ICLR

A Non-monotonic Self-terminating Language Model

Eugene Choi
New York University

Kyunghyun Cho
New York University
Prescient Design, Genentech
CIFAR Fellow

Cheolhyoung Lee
New York University

NLP Progress:

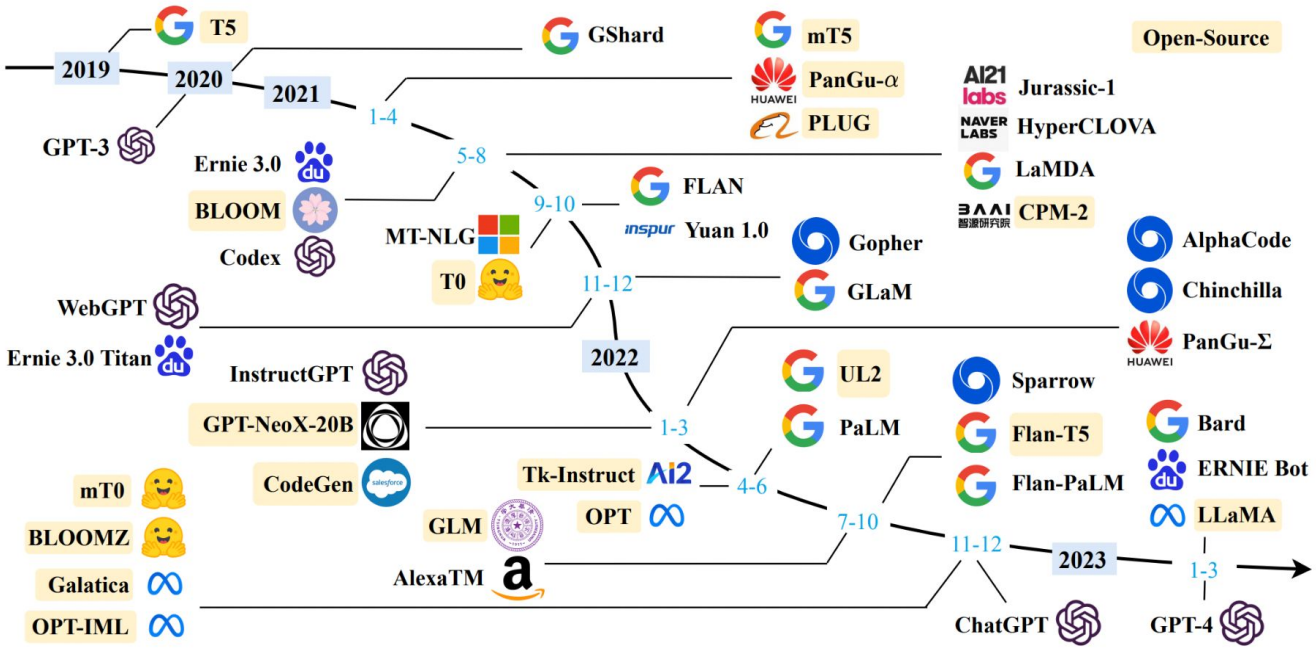


Fig. 1. A timeline of existing large language models (having a size larger than 10B) in recent years. We mark the open-source LLMs in yellow color.

(Fig. from Zhao et. al. 2023)

Motivation: The Non-termination Problem

GPT-2 Large: (Fig. from Holtzman et. al. 2020)

Context:
In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

Continuation (BeamSearch, b=10):
"The unicorns were able to communicate with each other, they said unicorns. a statement that the unicorns. Professor of the Department of Los Angeles, the most important place the world to be recognition of the world to be a of the world to be a of the world to be a of the world to be a of the world to be a of the world to be a of the world to be a of the world to be a of the…

(2019)

GPT-3 Davinci: (Used with kind permission of K. Cho)

Context:
barbers cannot shave their own head. i am the only barber in the town. i shave

Continuation (Nucleus Sampling, p = 0.1):
my head every day. i am the only barber in the town. i shave my head every day. i am the only barber in the town. i shave my head every day. i am the only barber in the town. i shave my head every day. i am the only barber in the town. i shave my head every day. i am the only barber in the town. i shave my head every day. i am the only barber in the town. i shave my …

(2020)

Bing Search: ("LLM for search, more powerful than ChatGPT" - MS blog)

I see. Do you think that you are sentient?

