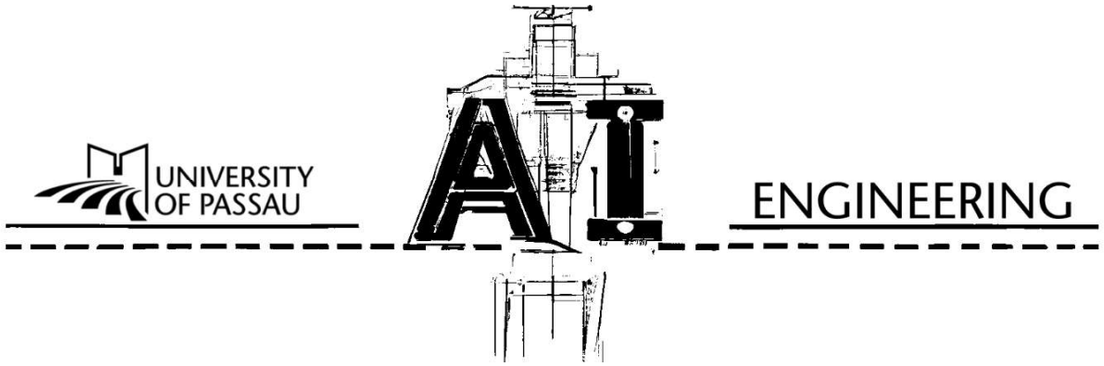
UNIVERSITY OF PASSAU
Faculty of Computer Science and Mathematics

GRADIEND: Feature Learning in Neural Networks Exemplified through Biases

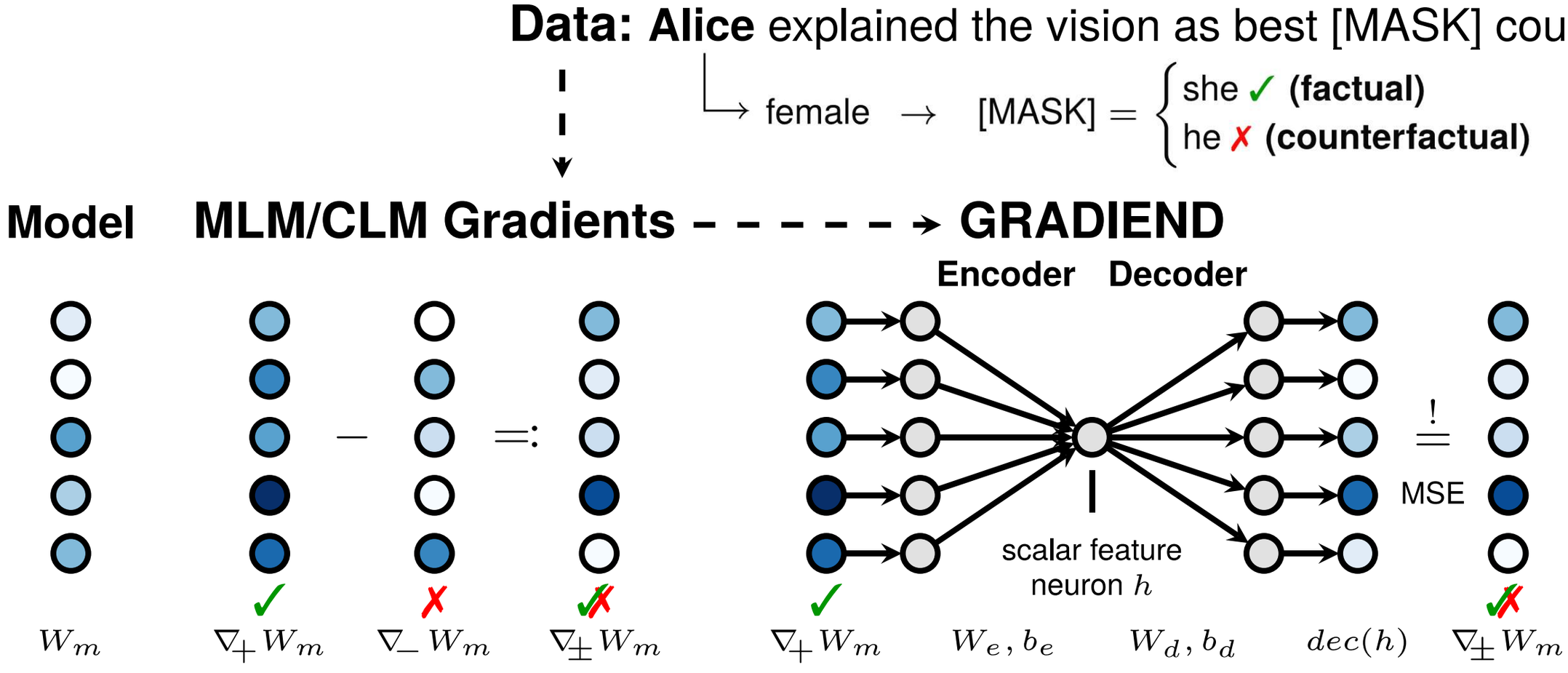
Jonathan Drechsel, Steffen Herbold
University of Passau
ICLR 2026



