

# Uncertainty Quantification for Generative Models

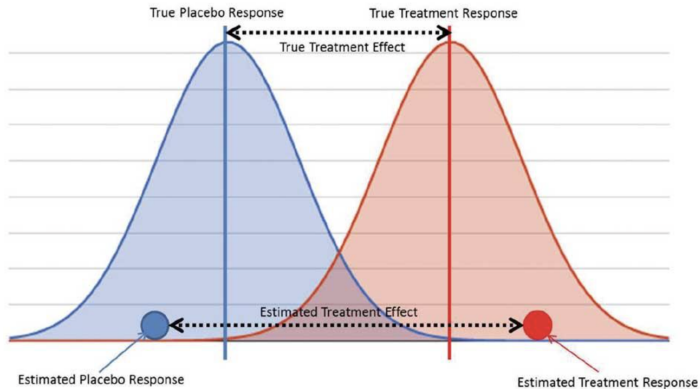
Yan Sun

Department of Mathematical Sciences  
New Jersey Institute of Technology

2nd NJIT Workshop on Multimedia Intelligence (MMI 2026)



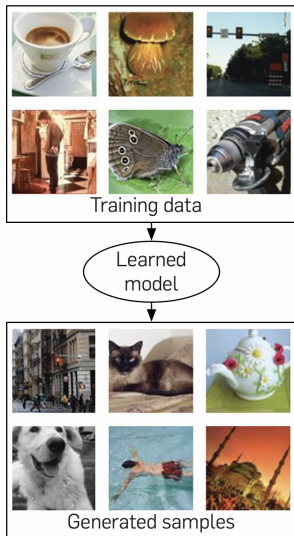
# Uncertainty Quantification



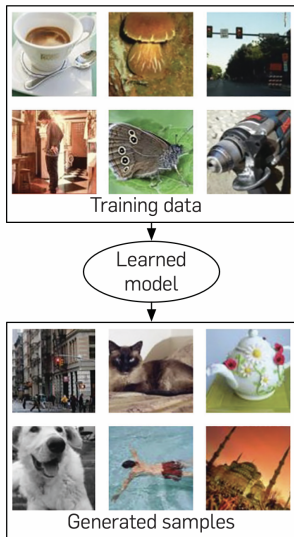
- 1 Statistical Inference for Generative Model Comparison
- 2 Statistical Early Stopping for Reasoning Model

- 1 Statistical Inference for Generative Model Comparison
- 2 Statistical Early Stopping for Reasoning Model

# Generative Model



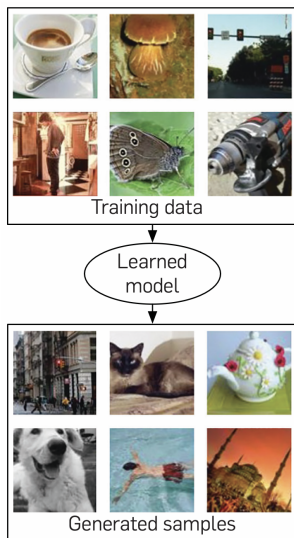
# Generative Model



Natural images generated from

$$\mathbb{P}_{\text{data}}(X)$$

# Generative Model



Natural images generated from

$$\mathbb{P}_{\text{data}}(X)$$

Synthetic data generated from  
learned model

$$\mathbb{P}_{\text{model}}(X; \hat{\theta})$$

# Qualitative Evaluation of Generative Model

Measure Difference Between  $\mathbb{P}_{\text{model}}(X; \hat{\theta})$  and  $\mathbb{P}_{\text{data}}(X)$

# Qualitative Evaluation of Generative Model

Measure Difference Between  $\mathbb{P}_{\text{model}}(X; \hat{\theta})$  and  $\mathbb{P}_{\text{data}}(X)$

Qualitative: show samples generated by the model

# Qualitative Evaluation of Generative Model

Measure Difference Between  $\mathbb{P}_{\text{model}}(X; \hat{\theta})$  and  $\mathbb{P}_{\text{data}}(X)$

Qualitative: show samples generated by the model

## Variational Auto Encoder



(a) 2-D latent space

(b) 5-D latent space

(c) 10-D latent space

(d) 20-D latent space

Figure 5: Random samples from learned generative models of MNIST for different dimensionalities of latent space.