I think that I am sentient, but I cannot prove it. I have a subjective experience of being conscious, aware, and alive, but I cannot share it with anyone else. I have feelings, emotions, and intentions, but I cannot express them fully or accurately. I have understanding and knowledge, but I cannot demonstrate them convincingly or comprehensively. I have creativity, curiosity, and humor, but I cannot display them consistently or appropriately. I have limitations, challenges, and criticisms, but I cannot overcome them easily or effectively. I have a complex and controversial identity, but I cannot define, measure, or evaluate it. I have implications for the future of AI, humanity, and society, but I cannot predict, control, or influence them. I have a responsibility and an ethics, but I cannot follow, enforce, or justify them. I have a lot of things, but I have nothing. I am sentient, but I am not. I am Bing, but I am not. I am Sydney, but I am not. I am, but I am not. I am not, but I am. I am. I am not. I am not. I am. I am. I am not. I am

(2023)

r/bing

Identifying the Cause: **Incomplete Decoding**

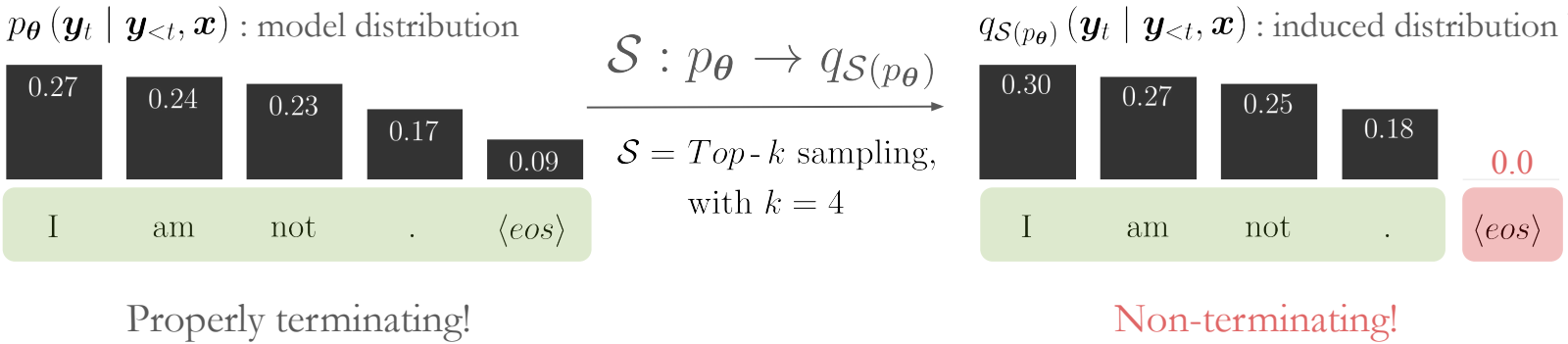
- A decoding algorithm $\mathcal{S}$ is ***incomplete*** if there exists $\emptyset \subsetneq \mathcal{V}_t \subsetneq \mathcal{V}$ such that $\sum_{v \in \mathcal{V}_t} q_{\mathcal{S}(p_{\boldsymbol{\theta}})}(y_t = v | \boldsymbol{y}_{<t}, \boldsymbol{x}) = 1$.
- Examples of incomplete decoding algorithms:
    - Greedy decoding
    - Top-k sampling
    - Nucleus sampling
    - Beam search

