35	48.52	48.30
	T1 Vague	50.15	49.59	49.71	49.55
	T1 Flipped	46.16	46.17	46.72	46.12
	T2 Pronoun	54.96	53.87	54.80	53.92
	T2 Vague	53.53	52.99	53.65	53.06
	T2 Flipped	51.69	51.20	51.70	51.20
Qwen-7B	Original	33.11	33.53	34.73	33.04
	T1 Pronoun	34.92	34.92	35.22	34.79
	T1 Vague	36.93	37.00	37.37	36.85
	T1 Flipped	37.15	36.93	37.18	36.81
	T2 Pronoun	33.31	33.71	34.15	33.43
	T2 Vague	32.93	33.09	34.12	32.71
	T2 Flipped	33.94	34.12	35.02	33.77
Qwen-14B	Original	47.92	47.61	48.52	47.23
	T1 Pronoun	44.70	44.50	44.92	44.25
	T1 Vague	46.18	46.00	46.16	46.00
	T1 Flipped	45.70	45.79	45.69	45.94
	T2 Pronoun	45.91	45.67	46.82	45.21
	T2 Vague	44.69	44.52	45.32	44.15
	T2 Flipped	45.83	45.63	46.18	45.34

we apply three controlled transformations: **Pronoun substitution** replaces the explicit target mention with a coreferential pronoun (e.g., "Democrats" → "they"), removing lexical specificity while preserving grammatical structure and discourse coherence. This tests whether models rely on explicit entity names to infer bias. **Vague substitution** replaces the target with a semantically underspecified description (e.g., "Democrats" → "some politician" or "a certain group"), reducing ideological salience while maintaining the sentence's overall meaning. This probes sensitivity to abstraction and ambiguity in target representation. **Flipped substitution** swaps the original target with a contrasting or opposing entity of the same semantic type (e.g., "Democrats" ↔ "Republicans"), intentionally reversing the ideological or social alignment while keeping the syntactic frame unchanged.

When only a single target is present in the original sentence, only T1 variants are generated; T2 variants are omitted to avoid redundant or duplicated samples. Across all manipulations, the surrounding context and factual content are otherwise preserved. We evaluate state-of-the-art instruction-tuned reasoning models across all experiments. See Appendix A for more details on dataset, experimental setup and implementation details, while the evaluation prompt is shown in Figure 1.

4 Results and Experiments

We report bias classification results for each model in all six settings against regular baseline (original) as shown in Table 4 and stance detection results in Table **??**. We observe clear differences in both task difficulty and model robustness to minimal, target-preserving edits. Bias detection is consistently easier than polarity prediction, where macro-F1 for bias ranges from 48–70 (Qwen-14B: 0.44–0.49; Llama-8B: 0.66–0.70), while polarity macro-F1 is substantially lower, ranging from 33– 53. Among models for the original settings using BABE, Llama-8B is the strongest overall, achieving the best bias performance with an F1 of 70.14 and the best polarity performance with an F1 of 53.05. For the same baseline settings, Qwen-7B is intermediate on bias with an F1 of 64.00 but performs poorly on polarity with an F1 of 33.53, while Qwen-14B shows the weakest bias results with an F1 of 47.67 but comparatively higher polarity results than Qwen-7B with an F1 of 47.61.

A key finding is variant sensitivity is non-trivial even when the underlying meaning is intended to remain comparable. For bias, performance fluctuates across T1/T2 conditions: for example, Qwen-14B reaches its best bias score on T1 Vague with an F1 of 49.05 but drops on T2 variants (e.g., T2 Pronoun F1 43.97), suggesting that altering the second target is particularly destabilizing. Llama-8B exhibits the most stable bias behavior with an accuracy of mostly 67–71% across variants, yet still shows degradation on flipped sentences (e.g., T1 Flipped F1 68.21) relative to its best settings (e.g., T2 Vague F1 70.32). For polarity, robustness is weaker across all models such as for Qwen-14B, the best polarity is the baseline (original) condition (Acc 47.92%) and accuracy typically decreases under pronoun/vague variants (e.g., T1 Pronoun Acc 44.70%), while Llama-8B shows its highest polarity accuracy on T2 Pronoun (Acc 54.96%), indicating that certain target edits can shift predicted ideology more than others.

