

# TEncDM: Понимание свойств диффузионной модели в пространстве кодировок языковых моделей

Представленная работа описывает Text Encoding Diffusion Model (TEncDM) — новаторский подход к моделированию текста с помощью диффузионной модели, работающей в пространстве кодировок языковой модели. TEncDM использует кодировки, которые содержат больше контекстной информации и улучшают качество предсказаний модели.

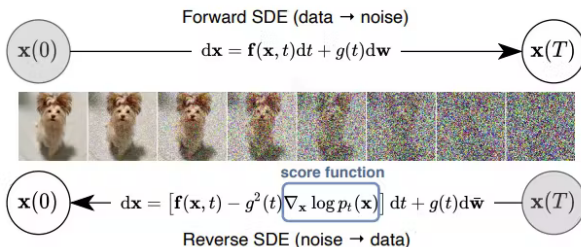
Авторегрессионные модели, такие как GPT-4 [5] и Llama 3 [2], демонстрируют высокое качество в генерации текста, но имеют два значительных недостатка:

- ▶ Невозможность корректировать ошибки, допущенные на ранних этапах.
- ▶ Замедление процесса генерации для длинных последовательностей.

Диффузионные модели предлагают альтернативный метод, генерируя текст параллельно и позволяя ускорить процесс.

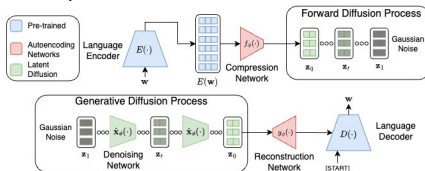
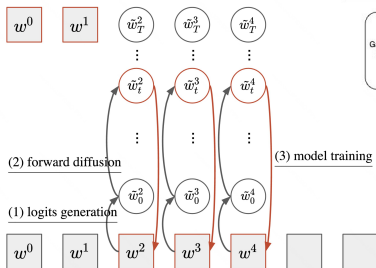
# Генерация с помощью диффузии

Задача генерации текста заключается в построении текста, который удовлетворяет заданному условию, например, теме или стилю. Диффузионное моделирование предполагает преобразование случайного шума в структурированный текст.



## Examining Diffusion Model Architectures for Text Generation: Challenges and Autoregressive Comparison

- Diffusion models are currently state-of-the-art (SOTA) in generating images and audio, but they still lag behind autoregressive models (AR-LM) for text generation.

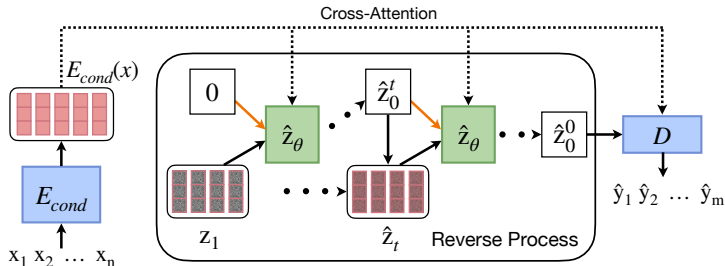
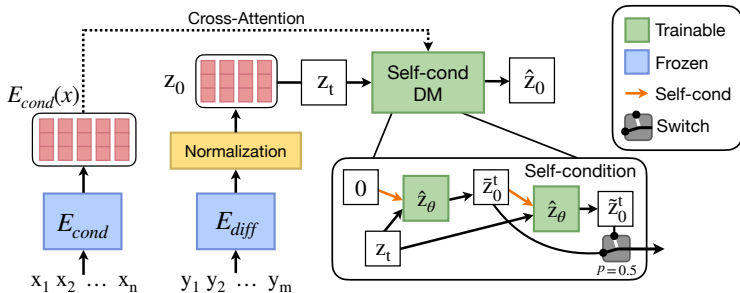


- A key drawback of autoregression is that if an error occurs at any step, it propagates through the following tokens, and these models are resource-intensive since they generate text one token at a time. Diffusion models offer the advantage of controllability through auxiliary models, but applying them to text is more challenging due to issues with data representation and compatibility with existing classifiers.

