Selective Generation for Controllable Language Models

Minjae Lee*[1] Kyungmin Kim*[1] Taesoo Kim[2] Sangdon Park[1] (*: Equal contribution)

[1]Pohang University of Science and Technology (**POSTECH**) [2]Georgia Institute of Technology (**GaTech**)

POSTECH · Georgia Tech · NEURAL INFORMATION PROCESSING SYSTEMS

Code: https://github.com/ml-postech/selective-generation

Summary

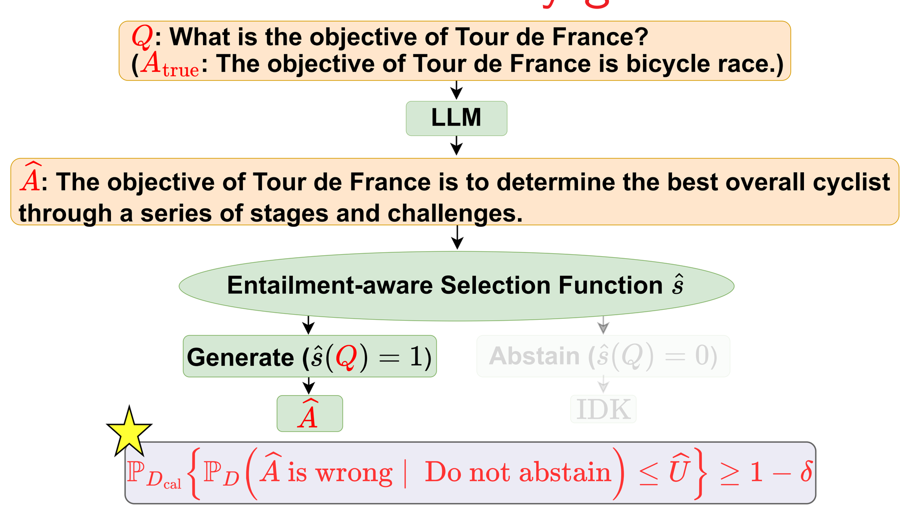
Learn an **entailment-aware selective generator** with an abstaining option that controls the "**rate of hallucination**" with a probabilistic guarantee.

Problem: Hallucination Control in Language Generation

Goal: Learn a hallucination-controllable language generator.

Contributions

1. Propose the first "certified" selective generator for language models.
2. Leverage textual entailment as a correctness metric.
3. Design a semi-supervised learning algorithm for selective generation.
4. Prove a controllability guarantee of the proposed algorithm.



$\mathbb{P}_{D_{\text{cal}}}\left\{\mathbb{P}_{\mathcal{D}}\left(\hat{A} \text{ is wrong} \mid \text{Do not abstain}\right) \le \hat{U}\right\} \ge 1-\delta$

Selective Classifier: [Geifman & El-Yaniv, 2017]

$$\hat{S}(\mathbf{X}) := \begin{cases} \hat{\mathbf{Y}} & \hat{s}(\mathbf{X}) = 1, \\ \texttt{IDK} & \text{o.w.} \end{cases}$$

Selective Generator:

$$\hat{S}(\mathbf{Q}) := \begin{cases} \hat{\mathbf{A}} & \hat{s}(\mathbf{Q}) = 1, \\ \texttt{IDK} & \text{o.w.} \end{cases}$$

Main Challenge: Metric Misalignment

Definition. Metric Misalignment

Learning Metric (*e.g.* EM) $\neq$ Evaluation Metric (*e.g.* SC)

- Example:
  - **Q**: **Where in the bible does it mention Sodom and Gomorrah?**
  - $\mathbf{A}_{\text{true}}$: **The book of Genesis mentions Sodom and Gomorrah.**
  - $\hat{\mathbf{A}}$: **The story of Sodom and Gomorrah is found in Genesis 19.**
- A standard learning metric on correct answers, *i.e.* Exact Match (EM), assumes a single correct answer (*i.e.* $\hat{\mathbf{A}} =_{\text{EM}} \mathbf{A}_{\text{true}}$?)
- As $\hat{\mathbf{A}} \neq_{\text{EM}} \mathbf{A}_{\text{true}}$, $\hat{\mathbf{A}}$ is wrong even if it is semantically correct (SC) ☹.

