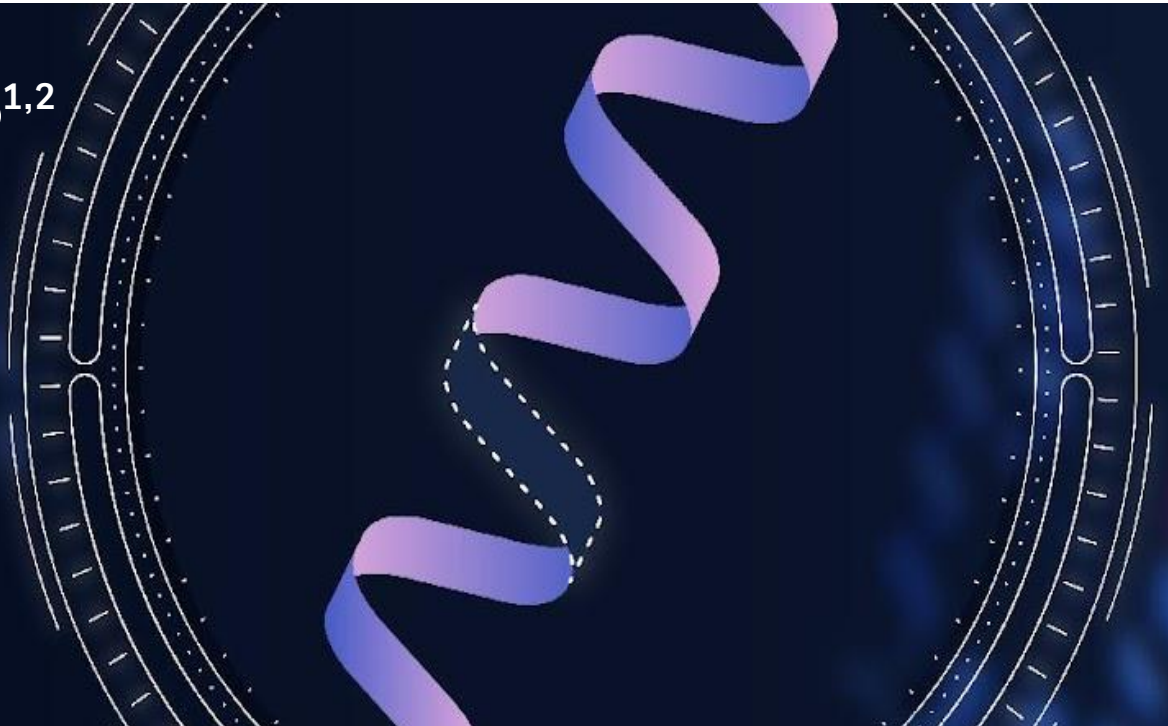
Prescient Design
A Genentech Accelerator

Genentech
A Member of the Roche Group

Efficient semantic uncertainty quantification in language models via diversity-steered sampling

Ji Won Park[1], Kyunghyun Cho[1,2]

[1] Prescient Design, Genentech
[2] Center for Data Science, NYU



Motivation

Prescient Design
A Genentech Accelerator

Estimating uncertainty in language generation is critical for building reliable AI systems but often relies on drawing large IID sample sets.

Q: What is a way to measure the Hubble constant?

A: Use **Cepheid variable stars.** Cluster 1

A: By analyzing the distances to **Cepheid variable stars.**

A: By comparing the brightness and redshift of Type Ia **supernovae.** Cluster 2

A: By measuring the distances to Type Ia **supernovae** in different galaxies.

A: The Hubble constant can be measured using the light curves of Type Ia **supernovae,** which have a constant maximum brightness allowing for distance estimation.

A: The method of standard candles involves using **supernovae** as "candles" to measure distances and then infer the Hubble constant from redshift data.

A: One way to measure the Hubble constant is by using the distance-luminosity relationship for Type Ia **supernovae.**

Motivation

Prescient Design
A Genentech Accelerator

For example, **semantic entropy** (Kuhn et al., 2023) groups generations into semantically equivalent clusters and computes the entropy over the clusters.