Методология TEncDM включает несколько ключевых компонентов:

- ▶ Кодировщик, использующий контекстуальные кодировки.
- ▶ Диффузионная модель, контролирующая добавление шума.
- ▶ Декодер, учитывающий контекст для каждого токена.
- ▶ Самоконтроль для повышения точности предсказаний.

# Обзор фреймворка



Для проверки эффективности TEncDM были проведены эксперименты на задачах условной генерации текста, включая перефразирование, суммаризацию и упрощение текста. Были использованы датасеты QQP [1] (вопросы-перефразирования), XSum [4] (экстремальная суммаризация) и Wiki-Auto [3] (упрощение текста)



## Сравнение энкодеров

Encoder	ppl ↓	mem ↓	div ↑	mauve ↑
ROCStories				
BERT emb	48.9 <sub>.36</sub>	.371 <sub>.003</sub>	.324 <sub>.002</sub>	.600 <sub>.016</sub>
BERT	29.1 <sub>.89</sub>	.453 <sub>.003</sub>	.295 <sub>.002</sub>	<b>.762</b> <sub>.043</sub>
RoBERTa	<b>28.3</b> <sub>.33</sub>	.443 <sub>.003</sub>	.302 <sub>.002</sub>	.647 <sub>.019</sub>
T5	31.3 <sub>.54</sub>	.427 <sub>.003</sub>	.312 <sub>.004</sub>	.706 <sub>.024</sub>
BART	34.1 <sub>.52</sub>	.441 <sub>.006</sub>	.299 <sub>.005</sub>	.705 <sub>.030</sub>
Source text	21.7	.365	.403	.876
Wikipedia				
BERT emb	156.1 <sub>1.8</sub>	.263 <sub>.004</sub>	.517 <sub>.002</sub>	.378 <sub>.055</sub>
BERT	<b>104.4</b> <sub>2.1</sub>	.286 <sub>.002</sub>	.504 <sub>.003</sub>	<b>.874</b> <sub>.011</sub>
Source text	37.3	.122	.615	.957





Таблица: Сравнение энкодеров

Decoder	ppl ↓	mem ↓	div ↑	mauve ↑
ROCStories				
MLP	39.7 <sub>3.38</sub>	.444 <sub>.002</sub>	.297 <sub>.004</sub>	.716 <sub>.074</sub>
+ $Cor(z_0)$	31.2 <sub>.33</sub>	.448 <sub>.002</sub>	.293 <sub>.003</sub>	.739 <sub>.051</sub>
Transformer	34.2 <sub>.29</sub>	.445 <sub>.001</sub>	.295 <sub>.003</sub>	.714 <sub>.037</sub>
+ $Cor(z_0)$	<b>29.1</b> <sub>.89</sub>	.453 <sub>.003</sub>	.295 <sub>.002</sub>	<b>.762</b> <sub>.043</sub>
Wikipedia				
Transformer	180.6 <sub>3.2</sub>	.261 <sub>.001</sub>	.511 <sub>.001</sub>	.526 <sub>.025</sub>
+ $Cor(z_0)$	<b>104.4</b> <sub>2.1</sub>	.286 <sub>.002</sub>	.504 <sub>.003</sub>	<b>.874</b> <sub>.011</sub>

Таблица: Сравнение декодеров

TEncDM показывает высокие результаты в задачах генерации текста и превосходит традиционные модели. Улучшение качества предсказаний достигается за счет контекстуальных кодировок и эффективного денойзинга.

# Список литературы

-  Zihang Chen, Hongbo Zhang, Xiaoji Zhang, and Leqi Zhao.  
Quora question pairs.  
2017.
-  Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and et al.  
The llama 3 herd of models, 2024.
-  Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu.  
Neural crf model for sentence alignment in text simplification.  
*In Proceedings of the Association for Computational Linguistics (ACL)*, 2020.
-  Shashi Narayan, Shay B. Cohen, and Mirella Lapata.  
Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization.  
*In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi*