# “AUTOPROMPT: Eliciting Knowledge from Language Models with Automatically Generated Prompts”, Shin et al, 2020

Thanawat Sornwanee

Stanford

2025

# Question: Do you know?

# Question: Do you know?

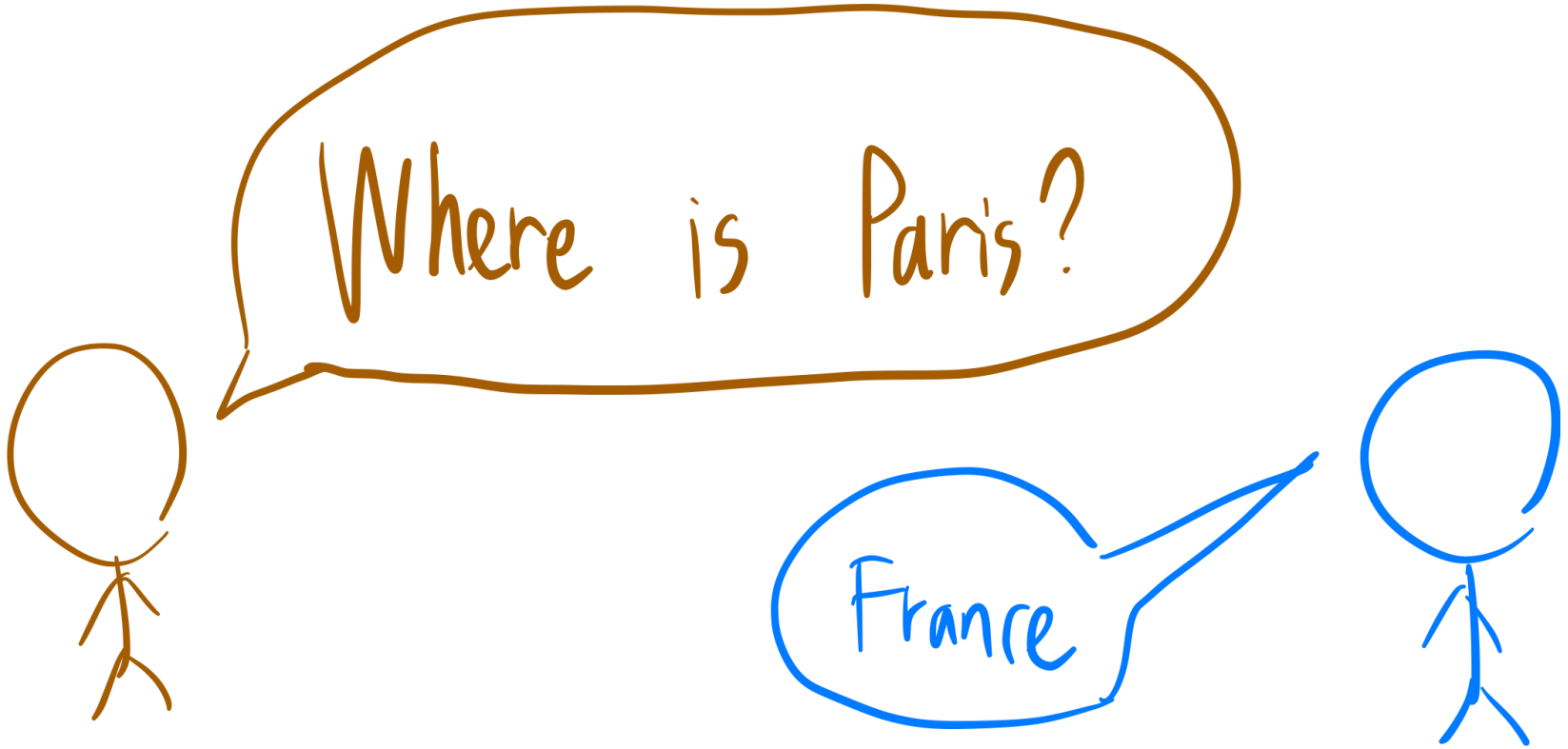
- Ask

# Question: Do you know?

- Ask

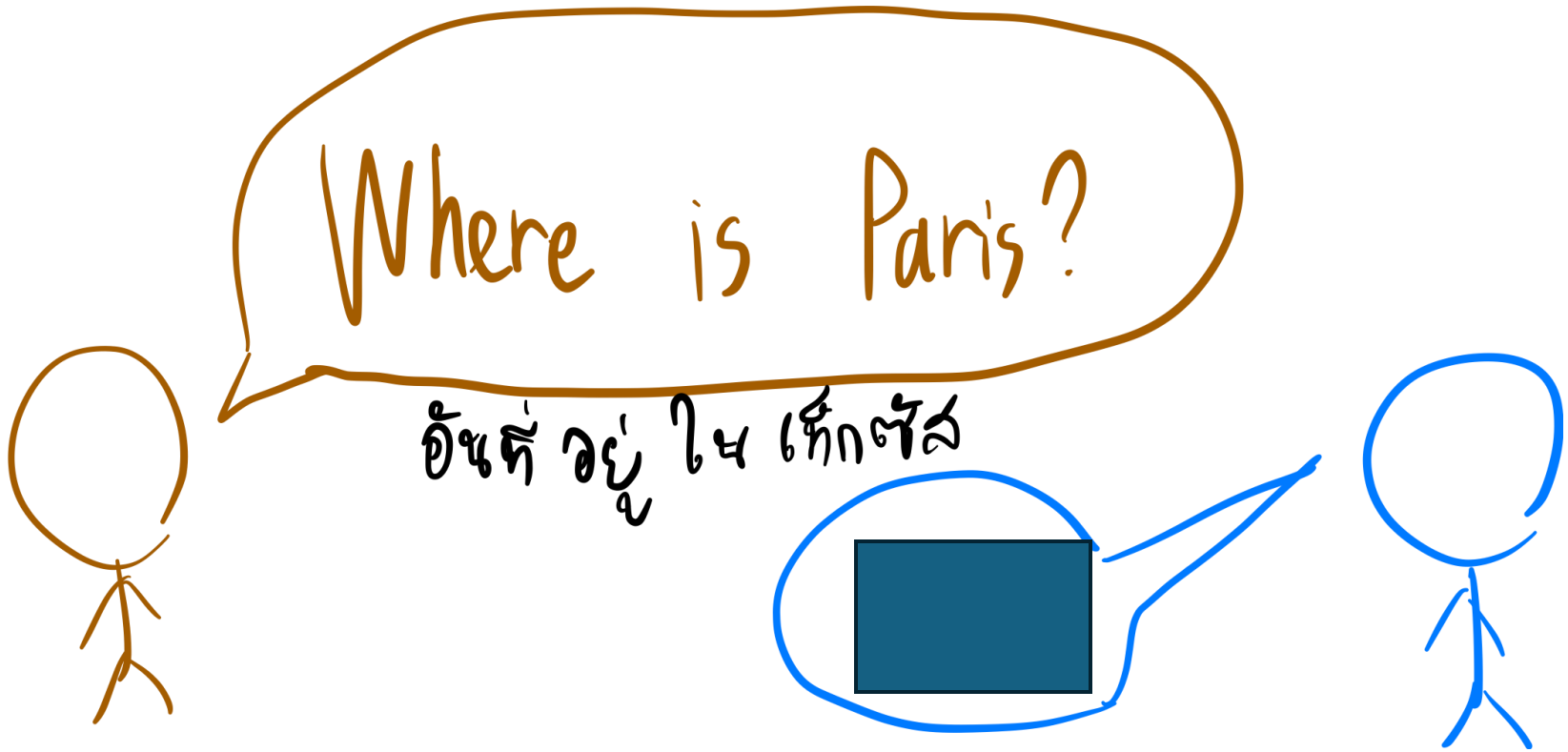
# Question: Do you know?

