Noisy Channel Language Model Prompting for Few-shot Text Classification

Sewon Min, Mike Lewis, Hannaneh Hajishirzi, Luke Zettlemoyer

University of Washington, Meta AI, Allen Institute for AI

sewon@cs.washington.edu / @sewon__min

PAPER

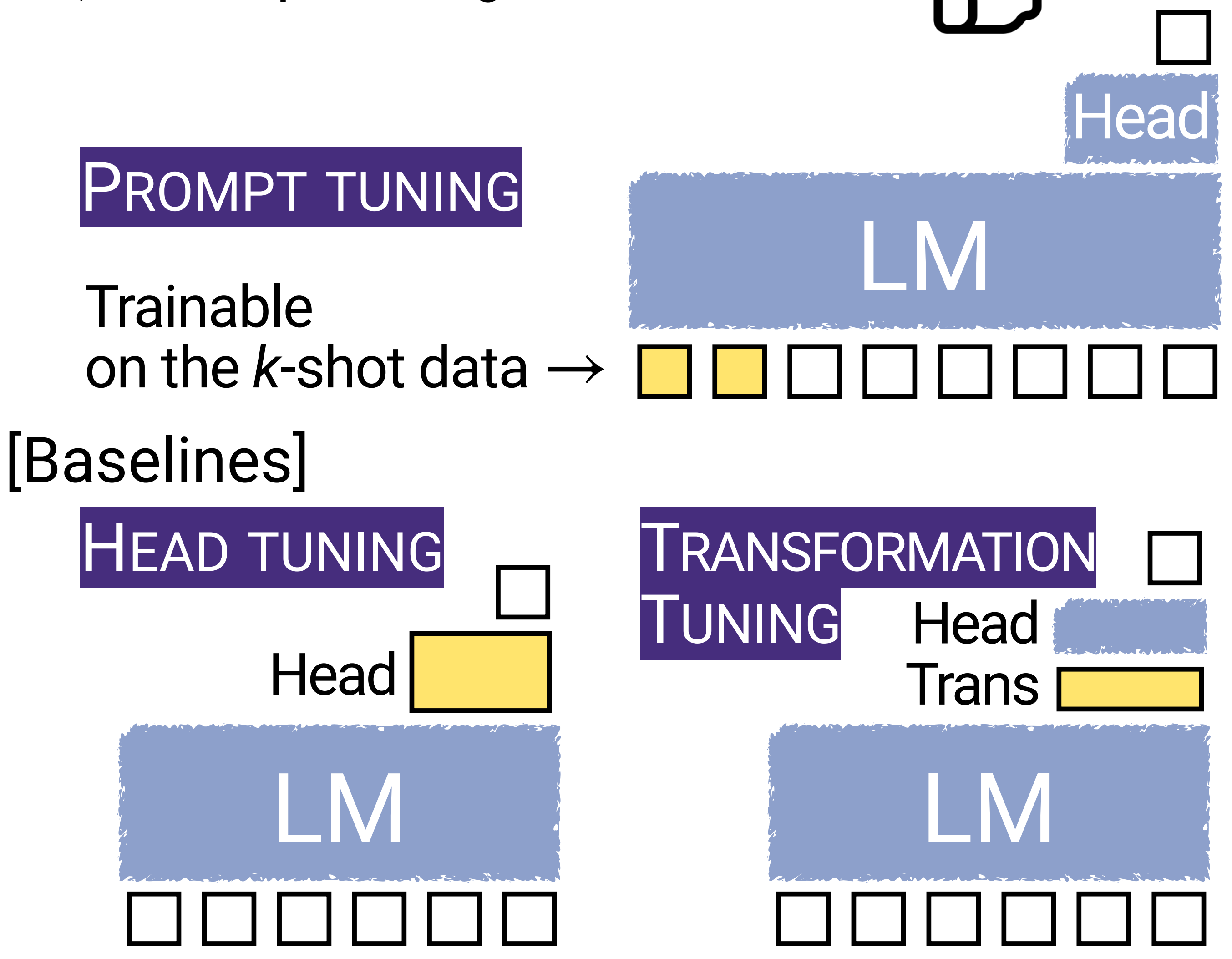
Background

LM Prompting (Brown et al 2020):
using a frozen LM for a downstream task

