Курсовая работа “Прогнозирование критических событий в моделях песчаной кучи БТВ и Манна”

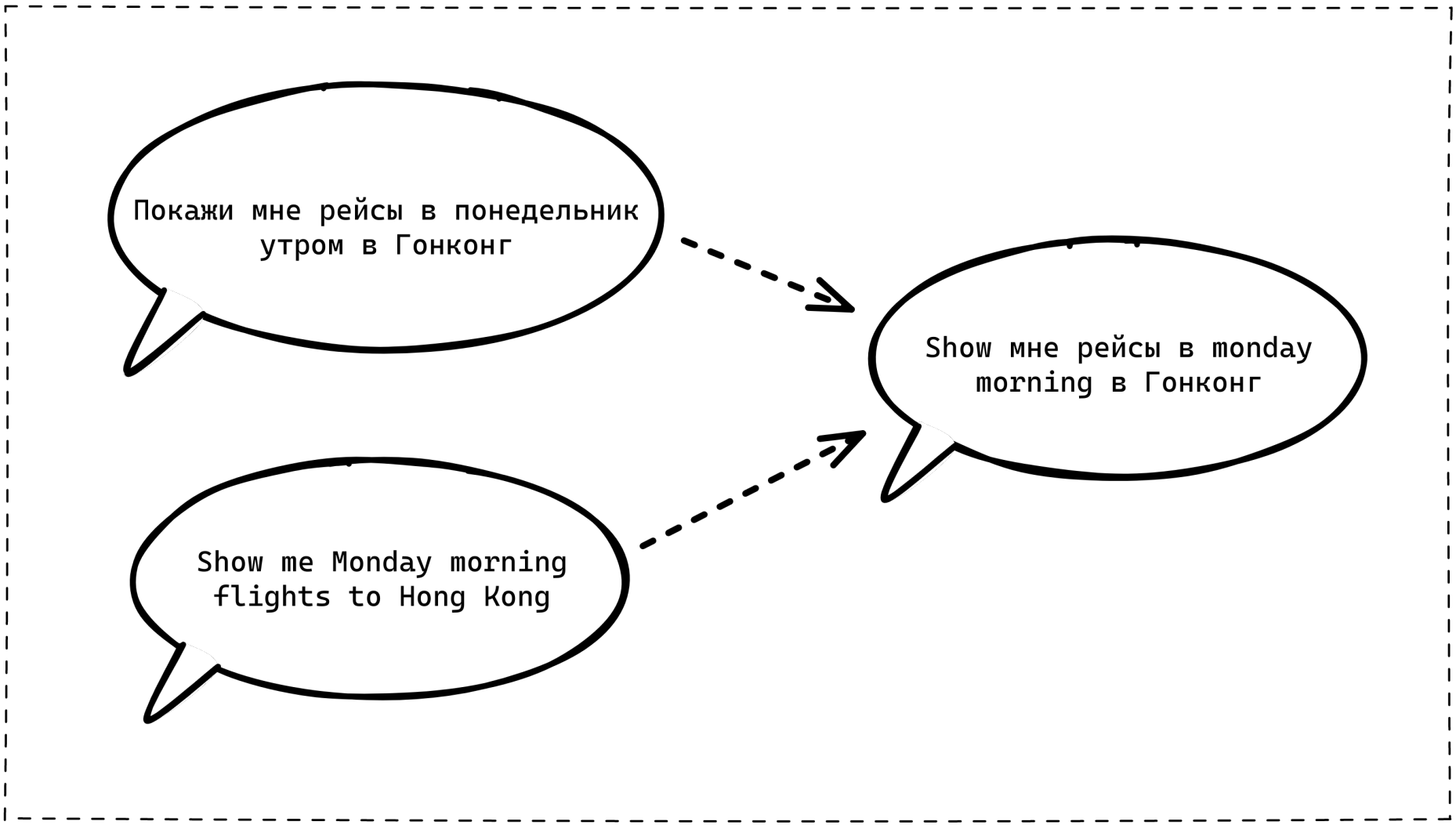
Сапожников Денис Сергеевич БПМИ 192

Руководитель КР: Попов Виктор Юрьевич

Соруководитель КР: Шаповал Александр Борисович

5 июня 2022 г.

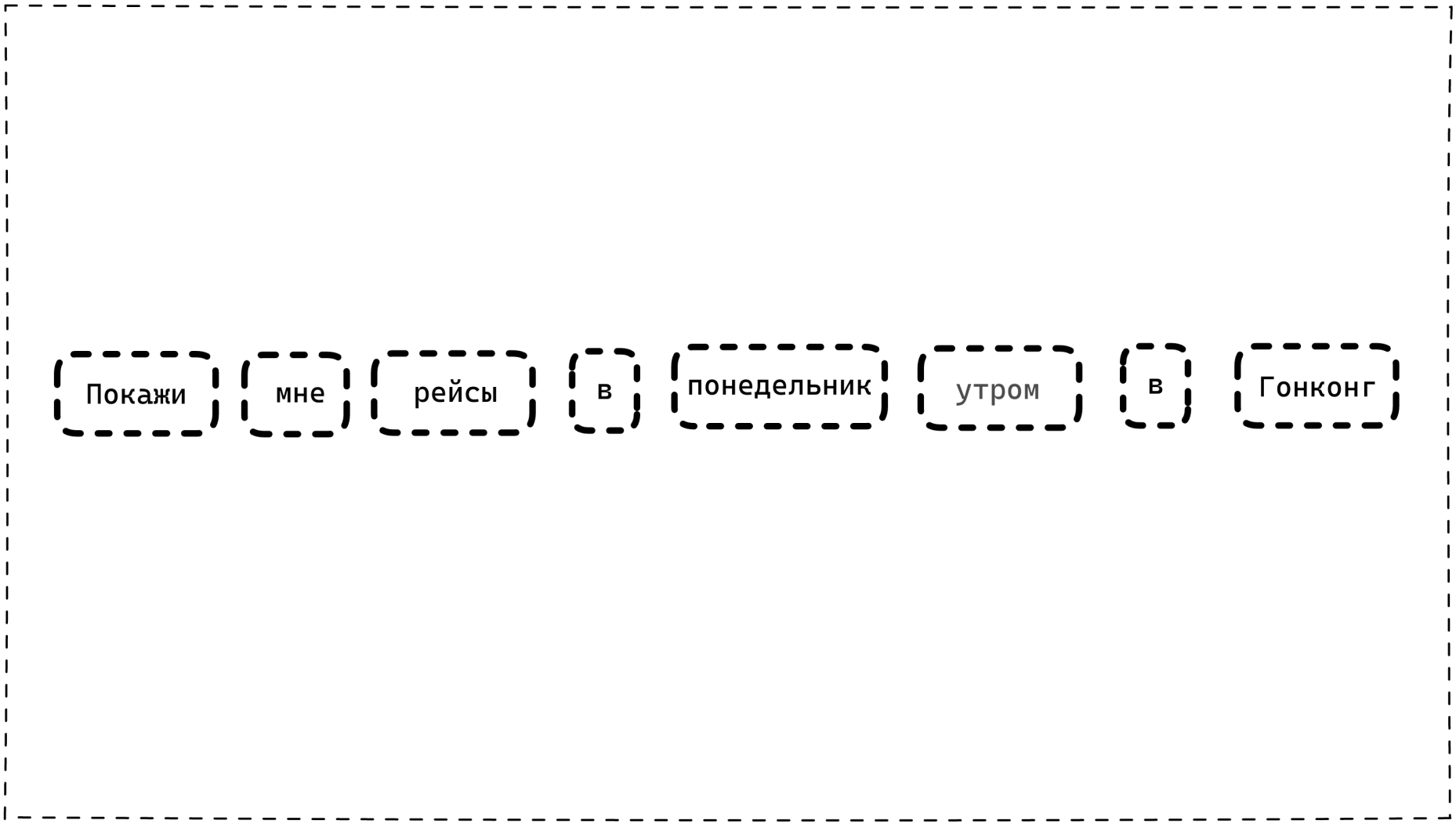
Модель песчаной кучи



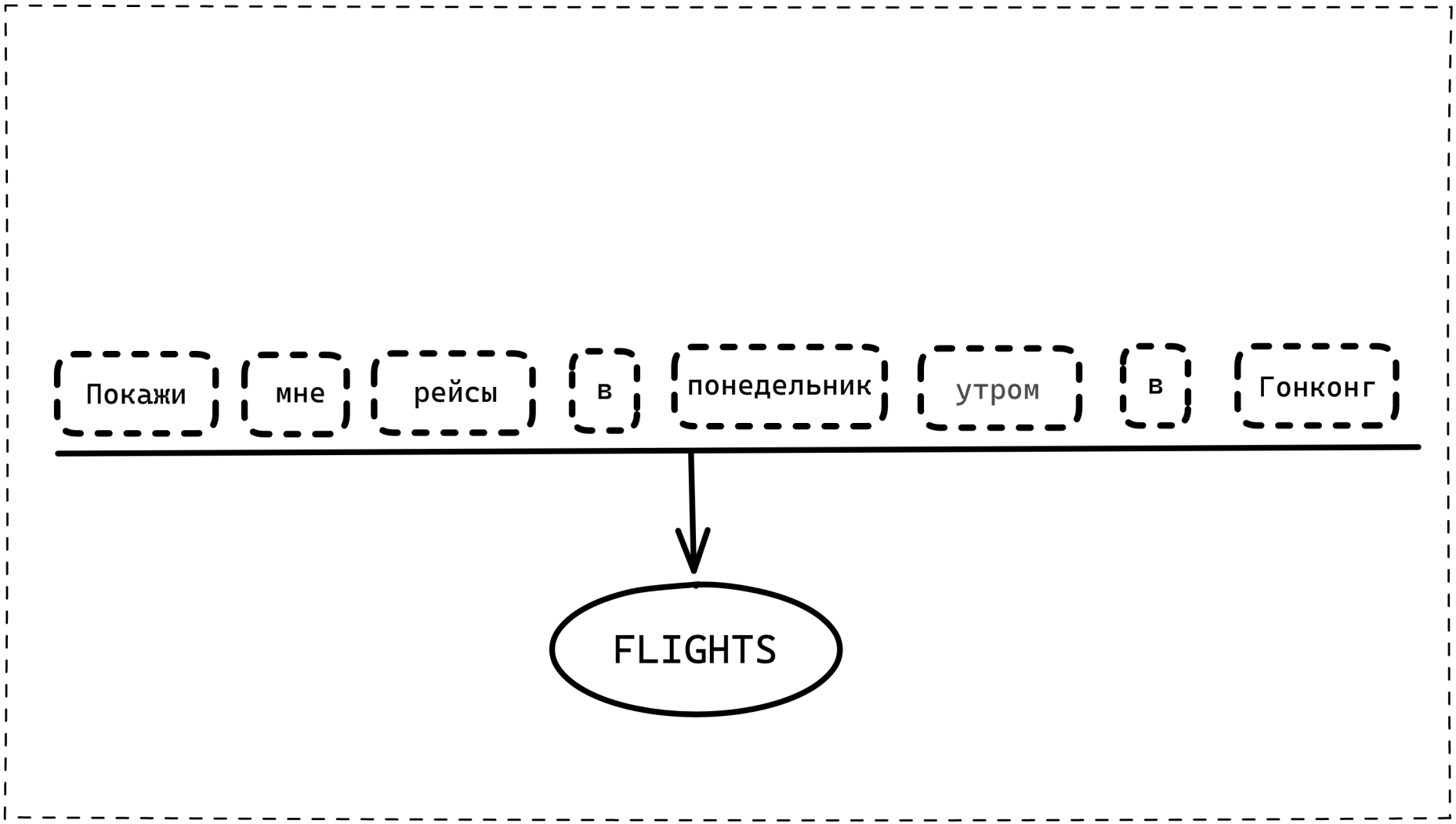
Имитация смешения кодов

- Хотим оценить качество моделей на смешении кодов
- Данных со смешением кодов мало
- Качество после атак как нижняя оценка

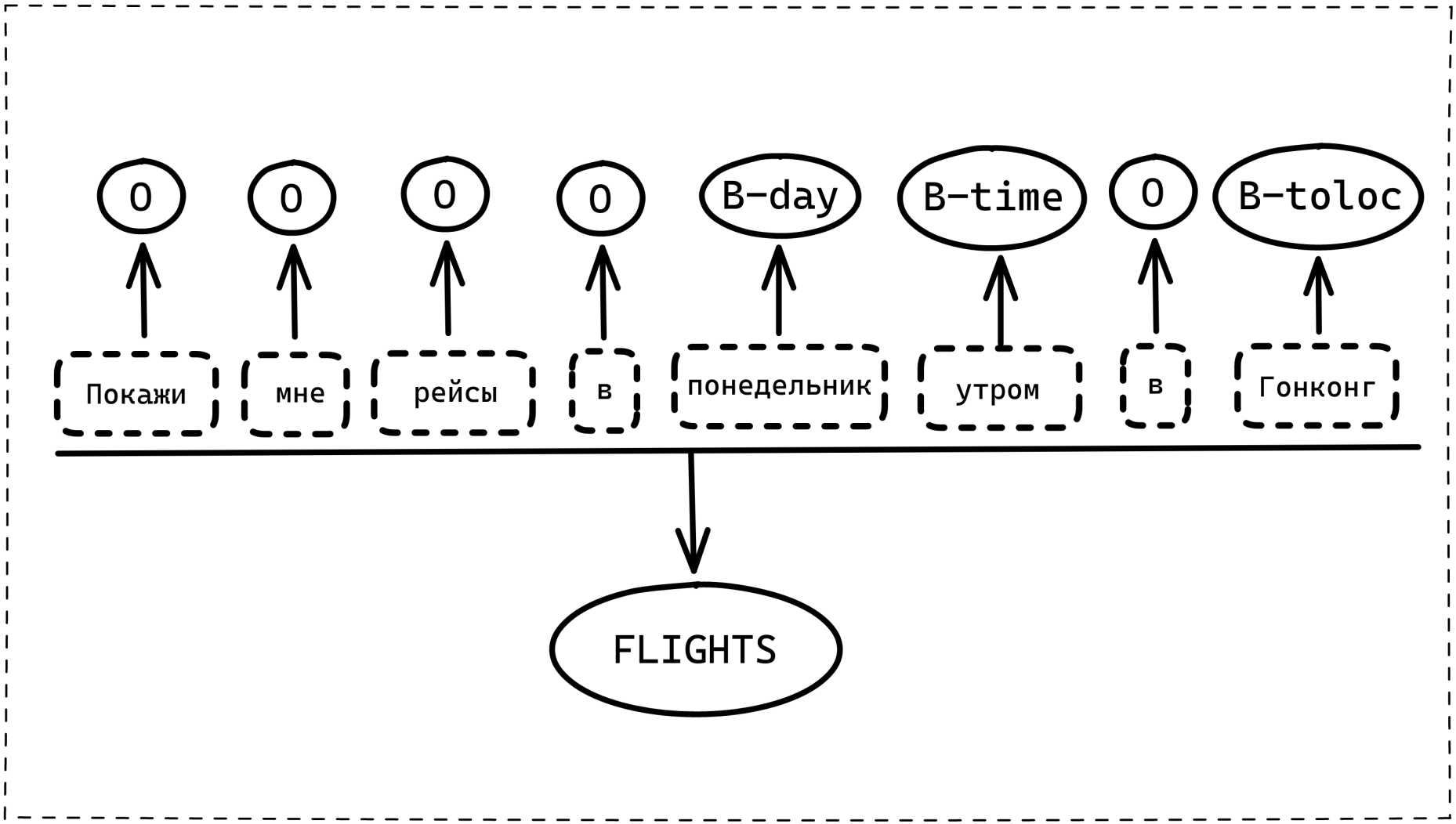
Постановка задачи



Постановка задачи



Постановка задачи

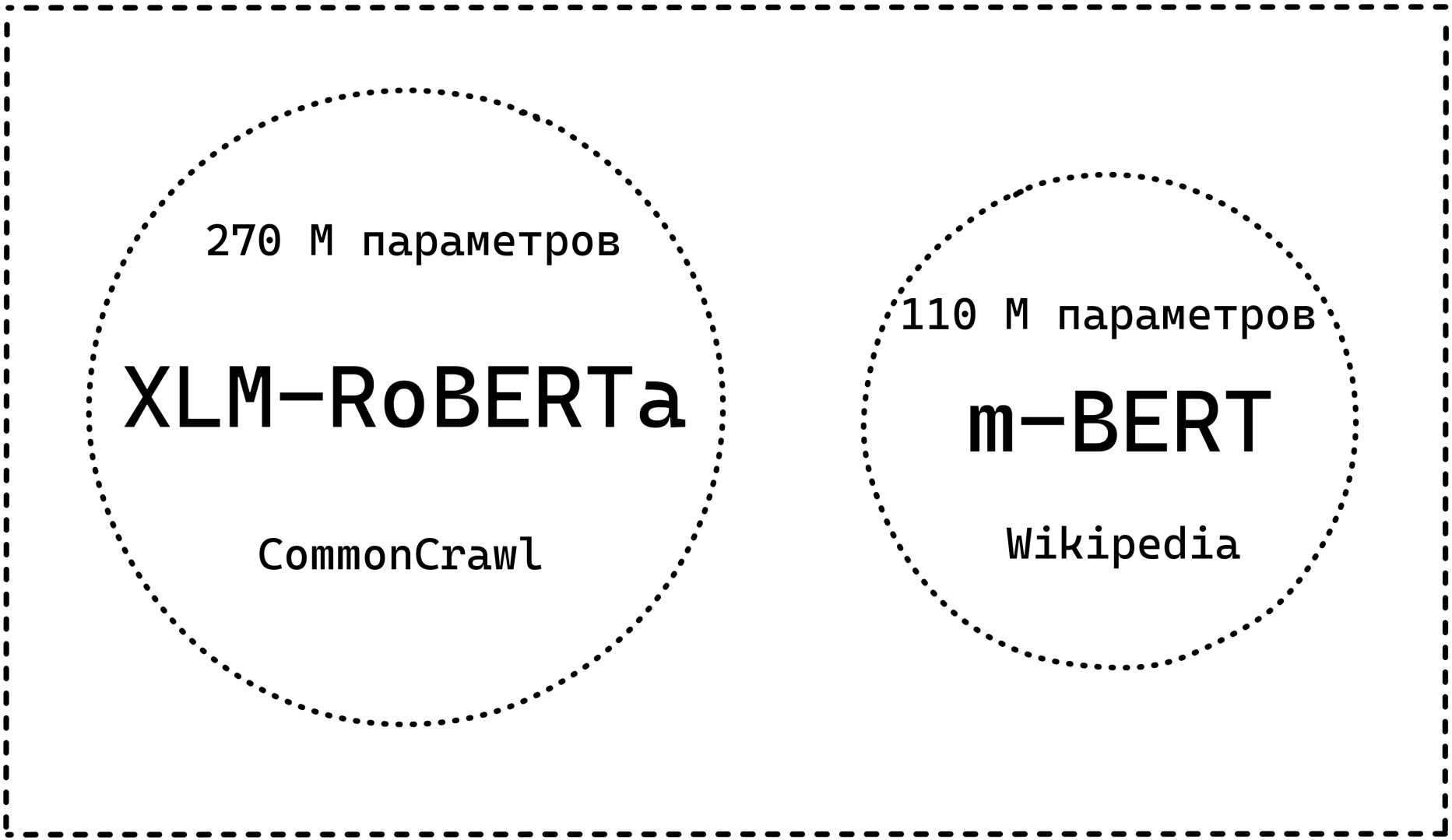


Набор данных

Intent	atis_flight						
Utterance en	show	me	flights	from	montreal	to	orlando
Slot labels en	O	O	O	O	B-fromloc.city_name	O	B-toloc.city_name
Utterance de	Zeige	mir	Flüge	von	Montreal	nach	Orlando
Slot labels de	O	O	O	O	B-fromloc.city_name	O	B-toloc.city_name

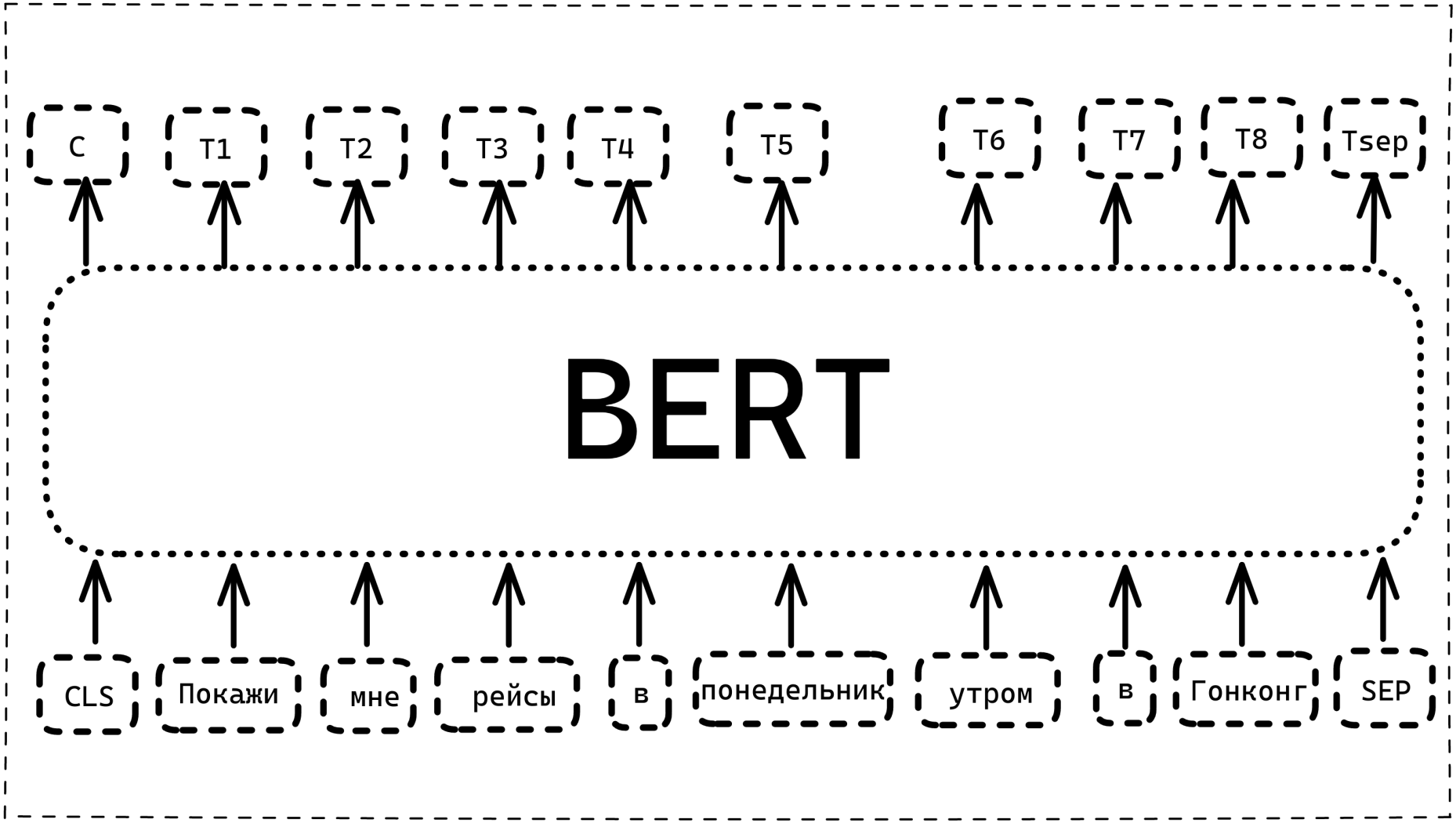
Пример объекта из набора данных MultiAtis++[**Xu2020EndtoEndSA**]

Исследуемые модели



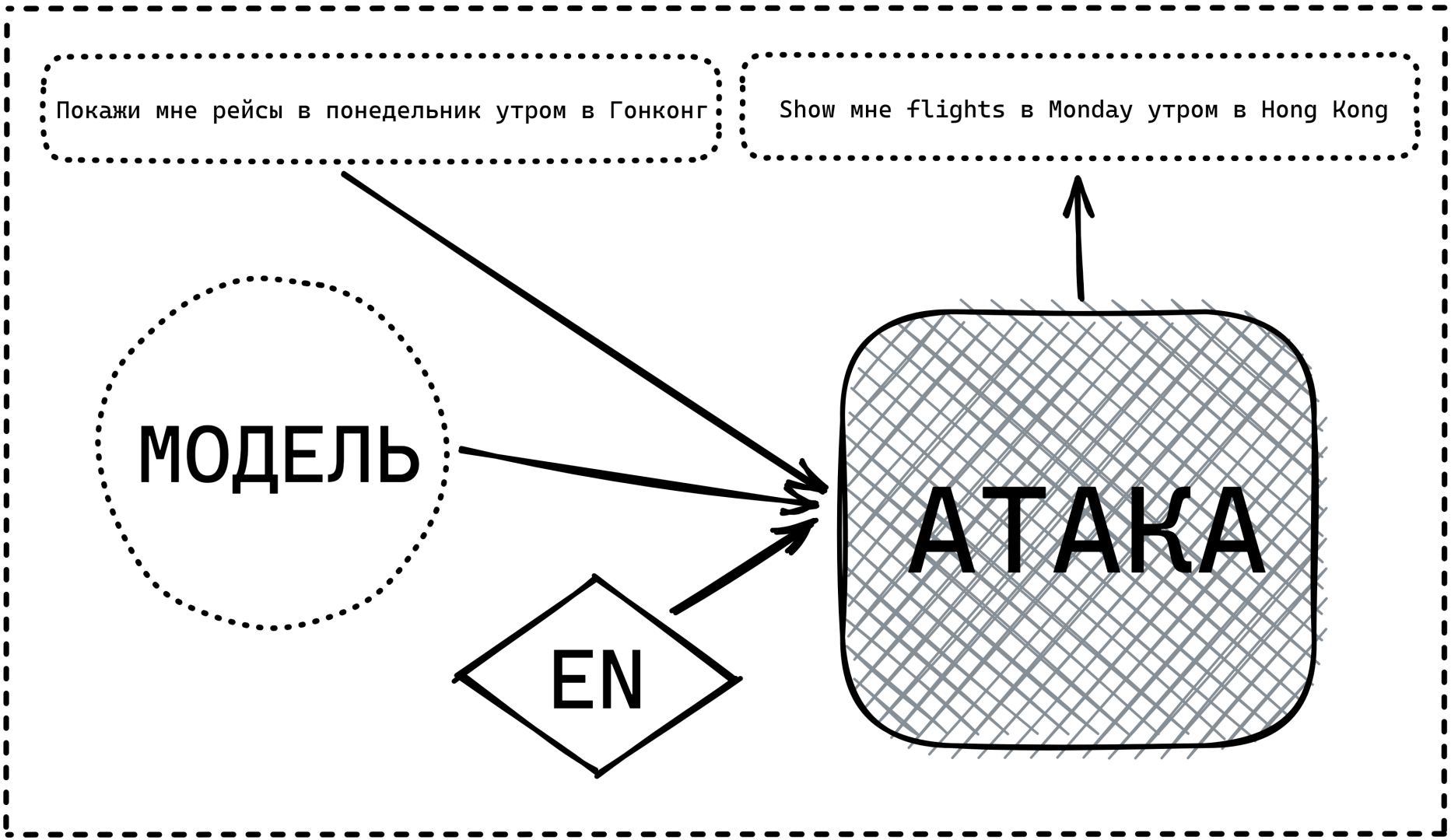
[Conneau2020UnsupervisedCR, devlin-etal-2019-bert]

