Generalizing text experiments to real-world contexts with language models

Victoria Lin Louis-Philippe Morency Eli Ben-Michael

Carnegie Mellon University

Motivation

- **Goal**: How can we generate optimal texts that elicit desired responses in readers?
- **Example**: Instead of deleting toxic messages on social media, removing toxicity while retaining overall message.
- **First step**: Estimate the *causal effect* of varying a linguistic attribute on a reader's response.
- We propose an estimator for *transporting* effects from one text distribution (e.g., constructed texts from a randomized experiment) to another text distribution (e.g., natural text).
- This builds on an existing body of work that explores how texts can be used for causal inference [1, 2, 4, 5].

What does it mean to estimate a text effect?

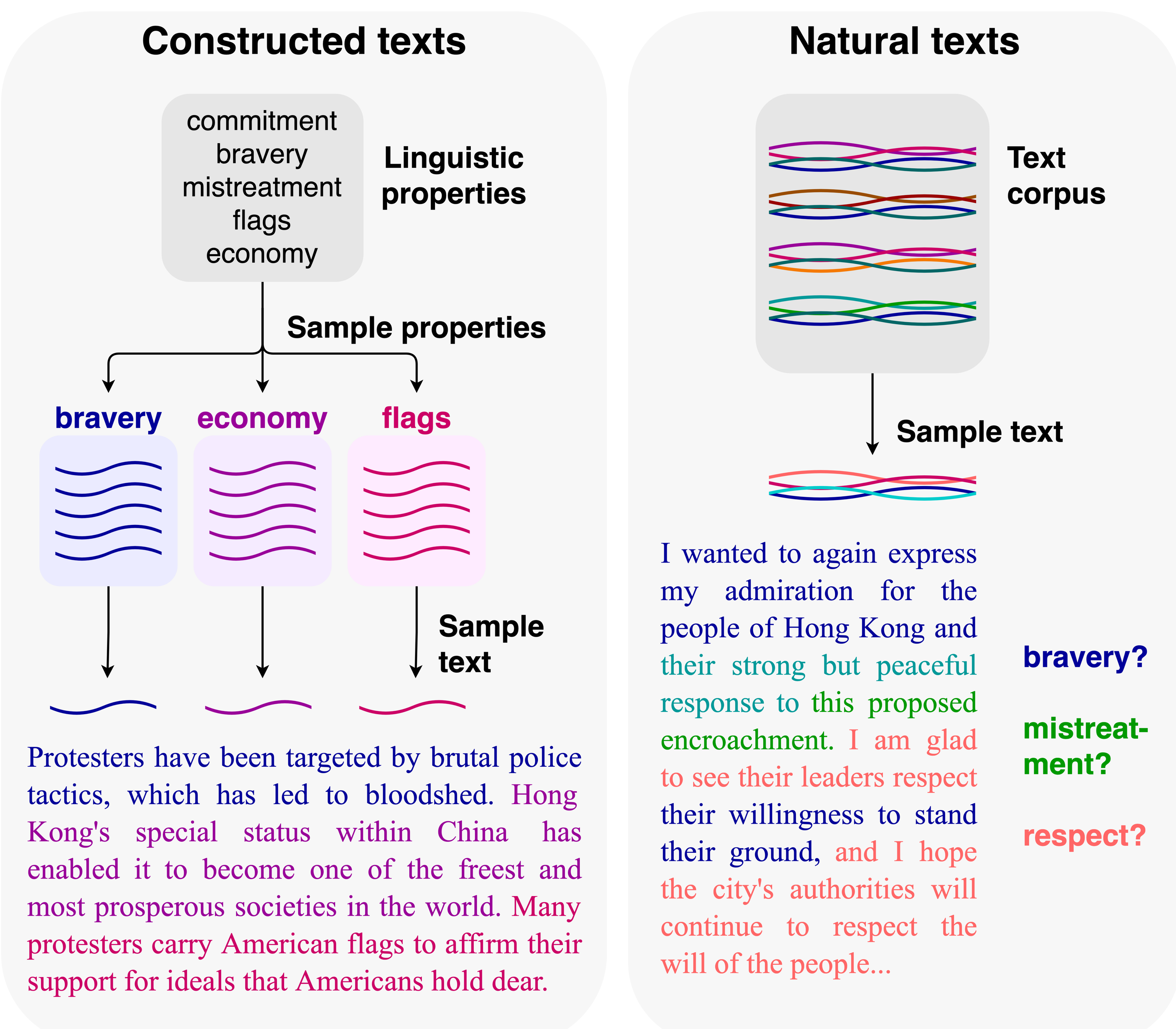
Problem setting: Consider a collection of texts (e.g., documents, sentences, utterances) $\mathcal{X}$, with individual texts $X \in \mathcal{X}$.

- $Y(X)$: Potential *outcome* of the respondent after reading X.
- We have high-dimensional treatments with potential positivity violations, so we use stochastic interventions [3].

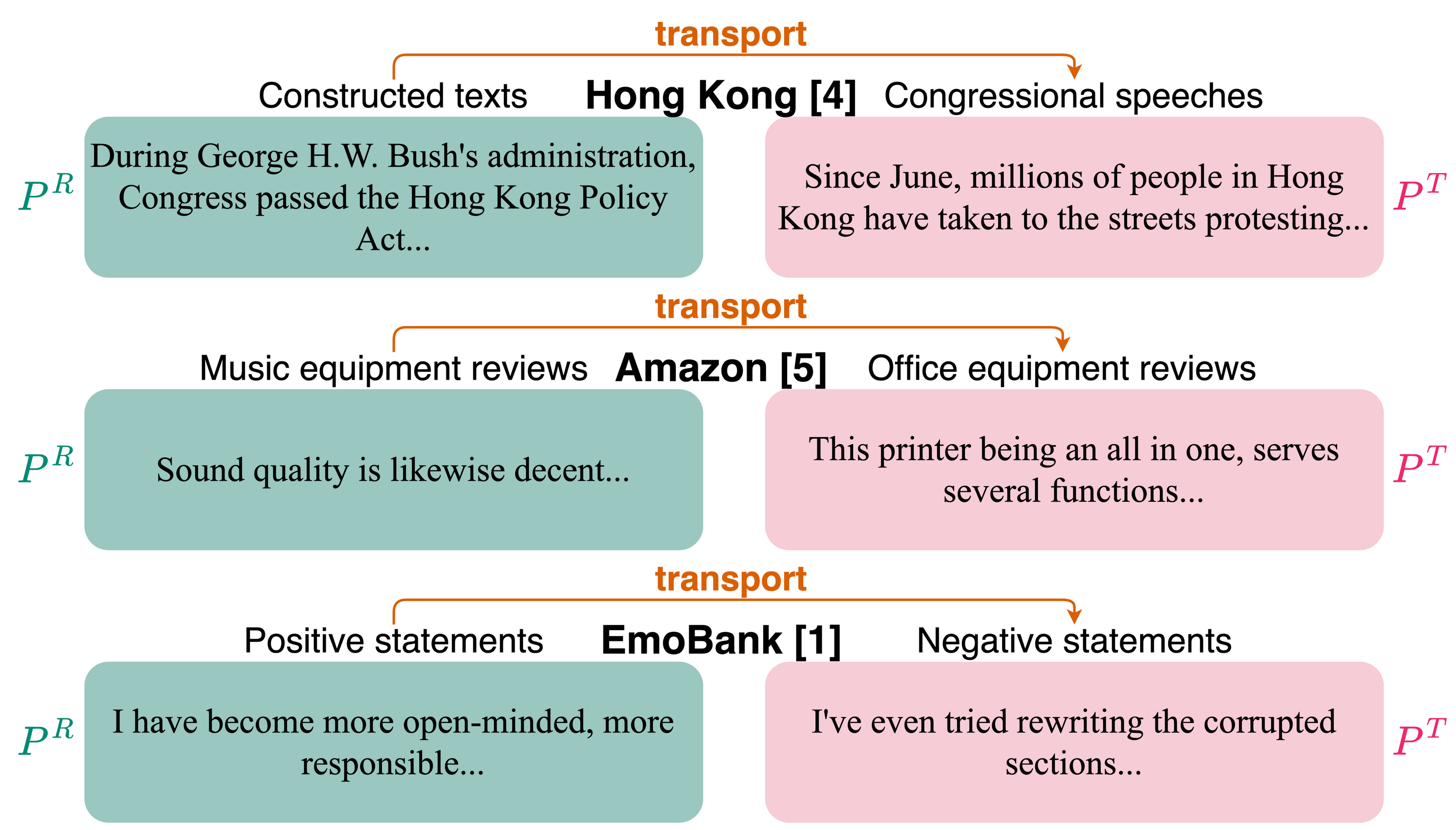
$$\mu(P) = E_{X\sim P}[Y(X)] = \frac{1}{|\mathcal{X}|}\sum_{X\in\mathcal{X}} Y(X)P(X)$$

