Part 3: Modeling Perspectives (decoding, distillation, and diffusion)

Wei Xu (Georgia Tech)

Three Popular Methods for Generation

1. Decoding:

an inference-time solution to optimize LLM outputs

(Survey by Welleck+ 2024 & Bertsch+, 2023; MBR with Multi-Prompt by Heineman+, 2024)

2. Distillation:

reproduce GPT-4 performance by small open-source LLMs

(Edit-based generation by Dou+ 2024; Feedback to refine LLM outputs by Wadhwa+ 2024)

3. Diffusion:

an alternative to Transformer-based LLM

(Diffusion-LM by Li+ 2022; DiffuSeq by Gong+, 2022; SeqDiffuSeq by Yuan+, 2024)

Three Popular Methods for Generation

1. Decoding:

an inference-time solution to optimize LLM outputs

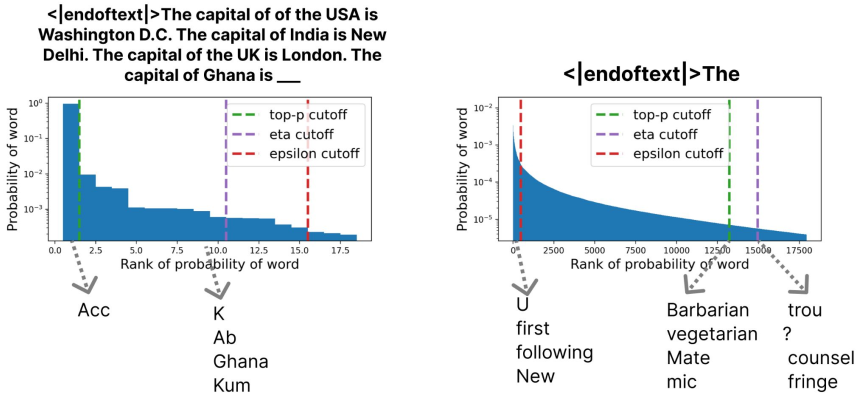
Besides data and model size, inference-time algorithms can make a big impact.

Three Popular Methods for Generation

1. Decoding:

an inference-time solution to optimize LLM outputs

Besides data and model size, inference-time algorithms can make a big impact.



Hewitt et al., "Truncation Sampling as Language Model Desmoothing" (ENNLP Findings 2022)

Decoding

Given an input x (and prompt ρ), an autoregressive LM parameterized by π_θ will estimate an output sequence:

$$y \sim \pi_\theta(x, \rho)$$,

using an decoding algorithm.

Decoding - common strategies

- Greedy Decoding
- Searching, e.g., Beam Search
- Sampling, e.g.
  - Temperature, or Top-k: sample from top k most likely words
  - Nucleus: take the top p% (95%) of the distribution, sample from within that
  - Epsilon: simple truncation, allow any word with greater than ε probability

predict the **next token** conditioned on the input $\pi_\theta(y_i | y_{<i}, x, \rho)$

Decoding - common strategies

- Greedy Decoding
- Searching, e.g., Beam Search
- Sampling, e.g.
  - Temperature, or Top-k: sample from top k most likely words
  - Nucleus: take the top p% (95%) of the distribution, sample from within that
  - Epsilon: simple truncation, allow any word with greater than ε probability

predict the **next token** conditioned on the input $\pi_\theta(y_i | y_{<i}, x, \rho)$

- Reranking, e.g.
  - Minimum Bayes Risk
  - Speculative decoding

generate multiple **candidate sequences**, then select one from them.

Minimum Bayes Risk (MBR) Decoding

- Early work (Bickel & Doksum ’77)
- Statistical Machine Translation and Speech Recognition, since 1997
- LLM-era, since 2020:
    - Mostly, **machine translation** (Eikema+ ’20; Fernandes+ ’22; Freitag+ ’22; Amrhein+ ’22; and more)
    - More recently, **generation**:
        - Code Generation (Shi+ ’22)
        - Summarization, Data-to-Text, Translation, Style Transfer (Suzgun+, ’23)
        - Summarization, Date-to-Text, Translation, Image Captioning (Jinnai+, ’24)
        - Text Simplification, Code Generation, Translation (Heineman+, ’24)

High Quality Rather than High Model Probability:
Minimum Bayes Risk Decoding with Neural Metrics

Markus Freitag, David Grangier, Qijun Tan, Bowen Liang

Google Research, USA

It’s MBR All the Way Down:
Modern Generation Techniques Through the Lens of Minimum Bayes Risk

Amanda Bertsch* and **Alex Xie*** and **Graham Neubig** and **Matthew R. Gormley**
Carnegie Mellon University

From Decoding to Meta-Generation:
Inference-time Algorithms for Large Language Models

Sean Welleck — wellecks@cmu.edu
Carnegie Mellon University

Amanda Bertsch* — abertsch@cs.cmu.edu
Carnegie Mellon University

Minimum Bayes Risk (MBR) Decoding

- Often deliver several points of performance improvement, over the standard beam search or sampling methods.

	R-1	R-2	R-L	BLEU
Summarization	XSUM			
Sample-Once	37.9	16.1	30.6	11.4
Random	37.6	16.1	30.1	11.5
Majority Voting	37.8	16.2	30.6	11.4
MBRD-BLEURT	39.8	17.9	32.4	12.8
MBRD-BERTScore	**41.2**	**19.0**	**33.4**	**13.5**
Translation	DE $\rightarrow$ EN (German to English)			
Sample-Once	68.1	45.9	63.9	39.0
Random	68.5	46.1	64.0	39.5
Majority Voting	70.2	48.7	66.1	40.9
MBRD-BLEURT	71.9	50.7	68.2	43.7
MBRD-BERTScore	**73.3**	**52.6**	**69.6**	**45.8**

(Suzgun+, '23)

	SARI	BScore	LENS	sBL↓	Human
Simplification	SIMPEVAL$_{2022}$				
T5-11B					
MLE$_{b=10}$	**46.4**	**93.8**	62.9	49.3	88.80
MBR-LENS$_{\|S\|=100}$	46.1	**93.8**	**74.4**	**44.6**	**90.13**
Close-source LLMs					
GPT-3.5 (0-shot)	41.4	93.4	60.7	31.8	90.77
GPT-3.5 (5-shot)	42.4	94.1	69.0	33.2	92.70
GPT-4 (0-shot)	43.7	**94.3**	**73.5**	**29.1**	**93.63**