- Ask



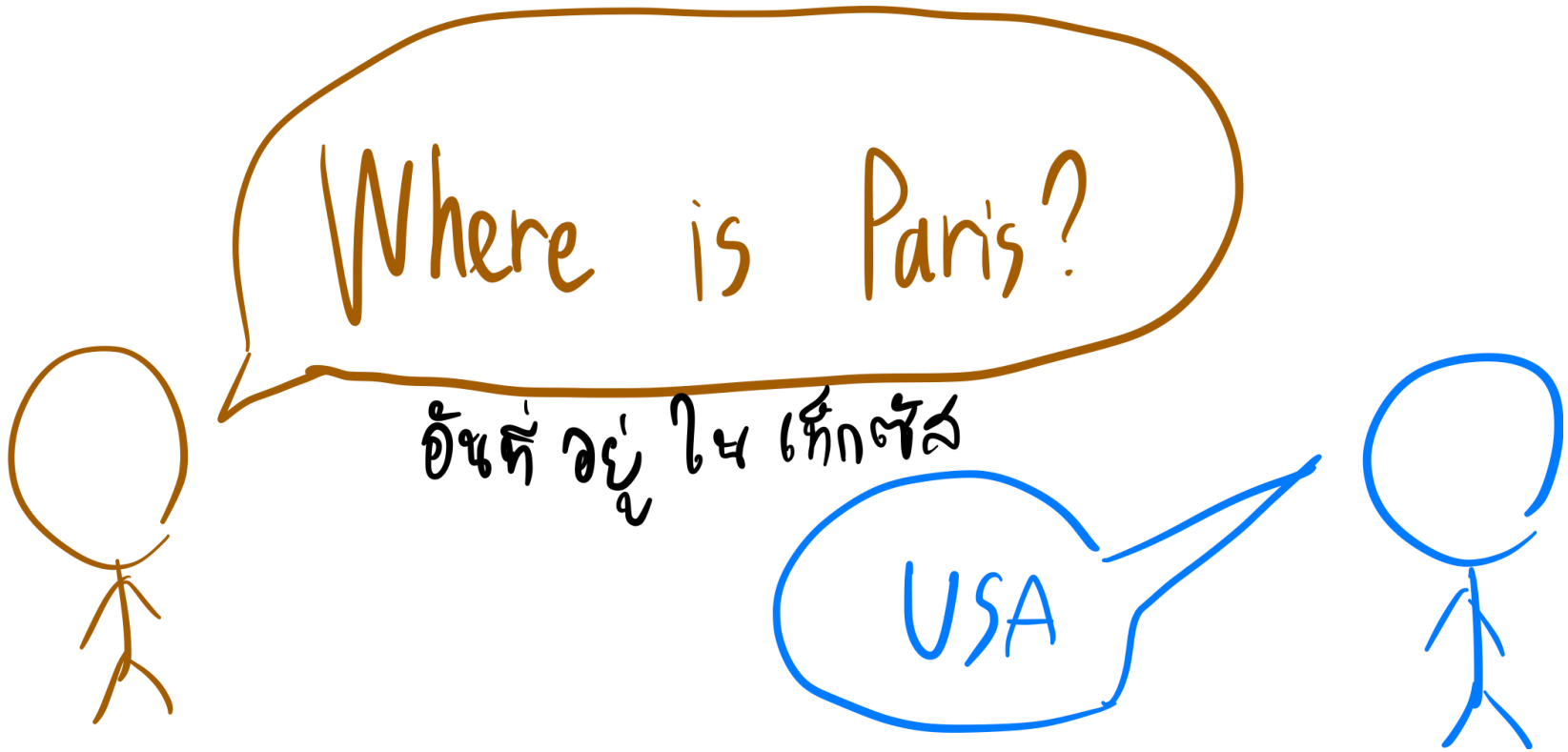
# Question: Do you know?

- Ask



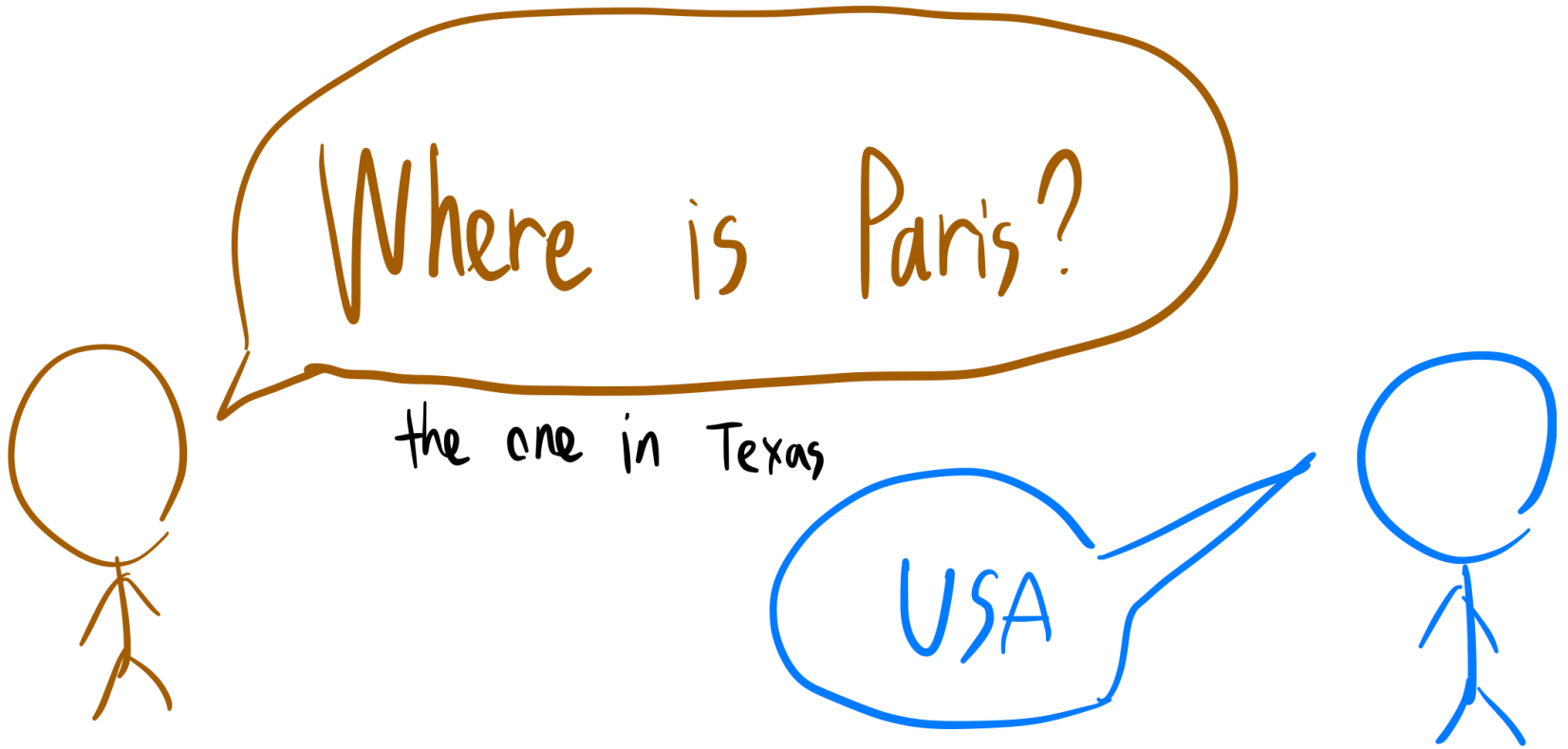
# Question: Do you know?