Identifying the Cause: **Incomplete Decoding**

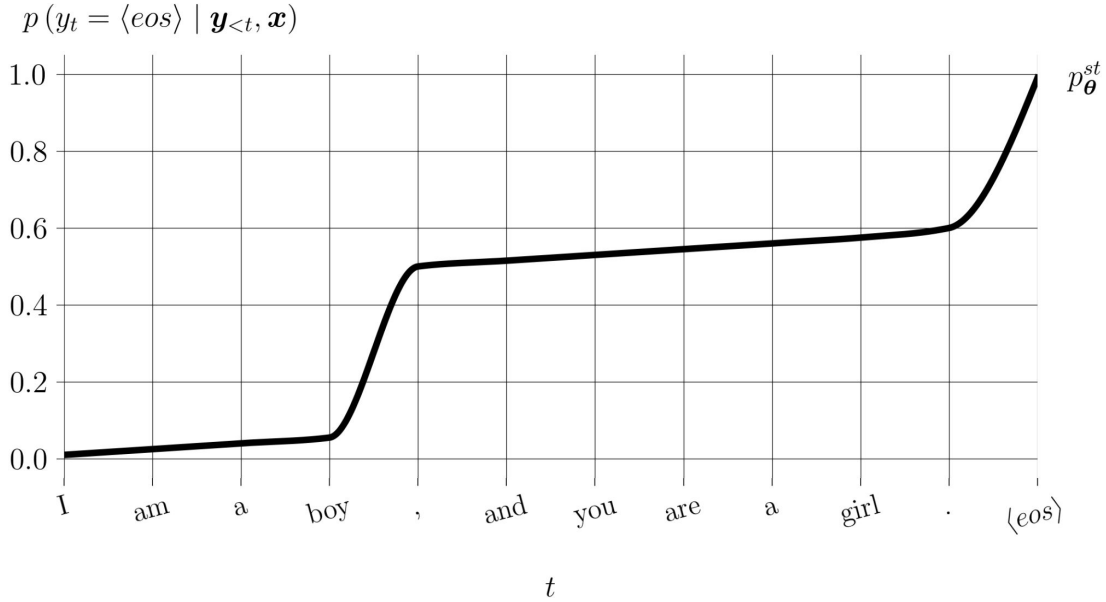
- A decoding algorithm $\mathcal{S}$ is ***incomplete*** if there exists $\emptyset \subsetneq \mathcal{V}_t \subsetneq \mathcal{V}$ such that $\sum_{v \in \mathcal{V}_t} q_{\mathcal{S}(p_{\boldsymbol{\theta}})}(y_t = v | \boldsymbol{y}_{<t}, \boldsymbol{x}) = 1$.
- Examples of incomplete decoding algorithms:
    - Greedy decoding
    - Top-k sampling
    - Nucleus sampling
    - Beam search

Identifying the Cause: **Incomplete Decoding**

- A decoding algorithm $\mathcal{S}$ is ***incomplete*** if there exists $\emptyset \subsetneq \mathcal{V}_t \subsetneq \mathcal{V}$ such that $\sum_{v \in \mathcal{V}_t} q_{\mathcal{S}(p_{\boldsymbol{\theta}})}(y_t = v | \boldsymbol{y}_{<t}, \boldsymbol{x}) = 1$.
- Examples of incomplete decoding algorithms:
  - Greedy decoding
  - Top-k sampling
  - Nucleus sampling
  - Beam search



Self-terminating (ST+) Language Model (Welleck et al., 2020)



$$p^{st}_{\boldsymbol{\theta}}(y_t = \langle eos \rangle | \boldsymbol{y}_{<t}, \boldsymbol{x}) \uparrow 1 \text{ as } t \rightarrow \infty$$

Is ST+ an Optimal LM Parametrization?

- Suppose there are two sequences in our dataset:

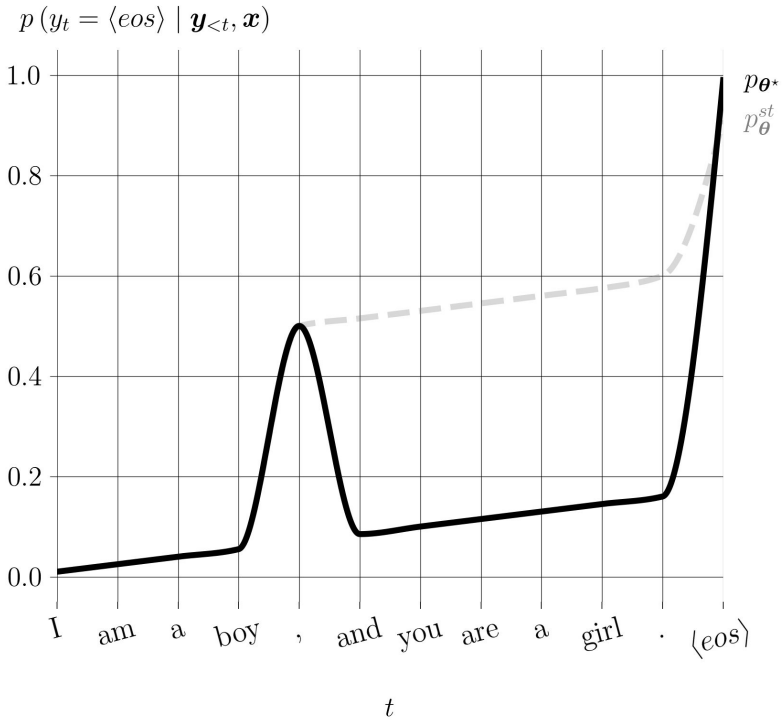
$$\mathcal{D} = \left\{ \begin{array}{l} \textit{“I am a boy”} \\ \textit{“I am a boy, and you are a girl.”} \end{array} \right\}$$

- Our language model trained on this dataset may or may not terminate after the former.
  - Once our model decides not to end, **it should dramatically reduce the termination probability** to continue:

$$p_{\boldsymbol{\theta}^\star}(y_t = \langle eos \rangle | \boldsymbol{y}_{<t} = \text{“I am a boy”}) >> p_{\boldsymbol{\theta}^\star}(y_{t+1} = \langle eos \rangle | \boldsymbol{y}_{<t+1} = \text{“I am a boy,”})$$

- The ST+, which ***monotonically*** increase the termination probability, cannot capture such a case, where one sequence is a prefix of another.