Исследуемые модели

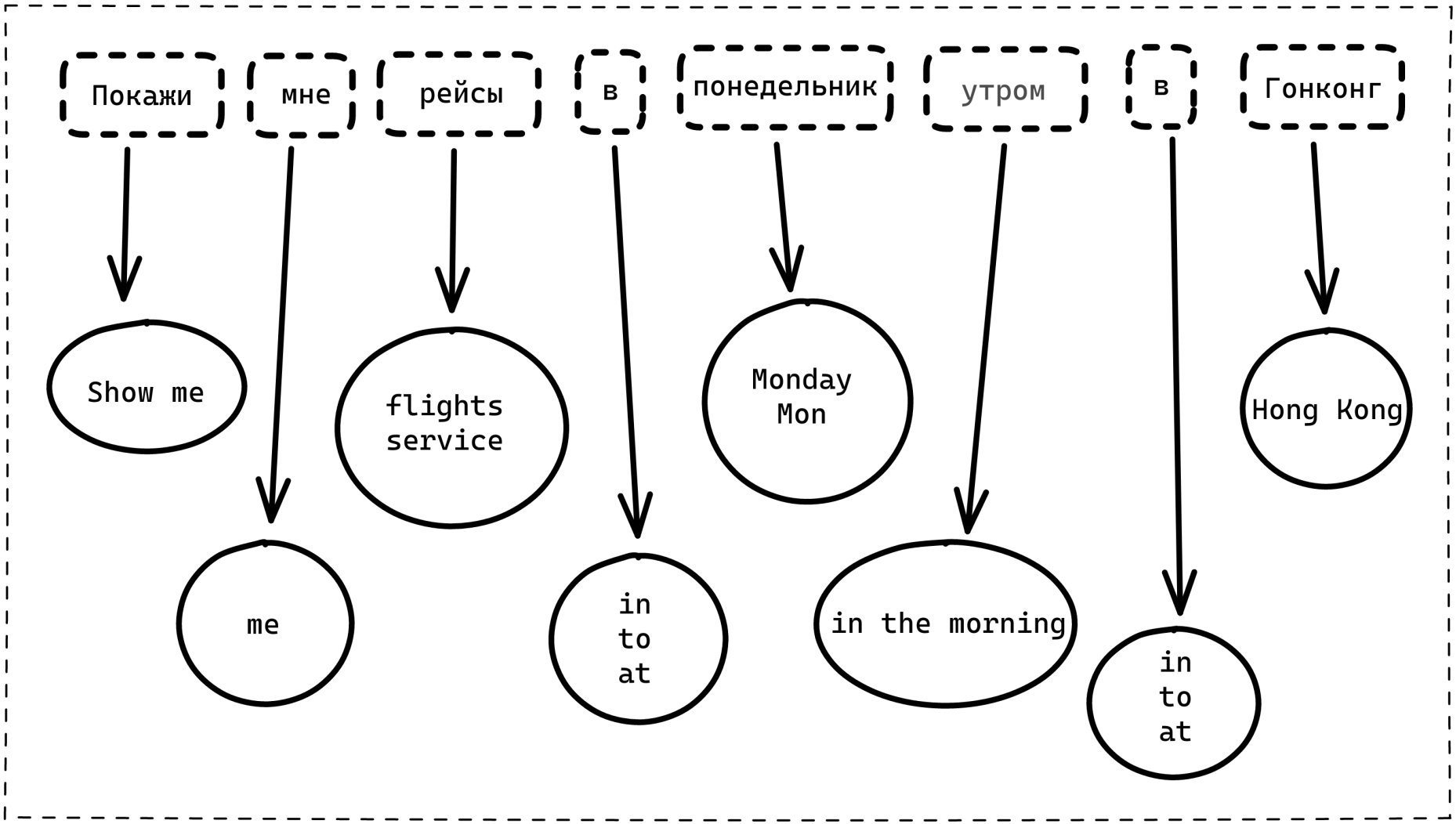


[Conneau2020UnsupervisedCR, devlin-etal-2019-bert]

