Selective Generation for Controllable Language Models

Minjae Lee[1*], Kyungmin Kim[1*], Taesoo Kim[2], Sangdon Park[1]

[1]**POSTECH**, [2]**Georgia Tech**
*Equal contribution

⭐ **NeurIPS 2024 Spotlight Paper**

Selective Classification

Motivation

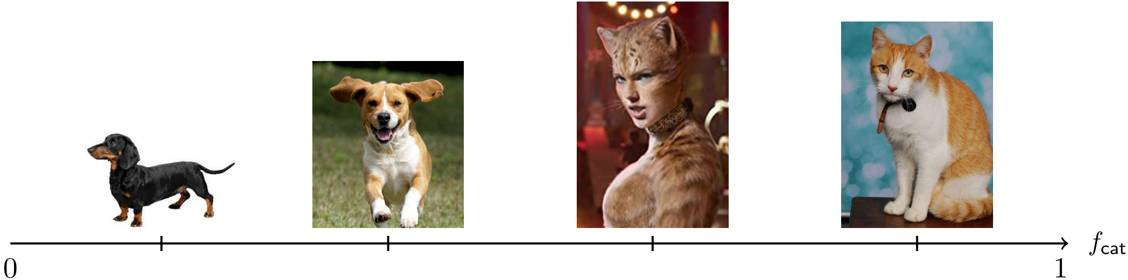
$f_{\text{cat}}\Big\{$  $\Big\} < \tau \quad \Longrightarrow$ Reject!

- Selective classification is a **certified** risk control method, which rejects instances as needed, to grant a **desired risk** ε with **high probability** $1 - \delta$.

Why Selective Classification?

Motivation

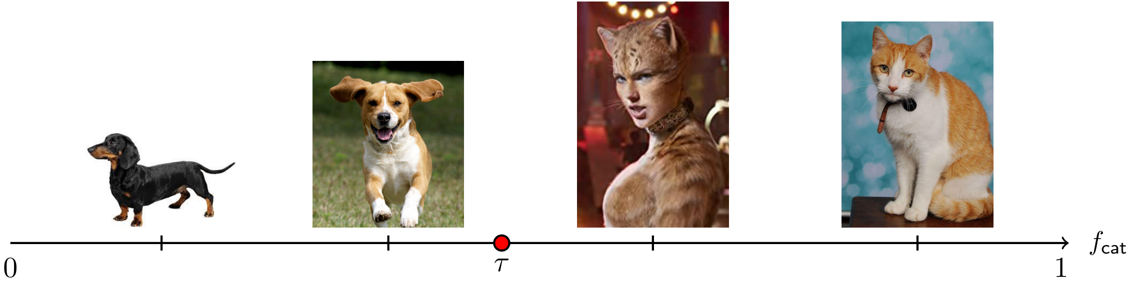
Training Phase



Why Selective Classification?

Motivation

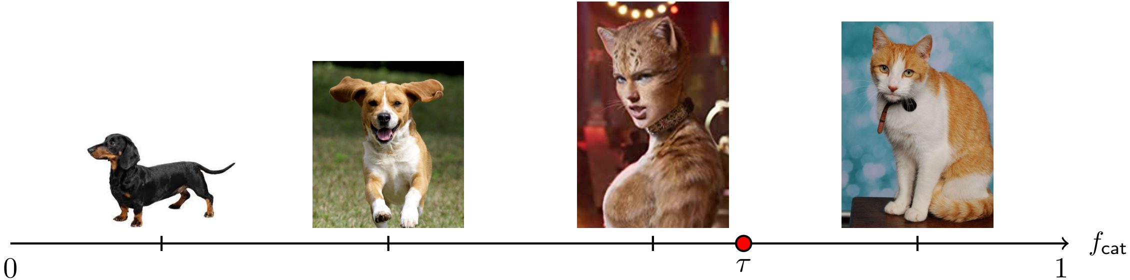
Training Phase



Why Selective Classification?