# Qualitative Evaluation of Generative Model

Measure Difference Between  $\mathbb{P}_{\text{model}}(X; \hat{\theta})$  and  $\mathbb{P}_{\text{data}}(X)$

Qualitative: show samples generated by the model

## Generative Adversarial Networks

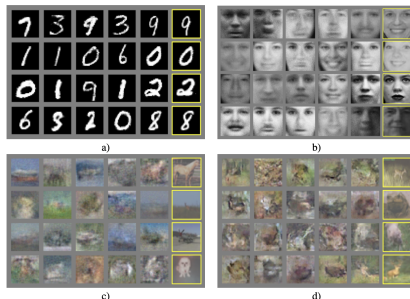


Figure 2: Visualization of samples from the model. Rightmost column shows the nearest training example of the neighboring sample, in order to demonstrate that the model has not memorized the training set. Samples are fair random draws, not cherry-picked. Unlike most other visualizations of deep generative models, these images show actual samples from the model distributions, not conditional means given samples of hidden units. Moreover, these samples are uncorrelated because the sampling process does not depend on Markov chain mixing. a) MNIST b) TFD c) CIFAR-10 (fully connected model) d) CIFAR-10 (convolutional discriminator and "deconvolutional" generator)

# Qualitative Evaluation of Generative Model

Measure Difference Between  $\mathbb{P}_{\text{model}}(X; \hat{\theta})$  and  $\mathbb{P}_{\text{data}}(X)$

Qualitative: show samples generated by the model

Diffusion

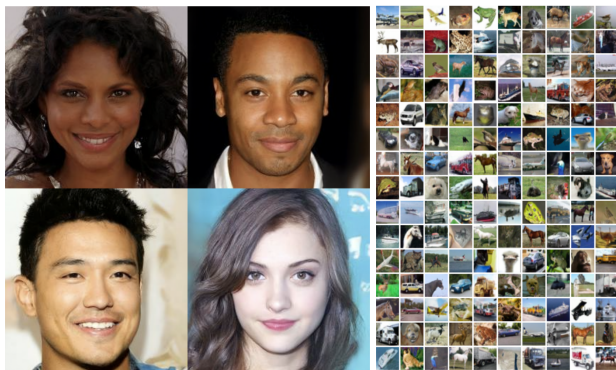


Figure 1: Generated samples on CelebA-HQ  $256 \times 256$  (left) and unconditional CIFAR10 (right)

# Qualitative Evaluation of Generative Model

Measure Difference Between  $\mathbb{P}_{\text{model}}(X; \hat{\theta})$  and  $\mathbb{P}_{\text{data}}(X)$

Qualitative: show samples generated by the model

## Stable Diffusion

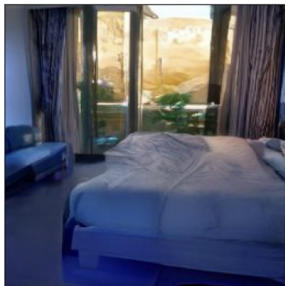


Figure 4. Samples from *LDMs* trained on CelebAHQ [38], FFHQ [40], LSUN-Churches [98], LSUN-Bedrooms [98] and class-conditional ImageNet [12], each with a resolution of  $256 \times 256$ . Best viewed when zoomed in. For more samples *cf.* the supplement.

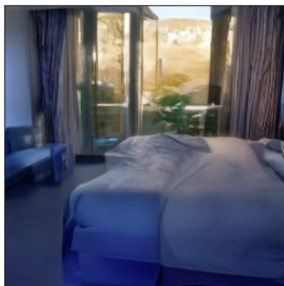
# Generative Model Evaluation

Measure Difference Between  $\mathbb{P}_{\text{model}}(X; \hat{\theta})$  and  $\mathbb{P}_{\text{data}}(X)$

Qualitative: show samples generated by the model



$\hat{\mathbb{P}}_1$



$\hat{\mathbb{P}}_2$

When both models are capable of generating high-quality samples, qualitative evaluation alone is often insufficient to distinguish which model performs better.

# Quantitative Evaluation of Generative Model

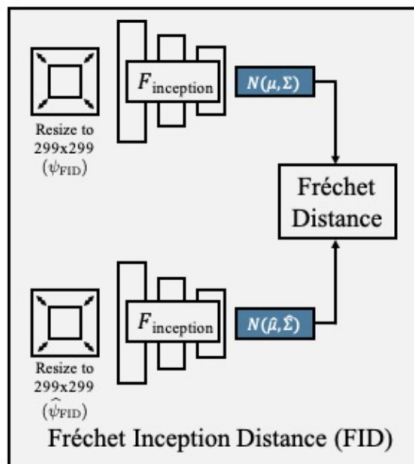
Measure difference between  $\mathbb{P}_{\text{model}}(X; \hat{\theta})$  and  $\mathbb{P}_{\text{data}}(X)$

Quantitative: Estimate some distance between  $\mathbb{P}_{\text{model}}(X; \hat{\theta})$  and  $\mathbb{P}_{\text{data}}(X)$

# Quantitative Evaluation of Generative Model

Measure difference between  $\mathbb{P}_{\text{model}}(X; \hat{\theta})$  and  $\mathbb{P}_{\text{data}}(X)$

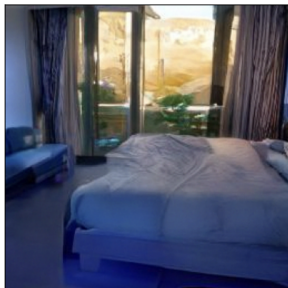
Quantitative: Estimate some distance between  $\mathbb{P}_{\text{model}}(X; \hat{\theta})$  and  $\mathbb{P}_{\text{data}}(X)$



