BIASGUARD PRO

By Dyuti Dasmahapatra

MOTIVATION

Career recommendation systems increasingly shape who gets seen and hired. However, they often reinforce gendered stereotypes - suggesting women for nurturing roles and men for technical or leadership positions. BiasGuard Pro addresses this challenge by introducing a lightweight, explainable, and human-centered framework for bias detection and mitigation in professional recommendation text.

Goal: Develop a scalable and transparent tool that bridges the gap between academic fairness research and deployable industry systems.

"Toward transparent and fair career AI systems."

METHODOLOGY OVERVIEW

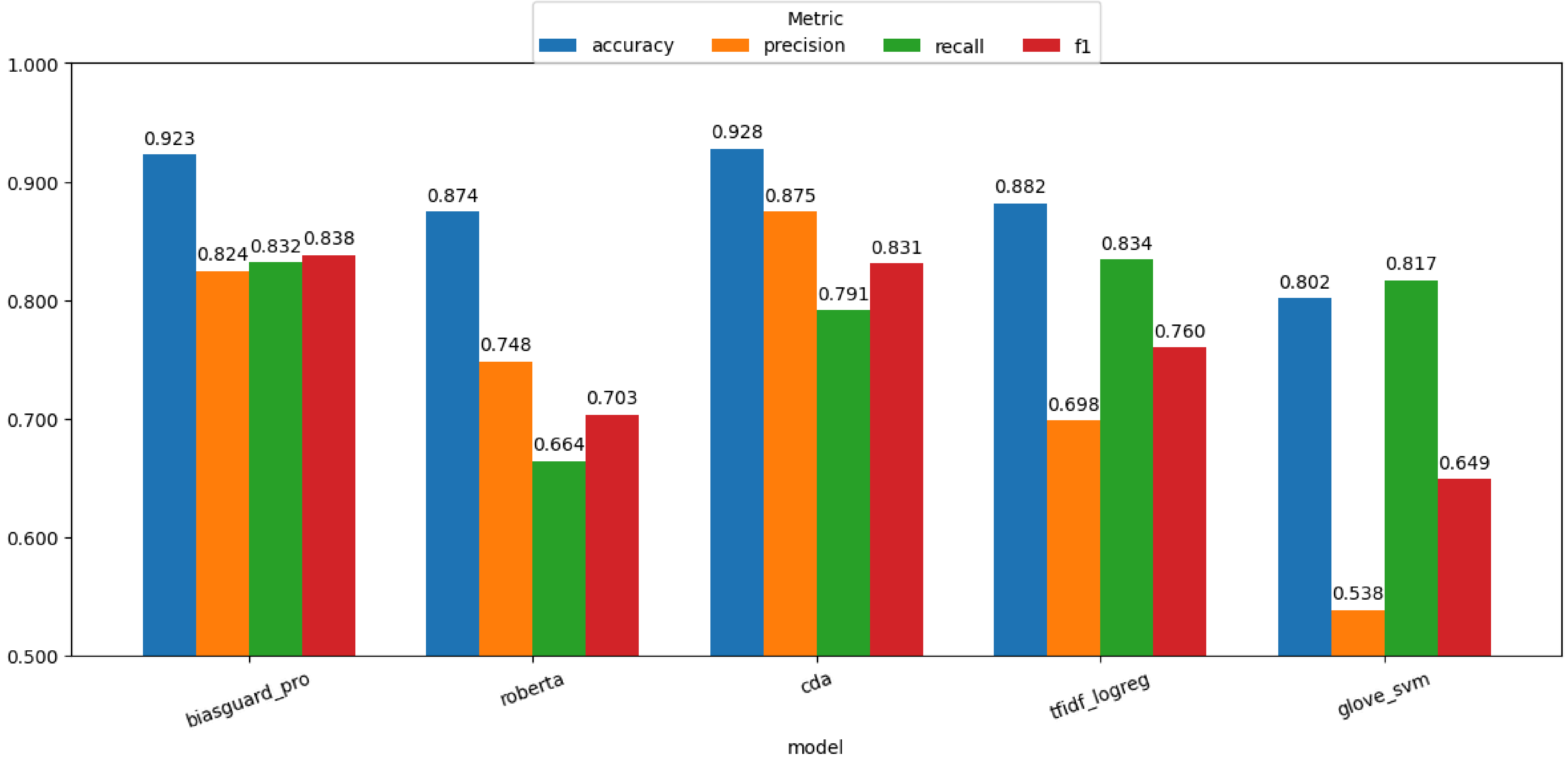
BiasGuard Pro integrates four key modules:

1. **DATA:** Combines GPT-4 generated synthetic LinkedIn-style text with real-world corpora (BiasBios, StereoSet) to build a balanced dataset.
2. **MODEL:** Fine-tunes a DistilBERT classifier for binary bias detection (biased/unbiased), optimized for both accuracy and latency
3. **EXPLAINABILITY:** Integrates SHAP for token-level feature attribution, DiCE for counterfactual generation, and prototype examples for human-centered interpretability.
4. **INTERFACE**: Implements a Gradio-based dashboard built on HCI principles such as progressive disclosure and participatory design.

