Федеральное государственное автономное образовательное учреждение
высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»
Физтех-школа Прикладной Математики и Информатики
Научно-образовательный центр когнитивного моделирования

Направление подготовки / специальность: 09.04.01 Информатика и вычислительная техника
Направленность (профиль) подготовки: Методы и технологии искусственного интеллекта

РАЗРАБОТКА МУЛЬТИМОДАЛЬНЫХ НЕЙРОСЕТЕВЫХ МЕТОДОВ РАСПОЗНАВАНИЯ ИЗОБРАЖЕНИЙ НА ОСНОВЕ ТРАНСФОРМЕРНЫХ АРХИТЕКТУР

(магистерская диссертация)

Студент:
Авшалумов Михаил Сергеевич

_____
(подпись студента)

Научный руководитель:
Юдин Дмитрий Александрович,
канд. техн. наук

_____
(подпись научного руководителя)

Консультант *(при наличии)*:
_____
_____
(подпись консультанта)

Москва 2025

Содержание

1 Введение **5**

2 Обзор современных подходов **8**

2.1 Модели распознавания изображений с открытым словарём 8

2.1.1 Модели детекции 8

2.1.2 Модели сегментации 11

2.2 Алгоритмы подбора текстовых промтов для моделей распознавания изображений с открытым словарем 12

2.3 Алгоритмы дообучения языковых моделей 14

2.3.1 Подходы, основанные на обучении с подкреплением 15

2.3.2 Альтернативные подходы 20

2.3.3 Сравнение подходов к дообучению языковых моделей 28

3 Формальная постановка задачи **31**

3.1 Формальная постановка задачи в общем виде 31

3.2 Дообучение языковой модели как марковский процесс 32

3.2.1 Состояния и переходы 32

3.2.2 Действия агента 33

3.2.3 Функция вознаграждения 33

3.3 Формальная постановка задачи в виде марковского процесса 35

4 Разработка метода улучшения качества работы нейросетевых трансформерных моделей распознавания изображений **36**

4.1 Общая структура разрабатываемого подхода 36

4.2 Первый этап обучения: SFT 38

4.3 Второй этап обучения: онлайн-режим 39

4.4 Второй этап обучения: оффлайн-режим 41

4.5 Формирование датасета 42

5 Вычислительные эксперименты **46**

5.1 Экспериментальная конфигурация 46

5.2 Технические детали реализации 47

5.3 Результаты экспериментов над предлагаемым подходом в задаче детекции 47

5.3.1 Количественный анализ результатов 47
5.3.2 Переформулирование текстовых запросов без обучения. 48
5.3.3 Влияние качества исходного датасета на качество результатов. . 49
5.3.4 Качественный анализ работы предложенного метода 50
5.3.5 Влияние выбора функции вознаграждения на качество результата 51
5.3.6 Возможность переноса обученных языковых моделей между различными моделями распознавания 52
5.4 Результаты экспериментов над предлагаемым подходом в задаче сегментации 52
5.4.1 Количественный анализ результатов 52
5.4.2 Качественный анализ работы предложенного метода 54

6 Заключение **57**

Аннотация

В последние годы наблюдается рост интереса к системам компьютерного зрения, способным обрабатывать произвольные текстовые запросы для распознавания объектов на изображениях. В данной работе рассматривается метод повышения эффективности таких систем посредством введения дополнительного этапа, нацеленного на преобразование исходных пользовательских запросов в формы, более точно соответствующие входным требованиям модели распознавания изображений. Данное преобразование осуществляется при помощи интеграции вспомогательной языковой модели, которая дообучается с использованием обучения с подкреплением. В качестве сигнала вознаграждения используется отклик целевой модели распознавания, что позволяет адаптивно настраивать преобразование текстовых запросов с целью повышения точности распознавания. Представленный подход проходит оценку на ряде современных архитектур, решающих задачи детекции и сегментации, и тестируется на различных наборах данных, включающих как датасеты, находящиеся в открытом доступе, так и собственные наборы данных, созданные специально для решаемой задачи. Результаты экспериментов демонстрируют устойчивое повышение точности по сравнению с базовыми методами без дополнительного преобразования запросов.

1 Введение

Модели распознавания изображений с открытым словарём (open-vocabulary image recognition) становятся всё более значимыми в области компьютерного зрения, поскольку предоставляют возможность распознавания объектов за пределами фиксированного словаря классов, используемого при обучении. Такие подходы позволяют формулировать запросы в виде произвольных текстовых описаний, расширяя применимость моделей в условиях открытого мира. Особенно это актуально для задач, в которых заранее определить полный список релевантных категорий невозможно или нецелесообразно.

