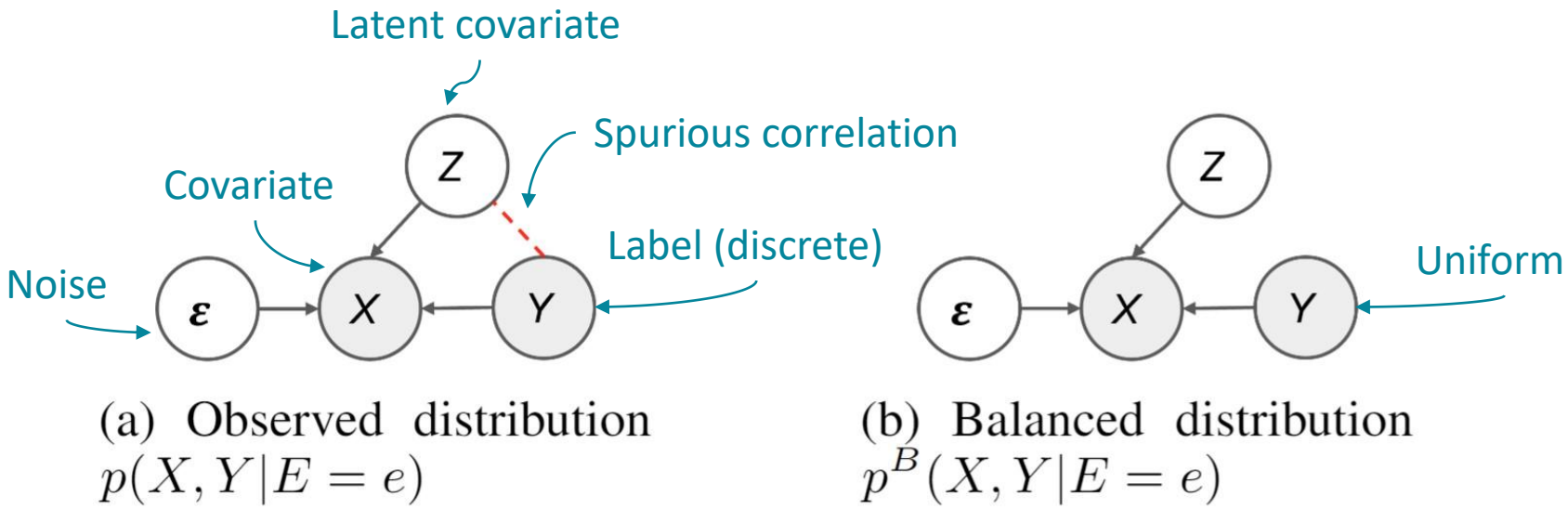
Causal Inference Aided Deep Learning

Xinyi Wang

Causal Balancing for Domain Generalization

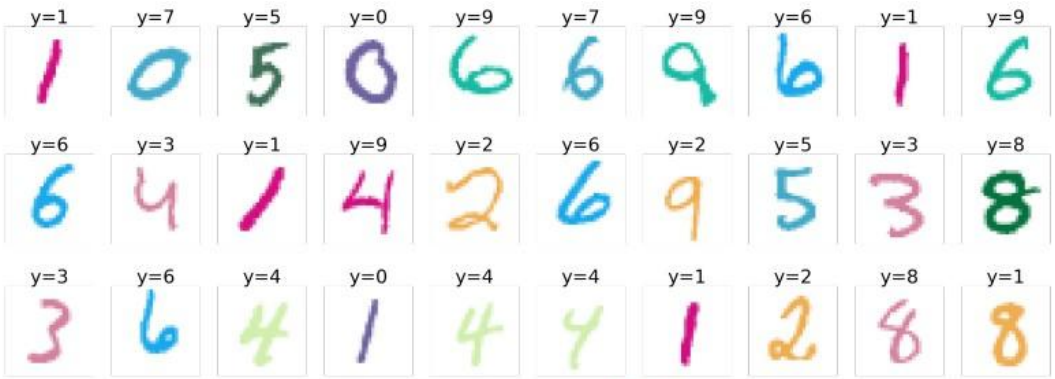


(a) Observed distribution $p(X, Y|E = e)$

(b) Balanced distribution $p^B(X, Y|E = e)$

- Spurious correlation $p(Z|Y)$ changes in different environments.
- Bayes optimal classifier trained on a balanced distribution is **minimax optimal** across all environment.
- Use an VAE to learn the observed distribution (shown to be identifiable).
- Then sample from the balanced distribution by construction balanced mini-batches.
- Train a robust classifier using the balanced mini-batches.

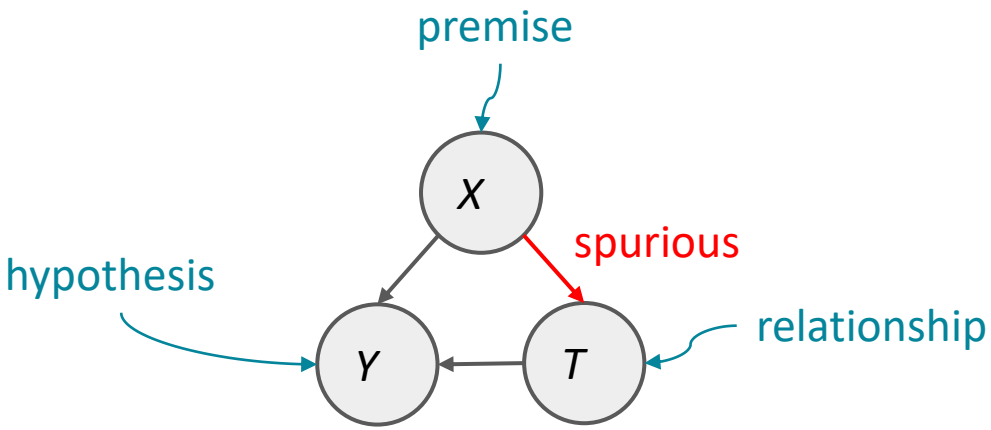
Balanced mini-batch



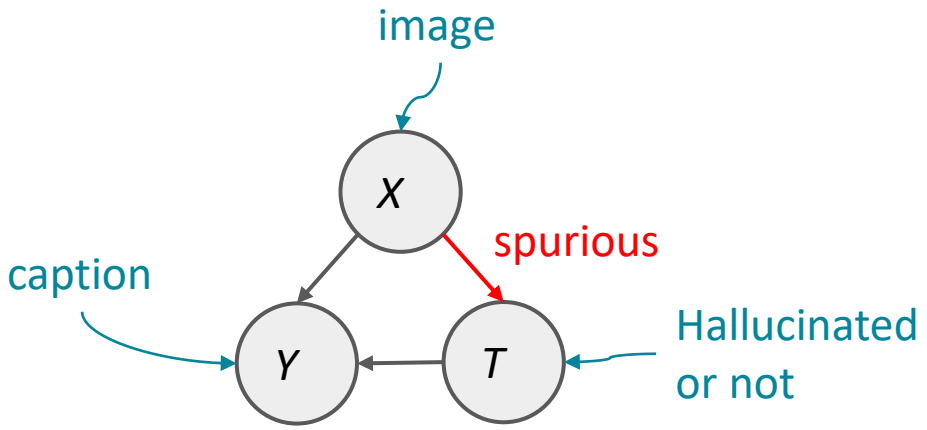
(a) A random mini-batch.

(b) A balanced mini-batch (obtained by our method).

Spurious correlation in NLP tasks



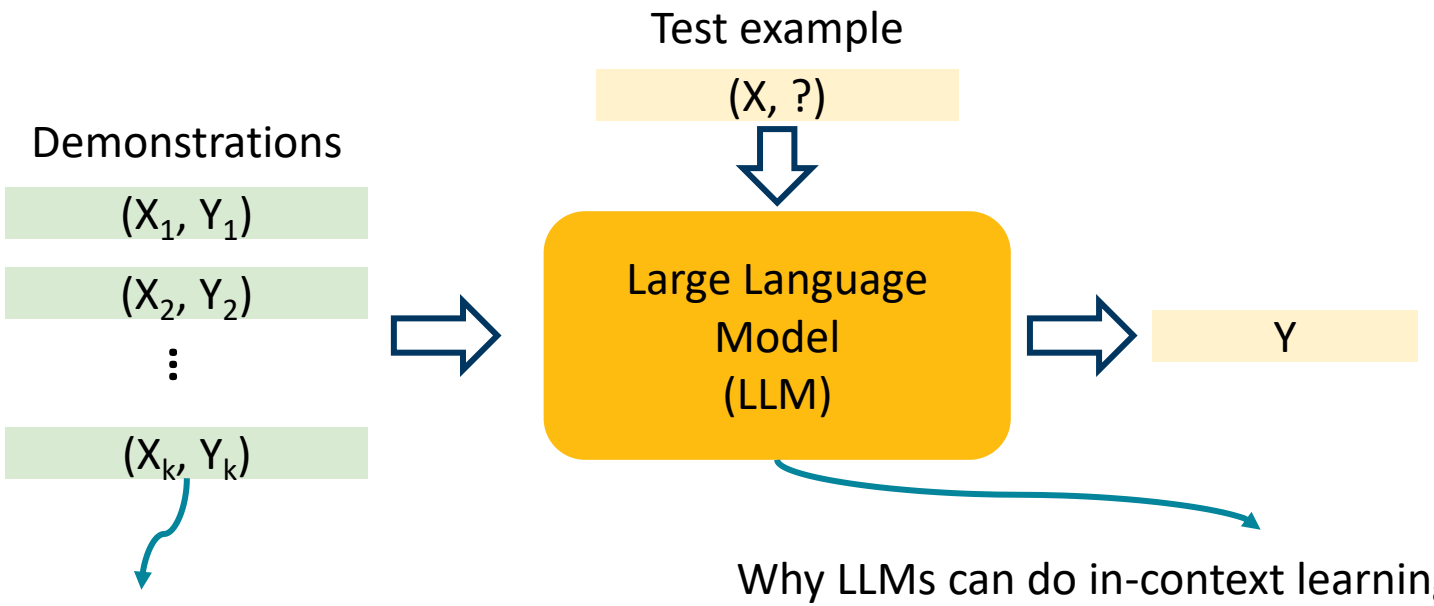
Observed distribution of natural language inference (NLI)



Observed distribution of image captioning

- Spurious correlations would not hold in OOD environments.
- **Idea 1**: learn a ***T***-invariant representation function of ***X*.**
- $$\arg\min_{\theta} \frac{1}{n}\sum_{i=1}^{n}\frac{n}{u_{t_i}}L_\theta(x_i,t_i,y_i)+\alpha\sum_{j=1}^{m}(1-\frac{u_j}{n})\text{WASS}(\{\Phi(x_i)\}_{i:t_i=j},\{\Phi(x_i)\}_{i:t_i\neq j})$$
- **Idea 2**: explicitly generate counterfactual examples at training.
- $$\arg\min_{\theta} \frac{1}{n}\sum_{i=1}^{n}[L_\theta(x_i,t_i,y_i)+\alpha(t_i)J_\phi(k_{t_i},x_i,y_\theta(x_i,k_{t_i}))]$$

In-context learning



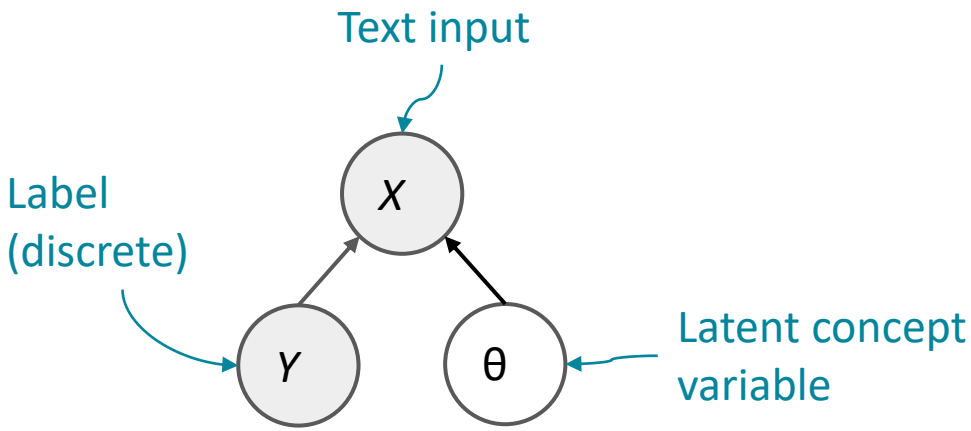
In-context learning is highly unstable:

- How do we choose the demonstrations if we have a set of annotated data? Similarity? (Liu et al. 2022; Su et al. 2022) Entropy of predicted labels? (Lu et al. 2022)

Why LLMs can do in-context learning:

- Pretraining distribution? HMM (Xie et al., 2022)? Long tailed? Burstiness (Chan et al. 2022)?
- Mimicking gradient descent? (von Oswald et al. 2022)
- Smaller models encoded in activation? (Akyurek et al. 2022)

Data generation direction matters

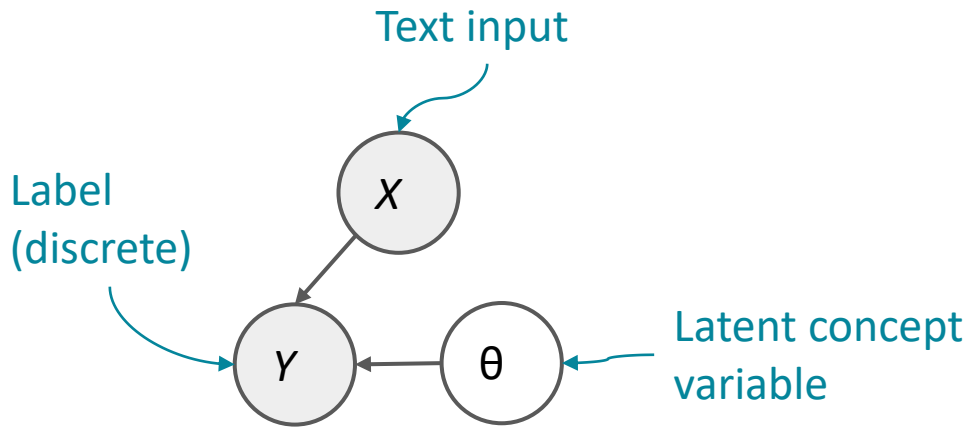


e.g. sentiment analysis, topic classification, emotion classification tasks

$$P_M^d(X|Y_1^d, X_1^d, ..., Y_k^d, X_k^d, Y)$$
$$= \int_\Theta P_M^d(X|\boldsymbol{\theta}, Y) P_M^d(\boldsymbol{\theta}|Y_1^d, X_1^d, ..., Y_k^d, X_k^d, Y) d\boldsymbol{\theta}$$

Latent concept variable learning (soft prompt tuning)

Demonstration selection



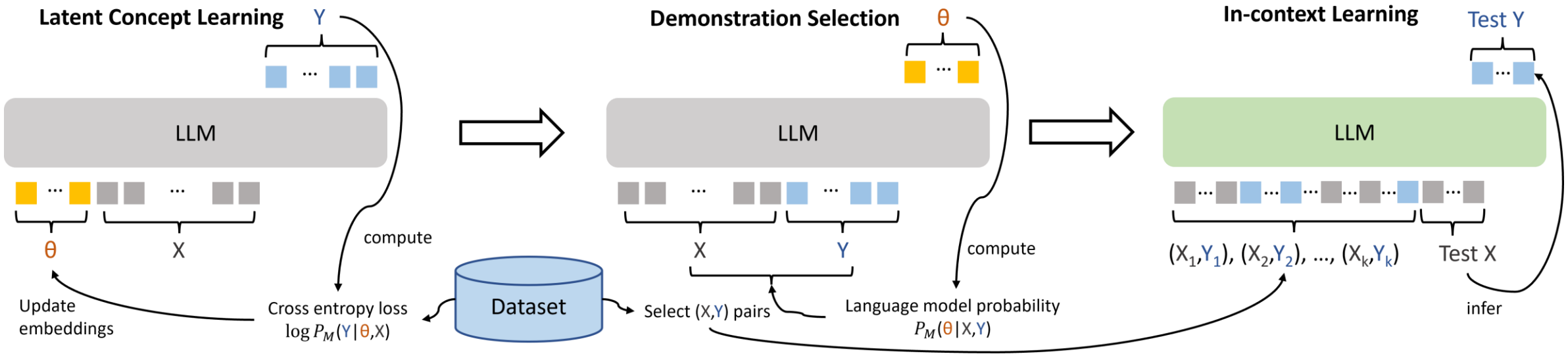
e.g. linguistic analysis, hate speech detection

$$P_M^d(Y|X_1^d, Y_1^d, ..., X_k^d, Y_k^d, X)$$
$$= \int_\Theta P_M^d(Y|\boldsymbol{\theta}, X) P_M^d(\boldsymbol{\theta}|X_1^d, Y_1^d, ..., X_k^d, Y_k^d, X) d\boldsymbol{\theta}$$

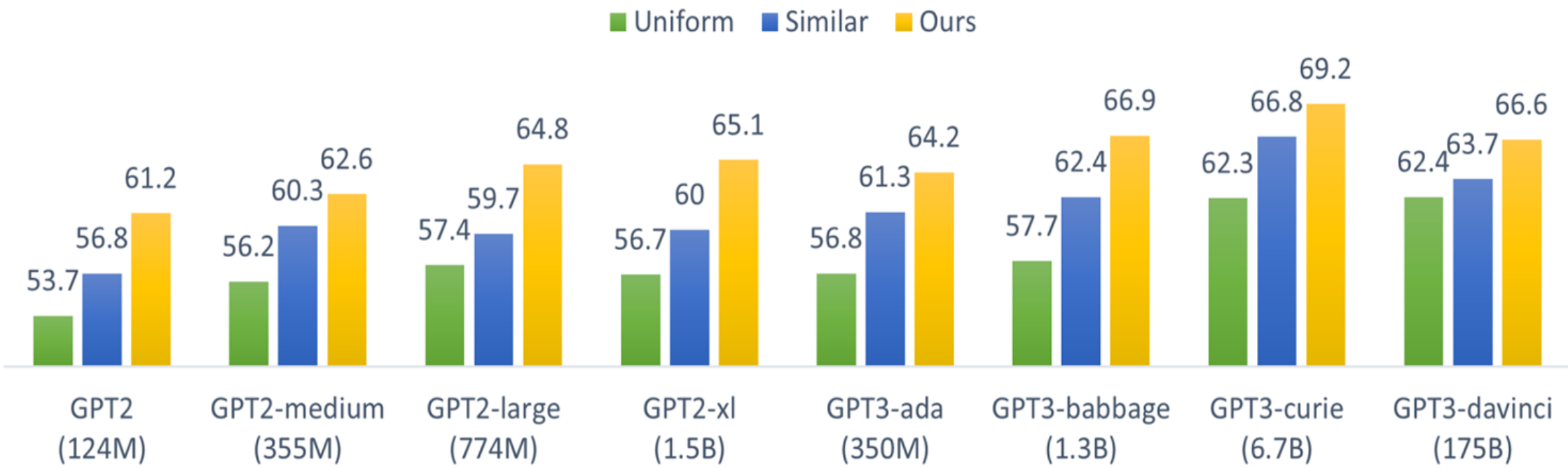
Latent concept variable learning (soft prompt tuning)

Demonstration selection

Algorithm overview



Empirical results



- Results are averaged over 8 text classification datasets, each experiment is repeated by 5 runs.
- We select the optimal demonstrations by GPT2-large, and use the same set of demonstrations for all other LLMs.

Thank you!

Questions?