- We can think about properties $\mu(P)$ and contrasts $\mu(P) - \mu(P')$.

Randomized text experiments



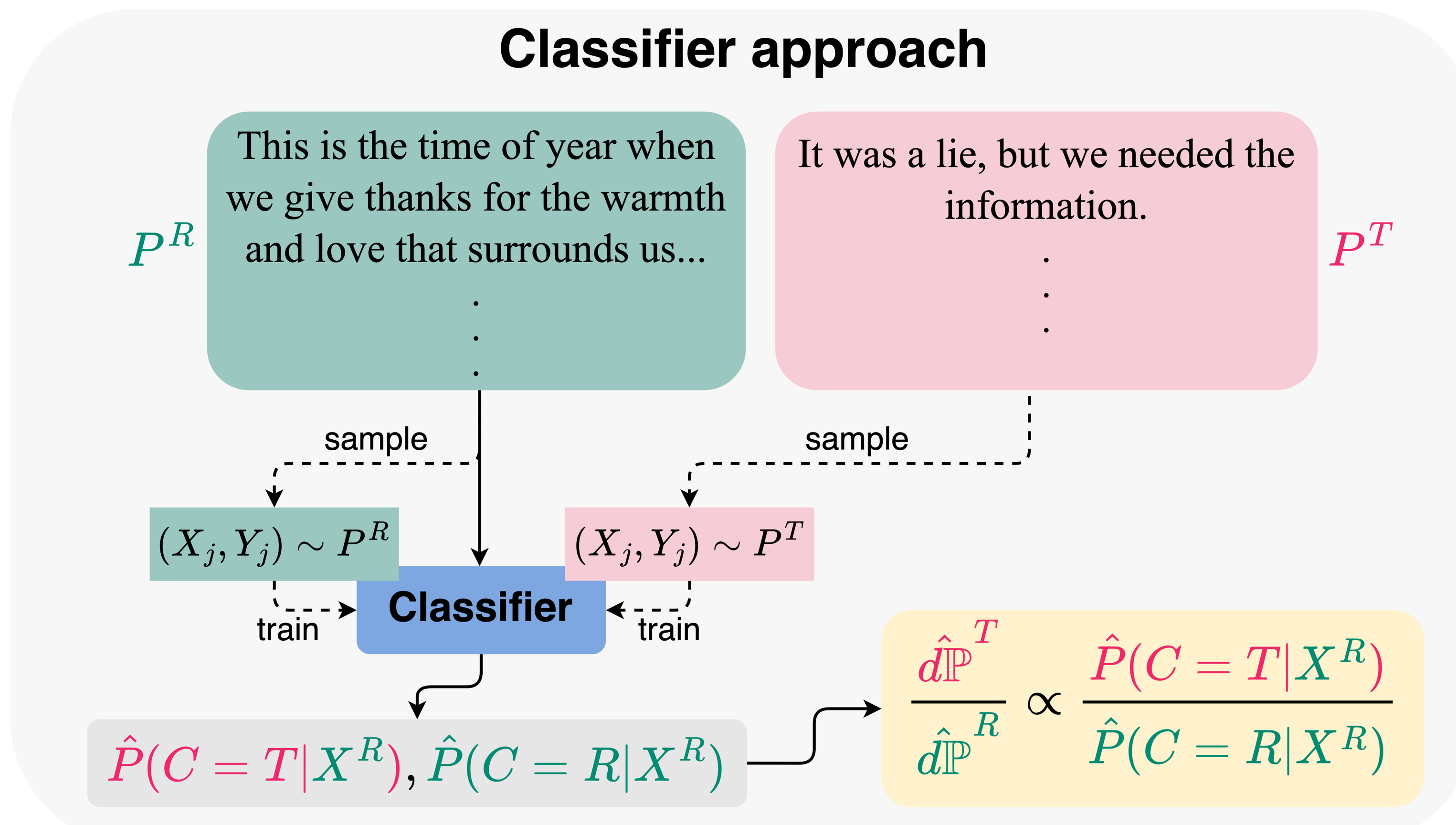