Motivation

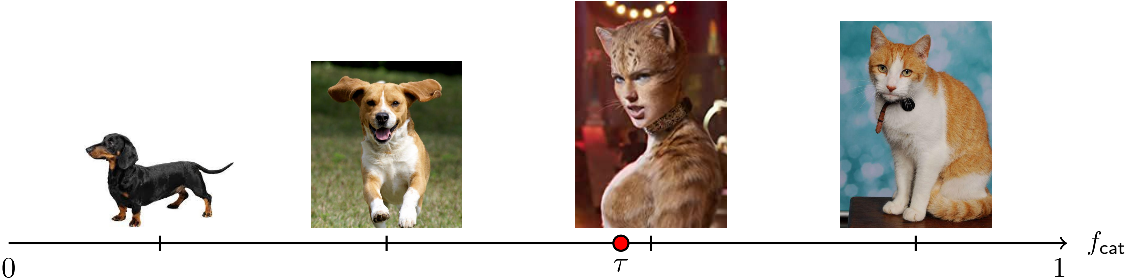
Training Phase



Why Selective Classification?

Motivation

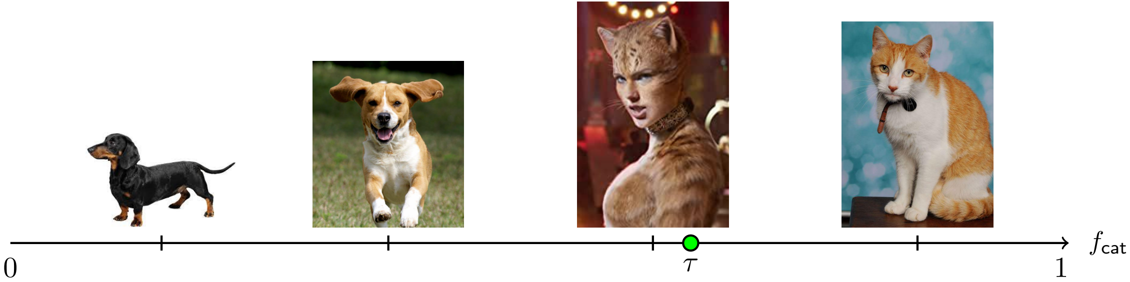
Training Phase



Why Selective Classification?

Motivation

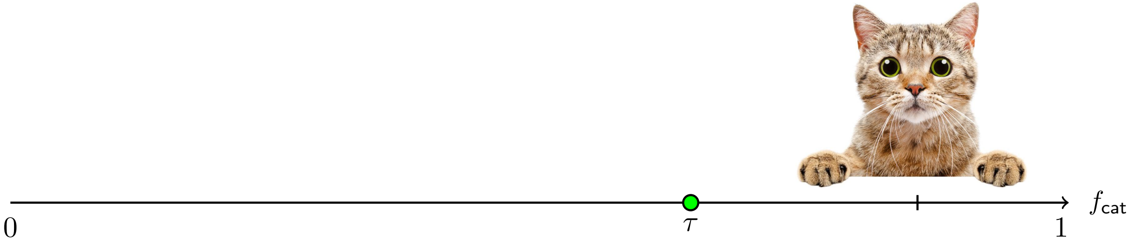
Training Phase



Why Selective Classification?

Motivation

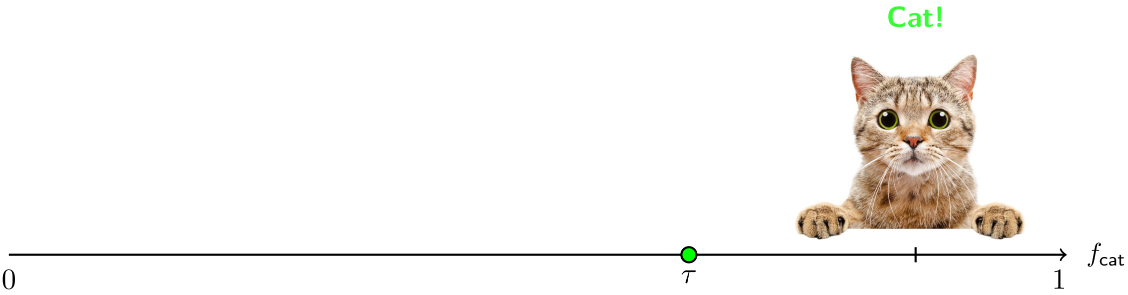
Inference Phase



Why Selective Classification?

Motivation

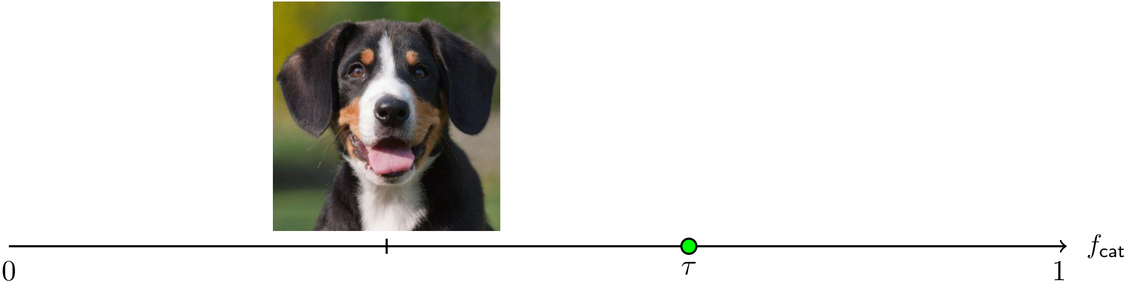
Inference Phase



Why Selective Classification?

Motivation

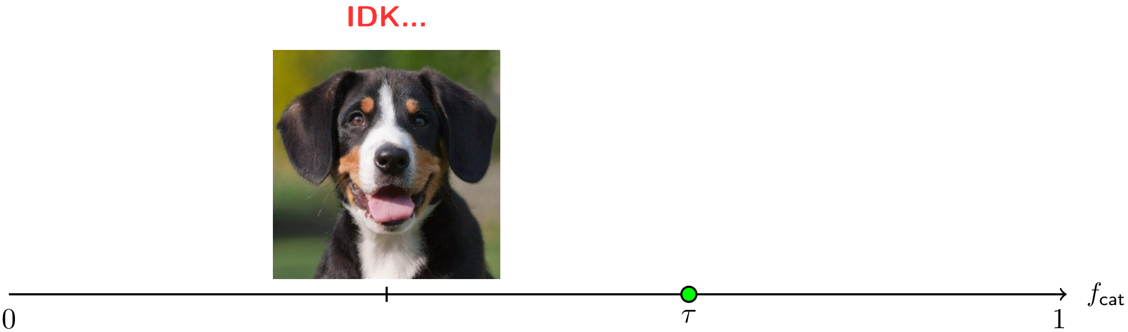
Inference Phase



Why Selective Classification?

Motivation

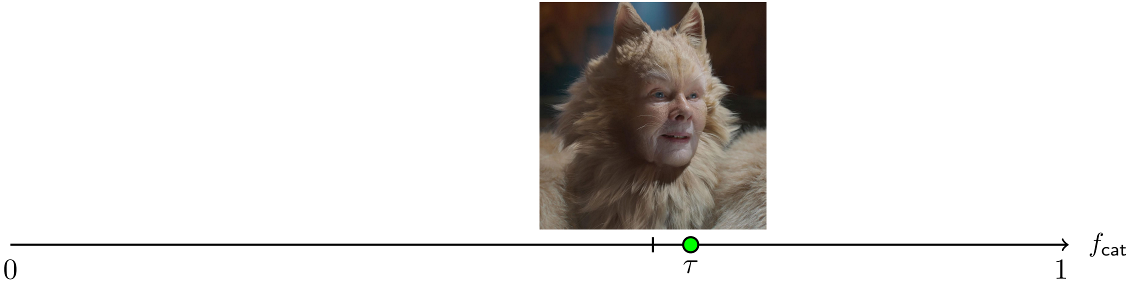
Inference Phase



Why Selective Classification?