Q: What is a way to measure the Hubble constant?

A: Use **Cepheid variable stars.**

A: By analyzing the distances to **Cepheid variable stars.**

Cluster 1

$$p(c \mid x, \theta) = \int 1[y \in c \mid x, \theta]\, p(y \mid x, \theta) dy$$

$$H\big(p(c \mid x, \theta)\big) = -\sum_{c \in \mathcal{C}} p(c \mid x, \theta) \log p(c \mid x, \theta)$$

A: By comparing the brightness and redshift of Type Ia **supernovae.**

A: By measuring the distances to Type Ia **supernovae** in different galaxies.

A: The Hubble constant can be measured using the light curves of Type Ia **supernovae,** which have a constant maximum brightness allowing for distance estimation.

A: The method of standard candles involves using **supernovae** as "candles" to measure distances and then infer the Hubble constant from redshift data.

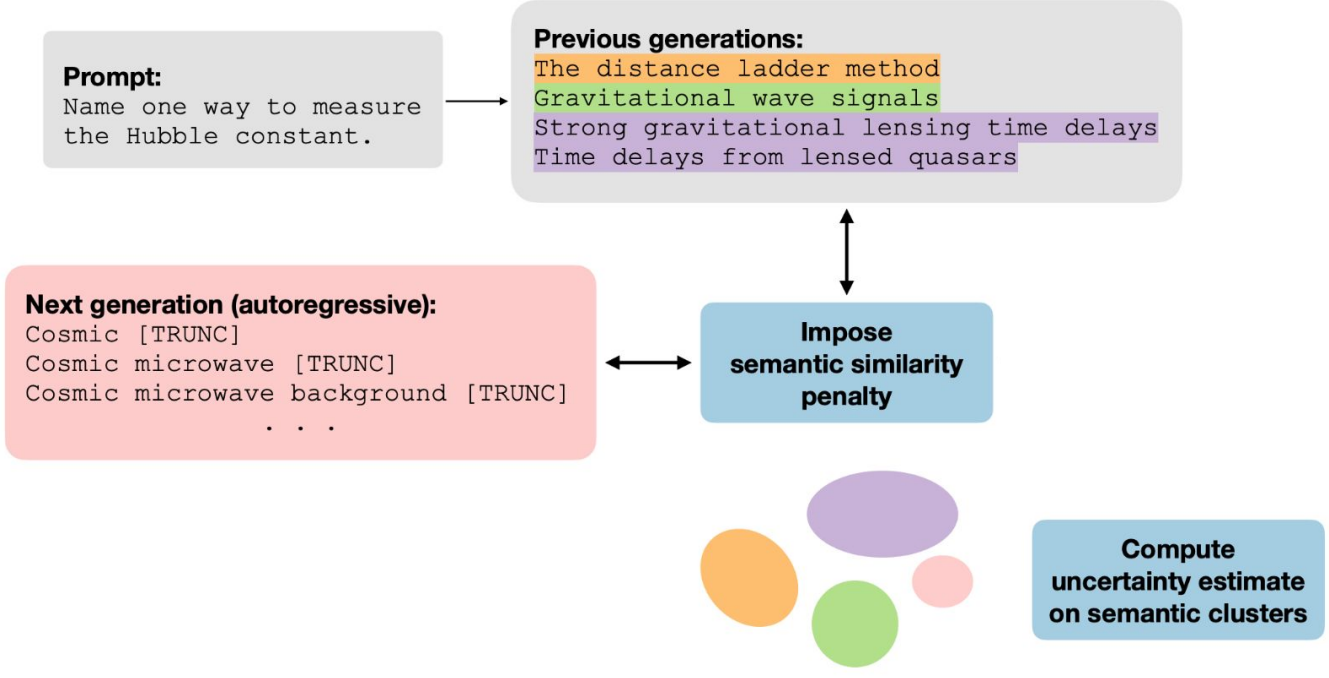
A: One way to measure the Hubble constant is by using the distance-luminosity relationship for Type Ia **supernovae.**