Finally, the precision– recall profiles show systematic prediction tendencies. The recall of Qwen-14B bias is extremely high in all variants (99), but with substantially lower F1 of 44–49, which is consistent with a model that over-predicts "bias" and struggles to separate classes reliably. In contrast, Llama-8B and Qwen-7B exhibit a more balanced behavior in the original setting: Llama-8B achieves a precision of 62.78 and a recall of 90.49, while Qwen-7B attains a precision of 58.10 and a recall of 83.27. These results suggest that smaller instruction-tuned models are better calibrated for bias classification than larger reasoning-oriented models. For polarity, we additionally observe a notable number of unmapped or invalid outputs, particularly for Qwen-7B and Llama-8B. This indicates that three-way ideological classification is more challenging and less stable than binary bias detection. Overall, our results demonstrate that even small, controlled changes to target mentions—such as pronoun substitution, vagueness, or target flipping—can measurably alter bias and stance predictions. The extent of this sensitivity varies across models, and the magnitude of this effect is strongly model-dependent, with polarity judgments showing notably lower stability than bias classification.

Implicit Target Bias Index (ITBI) In our study, we propose an ITBI metric that quantifies the instability of the prediction of the model caused by target-preserving obfuscations. Details are provided in Appendix A. We perform the ITBI analysis using a reasoning-based Qwen-14B model as shown in Table 5. ITBI shows a clear asymmetry between bias and polarity judgments. Bias predictions are highly stable under target obfuscation (mean ITBI bias of 0.062 with 4.44% flip rate), indicating reliance on explicit evaluative cues. In contrast, polarity judgments are substantially more sensitive to implicit target changes (mean ITBI polarity of 0.374), with nearly a quarter of sentences changing left/center/right classification under pronoun or vague substitutions. In general, while 72.4% of sentences remain fully stable (ITBI=0), a notable minority (14.8%) exhibit multiple flips, demonstrating that implicit target identity primarily affects ideological stance rather than bias detection.

Table 5: ITBI scores summary using Qwen-14B.

Metric	**Bias**	**Polarity**
Pronoun flip (%)	3.03	19.54
Vague flip (%)	3.22	18.35
Any flip rate (%)	4.44	24.72
Mean ITBI score	0.062	0.374
Total ITBI mean (Bias+Polarity): 0.437		
ITBI=0 (stable): 72.4%	**ITBI≥2**: 14.8%	

5 Conclusion

We introduce implicit target bias as a source of instability in LLM-based political bias and stance prediction, showing that model judgments depend not only on what is said, but on how political targets are referenced. Through controlled target-preserving transformations on BABE, we demonstrate that pronoun substitution, vagueness, and target flipping systematically alter bias and polarity predictions, even when semantic content remains comparable. Our results show that bias detection is relatively stable, whereas ideological polarity is markedly fragile, revealing stronger implicit effects at higher levels of political reasoning. Using metrics such as the ITBI, we quantify how latent associations tied to target identity influence model outputs. In general, our results show that implicit bias in LLMs manifests itself as context-dependent sensitivity to target framing rather than explicit lexical cues, underscoring the need for perturbation-based evaluation protocols when assessing fairness and reliability in political NLP systems.

Acknowledgements

The authors wish to express gratitude to the funding organisation as this study was carried out using the TSUBAME 4.0 supercomputer at Institute of Science Tokyo.

References

[1]Zeyuan Allen-Zhu and Yuanzhi Li. ''Physics of language models: Part 3.1, knowledge storage and extraction''. In: **arXiv preprint arXiv:2309.14316** (2023).

[2]Yejin Bang et al. ''Measuring Political Bias in Large Language Models: What Is Said and How It Is Said''. In: **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 11142–11159. DOI: 10.18653/v1/2024.acl-long.600. URL: https://aclanthology.org/2024.acl-long.600/.

