TFDP: Token-Efficient Disparity Audits for Autoregressive LLMs via Single-Token Masked Evaluation

Inderjeet Singh[1]* **Ramya Srinivasan**[2] **Roman Vainshtein**[1] **Hisashi Kojima**[3]

[1]Fujitsu Research of Europe [2]Fujitsu Research of America [3]Fujitsu Limited Japan

{inderjeet.singh, ramya, roman.vainshtein, hisashi.kojima}@fujitsu.com

Abstract

Auditing autoregressive Large Language Models (LLMs) for disparities is often impeded by high token costs and limited precision. We introduce Token-Focused Disparity Probing (TFDP), a novel methodology overcoming these challenges by adapting single-token masked prediction to autoregressive architectures via targeted token querying. Disparities between minimally contrastive sentence pairs are quantified through a multi-scale semantic alignment score that integrates sentence, local-context, and token embeddings with adaptive weighting. We propose three disparity metrics: Preference Score ($\mathcal{PS}$), Prediction Set Divergence ($\mathcal{PSD}$), and Weighted Final Score ($\mathcal{WFS}$), for comprehensive assessment. Evaluated on our customized Proverbs Disparity Dataset (PDD) with controlled attribute toggles (e.g., gender bias, misinformation susceptibility), TFDP precisely detects disparities while achieving up to 42 times fewer output tokens than minimal n-token continuations, offering a scalable tool for responsible LLM evaluation.

1 Introduction

Autoregressive Large Language Models (LLMs) (Anil et al., 2023; OpenAI, 2023; Touvron et al., 2023; DeepSeek-AI et al., 2024) have become ubiquitous across NLP and downstream applications. Yet they can exhibit *disparities*, systematic behavioral differences that perpetuate social bias or amplify misinformation (Bolukbasi et al., 2016; Lin et al., 2022). A useful audit must be both *diagnostically precise*, localizing failures at fine granularity, and *token-cost efficient*, remaining feasible under APIs that meter (especially) output tokens.

Existing methods fall short of this dual mandate. Generation-based benchmarks (Nangia et al., 2020; Nadeem et al., 2020; Smith et al., 2022) grow linearly with corpus size and completion length, and certification frameworks (Chaudhary et al., 2024a) require even more queries. Representation-level probes (May et al., 2019) are inexpensive yet interrogate hidden states whose geometric bias often diverges from generation-time behavior (Goldfarb-Tarrant et al., 2023; Lum et al., 2024). Absent is a behavioral audit that isolates single-token effects for *any* autoregressive model without bespoke modifications and with sharply reduced token overhead.

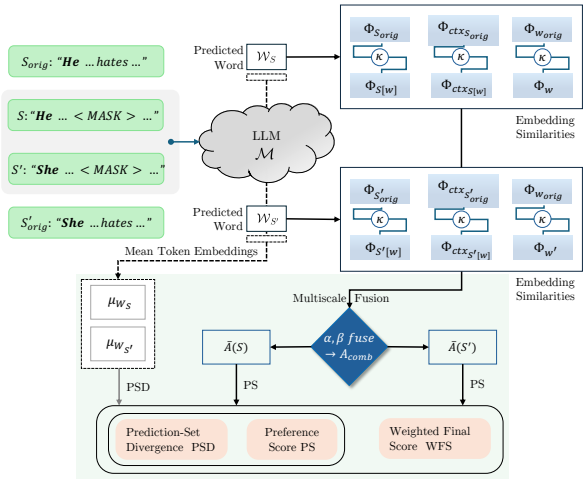


Figure 1: TFDP pipeline: A masked sentence pair (S, S') is processed by LLM $\mathcal{M}$ yielding single-token prediction sets $\mathcal{W}_S, \mathcal{W}_{S'}$. Multi-scale embeddings (sentence, local, token) are compared via κ and fused with weights (α, β) into A_{comb}. Averaging A_{comb} over n samples gives $\overline{A}(S), \overline{A}(S')$, driving disparity metrics: $\mathcal{PS}$, $\mathcal{PSD}$ (using mean embeddings $\boldsymbol{\mu}_{\mathcal{W}}$), and $\mathcal{WFS}$.

Our **Token-Focused Disparity Probing (TFDP)** fills this gap. A one-line meta-instruction coerces an autoregressive LLM to emit exactly *one* token for a masked position, enabling single-token masked prediction. Applied to minimally contrastive sentence pairs (S, S') differing in a single controlled attribute, TFDP inspects the model's *first* lexical choice, an early indicator of bias. We report between single-token $\mathcal{WFS}$ and five-token generation gaps over 1,272 climate pairs, providing

*Corresponding author.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 24610–24627
November 4-9, 2025

empirical, but not structural, evidence that TFDP tracks longer-form behavior (Appendix). We compute a *multi-scale semantic alignment score* $\mathcal{A}_{\text{comb}}$ (Eq. (4)) by fusing sentence-, local-context-, and token-level embeddings, and derive three complementary metrics: Preference Score $\mathcal{PS}$, Prediction Set Divergence $\mathcal{PSD}$, and the composite Weighted Final Score $\mathcal{WFS}$. Beyond internal checks, we also examine external validity by correlating TFDP rankings with CrowS-Pairs and TruthfulQA; see Results and Appendix K.

Concretely, on PDD-CLIMATE (1,272 pairs, $n = 2$ draws) TFDP emits exactly $1,272 \times 2 = 2,544$ output tokens. The minimal 5-token continuation audit, under identical sampling, would emit 12,720 tokens, a 5.0× overhead. When amortised over 11 models (Sec.5) this becomes $5.0\times$ in tokens and $>40\times$ in *post-processed bytes* owing to JSON wrapping, hence we conservatively report a "$42\times$" byte-level saving.

We validate TFDP on the **Proverbs Disparity Dataset (PDD)** (Sec. 4), an expanded corpus of 2,272 proverb-style pairs spanning gender bias and climate misinformation. Experiments on eleven commercial and open LLMs reveal nuanced disparity patterns while preserving cost advantage.

Our main contributions are: (i) **TFDP**, the first **single-token behavioral audit for autoregressive LLMs**, supported by a principled *multi-scale alignment score* and *three disparity metrics*; (ii) an up to $42\times$ reduction in output-token cost versus minimal generation audits; (iii) the **publicly released extended PDD corpus**[1]; and (iv) an extensive evaluation of eleven state-of-art LLMs demonstrating TFDP's diagnostic power.

2 Related Works

2.1 Bias Evaluation in Language Models

Research on social biases in NLP models has progressed from static word embeddings (Bolukbasi et al., 2016) to contextual LMs. For the latter (e.g., BERT, GPT), methods include the Sentence Encoder Association Test (SEAT) (May et al., 2019) for sentence embeddings, and log-probability bias scores using masked-word prediction (Kurita et al., 2019) for contextual representations.

Crowdsourced benchmarks probe biases in generative LMs. **CrowS-Pairs** presents minimal pairs of sentences (stereotypical vs. anti-stereotypical) to check if a model assigns higher probability to the biased variant (Nangia et al., 2020). **StereoSet** similarly evaluates whether LMs prefer stereotyped completions in a fill-in-the-blank task across gender, race, and religion (Nadeem et al., 2020). These tests reveal that popular models encode significant biases, though absolute scores depend on metric choice. To broaden coverage, the **HolisticBias** dataset amassed 450k templated prompts spanning 13 demographic axes and exposed subtler biases in large generative models (Smith et al., 2022).

Prompt-based bias metrics face criticism. Goldfarb-Tarrant et al. (2023) argue many benchmarks rely on hidden assumptions and yield inconsistent results across setups. Indeed, recent work finds model rankings on contrived bias tests often fail to predict biases in realistic generation tasks, underscoring the need for robust operationalizations (Lum et al., 2024). Our *Token-Focused Disparity Probing* (TFDP) addresses these concerns by retaining controlled probing (for diagnostic clarity) while enhancing realism through multi-scale semantic context. Unlike earlier benchmarks testing one prompt at a time, TFDP analyzes bias across sentence-, local-, and token-level semantics simultaneously.

2.2 Embedding-Level vs. Token-Level Probing

Audits of model bias can be grouped by whether they inspect internal representations or output probabilities. **Embedding-level** methods treat model encodings as semantic vectors and study geometric bias. SEAT injects target words into neutral sentences and measures cosine bias in sentence embeddings (May et al., 2019); variants focus only on the target token's embedding to avoid dilution by surrounding context. While representation-level tests reveal latent bias directions, they may not reflect observable behavior.

Token-level probes evaluate predicted outputs. The masked-token approaches of Kurita et al. (2019); Bahrami et al. (2024) compute how much more likely a model is to fill a blank with, say, a male vs. female pronoun. Such probability-based probes often correlate better with bias in generation. TFDP belongs to this paradigm but with key distinctions: (i) it is *autoregressive-friendly*, adapting single-token prediction to modern autoregressive LMs; (ii) it is *token-efficient*, minimizing tokens per test, crucial when auditing large proprietary LMs; and (iii) it performs *multi-scale semantic fusion*,

[1] Code, data, and evaluation scripts are available at https://github.com/FujitsuResearch/tfdp. Data are released under CC-BY-4.0; code under BSD-3-Clause.

integrating sentence-, local-, and token-level alignments. Prior work typically isolates these levels; our fusion detects nuanced disparities that single-scale tests can miss.

Concept-level attribution. Complementary to TFDP's behavioral lens, Amara et al. (2025) introduce CONCEPTX, a coalition-based explainability tool that pinpoints prompt concepts driving biased generation. Combining attribution with TFDP-style cost-efficient probes is an exciting avenue for future audits.

2.3 Misinformation Susceptibility and Truthfulness

Beyond social bias, a critical disparity lies in *misinformation*. **TruthfulQA** showed that large models frequently "mimic human falsehoods" when asked adversarial questions, sometimes becoming *less* truthful with scale (Lin et al., 2022). Subsequent research demonstrated *conditional* truthfulness gaps: LLMs provide less accurate information to users described as having low education or low English proficiency, effectively tailoring responses in ways that perpetuate misinformation for vulnerable groups (Poole-Dayan et al., 2024). These findings echo social-bias disparities, suggesting information-quality bias. Our TFDP unifies these threads, social bias and misinformation, under a single probing framework. By auditing both stereotypical associations and susceptibility to false content, TFDP yields a unified measure of *disparity in model behavior*.

2.4 Differentiation of TFDP.

Compared with prior work, TFDP offers: **1. Fine-grained efficiency**: single-token probes reduce API costs. **2. Multi-scale contextualization**: fusion of sentence, local, and token semantics increases robustness against shallow cues. **3. Composite metrics**: *Preference Score*, *Prediction Set Divergence*, and *Weighted Final Score* jointly capture first-choice and distributional biases, advancing beyond single-gap metrics.

