Mask-Predict: Parallel Decoding of Conditional Masked Language Models

1. В чем заключается проблема мультимодальности (multimodality problem), возникающая при неавторегрессионном декодировании последовательностей?
2. В чем заключается преимущество процедуры декодирования Conditional Masked Language Models перед авторегрессионным декодированием? С чем оно связано?
3. Как изменили стандартную архитектуру трансформера авторы Mask-Predict: Parallel Decoding of Conditional Masked Language Models?

On the Discrepancy between Density Estimation and Sequence Generation

1. Перед вами таблица с результатами экспериментов для трёх моделей перевода на одних и тех же данных. Предположите, из каких семейств эти модели и объясните, почему вы так думаете.

BLEU		LL	
RAW	Dist.	Raw	Dist
24.54	24.94	-1.77	-2.36
28.18	27.86	-1.44	-2.19
29.39	28.29	-1.35	-2.23

2. Есть ли корреляция между BLEU и логарифмом правдоподобия для моделей машинного перевода?
3. Определите вероятностную модель со скрытыми переменными (latent variable model), на которую опирается трансформер при неавторегрессионном переводе. Какую целевую функцию используют для её обучения?

Scaling Laws for Neural Language Models

1. Запишите любые три ключевых вывода, которые можно сделать из исследования Scaling Laws for Neural Language Models.
2. Как связана целевая функция языковой модели на тестовой выборке с размером модели и размером обучающей выборки?
3. Допустим мы знаем, что оптимальный размер языковой модели, которая не переобучается на данных размера D1 равен N1. Запишите оценку на размер модели на данных размера D2, если мы не хотим переобучаться.

The Curious Case of Neural Text Degeneration

1. Опишите любые три метода декодирования для авторегрессионных моделей, упомянутые в докладе The Curious Case of Neural Text Degeneration.
2. В чем заключается главная проблема top-k сэмплирования, и как Nucleus Sampling ее решает?
3. Какие проблемы возникают при жадном декодировании последовательностей в авторегрессионных моделях?
4. Опишите алгоритм вычисления метрики HUSE из работы The Curious Case of Neural Text Degeneration.

Electra: Pre-Training Text Encoders As Discriminators Rather Than Generators

1. Какую вспомогательную задачу решает ELECTRA для настройки модели?
2. Как выглядит целевая функция, которую оптимизирует модель ELECTRA?
3. При анализе предложенного решения авторы ELECTRA провели несколько сравнительных экспериментов, рассмотрев разные модификации настройки ELECTRA. Опишите один из таких экспериментов (один из трех), объясните чем мотивирована постановка этого эксперимента.