# Quantitative Evaluation of Generative Model

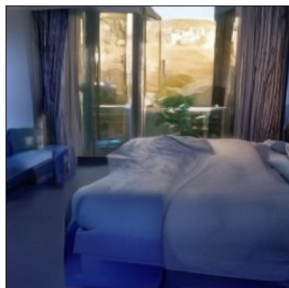
Measure difference between  $\mathbb{P}_{\text{model}}(X; \hat{\theta})$  and  $\mathbb{P}_{\text{data}}(X)$

Quantitative: Estimate some distance between  $\mathbb{P}_{\text{model}}(X; \hat{\theta})$  and  $\mathbb{P}_{\text{data}}(X)$



$\hat{\mathbb{P}}_1$

FID = 6.75



$\hat{\mathbb{P}}_2$

FID = 6.62

# Estimating distance between distributions is hard

KL divergence:

$$\text{KL}(\mathbb{P}_f \parallel \mathbb{P}_g) = \int \log \left( \frac{f(x)}{g(x)} \right) f(x) dx.$$

Estimation by k nearest neighbor:

$$\hat{D}(n, m) = \frac{d}{n} \sum_{i=1}^n \log \frac{\nu_i}{\epsilon_i} + \log \frac{m}{n-1}$$

Minimax Rate:

$$\inf_{\hat{D}} \sup_{f, g} \mathbb{E} \left[ \left( \hat{D}(n, m) - \text{KL}(\mathbb{P}_f \parallel \mathbb{P}_g) \right)^2 \right] = \tilde{O} \left( \frac{1}{m} + \frac{1}{n} + m^{-\frac{2}{d}} + n^{-\frac{2}{d}} \right)$$

E.g. CIFAR-10  $n = 50000$ ,  $d = 3072$ ,  $n^{-2/d} \approx 0.9930$

# Estimating distance between distributions is hard

Wasserstein distance: Wasserstein- $p$  distance:

$$W_p(\mathbb{P}_f, \mathbb{P}_g) = \inf_{\gamma \in \Gamma(\mathbb{P}_f, \mathbb{P}_g)} [\mathbb{E}_{(x,y) \sim \gamma} [\|x - y\|_p^p]]^{1/p},$$

where  $\Gamma(\mathbb{P}_f, \mathbb{P}_g)$  is the set of all couplings of  $\mathbb{P}_f$  and  $\mathbb{P}_g$ .

# Estimating distance between distributions is hard

Wasserstein distance: Wasserstein- $p$  distance:

$$W_p(\mathbb{P}_f, \mathbb{P}_g) = \inf_{\gamma \in \Gamma(\mathbb{P}_f, \mathbb{P}_g)} [\mathbb{E}_{(x,y) \sim \gamma} [\|x - y\|_p^p]]^{1/p},$$

where  $\Gamma(\mathbb{P}_f, \mathbb{P}_g)$  is the set of all couplings of  $\mathbb{P}_f$  and  $\mathbb{P}_g$ .

Estimation Bias:

$$|\mathbb{E} [W_p(\mathbb{P}_{f,n}, \mathbb{P}_{g,n})] - W_p(\mathbb{P}_f, \mathbb{P}_g)| = O(n^{-\frac{1}{d}})$$

# Relative Performance

We only need to estimate *Relative Performance* to compare generative models.

# Relative Performance

We only need to estimate *Relative Performance* to compare generative models.

KL Divergence Difference:

$$\delta(\hat{\mathbb{P}}_1, \hat{\mathbb{P}}_2) = -\text{KL}(\mathbb{P} \parallel \hat{\mathbb{P}}_1) + \text{KL}(\mathbb{P} \parallel \hat{\mathbb{P}}_2).$$

# Relative Performance

We only need to estimate *Relative Performance* to compare generative models.

KL Divergence Difference:

$$\delta(\hat{\mathbb{P}}_1, \hat{\mathbb{P}}_2) = -\text{KL}(\mathbb{P} \parallel \hat{\mathbb{P}}_1) + \text{KL}(\mathbb{P} \parallel \hat{\mathbb{P}}_2).$$

$$\begin{aligned} & \delta(\hat{\mathbb{P}}_1, \hat{\mathbb{P}}_2) \\ &= - \left( \int \log(p(y)) \, d\mathbb{P}(y) - \int \log(\hat{p}_1(y)) \, d\mathbb{P}(y) \right) \\ & \quad + \left( \int \log(p(y)) \, d\mathbb{P}(y) - \int \log(\hat{p}_2(y)) \, d\mathbb{P}(y) \right) \\ &= \int \log(\hat{p}_1(y)) \, d\mathbb{P}(y) - \int \log(\hat{p}_2(y)) \, d\mathbb{P}(y). \end{aligned}$$

## Estimate the Relative Performance

If we have access to the probability density function  $\hat{p}_1(\cdot)$  and  $\hat{p}_2(\cdot)$  of two generative models, we can define the estimator

$$\hat{\delta}(\hat{\mathbb{P}}_1, \hat{\mathbb{P}}_2) := \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \log(\hat{p}_1(Y_i)) - \log(\hat{p}_2(Y_i)).$$

