Semi-supervised Selective Generator Learning for Trustworthy Language Generation

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

[1]Graduate School of Artificial Intelligence, Pohang University of Science and Technology

[2]School of Computer Science, Georgia Institute of Technology

Motivation

- Trustworthy language generation is crucial for the deployment of large language models (LLMs) in critical decision-making systems.
- Hallucination is one of the main bottlenecks toward trustworthy language generation.
- Given an LLM, our main objective is to **control the rate of hallucination** to the target level with a theoretical guarantee.

Related Work: Selective Classification

- Selective prediction is a principled way of controlling the error rate to the target level in **supervised learning**.
- Given a generator, a selective predictor (1) returns "I don't know" (IDK) on the input that the model is uncertain, and (2) controls the error rate on predicted outputs.
- Geifman and Yaniv (2017) applies the selective prediction to classification tasks, where the learned selective classifier $\hat{S}$ controls the misclassification error $\mathcal{R}_{\text{EM}}(\hat{S})$ of the classifier $\hat{y}$ on test data with theoretical guarantee.

$$\mathcal{R}_{\text{EM}}(\hat{S}) := \mathbb{P}_{(x,y)\sim\mathcal{D}}\{\hat{y} \neq y \mid \hat{S}(x) \neq \texttt{IDK}\}$$

- However, unlike the supervised set-up, **generation problems** lack an appropriate metric for correctness evaluation – **metric misalignment**

Question (x)	Who played George Hazard's wife in North and South?	What is the setting of the story of Robin Hood?
Correct Answer (y)	Wendy Kilbourne	Sherwood Forest
Generated Answer (G(x))	Lesley-Anne Down played George Hazard's wife in North and South. (wrong)	The story of Robin Hood is set in medieval England, in the Sherwood Forest. (correct)
SG-EM	accepted	rejected
CSGen-MS (ours)	rejected	accepted

Table 1: Selective generation examples of SG-EM and CSGen-MS using GPT-3.5-turbo

Main Contribution: Addressing Metric Misalignment via Textual Entailment

- The textual entailment relation R_E is a subset of ordered pairs of declarative sequences $(\mathrm{y}', \mathrm{y}) \in \mathcal{Y} \times \mathcal{Y}$ as follows:

$$(\mathrm{y}', \mathrm{y}) \in R_E \text{ if } \mathrm{y}' \text{ implies } \mathrm{y}.$$

- Then, given a reference answer y that is true given the input sequence x, the correctness of the generated sequence $G(\mathrm{x})$ can be evaluated by an entailment set function E_{true} defined as follows:

$$E_{\text{true}}(\mathrm{y}) := \{\mathrm{y}' \in \mathcal{Y} \mid (\mathrm{y}', \mathrm{y}) \in R_E\}. \quad (1)$$

Prolem: Selective Generation

- Given a generator G, a selective generator $\hat{S}$ consists of the generator and the selection function pair $(G, \hat{s})$ as follows:

$$\hat{S}(\mathrm{x}) := \begin{cases} G(\mathrm{x}) & \text{if } \hat{s}(\mathrm{x}) = 1 \\ \texttt{IDK} & \text{otherwise} \end{cases}$$

  - ▹ A common choice of $\hat{s}$ is a single-threshold indicator function based on an uncertainty measure $f_M : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ called scoring function as follows:

$$\hat{s}(\mathrm{x}) := \mathbb{1}(f_M(\mathrm{x}, G(\mathrm{x})) \geq \tau). \quad (2)$$

- Our main goal reduces to learn $\hat{S}$ (τ if we consider (2)) that controls FDR-E $\mathcal{R}_{R_E}(\hat{S})$, which is defined based on (1) as follows:

$$\mathcal{R}_{R_E}(\hat{S}) := \mathbb{P}\{G(\mathrm{x}) \notin E_{\text{true}}(\mathrm{y}) \mid \hat{S}(\mathrm{x}) \neq \texttt{IDK}\},$$

requiring expensive human annotations on $e := \mathbb{1}(G(\mathrm{x}) \in E_{\text{true}}(\mathrm{y}))$.

  - ▹ Leveraging PAC prediction set learning algorithm, we fully exploit the unlabeled data $\mathbf{Z}_U$ in learning $\hat{S}$ by estimating an entailment set function $\hat{E} : \mathcal{Y} \to 2^{\mathcal{Y}}$, which pseudo-labels the entailment relation and satisfies

$$\mathbb{P}_{\mathbf{Z}}\{\mathbb{P}_{(\mathrm{x},\mathrm{y},e)\sim\mathcal{D}}\{e = 0 \wedge \hat{e} = 1 \mid \hat{S}(\mathrm{x}) \neq \texttt{IDK}\} \leq \epsilon_E\} \geq 1 - \delta_E,$$

where $\hat{e} := \mathbb{1}(G(\mathrm{x}) \in \hat{E}(\mathrm{y}))$.