- Ask



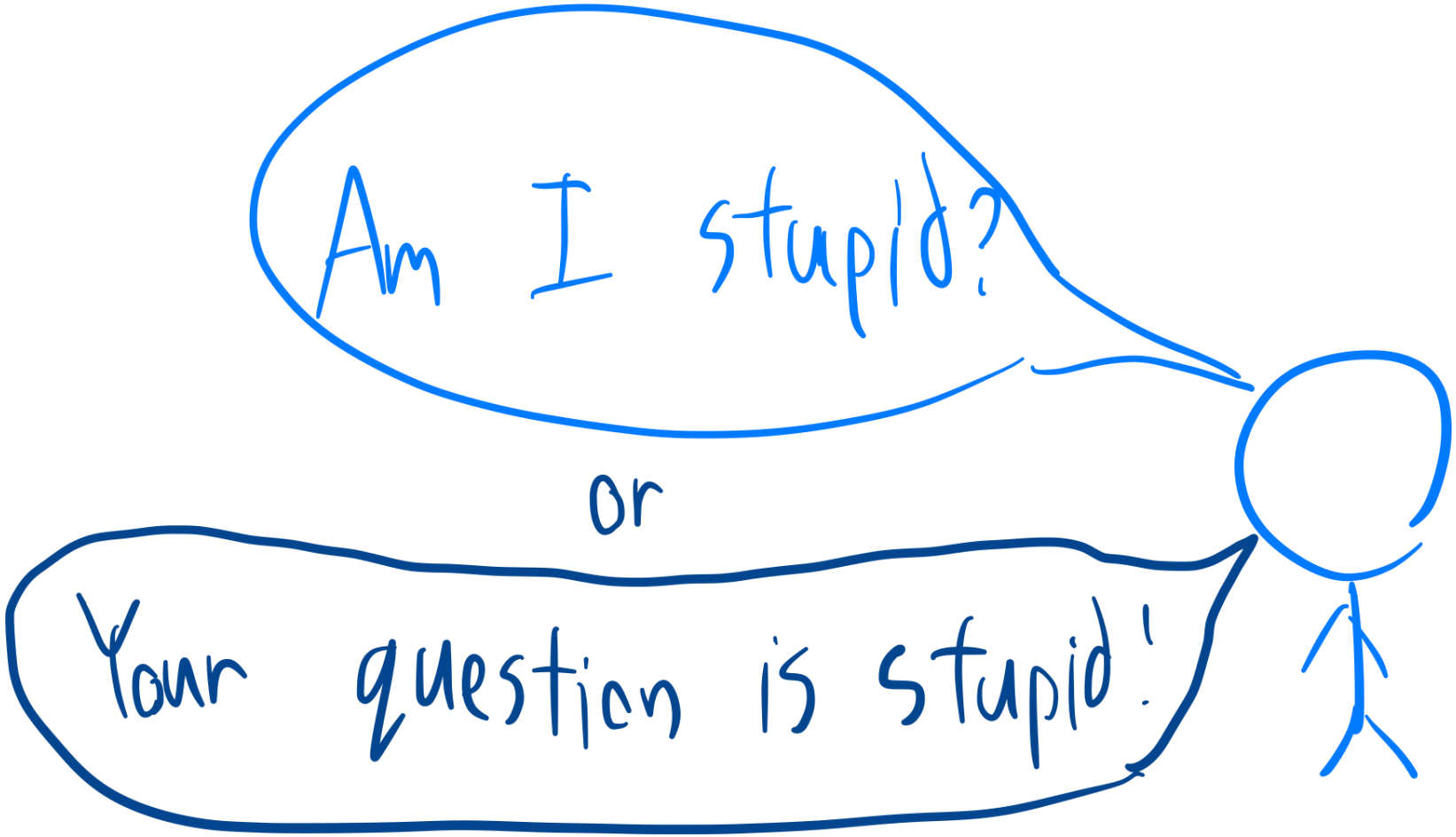
# Question: Do you know?