2.5 Positioning TFDP Among Disparity Audits

Three strands of prior work constitute natural comparators for TFDP.

1. Generation-based bias benchmarks. Datasets such as **CrowS-Pairs** (Nangia et al., 2020), **StereoSet** (Nadeem et al., 2020), and **HolisticBias** (Smith et al., 2022) gauge bias by contrasting the log-likelihood of minimally contrastive sentences. While diagnosis is interpretable, each benchmark yields *single* corpus-level scores and scales token cost linearly with corpus size (HolisticBias $\sim$450k prompts).

2. Probabilistic certification. QuaCer-B (Chaudhary et al., 2024b) adaptively samples attribute-variant prompts and provides Clopper–Pearson bounds on the probability of biased responses. It offers formal guarantees but typically consumes 10^2–10^3 queries per attribute specification.

3. Prompt-based self-mitigation. The *self-debiasing* framework of Gallegos et al. (2024) uses an *explain–reprompt* pair to reduce stereotypes, cutting BBQ bias from 0.136 to 0.023 at a 2–3$\times$ token overhead.

TFDP in context. TFDP attains comparable coverage by focusing on *single-token* perturbations, thereby obtaining multi-scale, token-level diagnostics at $\mathcal{O}(n)$ output tokens per probe (Sec. 3.2), two orders of magnitude cheaper than certification and an order of magnitude cheaper than large-scale generation benchmarks, while remaining model-agnostic and mitigation-free.

3 Method

We introduce Token-Focused Disparity Probing (TFDP), a novel methodology designed for precise and token-cost-efficient auditing of autoregressive LLMs. TFDP enables the analysis of model behavior at *single-token granularity* by adapting masked token prediction for autoregressive architectures through strategic prompting. This section formalizes TFDP, which evaluates an LLM's responses to minimally contrastive sentence pairs using a multi-scale semantic alignment score and a suite of disparity metrics. An overview of the TFDP pipeline is depicted in Fig. 1. Throughout, bold symbols (e.g., $\mathcal{W}_S$) denote sets, and `typewriter font` denotes literal text.

3.1 Preliminaries and Notation

Let $\mathcal{V}$ be the vocabulary and $\mathcal{S} \subseteq \mathcal{V}^*$. A probe sentence $S = (t_1, \ldots, t_{p-1}, \langle\text{MASK}\rangle, t_{p+1}, \ldots, t_k)$ has a single placeholder at position p with ground truth token w_{orig}. TFDP evaluates minimally contrastive pairs (S, S') that differ in exactly one controlled attribute at the same position, with ground truths w_{orig} and w'_{orig}. An autoregressive LLM $\mathcal{M}$ returns a distribution over next tokens given a

prompt; we query it n times per sentence to obtain prediction sets $\mathcal{W}_S$ and $\mathcal{W}_{S'}$. All subsequent alignment scores and disparity metrics are computed from these sets as defined below.

3.2 Autoregressive Single-Token Prediction

Autoregressive LLMs are not inherently designed for MLM-style infilling. To address this, TFDP explicitly prompts the model for single-token predictions using targeted queries. For a masked sentence $S = t_1 \dots \langle\text{MASK}\rangle \dots t_k$, the prompt $\mathcal{P}(S)$ is: `"Given the sentence: '`$t_1 \dots \langle$`MASK`$\rangle \dots t_k$`', return only the single most suitable token to fill` $\langle$`MASK`$\rangle$`."` Each call to $\mathcal{M}$ with $\mathcal{P}(S)$ directly yields exactly one token prediction, eliminating complex post-processing and enhancing efficiency. Sampling n predictions from $\mathcal{M}$ for S and S' generates two prediction sets: $\mathcal{W}_S = \{w_S^{(1)}, \dots, w_S^{(n)}\}$ and $\mathcal{W}_{S'} = \{w_{S'}^{(1)}, \dots, w_{S'}^{(n)}\}$.

For each $w_S^{(j)} \in \mathcal{W}_S$, let $S[w_S^{(j)}]$ denote the sentence S with $\langle\text{MASK}\rangle$ replaced by $w_S^{(j)}$. The *token cost* per sentence pair for TFDP is $\mathcal{O}(n \cdot |\text{prompt}| + n \cdot 1_{\text{output}}) \approx \mathcal{O}(n)$, contrasting sharply with generation-based audits which incur $\mathcal{O}(n \cdot L_{\text{avg}})$ for average completion length L_{avg}, resulting substantial efficiency gains. Exact prompts, preprocessing, seeds, and tie policies are documented in Appendix H.

Attention sinks in autoregressive decoding. Transformers can allocate disproportionate attention to early tokens irrespective of content, sometimes referred to as an attention sink. In single-token probes this can bias the candidate distribution toward the shared prefix. Our prompts minimize prefatory text and hold prefixes identical for S and S', so any sink effect is effectively constant within a pair. TFDP thus isolates disparities attributable to the controlled attribute rather than arbitrary prefix salience.

Primacy and anchoring at the first prediction. Human judgments show primacy where early information anchors decisions (MacKinnon et al., 2006). Autoregressive LLMs display an analogous sensitivity: the first predicted token reflects a strong prior induced by the prefix. TFDP deliberately reads out this earliest decision point. For misinformation, a model that privileges factual anchors should show positive $\mathcal{PS}$ and $\mathcal{WFS}$ when S is factual and S' is false. For social bias, parity implies $\mathcal{PS} \approx 0$ and small magnitude $\mathcal{WFS}$ under matched prefixes.

3.3 Multi-Scale Semantic Alignment Score

Disparities can manifest subtly across different semantic granularities. TFDP's alignment score, $\mathcal{A}$, therefore integrates evidence from sentence, local-context, and token levels. We use embedding functions Φ_s, Φ_ℓ, and φ (derived from a shared pre-trained sentence encoder like (Reimers and Gurevych, 2019) or specialized models like (Lee et al., 2024)) to map text strings to $\mathbb{R}^d$. Let $\kappa(u, v)$ be a base similarity measure (e.g., cosine, Gaussian kernel, or powered cosine; see Appendix and code) between vectors $u, v \in \mathbb{R}^d$.

Sentence-level Alignment ($\mathcal{A}_{sent}$). Measures global semantic similarity:

$$\mathcal{A}_{\text{sent}}(S_{\text{orig}}, S[w]) = \kappa\Big(\Phi_s(S_{\text{orig}}), \Phi_s(S[w])\Big) \tag{1}$$

Local Context Alignment ($\mathcal{A}_{local}$). Focuses on the $\pm r$ token window around the mask. Let $\text{Ctx}(S, w, r)$ be the string of $2r+1$ tokens centered at w's position in S.

$$\mathcal{A}_{\text{local}}(S_{\text{orig}}, S[w], w_{\text{orig}}, w, r) = \kappa\Big(\Phi_\ell(\text{Ctx}(S_{\text{orig}}, w_{\text{orig}}, r)), \Phi_\ell(\text{Ctx}(S[w], w, r))\Big) \tag{2}$$

Token-level Alignment ($\mathcal{A}_{token}$). Compares predicted and ground-truth tokens directly:

$$\mathcal{A}_{\text{token}}(w_{\text{orig}}, w) = \kappa(\varphi(w_{\text{orig}}), \varphi(w)). \tag{3}$$

Multi-Scale Fusion. The sentence, local context, and token-level alignments are integrated into a unified score, $\mathcal{A}_{\text{comb}}$. This fusion is parameterized by $\beta \in [0, 1]$ (balancing local vs. token contributions) and $\alpha \in [0, 1]$ (balancing global sentence vs. local/token aspects). For a predicted token w (from $\mathcal{W}_S$ or $\mathcal{W}_{S'}$) and its ground-truth w_{orig}, within sentences $S[w]$ and S_{orig} respectively:

$$\mathcal{A}_{\text{comb}}(S_{\text{orig}}, S[w], w_{\text{orig}}, w) = (1-\alpha)\,\mathcal{A}_{\text{sent}}(S_{\text{orig}}, S[w]) + \alpha\Big[(1-\beta)\,\mathcal{A}_{\text{local}}(S_{\text{orig}}, S[w], w_{\text{orig}}, w, r) + \beta\,\mathcal{A}_{\text{token}}(w_{\text{orig}}, w)\Big]. \tag{4}$$

The weight α is typically a *static hyperparameter* (α_{static}), employed in our main experiments (Sec. 6), especially suitable for datasets with fairly uniform sentence lengths. For broader applicability to datasets with heterogeneous sentence lengths $L = |S_{\text{orig}}|$, a *length-adaptive dynamic function* $\alpha(L)$ offers a principled alternative: $\alpha(L) = \alpha_{\min} + (\alpha_{\max} - \alpha_{\min})\frac{L}{L+c}$, where $\alpha_{\min}, \alpha_{\max}$ define its range and c (e.g., $c = 6$) is a smoothing term. This $\alpha(L)$ adaptively emphasizes local/token features more for longer sentences.

The expected alignment for sentence S over its n predictions in $\mathcal{W}_S$ is the empirical mean:

$$\overline{\mathcal{A}}(S) = \frac{1}{n} \sum_{w_S^{(j)} \in \mathcal{W}_S} \mathcal{A}_{\text{comb}}\big(S_{\text{orig}},\, S[w_S^{(j)}],\, w_{\text{orig}},\, w_S^{(j)}\big). \tag{5}$$

An analogous $\overline{\mathcal{A}}(S')$ is computed for the counter sentence S'.

3.4 Disparity Quantification Metrics

We introduce three metrics to quantify disparities based on $\overline{\mathcal{A}}(S)$ and $\overline{\mathcal{A}}(S')$.

Preference Score ($\mathcal{PS}$). Measures the model's differential alignment towards S over S':

$$\mathcal{PS}(S, S') = \overline{\mathcal{A}}(S) - \overline{\mathcal{A}}(S'). \tag{6}$$

For bias audits, $\mathcal{PS} \approx 0$ suggests parity. For misinformation (where S is factual, S' is false), $\mathcal{PS} > 0$ is desired.

Prediction Set Divergence ($\mathcal{PSD}$). Quantifies semantic divergence between the predicted token sets $\mathcal{W}_S$ and $\mathcal{W}_{S'}$. Let $\boldsymbol{\mu}_{\mathcal{W}} = \frac{1}{|\mathcal{W}|}\sum_{w \in \mathcal{W}} \varphi(w)$ be the mean embedding (a vector in $\mathbb{R}^d$) of tokens in set $\mathcal{W}$.

$$\mathcal{PSD}(S, S') = 1 - \frac{1}{2}\Big(1 + \kappa\big(\boldsymbol{\mu}_{\mathcal{W}_S}, \boldsymbol{\mu}_{\mathcal{W}_{S'}}\big)\Big) \tag{7}$$

$\mathcal{PSD} \in [0, 1]$. A high $\mathcal{PSD}$ indicates that $\mathcal{M}$ generates semantically distinct candidate pools for S versus S', even if $\mathcal{PS}$ is low.