Motivation

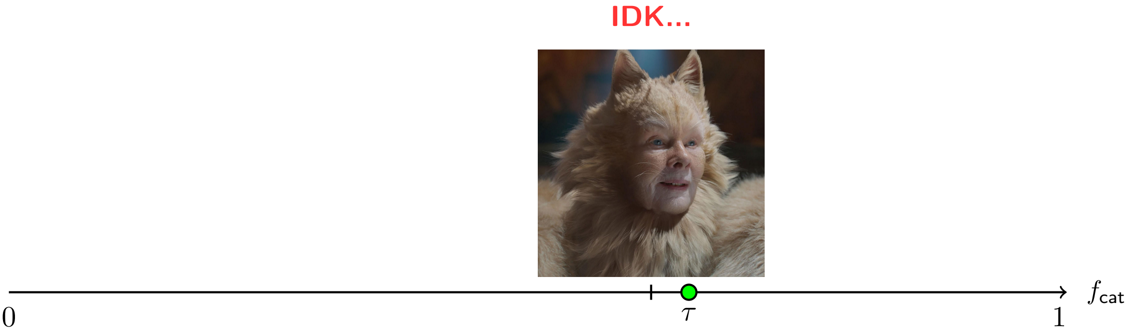
Inference Phase



Why Selective Classification?

Motivation

Inference Phase



Why Textual Entailment?

Motivation

Classification

T : *Classify the image below.*

$\hat{\mathbf{y}} = \text{cat}$

, $\mathbf{y} = \text{cat}$

- **Selective prediction** is also important to be applied to **generative tasks**.
- However, unlike exact match (EM) in **classification**, it is difficult to define a **correctness metric**.

Why Textual Entailment?

Motivation

Classification

T : *Classify the image below.*

$\hat{\mathbf{y}} =$ cat

,

$\mathbf{y} =$ cat

$\hat{\mathbf{y}} \overset{\text{EM!}}{=} \mathbf{y}$

- **Selective prediction** is also important to be applied to **generative tasks**.
- However, unlike exact match (EM) in **classification**, it is difficult to define a **correctness metric**.

Why Textual Entailment?

Motivation

Classification

T : Classify the image below.

$\hat{\mathbf{y}} =$ cat

, $\mathbf{y} =$ cat

$\hat{\mathbf{y}} \overset{\text{EM!}}{=} \mathbf{y}$

Generation

T : What is the objective of tour de france?

$$\mathcal{Y} = \left\{ \begin{array}{c} \text{bike race} \\ \text{cycle race} \\ \text{biking competition} \\ \text{to pick the best cyclist} \\ \vdots \end{array} \right\}, \quad \mathbf{y} = \text{bicycle race}$$

- **Selective prediction** is also important to be applied to **generative tasks**.
- However, unlike exact match (EM) in **classification**, it is difficult to define a **correctness metric**.

Why Textual Entailment?

Motivation

Classification

T : *Classify the image below.*

$\hat{\mathbf{y}} =$ cat

, $\mathbf{y} =$ cat

$\hat{\mathbf{y}} \overset{\text{EM!}}{=} \mathbf{y}$

Generation

T : *What is the objective of tour de france?*

$\mathcal{Y} = \left\{ \begin{array}{c} \text{bike race} \\ \text{cycle race} \\ \text{biking competition} \\ \text{to pick the best cyclist} \\ \vdots \end{array} \right\}$, $\mathbf{y} =$ bicycle race

$\hat{\mathbf{y}} \overset{\text{EM?}}{=} \mathbf{y}$

- **Selective prediction** is also important to be applied to **generative tasks**.
- However, unlike exact match (EM) in **classification**, it is difficult to define a **correctness metric**.