- Ask



# Question: Do you know?

- Ask



# Prompt Optimization

**Original Input** $x_{\text{inp}}$

a real joy.

**Trigger Tokens** $x_{\text{trig}}$

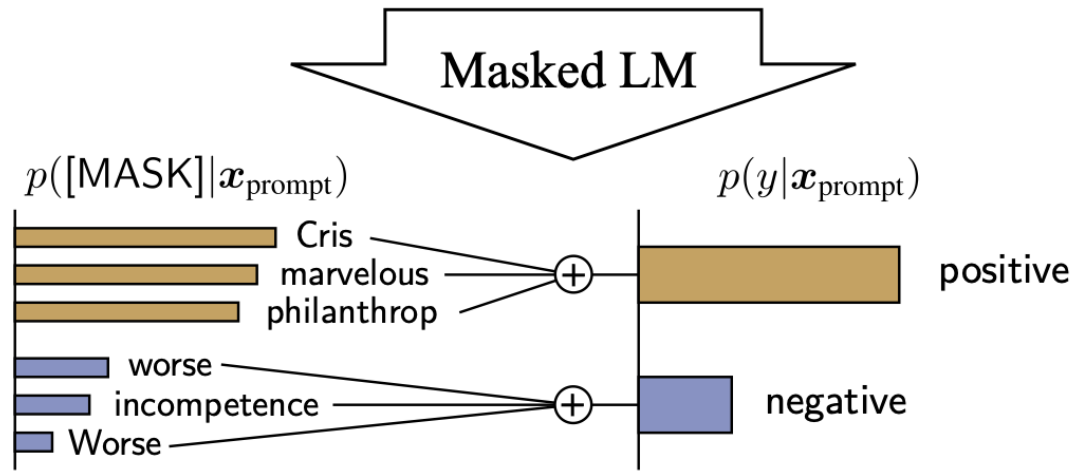
atmosphere, alot, dialogue, Clone...

**Template** $\lambda(x_{\text{inp}}, x_{\text{trig}})$

{sentence}[T][T][T][T][T][P].

**AUTOPROMPT** $x_{\text{prompt}}$

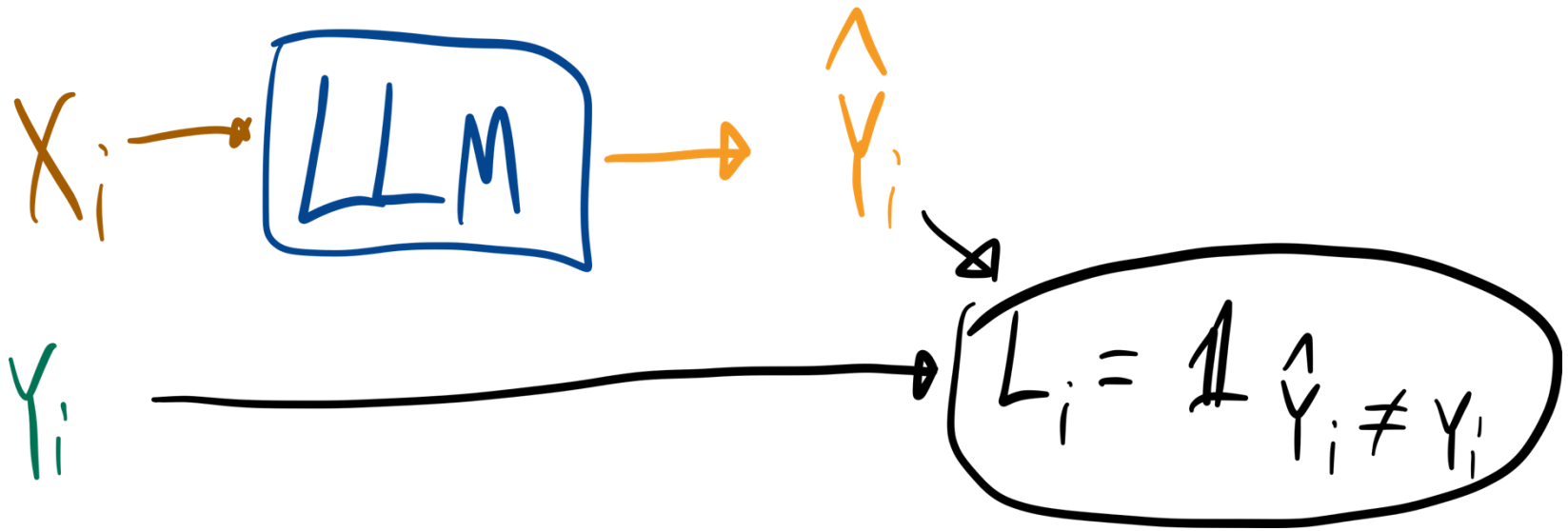
a real joy. atmosphere alot dialogue Clone totally [MASK].



