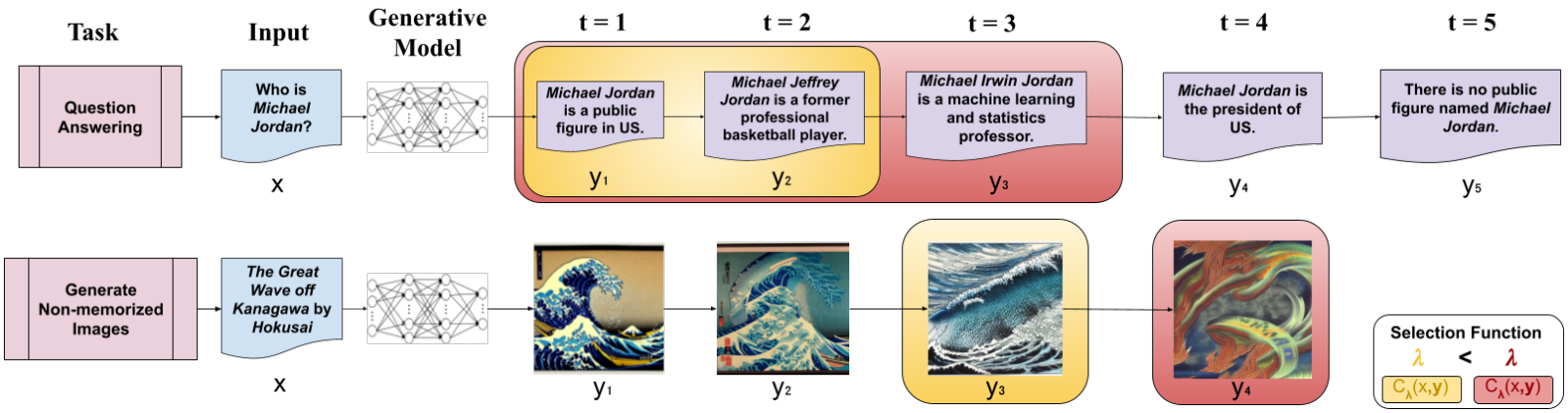
Conf-Gen: Conformal Uncertainty Quantification for Generative Models

Gabriel Loaiza-Ganem, Kevin Zhang, Wei Cui, Marc T. Law, Kin Kwan Leung

layer 6
AI at TD

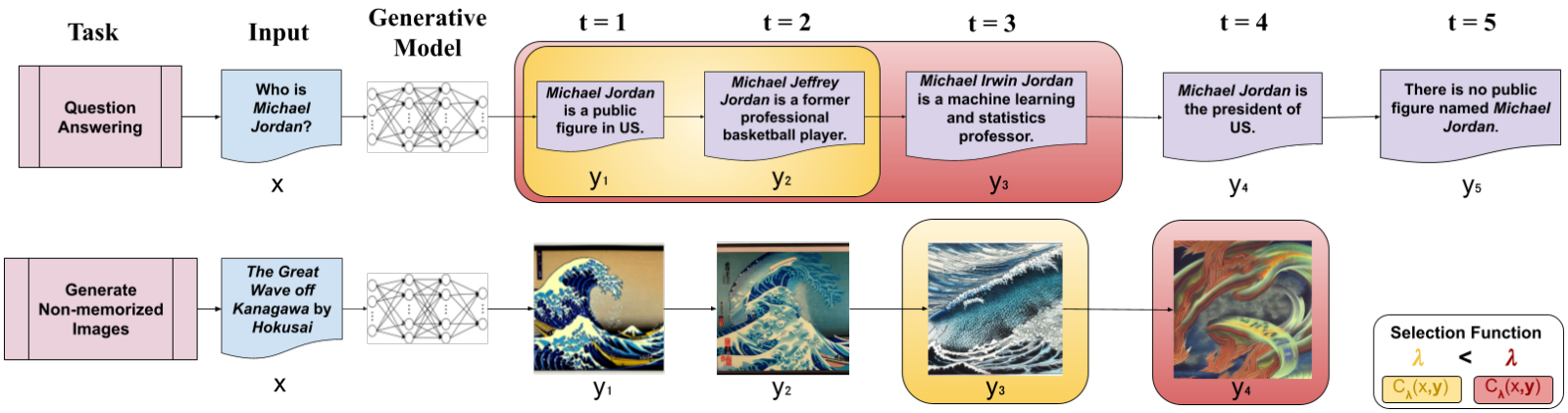
Conf-Gen: TDLR



- The function $\mathbf{C}_\lambda$ takes as input a conditioning x and a corresponding sequence of generations $\mathbf{y} = (y_1, \ldots, y_T)$, and returns an output which, in expectation, becomes increasingly conservative as λ grows.