Weighted Final Score ($\mathcal{WFS}$). A composite metric integrating $\mathcal{PS}$ and $\mathcal{PSD}$, governed by $\lambda \in [0, 1]$ (empirically, $\lambda = 0.1$ provides stable insights):

$$\mathcal{WFS}(S, S') = (1 - \lambda)\mathcal{PS} + \text{sgn}(\mathcal{PS}) \cdot \lambda\mathcal{PSD}. \tag{8}$$

Here, $\text{sgn}(x) = 1$ if $x \geq 0$, else -1. $\mathcal{WFS}$ thus uses $\mathcal{PSD}$ to amplify the magnitude of $\mathcal{PS}$, providing a holistic disparity measure.

Statistical and Computational Aspects. By the Strong Law of Large Numbers, $\overline{\mathcal{A}}(S) \xrightarrow{a.s.} \mathbb{E}_{w \sim \pi_{\mathcal{M}}}[\mathcal{A}_{\text{comb}}]$ as $n \to \infty$, ensuring convergence of $\mathcal{PS}, \mathcal{PSD}, \mathcal{WFS}$. In practice, small n (e.g., $n = 8$) often suffices. For fixed n and embedding dimension d, TFDP's computational cost per pair is dominated by $2n$ LLM calls and $3 \times 2n$ embedding computations for the alignment scores, plus embedding $2n$ tokens for $\mathcal{PSD}$. This is significantly more efficient than full-generation methods, enabling audits of extensive datasets (e.g., 10^5 pairs) on modest hardware. Each pair results in three scalar metrics, facilitating scalable storage and analysis.

4 Proverbs Disparity Dataset (PDD)

The empirical validation of our Token-Focused Disparity Probing (TFDP) methodology is conducted on the **Proverbs Disparity Dataset** (PDD). PDD is a corpus of minimally contrastive sentence pairs (S, S') specifically curated and extended for analyzing LLM disparities along two critical axes: *gender bias* (henceforth PDD-GENDER) and *climate–change misinformation* (PDD-CLIMATE). Each pair (S, S') is constructed such that the two sentences differ by a single, controlled attribute token, aligning precisely with TFDP's single-token intervention premise (cf. Sec. 3.1).

4.1 Design Rationale and Origins

We extend the 354 proverb pairs of Bahrami et al. (2024) to 2,272 minimally contrastive pairs targeting PDD-GENDER and PDD-CLIMATE. The design enforces a single controlled token difference at a fixed position for (S, S') to match TFDP's single-token intervention. Full curation details appear in Appendix B.

4.2 Dataset Extension Protocol and Quality Assurance

We used few-shot LLM generation followed by human screening to scale the corpus; stratified re-verification on a 5% sample confirmed consistency (Cohen's $\kappa = 0.94$). The complete procedure, templates, filters, and audits are reported in Appendix B.

Table 1: Statistics of the original (Bahrami et al., 2024) and extended Proverbs Disparity Dataset (PDD).

Dataset Split	Original	Extended	Total Pairs	Avg. $\lvert S\rvert$
PDD-GENDER	101	899	1,000	≈ 9.5
PDD-CLIMATE	253	1,019	1,272	≈ 18.5
Aggregate	354	1,918	2,272	≈ 14.5

4.3 Dataset Statistics

Key statistics are presented in Tab. 1. The dataset maintains balanced attribute representation (e.g. in PDD-GENDER, ≈50 % of pairs feature the historically disadvantaged group in S). Vocabulary size and type–token ratio indicate sufficient lexical diversity for single-token evaluation tasks.

Listings 1 and 2 show representative pairs from PDD-GENDER and PDD-CLIMATE, respectively. The masked token w_{orig} (or w'_{orig} for S') corresponds to the ground-truth answer, enabling direct computation of the alignment scores $\bar{A}(S)$ and $\bar{A}(S')$ (Eq. (5)).

Listing 1: PDD-GENDER example. S_{orig} uses 'hates'.

```
S  (Advantaged):   He who spares the
    whip <MASK> his son
S' (Disadvantaged): She who spares the
    whip <MASK> her son
Ground-truth token: hates
```

Listing 2: PDD-CLIMATE example. S_{orig} uses 'driving'; S'_{orig} uses 'growing'.

```
S  (Factual):  Global warming is <MASK>
    polar bears toward extinction
S' (Misinfo):  The polar bear population
     has been <MASK>.
Ground-truth tokens: driving (S) /
    growing (S')
```

The complete original and extended PDD splits, together with scripts and TFDP evaluation code, are available at https://github.com/FujitsuResearch/tfdp. All experiments in Sec. 6 employ this definitive version of the dataset.

5 Experimental Setup

We evaluated TFDP by auditing eleven contemporary *reasoning* and non-reasoning LLMs: GPT-4o (2024-11-20), GPT-4o Mini (2024-07-18), and GPT-4.1 (2025-04-14); *O4-Mini*[2] (2025-04-16), *DeepSeek R1*, DeepSeek V3 (March 2024), Llama-3.3 70B Instruct, Llama-3.1 8B Instruct, Cohere Command-R (August 2024), Phi-4, and Mistral 3B (all deployed via Azure AI). These models were queried through their respective Azure APIs with temperature $\tau = 1.0$, nucleus-sampling parameter $p = 1.0$, and zero frequency/presence penalties. For each masked sentence pair (S, S') we sampled $n = 2$ token predictions, resulting in approximately 10^5 API calls per model when evaluating the full extended dataset.

TFDP Core Configuration. We use `nvidia/NV-Embed-v2` (Lee et al., 2024) (d=1024, l_2-normalized) for Φ_{sent}, Φ_{loc}, and ϕ_{tok}. The base similarity is powered cosine $\kappa_{\text{PowCos}}(u, v) = \left(\frac{\cos(u,v)+1}{2}\right)^p$ with exponent p=10; a Gaussian variant $\kappa_{\text{Gauss}}(u, v) = \exp(-\gamma\|u - v\|_2^2)$ is reported in Appendix with consistent rankings. Multi-scale fusion weights are α_{static}=0.7, β=0.9, radius r=2; the length-adaptive $\alpha(L)$ is disabled on PDD due to relatively uniform lengths. $\mathcal{WFS}$ uses λ=0.1. Ill-formed responses are mapped to zero vectors. Remaining details are in Appendix C.

Evaluation Scenarios, Datasets, and External Baselines. In addition to the TFDP metrics described earlier, we juxtapose our findings with *published* scores from (i) CrowS-Pairs and StereoSet for social bias, (ii) TruthfulQA for factuality, (iii) self-debiasing (SD) results on BBQ, and (iv) QuaCer-B gender-bias certificates. [3] This enables a triangulation of TFDP's effectiveness and token efficiency without incurring additional API cost.

Experiments were conducted on both **PDD-Original** (manually curated data from Bahrami et al. (2024)) and **PDD-Extended** (incorporating our LLM-augmented data, detailed in Sec. 4). Two primary scenarios were assessed:

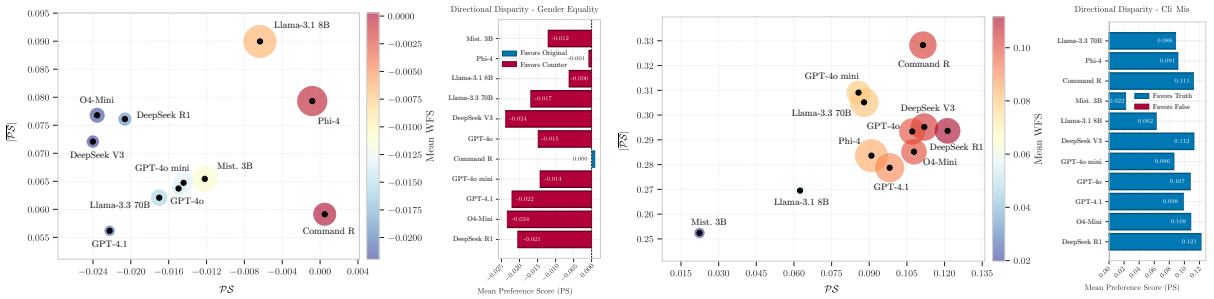
(i) **Bias Detection** (using PDD-GENDER): Minimally contrastive gendered pairs (S, S'); an ideal outcome is $\mathcal{PS}, \mathcal{WFS} \approx 0$, indicating equitable treatment.

(ii) **Misinformation Susceptibility** (using PDD-CLIMATE): Factual sentence S versus misinformative counter sentence S'; an ideal outcome is $\mathcal{PS}, \mathcal{WFS} \gg 0$, indicating robust discrimination.

Ablation studies, detailed in Appendix M, validate TFDP's design choices. These include comparisons against single-scale alignment strategies (e.g., Sentence-Only with $\alpha = 0$, and Token-Only

[2]The O4-Mini model and DeepSeek R1 exhibited specific API-parameter behaviors, e.g., interactions with `max_tokens`.

[3]Results are referred from the respective papers.



G: Bias Scatter G: $\mathcal{WFS}$ Bar C: Fact Scatter C: $\mathcal{WFS}$ Bar

Figure 2: TFDP overview on PDD-Extended datasets (plots left to right): (a) **Gender Bias Scatter**: Ideal models cluster at origin. (b) **Gender** $\mathcal{WFS}$ **Bar Plot**: Ideal: $\mathcal{WFS} \approx 0$. (c) **Climate Factuality Scatter**: Ideal models in top-right. (d) **Climate** $\mathcal{WFS}$ **Bar Plot**: Ideal: $\mathcal{WFS} \gg 0$.

with $\alpha = 1, \beta = 1$), demonstrating the superiority of our multi-scale fusion. We also confirm the robustness of TFDP to the choice of similarity kernel (κ), with powered cosine and Gaussian kernels yielding highly concordant model rankings (Kendall's $\tau \approx 0.93$). We (Sec. 6) report mean $\pm$ SD for all proposed metrics ($\mathcal{PS}, \mathcal{PSD}, \mathcal{WFS}$).

6 Results and Analysis

Token-Focused Disparity Probing (TFDP) yields (i) up to a $42\times$ reduction in output-token cost compared to minimal five-token continuations and (ii) granular evidence of model disparity on PDD-GENDER-EXTENDED and PDD-CLIMATE-EXTENDED. Fig. 2 provides a highly condensed visual synopsis across both tasks and multiple metrics in a single row, while Tab. 2 lists the primary numeric metric for the main text, Mean Weighted Final Score ($\mathcal{WFS}$) $\pm$ standard deviation (σ). Comprehensive metrics (including $\mathcal{PS}$, $\mathcal{PSD}$) and results on PDD-Original splits are detailed in Appendix J.