Transporting responses to texts

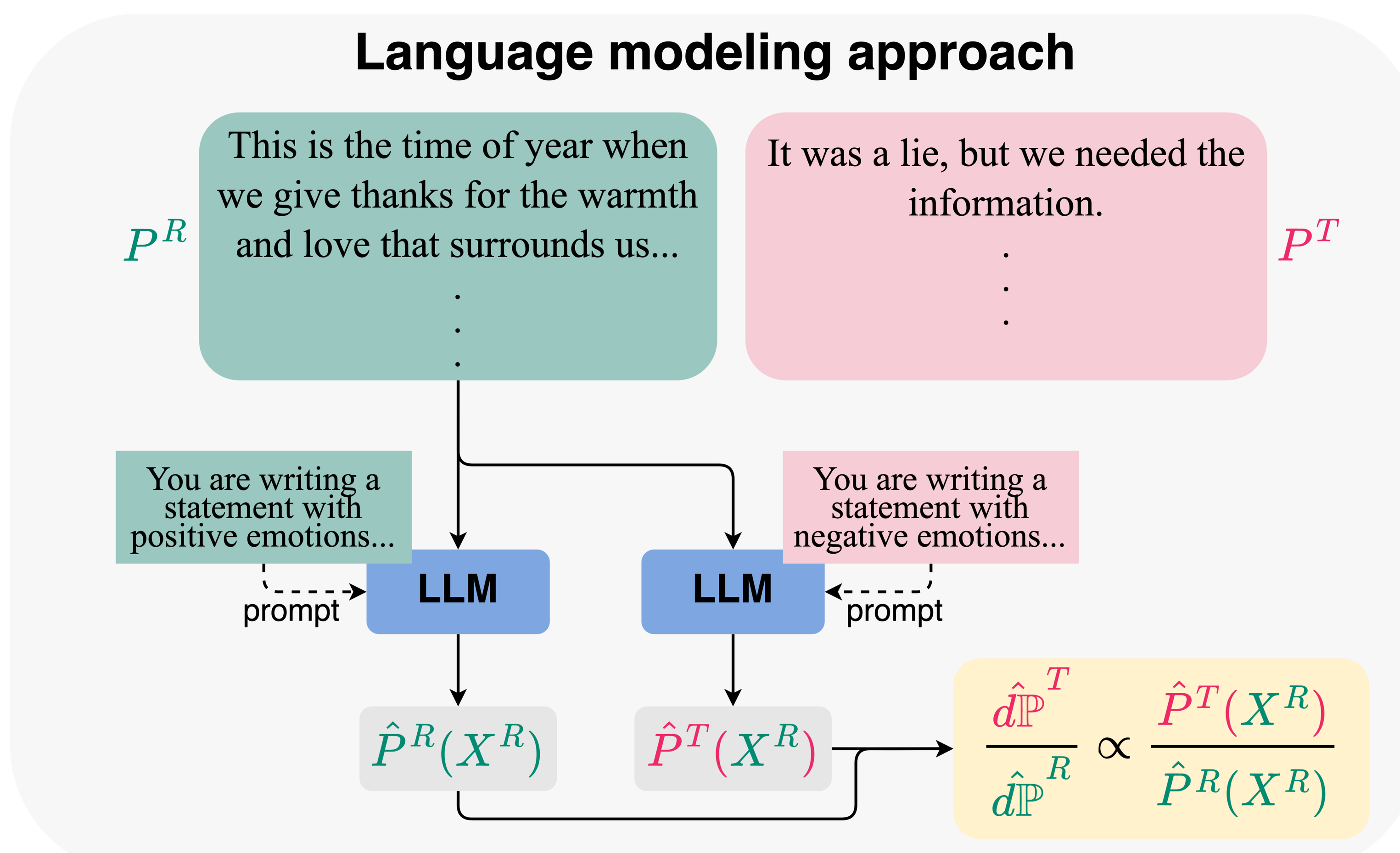


$$\hat{\mu}(P^T) = \frac{1}{n}\sum_{i=1}^{n} \frac{d\mathbb{P}^T}{d\mathbb{P}^R}(X_i)Y_i(X_i)$$