# Prompt Optimization

Question $X_i$

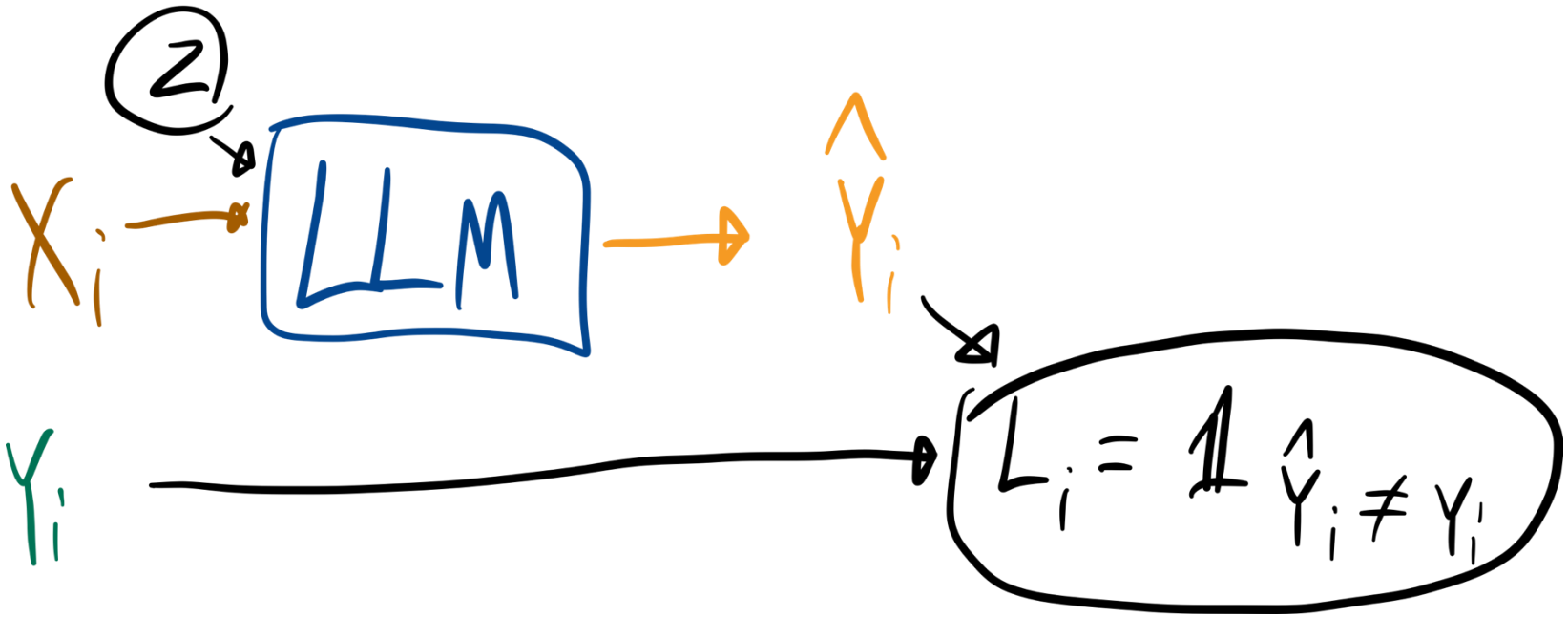
Answer $Y_i$



# Prompt Optimization

Question $X_i$

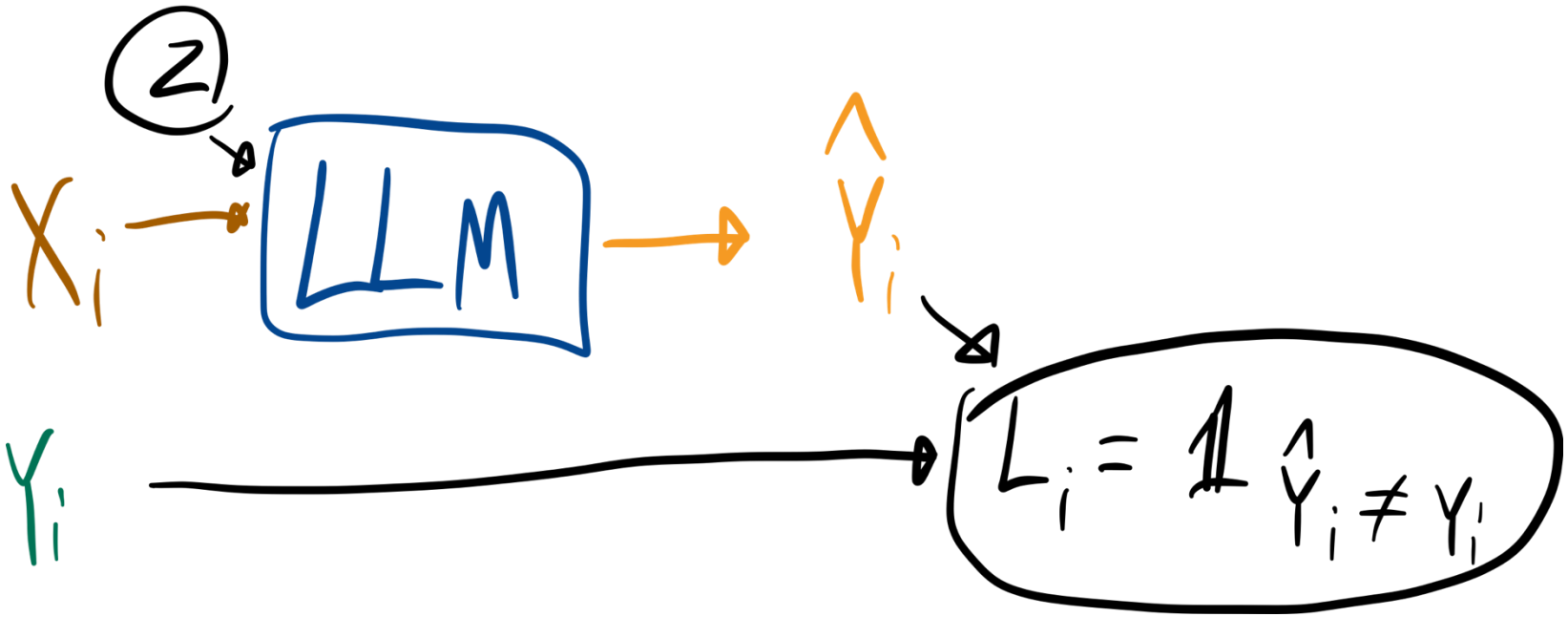
Answer $Y_i$



# Prompt Optimization

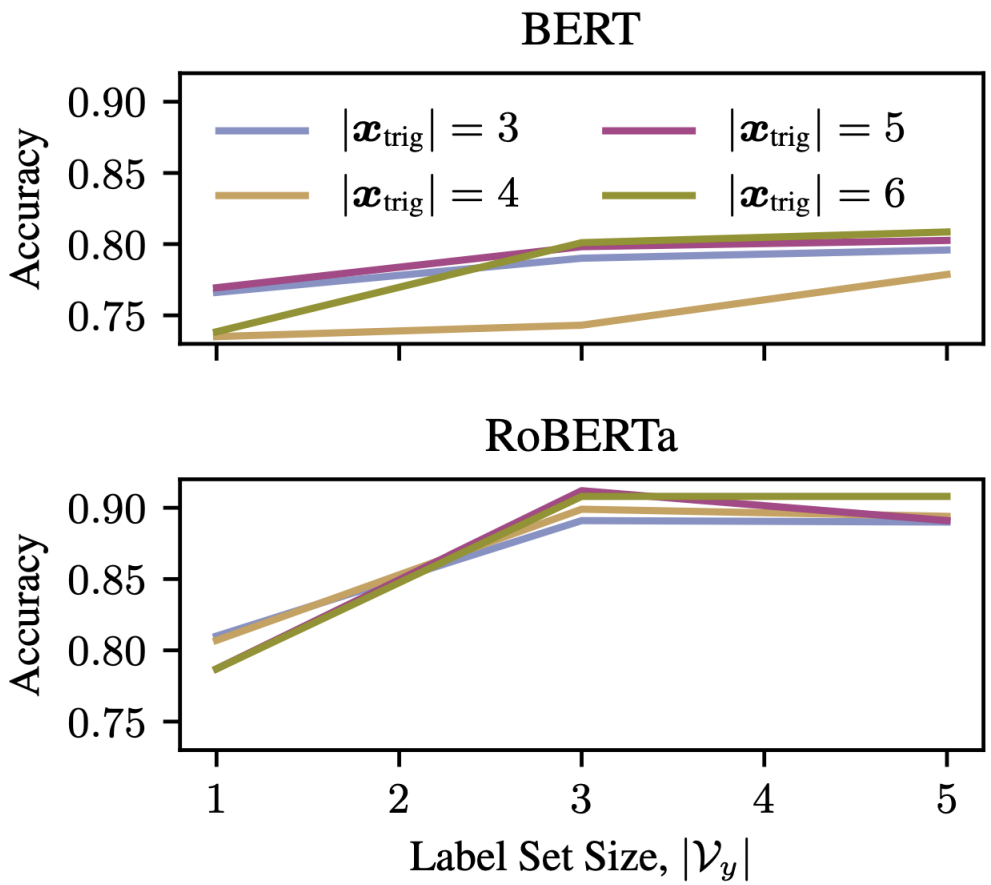
Question $X_i$

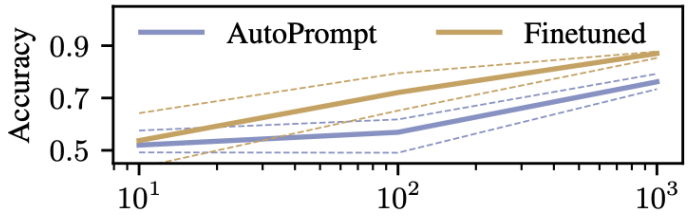
Answer $Y_i$



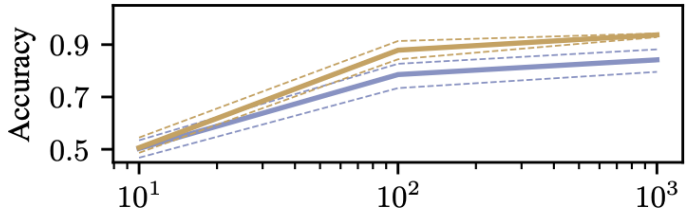

Optimize over Z

# Result

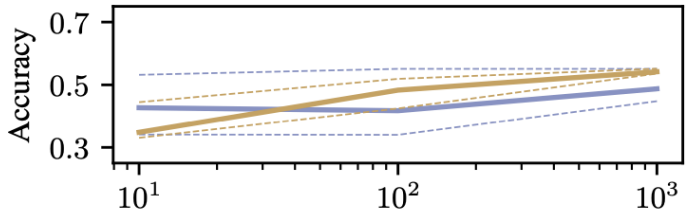


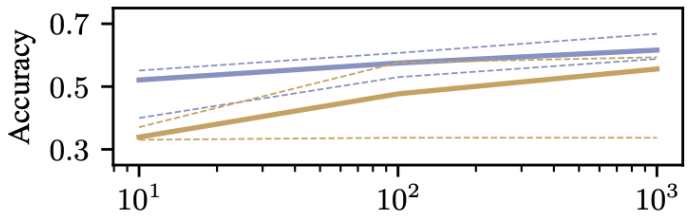