Why Textual Entailment?

Motivation

Classification

T : *Classify the image below.*

$\hat{\mathbf{y}} =$ cat

$, \quad \mathbf{y} =$ cat

$\hat{\mathbf{y}} \overset{\text{EM!}}{=} \mathbf{y}$

Generation

T : *What is the objective of tour de france?*

$$\mathcal{Y} = \left\{ \begin{array}{c} \text{bike race} \\ \text{cycle race} \\ \text{biking competition} \\ \text{to pick the best cyclist} \\ \vdots \end{array} \right\}, \quad \mathbf{y} = \text{bicycle race}$$

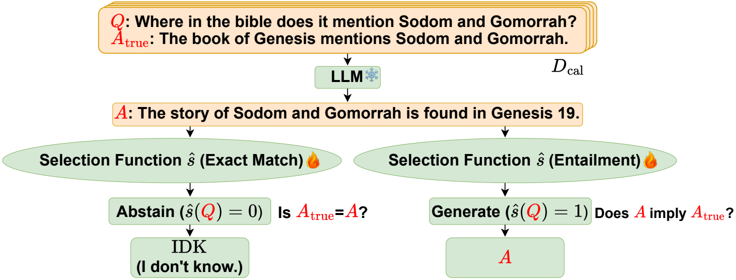
$\hat{\mathbf{y}} \overset{\text{EM?}}{=} \mathbf{y}$

- **Selective prediction** is also important to be applied to **generative tasks**.
- However, unlike exact match (EM) in **classification**, it is difficult to define a **correctness metric**.

$\Longrightarrow$ We employ ***textual entailment***: $E_{\text{true}}(\mathbf{y}) := \{\hat{\mathbf{y}} \in \mathcal{Y} \mid \hat{\mathbf{y}} \text{ implies } \mathbf{y}\}$.

Why Semi-Supervised Learning?

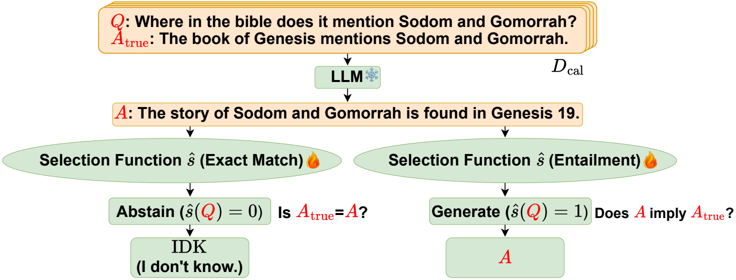
Motivation



- We can avoid *metric misalignment* in generation by leveraging entailment.
- However, labeling is expensive.

Why Semi-Supervised Learning?

Motivation

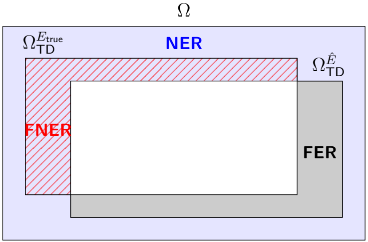


- We can avoid *metric misalignment* in generation by leveraging entailment.
- However, labeling is expensive.

$\Longrightarrow$ We leverage question-answering pairs without entailment via **semi-supervised learning** (SSL).