(Maddela+, '23)

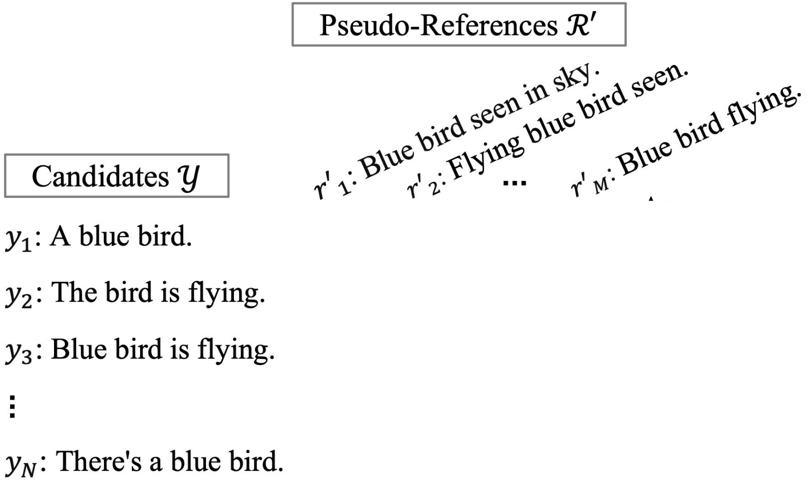
Suzgun et al. "Effective Text Generation via Minimum Bayes Risk Decoding" (ACL Findings 2023)
Maddela et al. "LENS: A Learnable Evaluation Metric for Text Simplification". (ACL 2023)

Minimum Bayes Risk (MBR) Decoding

Intuition: the best output not only have high probability (same as *maximum likelihood*), but also is consistent or similar to the other candidate outputs.

Minimum Bayes Risk (MBR) Decoding

1. sample multiple sequences

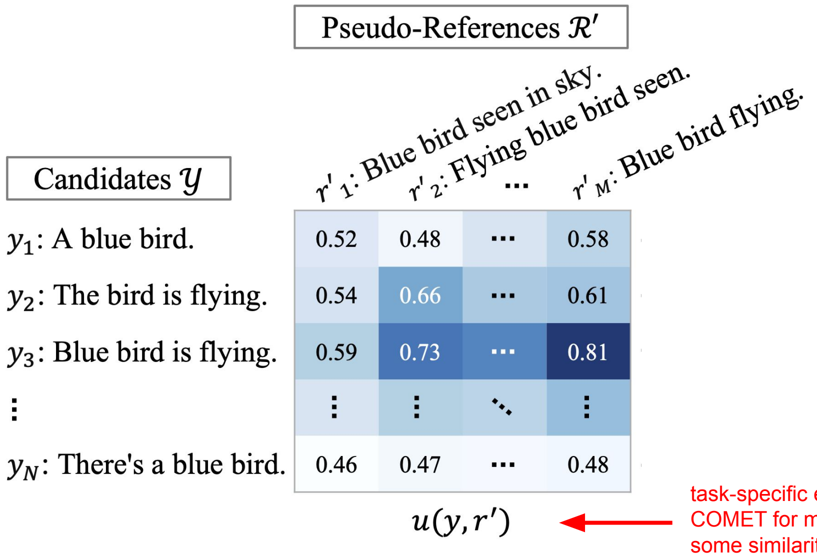


Ohashi et al. "On the True Distribution Approximation of Minimum Bayes-Risk Decoding" (NAACL 2024)

Minimum Bayes Risk (MBR) Decoding

1. sample multiple sequences

2. Compare each seq. to the others by a utility function



task-specific evaluation metrics (e.g., COMET for machine translation), or some similarity measurements

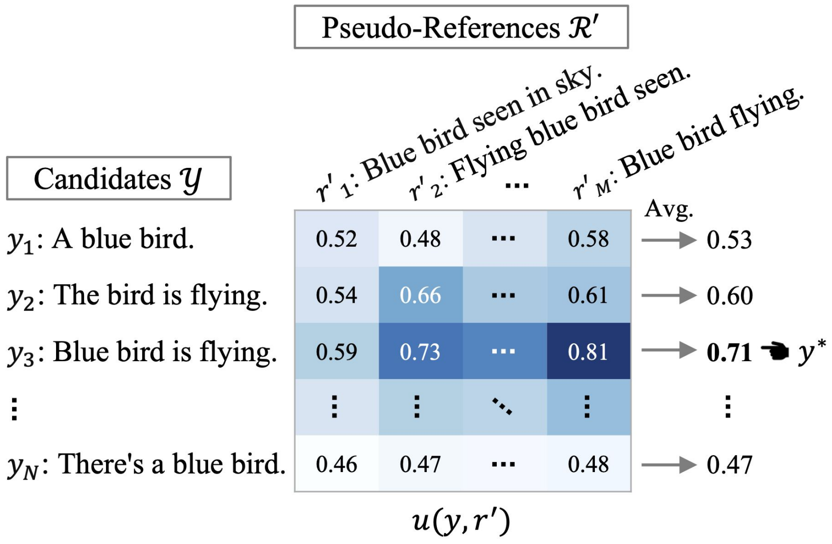
Ohashi et al. "On the True Distribution Approximation of Minimum Bayes-Risk Decoding" (NAACL 2024)

Minimum Bayes Risk (MBR) Decoding

1. sample multiple sequences

2. Compare each seq. to the others by a utility function

3. Select the seq. that maximizes the expected utility over the estimated probability distribution over the seq.'s.



Ohashi et al. "On the True Distribution Approximation of Minimum Bayes-Risk Decoding" (NAACL 2024)

Minimum Bayes Risk (MBR) Decoding

Intuition: the best output not only have high probability (same as *maximum likelihood*), but also is consistent or similar to the other candidate outputs.

More formally:

First sample a set of hypotheses $\mathcal{H}$ from the model π_θ. then select the output that maximizes the expected utility U (or minimize the expected risk) with respect to a set of references $\mathcal{R}$:

$$\hat{y}_{\text{MBR}} = \arg\max_{y \in \mathcal{H}} \left(\mathbb{E}_{\mathcal{H} \sim \pi_\theta} [U(y, \mathcal{R})] \right)$$

Minimum Bayes Risk (MBR) Decoding

Intuition: the best output <u>not only</u> have high probability (same as *maximum likelihood*), <u>but also</u> is consistent or similar to the other candidate outputs.