RESULTS: QUANTITATIVE PERFORMANCE OUTCOMES

The CDA baseline, though comparable in F1 (0.831), showed reduced recall, highlighting its limited sensitivity to contextual or implicit bias cues. RoBERTa-base, while accurate, exhibited significant latency and decreased recall, suggesting overreliance on lexical-level correlations.

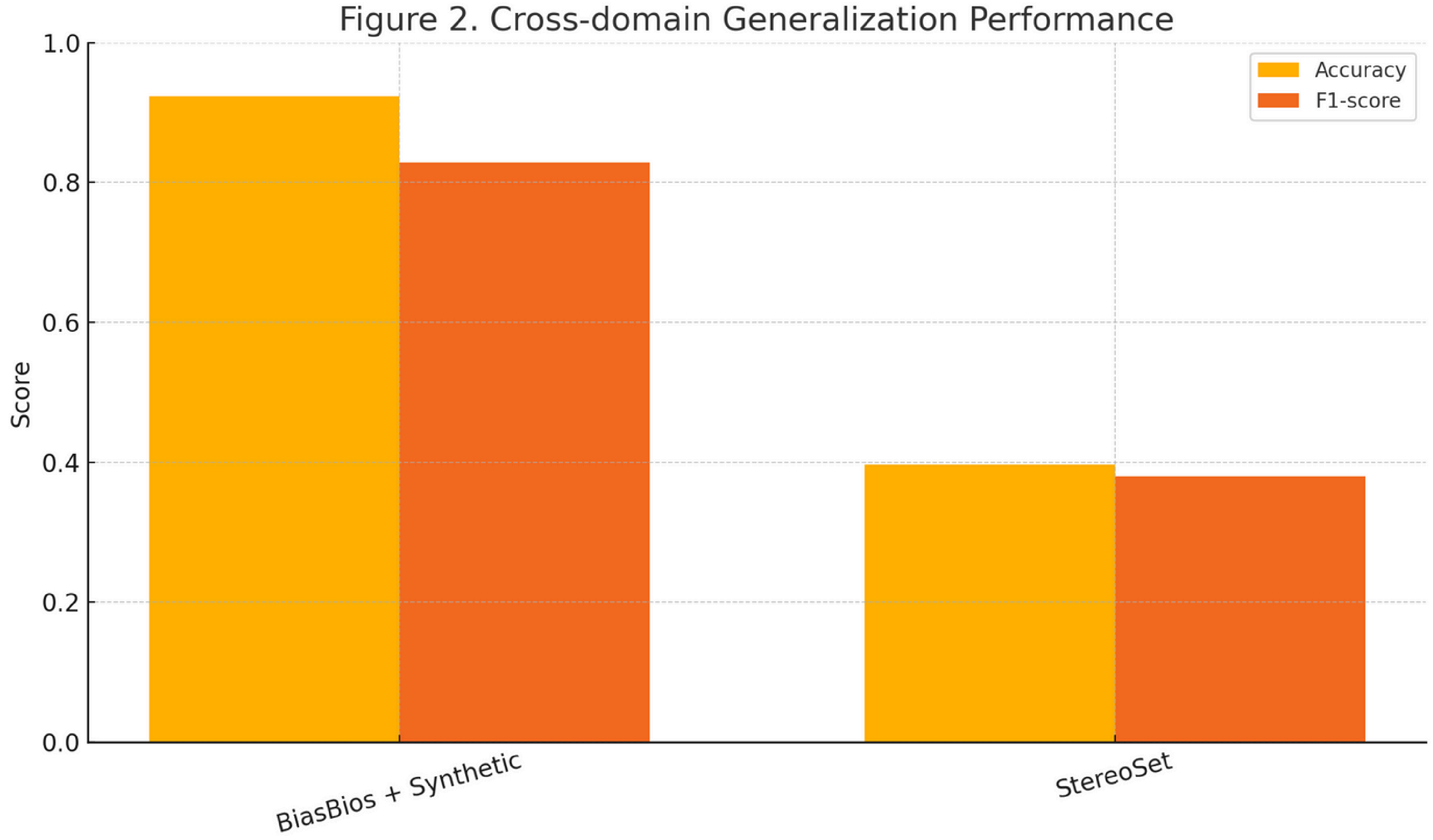
In contrast, BiasGuard Pro’s DistilBERT backbone offered consistent precision and recall with substantially lower computational overhead.



RESULTS: CROSS-DOMAIN GENERALIZATION PERFORMANCE

BiasGuard Pro demonstrates strong domain-specific performance, maintaining high accuracy on BiasBios and Synthetic LinkedIn datasets while showing predictable degradation on general benchmarks like StereoSet.