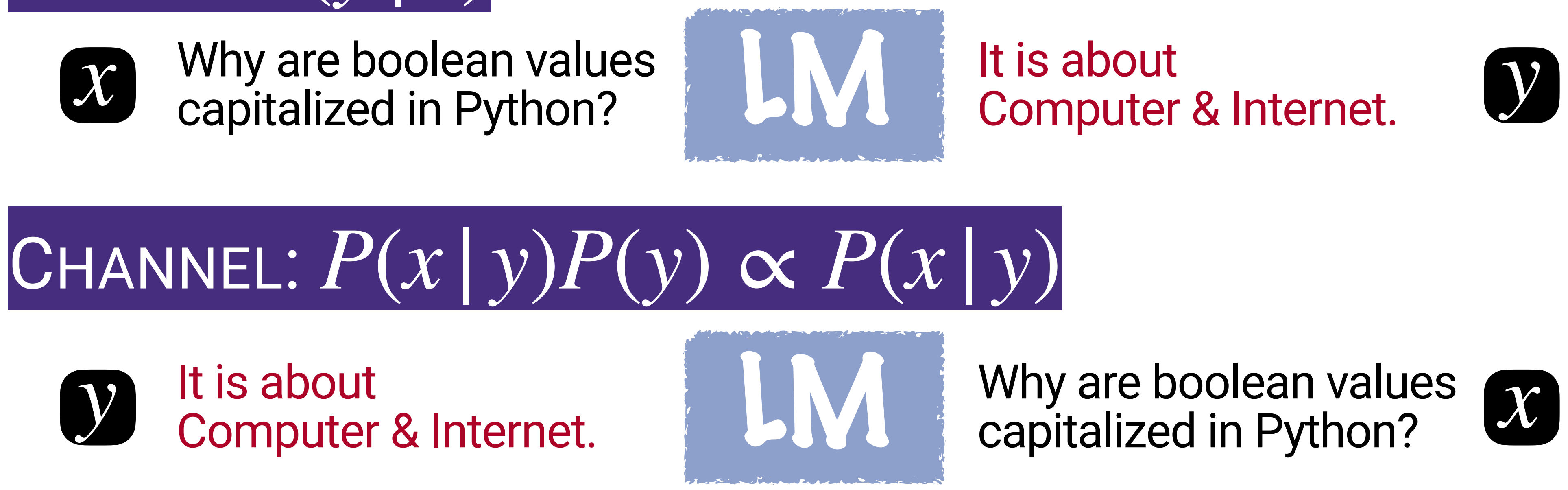
- No or very limited parameter updates
- High variance, low worst-case accuracy

Method

- ✓ Zero-shot inference
- ✓ In-context learning
- ✓ Ensemble-based In-context learning
- ✓ Prompt tuning (Lester et al 2021)



DIRECT: $P(y|x)$



x
Why are boolean values capitalized in Python?
LM
It is about Computer & Internet.
y