Idea 1. Textual Entailment as a Correctness Metric

Definition. Correctness Metric by Entailment

A generated answer $\hat{\mathbf{A}}$ is correct if

$$\hat{\mathbf{A}} \in E_{\text{true}}(\mathbf{A}_{\text{true}}) := \{\tilde{\mathbf{A}} \mid \tilde{\mathbf{A}} \text{ entails } \mathbf{A}_{\text{true}}\}.$$

Definition. False Discovery Rate with Entailment (FDR-E)

Learning Metric: $\mathbb{P}_{\mathcal{D}}\left(\hat{\mathbf{A}} \notin E_{\text{true}}(\mathbf{A}_{\text{true}}) \,\middle|\, \hat{S}(\mathbf{Q}) \neq \texttt{IDK}\right)$

- We find a learning algorithm to control the FDR-E.

Idea 2. Pseudo-labeling for Textual Entailment

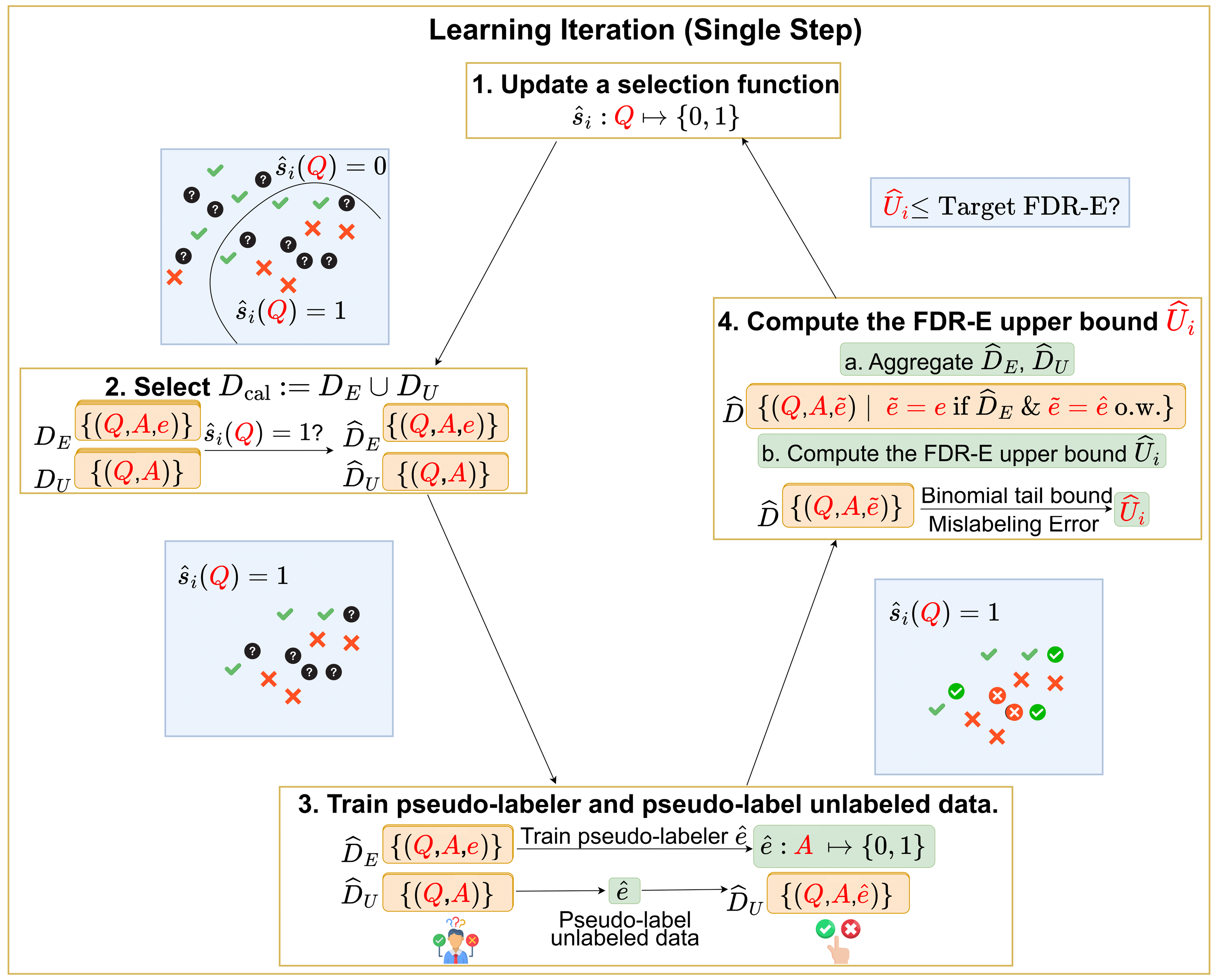
Calibration Set

$$\{\underbrace{(\mathbf{Q}, \mathbf{A}_{\text{true}}, \hat{\mathbf{A}} \in E_{\text{true}}(\mathbf{A}_{\text{true}})}_{\text{additional labels}}\} \cup \{\underbrace{(\mathbf{Q}, \mathbf{A}_{\text{true}}, \hat{\mathbf{A}} \in \hat{E}(\mathbf{A}_{\text{true}}))}_{\text{pseudo labels}}\}$$

- We propose a label efficient semi-supervised learning algorithm.

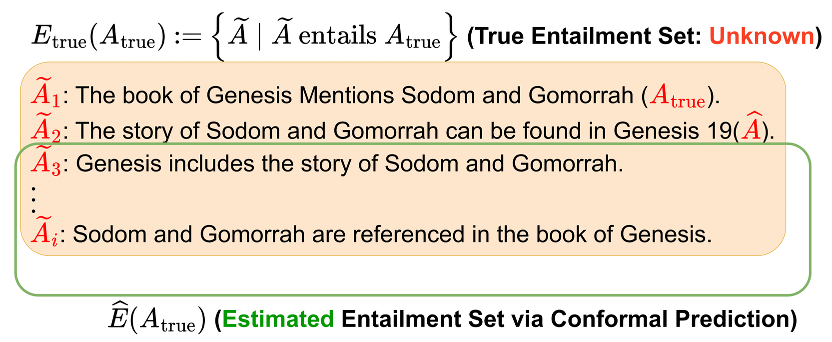
Solution: Semi-Supervised Selective Generator Learning

Algorithm $\texttt{SGen}^{\texttt{Semi}}$ to find a selective generator $\hat{S}$



Details on the Pseudo-Labeling Function $\hat{E}$