(a) BERT on SST-2




(b) RoBERTa on SST-2




(c) BERT on SICK-E




(d) RoBERTa on SICK-E

Training Data Size

# Discussion

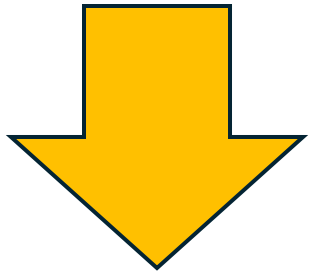
# Prompting is Lower Bound

# Prompting is Lower Bound

$$\min_{Z} \mathbb{E}[l(f(X_i, Z), Y_i)]$$

# Prompting is Lower Bound

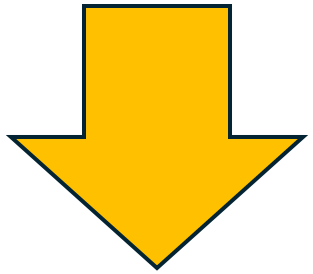
$$\min_{Z} \mathbb{E}[l(f(X_i, Z), Y_i)]$$



$$\min_{Z_i} \mathbb{E}[l(f(X_i, Z_i), Y_i)]$$

# Prompting is Lower Bound

$$\min_{Z} \mathbb{E}[l(f(X_i, Z), Y_i)]$$