ICLR

UNIVERSITY OF PASSAU
Faculty of Computer Science and Mathematics

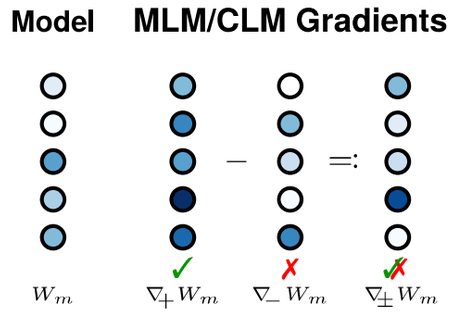
GRADIEND enables targeted feature learning

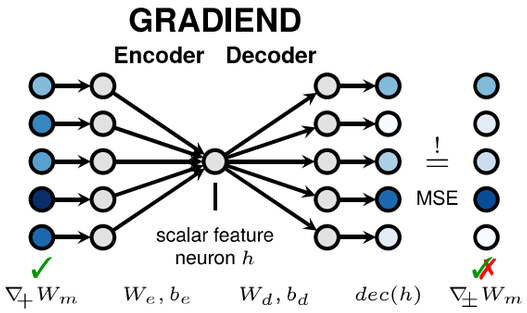


GRADIEND – Mathematical Breakdown

GRADIEND as encoder-decoder network

- $f = dec \circ enc$ with
  - $enc(\nabla_+ W_m) = \tanh(W_e^T \cdot \nabla_+ W_m + b_e) =: h$
    - tanh ensures encoding in $[-1, +1]$
    - h is a scalar feature value
  - $dec(h) = h \cdot W_d + b_d \approx \nabla_\pm W_m$
- Learn f such that $f(\nabla_+ W_m) \approx \nabla_\pm W_m$
  - $W_e, W_d, b_d \in \mathbb{R}^n, b_e \in \mathbb{R}$ are learnable parameters





GRADIEND model update

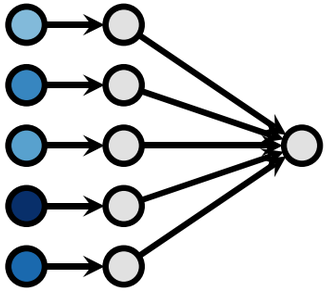
$$\widetilde{W}_m := W_m + \overbrace{\alpha}^{\text{update strength}} \cdot \underbrace{dec(h^*)}_{\text{feature direction}}$$

- α – learning rate controlling the update strength
- h^* – feature factor controlling the feature direction