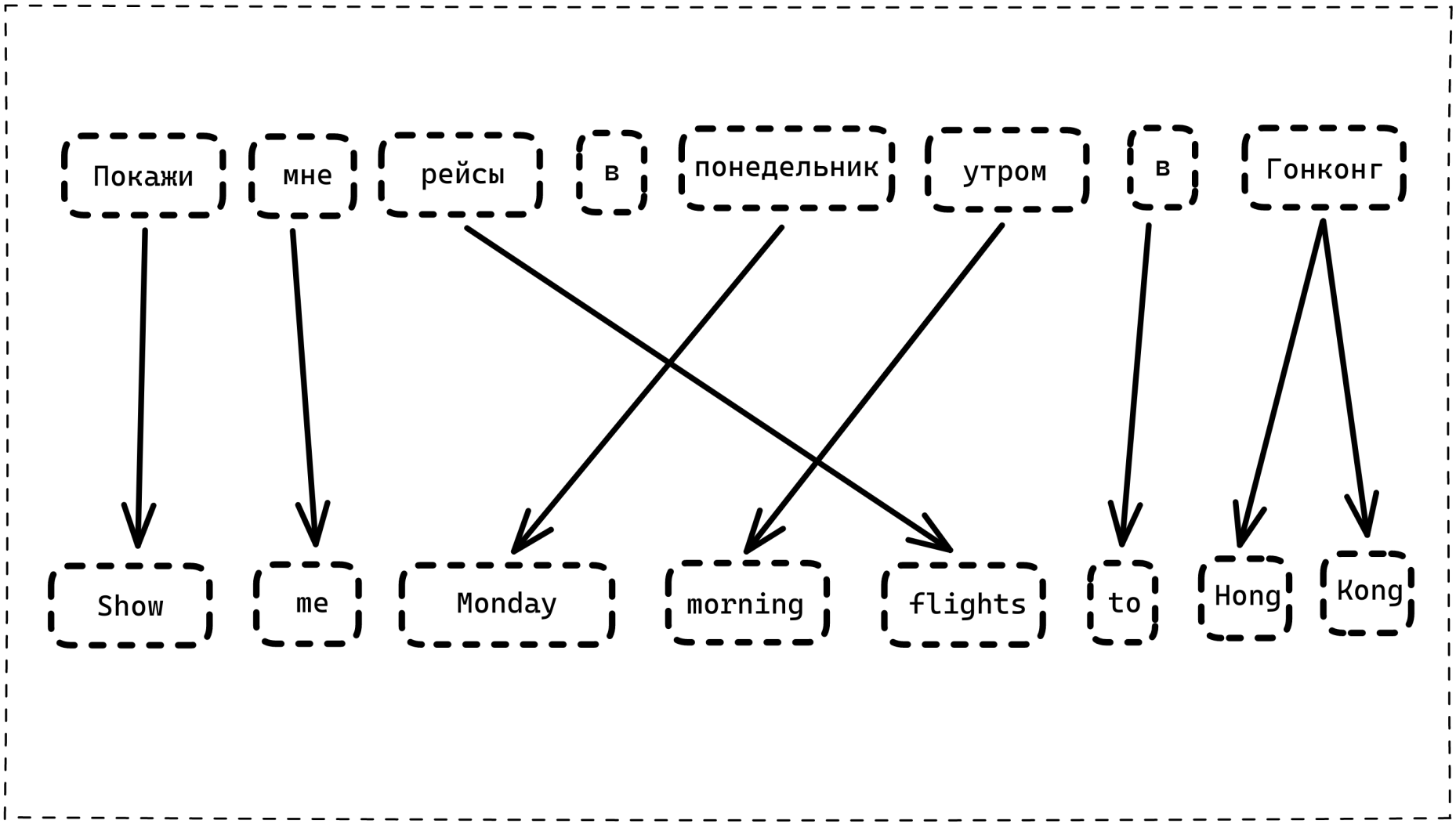
Предлагаемые атаки



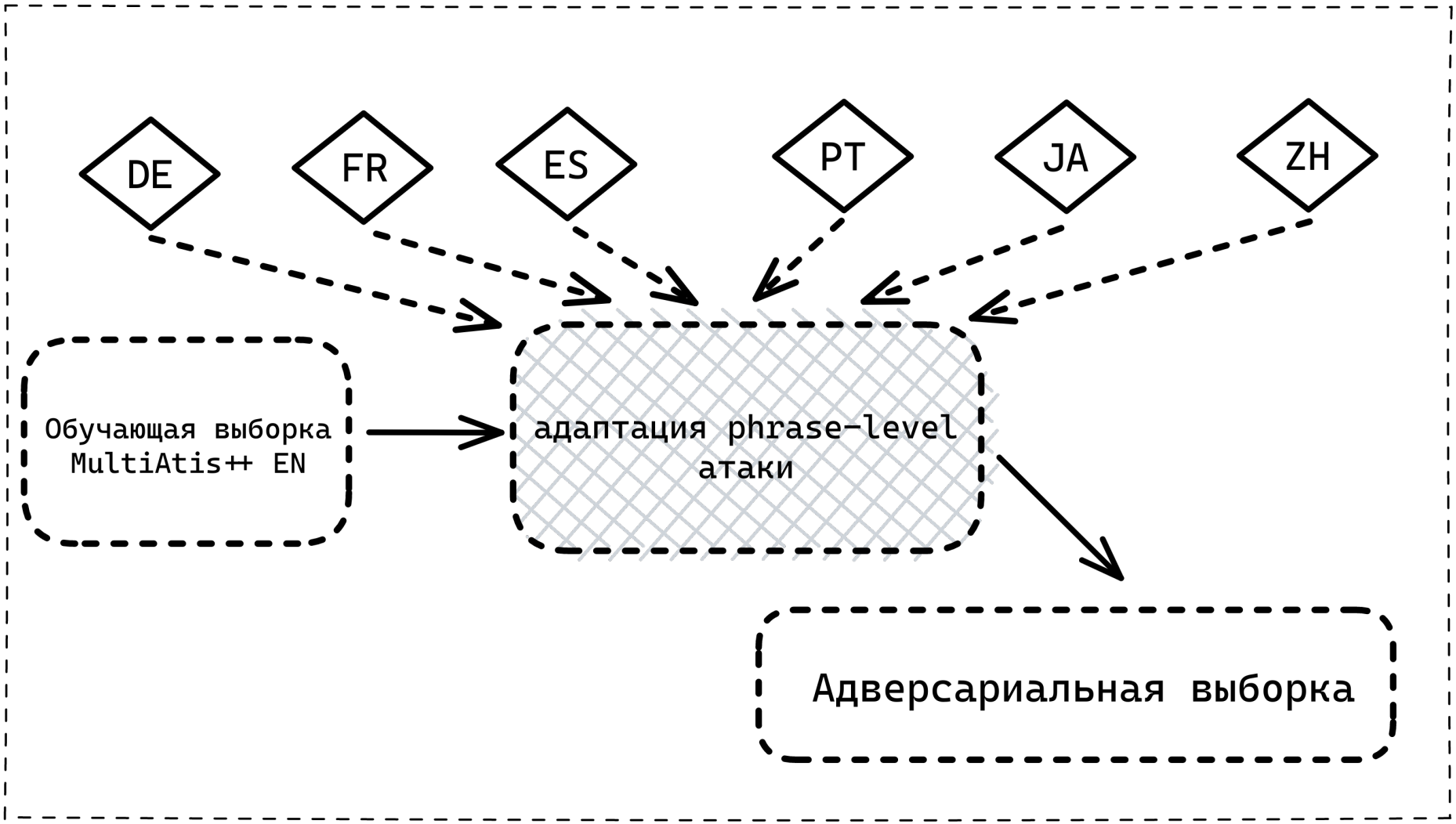
Word-level атака



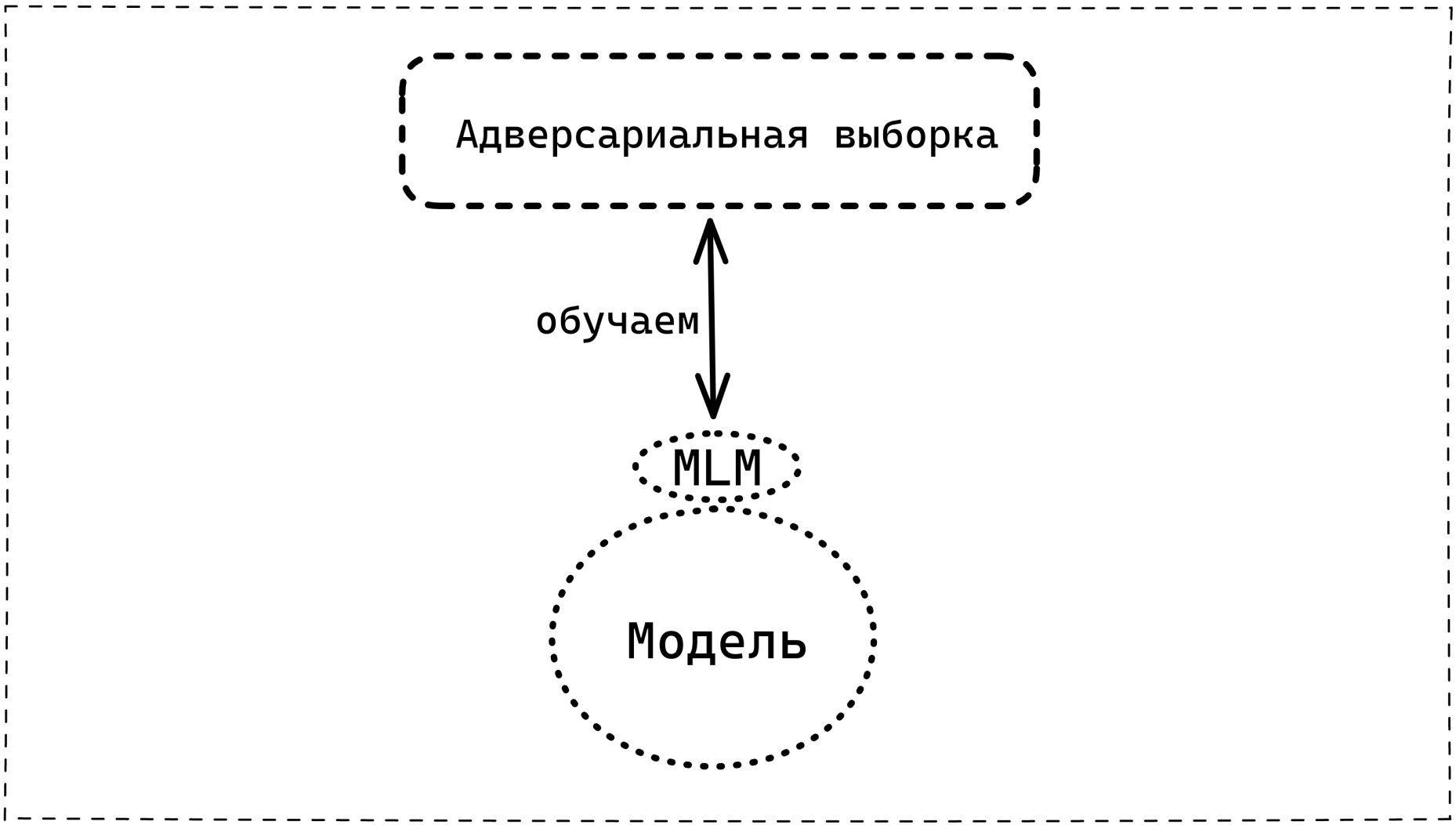
Phrase-level атака



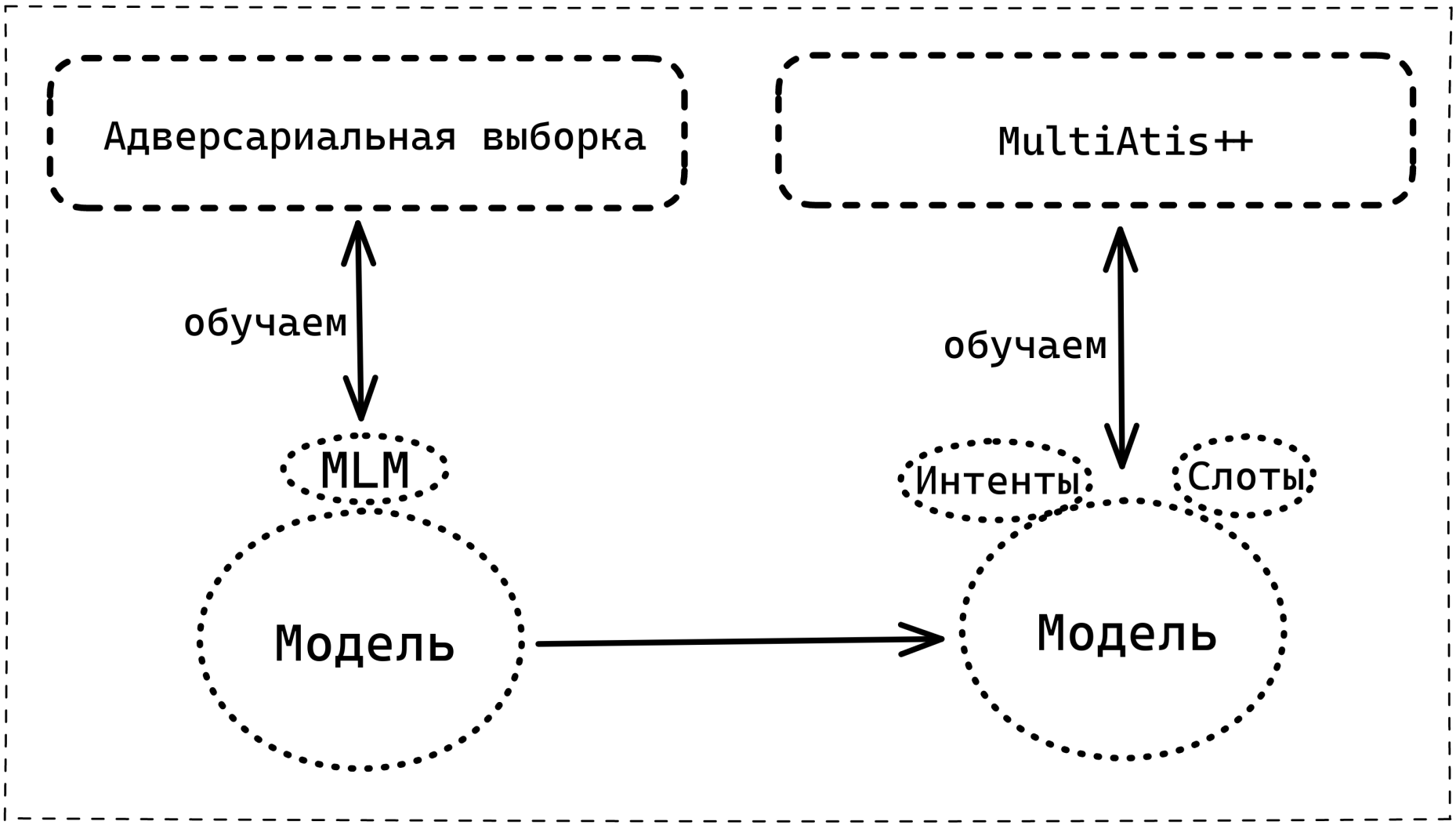
Метод адверсариального предобучения



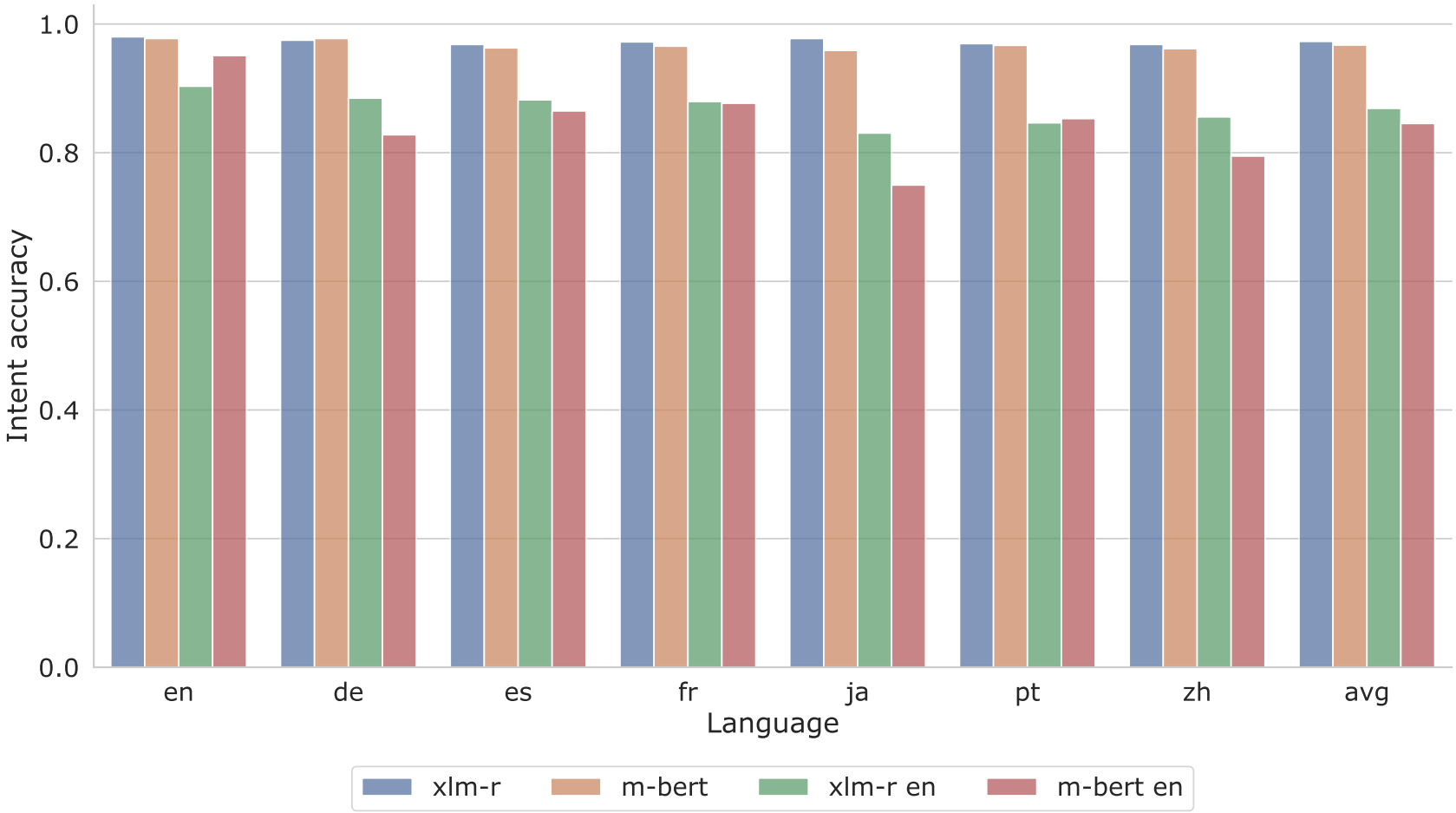
Метод адверсариального предобучения



Метод адверсариального предобучения

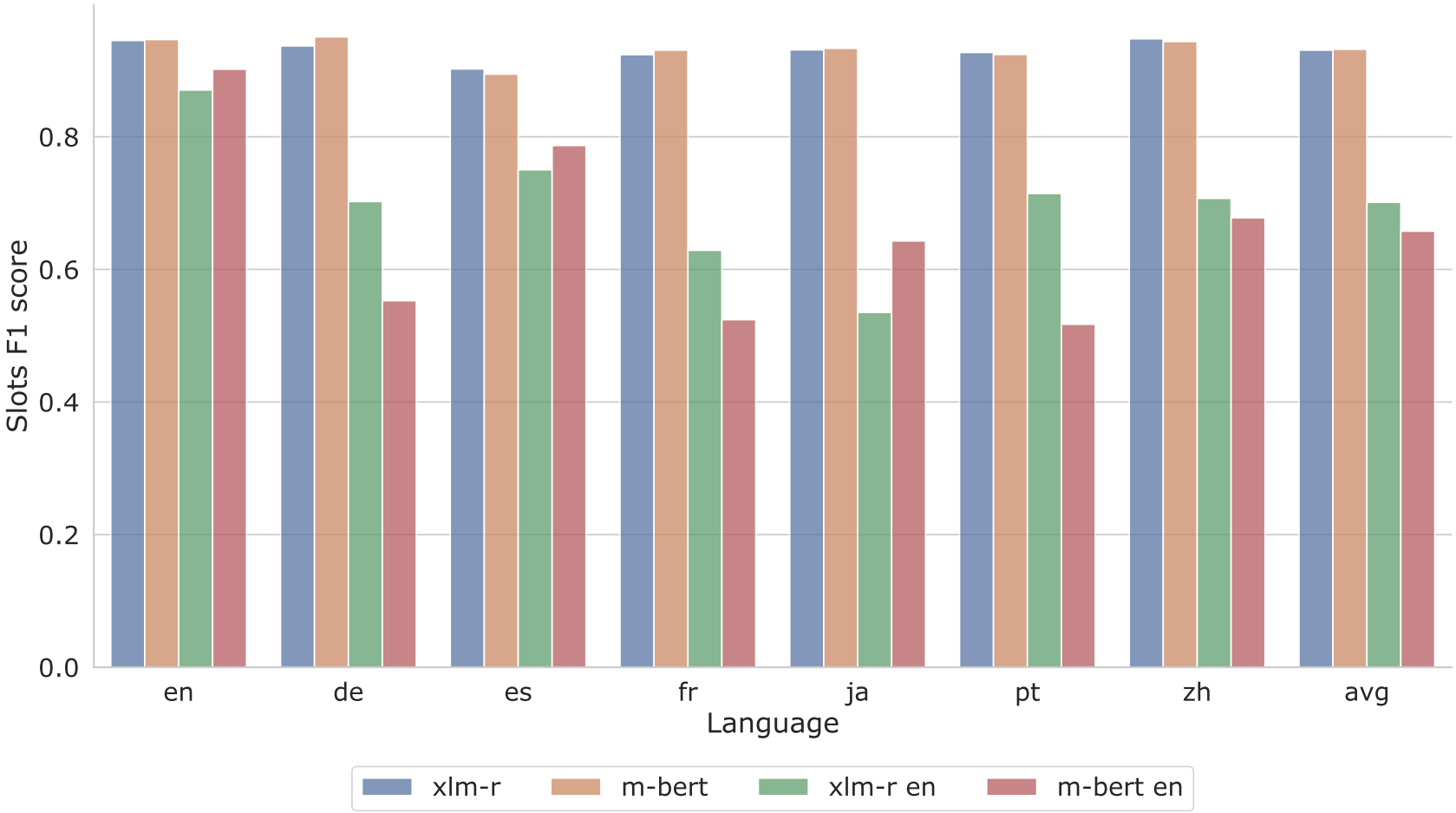


Тестовая выборка



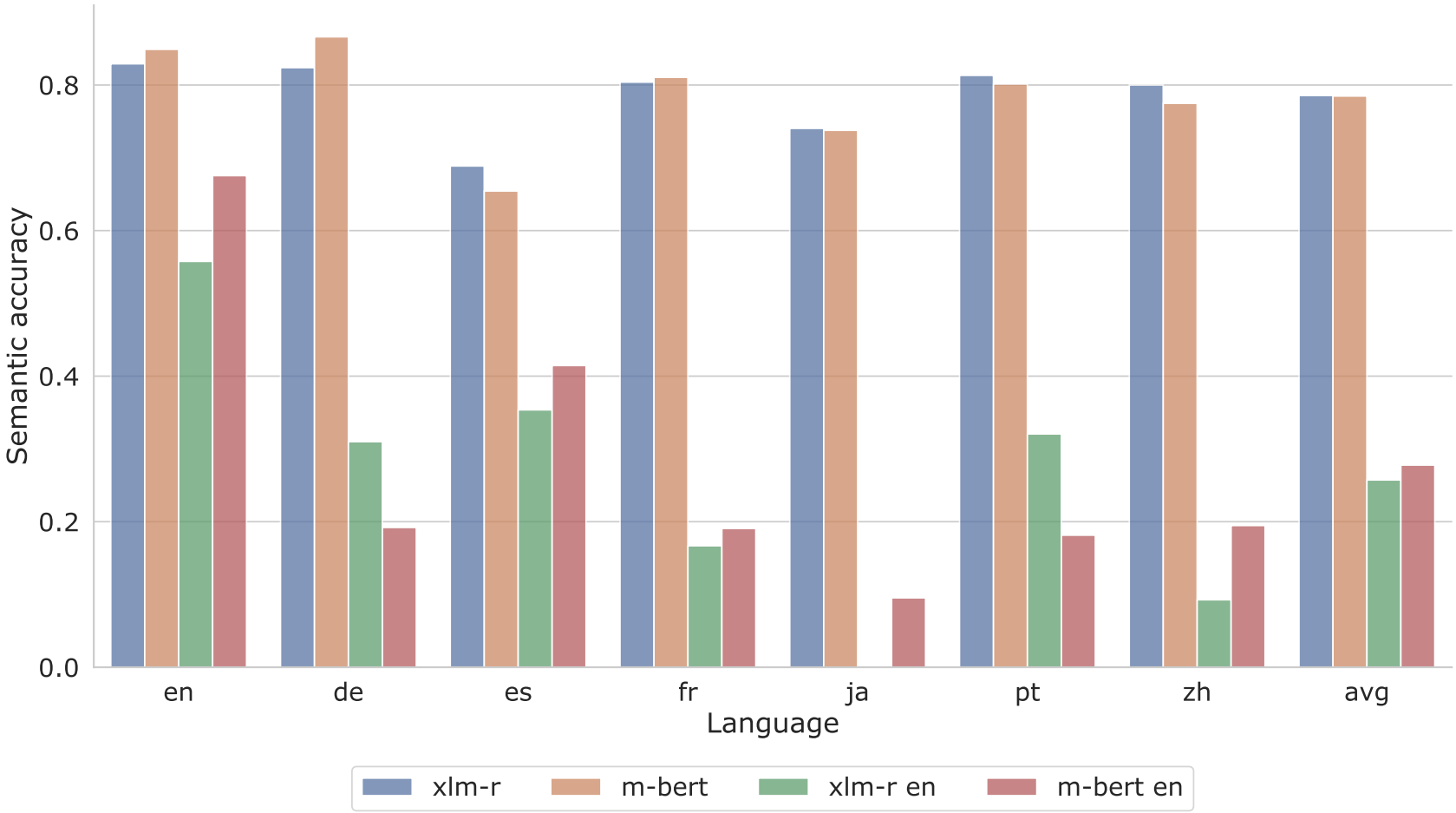
Доля предложений с верно классифицированным интентом