layer 6 AI at TD

Conf-Gen: TDLR



- The function $\mathbf{C}_\lambda$ takes as input a conditioning x and a corresponding sequence of generations $\mathbf{y} = (y_1, \ldots, y_T)$, and returns an output which, in expectation, becomes increasingly conservative as λ grows.
- Conf-Gen uses a calibration dataset to select $\hat{\lambda}$ in such a way that the selected outputs are mathematically guaranteed to achieve, in expectation, a user-specified target *admissibility*.

Conf-Gen: The Formal Guarantee

- For a dataset $(X^{(i)}, \mathbf{Y}^{(i)}, Y_{\mathsf{GT}}^{(i)})_{i=1}^{n}$ and a user-specified γ, Conf-Gen finds $\hat{\lambda}$ as the smallest λ such that the produced outputs achieve a large enough average admissibility, i.e.,

$$\hat{\lambda} := \inf \left\{ \lambda : \frac{1}{n} \sum_{i=1}^{n} A\left(X^{(i)}, \mathbf{C}_{\lambda}(X^{(i)}, \mathbf{Y}^{(i)}), Y_{\mathsf{GT}}^{(i)}\right) \geq \frac{n+1}{n}\gamma \right\},$$

  where Y_{GT} is an optional ground truth variable used to compute admissibility (e.g., the correct answer to X).

Conf-Gen: The Formal Guarantee

- For a dataset $(X^{(i)}, \mathbf{Y}^{(i)}, Y_{\mathsf{GT}}^{(i)})_{i=1}^{n}$ and a user-specified γ, Conf-Gen finds $\hat{\lambda}$ as the smallest λ such that the produced outputs achieve a large enough average admissibility, i.e.,

$$\hat{\lambda} := \inf \left\{ \lambda : \frac{1}{n} \sum_{i=1}^{n} A\left(X^{(i)}, \mathbf{C}_{\lambda}(X^{(i)}, \mathbf{Y}^{(i)}), Y_{\mathsf{GT}}^{(i)}\right) \geq \frac{n+1}{n}\gamma \right\},$$

  where Y_{GT} is an optional ground truth variable used to compute admissibility (e.g., the correct answer to X).
- Conf-Gen ensures that, under mild regularity conditions, for a test datapoint $(X^{(n+1)}, \mathbf{Y}^{(n+1)}, Y_{\mathsf{GT}}^{(n+1)})$,

$$\mathbb{E}\left[A\left(X^{(n+1)}, \mathbf{C}_{\hat{\lambda}}(X^{(n+1)}, \mathbf{Y}^{(n+1)}), Y_{\mathsf{GT}}^{(n+1)}\right)\right] \geq \gamma.$$

Conf-Gen Extends Conformal Risk Control

- Conformal risk control (CRC) is an existing framework upon which Conf-Gen is based.

Conf-Gen Extends Conformal Risk Control

- Conformal risk control (CRC) is an existing framework upon which Conf-Gen is based.
- CRC bounds the expected loss of a set-valued function $\mathcal{C}_\lambda$, which provides a similar conformal guarantee to Conf-Gen's.

Conf-Gen Extends Conformal Risk Control

- Conformal risk control (CRC) is an existing framework upon which Conf-Gen is based.
- CRC bounds the expected loss of a set-valued function $\mathcal{C}_\lambda$, which provides a similar conformal guarantee to Conf-Gen's.
- The regularity conditions are strictly weaker for Conf-Gen: in particular we weaken a strict monotonicity requirement to one in (conditional) expectation.

