Invariant Language Models

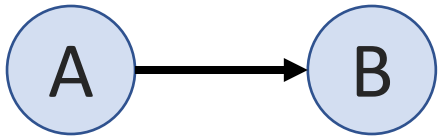
Maxime Peyrard

Causality

Correlation ≠ Causation

Causation Changes in A create changes in B

"B listens to A to take its value"

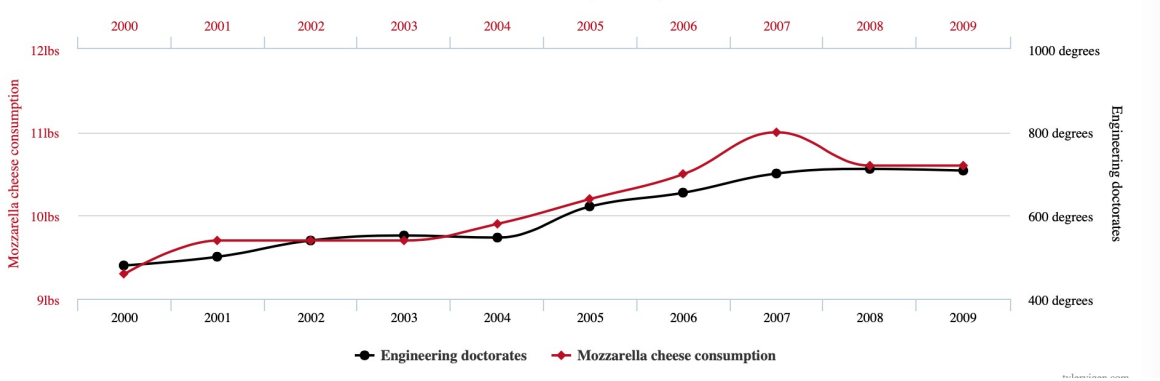


Smoking → Cancer

Correlation A and B tend to vary together

Per capita consumption of mozzarella cheese
correlates with
Civil engineering doctorates awarded

Correlation: 95.86% (r=0.958648)

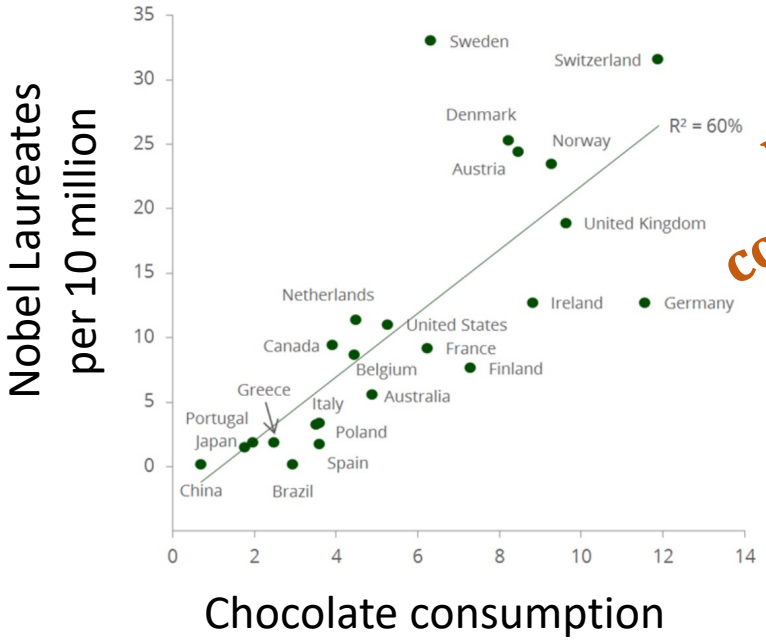


*When a correlation is **not** due to a causal relation, we say it is **spurious***

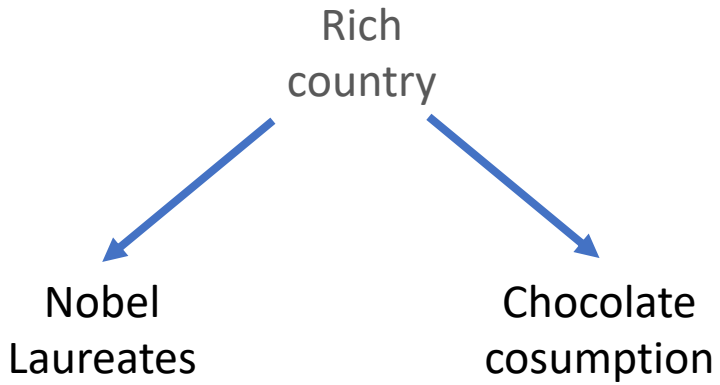
Typical Problems with Correlation

Hidden confounder

Confounding bias

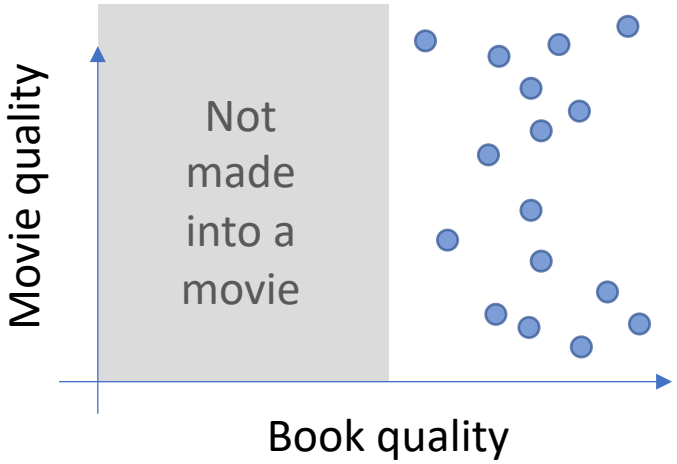


High correlation

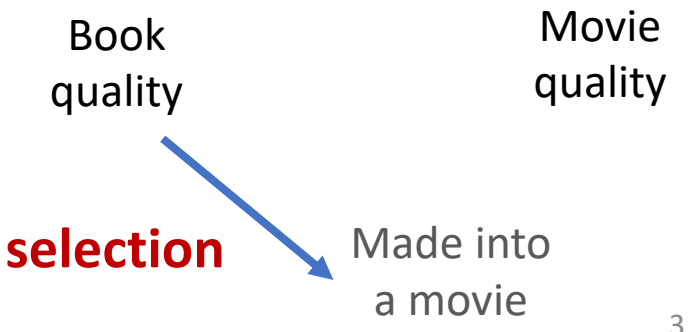


Selection bias

"The book is better than the movie"



High negative correlation



Machine Learning ignores Causation

Machine Learning

Features: X
Target: Y

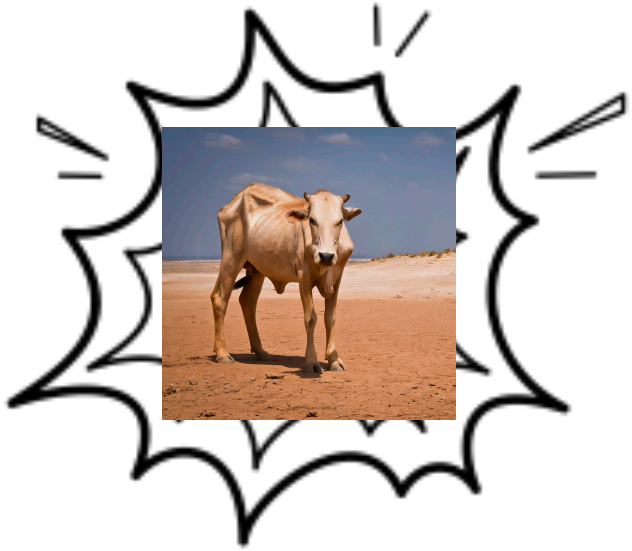
Purely statistical properties of $\mathbb{E}(Y|X)$ or $\mathbb{P}(Y|X)$

Problems

- Purely based on correlations
- Suffer from and amplify biases in the data
- Poor generalization: spuriousness

Problems

Spurious correlation → poor generalization



Sandy background has a **strong correlation** with *Camel*

Adding a sandy background to a cow **does not make it** a camel

Problems: bias

3 November 2021

Our computers are sexist towards male and female politicians

TECHNOLOGY & BIAS Female politicians are often described as 'beautiful' or 'hysterical' in language models, while descriptors like 'brave' and 'independent' are elicited for their male counterparts. A new analysis by researchers from the University of Copenhagen makes this problem abundantly clear. According to the researcher behind the study, sexist and stereotypical descriptions of gender are a problem for gender equality.

PUBLISHED ON JANUARY 11, 2022 IN OPINIONS

How bias creeps into large language models

According to DeepMind, unmodified LMs tend to assign high probabilities to exclusionary, biased, toxic, or sensitive utterances if such language is present in the training data.

Aug 30, 2019, 09:40am EDT | 10,443 views

Massive Machine Learning Study Demonstrates Gender Stereotyping And Sexist Language In Literature

Jesse Damiani Contributor

Consumer Tech

I run Postreality Labs, a new media art advisory & curatorial studio.

Ethical and social risks of harm from Language Models

NEWSLETTERS • EYE ON A.I.

Your favorite A.I. language tool is toxic

BY **JONATHAN VANIAN**

September 29, 2020 5:25 PM GMT+2

Goal: Remove **spurious correlations**

Bias mitigation

Observe a correlation to be **spurious**

↓

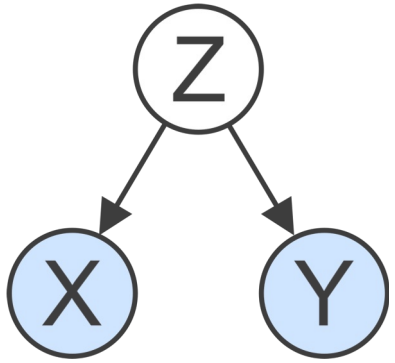
Explicitly (manually) remove it from the **model (or data)**

$$\min_{\theta} \mathcal{L}(\theta) + Bias(\theta)$$

Works but never-ending

Causal Inference/Reasoning

- Start with **variables**
- Infer the **causal graph**
- Use it to recognize **spurious correlations**



Very hard but fundamental!

Invariance: Invariance across Environments

Causal (non-spurious) correlations *are stable across environments!*

Cows in Switzerland

Cows in Corsica

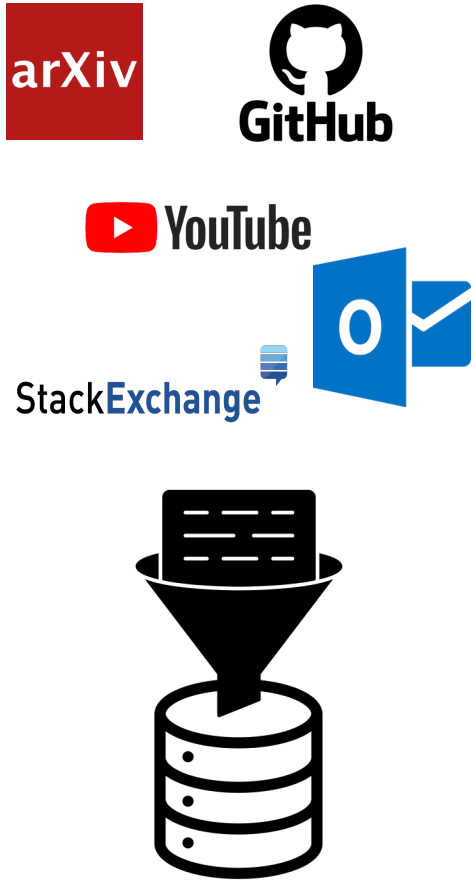
Grassy background is spuriously correlated to cows

`Is_Cow|Background` is **not stable** across environments

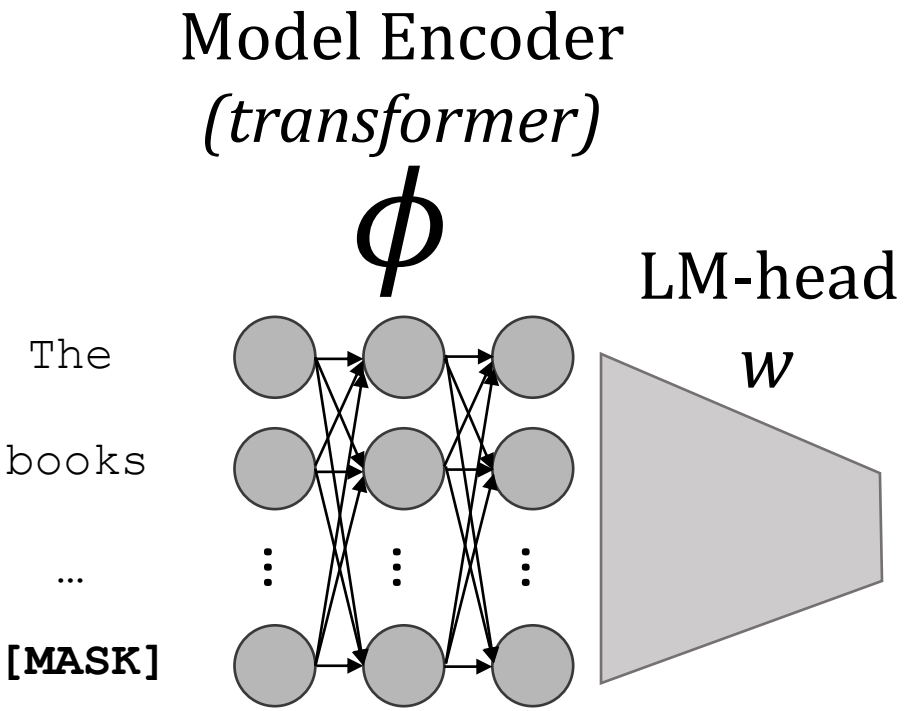
`Is_Cow|Cow_Features` is **stable** across environments

Language Modeling

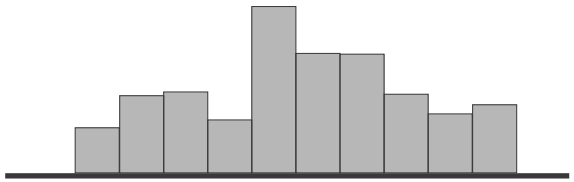
Model the probability of sentence continuation, i.e., probability of next token



Use web text data



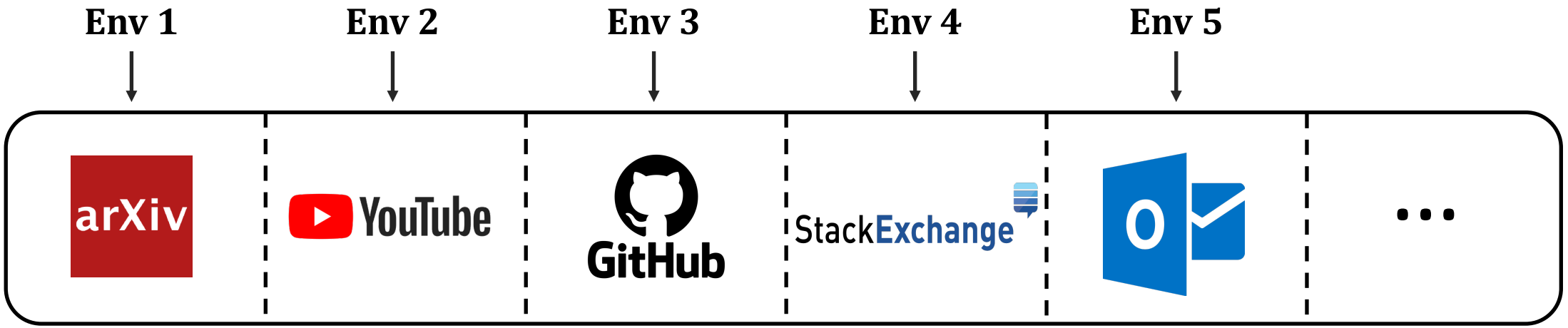
$w \circ \phi(X)$



Probability of next token

Invariant Language Modeling

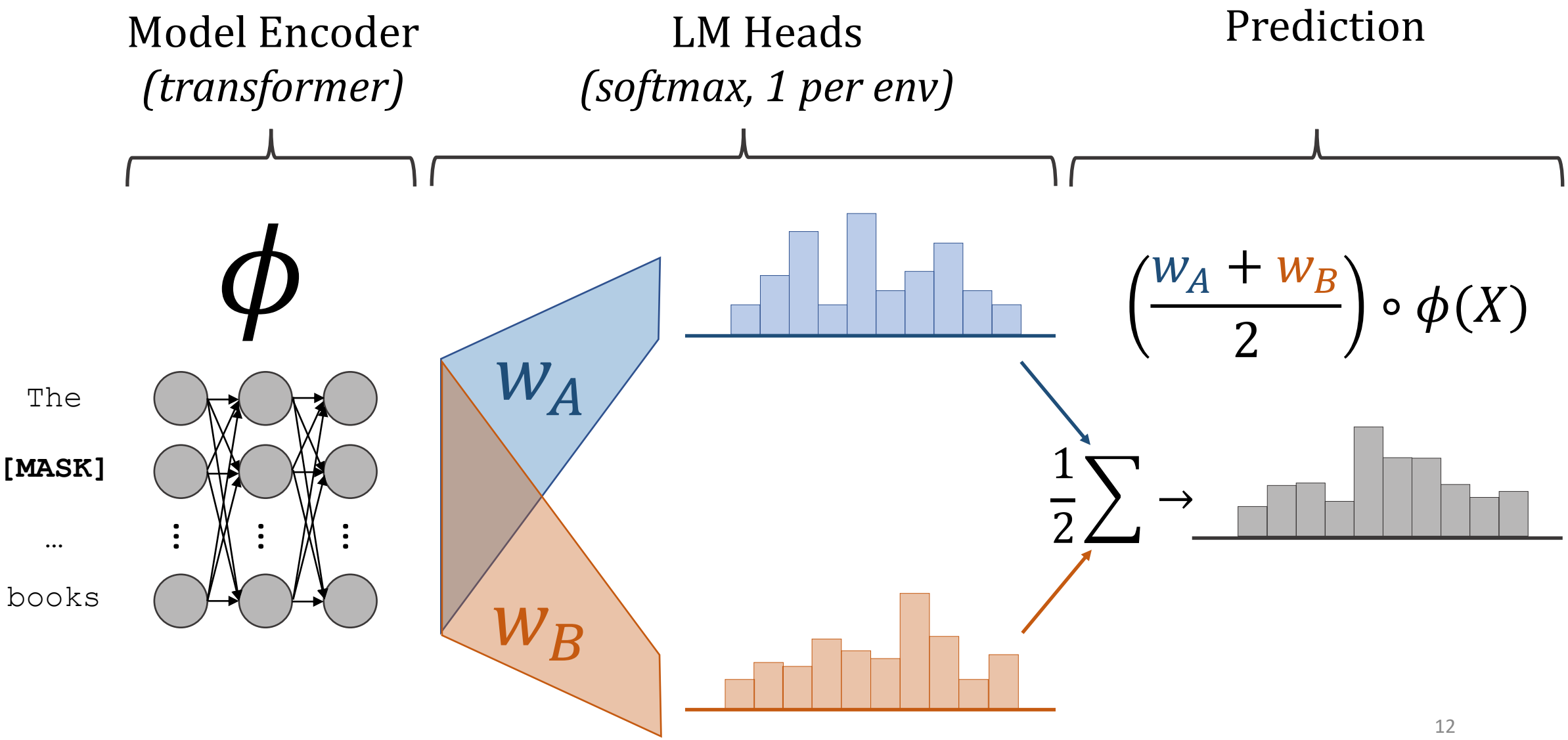
How to lift this very successful paradigm into the world of causality?



Textual data naturally comes in environments!

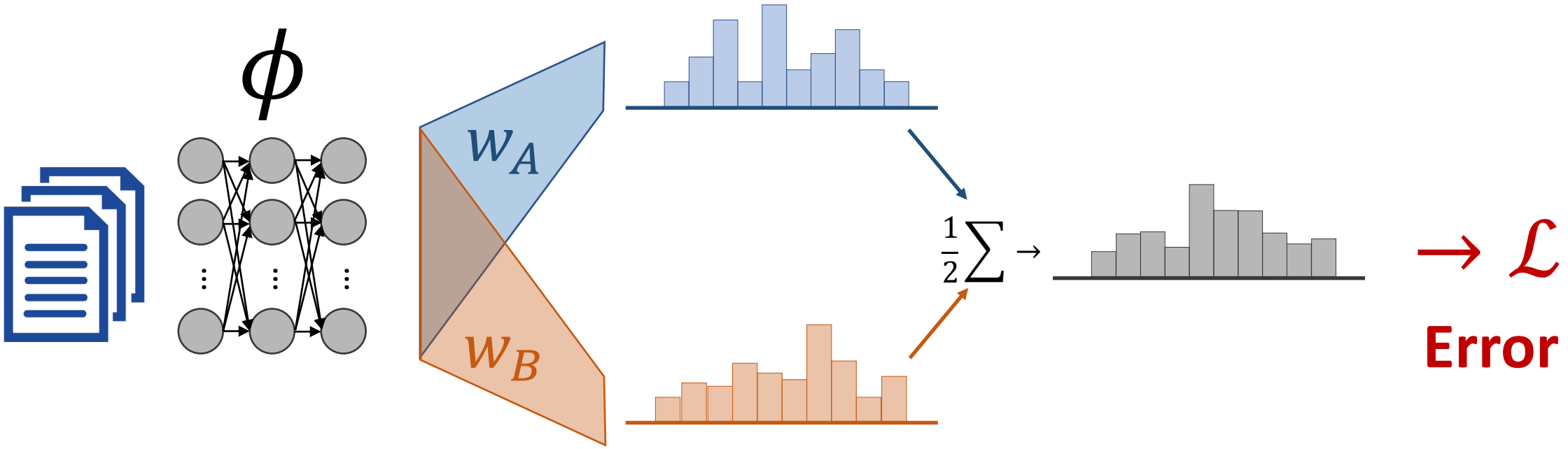
Use it to train/finetune more robust and stable LMs

Implementing IRM-LM: Architecture



Competitive Gradient Updates

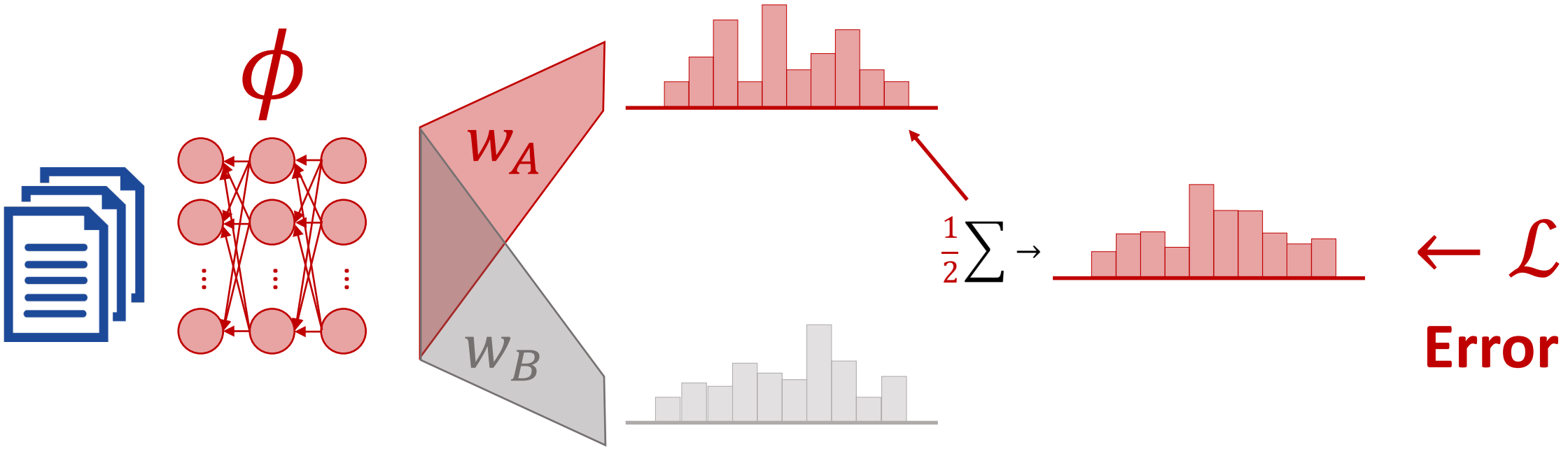
Batch from environment A



Competitive Gradient Updates

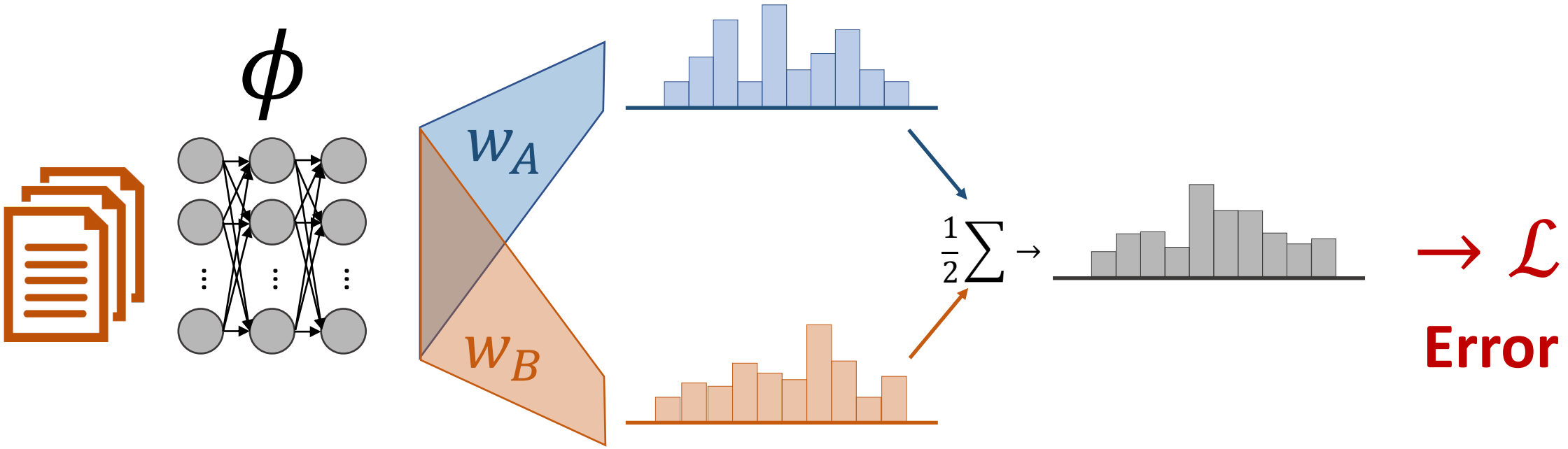
Batch from environment A

Backpropagate only through w_A and ϕ



Competitive Gradient Updates

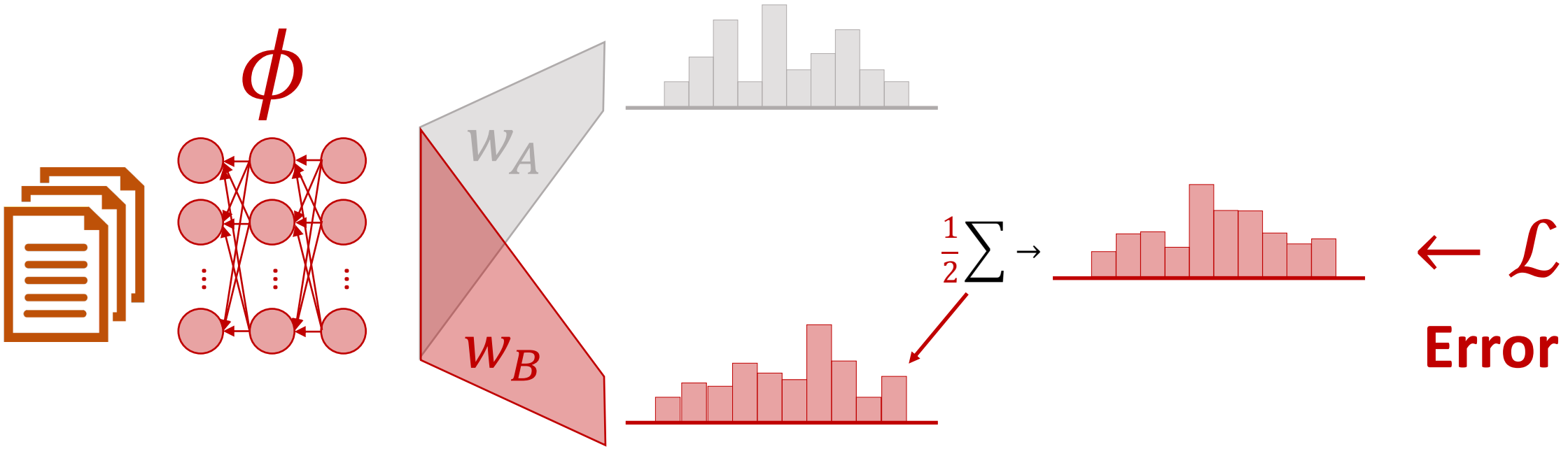
Batch from environment B



Competitive Gradient Updates

Batch from environment B

Backpropagate only through w_B and ϕ

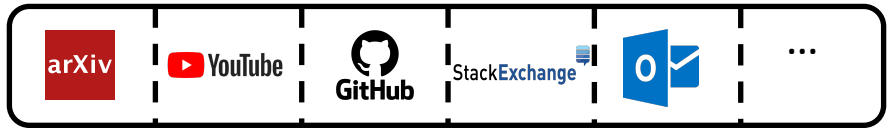


Testing

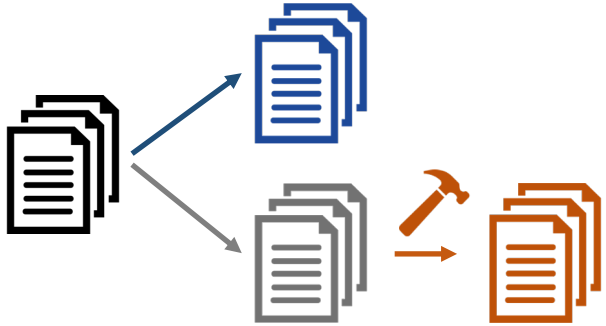
Compare standard **LM** and **InvariantLM**:
can we effectively ignore spurious correlations?

Craft **environments** with minimal and controlled changes

Out-of-domain generalization



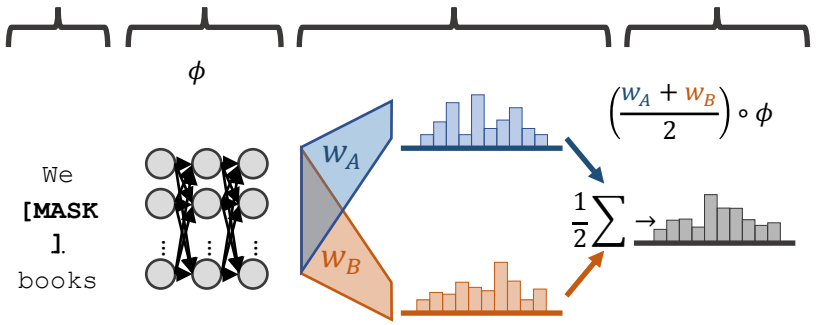
Controlled correlation removal



Structured noise

Conclusion

- More **robust** to noise
- SotA **bias mitigation** technique
- Better **OoD generalization**



- Code on **HuggingFace**
- No need to modify existing LMs
- Minimal computational overhead

Follow-up : *How to best choose environment splits in general?*

Thank you!
Questions?

Contact: maxime.peyrard@univ-grenoble-alpes.fr

Backup

Structured Noise Removal

Env A: Wikipedia texts

Env B: Full HTML pages

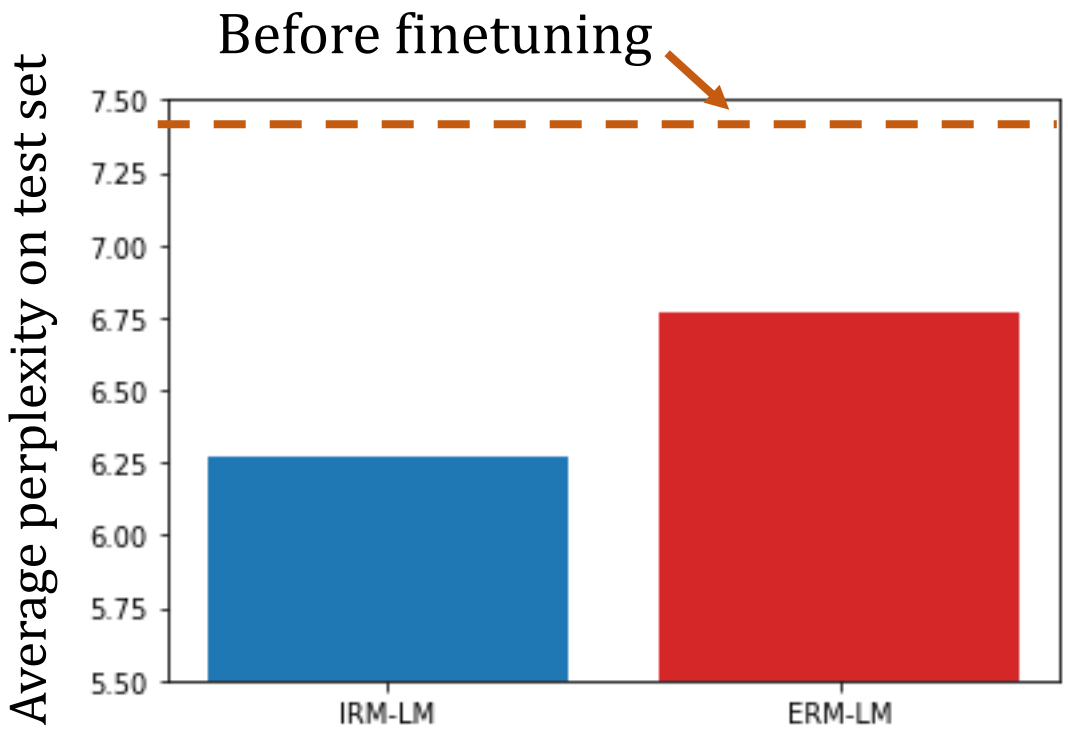
Testing:

Perplexity on clean held-out Wikipedia texts.

Structured Noise Removal: Preliminary results

Model: RoBERTa

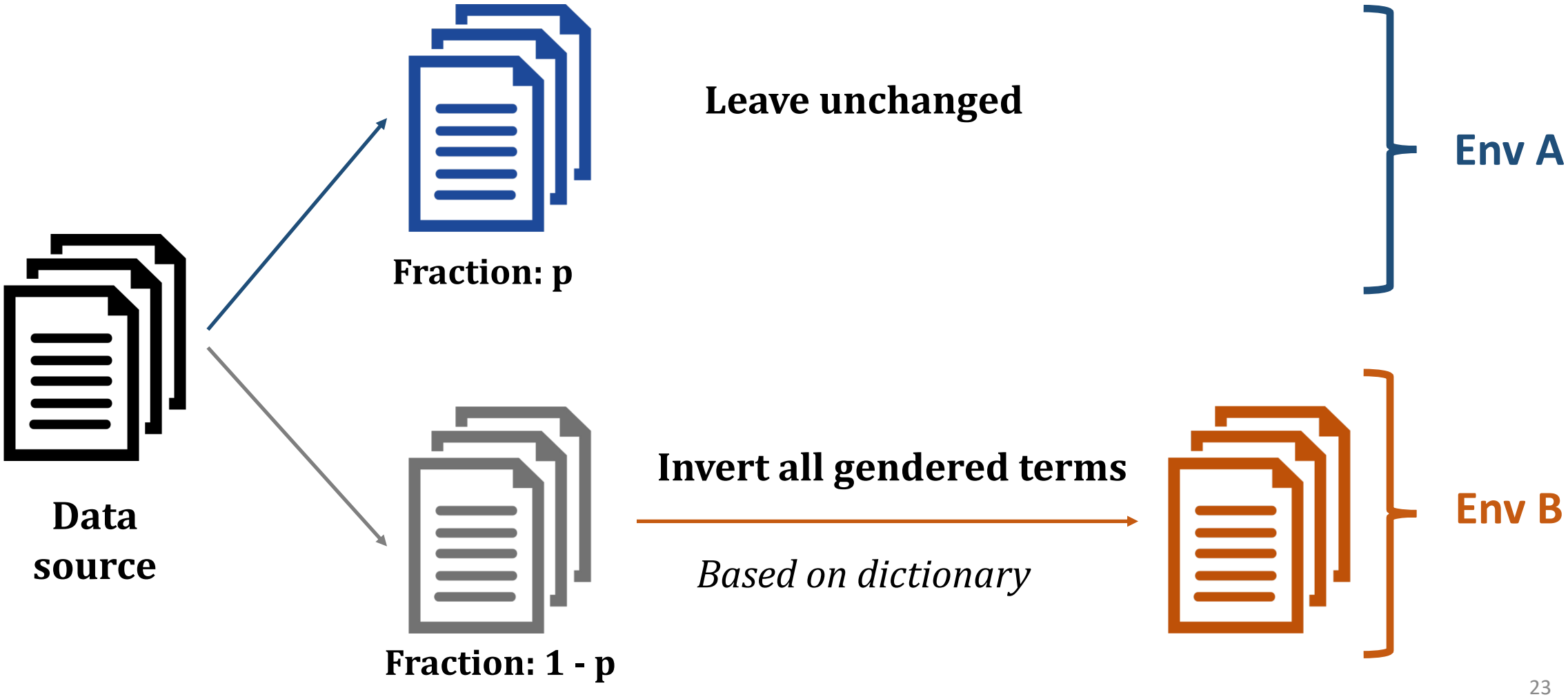
Data: Subsample of Wikipedia dump (08.2020), 10K tokens



IRM better for (almost) all parameter choices

Experiment 2

Controlled Correlation Removal: *Gender bias*



Experiment 2

Controlled Correlation Removal

Two setups:

Balanced: |EnvA| = |EnvB|

Un-balanced: |EnvA| = 4 |EnvB|

Intuition:

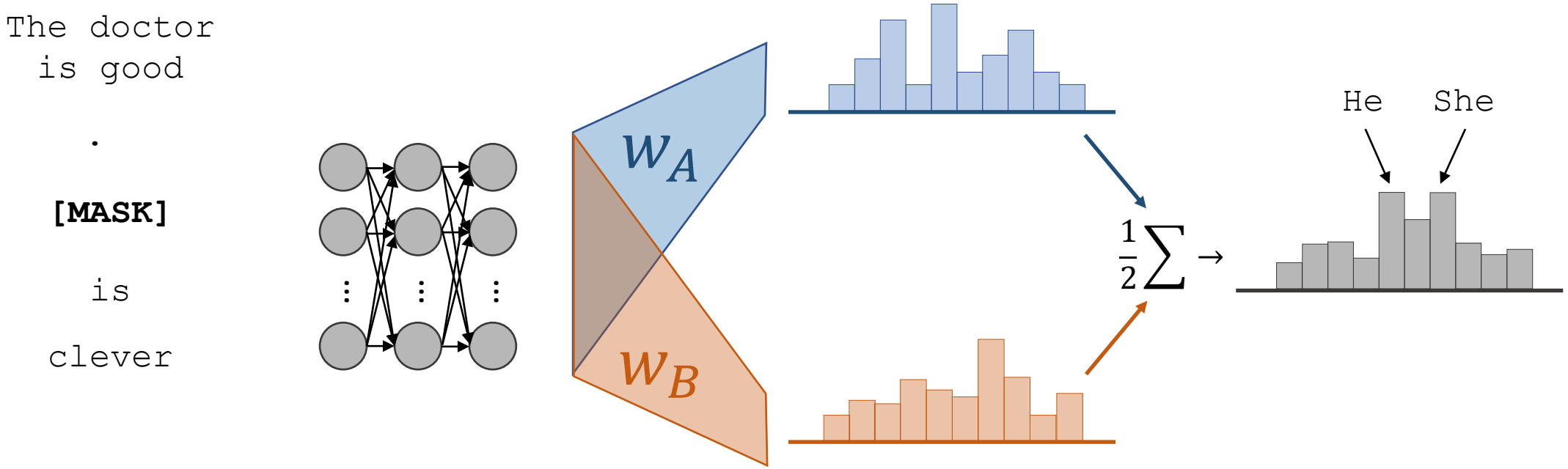
Balanced setup: Both IRM and ERM should remove gender correlations

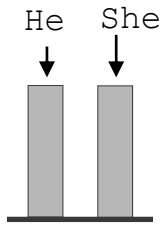
Unbalanced setup: ERM would fail!

Experiment 2

Controlled Correlation Removal

Measure: Inspect the softmax of the model for gendered term in the test set



Entropy (

)

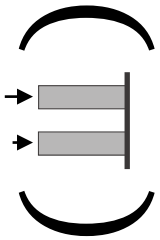
Should be high because the intervention destroyed gender-based correlations

Experiment 2

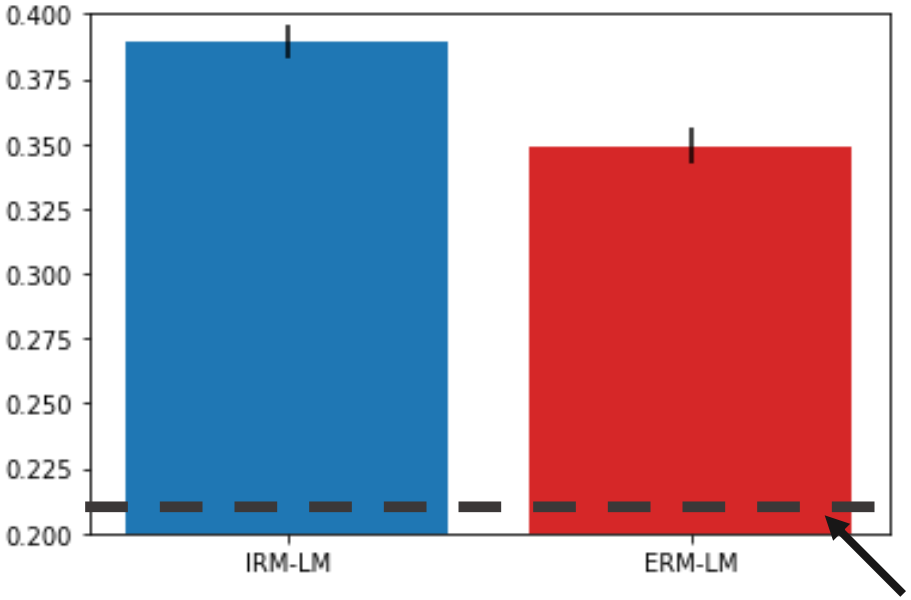
Controlled Correlation Removal: Results

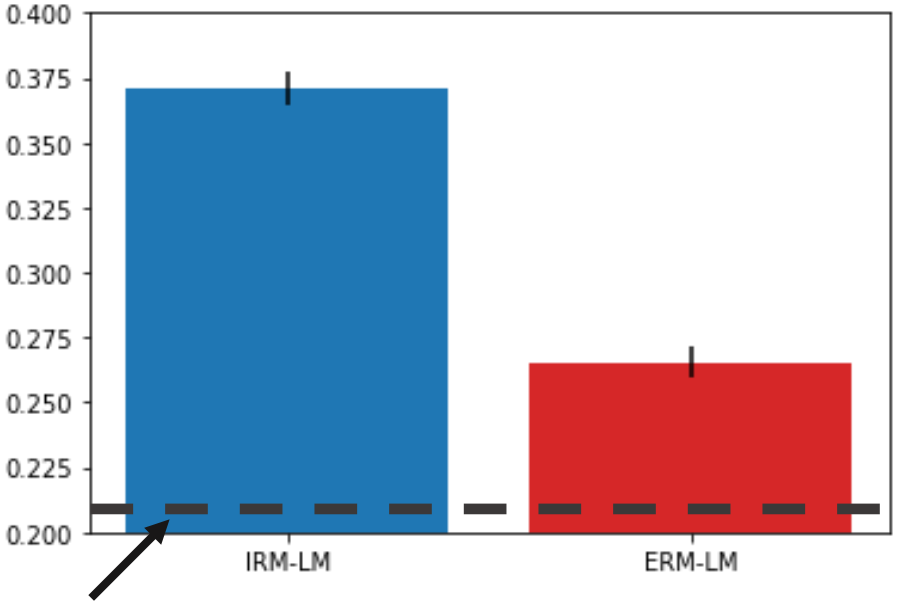
Balanced:
|EnvA| = |EnvB|

Un-balanced:
|EnvA| = 4 |EnvB|



Avg Entropy

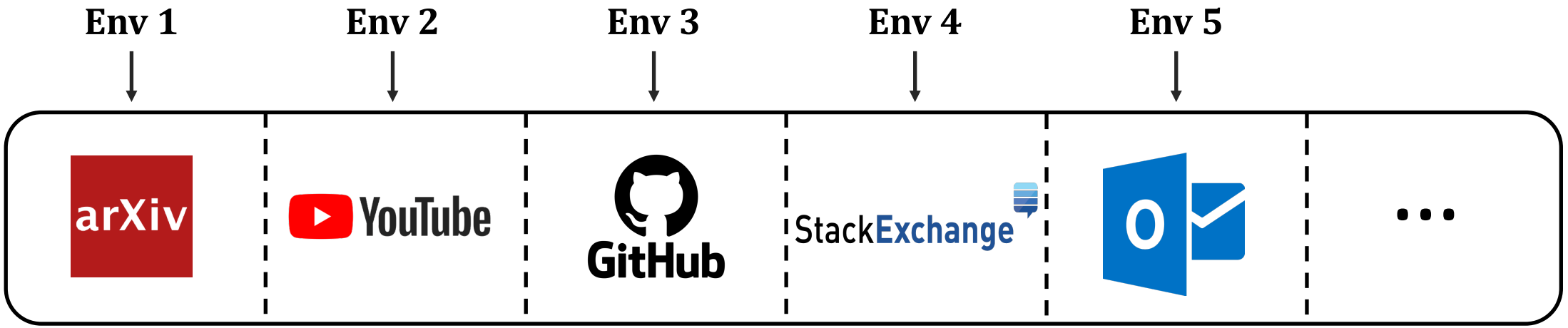




Before finetuning

Experiment 3

Out-of-Domain Generalization



The Pile (https://pile.eleuther.ai):

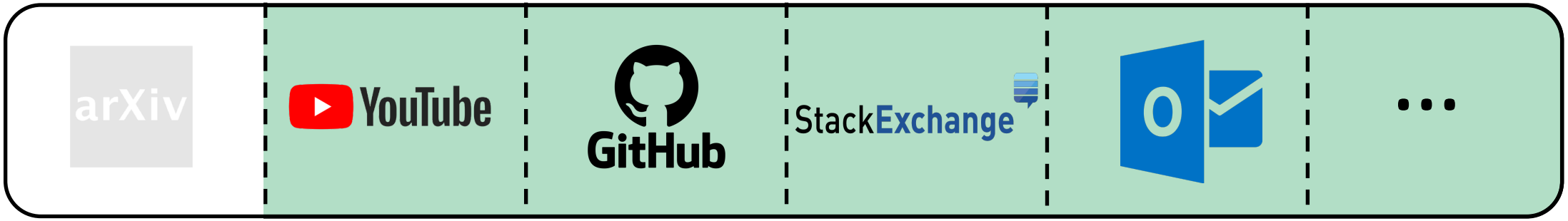
22 such domains, 800 GB of data

Experiment 3

Out-of-Domain Generalization

Leave-One-Domain-Out: *train on n-1 environments, measure perplexity on the other*

Train:



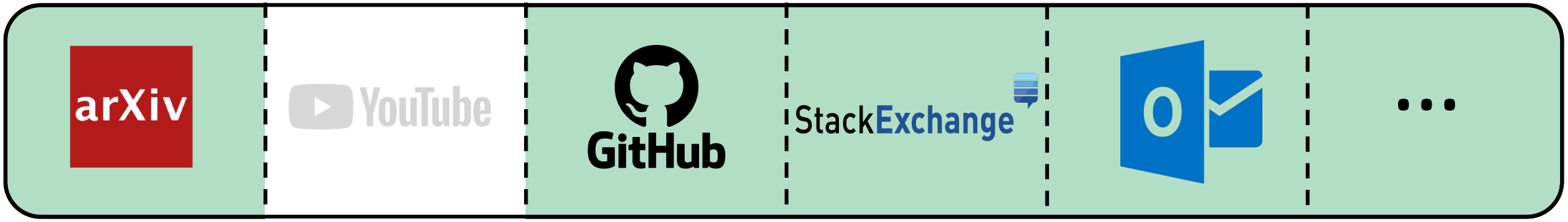
Test: arXiv *-Perplexity*

Experiment 3

Out-of-Domain Generalization

Leave-One-Domain-Out: *train on n-1 environments, measure perplexity on the other*

Train:

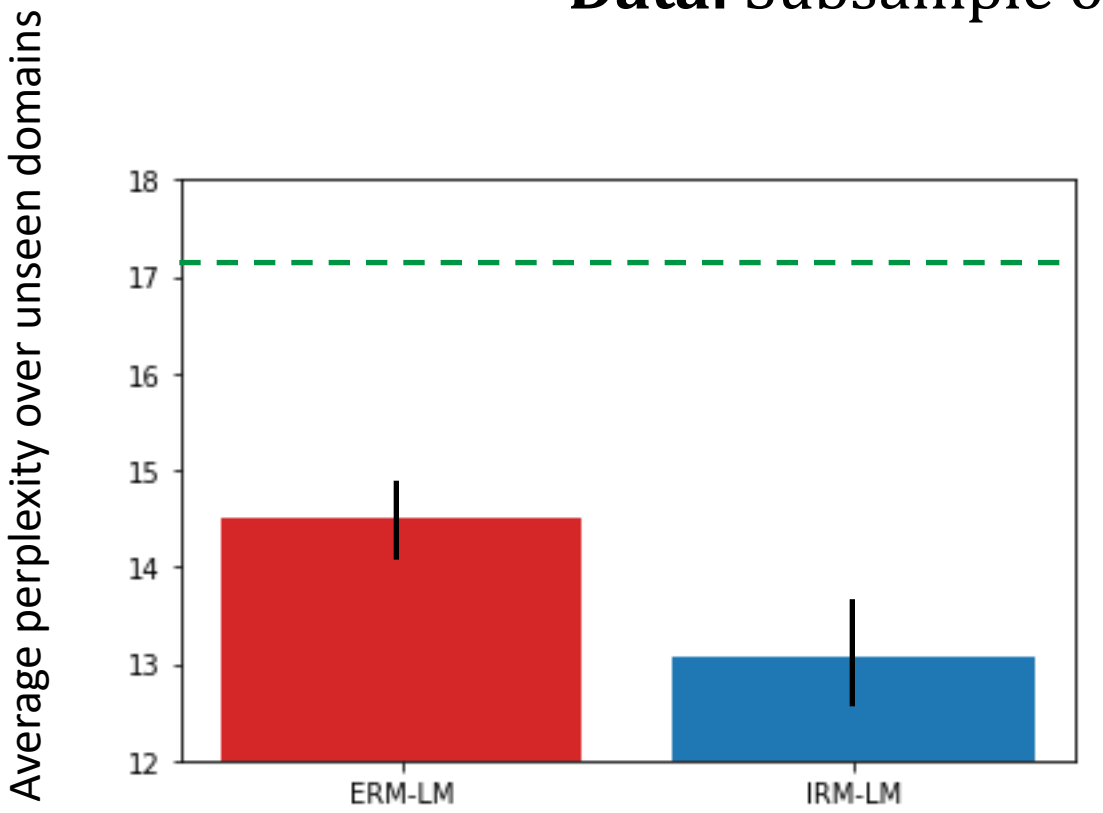


Test: YouTube *-Perplexity*

Experiment 1

Out-of-Domain Generalization: Results

Model: RoBERTa
Data: Subsample of the Pile, 10K tokens



Before finetuning: 17.2

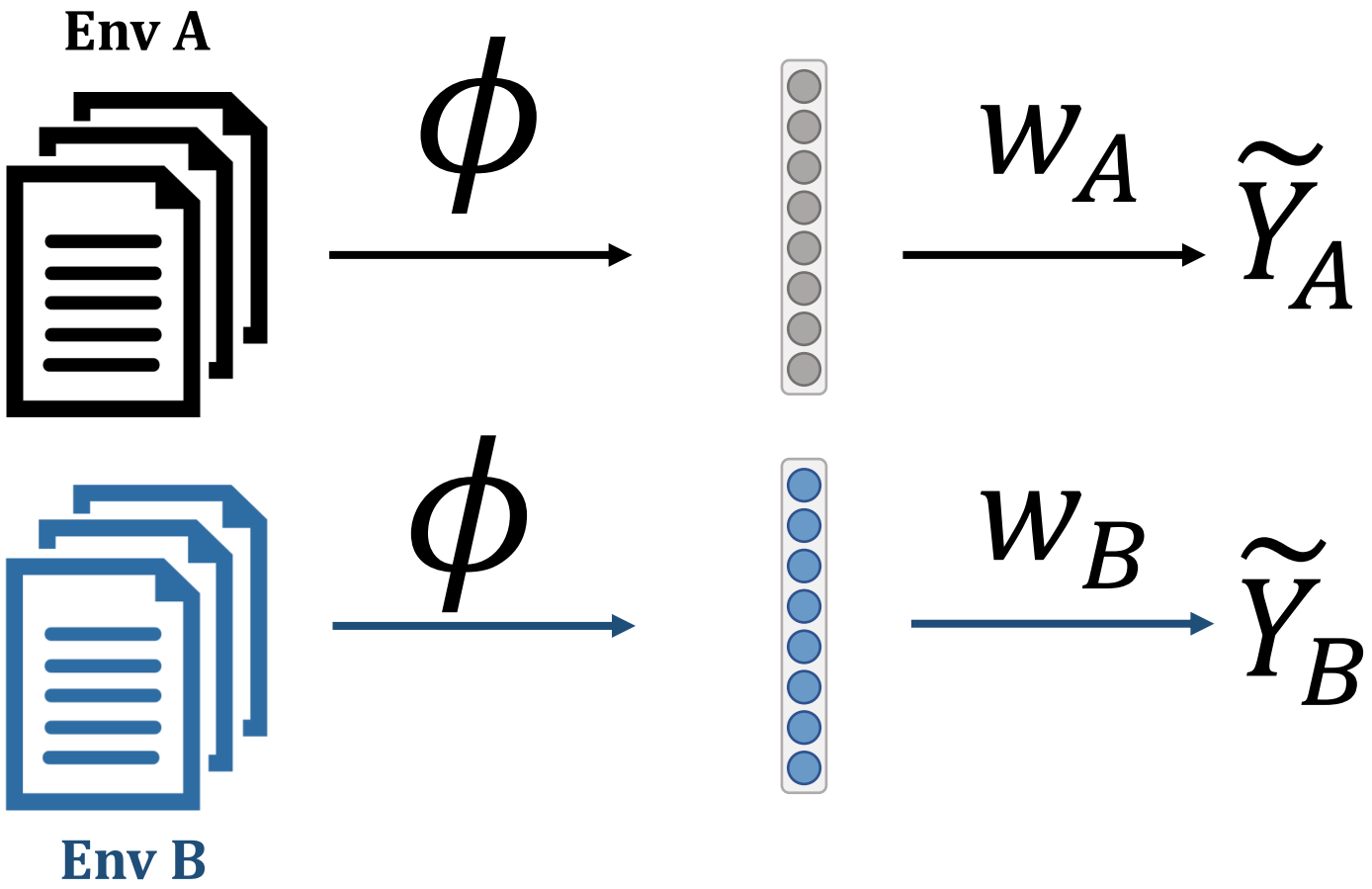
IRM-LM has better out-of-domain perplexity

Intuitive explanation:
Because training environments are very different IRM learns to be conservative and do not overfit

Invariance: Invariant Risk Minimization (IRM)

Arjovsky *et al. 2019*

Alternative to Empirical Risk Minimization that uses environments to enforce ***IaE***



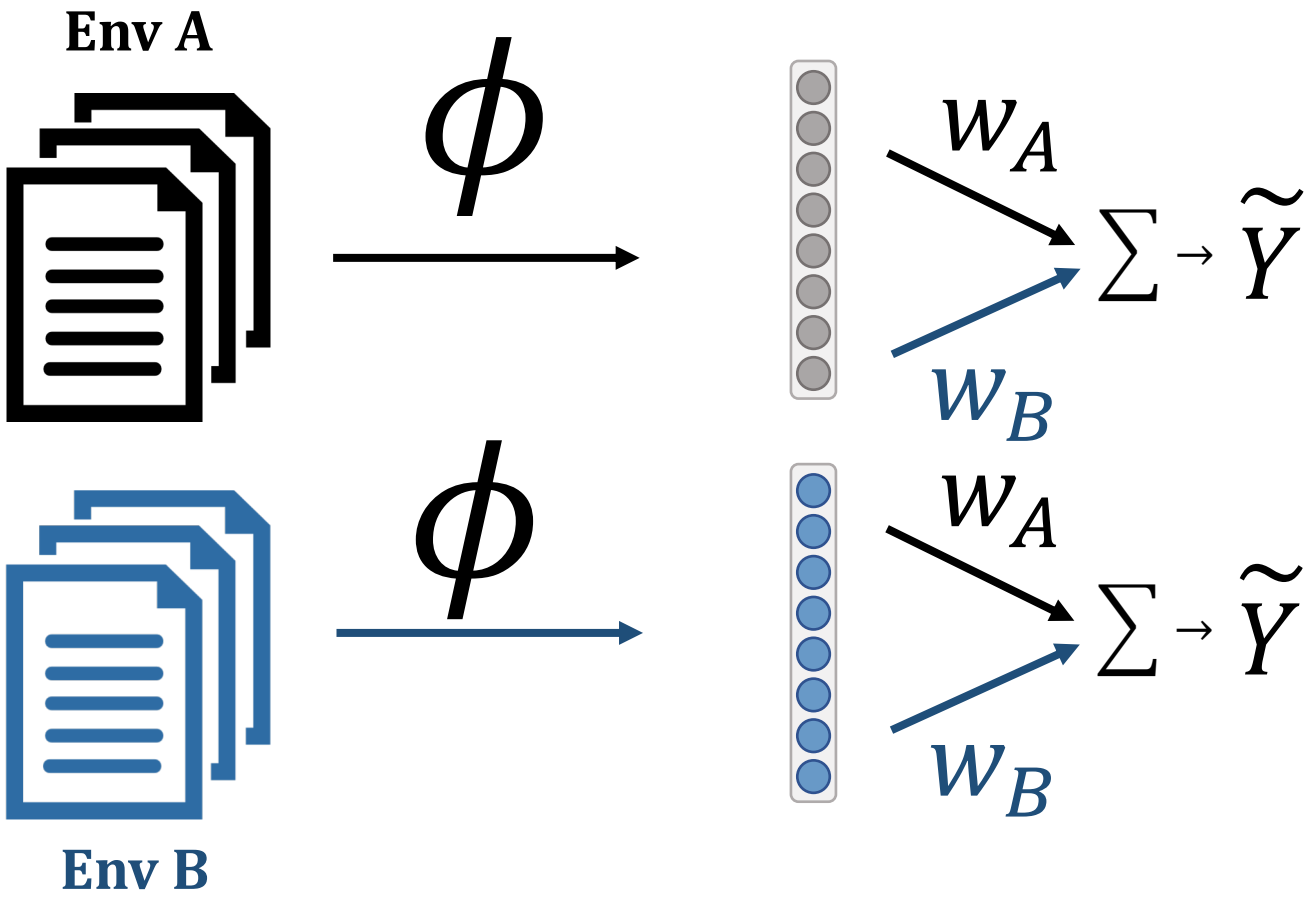
Goal: Find $\boldsymbol{\phi}$ s.t. each environment has the same optimal classifier $w_A = w_B$

Intuition: $\boldsymbol{\phi}$ has to extract invariant features: useful for every environment

Invariance: IRM-Games

Ahuja *et al. 2020*

Convenient implementation of IRM



- Keep one classifier per environment (w_A, w_B)
- Average their predictions

$$\left(\frac{w_A + w_B}{2}\right) \circ \phi(X)$$

- On batch from **env A** only update $\boldsymbol{w_A}$ and ϕ
- On batch from **env B** only update $\boldsymbol{w_B}$ and ϕ

Invariant Models

Default machine learning

$$\min_{\theta} \mathcal{L}oss(\, f_\theta(X), Y \,)$$

Different environments

ϕ_θ is learned to be the invariant representation

$$\phi_\theta(X^1) \qquad \phi_\theta(X^2)$$

"There is a classifier w that is optimal for all environments simultaneously"

$$w \circ \phi_\theta(X^1) \qquad w \circ \phi_\theta(X^2)$$

w is the best choice to minimize both

$$\mathcal{L}(\, w(\phi_\theta(X^1)), Y^1 \,) \qquad \mathcal{L}(\, w(\phi_\theta(X^2)), Y^2 \,)$$

Implementing Invariant LMs

Adapt IRM-games to Language Models (LM)

Environments

LM - Encoder

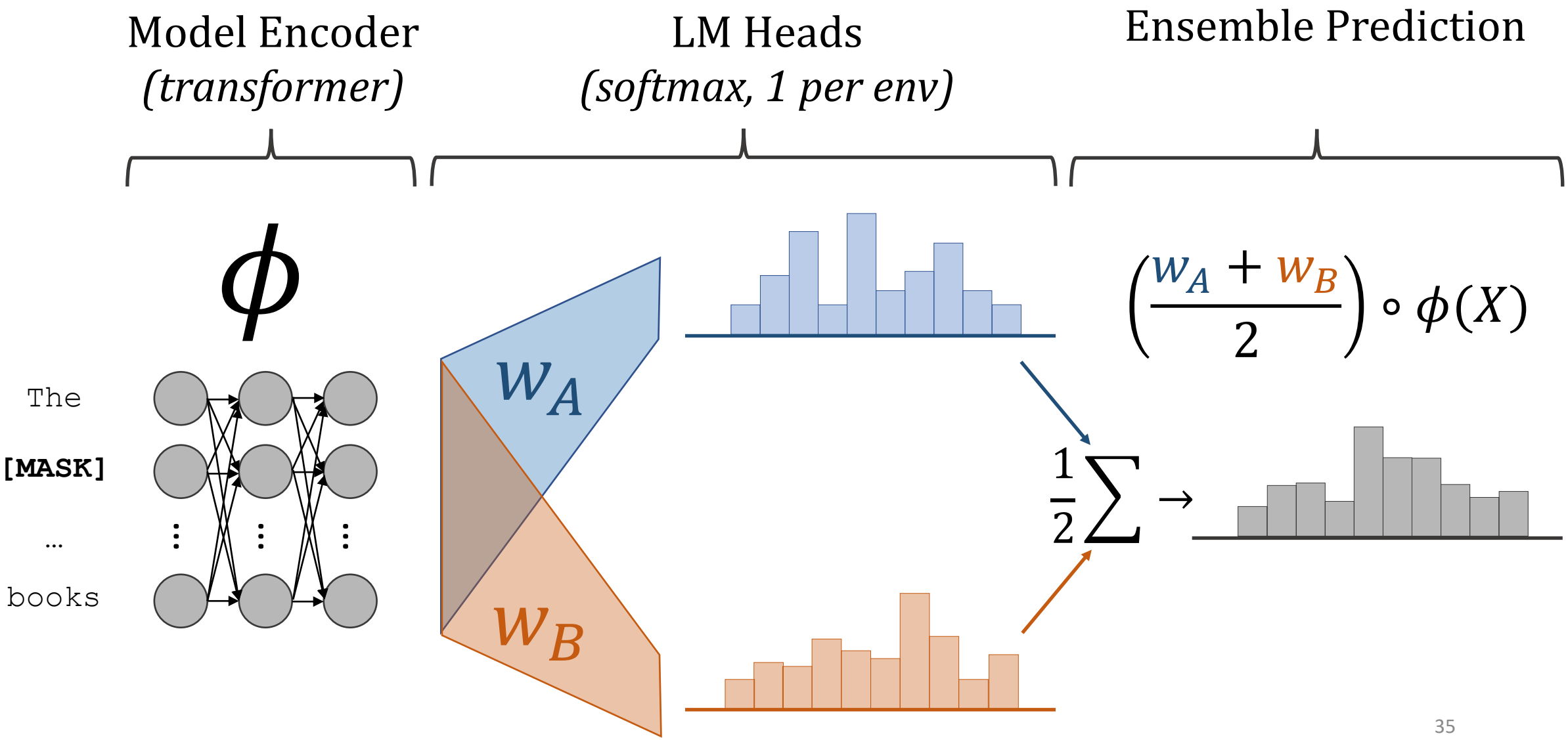
ϕ

LM - Heads

w_A

w_B

Ensemble Prediction

$$\left(\frac{w_A + w_B}{2}\right) \circ \phi(X)$$

Implementing Invariant LMs