Тестовая выборка



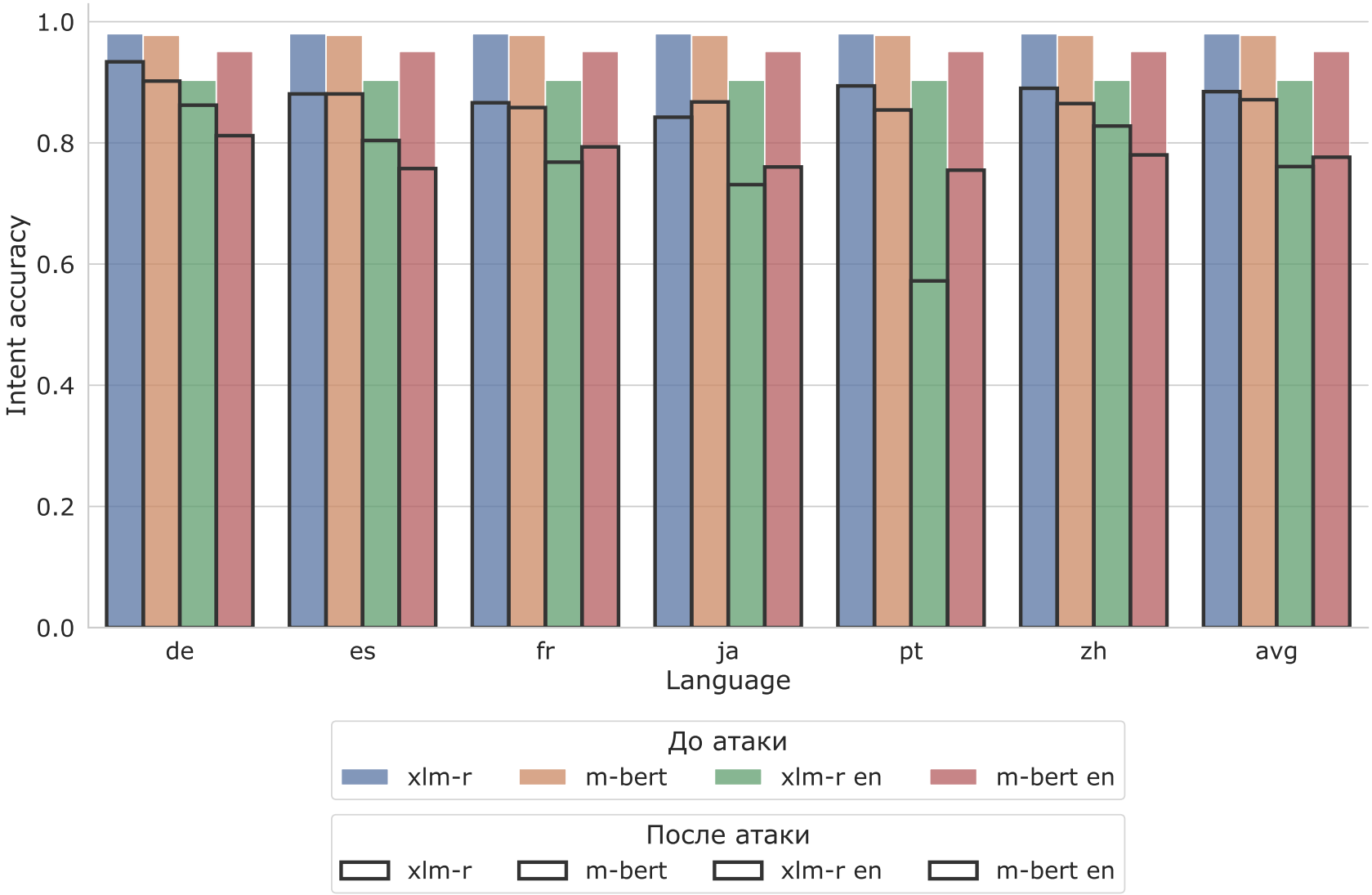
F1 мера по слотам

Тестовая выборка



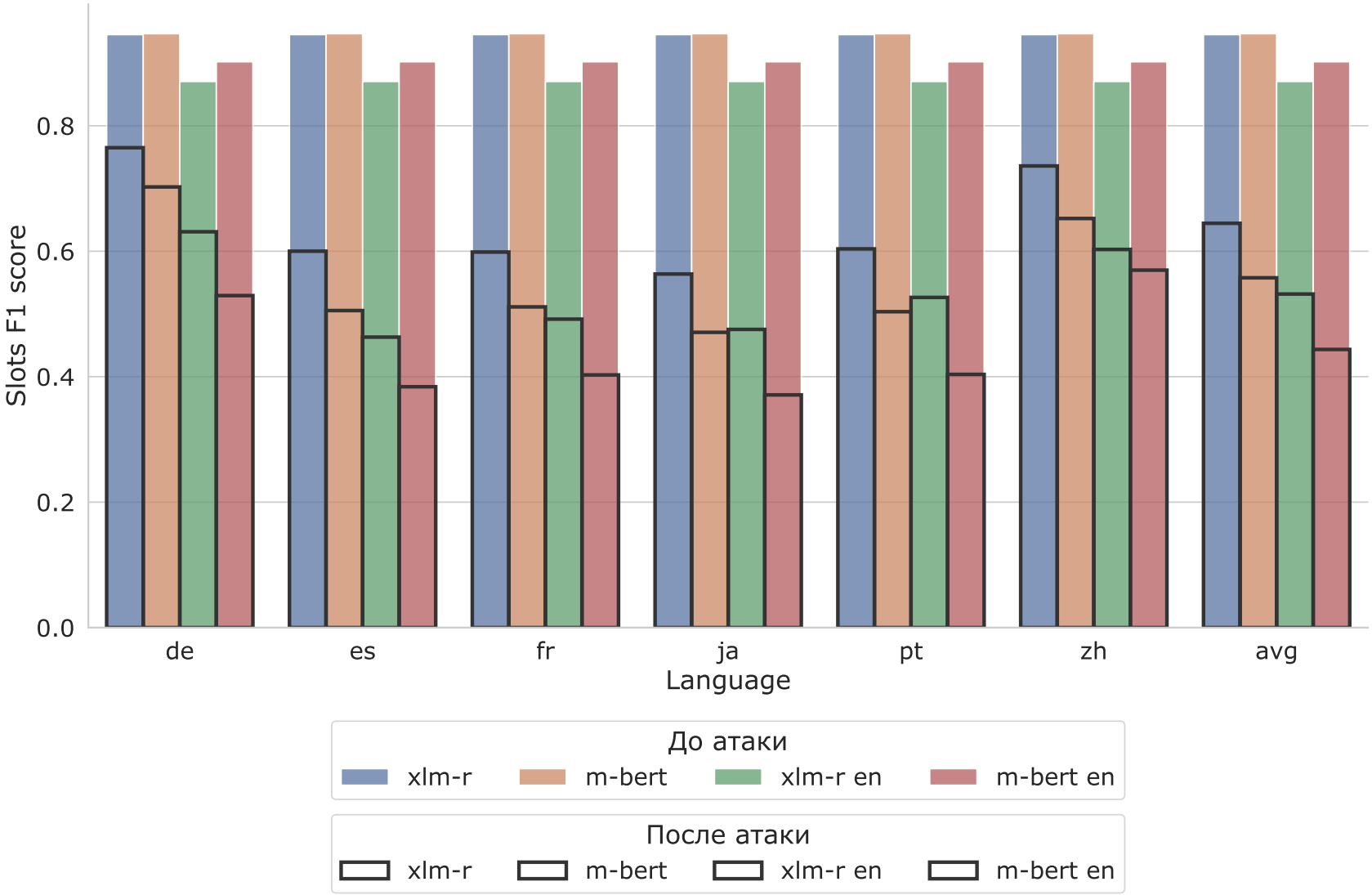
Доля полностью верно классифицированных предложений

Word-level атака



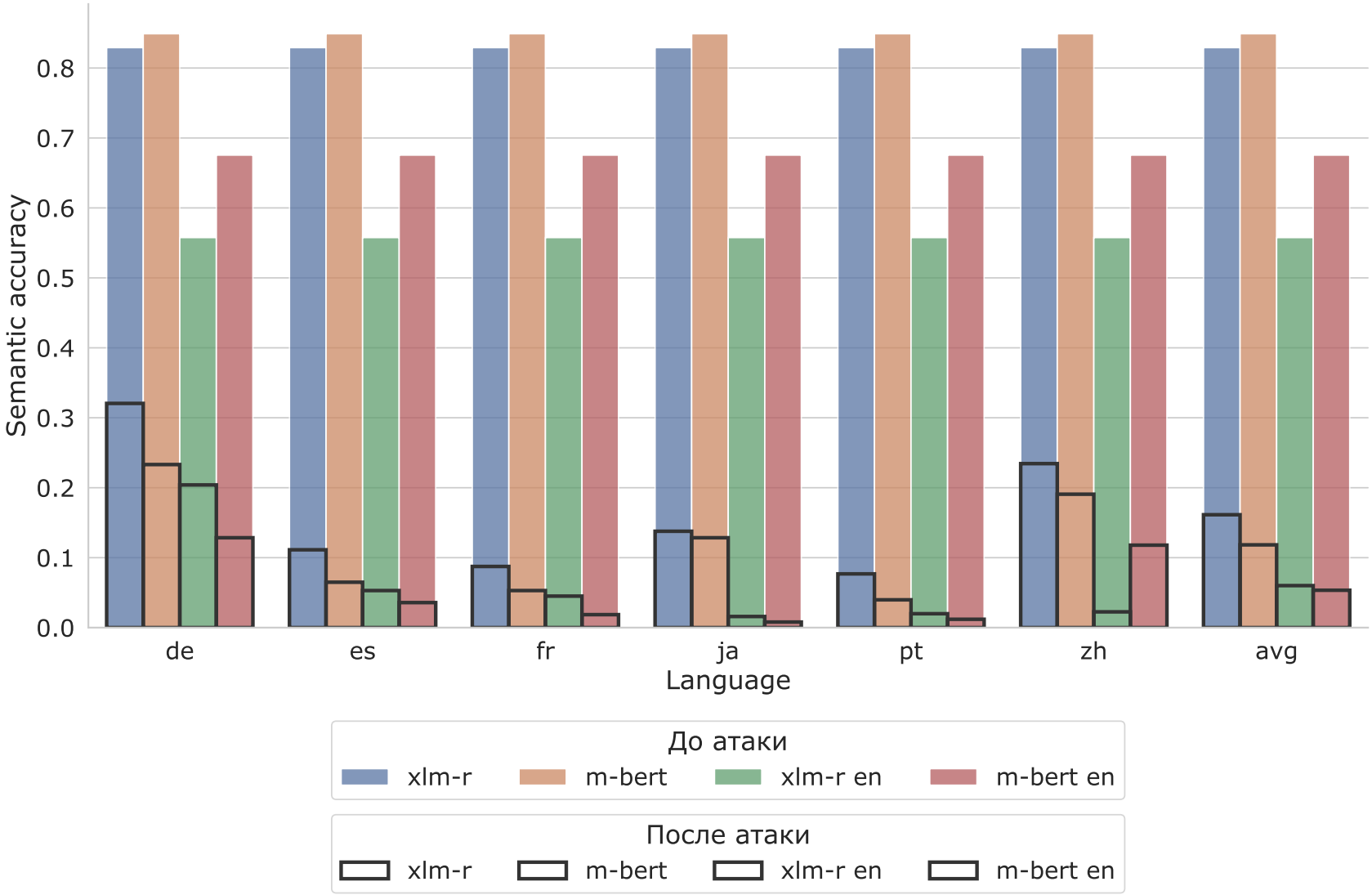
Доля предложений с верно классифицированным интентом