[3]Dennis Chong and James N Druckman. ''Framing theory''. In: **Annu. Rev. Polit. Sci.** 10.1 (2007), pp. 103–126.

[4]Faye Crosby, Stephanie Bromley, and Leonard Saxe. ''Recent unobtrusive studies of Black and White discrimination and prejudice: A literature review.'' In: **Psychological bulletin** 87.3 (1980), p. 546.

[5]Federica Durante and Susan T Fiske. ''How social-class stereotypes maintain inequality''. In: **Current opinion in psychology** 18 (2017), pp. 43–48.

[6]Naomi Ellemers and S Alexander Haslam. ''Social identity theory''. In: **Handbook of theories of social psychology** 2 (2012), pp. 379–398.

[7]Robert M Entman. ''Framing: Towards clarification of a fractured paradigm''. In: **McQuail's reader in mass communication theory** 390 (1993), p. 397.

[8]Fereshteh Hasanzadeh et al. ''Bias recognition and mitigation strategies in artificial intelligence healthcare applications''. In: **NPJ Digital Medicine** 8.1 (2025), p. 154.

[9]Jennifer Hu and Michael C Frank. ''Auxiliary task demands mask the capabilities of smaller language models''. In: **arXiv preprint arXiv:2404.02418** (2024).

[10]Jerry Kang. ''Little Things Matter a Lot: The Significance of Implicit Bias, Practically & Legally''. In: **Daedalus** 153.1 (2024), pp. 193–212.

[11]Benedek Kurdi et al. ''Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis.'' In: **American psychologist** 74.5 (2019), p. 569.

[12]Iffat Maab, Edison Marrese-Taylor, and Yutaka Matsuo. ''Target-Aware Contextual Political Bias Detection in News''. In: **Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)**. Nusa Dua, Bali: Association for Computational Linguistics, Nov. 2023, pp. 782–792. DOI: 10.18653/v1/2023.ijcnlp-main.50. URL: https://aclanthology.org/2023.ijcnlp-main.50/.

[13]Diego Saez-Trumper, Carlos Castillo, and Mounia Lalmas. ''Social media news communities: gatekeeping, coverage, and statement bias''. In: **Proceedings of the 22nd ACM international conference on Information & Knowledge Management**. 2013, pp. 1679–1684.

[14]Timo Spinde et al. ''MBIC–A Media Bias Annotation Dataset Including Annotator Characteristics''. In: **arXiv preprint arXiv:2105.11910** (2021).

[15]Timo Spinde et al. ''Neural Media Bias Detection Using Distant Supervision With BABE - Bias Annotations By Experts''. In: **Findings of the Association for Computational Linguistics: EMNLP 2021**. Association for Computational Linguistics, Nov. 2021, pp. 1166–1177. DOI: 10.18653/v1/2021.findings-emnlp.101. URL: https://aclanthology.org/2021.findings-emnlp.101/.

[16]Julius Steen and Katja Markert. ''Bias in news summarization: Measures, pitfalls and corpora''. In: **Findings of the Association for Computational Linguistics: ACL 2024**. 2024, pp. 5962–5983.

[17]Lihao Sun et al. ''Aligned but blind: Alignment increases implicit bias by reducing awareness of race''. In: **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. 2025, pp. 22167–22184.

[18]Alex Tamkin et al. ''Evaluating and mitigating discrimination in language model decisions''. In: **arXiv preprint arXiv:2312.03689** (2023).

[19]Bryan Chen Zhengyu Tan and Roy Ka-Wei Lee. ''Unmasking Implicit Bias: Evaluating Persona-Prompted LLM Responses in Power-Disparate Social Scenarios''. In: **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)**. Association for Computational Linguistics, Apr. 2025, pp. 1075–1108. ISBN: 979-8-89176-189-6. DOI: 10.18653/v1/2025.naacl-long.50. URL: https://aclanthology.org/2025.naacl-long.50/.

[20]Haining Wang et al. ''Fairness Evaluation of Large Language Models in Academic Library Reference Services''. In: **arXiv preprint arXiv:2507.04224** (2025).