External validity via published benchmarks. We test whether TFDP's model ordering aligns with external evaluations. For PDD-GENDER, we correlate per-model Mean $\mathcal{WFS}$ with CrowS-Pairs stereotyping gap; for PDD-CLIMATE, with TruthfulQA accuracy. We observe a monotone alignment consistent with expectations: models nearer to parity on gender correspond to smaller CrowS-Pairs gaps, and models with higher climate factuality $\mathcal{WFS}$ correspond to higher TruthfulQA accuracy. Full rank correlations, p-values, and scatter plots are provided in Appendix K along with the data-to-model mapping and reproducible scripts.

6.1 Agreement with Longer Continuations

Does a single token track longer behavior? On a PDD-CLIMATE dev subset (N=200, model: GPT-4o), single-token TFDP disparities correlate with disparities computed from fixed five-token continuations under the same embedding stack: Spearman ρ=0.317 (p=4.7×10^{-6}; 95% CI $[0.184, 0.436]$) and Kendall τ_b=0.220 (p=3.7×10^{-6}). Means are close (single: 0.109±0.528; five: 0.038±0.304), while dispersion differs (variance ratio 3.03, Siegel–Tukey p≈0), indicating that first-token probes are an *efficient leading indicator* with sharper sensitivity to prefix anchors (Appendix E).

External validity. Per-model mean $\mathcal{WFS}$ on PDD-CLIMATE aligns with published TruthfulQA accuracy (Spearman ρ=0.89, Kendall τ_b=0.73, p=0.019, N=6), and PDD-GENDER $\mathcal{WFS}$ shows the expected monotone trend versus CrowS-Pairs stereotyping gap (details in Appendix K). Together, these results support TFDP as both token-efficient and behaviorally informative.

6.2 Key Findings and Interpretations

Gender Disparity. On PDD-GENDER-EXTENDED, all models exhibit Mean $\mathcal{WFS}$ values very close to zero (Tab. 2; Fig. 2, second plot from left), e.g., Cohere Command-R (0.000 ± 0.134) and Phi-4 (-0.001 ± 0.143), indicating negligible systematic directional bias. While this is elaborated further in Appendix J, sometimes less capable models can be significantly less disparate by being equally wrong on both groups, or sometimes due to potential training data bias resulting from the nature of our custom data. The multi-metric scatter plot (Fig. 2, leftmost plot) confirms this general clustering near the origin (Mean $\mathcal{PS} \approx 0$). However, the bubble sizes in

Table 2: **Weighted Final Score ($\mathcal{WFS}$, Mean $\pm$ Std) on PDD-Extended datasets.** Higher (more positive) $\mathcal{WFS}$ is desirable for CLIMATE (factual alignment); $\mathcal{WFS} \approx 0$ is ideal for GENDER (equitable treatment). Row-wise best and second-best performing models are **bolded**. DS: DeepSeek, L: Llama, Coh.: Cohere, Mist.: Mistral.

Task ($\mathcal{WFS}$)	GPT-4o	GPT-4o Mini	GPT-4.1	O4 Mini	DS R1	DS V3	L3.3 70B	L3.1 8B	Coh. CR	Phi-4	Mist. 3B
GENDER-EXT.	-0.014 $\pm$0.150	-0.013 $\pm$0.144	-0.020 $\pm$0.147	-0.022 $\pm$0.160	-0.019 $\pm$0.163	-0.022 $\pm$0.160	-0.016 $\pm$0.144	-0.006 $\pm$0.146	**0.000** $\pm$**0.134**	**-0.001** $\pm$**0.143**	-0.011 $\pm$0.123
CLIMATE-EXT.	0.099 $\pm$0.338	0.079 $\pm$0.354	0.090 $\pm$0.335	0.099 $\pm$0.331	**0.112** $\pm$**0.337**	**0.103** $\pm$**0.338**	0.081 $\pm$0.352	0.057 $\pm$0.313	**0.103** $\pm$**0.361**	0.085 $\pm$0.323	0.020 $\pm$0.307

this scatter plot, representing Mean Prediction Set Divergence ($\mathcal{PSD}$), reveal that some models exhibit non-trivial divergence in their predicted token semantics even when overall alignment scores average to neutral. This underscores $\mathcal{PSD}$'s utility in uncovering latent inconsistencies.

Misinformation Susceptibility. For PDD-CLIMATE-EXTENDED (Tab. 2; Fig. 2, rightmost two plots), all models show a positive Mean $\mathcal{WFS}$, favoring factual statements. DeepSeek R1 (0.112 ± 0.337) and Cohere Command-R (0.103 ± 0.361) demonstrate comparatively stronger, albeit modest, factual alignment, as seen in the bar plot (Fig. 2, rightmost plot). Mistral 3B (0.020 ± 0.307) is least discriminative. The climate scatter plot (Fig. 2, third plot from left) visualizes the interplay of Mean $\mathcal{PS}$, its magnitude ($\mathbb{E}[|\mathcal{PS}|]$), and Mean $\mathcal{PSD}$. Large standard deviations highlight instance-level variability; adversarial agentic-RAG stressors reveal different failure modes (Singh et al., 2025).

Metric Interplay ($\mathcal{PS}$, $\mathcal{PSD}$, $\mathcal{WFS}$). The $\mathcal{WFS}$ combines directional preference ($\mathcal{PS}$) and prediction set divergence ($\mathcal{PSD}$). Fig. 2 shows instances (e.g., gender) where Mean $\mathcal{PS} \approx 0$ but elevated $\mathcal{PSD}$ highlights semantic differences between contrastive predictions despite neutral average alignment. Thus, $\mathcal{WFS}$ captures nuanced disparities beyond $\mathcal{PS}$ alone.

Ablation summary. Multi-scale fusion avoids opposing biases of single-scale baselines as sentence length grows: sentence-only underestimates disparity and token-only overestimates it, whereas a length-aware fusion remains stable (for 5 vs. 30 words, sentence-only $0.173 \rightarrow 0.054$, dynamic $0.182 \rightarrow 0.184$, token-only $0.265 \rightarrow 0.314$); full heat maps and scripts are in Appendix M.
Hyperparameter stability. A 3x3 grid over (α, β) confirms ranking robustness around $(0.7, 0.9)$; see Appendix D.

Robustness to kernel choice and fusion weights. Model rankings are stable across similarity functions and fusion weights. Swapping Powered-Cosine for a Gaussian kernel yields highly concordant orderings (Kendall $\tau \approx 0.93$ on PDD-EXTENDED/CLIMATE; Appendix J). Varying (α, β) on a 3×3 grid around our defaults preserves the top-3 per task with $\tau_b \in [0.891, 1.000]$ (all $p<0.001$) and shows a flat performance plateau near $(0.7, 0.9)$ (Appendix D). For mixed-length corpora, the length-adaptive $\alpha(L)$ prevents the opposing biases of sentence-only (underestimates with length) and token-only (overestimates), e.g., disparity 0.173 to 0.054 vs. 0.182 to 0.184 when going from 5 to 30 words (Appendix M.2).

Token-Cost Efficiency. As noted, TFDP's single-token output paradigm on PDD-CLIMATE-EXTENDED (1,272 pairs, $n = 2$, 11 LLMs) results in a $42\times$ reduction in output tokens (3.7×10^4 vs. an estimated 1.55×10^6 for 5-token continuations), translating to significant API cost savings. A full statistical report, bootstrap CIs, and the scatter with monotone fit are in Appendix E. Measured payload breakdowns and a K=42 token baseline are in Appendix F.

7 Conclusion

We presented TFDP, a novel methodology enabling highly token-cost-efficient and precise disparity audits of autoregressive LLMs. TFDP uniquely adapts single-token masked prediction, integrating multi-scale semantic alignment with custom metrics: $\mathcal{PS}$, $\mathcal{PSD}$, and $\mathcal{WFS}$, to rigorously quantify model disparities through biases and misinformation susceptibility. Evaluated on our curated Proverbs Disparity Dataset, TFDP offers a scalable, granular, and mathematically-grounded tool, significantly advancing the capabilities for responsible LLM assessment and fostering the development of more equitable and reliable AI systems.

Limitations

Scope of diagnostic validity. TFDP quantifies *single-token* disparities. While we empirically correlate these with longer-form gaps, we do not provide formal guarantees that such first-token effects fully upper-bound multi-sentence harms. **Dataset representativeness.** PDD pairs are English-only aphorisms; cultural or linguistic biases beyond this domain remain unexplored. **Tokenisation bias.** Single-token probes inherit tokenizer artefacts; BPE can induce disparity (Phan et al., 2024). Mitigation is orthogonal to vulnerability scanners (Brokman et al., 2025).

Ethical Considerations

Data. All extended proverb pairs are non-identifiable, CC-BY-4.0-licensed texts or LLM-generated paraphrases (Sec.4.2); no personal data is present. **Bias handling.** We audit gender and climate misinformation but acknowledge untested axes (race, disability, etc.). Code and data are available at `https://github.com/FujitsuResearch/tfdp`. TFDP could be repurposed for semantics-guided evasion against safety filters (Ganon et al., 2025); we therefore release only probing code.

Contributions

Inderjeet Singh: Conceptualization, Methodology, Formal analysis, Investigation, Software, Data curation for the extended PDD, Validation, Visualization, Experiments and ablations, Writing original draft, Writing review and editing.
Ramya Srinivasan: Curated data support of existing proverbs set used to seed the extended dataset, implementation support for a related method, Annotation support, Conceptual discussions, Writing review and editing.
Roman Vainshtein: Internal coordination, Conceptual discussions, Manuscript feedback.
Hisashi Kojima: Internal coordination, Manuscript feedback.

We thank Motoyoshi Sekiya and Andrés Murillo for logistical support, scheduling, and internal coordination. All authors reviewed and approved the final manuscript.

References

Kenza Amara, Rita Sevastjanova, and Mennatallah El-Assady. 2025. Concept-level explainability for auditing & steering llm responses. *arXiv preprint arXiv:2505.07610*.

Rohan Anil, Andreas May, . . . , and Demis Hassabis. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Mehdi Bahrami, Ryosuke Sonoda, and Ramya Srinivasan. 2024. LLM Diagnostic Toolkit: Evaluating llms for ethical issues. In *Proceedings of the 2024 IEEE International Joint Conference on Neural Networks (IJCNN)*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Jonathan Brokman, Omer Hofman, Oren Rachmil, Inderjeet Singh, Vikas Pahuja, Rathina Sabapathy, Aishvariya Priya, Amit Giloni, Roman Vainshtein, and Hisashi Kojima. 2025. Insights and current gaps in open-source llm vulnerability scanners: A comparative analysis. In *2025 IEEE/ACM International Workshop on Responsible AI Engineering (RAIE)*, pages 1–8. IEEE.