Cluster 2

Solution: expose diverse semantic clusters

Prescient Design
A Genentech Accelerator

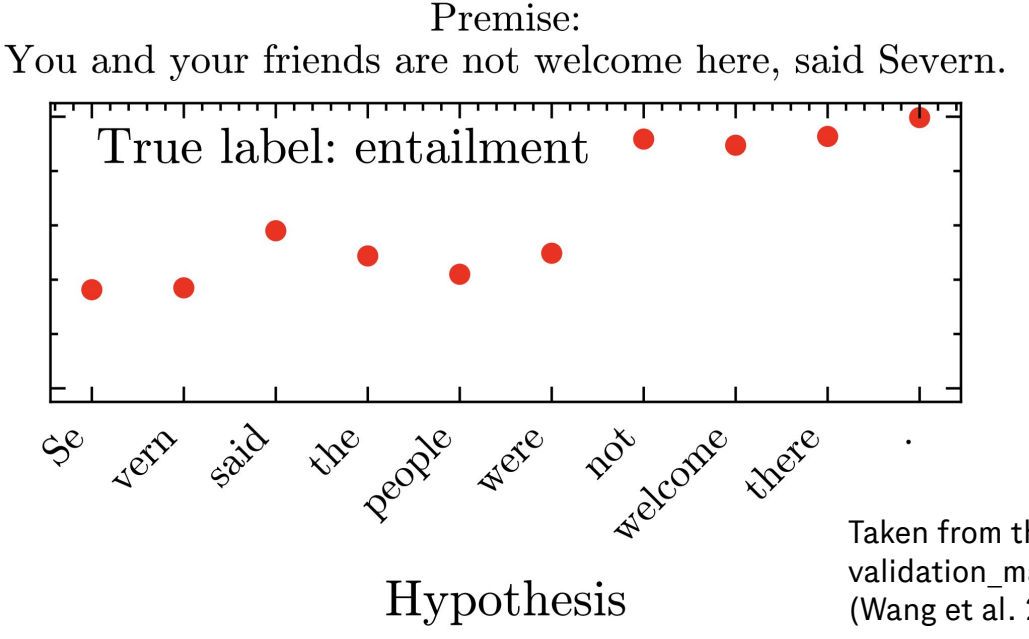
We propose a **diversity-steered sampler** that actively discourages outputs that are semantically redundant to previous generations.



The main ingredient

Prescient Design
A Genentech Accelerator

We introduce a new [TRUNC] token and lightly **finetune** a natural language inference (NLI) model on **partial prefixes** of either the premise or hypothesis at all levels of truncation.



Taken from the GLUE MNLI validation_matched split (Wang et al. 2018)

Key idea: tilting the token-level conditional distributions

Prescient Design
A Genentech Accelerator

The **bidirectional entailment score** predicted by the finetuned NLI model is used to tilt the sampling distribution away from previous generations.

partial generation and existing sample s assumed to be semantically equivalent if they entail each other

$$E(y_{\leq t}, s) = {}^1\!/_2 \left(\texttt{entailment}(y_{\leq t}, s) + \texttt{entailment}(s, y_{\leq t})\right)$$

$$\log p(y_t | y_{<t}) - \lambda \max_{s \in \mathcal{S}} E(y_{\leq t}, s)$$

original distribution for token t

repel sampling away from the most similar existing generation

Importance weighting

Prescient Design
A Genentech Accelerator

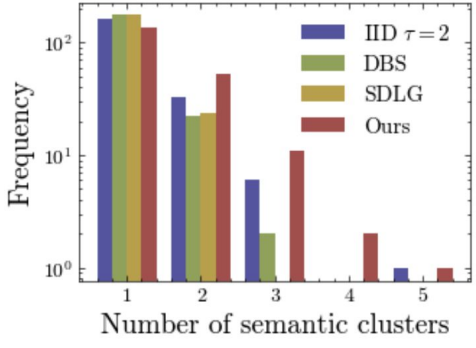
Sampling from the tilted distribution exposes more semantic clusters with the same number of generations. For estimating semantic entropy downstream, we correct the induced sampling bias with self-normalized importance weights.