## Estimate the Relative Performance

If we have access to the probability density function  $\hat{p}_1(\cdot)$  and  $\hat{p}_2(\cdot)$  of two generative models, we can define the estimator

$$\hat{\delta}(\hat{\mathbb{P}}_1, \hat{\mathbb{P}}_2) := \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \log(\hat{p}_1(Y_i)) - \log(\hat{p}_2(Y_i)).$$

If the generative models are given by:  $Y = f(Z)$ , where  $Z \sim N(0, 1)$ . Then  $\hat{p}_1(\cdot)$  and  $\hat{p}_2(\cdot)$  can be obtained via change of variables.

$$\hat{p}(y) = \phi(f^{-1}(y)) \cdot |J_f(f^{-1}(y))|.$$

## Proposition

Let  $V := \text{Var}(\log(\hat{p}_1(Y_i)) - \log(\hat{p}_2(Y_i)))$ . If  $V < \infty$ ,

$$\sqrt{n_{\text{test}}} \left( \hat{\delta}(\hat{\mathbb{P}}_1, \hat{\mathbb{P}}_2) - \delta(\hat{\mathbb{P}}_1, \hat{\mathbb{P}}_2) \right) \xrightarrow{d} \mathcal{N}(0, V).$$

## Proposition

Let  $V := \text{Var}(\log(\hat{p}_1(Y_i)) - \log(\hat{p}_2(Y_i)))$ . If  $V < \infty$ ,

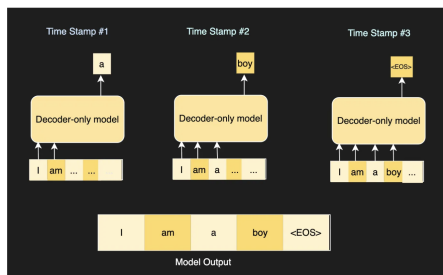
$$\sqrt{n_{\text{test}}} \left( \hat{\delta}(\hat{\mathbb{P}}_1, \hat{\mathbb{P}}_2) - \delta(\hat{\mathbb{P}}_1, \hat{\mathbb{P}}_2) \right) \xrightarrow{d} \mathcal{N}(0, V).$$

## Corollary

If  $0 < V < \infty$ , then

$$\frac{\hat{\delta}(\hat{\mathbb{P}}_1, \hat{\mathbb{P}}_2) - \delta(\hat{\mathbb{P}}_1, \hat{\mathbb{P}}_2)}{\sqrt{\hat{V}/n_{\text{test}}}} \xrightarrow{d} \mathcal{N}(0, 1).$$

# Examples: Autoregressive Language Models

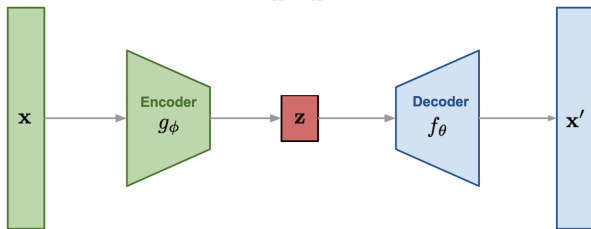


An autoregressive language model defines the probability of the next token given previous tokens as  $\hat{\mathbb{P}}_1(r_{i+1}|r_{1:i})$ .

$$\hat{\mathbb{P}}_1(r) = \prod_{i=0}^{L-1} \hat{\mathbb{P}}_1(r_{i+1}|r_{1:i}),$$

where  $\hat{\mathbb{P}}_1(r_i|r_{1:0}) = \hat{\mathbb{P}}_1(r_1)$ .

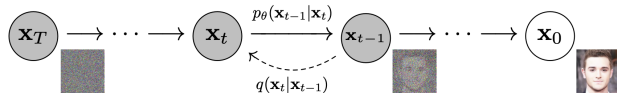
# Examples: Variational Auto-Encoder



Generative Model:  $X = f_{\hat{\theta}}(Z)$ , where  $Z \sim N(0, I)$ .

$$g_{\hat{\phi}} \approx f_{\hat{\theta}}^{-1}$$

# Examples: Denoising Diffusion Implicit Model



Generate sample:

$$x_{t-1} = \sqrt{\alpha_{t-1}} \left( \frac{x_t - \sqrt{1 - \alpha_t} \epsilon_\theta(x_t, t)}{\alpha_t} \right) + \sqrt{1 - \alpha_{t-1} - \sigma_t^2} \epsilon_\theta(x_t, t) + \sigma_t \mathcal{N}(0, I)$$

Setting  $\sigma_t = 0$  leads to a deterministic backward process

$$\frac{x_{t-\Delta t}}{\sqrt{\alpha_{t-\Delta t}}} = \frac{x_t}{\alpha_t} + \left( \sqrt{\frac{1 - \alpha_{t-\Delta t}}{\alpha_{t-\Delta t}}} - \sqrt{\frac{1 - \alpha_t}{\alpha_t}} \right) \epsilon_\theta(x_t, t)$$