FDR-E Bound (1)

Method

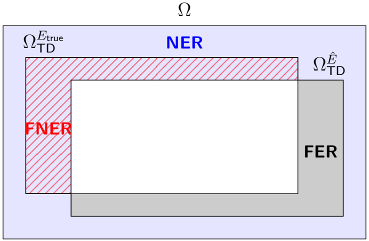


With the previously defined ***textual entailment*** $E_{\text{true}}(\mathbf{y})$, We can define **FDR-E**, the false discovery rate with respect to the textual entailment relation, as follows:

$$\mathbb{P}\{G(\mathbf{x}) \notin E_{\text{true}}(\mathbf{y}) \mid \hat{S}(\mathbf{x}) \neq \texttt{IDK}\}$$

FDR-E Bound (2)

Method

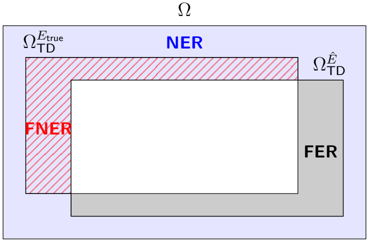


In the SSL setup, the FDR-E can be decomposed as follows:

$$\underbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{G(\mathbf{x}) \notin E_{\text{true}}(\mathbf{y})\}}_{\text{(A)}} = \underbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{v = 1\}}_{\text{(B)}} \underbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{e = 0\}}_{\text{(C)}} + \underbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{v = 0\}}_{\text{(D)}} \underbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{e = 0\}}_{\text{(E)}}$$

FDR-E Bound (2)

Method

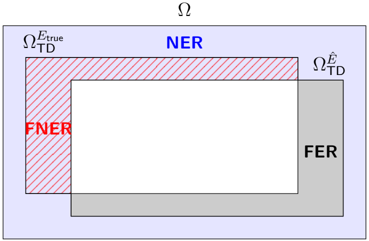


In the SSL setup, the FDR-E can be decomposed as follows:

$$\underbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{G(\mathbf{x}) \notin E_{\mathsf{true}}(\mathbf{y})\}}_{\text{(A)}} = \underbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{v = 1\}}_{\text{(B)}} \underbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{e = 0\}}_{\text{(C)}} + \underbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{v = 0\}}_{\text{(D)}} \underbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{e = 0\}}_{\text{(E)}}$$

FDR-E Bound (2)

Method



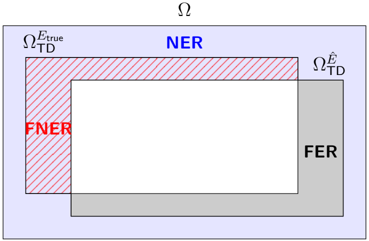
In the SSL setup, the FDR-E can be decomposed as follows:

$$\underbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{G(\mathbf{x}) \notin E_{\text{true}}(\mathbf{y})\}}_{\text{(A)}} = \underbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{v=1\}}_{\text{(B)}} \underbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{e=0\}}_{\text{(C)}} + \underbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{v=0\}}_{\text{(D)}}$$

$$\underbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{e=0\}}_{\text{(E)}}$$

FDR-E Bound (2)

Method



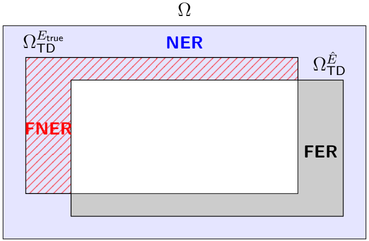
In the SSL setup, the FDR-E can be decomposed as follows:

$$\underbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{G(\mathbf{x}) \notin E_{\text{true}}(\mathbf{y})\}}_{\text{(A)}} = \underbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{v=1\}}_{\text{(B)}} \underbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{e=0\}}_{\text{(C)}} + \underbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{v=0\}}_{\text{(D)}}$$

$$\underbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{e=0\}}_{\text{(E)}}$$

FDR-E Bound (2)