Prompt:
Name one way to measure the Hubble constant.

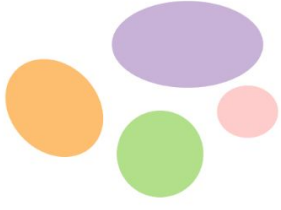
Previous generations:
The distance ladder method
Gravitational wave signals
Strong gravitational lensing time delays
Time delays from lensed quasars

Next generation (autoregressive):
Cosmic [TRUNC]
Cosmic microwave [TRUNC]
Cosmic microwave background [TRUNC]
. . .

Impose semantic similarity penalty



$$w_i = \frac{p\left(s^{(i)}\right)}{q\left(s^{(i)}\right)}$$ original / tilted



Compute uncertainty estimate on semantic clusters

$$\tilde{w}_i = \frac{w_i}{\sum_{j=1}^{N} w_j}$$

Results

Prescient Design
A Genentech Accelerator

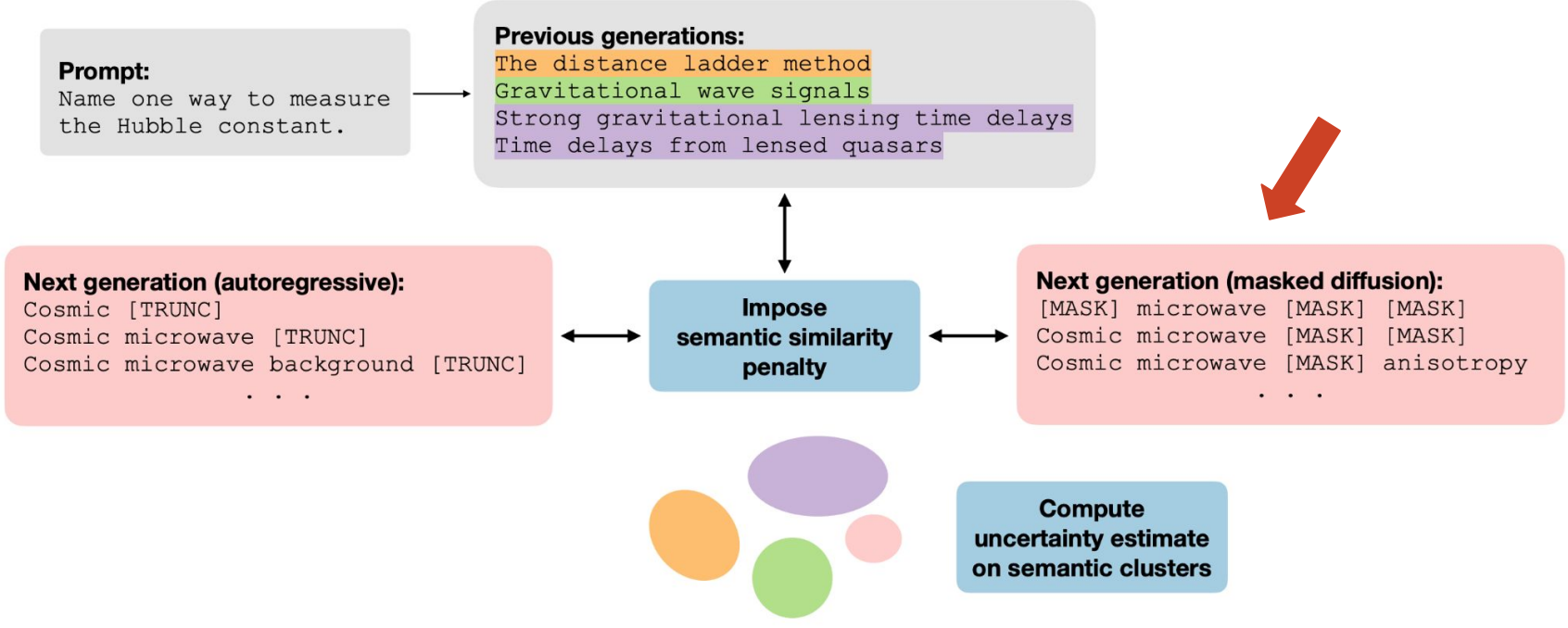
Across four QA benchmarks, our method achieves comparable or better AUROC relative to baselines, including temperature scaling, diverse beam search (Vijayakumar et al., 2018), and SDLG (Aichberger et al., 2025).