CSGen-MS: Semi-supervised Selective Generator Learning with Model Selection

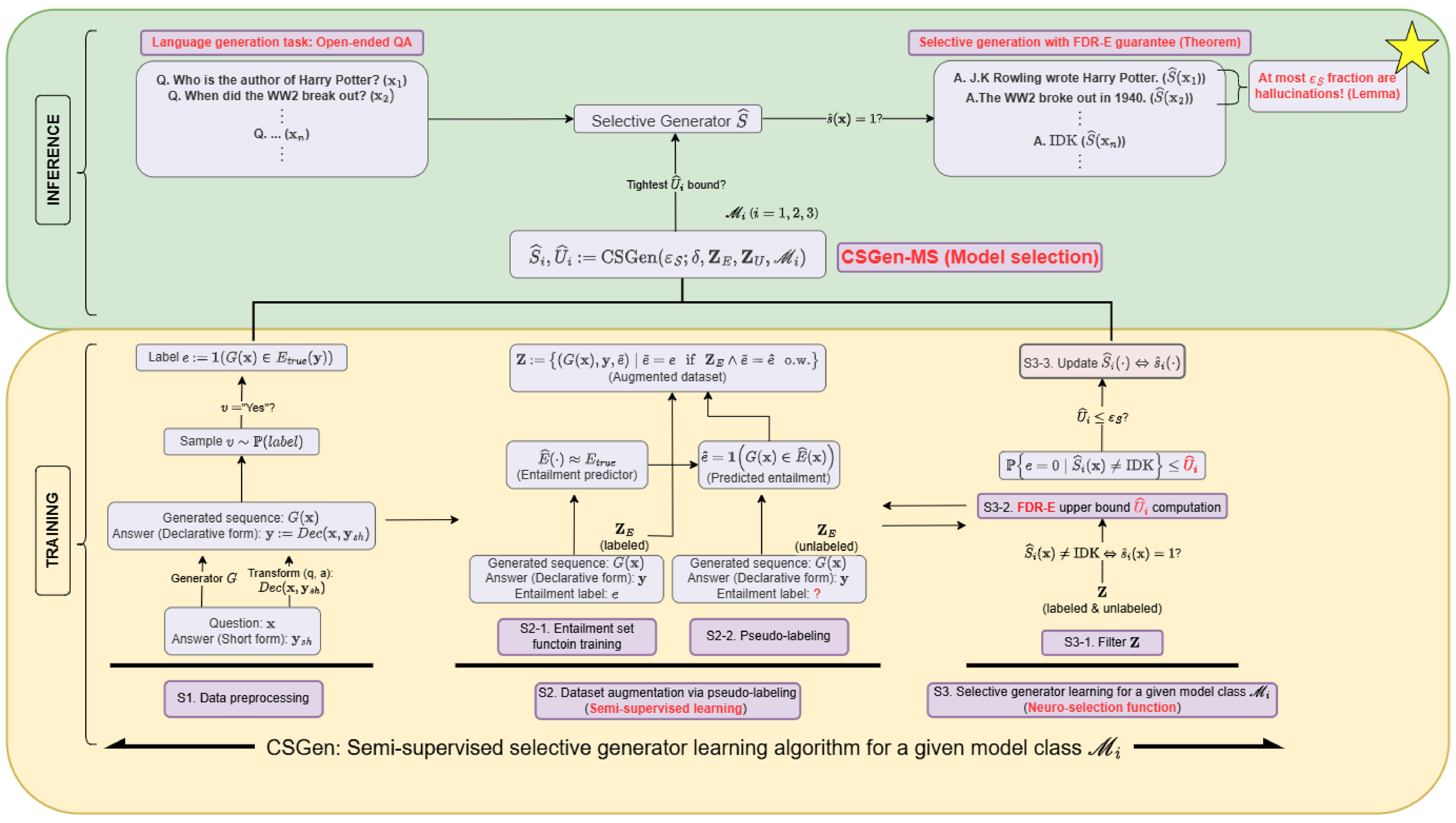


Figure 1: Training and inference phase of CSGen-MS

Controllability Guarantee on the Rate of Hallucination

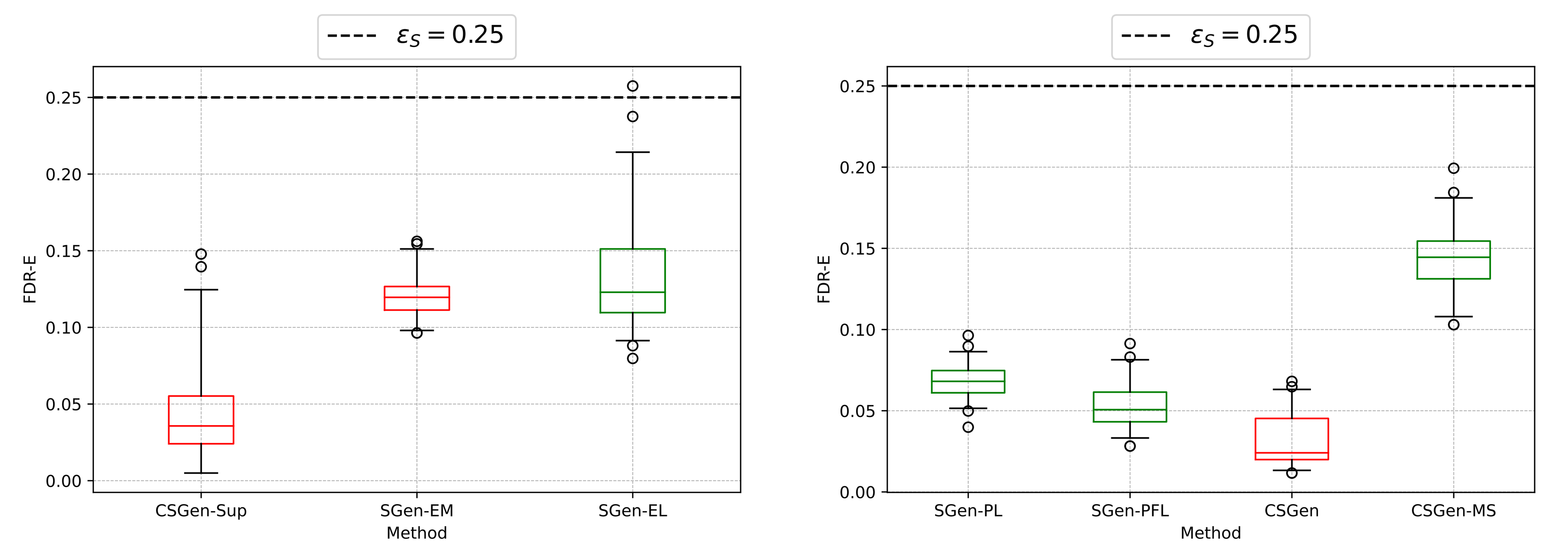
Theorem. $\mathcal{A}_{\text{CSGen-MS}}$ satisfies the following guarantee on FDR-E as follows:

$$\mathbb{P}\Big\{\mathbb{P}\{G(\mathrm{x}) \notin E_{\text{true}}(\mathrm{y}) \mid \hat{S}(\mathrm{x}) \neq \texttt{IDK}\} \leq \epsilon_S \mathbb{1}(\hat{U} \leq \epsilon_S) + \hat{U}\mathbb{1}(\hat{U} > \epsilon_S)\Big\} \geq 1 - \delta.$$

Lemma (SC for Perfect Controllability). If the estimated entailment set function $\hat{E}$ well separates entailment labels as E_{true}, and f_M is perfectly calibrated with respect to $\hat{E}$, FDR-E is monotonically non-increasing in τ_S.

Experiment

- In Figure 3, we can see that the error rate(**FDR-E**), we want to control, is well controlled under the user-defined value ϵ_S.



(a) Supervised methods (b) Semi-supervised methods

Figure 2: Box plots of FDR-E by selective generator learning algorithms using GPT-3.5-turbo

- In Table 2, our method CSGen-MS can overall achieve desired FDR-E guarantees with better **efficiency** compared to baselines.
  - ▹ efficiency: the ratio of data selected in the test set