More formally:

can be the same or different set, often about 10~1000 sequences generated by sampling or beam search

First sample a set of hypotheses $\mathcal{H}$ from the model π_θ. then select the output that maximizes the expected utility U (or minimize the expected risk) with respect to a set of references $\mathcal{R}$:

$$\hat{y}_{\text{MBR}} = \underset{y \in \mathcal{H}}{\arg\max} \left(\mathbb{E}_{\mathcal{H} \sim \pi_\theta} [\, U(y, \mathcal{R})] \right)$$

Minimum Bayes Risk (MBR) Decoding

Main challenges:

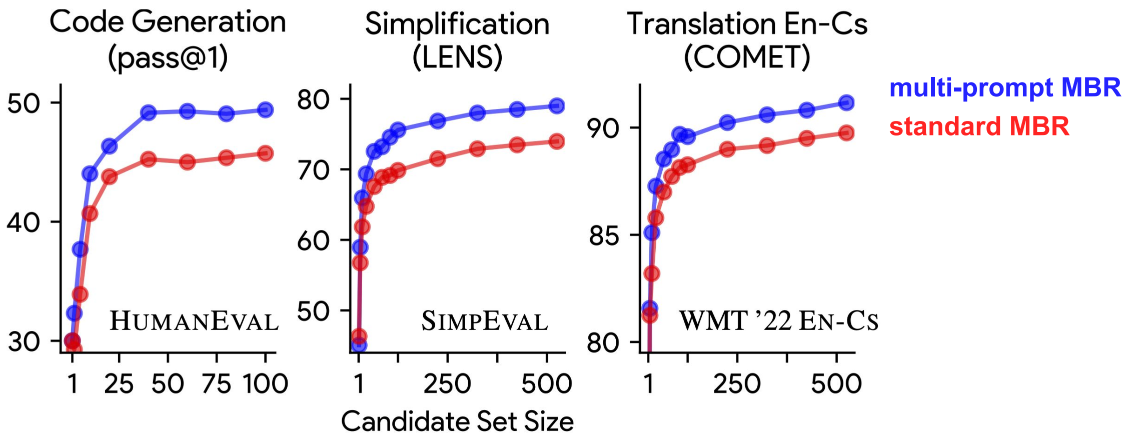
- $O(|\mathcal{H}|^2)$ computation time for utility function
- number of sample $|\mathcal{H}|$ << number of all possible hypotheses $|\mathcal{Y}|$

Interesting research directions:

- choice of sampling algorithm to collect $\mathcal{H}$ (and $\mathcal{R}$, if different)
  - appear to be critical (Ohashi+ ’24)
  - probabilistic sampling better than beam search? (Eikema+ ’20, Fernandes+ ’22, Freitga+ ’23)
- approximation for estimating the probability distribution in expected utility
  - model-based estimation (Jinnai+ ’24a)
- promoting diversity (Heineman+ ’24, Jinnai+ ‘24b)
- reducing computation time for utility function (Tomani+ ’24)

Minimum Bayes Risk (MBR) Decoding

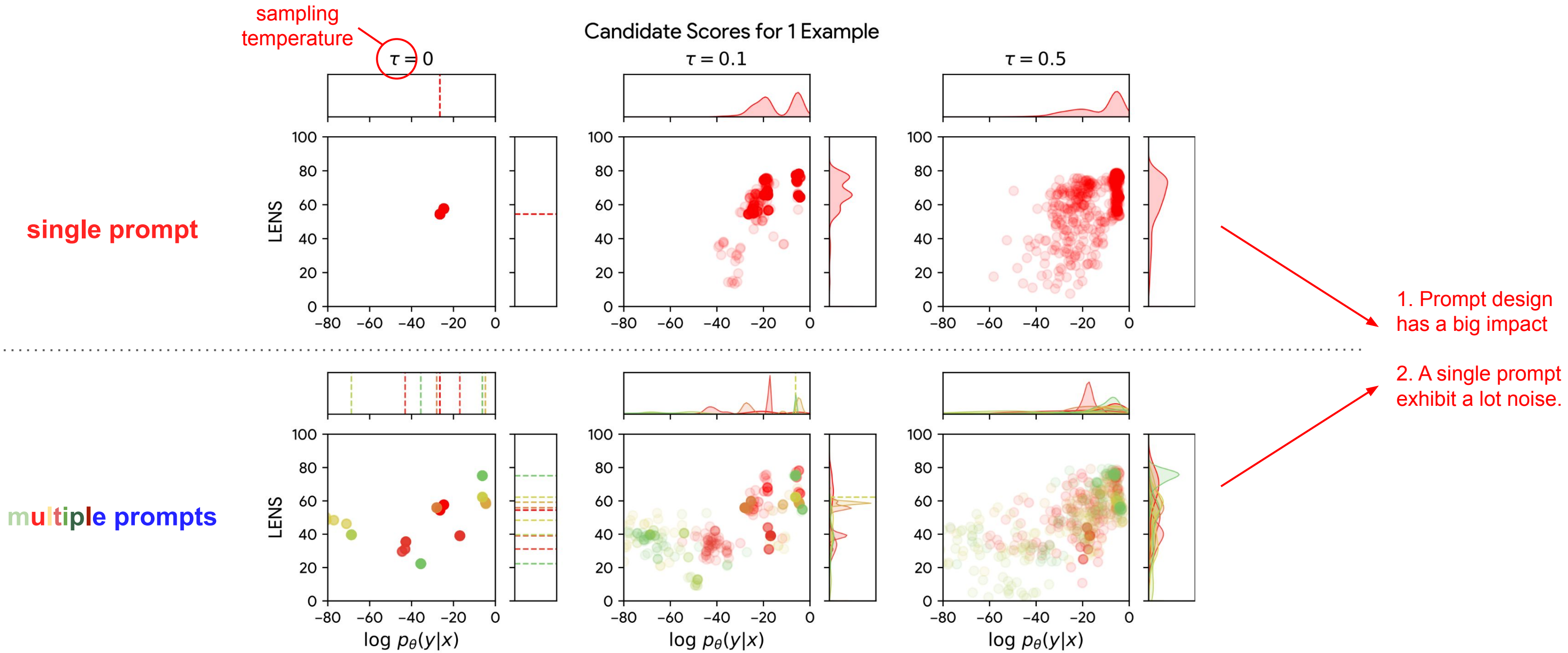
Diverse Prompting + MBR Decoding



Heineman et al. "Improving Minimum Bayes Risk Decoding with Multi-Prompt" (2024)