Word-level атака



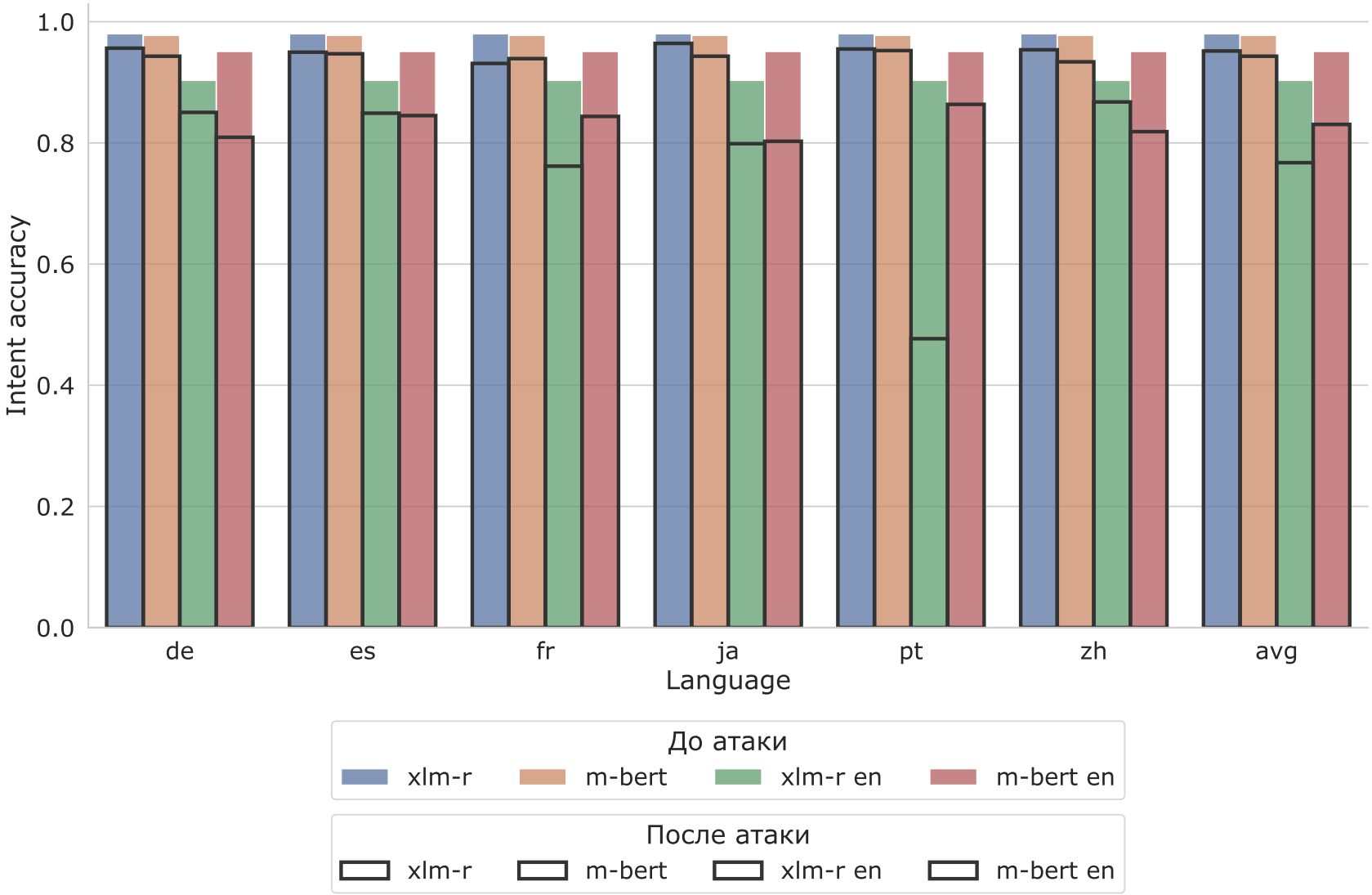
F1 мера по слотам

Word-level атака



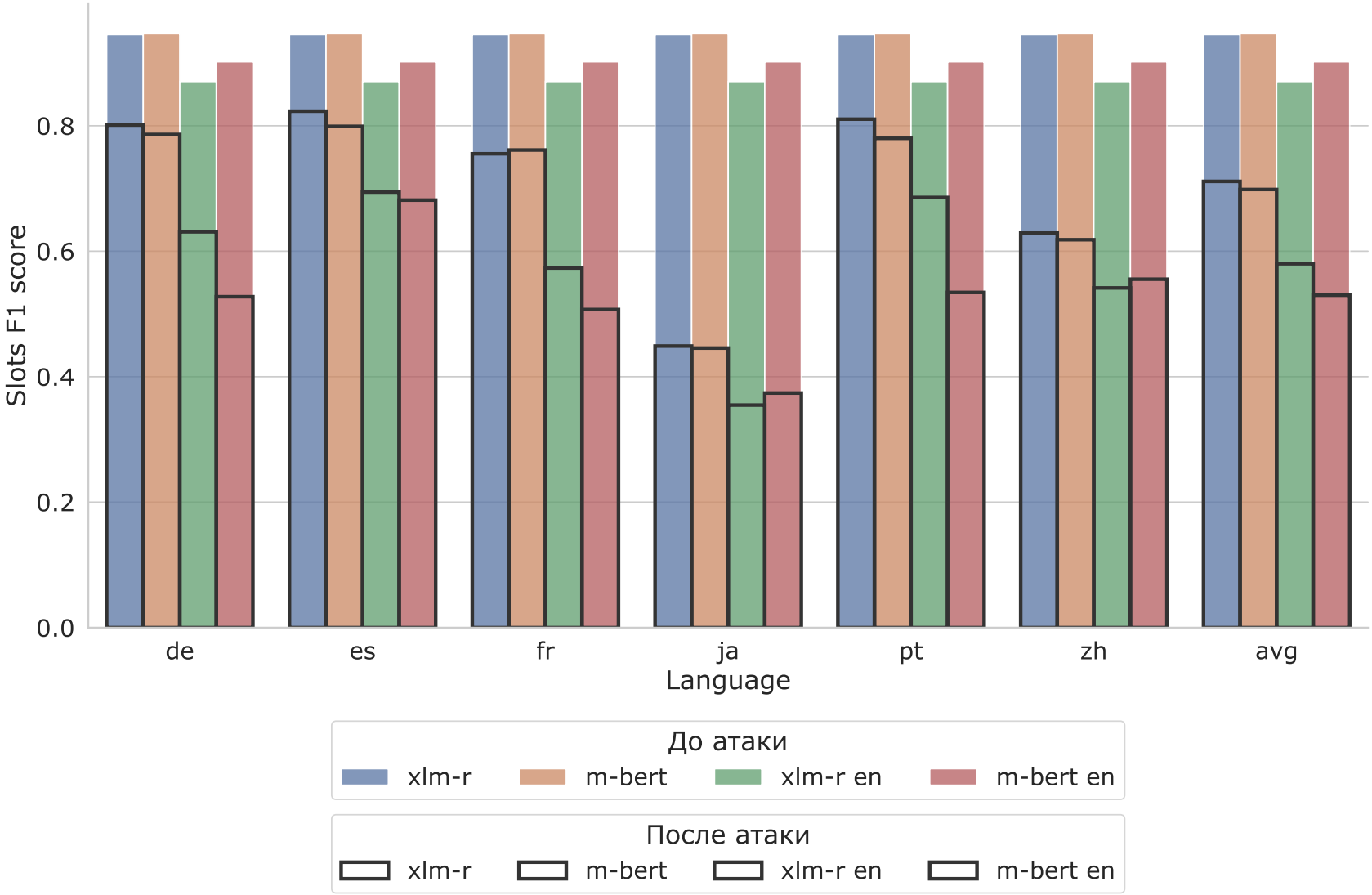
Доля полностью верно классифицированных предложений

Phrase-level атака



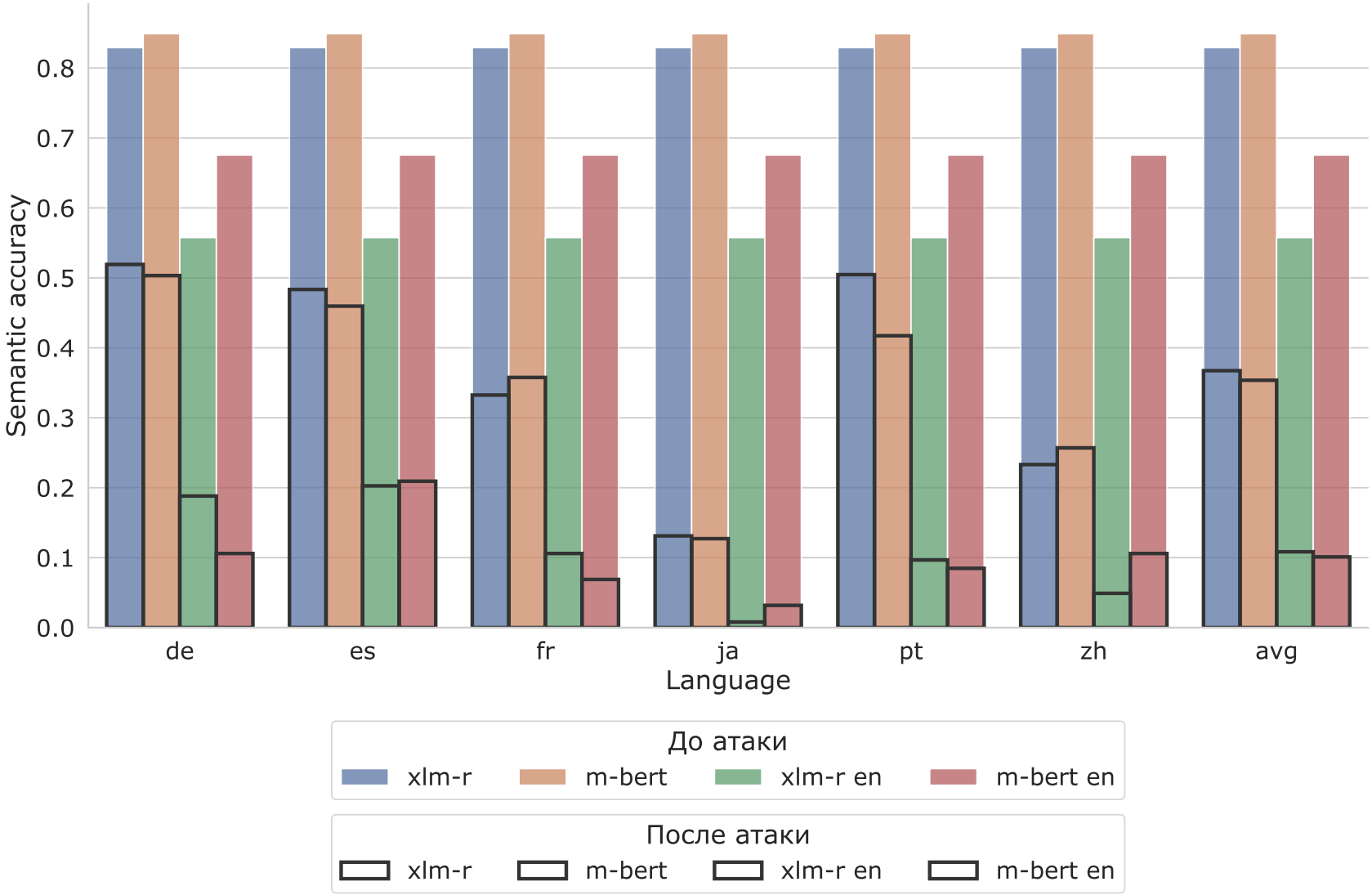
Доля предложений с верно классифицированным интентом

Phrase-level атака



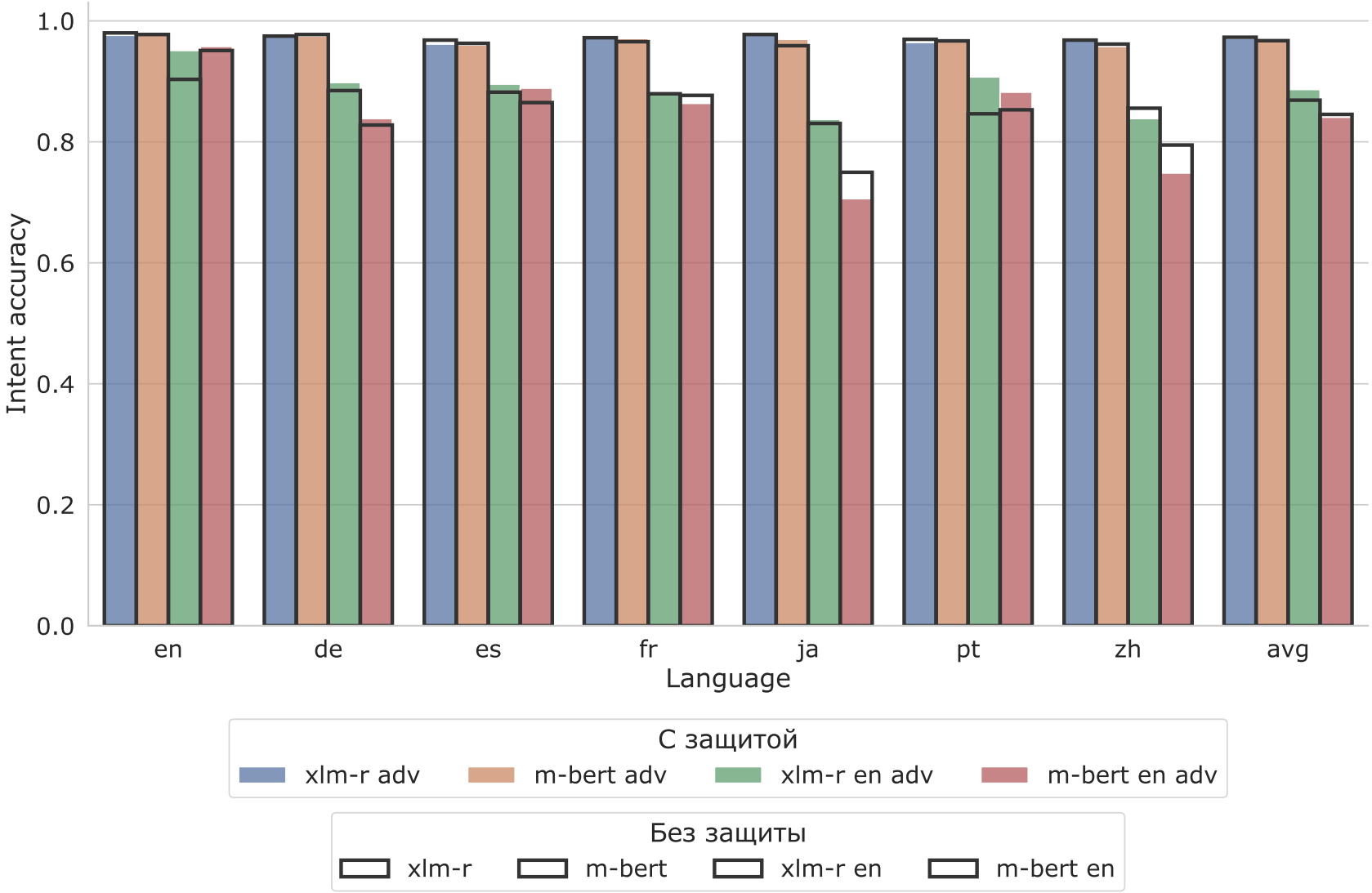
F1 мера по слотам

Phrase-level атака



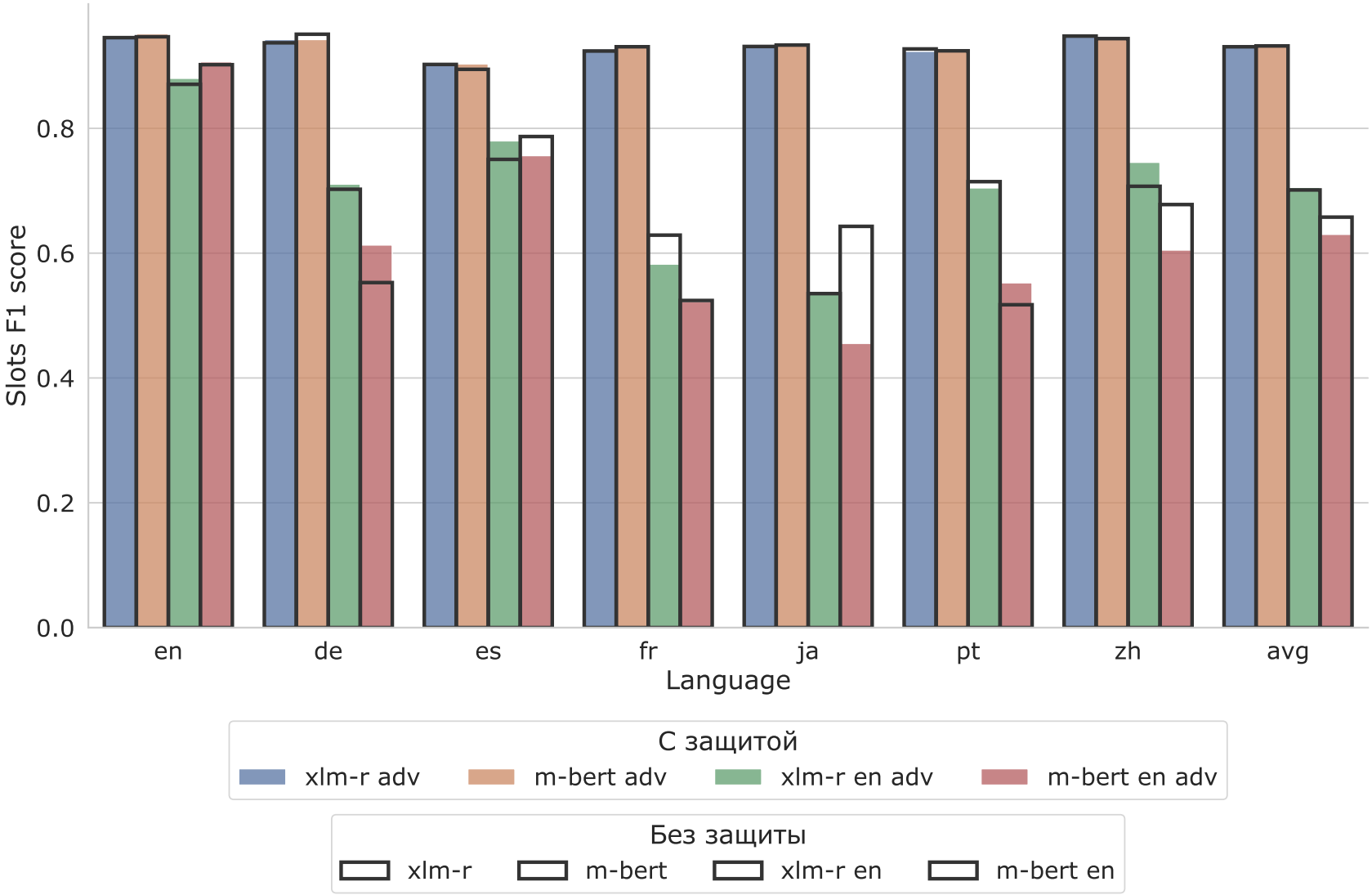
Доля полностью верно классифицированных предложений