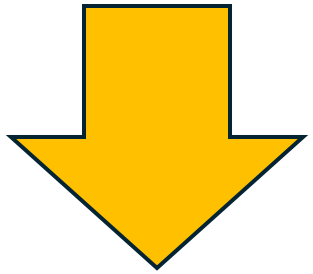


$$\min_{Z_i} \mathbb{E}[l(f(X_i, Z_i), Y_i)]$$

$$Z \perp Y, X$$

# Prompting is Lower Bound

$$\min_{Z} \mathbb{E}[l(f(X_i, Z), Y_i)]$$



$$\min_{Z_i} \mathbb{E}[l(f(X_i, Z_i), Y_i)]$$

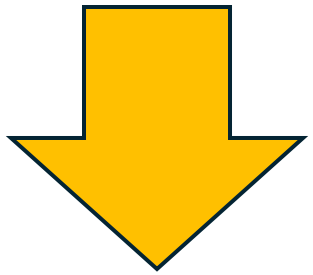
$$Z \perp Y|X?$$

# Adversarial

$$\max_{Z} \mathbb{E}[l(f(X_i, Z), Y_i)]?$$

# Adversarial

$$\max_{Z} \mathbb{E}[l(f(X_i, Z), Y_i)] \, ?$$



$$\max_{Z} \min_{Z'} \mathbb{E}[l(f(X_i, Z, Z'), Y_i)] \, ?$$