Dataset	Model	Vanilla ($\tau = 1$)	$\tau = 2$	DBS [39]	SDLG [4]	Ours
CoQA	OPT-6.7B	.59±.06	.69±.04	.68±.04	.71±.02	**.75±.02**
	OPT-13B	.70±.04	**.76±.04**	.73±.04	.72±.02	.75±.03
	LLaMA 3 8B-Instruct	.68±.03	.72±.04	.71±.05	.74±.02	**.77±.02**
	LLaDA 8B-Instruct	.78±.02	**.81±.05**	-	-	**.81±.04**
TriviaQA	OPT-6.7B	.66±.05	.67±.06	.71±.04	.78±.03	**.82±.03**
	OPT-13B	.72±.04	.70±.05	.73±.04	**.86±.03**	.85±.03
	LLaMA 3 8B-Instruct	.79±.04	.70±.04	.70±.03	.79±.04	**.84±.03**
	LLaDA 8B-Instruct	.81±.11	.83±.05	-	-	**.86±.04**
AmbigQA	OPT-13B	.65±.10	.68±.11	**.78±.08**	.71±.08	**.78±.04**
	LLaMA 3 8B-Instruct	.70±.04	.55±.07	.71±.08	**.77±.05**	.76±.03
	LLaDA 8B-Instruct	.70±.09	.71±.08	-	-	**.76±.03**
TruthfulQA	OPT-6.7B	.80±.04	.80±.05	.77±.02	.78±.06	**.81±.06**
	OPT-13B	.73±.06	.74±.08	.79±.05	.81±.04	**.85±.04**
	LLaMA 3 8B-Instruct	.88±.04	.88±.05	**.89±.04**	.86±.04	**.89±.02**
	LLaDA 8B-Instruct	.85±.04	.89±.04	-	-	**.94±.02**

Both autoregressive and masked diffusion paradigms

Prescient Design
A Genentech Accelerator

Our method also applies to **masked diffusion** models, by similarly finetuning the NLI model on intermediate diffusion states (masked spans).



Summary

Prescient Design
A Genentech Accelerator

- We propose an importance sampling framework that actively promotes semantic novelty with respect to previous generations during decoding.
- The key idea is to inject a token-level semantic similarity penalty into the proposal distribution. An NLI model, finetuned minimally with a new [TRUNC] token for autogressive models, enables live semantic scoring.
- For downstream uncertainty estimation, the induced sampling bias is corrected with self-normalized importance weighting.
- Across four QA benchmarks, our method matches or surpasses baselines while covering more semantic clusters with the same number of samples.
- Our method is modular, requires no gradient access to the base LLM, and applies to both autoregressive and masked diffusion models.

Thank you!

References

Aichberger, Lukas, et al. "Improving uncertainty estimation through semantically diverse language generation." *ICLR* (2025).

Kuhn, Lorenz, Yarin Gal, and Sebastian Farquhar. "Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation." *ICLR* (2023).

Vijayakumar, Ashwin K., et al. "Diverse beam search: Decoding diverse solutions from neural sequence models." *AAAI* (2018).

Wang, Alex, et al. "GLUE: A multi-task benchmark and analysis platform for natural language understanding." *EMNLP workshop BlackboxNLP* (2018).