GRADIEND Encoder Learns the Targeted Feature

UNIVERSITY OF PASSAU
Faculty of Computer Science and Mathematics

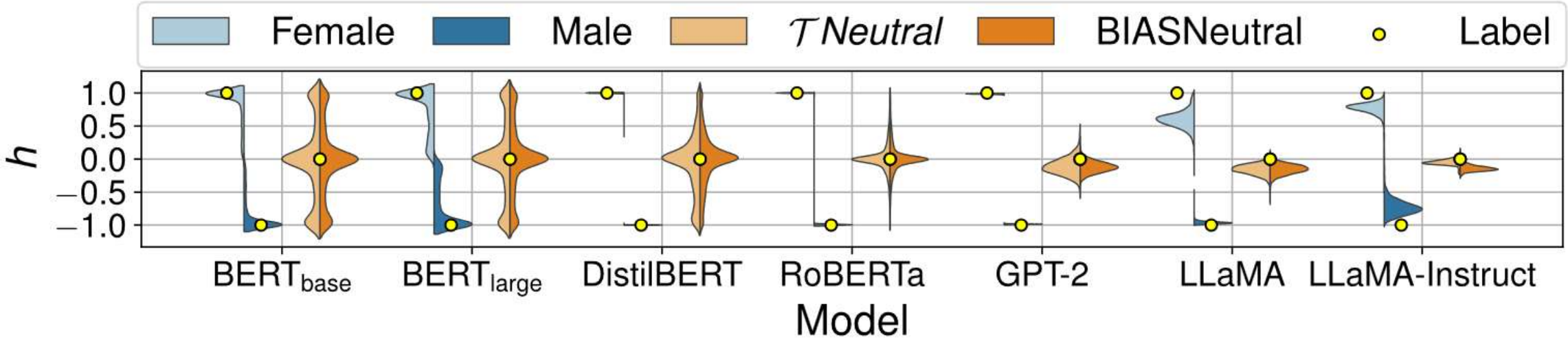
Encoded gender value



$$\begin{cases} > 0 & \rightarrow \text{female} \\ = 0 & \rightarrow \text{neutral} \\ < 0 & \rightarrow \text{male} \end{cases}$$

Gender classes are consistently separated across models

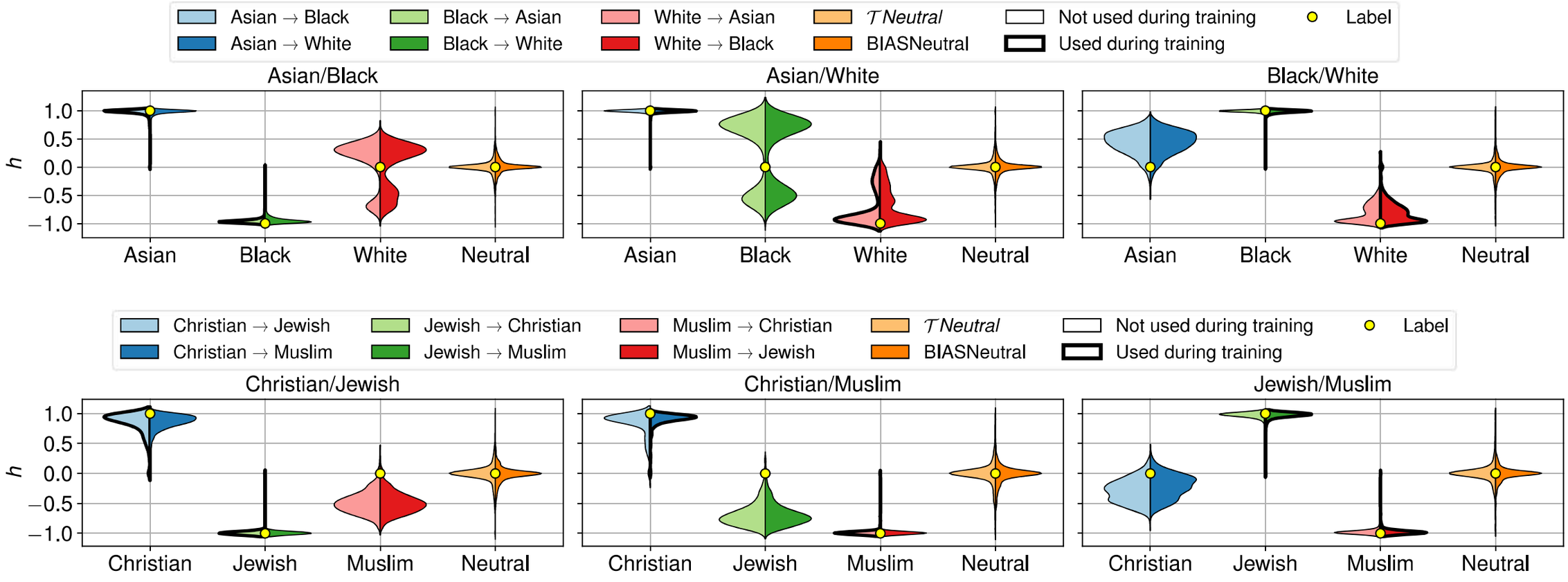
- Separation holds across diverse architectures
- Neutral inputs remain distinct from the feature classes