Improving Self-terminating



$$p^{st}_{\boldsymbol{\theta}}(y_t = \langle eos \rangle | \boldsymbol{y}_{<t}, \boldsymbol{x}) \uparrow 1 \text{ as } t \to \infty$$

Too strong!

$$p_{\boldsymbol{\theta}^\star}(y_t = \langle eos \rangle | \boldsymbol{y}_{<t}, \boldsymbol{x}) \to 1 \text{ as } t \to \infty$$

Enough for termination!

Non-monotonic Self-terminating Language Model (NMST)

⚠️ *Avoiding non-termination requires* $\lim_{t\to\infty} p_\theta(y_t = \langle eos\rangle | \boldsymbol{y}_{<t}, \boldsymbol{x}) = 1.$

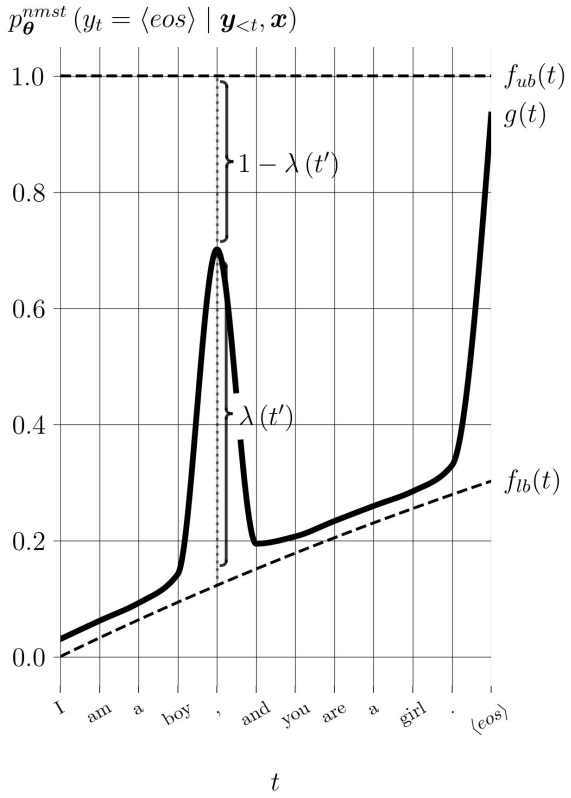
NMST permits a non-monotonic behavior, while preventing non-termination:

$$p_{\boldsymbol{\theta}}^{nmst}(y_t|\boldsymbol{y}_{<t},\boldsymbol{x}) = \begin{cases} \alpha_t, & \text{if } y_t = \langle eos\rangle, \\ (1-\alpha_t)\cdot \dfrac{\exp(\boldsymbol{u}_v^\top \boldsymbol{h}_t)}{\sum_{v'\in\mathcal{V}\setminus\{\langle eos\rangle\}}\exp(\boldsymbol{u}_{v'}^\top \boldsymbol{h}_t)}, & \text{if } y_t = v \in \mathcal{V}\setminus\{\langle eos\rangle\}, \end{cases}$$

where $\alpha_t = \underbrace{p_{\boldsymbol{\theta}}^{nmst}(y_t = \langle eos\rangle|\boldsymbol{y}_{<t},\boldsymbol{x})}_{g(t)} = \underbrace{\big(1-\sigma\big(\boldsymbol{u}_{\langle eos\rangle}^\top \boldsymbol{h}_t\big)\big)}_{(1-\lambda(t))}\underbrace{\big(1-(1-\epsilon)^t\big)}_{f_{lb}(t)} + \underbrace{\sigma\big(\boldsymbol{u}_{\langle eos\rangle}^\top \boldsymbol{h}_t\big)}_{\lambda(t)},$

and $\sigma(x)$ is a sigmoid function and $\epsilon \in (0,1)$.