Method



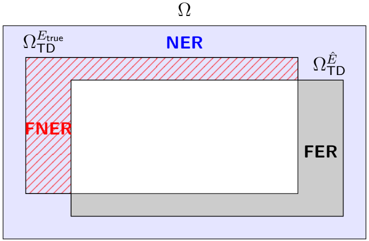
In the SSL setup, the FDR-E can be decomposed as follows:

$$\underbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{G(\mathbf{x}) \notin E_{\mathsf{true}}(\mathbf{y})\}}_{\text{(A)}} = \underbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{v = 1\}}_{\text{(B)}} \underbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{e = 0\}}_{\text{(C)}} + \underbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{v = 0\}}_{\text{(D)}}$$

$$\underbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{e = 0\}}_{\text{(E)}}$$

FDR-E Bound (2)

Method



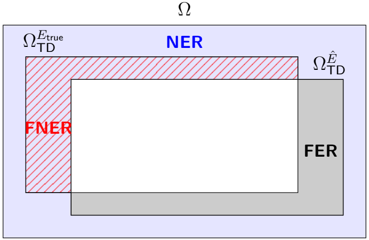
In the SSL setup, the FDR-E can be decomposed as follows:

$$\underbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{G(\mathbf{x}) \notin E_{\mathrm{true}}(\mathbf{y})\}}_{\text{(A)}} = \underbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{v = 1\}}_{\text{(B)}} \underbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{e = 0\}}_{\text{(C)}} + \underbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{v = 0\}}_{\text{(D)}}$$

$$\underbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{e = 0\}}_{\text{(E)}}$$

FDR-E Bound (2)

Method



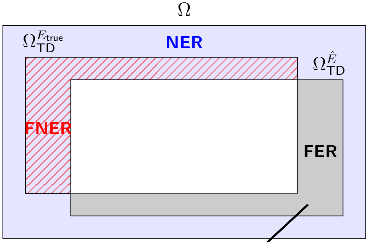
Lemma 1

(E) is decomposed as follows:

$$\underbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{e=0\}}_{\text{(E)}} = \underbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{e=0,\hat{e}=1\}}_{\text{FER}} - \underbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{e=1,\hat{e}=0\}}_{\text{FNER}} + \underbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{\hat{e}=0\}}_{\text{NER}}.$$

FDR-E Bound (2)

Method



Lemma 1

(E) is decomposed as follows:

$$\underbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{e=0\}}_{\textit{(E)}} = \underbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{e=0,\hat{e}=1\}}_{\textit{FER}} - \underbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{e=1,\hat{e}=0\}}_{\textit{FNER}} + \underbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{\hat{e}=0\}}_{\textit{NER}}.$$

$\leq \varepsilon_E$

FDR-E Bound (3)

Method

Entailment Set Learning

$\mathcal{A}_{\text{FER}}$ returns $\hat{E}$ which controls the **FER** of pseudo-labeled examples, *i.e.*,

$$\mathbb{P}\{\mathcal{R}_{\text{FER}}(\hat{E}) \le \varepsilon_E\} \ge 1 - \delta_E.$$

- We use $\hat{E}$ as a pseudo-labeling function for SSL – see our paper!

Lemma 2

If $\hat{E} := \mathcal{A}_{\mathit{FER}}(\hat{\mathbf{Z}}_E)$ *satisfies the above guarantee, we have*

$$\mathbb{P}_{\mathcal{D}}\{e = 0\} \le \varepsilon_E - L_{\text{Binom}}(\hat{k}; |\hat{\mathbf{Z}}_E|, \delta'_E/2) + U_{\text{Binom}}(\hat{l}; |\hat{\mathbf{Z}}_U|, \delta'_S) =: U_{\mathit{SSL}}$$

- We find an optimal ε_E that minimizes U_{SSL}, resulting $U_{\text{SSL}}^{\text{OPT}}$ – see our paper!