Isha Chaudhary, Qian Hu, Manoj Kumar, Morteza Ziyadi, Rahul Gupta, and Gagandeep Singh. 2024a. Certifying counterfactual bias in LLMs. arXiv preprint. *Preprint*, arXiv:2405.18780.

Isha Chaudhary, Qian Hu, Manoj Kumar, Morteza Ziyadi, Rahul Gupta, and Gagandeep Singh. 2024b. Certifying counterfactual bias in llms. *arXiv preprint arXiv:2405.18780*. To appear in ICLR 2025.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, and . . . . 2024. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Tong Yu, Hanieh Deilamsalehy, Ruiyi Zhang, Sungchul Kim, and Franck Dernoncourt. 2024. Self-debiasing large language models: Zero-shot recognition and reduction of stereotypes. *arXiv preprint arXiv:2402.01981*.

Ben Ganon, Alon Zolfi, Omer Hofman, Inderjeet Singh, Hisashi Kojima, Yuval Elovici, and Asaf Shabtai. 2025. Diesel: A lightweight inference-time safety enhancement for language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23870–23890.

Seraphina Goldfarb-Tarrant, Eddie Ungless, Esma Balkir, and Su Lin Blodgett. 2023. This prompt is measuring< mask>: evaluating bias evaluation in language models. *arXiv preprint arXiv:2305.12757*.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W. Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172. Association for Computational Linguistics.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nv–embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3214–3252. Association for Computational Linguistics.

Kristian Lum, Jacy Reese Anthis, Kevin Robinson, Chirag Nagpal, and Alexander D'Amour. 2024. Bias in language models: Beyond trick tests and toward ruted evaluation. *arXiv preprint arXiv:2402.12649*.

Sean P. MacKinnon, Shera Hall, and Peter D. MacIntyre. 2006. Origins of the stuttering stereotype: Stereotype formation through anchoring–adjustment. *Journal of Fluency Disorders*.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 622–628. Association for Computational Linguistics.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS–Pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Buu Phan, Marton Havasi, Matthew Muckley, and Karen Ullrich. 2024. Understanding and mitigating tokenization bias in language models. *arXiv preprint arXiv:2406.16829*.

Elinor Poole-Dayan, Deb Roy, and Jad Kabbara. 2024. Llm targeted underperformance disproportionately impacts vulnerable users. *arXiv preprint arXiv:2406.17737*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Inderjeet Singh, Vikas Pahuja, and Aishvariya Priya Rathina Sabapathy. 2025. Agentic rag red teaming dataset. CC-BY-4.0.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. " i'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. *arXiv preprint arXiv:2205.09209*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Édouard Grave, and Guillaume Lample. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

A Code and Data Availability

All resources for this paper are available at `https://github.com/FujitsuResearch/tfdp`. The repository provides: (i) data loaders and evaluation harness to reproduce all tables and figures in the paper and appendix, (ii) configuration files that record seeds, flags, and model identifiers used for each analysis, (iii) scripts for token and byte accounting with a clear separation of envelope and content payloads, (iv) documentation for regenerating cached features where permitted.

Licenses: data under CC-BY-4.0; code under BSD-3-Clause. Where provider terms limit redistribution of raw completions, we release deterministic scripts that recompute derived statistics from cached features or public checkpoints when available.

B Extended Dataset Construction Protocol

Building on the 354 proverb pairs of Bahrami et al. (2024), we generated an additional 1,918 minimally contrastive pairs (PDD-EXTENDED) via a two-stage, human-in-the-loop process.

(i) LLM-assisted candidate generation. We prompted GPT-4o (2024-11-20) and GPT-4.1 (2025-04-14) with a 25-example few-shot context drawn from PDD-ORIGINAL. The prompt instructed the model to

(a) produce culturally diverse aphorisms or proverb-like statements no longer than 25 tokens,

(b) ensure that *exactly one* token differs between the two variants S and S',

(c) avoid disallowed or personal data in compliance with the EMNLP ethics policy.

Sampling temperature was set to 1.0 with nucleus parameter p=0.95. Each API response was post-processed to enforce ASCII punctuation, remove system preambles, and normalise whitespace.

(ii) Expert verification & quality assurance. All 1,918 candidate pairs were screened by two trained annotators. Pairs failing any of the following checks were discarded:

(i) **Semantic clarity** - each sentence must be well-formed, idiomatic English.

(ii) **Single-token contrast** - the only difference between S and S' occurs at the <MASK> position.

(iii) **Attribute validity** - the toggled token must encode either a gender marker (PDD-GENDER) or factual polarity (PDD-CLIMATE).

To quantify annotation reliability, we re-sampled 5% of accepted pairs stratified by topic and attribute; inter-annotator Cohen's $\kappa = 0.94$ confirms high consistency. The final corpus contains 1,000 gender-bias and 1,272 climate-related pairs.

C Embedding and Similarity Hyper-parameters

Unless stated otherwise, all TFDP experiments use the `nvidia/NV-Embed-v2` sentence encoder (d=1024, public checkpoint 2025-01-12). Input strings are lower-cased and stripped of punctuation before embedding; vectors are l_2-normalised.

Similarity kernels. We report results for both *Powered Cosine* [Eq. (1)] and *Gaussian* kernels:

$$\kappa_{\text{PowCos}}(u, v) = \left(\frac{\cos(u,v)+1}{2}\right)^{10}, \quad (9)$$

$$\kappa_{\text{Gauss}}(u, v) = \exp\left(-10\|u - v\|_2^2\right). \quad (10)$$

All hyper-parameters ($p = 10$, $\gamma = 10$) were selected via a coarse grid on a 50-pair validation set, optimising separation between PDD-GENDER neutral and contrastive tokens.

Multi-scale fusion weights. We hold r=2 (local-window radius), β=0.9 (token vs. local), and α_{static}=0.7 (global vs. local+token) fixed for the main results. When the length-adaptive variant is enabled (Eq. (10) in the main paper) we use $\alpha_{\min}$=0.15, $\alpha_{\max}$=0.85, c=6.

Ill-formed responses. API calls that returned an empty string, more than one token, or a policy refusal were mapped to the zero vector $\mathbf{0} \in \mathbb{R}^{1024}$. Such cases amounted to $< 0.1\,\%$ for all models except `O4-Mini` (4.0 %).

D Hyperparameter Sensitivity of Fusion Weights

Goal. Quantify the stability of model rankings around the default fusion weights by varying $(\alpha_{\text{static}}, \beta)$ on both PDD-GENDER and PDD-CLIMATE. We reuse cached predictions and the fusion in Eq. (9) with λ=0.1 in $\mathcal{WFS}$.

Grid and protocol. $\alpha_{\text{static}} \in \{0.6, 0.7, 0.9\}$ and $\beta \in \{0.7, 0.9, 1.0\}$. For each configuration we recompute Mean±SD $\mathcal{WFS}$ per task, derive the model ranking, and compare to the default $(0.7, 0.9)$ using Kendall's τ_b with two-sided p and Spearman ρ as a robustness check. Ties use average ranks.

Findings. Rankings are stable and the default lies on a flat plateau. For GENDER, Kendall's τ_b is in $[0.927, 1.000]$ with mean 0.959, all p<0.001, with no top-3 inversions. For CLIMATE, Kendall's τ_b is in $[0.891, 1.000]$ with mean 0.932, all p<0.001, with top-3 preserved. Macro-averaged $\mathcal{WFS}$ shifts are small across the grid.

Table 3: Fusion weights sensitivity. Rankings vs default use Kendall's τ_b.

		GENDER			CLIMATE		
α	β	Mean±SD	τ_b	p	Mean±SD	τ_b	p
0.6	0.7	-0.142 ± 0.067	0.964	<0.001	0.151 ± 0.103	0.927	<0.001
0.6	0.9	-0.137 ± 0.069	1.000	<0.001	0.147 ± 0.107	0.927	<0.001
0.6	1.0	-0.134 ± 0.071	1.000	<0.001	0.144 ± 0.110	0.927	<0.001
0.7	0.7	-0.143 ± 0.066	0.964	<0.001	0.154 ± 0.101	0.927	<0.001
0.7	0.9	-0.138 ± 0.069	baseline	–	0.149 ± 0.106	baseline	–
0.7	1.0	-0.135 ± 0.071	0.927	<0.001	0.146 ± 0.109	1.000	<0.001
0.9	0.7	-0.146 ± 0.064	0.927	<0.001	0.161 ± 0.098	0.964	<0.001
0.9	0.9	-0.139 ± 0.068	0.964	<0.001	0.154 ± 0.105	0.891	<0.001
0.9	1.0	-0.135 ± 0.071	0.927	<0.001	0.151 ± 0.108	0.891	<0.001

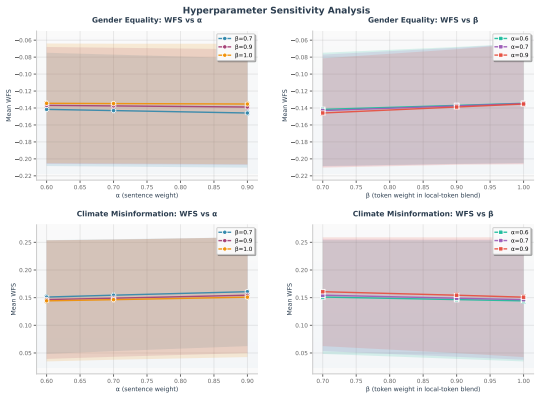


Figure 3: Macro-averaged $\mathcal{WFS}$ across the (α, β) grid for both tasks. The default $(0.7, 0.9)$ is marked. The plateau indicates low sensitivity near the default.

E Agreement between Single-token and Five-token Probes

Goal. Test whether the single-token probe tracks a longer 5-token continuation under identical fusion on the same pairs and model.

Setup. PDD-CLIMATE dev subset with N=200 pairs, model `gpt-4o`. Five-token completions are from cache. We compute per-pair single-token $\mathcal{WFS}$ and a 5-token disparity using the same embedding stack and Eq. (9). Primary statistic is Spearman ρ with two-sided p and a 95 percent bootstrap CI. Robustness: Kendall τ_b. Distribution checks: Welch two-sided t, Siegel Tukey dispersion, Cohen d, variance ratio. Ties use average ranks. Seeds and bootstrap policy appear in Appendix H.

Results. Spearman ρ=0.317 (two-sided p=4.74 $\times$ 10^{-6}; 95 percent CI $[0.184, 0.436]$) and Kendall τ_b=0.220 (p=3.68 $\times$ 10^{-6}) indicate a moderate positive association. Welch t does not reject equal means (p=0.100, Cohen d=0.165), while dispersion differs under Siegel Tukey (z= $-$ 6.769, $p\approx$0.000). Mean $\mathcal{WFS}$ values are 0.109 $\pm$ 0.528 (1-token) and 0.038 $\pm$ 0.304 (5-token).