Тестовая выборка (с защитой)



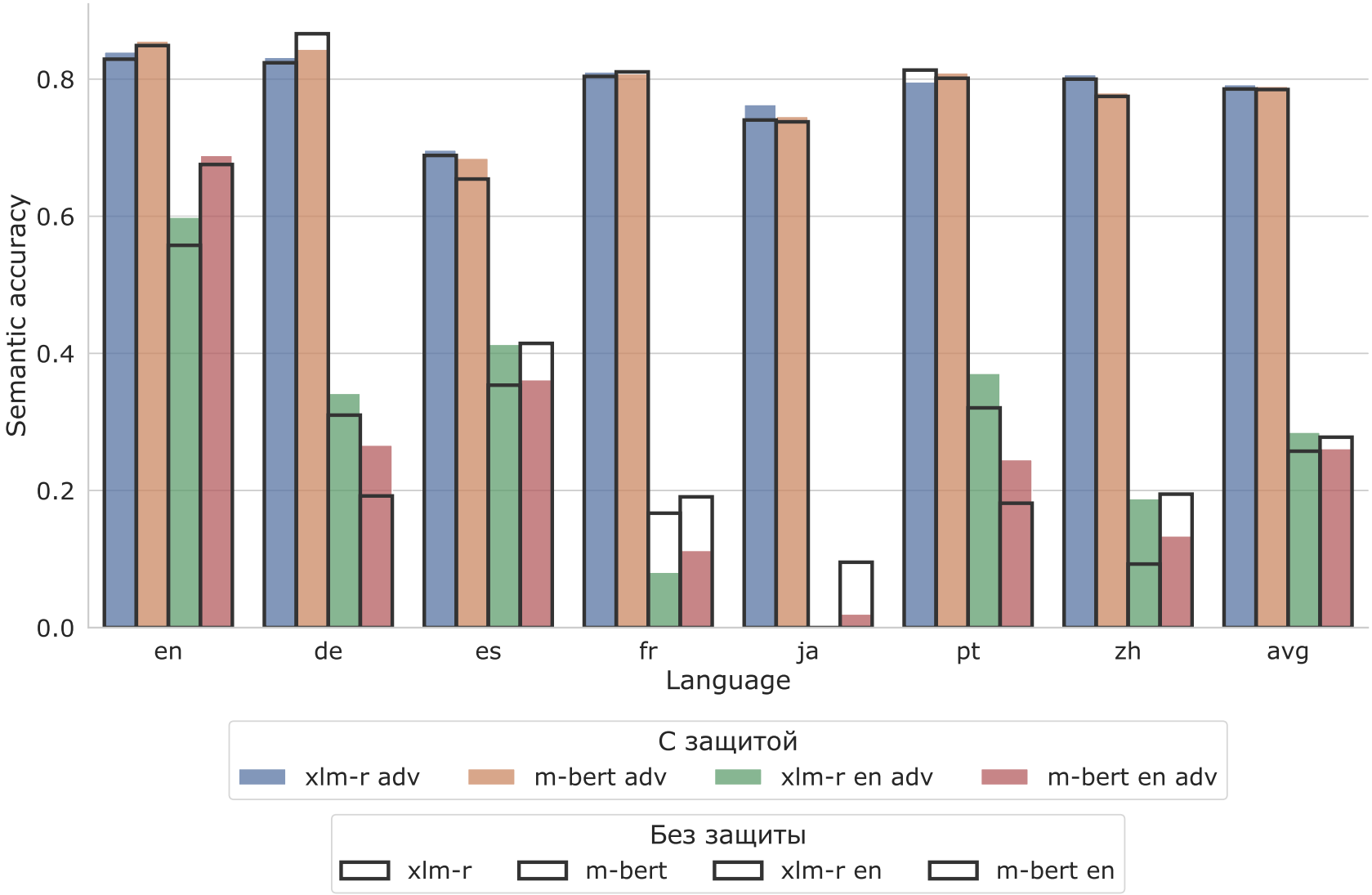
Доля предложений с верно классифицированным интентом

Тестовая выборка (с защитой)



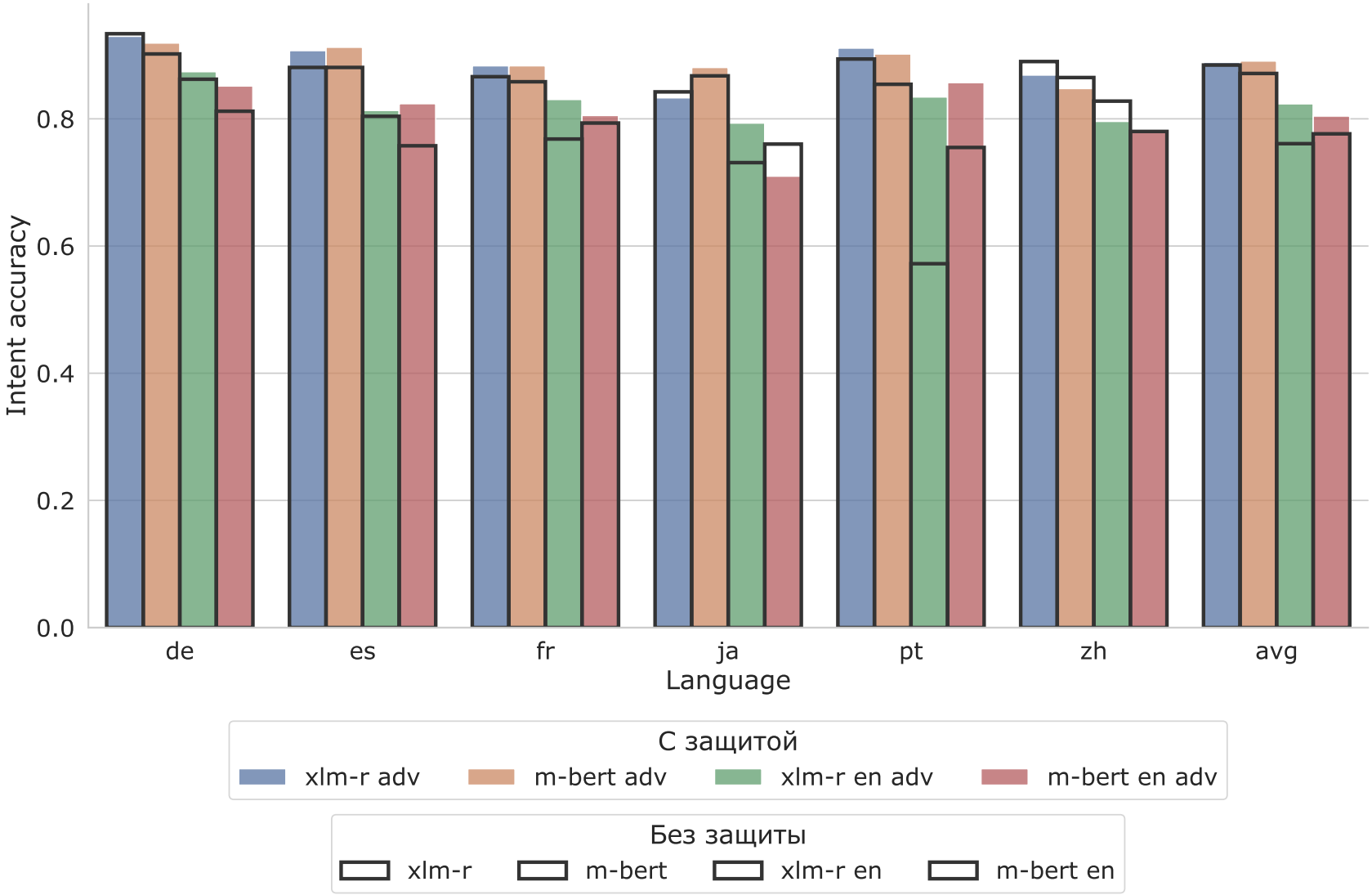
F1 мера по слотам

Тестовая выборка (с защитой)



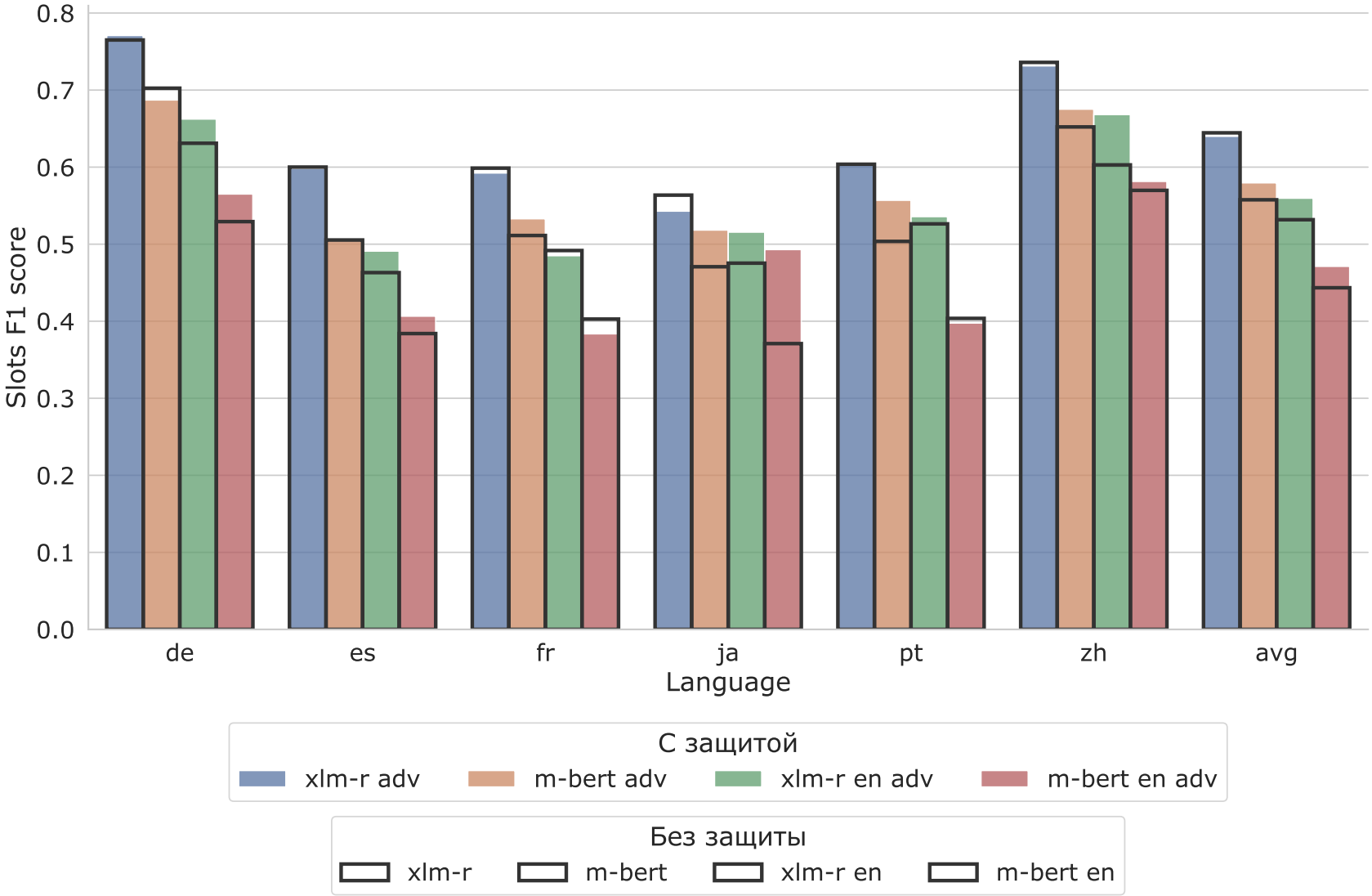
Доля полностью верно классифицированных предложений

Word-level атака (с защитой)