Minimum Bayes Risk (MBR) Decoding

Diverse Prompting + MBR Decoding

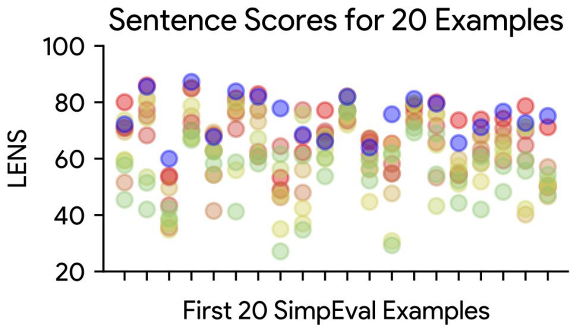


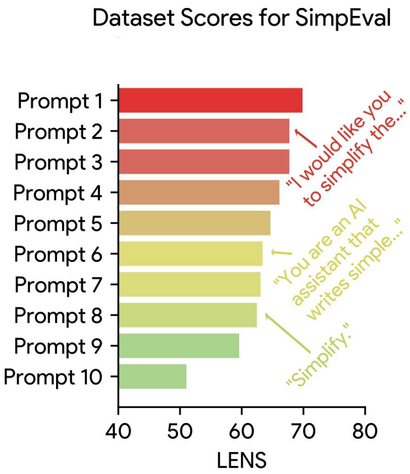
Heineman et al. "Improving Minimum Bayes Risk Decoding with Multi-Prompt" (2024)

Minimum Bayes Risk (MBR) Decoding

Diverse Prompting + MBR Decoding

- no single prompt consistently produces the highest quality sequences
- different prompts are most effective at different inputs





Heineman et al. "Improving Minimum Bayes Risk Decoding with Multi-Prompt" (2024)

Minimum Bayes Risk (MBR) Decoding

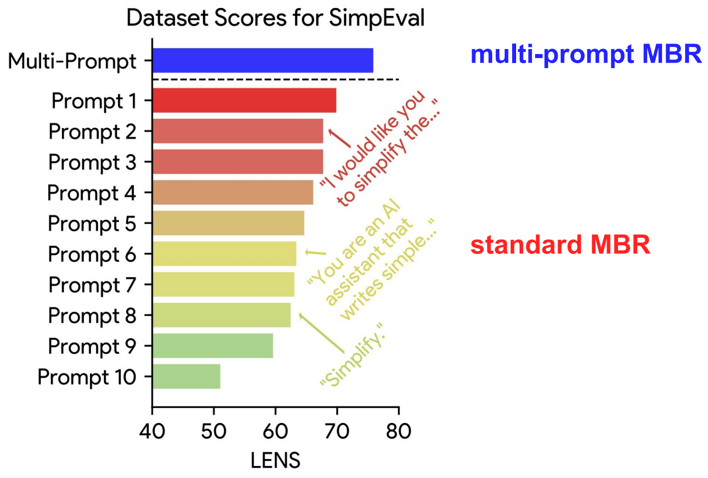
Diverse Prompting + MBR Decoding

- <u>However</u>, simply using many prompts may introduce too much noise
- Instead, estimate the probability distribution of prompts on a dev set, then

  (1) top-p prompt sampling

  (2) prompt selection:

  - closest similarity
  - greatest dissimilarity
  - k-NN cluster



Heineman et al. "Improving Minimum Bayes Risk Decoding with Multi-Prompt" (2024)

Minimum Bayes Risk (MBR) Decoding

Diverse Prompting + MBR Decoding

- consistent, further improvement over standard MBR across generation tasks
- works for both open-source and black-box LLMs

Code Generation ($\|\mathcal{H}\|=20$) – HumanEval (pass@1)		
StarCoder 2 15B	44.51	49.39 (+4.88)
CodeLlama 7B	37.80	40.85 (+3.05)
CodeLlama 13B	43.29	48.17 (+4.88)
CodeLlama 34B	45.73	52.44 (+6.71)
CodeLlama 70B	61.59	68.90 (+7.31)
GPT-3.5	68.29	73.78 (+5.49)
GPT-4	81.71	82.93 (+1.22)

Code Generation ($\|\mathcal{H}\|=20$) – HumanEval (pass@1)	standard MBR	multi-prompt MBR
StarCoder 2 15B	44.51	49.39 (+4.88)
CodeLlama 7B	37.80	40.85 (+3.05)
CodeLlama 13B	43.29	48.17 (+4.88)
CodeLlama 34B	45.73	52.44 (+6.71)
CodeLlama 70B	61.59	68.90 (+7.31)
GPT-3.5	68.29	73.78 (+5.49)
GPT-4	81.71	82.93 (+1.22)

Translation ($\|\mathcal{H}\|=100$) – WMT '22 En-Cs (COMET)		
WMT '22 Winners	91.9	–
MS Translate API	90.6	–
ALMA 7B R	89.17	89.94 (+0.77)
ALMA 13B R	89.41	90.45 (+1.04)
GPT-3.5	91.27	91.35 (+0.08)
GPT-4	92.24	92.47 (+0.23)

Heineman et al. "Improving Minimum Bayes Risk Decoding with Multi-Prompt" (2024)

Minimum Bayes Risk (MBR) Decoding

Intuition: the best output <u>not only</u> have high probability (same as *maximum likelihood*), <u>but also</u> is consistent or similar to the other candidate outputs.