Современные модели демонстрируют высокую степень гибкости и адаптивности, обеспечивая корректную обработку ранее не встречавшихся объектов. Современные модели показывают устойчивые результаты при распознавании объектов вне обучающего распределения, позволяя использовать семантически богатые текстовые описания в качестве запросов, что открывает возможности для распознавания объектов, не представленных в исходной обучающей выборке.

Тем не менее, несмотря на свою универсальность, подобные модели остаются крайне чувствительными к формулировке входного текстового запроса. Даже незначительные изменения в синтаксисе или лексике могут приводить к радикально различным результатам, вплоть до полного пропуска целевого объекта. Такая нестабильность является следствием внутренних языковых смещений, сформированных в процессе обучения на крупных корпусах данных с неравномерным распределением аннотаций и описаний. Эти смещения обуславливают предпочтение моделей к определённым типам формулировок, что снижает предсказуемость поведения модели и усложняет её применение в критически важных сценариях.

В связи с этим особую значимость приобретает задача адаптации или автоматической трансформации пользовательских запросов таким образом, чтобы они в наибольшей степени соответствовали внутренним представлениям конкретной визуальной модели. Такая задача может быть интерпретирована как форма кросс-модальной донастройки промптов — процесса, при котором текстовый ввод оптимизируется с учётом архитектуры и характеристик системы распознавания.

В данной работе описывается подход, направленный на повышение качества работы моделей с открытым словарём путём автоматического преобразования входных текстовых запросов на естественном языке. Для этого используется вспомогатель-

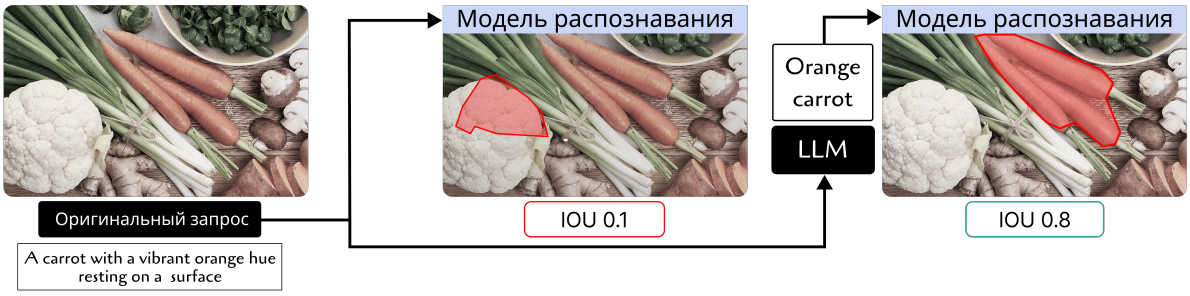


Рис. 1: Принципиальная схема предлагаемого подхода

ная языковая модель, обученная формулировать альтернативные версии исходного запроса таким образом, чтобы они соответствовали индивидуальным особенностям конкретной модели распознавания.

Адаптация осуществляется в виде циклического процесса, в котором отклик визуальной модели (например, значения IoU, значение уверенности или семантическое ранжирование) служит сигналом вознаграждения. Таким образом, языковая модель обучается находить те формулировки, которые наиболее точно соответствуют внутреннему пространству признаков модели. Оптимизация может происходить как в оффлайн-режиме — на основе агрегированной статистики по ранее полученным откликам, так и в онлайн-режиме — с динамической адаптацией к текущим входным данным. Принципиальная схема предлагаемого подхода изображена на Рис. 1

В Разделе 2 будут введены основные понятия, необходимые для понимания постановки задачи, приведен обзор существующих методов распознавания изображений с отрытым словарём, а также описаны способы улучшения качества их работы с внедрением дополнительных компонент и сформулирована формальная постановка задачи.

В Разделе 4 будут рассмотрены предложенные подходы к решению поставленной проблемы, описана их архитектура, новизна и предложенные улучшения.

В Разделе 5 будут приведены используемые наборы данных, описаны эксперименты, проведенные для оценки качества предложенных методов.

В Разделе 5.3 будут приведены результаты экспериментов с предлагаемыми методами и анализ полученных результатов.