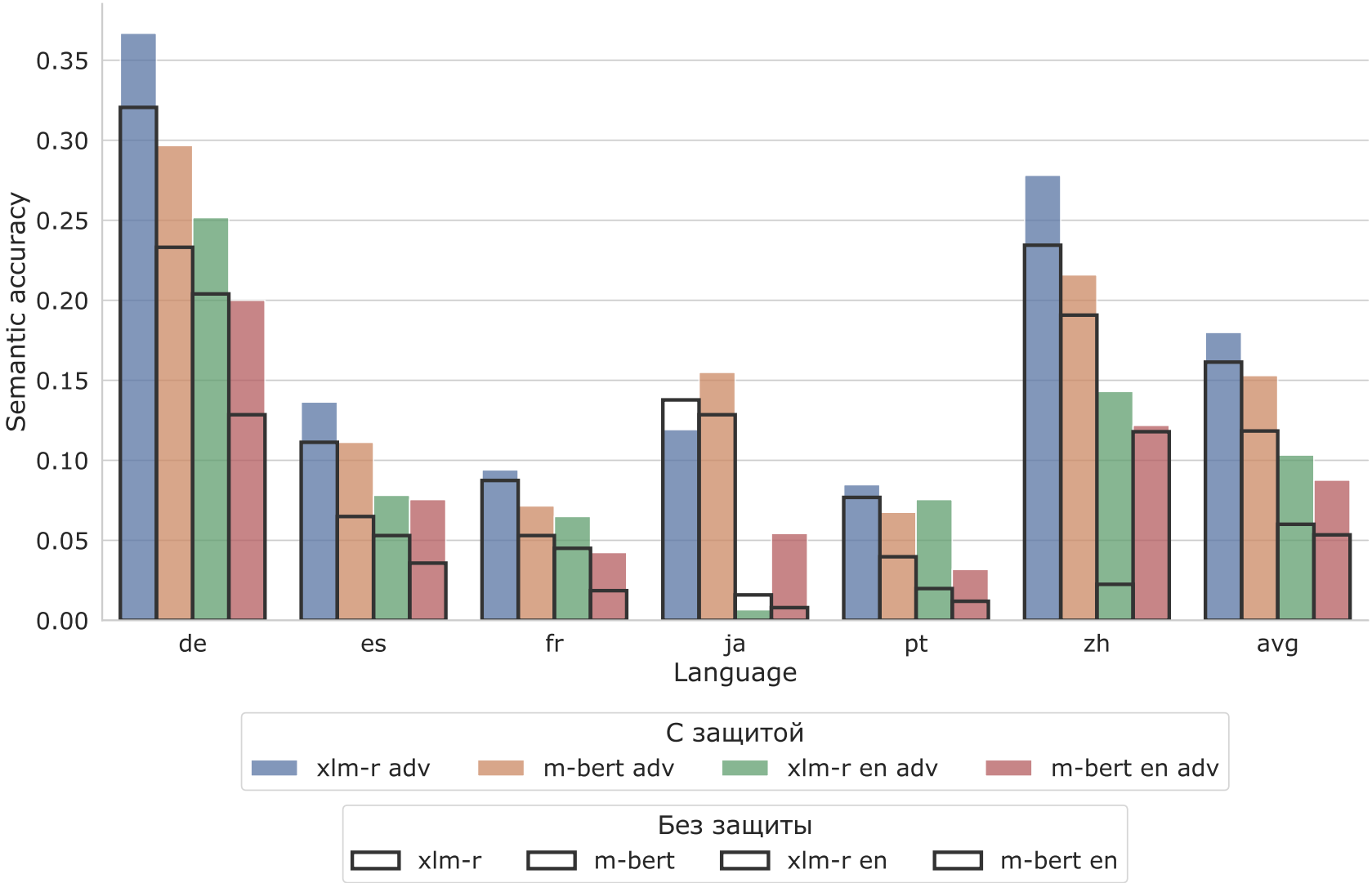
Доля предложений с верно классифицированным интентом

Word-level атака (с защитой)



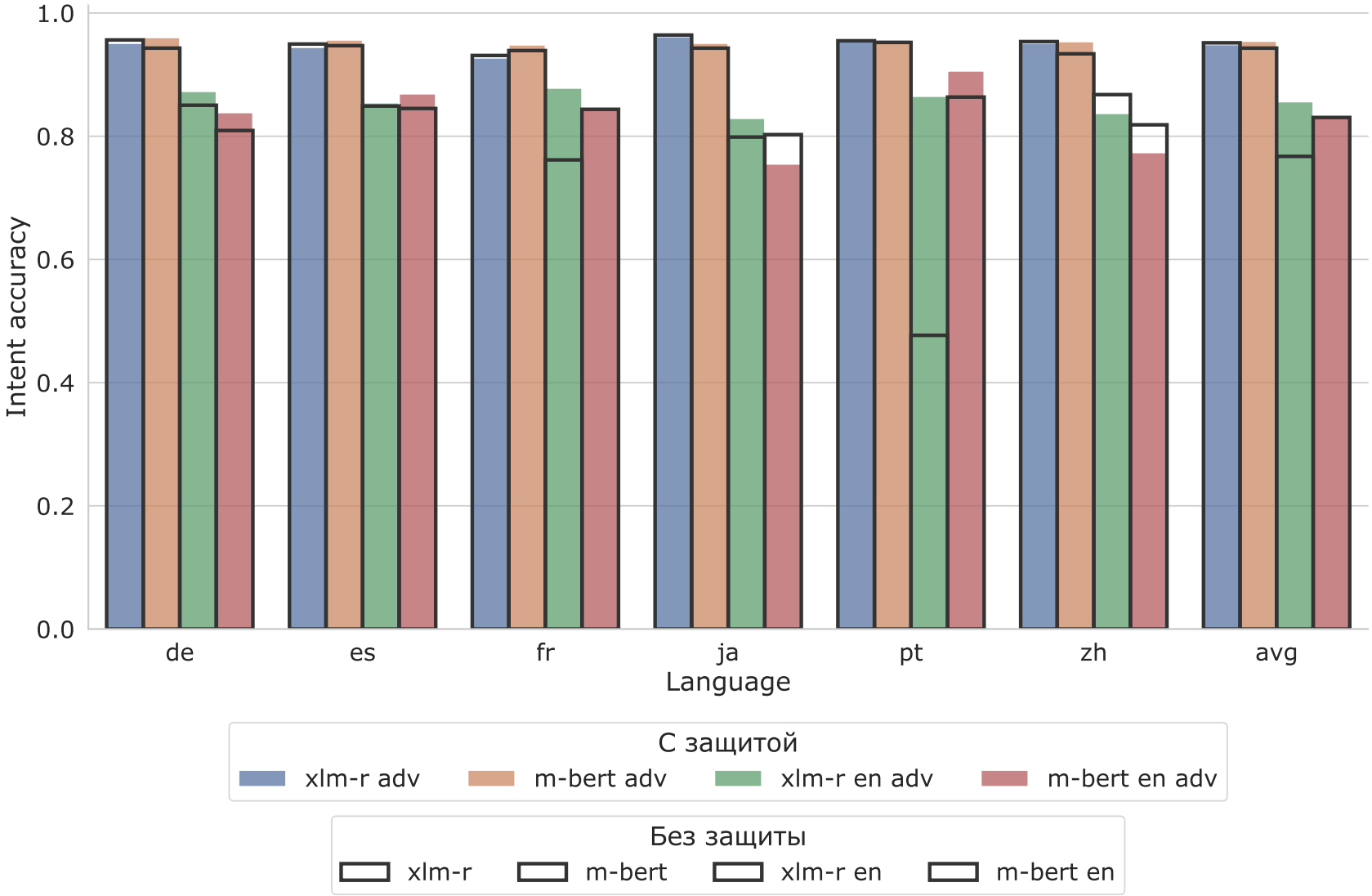
F1 мера по слотам

Word-level атака (с защитой)



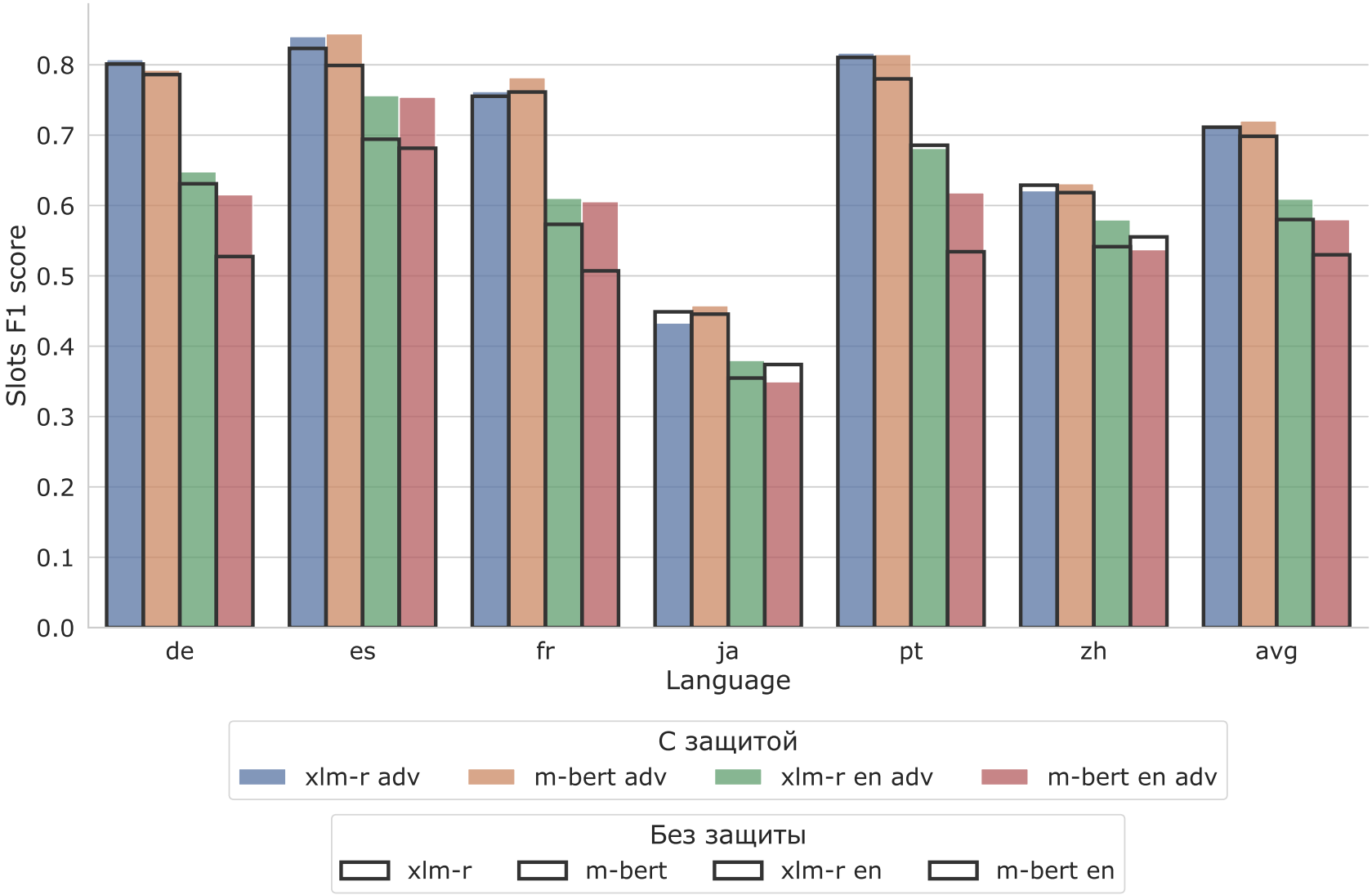
Доля полностью верно классифицированных предложений

Phrase-level атака (с защитой)



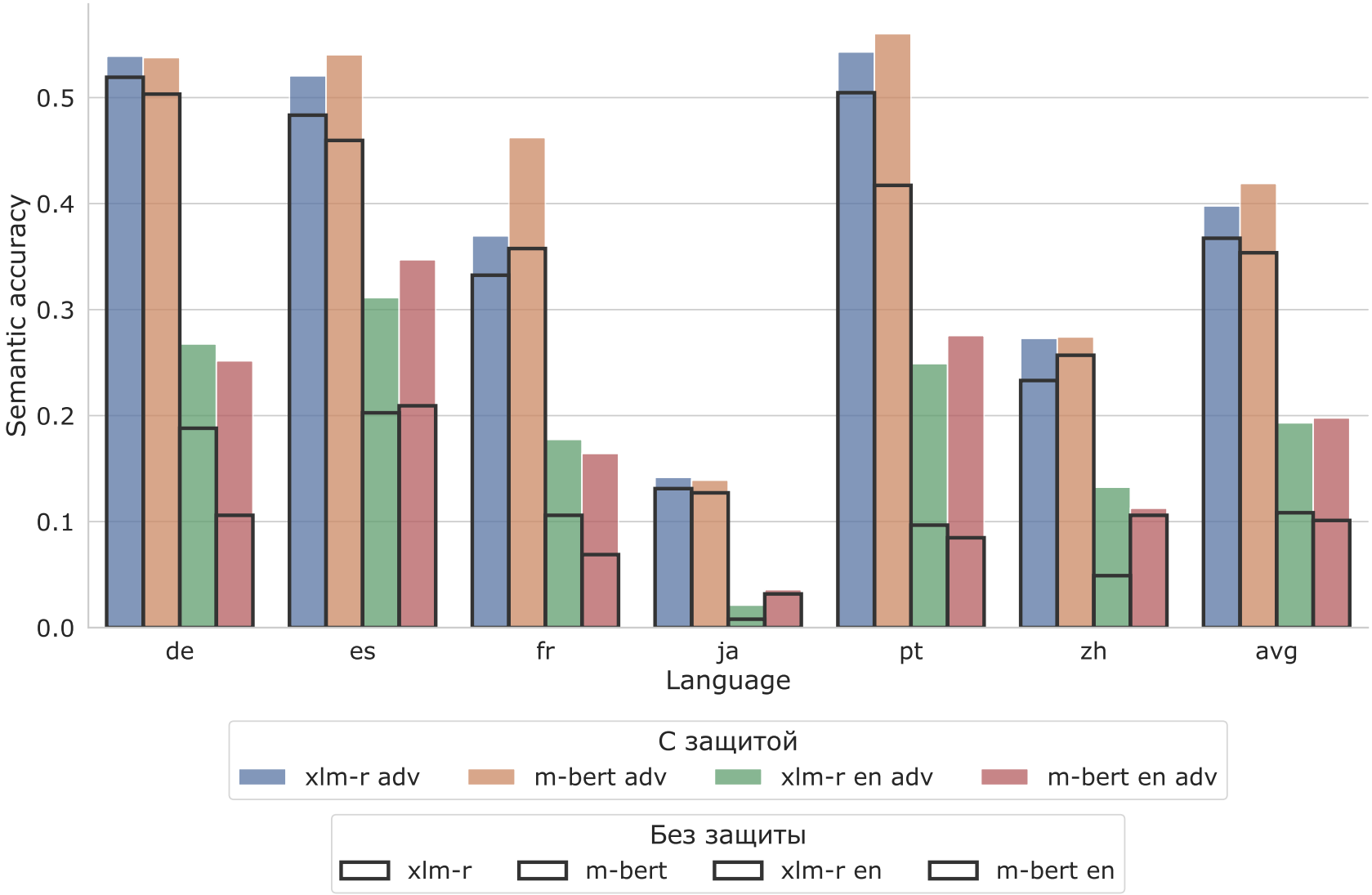
Доля предложений с верно классифицированным интентом

Phrase-level атака (с защитой)



F1 мера по слотам

Phrase-level атака (с защитой)



Доля полностью верно классифицированных предложений

Заключение

- Решили задачу классификации интентов и заполнения слотов

Заключение

- Решили задачу классификации интентов и заполнения слотов
- Провели анализ качества моделей после двух предложенных атак

Заключение

- Решили задачу классификации интентов и заполнения слотов
- Провели анализ качества моделей после двух предложенных атак
- Провели анализ качества моделей после предложенного метода защиты

Спасибо за внимание!