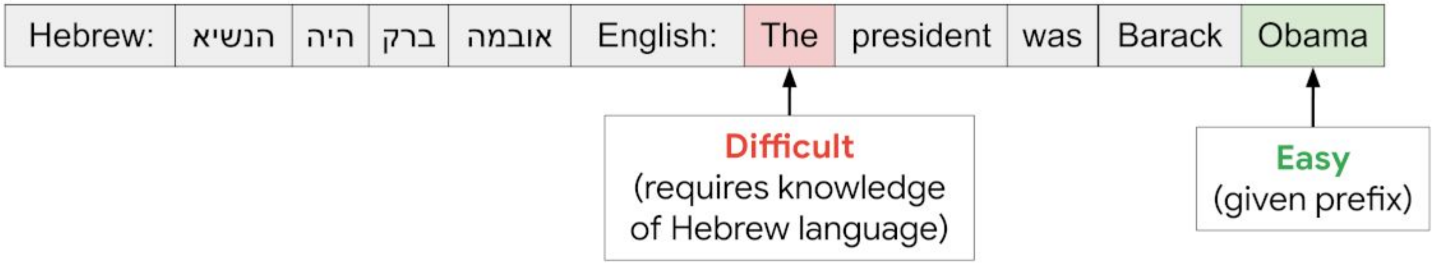
A number of widely used techniques with LLMs can be viewed as special cases of MBR

- self-consistency (Wang+ ’23)
- range voting (Borgeaud+ ’20)
- output ensembling (Denero+ ’10, Lorenzo+ ’23)
- density estimation (Kobayashi ’18)

Bertsch et al. “It’s MBR All the Way Down: Modern Generation Techniques Through the Lens of Minimum Bayes Risk” (Big Picture Workshop 2024)

Speculative Decoding

Intuition: Some tokens in the sequence are easier (can use a small LLM) to generate than others (ideally, use a larger LLM).



How to combine small and large LLMs to do this more efficiently?

Leviathan et al. “Fast Inference from Transformers via Speculative Decoding” (ICML 2023)

Speculative Decoding

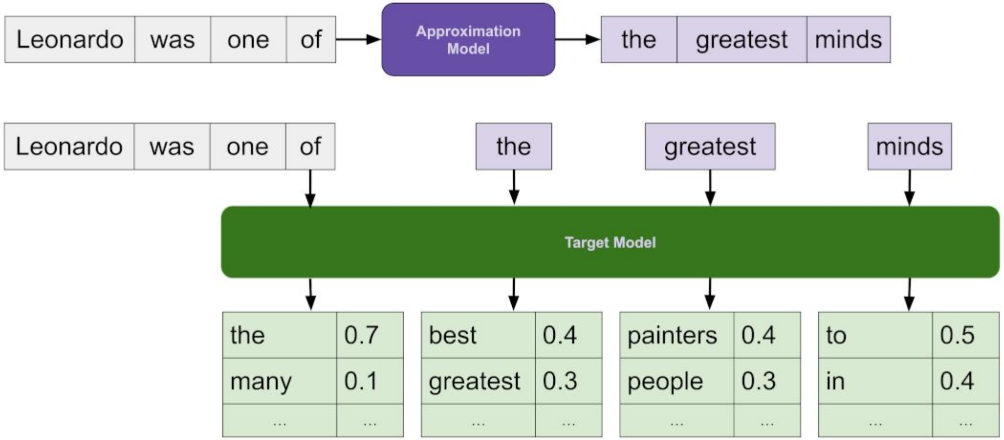
1. generate γ tokens by a small (approximation) model

Leviathan et al. “Fast Inference from Transformers via Speculative Decoding” (ICML 2023)

Speculative Decoding

1. generate γ tokens by a small (approximation) model

2. use a large (target) model to generate next-token distributions for all γ+1 prefixes



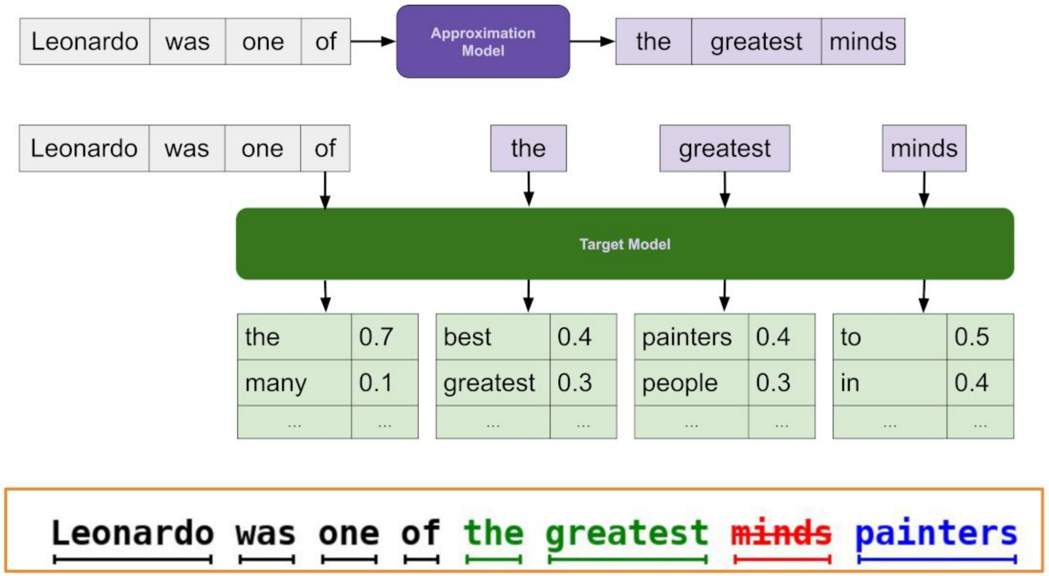
Leviathan et al. "Fast Inference from Transformers via Speculative Decoding" (ICML 2023)

Speculative Decoding

1. generate γ tokens by a small (approximation) model

2. use a large (target) model to generate next-token distributions for all γ+1 prefixes

3. Decide which tokens to **accept** or **reject** (with a probability) based on the large model, and **sample** one more token from the large model



Leviathan et al. "Fast Inference from Transformers via Speculative Decoding" (ICML 2023)