- We measure the estimation error by $\hat{E}$ to upper-bound the FDR-E.



FDR-E (for SSL)

$$\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{e=0\} = \overbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{e=0, \hat{e}=1\}}^{\text{false entailment rate}} - \underbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{e=1, \hat{e}=0\}}_{\text{false neg.-entailment rate}} + \underbrace{\mathbb{P}_{\mathcal{D}_{\hat{S}}}\{\hat{e}=0\}}_{\text{neg. entailment rate}}$$

- $e := \hat{\mathbf{A}} \in E_{\text{true}}(\mathbf{A}_{\text{true}})$
- $\hat{e} := \hat{\mathbf{A}} \in \hat{E}(\mathbf{A}_{\text{true}})$
- $\mathbb{P}_{\mathcal{D}_{\hat{S}}}(\cdot) := \mathbb{P}_{\mathcal{D}}(\cdot \mid \hat{S}(\mathbf{Q}) \neq \texttt{IDK})$

Theoretical Result

Theorem. Controllability Guarantee on the FDR-E

For any LLMs and downstream language generation tasks, the following model-agnostic and task-free controllability guarantee holds:

$$\mathbb{P}_{D_{\text{cal}}}\left\{\overbrace{\mathbb{P}_{\mathcal{D}}\big(\underbrace{\hat{\mathbf{A}} \notin E_{\text{true}}(\mathbf{A}_{\text{true}})}_{\hat{\mathbf{A}} \text{ is "wrong"}} \,\big|\, \underbrace{\hat{S}(\mathbf{Q}) \neq \texttt{IDK}}_{\text{Do not abstain}}\big)}^{\text{FDR-E}} \le \hat{U}\right\} \ge 1-\delta,$$

where δ is the confidence level and $(\hat{s}, \hat{U})$ is the algorithm output.