Conf-Gen Extends Conformal Risk Control

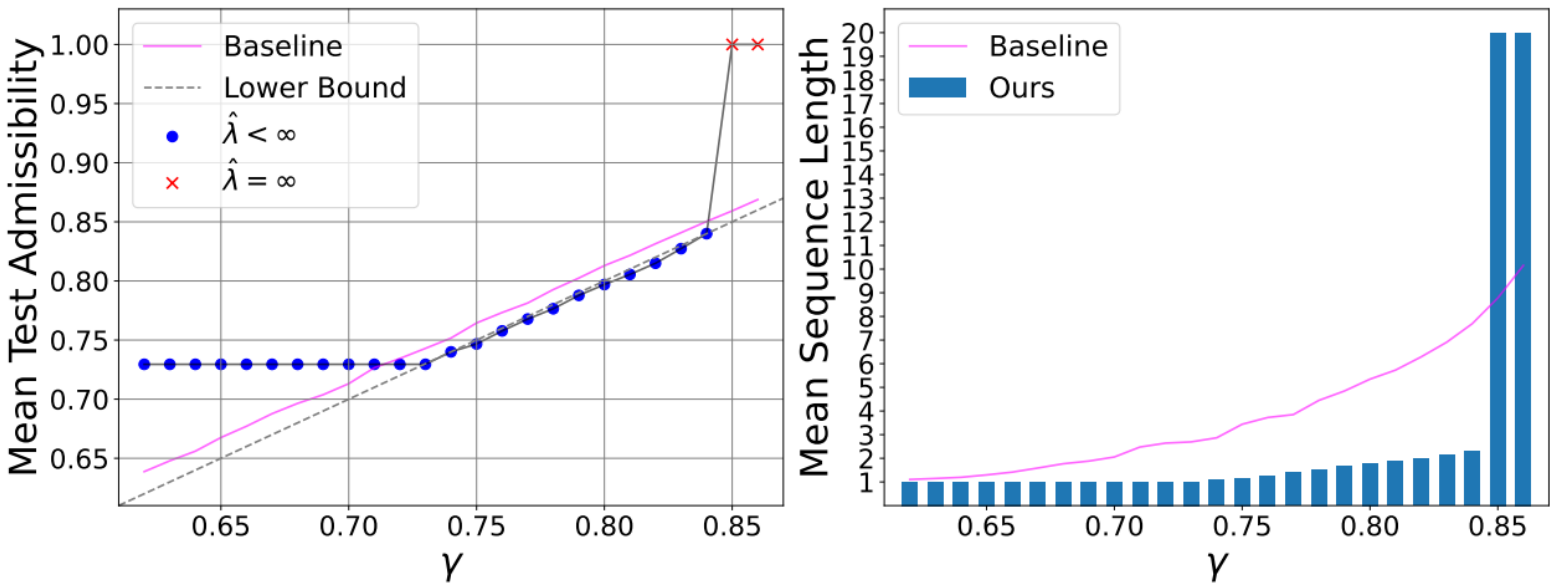
- Conformal risk control (CRC) is an existing framework upon which Conf-Gen is based.
- CRC bounds the expected loss of a set-valued function $\mathcal{C}_\lambda$, which provides a similar conformal guarantee to Conf-Gen's.
- The regularity conditions are strictly weaker for Conf-Gen: in particular we weaken a strict monotonicity requirement to one in (conditional) expectation.
- CRC was designed with supervised learning problems in mind, and both applications and implementations reflect this; Conf-Gen can be seen as extending CRC to generative models.

Question Answering

- X is a question.
- $\mathbf{Y} = (Y_1, \ldots, Y_T)$ is a sequence of LLM-generated responses.
- $\mathbf{C}_\lambda(x, \mathbf{y}) = \mathbf{y}_{:\tau(x,\mathbf{y},\lambda)}$, where $\tau(x, \mathbf{y}, \lambda)$ is the first t such that $\max(S_1^\uparrow, \ldots, S_t^\uparrow) > \lambda$, where $S_t^\uparrow = S^\uparrow(x, y_t)$ is the LLM's own self-confidence assessment.
- Y_{GT} is an optional ground truth response, and $A(x, \mathbf{y}, y_{\mathsf{GT}})$ is 1 if any response in $\mathbf{y}$ is a correct answer to x, and 0 otherwise (e.g., LLM as a judge or human evaluation).
- Conformal guarantee: with probability at least γ, at least one selected response is correct.
- Note: At test time, we don't have to generate T responses, we can do it sequentially and stop.