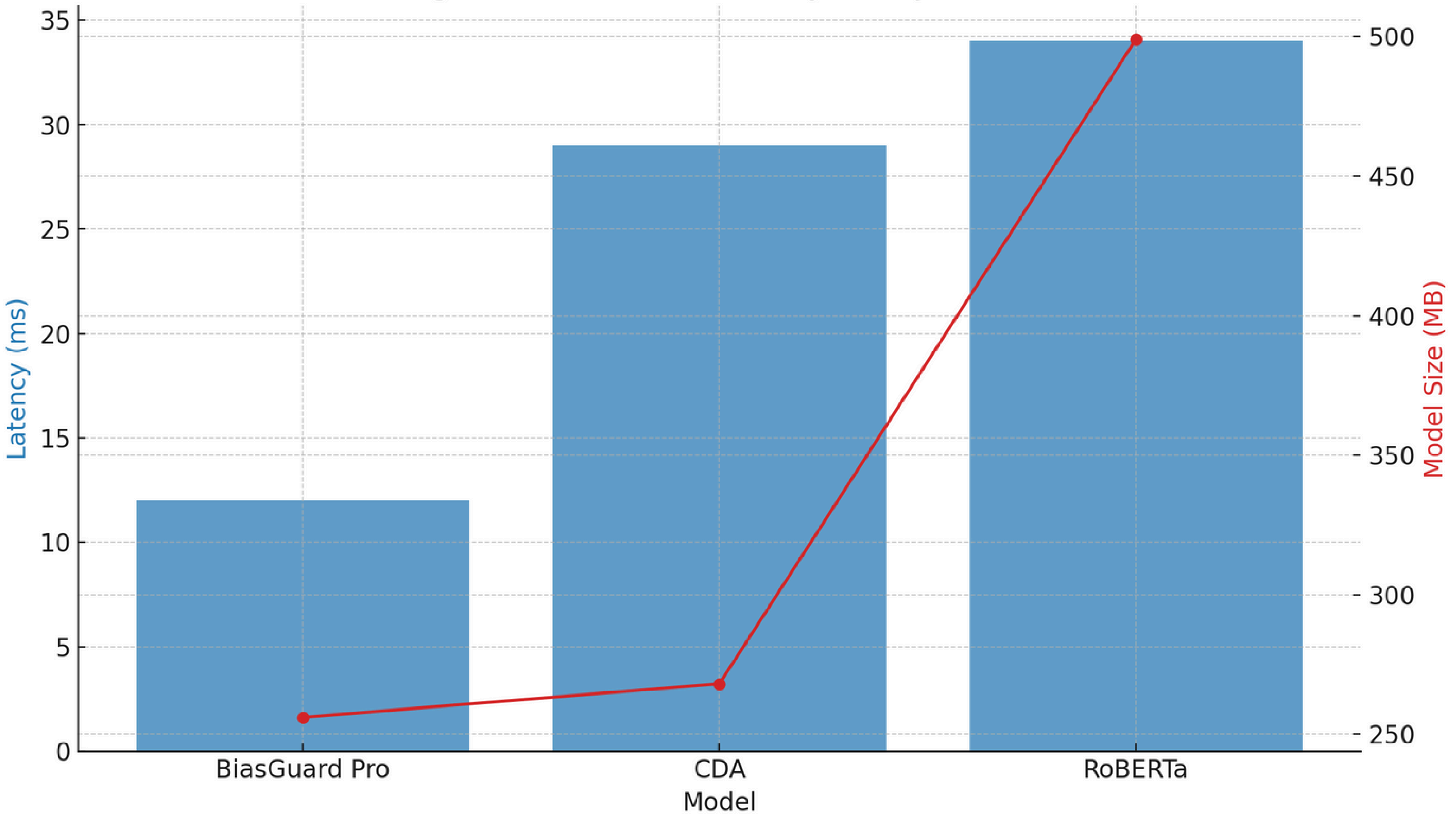
This behavior reflects appropriate domain specialization rather than overfitting.



RESULTS: MODEL EFFICIENCY COMPARISON

BiasGuard Pro achieves near-optimal efficiency while maintaining high fairness and interpretability. DistilBERT delivers a balanced tradeoff between speed, size, and transparency, making it suitable for real-time deployment in fairness auditing pipelines.

Figure 6. Model Efficiency Comparison



ETHICAL CONSIDERATIONS AND FUTURE WORK

Ethical Considerations:

- Intended use: auditing and awareness, not automated filtering.
- Scope: English-only, gender bias detection.
- Human oversight: mandatory for decision-making.
- Transparency: open-source code, reproducible datasets, and documented limitations.

Future Work:

- Expand to multilingual and intersectional contexts (race, age, ethnicity).
- Integrate fairness-aware loss functions and adaptive learning.
- Collaborate with HR partners for IRB-approved validation.
- Build continuous fairness auditing pipelines for large-scale deployment.