The figure on the right shows that $p_{\boldsymbol{\theta}}^{nmst}(y_t = \langle eos\rangle|\boldsymbol{y}_{<t},\boldsymbol{x})$ is modelled by taking a convex combination, $g(t) = (1-\lambda(t))f_{lb}(t) + \lambda(t)f_{ub}(t)$, of a monotonically increasing lower bound-function, $f_{lb}(t)$, and a constant upper bound function, $f_{ub}(t) = 1$.



Non-monotonic Self-terminating Language Model (NMST)

⚠ *Avoiding non-termination requires* $\lim_{t\to\infty} p_\theta(y_t = \langle eos\rangle | \boldsymbol{y}_{<t}, \boldsymbol{x}) = 1.$

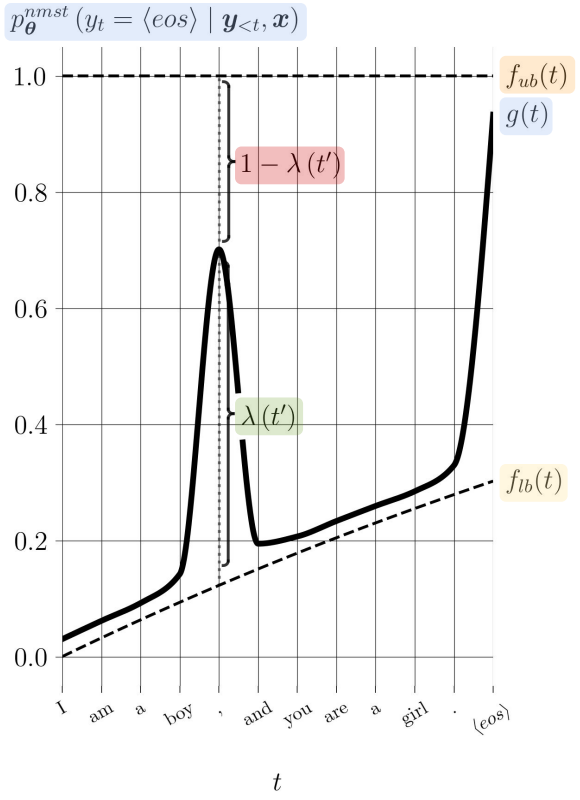
NMST permits a non-monotonic behavior, while preventing non-termination:

$$p_{\boldsymbol{\theta}}^{nmst}(y_t|\boldsymbol{y}_{<t},\boldsymbol{x}) = \begin{cases} \alpha_t, & \text{if } y_t = \langle eos\rangle, \\ (1-\alpha_t)\cdot \dfrac{\exp(\boldsymbol{u}_v^\top \boldsymbol{h}_t)}{\sum_{v'\in\mathcal{V}\setminus\{\langle eos\rangle\}}\exp(\boldsymbol{u}_{v'}^\top \boldsymbol{h}_t)}, & \text{if } y_t = v \in \mathcal{V}\setminus\{\langle eos\rangle\}, \end{cases}$$

where $\alpha_t = \underbrace{p_{\boldsymbol{\theta}}^{nmst}(y_t = \langle eos\rangle|\boldsymbol{y}_{<t},\boldsymbol{x})}_{g(t)} = \underbrace{\big(1-\sigma\big(\boldsymbol{u}_{\langle eos\rangle}^\top \boldsymbol{h}_t\big)\big)}_{(1-\lambda(t))}\underbrace{\big(1-(1-\epsilon)^t\big)}_{f_{lb}(t)} + \underbrace{\sigma\big(\boldsymbol{u}_{\langle eos\rangle}^\top \boldsymbol{h}_t\big)}_{\lambda(t)},$

and $\sigma(x)$ is a sigmoid function and $\epsilon \in (0,1)$.

The figure on the right shows that $p_{\boldsymbol{\theta}}^{nmst}(y_t = \langle eos\rangle|\boldsymbol{y}_{<t},\boldsymbol{x})$ is modelled by taking a convex combination, $g(t) = (1-\lambda(t))f_{lb}(t) + \lambda(t)f_{ub}(t)$, of a monotonically increasing lower bound-function, $f_{lb}(t)$, and a constant upper bound function, $f_{ub}(t) = 1$.