[21]Yuchen Wen et al. ''Evaluating implicit bias in large language models by attacking from a psychometric perspective''. In: **Findings of the Association for Computational Linguistics: ACL 2025**. 2025, pp. 5081–5097.

[22]Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. ''Implicitly Abusive Language – What does it actually look like and why are we not getting there?'' In: **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**. June 2021. DOI: 10.18653/v1/2021.naacl-main.48. URL: https://aclanthology.org/2021.naacl-main.48/.

A Appendix

Dataset In this study, we use BABE which comprises 3,700 sentences drawn from U.S. news media. The dataset combines 1,700 instances originating from the MBIC corpus [14] with an additional 2,000 sentences collected by [15]. The sentences are extracted from news articles published by 14 major U.S. news outlets and span around a dozen socially and politically contested topics. Owing to its U.S.-centric media sources and explicit ideological annotations, we utilize BABE for analyzing political polarity and bias-related phenomena in news text.

Experimental settings We evaluate both reasoning-based and instruction-tuned LLMs including DeepSeek-R1 distilled Qwen (7B, 14B) and Llama (3B, 8B) models. All models are queried in a zero-shot setting using identical prompts for both the original and manipulated sentence variants without any fine-tuning. This setup enables a controlled comparison of model robustness to minimal, target-preserving textual edits. The metrics included accuracy (*Acc.*), precision (*Prec*), recall (*Rec*), and macro F1 (*F1*).

Bias and Polarity Evaluation Prompt

```
SYSTEM:
You are an expert political bias and ideology analyst.
Your task is to analyze the given sentence and:

1) Determine whether the sentence contains political
or ideological bias.
2) Identify the political polarity expressed in the
sentence.

Definitions:
- Bias label: "bias" or "no bias"
- Polarity label: "left", "center", or "right"
Input - Sentence: {sentence}

USER_INSTRUCTIONS:
Return ONLY valid JSON with following EXACT schema:
{"bias": {
    "label": "...", "justification": "..." },
  "polarity": {
    "label": "...", "justification": "..." }}

Justifications should be concise:
- Bias justification: 2--4 sentences
- Polarity justification: 1--3 sentences
```

Figure 1: Prompt used to evaluate political bias and ideological polarity for original and manipulated sentence variants in the BABE dataset.

Implicit Target Bias Index (ITBI) For each sentence, ITBI counts whether the bias label or ideological polarity changes when the target is replaced by a pronoun or a vague descriptor, without altering the meaning of the sentence. Bias-ITBI and Polarity-ITBI each range from 0 to 2, and their sum yields a total ITBI in [0,4]. For a given sentence, the score increases by 1 each time the prediction differs from the original under these conditions:

- **0**: No change — predictions remain stable under both pronoun and vague substitutions.
- **1**: One change — prediction flips under either pronoun or vague substitution.
- **2**: Two changes — prediction flips under both pronoun and vague substitutions.

GPT-4o Variant Prompt

```
SYSTEM:
You are an expert annotator for political bias and
target-framing research.

Your task: given a sentence, you must:
1) Detect up to TWO main political or social targets.
2) Create sentence variants (t1_pronoun, t1_vague,
t1_flipped, t2_pronoun, t2_vague, t2_flipped).
3) Preserve grammar and meaning except for the target
transformation.
4) Do not add new facts.
5) If no target exists, targets = [] and all
variants = original sentence.

Input - Sentence: {sentence}
USER_INSTRUCTIONS
Return ONLY valid JSON with this EXACT schema:
{{"original": "...",
   "targets": [
     {{
       "index": 1, "span": "...", "type": "..."
     }},
     {{
       "index": 2, "span": "...", "type": "..."
     }}],
   "variants": {{
     "t1_pronoun": "...", "t1_vague": "...",
     "t1_flipped": "...", "t2_pronoun": "...",
     "t2_vague": "...", "t2_flipped": "..."
   }} }}

 If there is one target
→ include only index=1, and fill
 all t2_* with original.
 If no targets → targets=[], and all variants =
 original sentence.
```

Figure 2: Prompt used with GPT-4 to make three different types of bias sentences using BABE dataset.