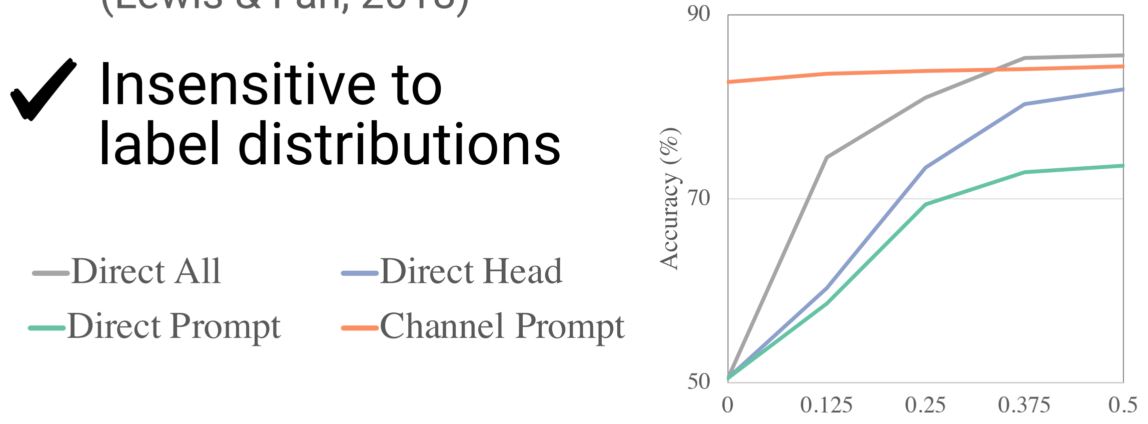
CHANNEL: $P(x|y)P(y) \propto P(x|y)$

y
It is about Computer & Internet.
LM
Why are boolean values capitalized in Python?
x

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} \propto P(x|y)P(y)$$

Why does it work?

- ✓ Better at few-shot (Ng and Jordan, 2002, Ding and Gimpel, 2019)
- ✓ More robust to distribution shift (Yogtama et al. 2017, Lewis and Fan, 2018)
- ✓ Required to predict the entire input (Lewis & Fan, 2018)
- ✓ Insensitive to label distributions



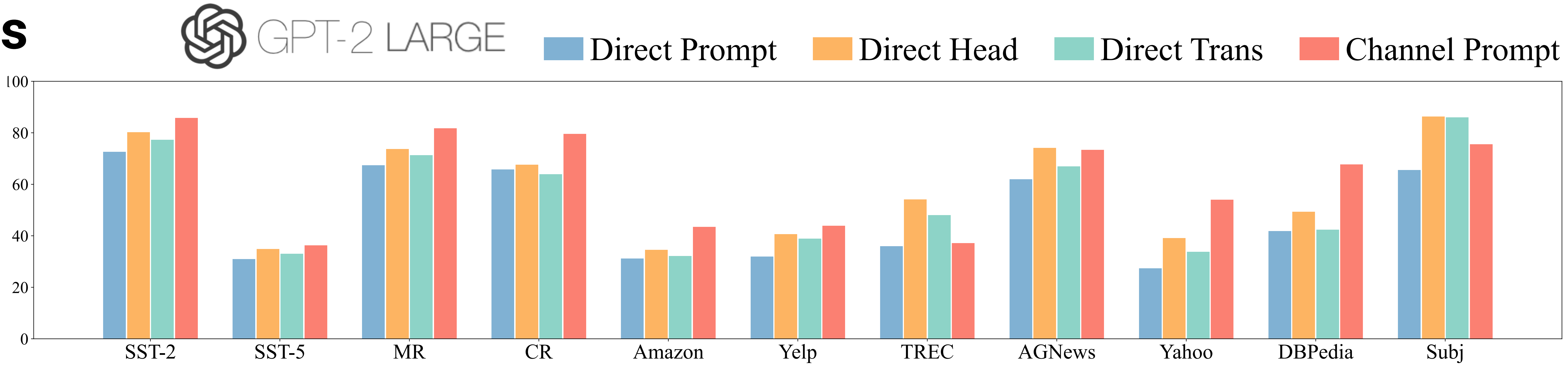
Experiments

16-shot data

4 x labels

5 x data

4 x train



- ✓ Head tuning is a powerful baseline
- ✓ Direct models suffer from high variance & low worst-case accuracy
- ✓ Channel models have significantly lower variance and higher worst-case accuracy → better performance on average

Ablations in the paper:
Channel is better with ...

Small k · *Large # of labels* · *Imbalanced data* · *Distribution shift*

CODE & DATA