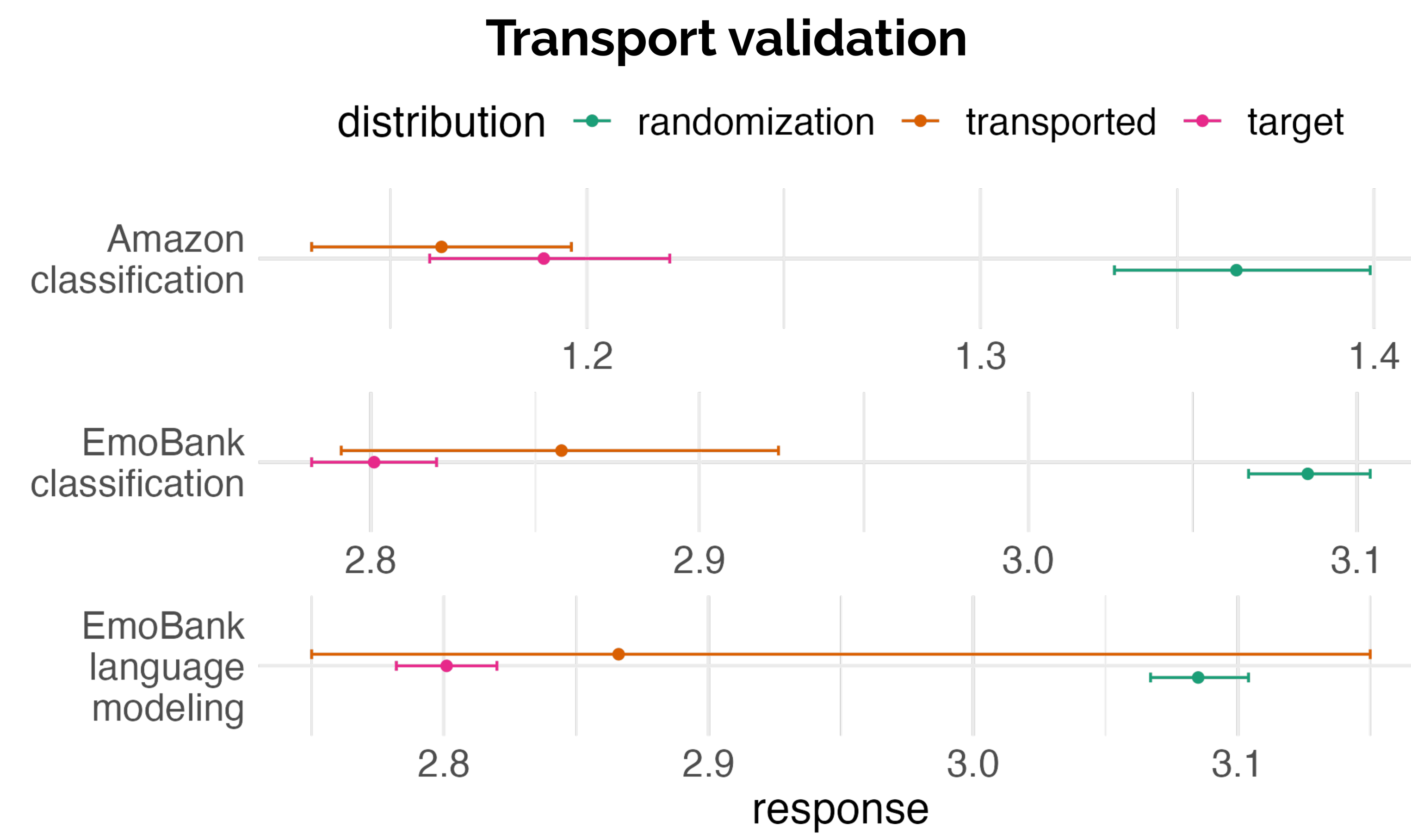
Pros: Unbiased, can explicitly compute variance, asymptotically normal under conditions.