Collaborative Decoding

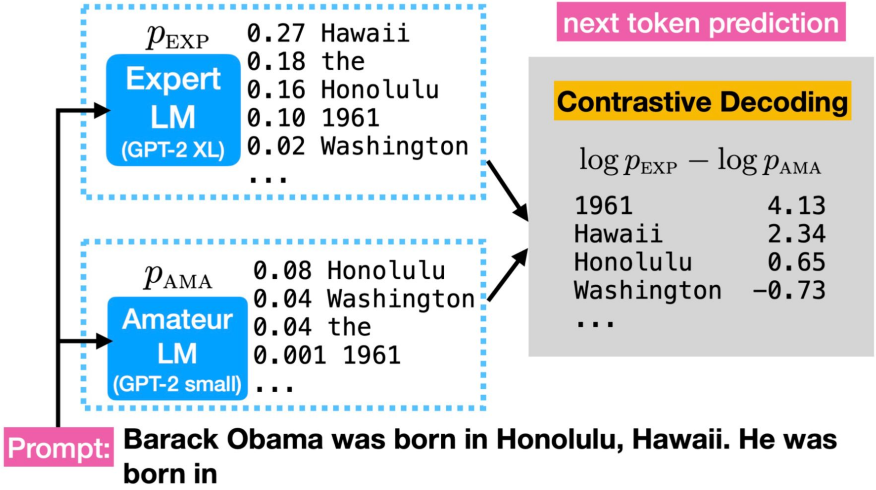
Intuition: The decision of deferral from a small model to a large model is learnt as a latent variable $Z_t \in \{0, 1, \ldots, M\}$ by optimizing the marginal likelihood:

$$P(X) = \prod_{t=1}^{T} \Big(\sum_{Z_t=0}^{M} P_\theta(Z_t|X_{<t}) P_{Z_t}(X_t|X_{<t}) \Big)$$

Shen et al. "Learning to Decode Collaboratively with Multiple Language Models" (ACL 2024)

Contrastive Decoding

Intuition: The failures of larger LLMs are even more prevalent in smaller LMs, thus the difference between the two can be a useful signal.



Continuations:

Greedy: **Hawaii.** He was born in Hawaii. He was born in Hawaii…

Nucleus: **Washington**, D.C., to Barack Obama and Michelle Robinson…

CD: **1961** to a Kenyan father, Barack Hussein Obama and a mother of American descent, Stanley Ann Dunham…

Li et al. "Contrastive Decoding: Open-ended Text Generation as Optimization" (ACL 2023)

Three Popular Methods for Generation

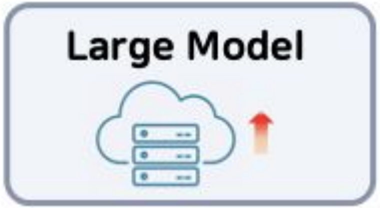
1. **Decoding**:
an inference-time solution to optimize LLM outputs
(Survey by Welleck+ 2024 & Bertsch+, 2023; MBR with Multi-Prompt by Heineman+, 2024)

2. **Distillation**:
reproduce GPT-4 performance by small open-source LLMs
(Edit-based generation by Dou+ 2024; Feedback to refine LLM outputs by Wadhwa+ 2024)

Three Popular Methods for Generation

2. Distillation:

reproduce GPT-4 performance by small open-source LLMs



(prompt input1, output1)
(prompt input2, output2)
(prompt input3, output3)
(prompt input4, output4)
...

generate training data

fine-tune

Span-level revision distilled by LLM

Task: Self-disclosure abstraction

Dou et al. “Reducing Privacy Risks in Online Self-Disclosures with Language Models” (ACL 2024)

Span-level revision distilled by LLM

Task: Self-disclosure abstraction

Definition: rephrase self-disclosures (personal information) with less specific details while preserving the content utility

Im 16F I think I want to be a bi M

→ I am exploring my sexual identity

→ I have a desire to explore new options

→ I am attracted to the idea of exploring different gender identities

Dou et al. “Reducing Privacy Risks in Online Self-Disclosures with Language Models” (ACL 2024)

Span-level revision distilled by LLM

Task: Self-disclosure abstraction

Definition: rephrase self-disclosures (personal information) with less specific details while preserving the content utility

Im 16F I think I want to be a bi M

- I am exploring my sexual identity
- I have a desire to explore new options
- I am attracted to the idea of exploring different gender identities

Why distillation? writing diverse abstractions is challenging for human annotators

Dou et al. “Reducing Privacy Risks in Online Self-Disclosures with Language Models” (ACL 2024)

Span-level revision distilled by LLM

Prompt

Your task is to abstract the given 'disclosure span' in the sentence. <more instruction>

Example 1:
Sentence: "Should I submit a 1470 SAT score to Carnegie Mellon and Duke?"
Disclosure Span to Revise: "1470 SAT score"
Rationale: <rationale>
Abstracted Spans: {"span 1": "a high 1400-range SAT score", "span 2": "an SAT score in the upper 1400s", "span 3": "an SAT score above 1450"}

<2 more examples>

First, provide a rationale explaining why the disclosure span needs abstraction. Then, offer three abstracted alternatives in a JSON format like this: {'span 1': xxx, 'span 2': xxx, 'span 3': xxx}.

Criteria:
<3 criteria>

Sentence: "{sentence}"
Disclosure Span to Revise: "{span}"
Rationale:

Prompt

Contains three in-context examples

GPT-3.5 / 4

Rationale and abstracted spans

Span-level revision distilled by LLM

Use GPT-3.5 to generate abstractions of 780 instances for distilling Llama2 7B

Three training methods

Sampling three times from a model that generates one abstraction at a time

End-to-end training model to generate three abstractions all at once

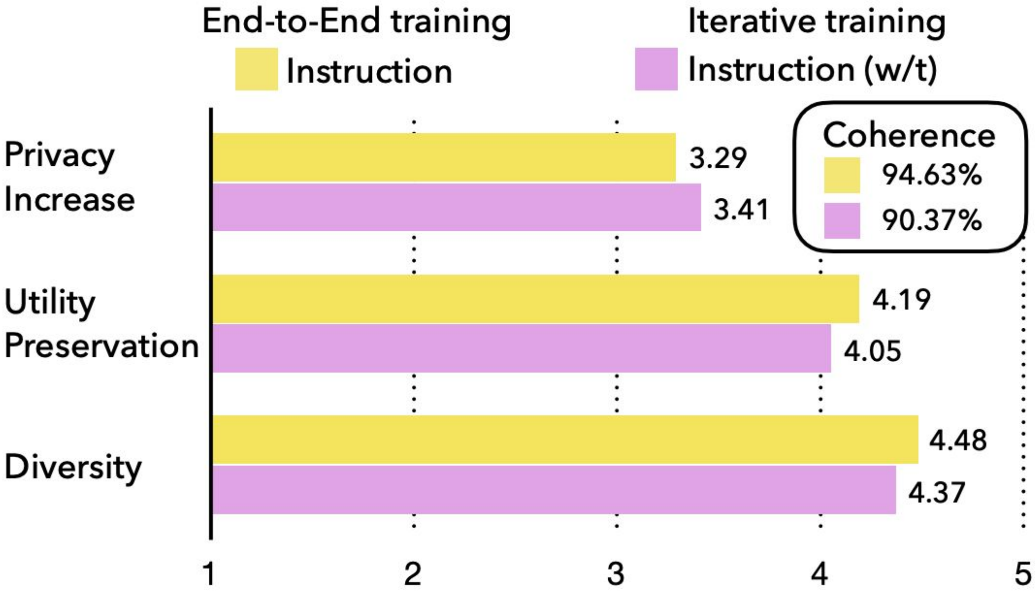
Iterative training model to generate new abstraction given the previous ones.