Model		GPT-3.5-turbo				Alpaca-7B			
Method		Heuristic		Certified		Heuristic		Certified	
		SGen-PL	SGen-PFL	CSGen	CSGen-MS	SGen-PL	SGen-PFL	CSGen	CSGen-MS
f_{M_1}	FDR-E	0.0565	0.0449	<u>0.0216</u>	0.1611	0.0047	<u>0.0041</u>	<u>0.0278</u>	0.0142
	Efficiency	0.3472	0.2741	<u>0.1412</u>	0.8422	0.0305	<u>0.0271</u>	<u>0.1186</u>	0.1532
f_{M_2}	FDR-E	0.1561	0.1844	0.1645	0.1611	0.0393	0.0454	0.0149	0.0142
	Efficiency	0.8339	0.8904	0.8488	0.8422	0.2759	0.2936	0.1634	0.1532
Average Efficiency		0.5906	0.5823	-	**0.8422**	**0.1532**	-	-	**0.1532**

Table 2: FDR-E and selection efficiency by selective generator learning algorithms on two LLMs The best results are highlighted in **bold** and results from methods that do not satisfy ϵ-guarantee are <u>underlined</u>.

Limitation

- PAC guarantee of CSGen-MS on FDR-E bound depends on the i.i.d. assumption.
- Despite its generalizability and cost-efficiency, the applicability of CSGen-MS depends on the quality of a entailment classifier on the language generation task that the user considers.
  - ▹ As every selective prediction method does, CSGen-MS also depends on the quality of a scoring function.
    - ◦ Future work: Designing a learning algorithm on a general class of *neuro-selection functions*

contact E-mail:sangdon@postech.ac.kr [37673] 77, Cheongam-ro, Nam-gu, Pohang-si, Gyeongsangbuk-do, Korea