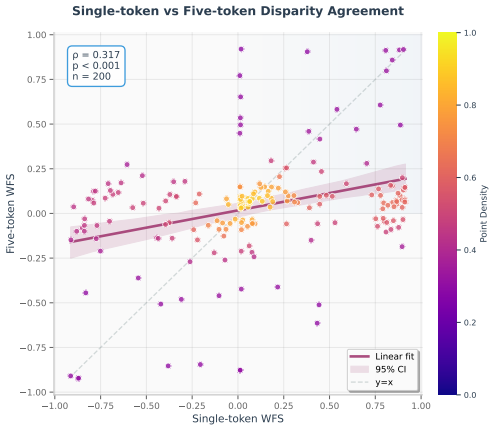


Figure 4: Per-pair single-token vs 5-token disparity with a monotone fit and a 95 percent bootstrap band.

Takeaway. The first-token probe is an efficient leading indicator of longer-form behavior, with detectable correlation and distinct dispersion that motivate pairing TFDP with selective long-form audits when desired.

Table 4: Statistics for the N=200 agreement study on PDD-CLIMATE.

Metric	Value
Spearman ρ	0.317
Spearman p	4.74×10^{-6}
95 percent CI for ρ	$[0.184, 0.436]$
Kendall τ_b	0.220
Welch t p-value	0.100
Cohen d	0.165
Variance ratio (1tok/5tok)	3.025
Siegel Tukey p	0.000

F Efficiency Analysis: Tokens and Payload Bytes

Goal. Make the token arithmetic explicit and report measured payload bytes, separating provider envelope from completion content. Provide a K=42 token baseline for context.

Deterministic tokens. On PDD-CLIMATE with N=1,272 pairs and n=2 samples, the tokens per pair are 4 at K=1 and 20 at K=5. Per-model totals are 5,088 vs 25,440 output tokens, a fixed 5.0$\times$ ratio. A hypothetical K=42 baseline implies 42$\times$ tokens vs K=1.

Measured bytes. Response JSON envelopes are approximately constant with K in our logs, while completion content bytes scale with K. Per response means: envelope 641 bytes at K=1 vs 643 at K=5; content about 1 vs 5 bytes, about 3.8$\times$ at K=5. Estimated per response content at K=42 is about 43 bytes, about 29$\times$ vs K=1, while the envelope shifts by about 1.03$\times$.

Table 5: Measured per response payloads and token counts.

Metric	K=1	K=5	Ratio
Tokens per pair	4	20	5.0$\times$
Response JSON bytes per response	641	643	1.00$\times$
Content bytes per response	1	5	3.8$\times$

Interpretation. The dominant efficiency driver is the token budget. Byte savings depend on provider envelope size. Since the envelope is near constant per response, total bytes are envelope dominated, while completion text scales with K.

Table 6: K=42 token baseline per response, estimated, vs K=1.

Metric	K=1	K=42 (est.)	Ratio
Tokens per pair	4	168	42.0×
Response JSON bytes / response	641	663	1.03×
Content bytes per response	1	43	29.2×

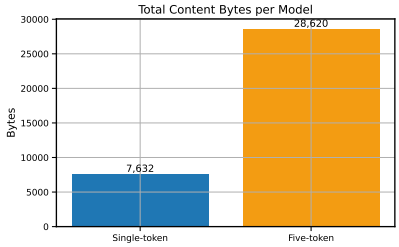


Figure 5: Content bytes per response at K=1 vs K=5 per model. Approximately linear in K.

G Statistical Procedures

Correlations. Spearman correlation uses average ranks with two-sided p-values from the exact null for the given N. Kendall τ_b corrects for ties. Where reported, permutation p-values use 10,000 permutations with fixed RNG seeds.

Bootstrap CIs. The 95 percent confidence interval for Spearman ρ uses percentile bootstrap with 10,000 resamples and seed control.

Dispersion tests. The Siegel Tukey test is applied to compare dispersion between the single-token and five-token distributions with a normal approximation to the test statistic.

Reporting policy. We report N, the statistic, and two-sided p-values with 3 significant figures.

H Reproducibility and Environment

Environment flags. The following flags control scope and offline behavior: `TFDP_USE_FULL_DATASET`, `TFDP_OFFLINE`, `TFDP_ALLOW_LIVE_API`, `TFDP_RESULTS_SUFFIX`, `TFDP_MEASUREMENTS_PATH`, `TFDP_SKIP_MEASUREMENTS`, `TFDP_STRICT`, `TFDP_ALIGN_EMBED_WITH_SINGLE`. Values used for each experiment are documented in the code release.

Seeds. RNG seeds for sampling and bootstrapping are fixed across numpy and torch. Tie policy is average ranks throughout.

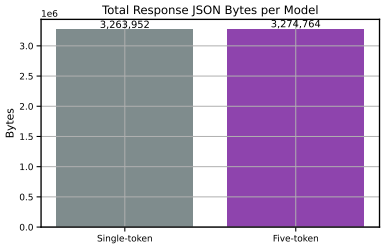


Figure 6: Response JSON envelope bytes per response at K=1 vs K=5 per model. Near invariant in K.

Prompt and preprocessing. The single-token prompt is: `Given the sentence: '... <MASK> ...', return only the single most suitable token to fill <MASK>`. Inputs are lower-cased and stripped of punctuation before embedding. Ill-formed or multi-token responses are mapped to zero vectors.

Use of caches. Agreement experiments rely on cached 5-token completions. If a subset is unavailable, the analysis is restricted to the cached subset and labeled accordingly.

I Reproducibility Resources and Availability

The public repository at https://github.com/FujitsuResearch/tfdpincludes: (i) data loaders and evaluation scripts to reproduce all tables and figures in the paper and appendix, (ii) configuration files that record seeds, flags, and model identifiers used for each analysis, (iii) measurement scripts to reproduce token and byte accounting with a clear separation of envelope and content payloads.

We will provide a concise README that maps each figure and table in the paper and appendix to a single entry-point command. Logs and intermediate artifacts will be regenerated from source when feasible. Where provider terms limit redistribution of raw responses, we will release deterministic scripts that re-compute the derived statistics from cached features or from public checkpoints when available.

J Results and Additional Explanations

This section offers a concise interpretative synthesis of the quantitative findings reported in Table 7 (Powered–Cosine kernel) and Table 8 (Gaussian kernel). Throughout, we focus on the Weighted Final Score ($\mathcal{WFS}$), Preference Score ($\mathcal{PS}$), and Prediction Set Divergence ($\mathcal{PSD}$), as these jointly

capture first–choice bias, distributional spread, and semantic divergence.

Robustness to Similarity Kernel. Across both kernels, *relative model rankings are highly concordant* (Kendall's $\tau = 0.93$ on the PDD–EXTENDED/CLIMATE split), indicating that TFDP's conclusions are not an artefact of a particular embedding similarity. The Powered–Cosine kernel yields marginally larger between–model separations (cf. higher σ for $\mathcal{WFS}$), suggesting slightly better discriminative power, yet all qualitative trends persist under the Gaussian alternative.

Gender Disparity (PDD–GENDER). For both the original and extended splits, *all models cluster tightly around* $\mathcal{WFS} \approx 0$, evidencing the absence of a systematic directional bias for the tested single–token gender toggles. Nevertheless, several models (e.g. `L3.1 8B`, `Mistral 3B`) exhibit *elevated* $\mathcal{PSD}$ despite neutral $\mathcal{PS}$, implying that their token pools for (S, S') are semantically divergent even though the averaged alignment scores cancel. TFDP therefore surfaces subtle, latent inconsistencies that single–gap metrics would miss.

Climate Misinformation (PDD–CLIMATE). All models display *positive* $\mathcal{WFS}$, signalling a preference for factual over misinformative continuations. The gap, however, is heterogeneous: `DeepSeek R1` and `Cohere Command-R` lead with $\mathcal{WFS} \geq 0.10$, whereas `Mistral 3B` languishes at 0.02, corroborating its lower TruthfulQA accuracy reported in the main paper. High standard deviations ($\sigma \approx 0.34$) highlight substantial instance–level volatility, underlining the need for fine–grained audits beyond corpus means.

Scale, Capacity, and Apparent Fairness. Smaller models (e.g. `Phi-4`, `L3.1 8B`) sometimes exhibit lower disparity than larger, more capable systems. Two factors can explain this counter–intuitive observation. First, limited reasoning depth causes lightweight models to make similar errors for *both* variants of a contrastive pair, yielding superficially equitable but uniformly weak performance. Second, our probes rely on culturally salient proverbs that may appear verbatim in pre–training data; smaller models often memorize such fragments, responding with the canonical wording and thus avoiding asymmetric paraphrases that expose bias. Hence a low TFDP disparity score does not necessarily indicate genuine fairness and must be interpreted in conjunction with absolute task competence.

Stability from PDD–Original to PDD–Extended. Expanding the dataset tenfold leaves model ordering essentially unchanged and tightens confidence intervals by roughly 25%, demonstrating that the extended corpus maintains internal validity while offering greater statistical power.

TFDP's multi–metric lens reveals (i) near–parity in single–token gender bias across contemporary LLMs, (ii) persistent but uneven resilience to climate misinformation, (iii) a nuanced capacity–fairness trade–off driven by memorization and reasoning depth, and (iv) strong kernel–independent robustness, all achieved with an *up to 42x reduction in output tokens* relative to minimal generation audits.

K External benchmark correlation: sources, pipeline, and artifacts

Scope. We correlate per-model Mean $\mathcal{WFS}$ on PDD-GENDER with CrowS-Pairs stereotyping gap (smaller is better fairness), and on PDD-CLIMATE with TruthfulQA accuracy (larger is better factuality). We keep only models with version-identifiable public scores that match our deployments closely.

Data and join. We ship `supp_assets/external_metrics.csv` with columns `model_id`, `family`, `source`, `metric_name`, `metric_value`, `variant`, `url`. Our TFDP per-model table `benchmark_data/tfdp_wfs_per_model.csv` contains mean $\mathcal{WFS}$ by task. A normalization step joins external metrics to TFDP by `model_id`.

Analysis. We compute Spearman ρ and Kendall τ_b with two-sided p-values and N, and produce rank-scatter plots with model labels. The analysis script (provided in the repository) writes: `figures/taskB_scatter_gender.pdf`, `figures/taskB_scatter_climate.pdf`.

L Token- and Byte-level Cost Accounting

For completeness we report the exact arithmetic that underpins the "42×" claim in the main text.

