ICLR 2026

FlowNIB

An Information Bottleneck Analysis of Bidirectional vs. Unidirectional Language Models

Md Kowsher · Nusrat Jahan Prottasha · Shiyun Xu · Shetu Mohanto
Niloofar Yousefi · Ozlem Garibay · Chen Chen

University of Central Florida · University of Pennsylvania · Delineate Inc.

💻 github.com/Kowsher/BidiVsUniLM

Motivation: Why Do Bidirectional Models Win?

Well-Known Empirical Fact

- **BERT outperforms GPT on ALL GLUE benchmarks at comparable size**
- BERT RTE: 66.4% vs GPT: 56.0%
- Bidirectional models dominate NLU tasks across the board

Even smaller bidirectional models beat larger unidirectional ones

Open Theoretical Question

WHY do bidirectional models preserve more useful information?

No formal information-theoretic framework existed to measure and compare the layer-wise information capacity across architectures.

Our Answer: Information Bottleneck theory + FlowNIB framework

Information Bottleneck Framework

IB Principle: Good representations compress input (small **I(X;Z)**) while preserving task-relevant content (large **I(Z;Y)**)

Unidirectional (→)

Conditions only on past tokens $x_1 \ldots x_t$

Bidirectional (↔)

Conditions on past AND future tokens

Key Inequality

$I(X;Z\leftrightarrow) \geq I(X;Z\rightarrow)$
$I(Z\leftrightarrow;Y) \geq I(Z\rightarrow;Y)$

Theorem: Since conditioning reduces entropy, $H(X|Z\leftrightarrow) \leq H(X|Z\rightarrow) \implies$ **Bidirectional layers carry strictly more mutual information**

FlowNIB: Flow Neural Information Bottleneck

1. **Fine-tune LM:** Fine-tune the language model on dataset with inputs X and targets Y
2. **Cache Activations:** Run one forward pass to cache (X, Y, Zℓ) for every layer ℓ
3. **Train Critics:** Fit two MINE critics per layer on the fixed cache
4. **Dynamic Schedule α(t):** α decays 1→0: early steps focus on I(X;Z), later on I(Z;Y)
5. **Compute OIC:** Select Optimal Information Coordinate summarizing joint capacity

Loss: $\mathcal{L}(t) = -[\ \alpha(t) \cdot \hat{I}(X;Z\ell) + (1-\alpha(t)) \cdot \hat{I}(Z\ell;Y)\ ]$ with $\alpha(t)$ monotonically decreasing from 1 → 0

Optimal Information Coordinate (OIC)

Definition

Each epoch t yields a coordinate:

$x_t = I(t)(X;Z\ell)$ and $y_t = I(t)(Z\ell;Y)$

The OIC is selected at epoch t* where:

$t^*(\gamma) = \text{argmax}[\ \gamma \cdot x_t + (1-\gamma) \cdot y_t\ ]$

Balance weight:

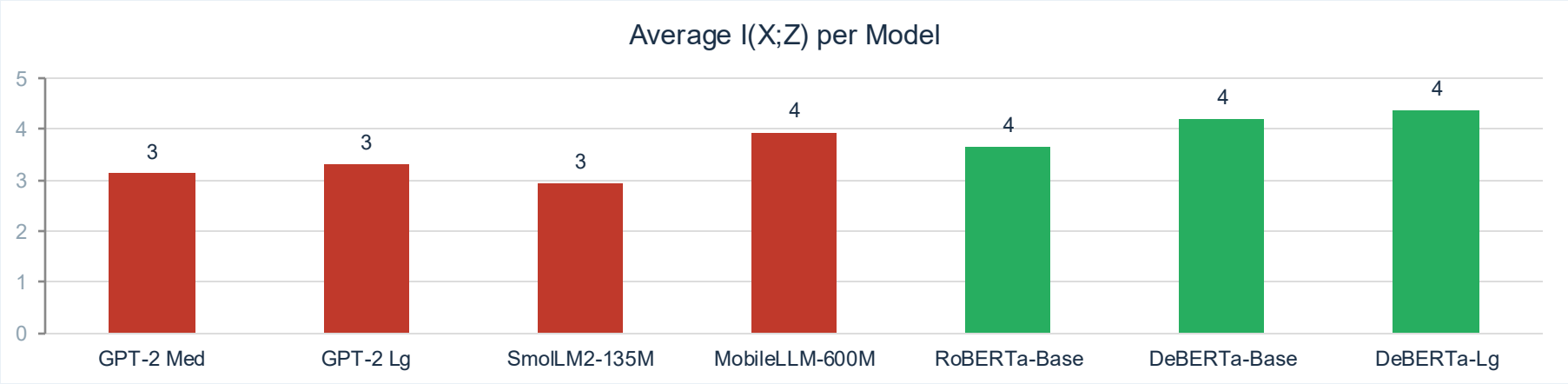
$\gamma\bigstar = Ry / (Rx + Ry)$

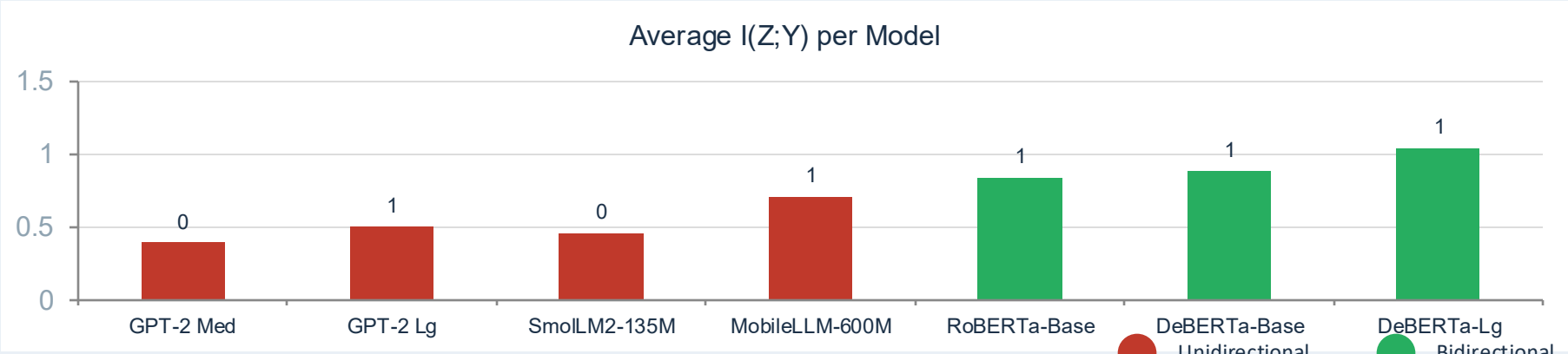
The OIC jointly maximizes information about both the input AND the target.

Why OIC Matters

- ✓ Both MI values come from ONE shared critic — making them directly comparable across layers
- ✓ Eliminates scale artifacts from independently trained critics
- ✓ Higher OIC = richer representation = better downstream accuracy
- ✓ Bidirectional models consistently achieve higher OIC than unidirectional models

Mutual Information Results: Bidirectional ≫ Unidirectional





Downstream Task Performance: Classification

Model	Method	SST-2	MRPC	RTE	CoLA	HellaSwag	Avg.
▶ BIDIRECTIONAL							
DeBERTa-v3-Lg	Masking	96.1	94.0	89.9	93.0	73.4	**84.7**
RoBERTa-Lg	Masking	96.2	91.3	87.8	95.8	71.4	**84.0**
DeBERTa-v3-Base	Masking	95.8	91.2	85.0	87.4	69.5	**81.5**
▶ UNIDIRECTIONAL							
MobileLLM-600M	Generation	95.1	87.9	72.3	86.3	48.5	76.6
GPT-2 Large	Generation	94.2	87.2	67.3	83.9	39.5	72.1
SmolLM2-360M	Generation	94.7	85.3	71.1	84.5	43.7	74.4

Key Finding: DeBERTa-v3-Base (184M params) outperforms MobileLLM-600M (600M params) by ~5% average accuracy — smaller bidirectional beats larger unidirectional!

Computational Efficiency: Bidirectional is Faster Too

RoBERTa-Base

125M — Bidirectional

2.11s

per step

MobileLLM-125M

125M — Unidirectional

3.83s

per step

GPT-2 Medium

345M — Unidirectional

6.04s

per step

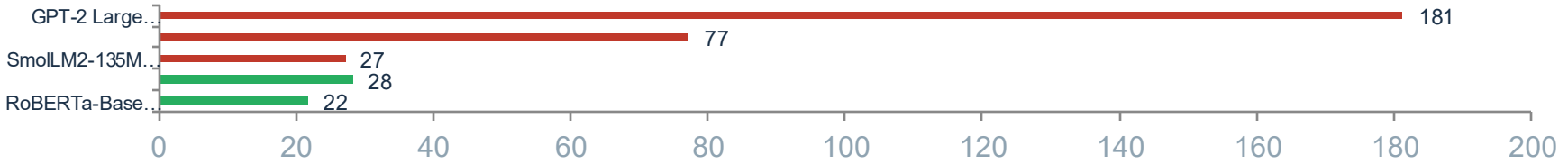
ModernBERT-Base

149M — Bidirectional

1.15s

per step

GFLOPs Comparison (single forward pass)



Ablation Study & Key Findings

Step Size δ Controls IB Trade-off

$\delta = 10^{-3}$ to 10^{-4} achieves best balance: gradually reduces I(X;Z) while increasing I(Z;Y) toward the IB frontier

Higher Effective Dimensionality

Bidirectional models show higher deff(Z) at every layer (DeBERTa-Lg: 8.73→1.98 vs MobileLLM-600M: 5.38→1.44)

OIC Predicts Task Performance

Models and layers with higher OIC consistently achieve higher downstream accuracy — MI is a strong performance proxy

Extends Beyond NLU

On ETTh1/ETTh2 time-series forecasting: bidirectional attention achieves lower MSE and higher I(Zℓ;Y) at all horizons

Layer-wise Linear Probing confirms: bidirectional representations are more linearly decodable at every depth, including early layers.

Contributions & Conclusion

01

Theoretical Foundation

Formal proof that bidirectional representations carry ≥ mutual information about input AND output versus unidirectional ones

02

FlowNIB Framework

Lightweight post-hoc framework with curriculum schedule $\alpha(t)$ that jointly estimates $I(X;Z\ell)$ and $I(Z\ell;Y)$ comparably across all layers

03

Optimal Information Coordinate

Novel diagnostic metric (OIC) strongly correlated with downstream accuracy — higher OIC = better performance

04

Empirical Validation

15 diverse datasets (GLUE, commonsense, regression, time-series): smaller bidirectional models outperform larger unidirectional ones