1. **Language Models are Unsupervised Multitask Learners**
    1.1. Каким способом реализуется one-shot learning для различных узких задач в модели GPT-2? Как бы вы сформулировали входные данные для модели, если нужно предсказать прошедшее время для глагола _verb_, и вы знаете несколько пар <verb_in_present, verb_in_past>?
    1.2. В чем особенность механизма Attention, применяемого в модели GPT-2 при обучении? Какое свойство достигается при использовании данного механизма?
    1.3. Какой смысл метрики perplexity? Как метрика связана с правдоподобием тестовой выборки? Запишите формулу. Допустим, у вас есть две модели с результатами perplexity 500 и 100, какая из них лучше?
2. **Stochastic Beams and Where to Find Them:**
    2.1. Как получить реализацию гумбелевской случайной величины с параметром **x**, используя реализацию равномерной случайной величины? Приведите и поясните формулу.
    2.2. Что такое “Gumbel-max trick”? На каком свойстве распределения Гумбеля он основывается?
    2.3. Опишите алгоритм сэмплирования k значений категориальной случайной величины без возвращения с использованием распределения Гумбеля.
    2.4. Как в дереве связаны зашумленные вероятности (“perturbed probabilities”) в узле и в его детях?
3. **Breaking the Softmax Bottleneck: A High-Rank RNN Language Model**
    3.1. Опишите суть проблемы softmax bottleneck. Отвечая на вопрос, необходимо упомянуть связь матричных рангов
    3.2. Написать и пояснить формулу вероятности для смеси софтмаксов.
    3.3. В чем различие смеси софтмаксов и смеси контекстов? Почему смесь контекстов показывает худшее качество?
4. **Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned**
    4.1. Какие метрики для классификации голов используются и как они вычисляются.
    4.2. Выпишите предложенную в статье классификацию attention heads. Для одного из них напишите метод классификации.
    4.3. Выпишите функцию потерь, которая используется для обучения трансформера для разреживания.
5. **Ordered Neurons: Integrating Tree Structures into Recurrent Neural Networks**
    5.1. В чём заключается мотивация метода “Ordered Neurons” для задач обработки естественного языка?
    5.2. Опишите предлагаемую в статье функцию активации cumax(), для какой дискретной случайной величины она является математическим ожиданием?
    5.3. В чём заключаются отличия стандартной ячейки LSTM и предложенной ячейки ON-LSTM?