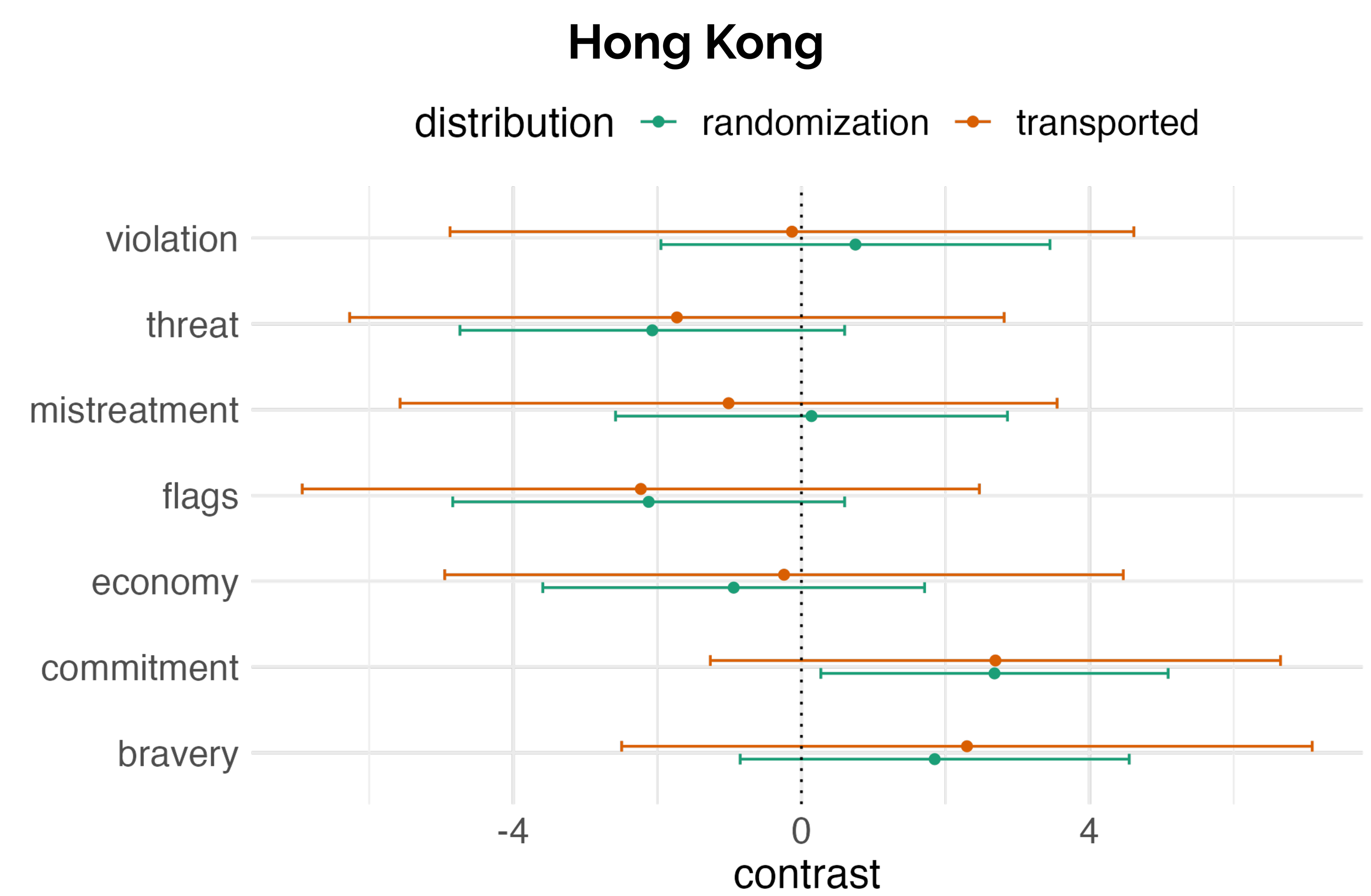
Estimation

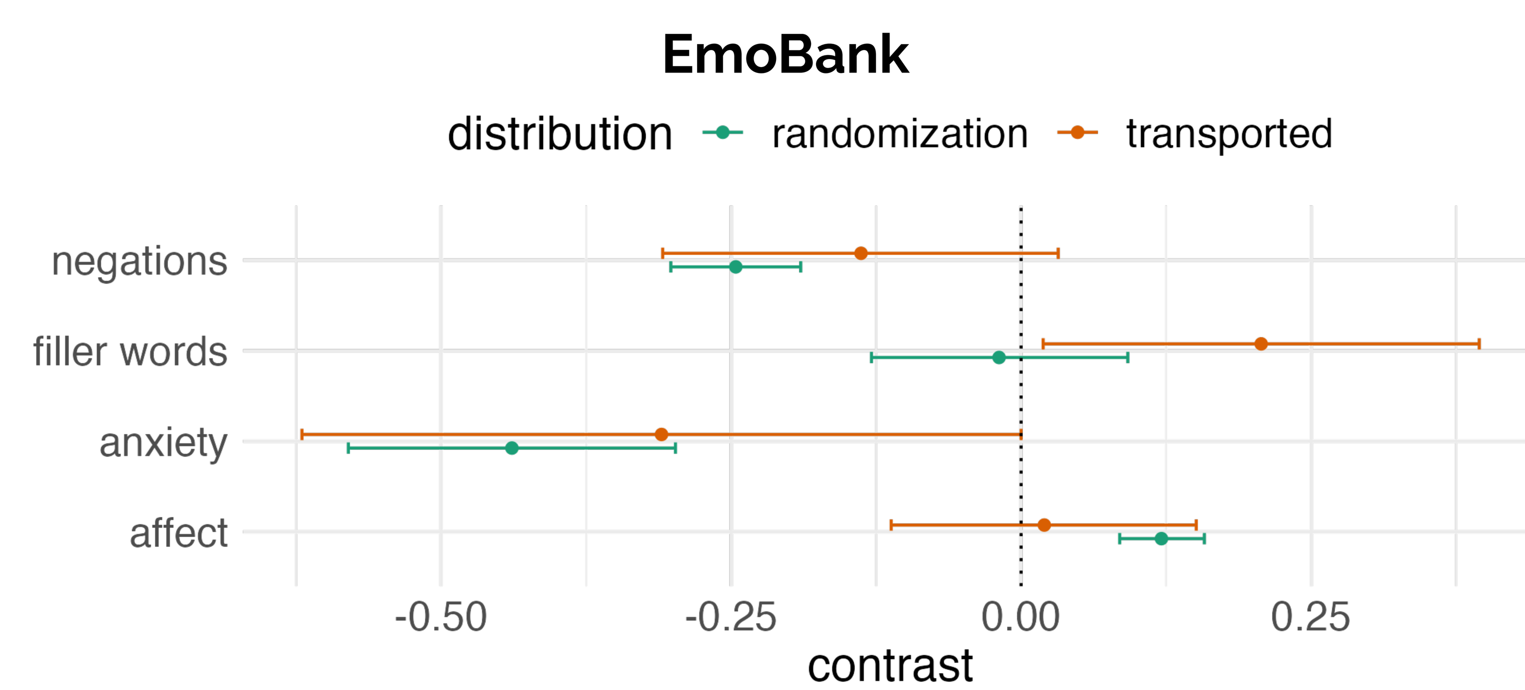




Empirical studies







References

[1] Buechel Hahn. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. *EACL 2017*.
[2] Egami et al. How to make causal inferences using texts. *arXiv preprint arXiv:1802.02163*, 2018.
[3] Feder et al. Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond. *TACL*, 10:1138–1158, 10 2022.
[4] Fong Grimmer. Causal inference with latent treatments. *American Journal of Political Science*, 67(2):374–389, 2023.
[5] McAuley Leskovec. Hidden factors and hidden topics: Understanding rating dimensions with review text. *RecSys 2013*.
[6] Papadogeorgou et al. Causal Inference with Spatio-Temporal Data: Estimating the Effects of Airstrikes on Insurgent Violence in Iraq. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 84(5):1969–1999, 11 2022.
[7] Pryzant et al. Causal effects of linguistic properties. *ACL 2021*.
[8] Veitch et al. Adapting text embeddings for causal inference. *UAI 2020*.