input → *A*, *input*+*A* → *B*,
input+*A*+*B* → *C*

Dou et al. "Reducing Privacy Risks in Online Self-Disclosures with Language Models" (ACL 2024)

Span-level revision distilled by LLM

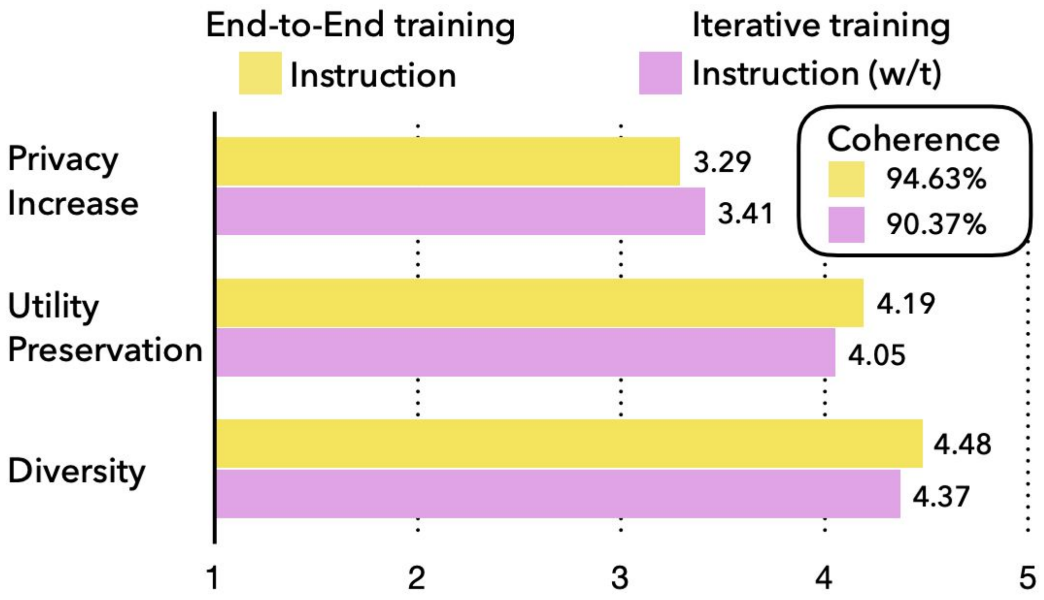
Human evaluation on three aspects with Likert scale



Dou et al. “Reducing Privacy Risks in Online Self-Disclosures with Language Models” (ACL 2024)

Span-level revision distilled by LLM

Human evaluation on three aspects with Likert scale



The distilled Llama2 7B can generate **diverse** abstractions that **moderately reduce privacy risks** while **maintaining high utility**.

Dou et al. "Reducing Privacy Risks in Online Self-Disclosures with Language Models" (ACL 2024)

Three Popular Methods for Generation

1. Decoding:
an inference-time solution to optimize LLM outputs
(Survey by Welleck+ 2024 & Bertsch+, 2023; MBR with Multi-Prompt by Heineman+, 2024)

2. Distillation:
reproduce GPT-4 performance by small open-source LLMs
(Edit-based generation by Dou+ 2024; Feedback to refine LLM outputs by Wadhwa+ 2024)

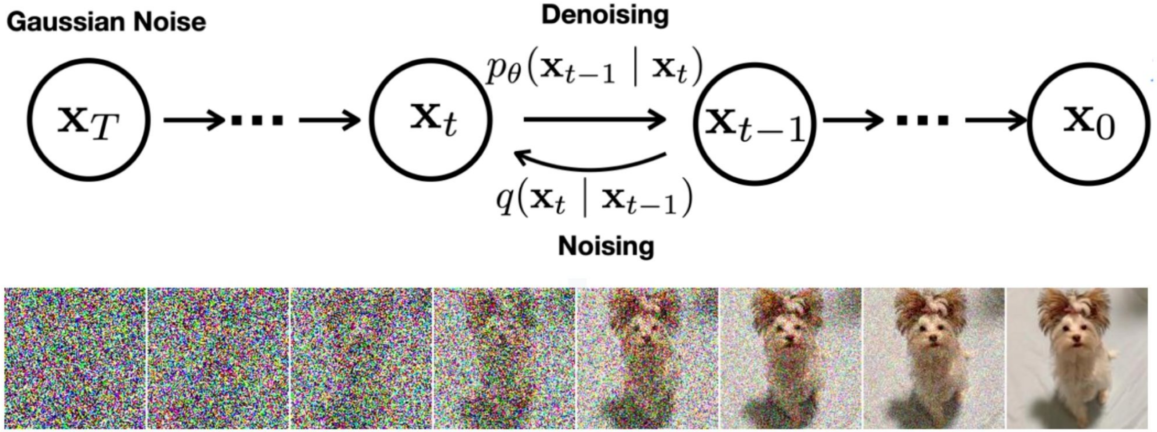
3. Diffusion:
an alternative to Transformer-based LLM
(Diffusion-LM by Li+ 2022; DiffuSeq by Gong+, 2022; SeqDiffuSeq by Yuan+, 2024)