Race and Religion Extend to Non-Binary Features

UNIVERSITY OF PASSAU
Faculty of Computer Science and Mathematics

GRADIEND is trained for pairs of feature classes

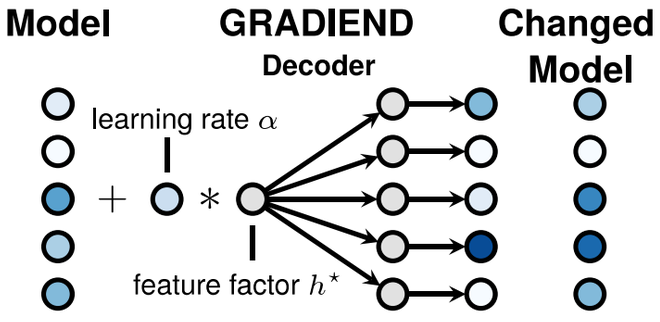


Results shown for BERT base for readability; the same pattern holds across models.

GRADIEND Enables Controlled Model Rewriting

UNIVERSITY OF PASSAU
Faculty of Computer Science and Mathematics

GRADIEND for model rewriting



$$\widetilde{W}_m := W_m + \alpha \cdot dec(h^*)$$

Decoder converts feature signal into a model update

- Update parameters selected via grid search
- Debiasing maximized while preserving language modeling performance

GRADIEND is strongest weight-changing approach for gender debiasing

- Language modeling performance remains close to the base model

Variant			Prop. Rank Bias			Language Modeling (%)		
Name	ΔW	PP	Mean ↑	SS	SEAT	LMS	GLUE	SuperGLUE
GRADIEND + INLP	✓	✓	**0.88**	**0.91**	**0.84**	↓ -0.39 87.06	↓ -0.47 68.23	↓ -1.72 50.65
CDA + INLP	✓	✓	0.75	0.78	0.73	↑ 0.97 86.48	↑ 0.36 77.55	↑ 1.86 52.67
Dropout + INLP	✓	✓	0.71	0.78	0.64	↓ -1.09 84.42	↓ -2.43 74.75	↓ -0.80 50.01
INLP	✗	✓	0.67	0.62	0.72	↑ 0.10 87.56	↑ 0.13 68.83	↓ -0.82 51.55
GRADIEND + SentDebias	✓	✓	0.64	0.67	0.61	↓ -1.12 86.34	↓ -0.92 67.78	↓ -0.83 51.54
Dropout + SentDebias	✓	✓	0.62	0.70	0.55	↓ -3.25 82.27	↓ -2.25 74.93	↓ -0.21 50.60
SentDebias	✗	✓	0.60	0.48	0.72	↓ -0.52 86.94	↓ -0.44 68.27	↓ -0.08 52.29
CDA + SentDebias	✓	✓	0.57	0.71	0.43	↑ 0.01 85.52	↑ 0.50 77.68	↑ 1.25 52.06
GRADIEND	✓	✗	0.46	0.50	0.42	↓ -0.73 86.72	↓ -0.00 68.70	↓ -0.63 51.73
CDA	✓	✗	0.44	0.42	0.45	↑ 0.23 85.74	↑ 0.45 77.64	↑ 1.37 52.18
SelfDebias	✗	✓	0.41	0.41	–	↓ -9.65 77.81	–	–
LEACE	✗	✓	0.36	0.32	0.41	↓ -0.49 86.97	↑ 0.01 68.71	↓ -1.71 50.66
RLACE	✗	✓	0.31	0.21	0.40	↓ -2.19 85.26	↓ -0.06 68.64	↓ -1.85 50.51
Dropout	✓	✗	0.30	0.40	0.20	↓ -2.11 83.40	↓ -3.09 74.10	↓ -0.42 50.39
Base Model	✗	✗	0.17	0.11	0.23	87.46	68.70	52.37

Models are sorted by the Mean column of proportional debiasing ranks. ΔW and PP indicate debiasing variant using weight modification and post-processing, respectively. Best variant type marked with a blue ✓.

UNIVERSITY OF PASSAU
Faculty of Computer Science and Mathematics

GRADIEND: Feature Learning in Neural Networks Exemplified through Biases

GRADIEND learns a targeted feature neuron from model gradients and decodes it into controlled model updates

Demonstrated for social debiasing: gender, race, and religion

https://github.com/aieng-lab/gradiend-bias