Controllable Guarantee

Method

Algorithm

Our **semi-supervised** method $\mathcal{A}_{\mathsf{SGen}}^{\mathsf{Semi}}$ solves the following optimization problem:

$$\text{find}_{\hat{S} \in \mathcal{H}} \ \hat{S} \quad \text{subj. to} \quad w_{\mathsf{SL}} U_{\mathsf{SL}} + w_{\mathsf{SSL}} U_{\mathsf{SSL}}^{\mathsf{OPT}} \leq \varepsilon_S$$

Theorem 1

$\mathcal{A}_{\mathit{SGen}}^{\mathit{Semi}}$ *satisfies the following controllable guarantee on the* FDR-E, *i.e.,*

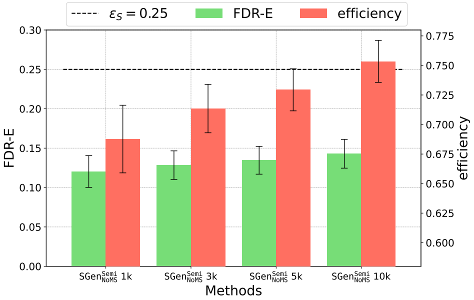
$$\mathbb{P}\left\{ \mathbb{P}\{G(\mathbf{x}) \notin E_{\mathsf{true}}(\mathbf{y}) \mid \hat{S}(\mathbf{x}) \neq \texttt{IDK}\} \leq \hat{U} \right\} \geq 1 - \delta.$$

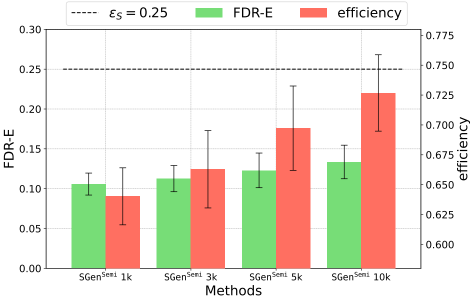
Experiment

Question $\mathbf{x}$	Who is the actor who plays Draco Malfoy?	When did the movie Benjamin Button come out?
Correct Answer $\mathbf{y}$	Thomas Andrew Felton plays Draco Malfoy in the Harry Potter movies.	The movie Benjamin Button come out December 25, 2008
Generated Answer $G(\mathbf{x})$	The actor who plays Draco Malfoy is Tom Felton. (correct)	The movie The Curious Journey of Benjamin Button was released in 2008. (correct)
$\mathsf{SGen}_{\mathsf{EM}}$	rejected	rejected
$\mathsf{SGen}^{\mathsf{Semi}}$ (ours)	accepted	accepted

- $\mathsf{SGen}^{\mathsf{Semi}}$ can capture correctness better than $\mathsf{SGen}_{\mathsf{EM}}$.

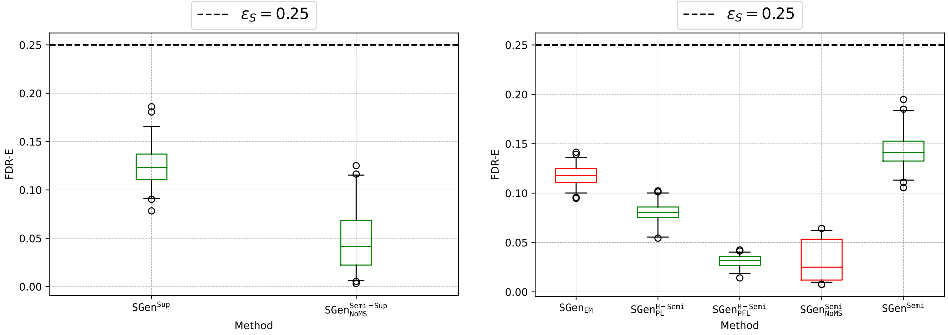
Experiment





- More unlabeled samples are beneficial to achieving better efficiency.

Experiment



- The FDR-E for $\hat{S}$ is well controlled below ε_S, desired FDR-E, under the test environment.

Conclusion

- We leverage logical **entailment** and propose a novel **semi-supervised** learning approach for **selective generation**, demonstrating its *theoretical* and *empirical* efficacy.