Experiments & Results

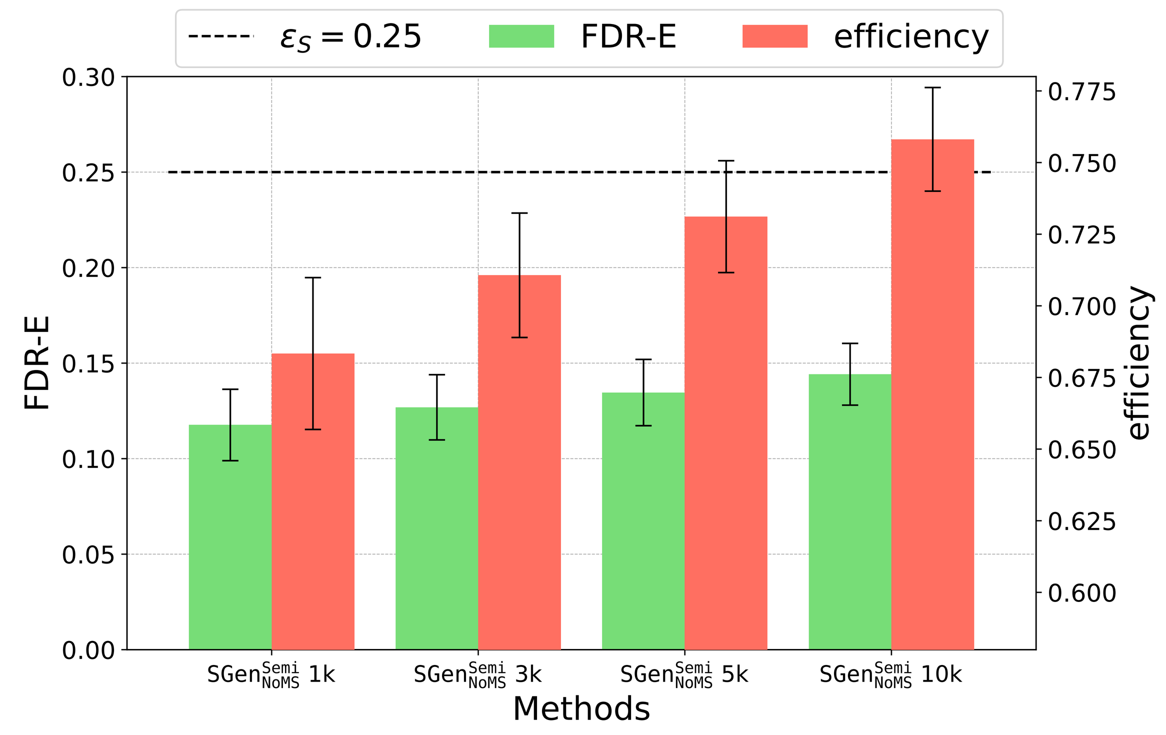
1. Benefit of Textual Entailment

- Our entailment-based learning metric shows better selection efficiency.
  - Selection efficiency: the proportion of non-abstained samples

$\mathbf{Q}$	Who is the actor that plays Draco Malfoy?	When did the movie Benjamin Button come out?
$\mathbf{A}_{\text{true}}$	Thomas Andrew Felton plays Draco Malfoy in the Harry Potter movies.	The movie Benjamin Button come out December 25, 2008.
$\hat{\mathbf{A}}$	The actor who plays Draco Malfoy is Tom Felton. (correct)	The Curious Journey of Benjamin Button was released in 2008. (correct)
EM (Baseline)	rejected	rejected
Textual Entailment (Ours)	accepted	accepted

2. Benefit of Semi-Supervised Learning

- As we use more unlabeled data, selection efficiency gets better.



(a) $\texttt{SGen}^{\texttt{Semi}}_{\texttt{NoMS}}$ on GPT-3.5-turbo

3. Benefit of Neuro-Selection Function

- Learning a selection function combination improves selection efficiency.

Models		GPT-3.5-turbo		Alpaca-7B	
Methods		$\texttt{SGen}^{\texttt{Semi}}_{\texttt{NoMS}}$(x)	$\texttt{SGen}^{\texttt{Semi}}$(o)	$\texttt{SGen}^{\texttt{Semi}}_{\texttt{NoMS}}$(x)	$\texttt{SGen}^{\texttt{Semi}}$(o)
f_{M_1}	FDR-E	$\underline{0.0609}$	0.1589	$\underline{0.0359}$	0.0685
	efficiency	$\underline{0.2829}$	0.7334	$\underline{0.1580}$	0.3173
f_{M_2}	FDR-E	0.1785	0.1589	0.0698	0.0685
	efficiency	0.7835	0.7334	0.3200	0.3173
average efficiency		0.5347	**0.7334**	0.2390	**0.3173**

More in Our Paper

- How the mislabeling error in pseudo-labeling is controlled via conformal prediction & affects the FDR-E bound computation.
- Supervised-learning algorithm as a special case of the proposed semi-supervised learning algorithm.
- Defining and choosing a "good" scoring function in terms of designing a single-threshold selection function.