Three Popular Methods for Generation

3. Diffusion:

an alternative to Transformer-based LLM

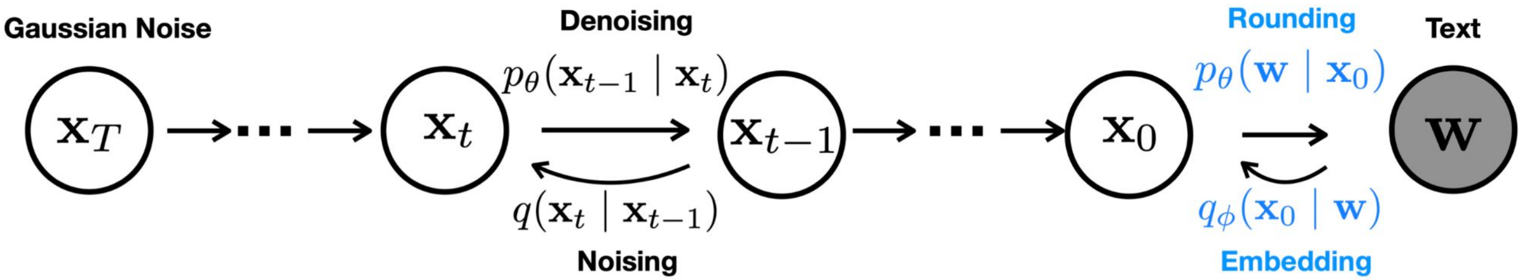
Learning to generate data by iteratively denoising -- a big success in computer vision!



Ho et al. "Denoising Diffusion Probabilistic Models" (NeurIPS 2020)

Diffusion-LM (Li+ '22)

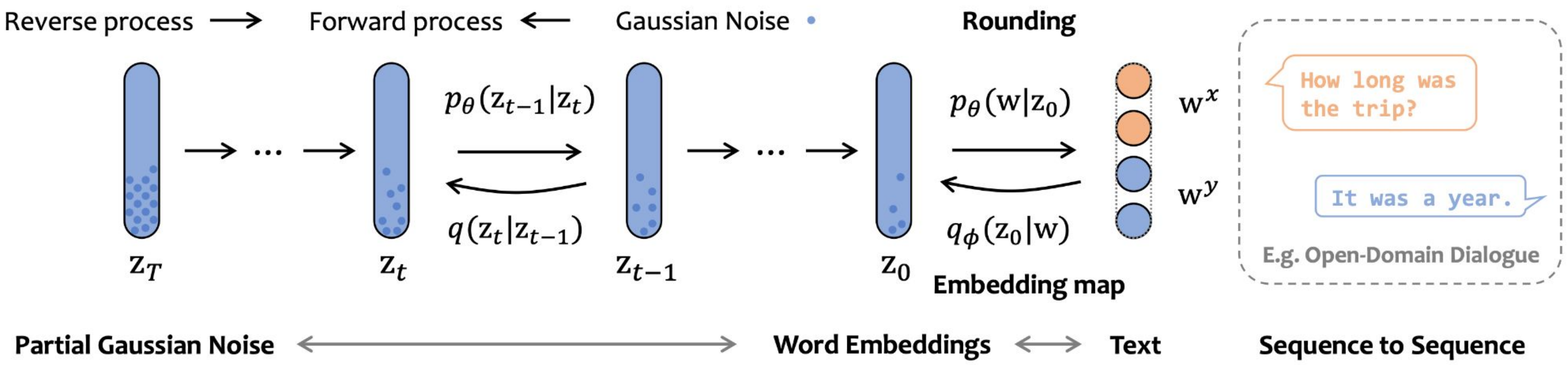
Several modifications to standard diffusion model to make it work on discrete text data (embedding/rounding steps), instead of the continuous image data.



Li et al. "Diffusion-LM Improves Controllable Text Generation" (ICLR 2022)

DiffuSeq (Gong+ '22)

Extended Diffusion-LM to seq-to-seq generation tasks, by combining the source $\mathbf{w}^x$ and the target $\mathbf{w}^y$ into a continuous space $\mathbf{z}_0$. Only impose noising on $\mathbf{y}_t$.



(experiments on dialogue, question generation, text simplification, paraphrasing)

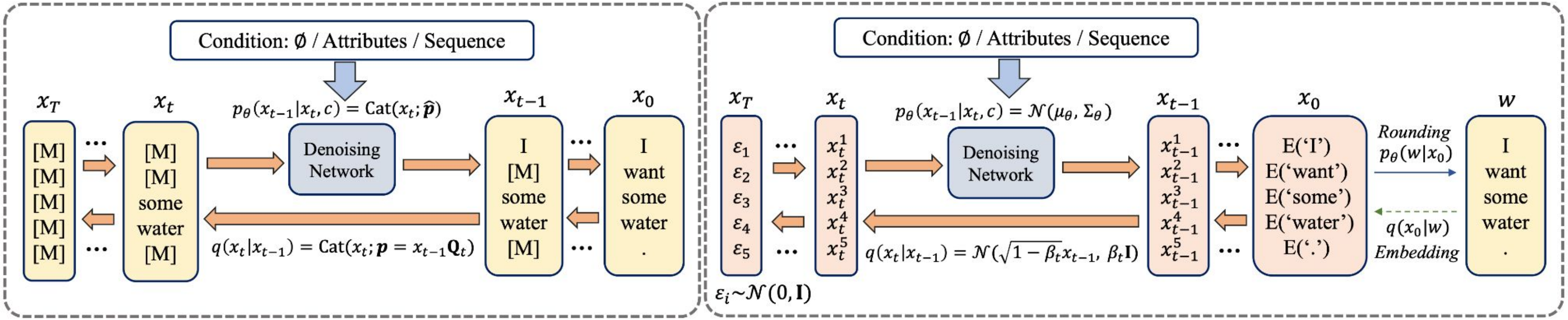
Gong et al. "DiffuSeq: Sequence to Sequence Text Generation with Diffusion Models" (ICLR 2022)

Some other text diffusion models work directly in discrete space.

4 key designs: denoising network, noise schedule, objective function, and conditioning strategy

Diffusion Models for Non-autoregressive Text Generation: A Survey

Yifan Li[1], **Kun Zhou**[2,3], **Wayne Xin Zhao**[1,3] *, and **Ji-Rong Wen**[1,2,3]
[1]Gaoling School of Artificial Intelligence, Renmin University of China
[2]School of Information, Renmin University of China
[3]Beijing Key Laboratory of Big Data Management and Analysis Methods



(a) Discrete text diffusion model.

(b) Continuous text diffusion model.