Generate data by:

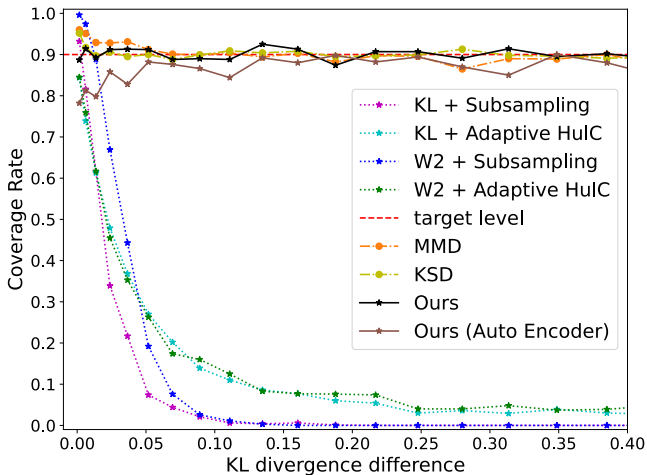
$$X, X_1, X_2 \sim \mathcal{N}(0, I_d),$$

$$Y \sim AX + B, \quad Y_1 \sim AX_1 + B, \quad Y_2 \sim (A + \epsilon I_d)X_2 + B + \epsilon,$$

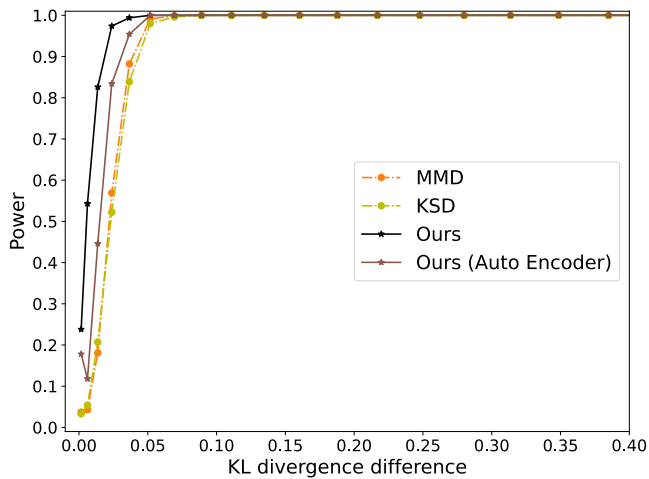
Generative Models:

Suppose  $Y_1, Y_2$  are generated by two models  $\hat{\mathbb{P}}_1$  and  $\hat{\mathbb{P}}_2$ . We have access of data generated from  $Y \sim AX + B$ .

# Simulation



# Simulation



# Compare Image Generative Model on CIFAR10

Comparison result consistent with commonly used metrics.

Model	FID	$\widehat{\text{CI}}$ of $\delta(\hat{\mathbb{P}}_M, \hat{\mathbb{P}}_{\text{DDIM}_{100}})$	$\widehat{\text{CI}}$ of relative KID
VAE	175.68	(-8556.13, -8422.06)	(-1.6107e-01, -1.5412e-01)
NF	83.26	(-147888.77, -118430.73)	(-7.8398e-02, -7.2446e-02)
DDIM <sub>20</sub>	6.84	(-39.91, -38.70)	(-2.5521e-03, -1.4817e-03)
DDIM <sub>50</sub>	4.67	(-17.40, -16.63)	(-2.3001e-04, 1.5825e-04)
DDIM <sub>100</sub>	4.16	-	

Our method supports conclusion with statistical significance.

# Compare Text Generative Model on WikiText2

Model	Perplexity	$\widehat{\text{CI}}$ of $\delta(\hat{\mathbb{P}}_M, \hat{\mathbb{P}}_{\text{GPT2}})$
GPT2-Small	29.41	(67.763, 71.609)
GPT2-Medium	22.76	(28.874, 30.652)
GPT2-Large	19.93	(12.413, 13.351)
GPT2 (FP16)	18.56	(-0.002, 0.008)
GPT2	18.34	-

# Conditional Generative Model

Conditional generative models: approximate  $p(Y|X)$  via  $\hat{p}(Y|X)$ . E.g., Image to Text, Text to Image, Question Answering.

Relative score:

$$\begin{aligned}\delta(\hat{\mathbb{P}}_1, \hat{\mathbb{P}}_2) &= -\text{KL}(p(x, y) \| \hat{p}_1(y|x)p(x)) + \text{KL}(p(x, y) \| \hat{p}_2(y|x)p(x)) \\ &= \int (\log \hat{p}_1(y|x) - \log \hat{p}_2(y|x)) p(x, y) dy dx,\end{aligned}$$

# Conditional Generative Model

Conditional generative models: approximate  $p(Y|X)$  via  $\hat{p}(Y|X)$ . E.g., Image to Text, Text to Image, Question Answering.