Question Answering

Results on TriviaQA:



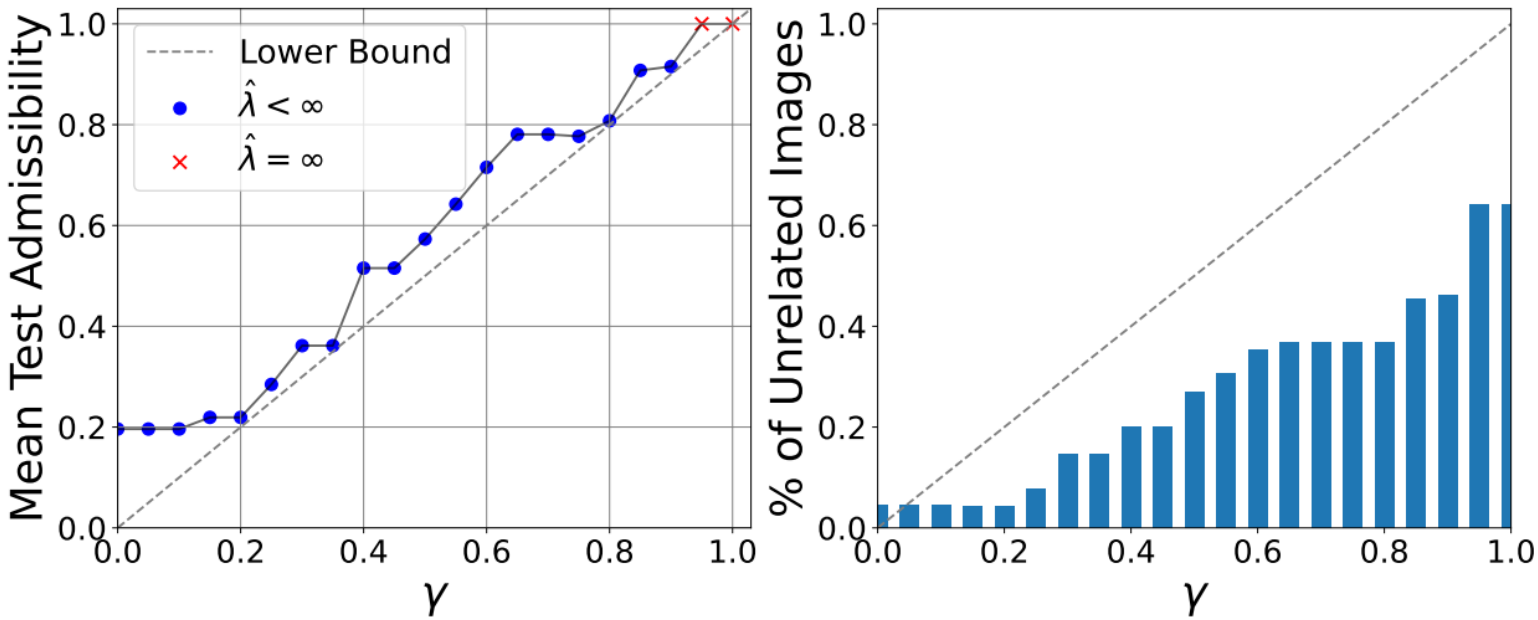
layer 6
AI at TD

Generation of Non-Memorized Images

- X is a prompt describing an image Y_{GT} that was memorized by Stable Diffusion.
- $\mathbf{Y} = (Y_1, \ldots, Y_T)$ is a sequence of images generated with Stable Diffusion. Y_1 is generated using X, Y_2 is generated by changing 1 token in X, Y_3 is generated by changing 2 tokens in X, etc.
- $\mathbf{C}_\lambda(x, \mathbf{y}) = y_{\tau(x,\mathbf{y},\lambda)}$, where $\tau(x, \mathbf{y}, \lambda)$ is the first t such that $S_t^{\uparrow} > \lambda$, where $S_t^{\uparrow} = S^{\uparrow}(x, y_t)$ is the negative norm of the classifier free guidance term used during generation of y_t.
- $A(x, y, y_{\mathsf{GT}})$ is the fraction of 10 human evaluators who assess y as not being a memorized version of y_{GT}.
- Conformal guarantee: a fraction of at least γ humans would assess the generated image as not memorized.

Generation of Non-Memorized Images

Results on memorized images by Stable Diffusion:



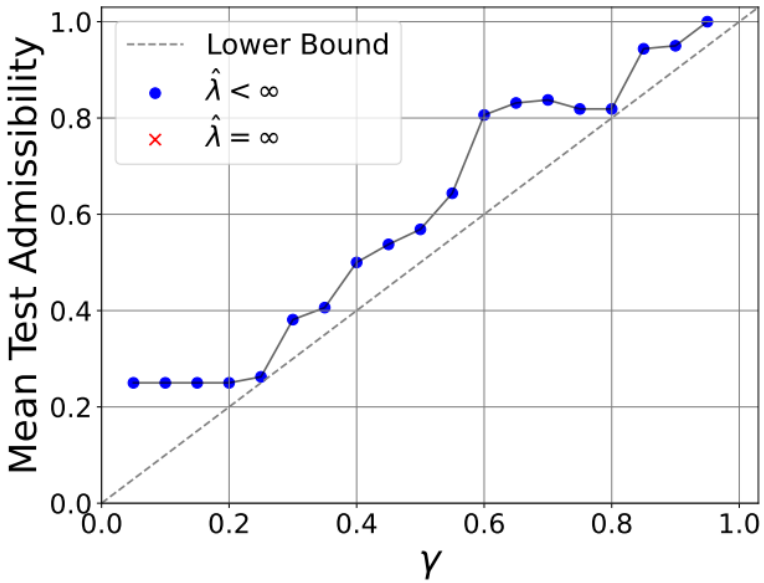
layer 6
AI at TD

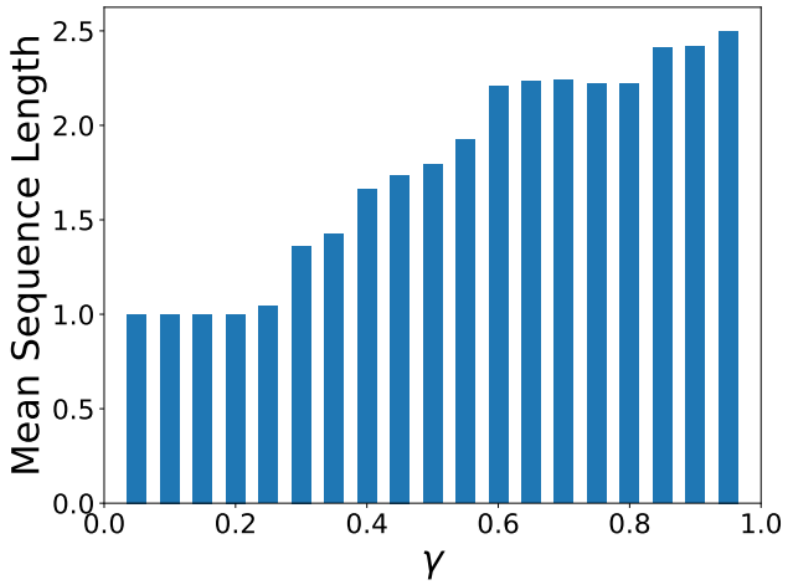
Conversational AI Chatbot

- $\mathbf{X} = (X_1, \ldots, X_T)$ is a sequence of questions asked by a user to a chatbot.
- $\mathbf{Y} = (Y_1, \ldots, Y_T)$ is a sequence of hypothetical responses.
- At step t, the chatbot decides between showing response Y_t to the user and ending the conversation, or asking for additional clarification, which means the user provides X_{t+1}.
- $\mathbf{C}_\lambda(\mathbf{x}, \mathbf{y}) = y_{\tau(\mathbf{x},\lambda)}$, where $\tau(\mathbf{x}, \lambda)$ is the first t such that $S_t^\uparrow > \lambda$, where $S_t^\uparrow = S^\uparrow(x_t)$ is the LLM's self-assessment that the question x_t is sufficiently unambiguous to be answered.
- We understand $\mathbf{C}_\lambda(\mathbf{x}, \mathbf{y}) = y_{\tau(\mathbf{x},\lambda)}$ as meaning that the chatbot decided to provide an answer at turn $\tau(\mathbf{x}, \lambda)$.
- $A(\mathbf{x}, \mathbf{y})$ is a binary label indicating whether the last question of $\mathbf{x}$ is sufficiently unambiguous to be answered.
- Conformal guarantee: with probability at least γ, the chatbot has asked enough clarifying questions before providing a response.

Conversational AI Chatbot

Results on the ClariQ dataset:





layer 6 AI at TD

More About Conf-Gen

- We beat an existing baseline at question answering, and provide 4 new tasks where conformal ideas can be applied.
- All instances of conformal ideas applied to Gen AI are instances of Conf-Gen.
- More theory in the paper!
- Python package!

Thanks!