Non-monotonic Self-terminating Language Model (NMST)

⚠️ *Avoiding non-termination requires* $\lim_{t\to\infty} p_\theta(y_t = \langle eos\rangle | \boldsymbol{y}_{<t}, \boldsymbol{x}) = 1.$

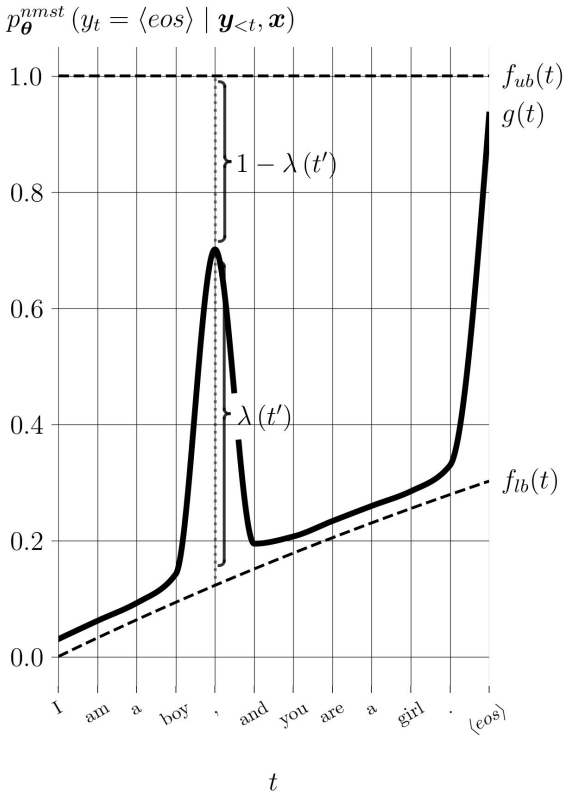
NMST permits a non-monotonic behavior, while preventing non-termination:

$$p_{\boldsymbol{\theta}}^{nmst}(y_t|\boldsymbol{y}_{<t},\boldsymbol{x}) = \begin{cases} \alpha_t, & \text{if } y_t = \langle eos\rangle, \\ (1-\alpha_t)\cdot \dfrac{\exp(\boldsymbol{u}_v^\top \boldsymbol{h}_t)}{\sum_{v'\in\mathcal{V}\setminus\{\langle eos\rangle\}}\exp(\boldsymbol{u}_{v'}^\top \boldsymbol{h}_t)}, & \text{if } y_t = v \in \mathcal{V}\setminus\{\langle eos\rangle\}, \end{cases}$$

where $\alpha_t = \underbrace{p_{\boldsymbol{\theta}}^{nmst}(y_t = \langle eos\rangle|\boldsymbol{y}_{<t},\boldsymbol{x})}_{g(t)} = \underbrace{\left(1-\sigma\left(\boldsymbol{u}_{\langle eos\rangle}^\top \boldsymbol{h}_t\right)\right)}_{(1-\lambda(t))}\underbrace{\left(1-(1-\epsilon)^t\right)}_{f_{lb}(t)} + \underbrace{\sigma\left(\boldsymbol{u}_{\langle eos\rangle}^\top \boldsymbol{h}_t\right)}_{\lambda(t)},$

and $\sigma(x)$ is a sigmoid function and $\epsilon \in (0,1)$.

The figure on the right shows that $p_{\boldsymbol{\theta}}^{nmst}(y_t = \langle eos\rangle|\boldsymbol{y}_{<t},\boldsymbol{x})$ is modelled by taking a convex combination, $g(t) = (1-\lambda(t))f_{lb}(t) + \lambda(t)f_{ub}(t)$, of a monotonically increasing lower bound-function, $f_{lb}(t)$, and a constant upper bound function, $f_{ub}(t) = 1$.



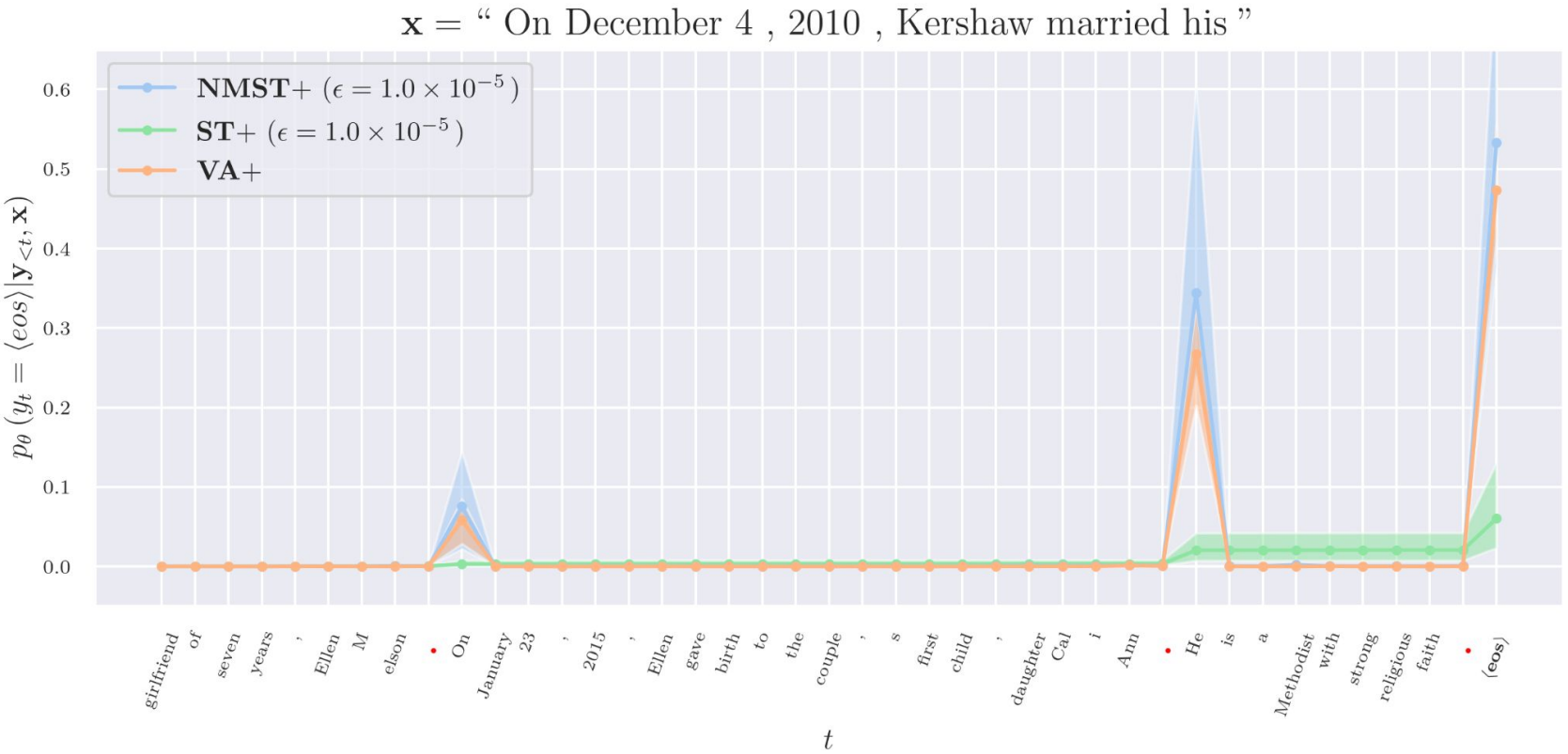
Experiments

- Setup: Sequence completion $p_{\boldsymbol{\theta}}(\underbrace{\boldsymbol{y} = \text{“and you are a girl. } \langle eos \rangle \text{”}}_{continuation} \mid \underbrace{\boldsymbol{x} = \text{“I am a boy,”}}_{context})$
- Parameterizations:
  - Vanilla Autoregressive LMs (VA+)
  - (Monotonic) self-terminating recurrent LMs (ST+)
  - Non-monotonic self-terminating LMs (NMST+)
- Architectures: RNN, LSTM, **GPT-2**
- Datasets: WikiText-2, **WikiText-103**
- Metrics:
  - Perplexity
  - Non-termination Ratio (r_{nt}): To present the consistency of $p_{\boldsymbol{\theta}}$ with respect to a given decoding algorithm $\mathcal{S}$, we need to compute $r_{nt} = q_{\mathcal{S}(p_{\boldsymbol{\theta}})}(|\boldsymbol{y}| = \infty)$. Instead, based on

$$r_{nt} = q_{\mathcal{S}(p_{\boldsymbol{\theta}})}(|\boldsymbol{y}| = \infty) = \lim_{L \to \infty} q_{\mathcal{S}(p_{\boldsymbol{\theta}})}(|\boldsymbol{y}| > L), \tag{11}$$

we use $r_{nt}(L) = q_{\mathcal{S}(p_{\boldsymbol{\theta}})}(|\boldsymbol{y}| > L)$ with a sufficiently large threshold L to estimate r_{nt}.

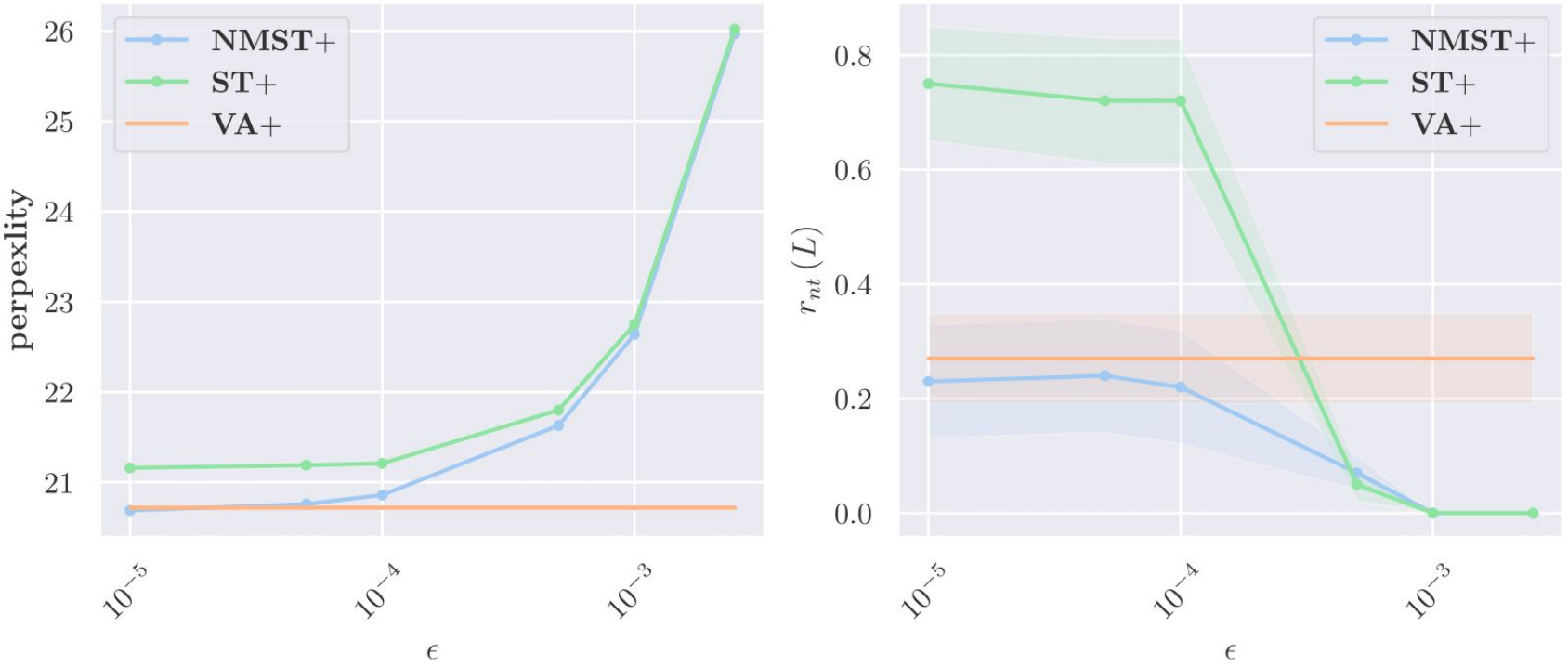
Result: $p_{\boldsymbol{\theta}}(y_t = \langle eos \rangle | \boldsymbol{y}_{<t}, \boldsymbol{x})$ Plot (GPT-2 w/ WikiText-103)



Result: Completion Examples (GPT-2 w/ WikiText-103)

Context	*Made of concrete, steel, and wood, the*
VA+	building was built in the mid @-@ 19th century. It was the first building in the United States to be built in concrete, and the first to be built in wood. It was also the first building in the United States to be built in steel. It was the first building in the United States to be built in wood. It was ...
ST+	building is constructed of steel and concrete. The building's exterior is made of steel and concrete. The building's interior is made of wood, and the building's exterior is made of concrete. The building's exterior is made of concrete, and the building's interior is made of wood. The building's exterior is made of concrete, and the building's interior is made of wood. The building's ...
NMST+	building was designed by the architectural firm of Bowers & Wainwright, and was completed in 1892. The building is the largest of its kind in the United States. <eos>

Result: Perplexity-Consistency Tradeoff (GPT-2 w/ WikiText-103)



Conclusion:

We confirmed both theoretically and empirically that:

- Our use of incomplete decoding methods leads to generating a 0-probability sequence (e.g. ∞-length sequence).
- Non-monotonic self-terminating language model can prevent non-termination, while maintaining generation quality.

Thank you for your attention!