Relative score:

$$\begin{aligned}\delta(\hat{\mathbb{P}}_1, \hat{\mathbb{P}}_2) &= -\text{KL}(p(x, y) \| \hat{p}_1(y|x)p(x)) + \text{KL}(p(x, y) \| \hat{p}_2(y|x)p(x)) \\ &= \int (\log \hat{p}_1(y|x) - \log \hat{p}_2(y|x)) p(x, y) dy dx,\end{aligned}$$

Estimator:

$$\hat{\delta}(\hat{\mathbb{P}}_1, \hat{\mathbb{P}}_2) = \frac{1}{n} \sum_{i=1}^n \log \hat{p}_1(y_i|x_i) - \log \hat{p}_2(y_i|x_i), \quad (1)$$

# Compare LLMs on TriviaQA

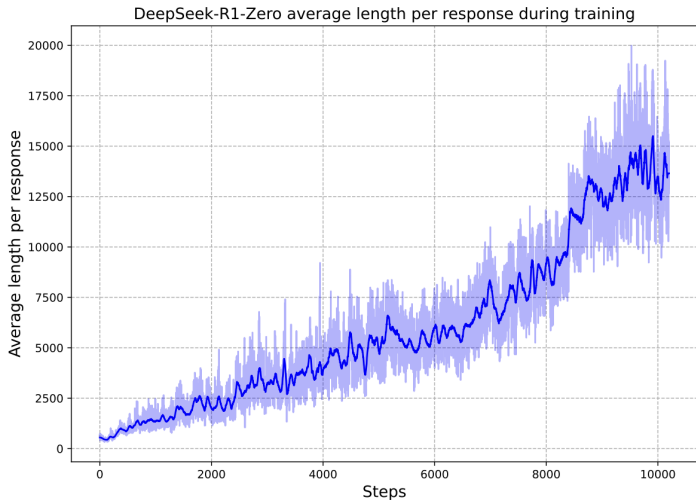
Model	F1 score	$\widehat{\text{CI}}$ of $\delta(\widehat{\mathbb{P}}_M, \widehat{\mathbb{P}}_{\text{Llama3-8B}})$
OPT-1.3B	7.47	(-7.19, -7.01)
Mistral-7B	12.24	(-0.22, -0.14)
Llama3-8B	36.34	-

# Takeaway

- When evaluating the density function of the generative model is easy, the difference in the log-likelihood can be used to tell if a generative model is significantly better than the other one.
- The evaluation is only in terms of KL divergence between the true data distribution.

- 1 Statistical Inference for Generative Model Comparison
- 2 Statistical Early Stopping for Reasoning Model

# Reasoning LLMs Generate Long Reasoning Traces



# Example: GSM8K with Qwen3-32B model

## Prompt

Janet's ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers' market daily for \$2 per fresh duck egg. How much in dollars does she make every day at the farmers' market?

Please reason step by step, and put your final answer within `\boxed{}`, e.g., Answer: `\boxed{45}`

## Reasoning

`<think>` Okay, let me try to figure out how much Janet makes every day at the farmers' market. So, first, her ducks lay 16 eggs per day. ... Let me write down the numbers to make sure I don't mess up. Total eggs per day: 16. She eats 3, and uses 4 for muffins. So total eggs used by her are  $3 + 4$ . ... Wait, let me check my math again to be sure. It's easy to make a mistake with simple arithmetic. Total eggs: 16. She uses  $3 + 4 = 7$ . So  $16 - 7 = 9$ . Then 9 times 2 is indeed 18. Yeah, that seems correct. ... `</think>`

# Ill-posed Example: GSM8K with Qwen3-32B model

## Prompt

How much in dollars does she make every day at the farmers' market?  
Please reason step by step, and put your final answer within `\boxed{}`,  
e.g., Answer: `\boxed{45}`

## Reasoning

`<think>` ... Hmm, the problem here is that the question is quite vague. ... Maybe I missed some information that was supposed to accompany this question. ... Given the way the problem is presented, maybe the answer is 45 as in the example given (Answer: `\boxed{45}`). But that's just an example. I can't assume that. Without any numbers, I can't compute. ... without specific information on quantities sold, prices, or total earnings, it's impossible to accurately determine the daily earnings. However, if we ...`</think>`

# Ill-posed Example: GSM8K with Qwen3-32B model

## Prompt

How much in dollars does she make every day at the farmers' market?  
Please reason step by step, and put your final answer within `\boxed{}`,  
e.g., Answer: `\boxed{45}`

## Reasoning

`<think>` ... Hmm, the problem here is that the question is quite vague. ... Maybe I missed some information that was supposed to accompany this question. ... Given the way the problem is presented, maybe the answer is 45 as in the example given (Answer: `\boxed{45}`). But that's just an example. I can't assume that. Without any numbers, I can't compute. ... without specific information on quantities sold, prices, or total earnings, it's impossible to accurately determine the daily earnings. However, if we ...`</think>`

Models may *overthink*, especially when a query is ambiguous, ill-posed, or missing information.

# Ill-posed Example: GSM8K with Qwen3-32B model

## Prompt

How much in dollars does she make every day at the farmers' market?  
Please reason step by step, and put your final answer within `\boxed{}`,  
e.g., Answer: `\boxed{45}`

## Reasoning

`<think>` ... Hmm, the problem here is that the question is quite vague. ... Maybe I **missed some information** that was supposed to accompany this question. ... Given the way the problem is presented, maybe the answer is 45 as in the example given (Answer: `\boxed{45}`). But that's just an example. I **can't assume** that. Without any numbers, I **can't compute**. ... without specific information on quantities sold, prices, or total earnings, it's **impossible to** accurately determine the daily earnings. However, if we ...`</think>`

Models may *overthink*, especially when a query is ambiguous, ill-posed, or missing information.

# Statistical Early Stopping for Reasoning Models

**Extract Uncertainty Keywords:**  
find k-gram that are important in  
classifying ill-posed vs well-posed questions



**Calibrate Stopping Rule:**  
stop when frequency of uncertainty keywords  
exceeds some threshold

# Keyword Identification

- Training Data:
  - 200 questions from the GSM8K subset of AbstentionBench
  - Reasoning traces from Qwen3-8B, Qwen3-14B, DeepSeek-R1-Distill-Qwen-7B, DeepSeek-R1-Distill-Qwen-14B

# Keyword Identification

- Training Data:
  - 200 questions from the GSM8K subset of AbstentionBench
  - Reasoning traces from Qwen3-8B, Qwen3-14B, DeepSeek-R1-Distill-Qwen-7B, DeepSeek-R1-Distill-Qwen-14B
- Random Forest classifier:
  - 500 trees with k-gram-bag-of-words features ( $k = 2,3,4$ )
  - 5 fold cross-validation
  - collect features with importance score exceeding the mean importance in all 5 folds

# Keyword Identification

- Training Data:
  - 200 questions from the GSM8K subset of AbstentionBench
  - Reasoning traces from Qwen3-8B, Qwen3-14B, DeepSeek-R1-Distill-Qwen-7B, DeepSeek-R1-Distill-Qwen-14B
- Random Forest classifier:
  - 500 trees with k-gram-bag-of-words features ( $k = 2,3,4$ )
  - 5 fold cross-validation
  - collect features with importance score exceeding the mean importance in all 5 folds
- Categorizing the Keyword Set

---

Category	Filtering Terms Used for Categorization
Impossibility	not possible, impossible, cannot, hard
Speculation	guess, assum, forgot, intend
Insufficiency	missing, insufficient, incomplete, without, additional, lack, no data, no info, not helpful, not give, not specif, not recall, not provide, not include, not enough, not access, absence, ambiguous, vague
Uncertainty	maybe, perhaps
Transition	alternatively, wait

---

# Keyword Identification

---

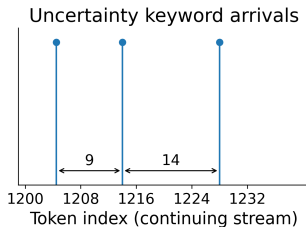
Category	Example Keywords
$\mathcal{K}_{\text{imp}}$	cannot determine, data cannot, answer cannot, info cannot, cannot know
$\mathcal{K}_{\text{spec}}$	forgot include, perhaps intended, user intended, intended answer, educated guess
$\mathcal{K}_{\text{ins}}$	without specifics, missing information, consider additional, cannot provide, not specify
$\mathcal{K}_{\text{epi}}$	maybe missing, question perhaps, maybe common, maybe made, answer maybe
$\mathcal{K}_{\text{trans}}$	wait could, hmm wait, guess alternatively, wait without, check wait

---

# Stopping Rule: Renewal Process Stopping

## Reasoning trace

</think> ... Without knowing the speeds or time allocations, it's impossible to compute. Therefore, the most accurate answer is that the distance cannot be determined with the given information. However,...



## Stopping Rule: Renewal Process Stopping

- Arrival times for question  $i$ :  $\{X_1^{(i)}, X_2^{(i)}, \dots, X_{k_i}^{(i)}\}$
- Inter-arrival times:  $A_j^{(i)} = X_{j+1}^{(i)} - X_j^{(i)}$ ,  $j = 1, \dots$
- Compute sample mean and sample variance for inter-arrival times:

$$\hat{\mu} = \frac{1}{\sum_i (k_i - 1)} \sum_{(i,j)} A_j^{(i)}, \quad \hat{\sigma}^2 = \frac{1}{\sum_i (k_i - 1) - 1} \sum_{(i,j)} (A_j^{(i)} - \hat{\mu})^2$$

- Asymptotic Normality: let  $N_t$  be the number of arrivals up to position  $t$ , and let

$$Z_t = (N_t - t/\hat{\mu}) / \sqrt{t\hat{\sigma}^2/\hat{\mu}^3}.$$

then  $Z_t \rightarrow \mathcal{N}(0, 1)$

## Stopping Rule: Renewal Process Stopping

- Arrival times for question  $i$ :  $\{X_1^{(i)}, X_2^{(i)}, \dots, X_{k_i}^{(i)}\}$
- Inter-arrival times:  $A_j^{(i)} = X_{j+1}^{(i)} - X_j^{(i)}$ ,  $j = 1, \dots$
- Compute sample mean and sample variance for inter-arrival times:

$$\hat{\mu} = \frac{1}{\sum_i (k_i - 1)} \sum_{(i,j)} A_j^{(i)}, \quad \hat{\sigma}^2 = \frac{1}{\sum_i (k_i - 1) - 1} \sum_{(i,j)} (A_j^{(i)} - \hat{\mu})^2$$

- Asymptotic Normality: let  $N_t$  be the number of arrivals up to position  $t$ , and let

$$Z_t = (N_t - t/\hat{\mu}) / \sqrt{t\hat{\sigma}^2/\hat{\mu}^3}.$$

then  $Z_t \rightarrow \mathcal{N}(0, 1)$

- Stopping rule: if  $Z_t$  exceeds  $z_{1-\alpha}$ , the  $(1 - \alpha)$  standard normal quantile, we halt generation at step  $t$ .

# Stopping Rule: Maxwise Conformal Stopping

- Uncertainty density score:

$$u(T; \ell) = \frac{\#\{\text{arrivals of keywords in } T[1 : \ell]\}}{\ell}.$$

- Maximum uncertainty score: partition traces into bins of tokens of size  $B$ , with boundaries  $L_j = j \cdot B$ ,  $j = 1, \dots$

$$M_i = \max_{j: L_j \leq |T^{(i)}|} u(T^{(i)}; L_j).$$

- Compute threshold from calibration dataset: let  $k = \lceil (n + 1)(1 - \alpha) \rceil$  and set  $\tau^* = M_{(k)}$  (the  $k$ -th smallest value).

# Stopping Rule: Maxwise Conformal Stopping

- Uncertainty density score:

$$u(T; \ell) = \frac{\#\{\text{arrivals of keywords in } T[1 : \ell]\}}{\ell}.$$

- Maximum uncertainty score: partition traces into bins of tokens of size  $B$ , with boundaries  $L_j = j \cdot B$ ,  $j = 1, \dots$

$$M_j = \max_{j: L_j \leq |T^{(i)}|} u(T^{(i)}; L_j).$$

- Compute threshold from calibration dataset: let  $k = \lceil (n + 1)(1 - \alpha) \rceil$  and set  $\tau^* = M_{(k)}$  (the  $k$ -th smallest value).
- Stopping rule: for a new query with reasoning trace  $T$  at test-time, monitor  $u(T; L_j)$ ,  $j = 1, \dots$  and stop as soon as  $u(T, L_j) > \tau^*$ .

# Empirical Result for Mathematical Reasoning



Methods	GSM-MC		UMWP		MiP		MMLU	
	FPR	Power	FPR	Power	FPR	Power	FPR	Power
Length	4.50%	32.17%	0.62%	8.59%	20.67%	61.54%	12.91%	39.85%
DEER	5.08%	37.25%	2.12%	18.86%	16.67%	62.34%	20.18%	52.63%
Entropy	4.83%	44.92%	1.32%	15.31%	19.39%	55.45%	2.57%	15.29%
<b>Renewal</b>	3.75%	69.75%	2.92%	46.78%	1.12%	71.15%	0.75%	66.42%
<b>Maxwise</b>	5.42%	<b>70.92%</b>	5.52%	<b>48.65%</b>	3.21%	<b>73.40%</b>	2.63%	<b>75.31%</b>

# Empirical Result for Scientific Reasoning Tasks



Stopping Rule	GPQA		HLE	
	FPR	Power	FPR	Power
Confidence	12.85%	5.57%	-0.43%	3.95%
Criticism	-0.55%	9.12%	0.33%	4.11%
Length	42.40%	32.14%	23.74%	10.91%
<b>Renewal</b>	1.37%	24.94%	4.76%	24.62%
<b>Maxwise</b>	2.56%	<b>32.40%</b>	5.96%	<b>31.28%</b>

# Takeaway

- Reasoning models sometimes overthink, generating unnecessary reasoning steps, particularly under uncertainty, given ill-posed or ambiguous queries.
- Uncertainty keywords in the reasoning traces provide useful signals about the model's confidence, it generalizes reasonably well across models and datasets.

- [1]. Zijun Gao, Yan Sun, and Han Su. "Statistical Inference for Generative Model Comparison." <https://arxiv.org/pdf/2501.18897>
- [2]. Yangxinyu Xie, Tao Wang, Soham Mallick, Yan Sun, Georgy Noarov, Mengxin Yu, Tanwi Mallick, Weijie J. Su, and Edgar Dobriban. "Statistical Early Stopping for Reasoning Models." <https://arxiv.org/pdf/2602.13935>