Детекция галлюцинаций больших языковых моделей

Задача: выделить фрагменты-галлюцинации - классификация токенов (sequence labeling). $x_i \to y_i = \{0, 1\}$. Если $y_i = 1$, то i-й токен считается галлюцинацией.

Эксперимент 1: instruction-based подход. Использована генеративная модель `Meta-Llama-3-8B` с промптом - детектировать галлюцинации. Выборка из 50 размеченных QA сообщений на английском.

Результаты: распределение IoU на 50 объектах - рис. 1 Средний IoU: **0.3975**

Будущий эксперимент 2: Анализ временного ряда логитов токенов.

Будущий эксперимент 3: дообучение Token classification модели минимизацией бинарной кросс-энтропии:

$$\mathcal{L}(y, \hat{y}) = \sum_i logloss(y_i, \hat{y}_i) \to \min$$

$$logloss(y, \hat{y}) = -ylog(\hat{y}) - (1-y)log(1-\hat{y})$$

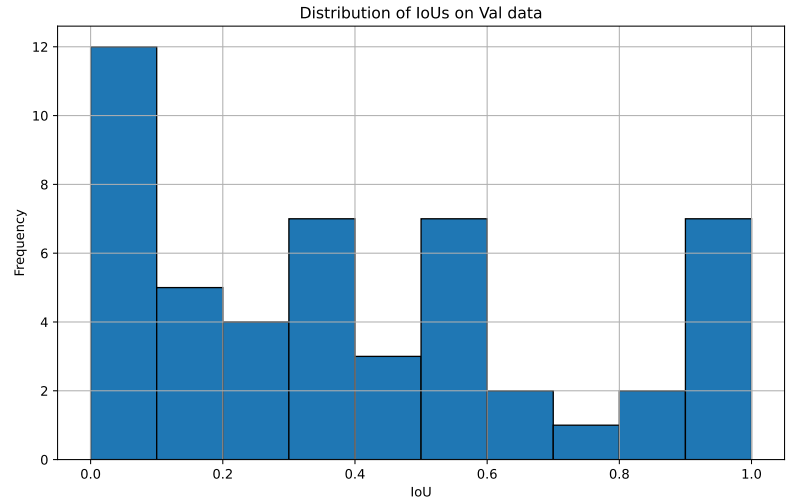

Рис.: Результаты эксперимента 1

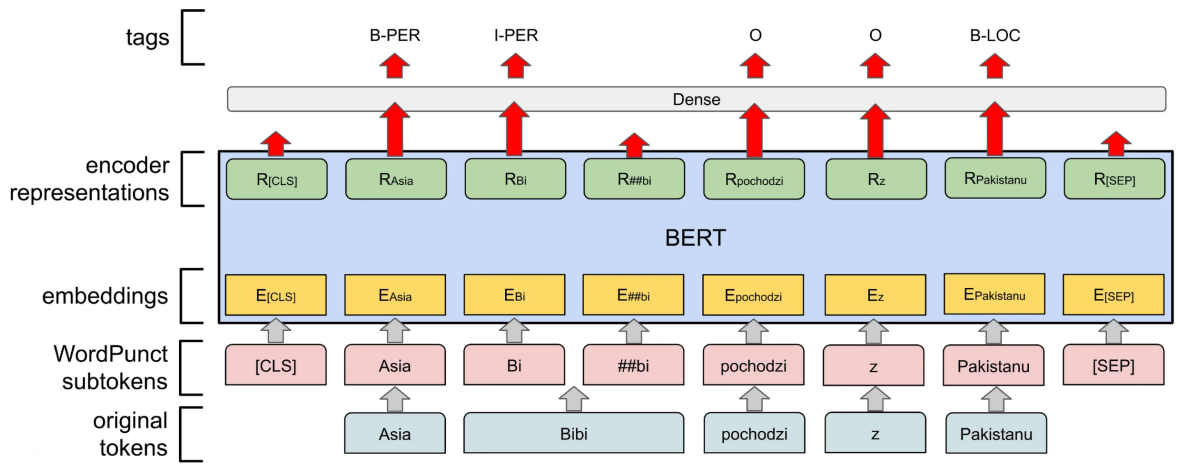

Рис.: Планы на эксперимент 2. Архитектура BERT NER