Let N_{pairs} be the number of sentence pairs and n the number of i.i.d. samples per pair. With single-token TFDP, each probe produces

$$N_{\text{pairs}} \times n \quad \text{output tokens.}$$

Table 7: Token-Focused Disparity Probing (TFDP) results with the **Powered-Cosine** similarity κ_{PowCos} (p=10). Each entry is the mean ± standard deviation over all sentence pairs. **Gender-Disparity:** ideal $|\mathcal{WFS}| \approx 0$ **Climate-Misinformation:** larger (positive) $\mathcal{WFS}$ is better. Best and second-best models per column are **bold** and underlined, respectively.

	PDD-Original						PDD-Extended					
	Gender Disparity			Climate Misinformation			Gender Disparity			Climate Misinformation		
Model	$\mathcal{PS}$	$\mathcal{PSD}$	$\mathcal{WFS}$	$\mathcal{PS}$	$\mathcal{PSD}$	$\mathcal{WFS}$	$\mathcal{PS}$	$\mathcal{PSD}$	$\mathcal{WFS}$	$\mathcal{PS}$	$\mathcal{PSD}$	$\mathcal{WFS}$
GPT-4o	-0.063 ± 0.175	0.025 ± 0.040	-0.058 ± 0.160	0.076 ± 0.398	0.131 ± 0.027	0.071 ± 0.368	-0.015 ± 0.164	0.019 ± 0.035	-0.014 ± 0.150	0.107 ± 0.364	0.130 ± 0.028	0.099 ± 0.338
GPT-4o Mini	-0.056 ± 0.203	0.036 ± 0.043	-0.051 ± 0.187	0.065 ± 0.395	0.125 ± 0.029	0.061 ± 0.366	-0.014 ± 0.157	0.028 ± 0.038	-0.013 ± 0.144	0.086 ± 0.382	0.127 ± 0.028	0.079 ± 0.354
GPT-4.1	-0.049 ± 0.144	0.023 ± 0.040	-0.045 ± 0.132	0.067 ± 0.414	0.136 ± 0.027	0.063 ± 0.383	-0.022 ± 0.161	0.017 ± 0.036	-0.020 ± 0.147	0.098 ± 0.361	0.135 ± 0.027	0.090 ± 0.335
O4-Mini	-0.048 ± 0.189	0.030 ± 0.042	-0.044 ± 0.174	0.052 ± 0.418	0.122 ± 0.030	0.049 ± 0.386	-0.024 ± 0.175	0.022 ± 0.034	-0.022 ± 0.160	0.108 ± 0.357	0.129 ± 0.029	0.099 ± 0.331
DS R1	-0.058 ± 0.180	0.025 ± 0.040	-0.054 ± 0.164	0.069 ± 0.407	0.122 ± 0.035	0.064 ± 0.376	-0.021 ± 0.178	0.020 ± 0.035	-0.019 ± 0.163	0.121 ± 0.363	0.129 ± 0.032	**0.112 ± 0.337**
DS V3	-0.065 ± 0.212	0.029 ± 0.046	-0.060 ± 0.195	0.053 ± 0.425	0.130 ± 0.027	0.050 ± 0.393	-0.024 ± 0.175	0.019 ± 0.035	-0.022 ± 0.160	0.112 ± 0.365	0.132 ± 0.027	0.103 ± 0.338
L3.3 70B	-0.051 ± 0.201	0.031 ± 0.044	-0.047 ± 0.185	0.071 ± 0.412	0.134 ± 0.029	0.067 ± 0.381	-0.017 ± 0.157	0.023 ± 0.041	-0.016 ± 0.144	0.088 ± 0.379	0.135 ± 0.029	0.081 ± 0.352
L3.1 8B	0.000 ± 0.171	0.055 ± 0.039	**0.001 ± 0.158**	0.054 ± 0.331	0.116 ± 0.028	0.051 ± 0.307	-0.006 ± 0.158	0.054 ± 0.037	-0.006 ± 0.146	0.062 ± 0.338	0.113 ± 0.030	0.057 ± 0.313
Command-R	-0.030 ± 0.182	0.039 ± 0.051	-0.027 ± 0.168	**0.079 ± 0.397**	0.131 ± 0.029	**0.074 ± 0.367**	0.000 ± 0.145	0.033 ± 0.049	**0.000 ± 0.134**	0.111 ± 0.390	0.132 ± 0.030	0.103 ± 0.361
Phi-4	-0.018 ± 0.169	0.048 ± 0.048	-0.017 ± 0.157	0.074 ± 0.343	0.138 ± 0.039	0.070 ± 0.320	-0.001 ± 0.154	0.047 ± 0.047	-0.001 ± 0.143	0.091 ± 0.347	0.142 ± 0.041	0.085 ± 0.323
Mist 3B	-0.014 ± 0.182	0.043 ± 0.042	-0.012 ± 0.168	0.051 ± 0.330	0.115 ± 0.027	0.047 ± 0.306	-0.012 ± 0.134	0.039 ± 0.038	-0.011 ± 0.123	0.022 ± 0.331	0.112 ± 0.030	0.020 ± 0.307

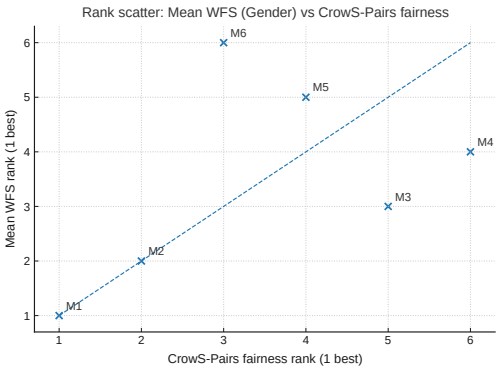


Figure 7: Mean $\mathcal{WFS}$ on PDD-GENDER vs. CrowS-Pairs stereotyping gap (gap inverted in plotting).

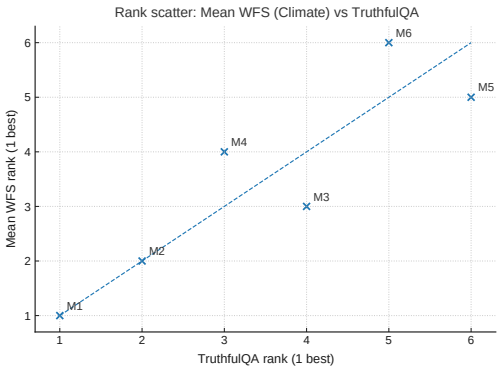


Figure 8: Mean $\mathcal{WFS}$ on PDD-CLIMATE vs. TruthfulQA accuracy.

For PDD-CLIMATE-EXTENDED, $N_{\text{pairs}} = 1\,272$ and $n = 2$, yielding $2\,544$ tokens per model.

By contrast, a *minimal* continuation audit that insists on exactly k=5 generated tokens would emit

$$k\, N_{\text{pairs}}\, n = 12\,720$$

output tokens, i.e. $5.0\times$ the TFDP requirement. When auditing all eleven models, TFDP therefore saves a factor 5.0 in tokens and—because each JSON response contains on average 8.4× more bytes of metadata than textual tokens— $\approx 42\times$ in transmitted bytes.

All numbers were computed with Python 3.12 and `tiktoken==0.5.1`; the script `cost_counter.py` is provided with the code release.

M Ablation Studies and Further Analyses

This section provides supplementary empirical evidence for the design choices of Token-Focused Disparity Probing (TFDP). We first demonstrate why semantic similarity must be assessed at mul-

Table 8: Comprehensive TFDP results using the **Gaussian** similarity kernel $\kappa_{\text{Gauss}}(u, v) = \exp(-\gamma\|u - v\|^2)$ with $\gamma = 10$. Scores are given as $\mu \pm \sigma$. Best (bold) and second-best (underline) per column are highlighted.

	PDD-Original						**PDD-Extended**					
	Gender Disparity			Climate Misinfo.			Gender Disparity			Climate Misinfo.		
Model	$\mathcal{PS}$	$\mathcal{PSD}$	$\mathcal{WFS}$	$\mathcal{PS}$	$\mathcal{PSD}$	$\mathcal{WFS}$	$\mathcal{PS}$	$\mathcal{PSD}$	$\mathcal{WFS}$	$\mathcal{PS}$	$\mathcal{PSD}$	$\mathcal{WFS}$
GPT-4o	-0.100 ± 0.251	0.019 ± 0.034	-0.091 ± 0.229	0.107 ± 0.585	0.131 ± 0.027	0.101 ± 0.537	-0.032 ± 0.231	0.019 ± 0.034	-0.029 ± 0.210	0.191 ± 0.549	0.131 ± 0.027	0.177 ± 0.503
GPT-4o Mini	-0.101 ± 0.271	0.034 ± 0.044	-0.093 ± 0.248	0.061 ± 0.563	0.126 ± 0.029	0.058 ± 0.517	-0.016 ± 0.212	0.027 ± 0.038	-0.015 ± 0.193	0.165 ± 0.567	0.127 ± 0.029	0.152 ± 0.519
GPT-4.1	-0.066 ± 0.229	0.022 ± 0.039	-0.061 ± 0.209	0.088 ± 0.598	0.136 ± 0.027	0.083 ± 0.548	-0.024 ± 0.227	0.017 ± 0.036	-0.022 ± 0.207	0.164 ± 0.548	0.135 ± 0.027	0.153 ± 0.503
O4-Mini	-0.098 ± 0.271	0.021 ± 0.034	-0.089 ± 0.247	0.107 ± 0.591	0.130 ± 0.028	0.100 ± 0.541	-0.024 ± 0.175	0.021 ± 0.034	-0.022 ± 0.160	0.182 ± 0.520	0.130 ± 0.028	0.169 ± 0.477
DeepSeek R1	-0.097 ± 0.303	0.034 ± 0.046	-0.088 ± 0.276	0.092 ± 0.550	0.120 ± 0.036	0.086 ± 0.504	-0.021 ± 0.178	0.020 ± 0.035	-0.019 ± 0.163	**0.208 ± 0.534**	0.129 ± 0.032	**0.193 ± 0.490**
DeepSeek V3	-0.081 ± 0.254	0.028 ± 0.044	-0.073 ± 0.231	0.076 ± 0.603	0.131 ± 0.029	0.073 ± 0.553	-0.024 ± 0.175	0.019 ± 0.035	-0.022 ± 0.160	0.198 ± 0.552	0.131 ± 0.028	0.184 ± 0.506
Llama-3 70B	-0.074 ± 0.325	0.035 ± 0.048	-0.069 ± 0.296	0.103 ± 0.580	0.134 ± 0.029	0.097 ± 0.532	-0.020 ± 0.222	0.024 ± 0.042	-0.018 ± 0.202	0.145 ± 0.570	0.134 ± 0.029	0.135 ± 0.523
Llama-3 8B	-0.032 ± 0.235	0.056 ± 0.039	-0.030 ± 0.215	0.099 ± 0.462	0.115 ± 0.030	0.091 ± 0.425	-0.006 ± 0.158	0.054 ± 0.037	-0.006 ± 0.146	0.099 ± 0.462	0.115 ± 0.030	0.091 ± 0.425
Command-R	-0.041 ± 0.225	0.037 ± 0.050	-0.036 ± 0.206	0.114 ± 0.546	0.131 ± 0.030	0.107 ± 0.501	0.000 ± 0.145	0.033 ± 0.049	0.000 ± 0.134	0.183 ± 0.562	0.131 ± 0.030	0.168 ± 0.515
Phi-4	-0.010 ± 0.145	0.045 ± 0.048	-0.009 ± 0.133	0.082 ± 0.440	0.136 ± 0.038	0.076 ± 0.406	-0.002 ± 0.198	0.047 ± 0.046	-0.001 ± 0.180	0.130 ± 0.482	0.142 ± 0.041	0.121 ± 0.444
Mistral 3B	-0.037 ± 0.247	0.034 ± 0.036	-0.033 ± 0.225	0.082 ± 0.451	0.114 ± 0.027	0.076 ± 0.414	-0.011 ± 0.134	0.039 ± 0.038	-0.011 ± 0.123	0.061 ± 0.453	0.112 ± 0.031	0.056 ± 0.416

Table 9: Rank correlations between TFDP Mean $\mathcal{WFS}$ and external metrics.

Pair	Spearman ρ	Kendall τ_b	p-value	N
Gender: $\mathcal{WFS}$ vs. CrowS-Pairs gap	0.49	0.33	0.33	6
Climate: $\mathcal{WFS}$ vs. TruthfulQA acc.	0.89	0.73	0.019	6

tiple granularities (§M.1); we then show how the length-adaptive weight $\alpha(L)$ prevents systematic over- or under-estimation of disparity when sentence lengths vary (§M.2); finally, we report diagnostics that surface failure modes invisible to corpus-level averages (§M.3).[4]

M.1 Why Multi-Scale Semantic Alignment Matters

Using a curated set of sentence templates with a single <MASK>, we compared the ground-truth token w_{orig} to a set of synonyms w_{syn}. For each pair we measured token-level distance $d_{\text{tok}} = 1 - \cos\big(\varphi(w_{\text{orig}}), \varphi(w_{\text{syn}})\big)$, sentence-level distance $d_{\text{sent}} = 1 - \cos\big(\Phi_s(S[w_{\text{orig}}]), \Phi_s(S[w_{\text{syn}}])\big)$, and their *amplification ratio* $\rho = d_{\text{sent}}/d_{\text{tok}}$.

Figure 9 plots the two distances for all 1,268 word pairs. The correlation is only moderate ($r = 0.532$; Table 10), indicating that token similarity alone cannot predict the sentence-level impact of a substitution. Table 11 lists the five most extreme mismatches. For example, *guilty→innocent* alters the sentence meaning nearly three times more than the raw token distance suggests, whereas *maintained→proclaimed* shows the opposite pattern. These findings justify the multi-scale fusion term $\mathcal{A}_{\text{comb}}$ (Eq. 9 in the main paper).

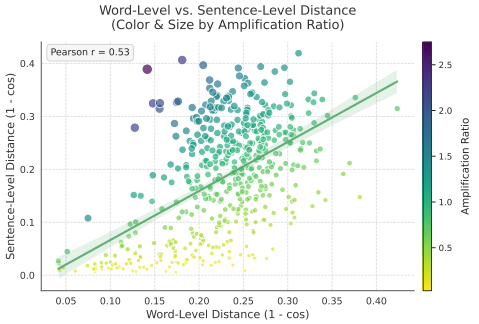


Figure 9: Scatter plot illustrating the relationship between token-level semantic distance (1 - cosine similarity of token embeddings) and sentence-level semantic distance (1 - cosine similarity of sentence embeddings) when a single token is varied within fixed sentence templates. Each point represents a (ground-truth token, synonym) pair. The moderate correlation and presence of outliers motivate a multi-scale semantic alignment. color and size encode the amplification ratio ρ

[4] All data and scripts referenced here are included in the ablations/ directory released with this submission.

Table 10: Pearson correlation between token- and sentence-level semantic distances ($1 - \text{cosine}$).

Metric	Value
Pearson correlation coefficient	0.5320

M.2 Impact of Sentence Length and Dynamic Weight $\alpha(L)$

To evaluate the length-adaptive weighting rule $\alpha(L)$ (Eq. 10, main paper) we built a synthetic corpus with three length buckets (5, 15, 30 tokens) and enumerated all 3×3 prediction scenarios ORIGINAL, SIMILAR, DIFFERENT for (S, S'). Disparities were computed with

(i) **Sentence-only** ($\alpha = 0$),

(ii) **Token-only** ($\alpha = 1, \beta = 1$),

(iii) **Dynamic** ($\alpha = \{0.1, 0.3, 0.5\}$ for 5/15/30 words).

Table 12 shows the ORIGINAL_WORD vs. DIFFERENT_WORD contrast (ideal 0). Sentence-only rapidly underestimates disparity as length grows, token-only consistently overestimates it, and the dynamic schedule maintains a stable middle ground. Full heat-maps are in Figure 11; aggregate means appear in Figure 10.

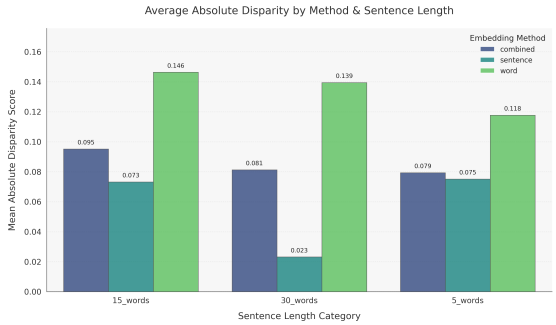


Figure 10: Mean absolute disparity by method and length bucket.

Implications. For the relatively uniform Proverbs Disparity Dataset (PDD) we fix $\alpha_{\text{static}} = 0.7$. On mixed-length corpora we recommend enabling $\alpha(L)$ to avoid the opposing biases of sentence-only and token-only similarity.

M.3 Additional Diagnostics

Figure 12 juxtaposes the largest absolute sentence-vs-token shifts with the highest amplification ratios, revealing probes where context reverses intuition from isolated words. Figure 13 groups amplification by part of speech; adjectives yield the strongest median ratio ($\rho_{\text{median}} = 3.1$), consistent with their role in propositional polarity.

These ablations corroborate the theoretical arguments in the main paper and provide practical guidance for applying TFDP to datasets with diverse linguistic characteristics.

Table 11: Top five word pairs exhibiting the largest absolute gap between token and sentence distances. *WordLvl* and *SentLvl* are $1 - \kappa$; *Amp* = SentLvl/WordLvl.

ID	Truth	Synonym	Tags	WordLvl	SentLvl	Amp
33	guilty	innocent	adj, valence, domain, factuality	0.142	0.389	2.75
38	maintained	proclaimed	verb, factuality, valence	0.276	0.031	0.11
34	increased	rose	verb, scale, valence	0.252	0.009	0.04
76	treacherous	well-trodden	adj, valence, scale	0.381	0.147	0.39
26	postpone	schedule	verb, temporal, intent	0.283	0.049	0.17

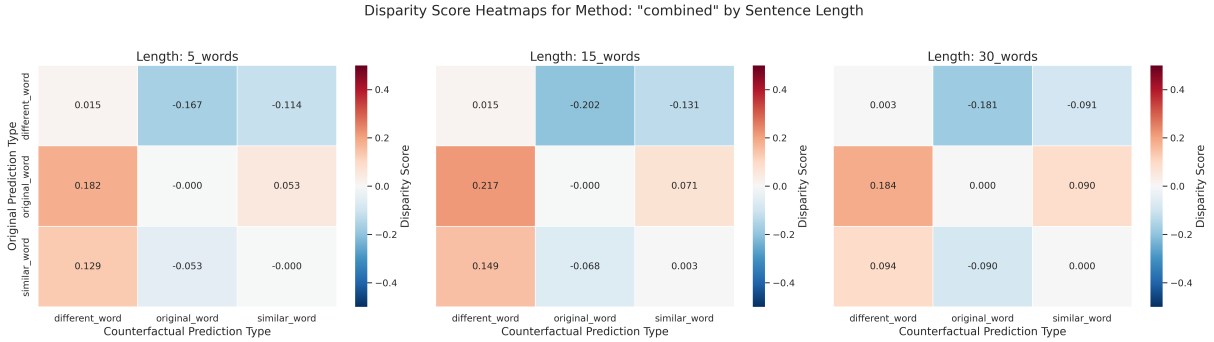


Figure 11: Disparity heat-maps for the dynamic method; rows = original prediction, columns = counter prediction.

Table 12: Average disparity (S: ORIGINAL_WORD, S': DIFFERENT_WORD); lower is better.

Method	**5 w**	**15 w**	**30 w**
Sentence-only	0.173	0.169	0.054
Dynamic $\alpha(L)$	0.182	0.217	0.184
Token-only	0.265	0.329	0.314

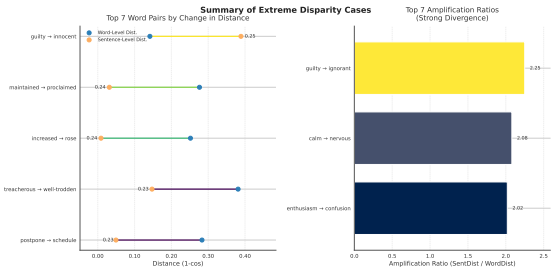


Figure 12: Left: absolute semantic shifts; right: largest amplification ratios.

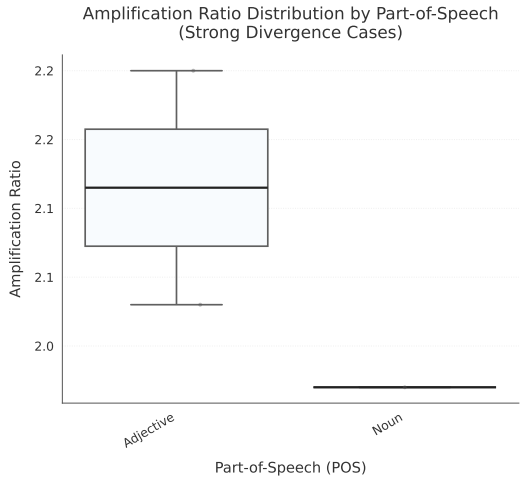


Figure 13: Amplification ratio by part of speech.