В результате выполнения научно-исследовательской работы были получены следующие результаты:

1. Предложен механизм адаптации текстовых запросов для моделей распознавания изображений с открытым словарём, использующий вспомогательную язы-

ковую модель и сигналы обратной связи от моделей детекции и сегментации для повышения точности распознавания.

2. Предложены три способа обучения при помощи обучения с подкреплением, позволяющие учитывать как накопленные знания, так и текущие входные данные.

3. Представлен новый датасет GOLD-A, ориентированный на устранение ограничений классического набора данных, включая избыточную однородность изображений и повторяющиеся структуры расположения объектов.

4. Проведены эксперименты с предложенными подходами и архитектурами, и проведен их анализ

Полученные результаты показывают, что предложенный подход решает поставленную задачу. Разработанный алгоритм позволяет получить значительные улучшения в качестве работы моделей распознавания изображений с открытым словарём, хотя и имеет некоторые ограничения и перспективу для улучшения.

2 Обзор современных подходов

2.1 Модели распознавания изображений с открытым словарём

Это современные подходы в области компьютерного зрения, способные идентифицировать и интерпретировать объекты, не ограничиваясь заранее определённым набором категорий. В отличие от традиционных моделей, обученных на фиксированных классах, эти системы могут распознавать новые объекты, используя текстовые описания, что обеспечивает гибкость и адаптивность в разнообразных сценариях. Ключевыми направлениями в этой области являются детекция и сегментация с открытым словарём. Ниже будут рассмотрены классические и наиболее передовые модели сегментации и детекции с открытым словарём, проанализированы их положительные и отрицательные стороны.

2.1.1 Модели детекции

OV-DETR [39] это одна из самых первых моделей детекции с открытым словарём на основе трансформерных архитектур, способных находить объекты по произвольному названию класса. Расширяя классическую модель детекции DETR [22] авторы формулируют задачу детекции как бинарное сопоставление между входным запросом и объектами на изображении. Используя эмбеддинги CLIP [27] в качестве входных данных на входе трансформерного декодера, OV-DETR устраняет необходимость обучения на всех целевых классах и показывает значительные улучшения на LVIS [20] и COCO [6]. Однако, несмотря на способность обнаруживать объекты новых классов, эта модель демонстрирует более низкую точность на них по сравнению с базовыми классами. Это связано с отсутствием аннотированных данных для новых классов и потенциальным смещением модели в сторону известных категорий.

В работе **ViLD** [29] авторы решают задачу детекции с открытым словарём через дистилляцию знаний от предобученной модели (CLIP или ALIGN), что позволяет наделить детектор, построенный на основе R-CNN, языковыми знаниями, гарантируя, что полученные представления регионов включают семантику естественноязыковых описаний. Модель-учитель кодирует названия классов и регионы изображений, а ученик-детектор обучается сопоставлять свои эмбеддинги с ними. ViLD использует статическую текстовую шаблонную формулировку вида «a photo of [CATEGORY]

in the scene», позволяя адаптировать знания моделей к новым классам. Такой подход позволяет ViLD распознавать категории, отсутствующие в обучающем наборе, и превосходить полностью обученные аналоги по новым классам на LVIS. Однако этот подход все же не воспринимает полностью свободных текстовых формулировок, требуя шаблона, что ограничивает его применения в реальных задачах. Кроме того, как и другие модели на основе DETR, ViLD испытывает трудности при обнаружении мелких объектов, что может ограничивать её применение в задачах, требующих высокой точности на деталях.

Подход **Grounded Language-Image Pre-training(GLIP)** [33] объединяет задачи детекции объектов и привязки фраз к регионам изображения. Авторы рассматривают задачу детекции объектов как задачу привязки описаний к изображениям (grounding), используя дополнительные данные для обучения привязке, используя обширный корпус изображений с текстами и псевдоразметкой, чтобы найти семантическое соответствие между фразами и областями изображений. Благодаря этому модель эффективно обучается на совмещённых данных, демонстрируя выдающиеся результаты в zero-shot режиме. Однако GLIP, так же как и ViLD, требует использования текстовых шаблонов, представляющих классы объектов. Неправильно сформулированные или неоднозначные шаблоны могут снизить точность обнаружения. Кроме того, шаблоны, не соответствующие визуальному контексту, могут привести к ошибкам в распознавании.

DetCLIP [31] предлагает новый подход к предобучению моделей с открытым словарём, основанный на словаре концептов с текстовыми описаниями, собранном из различных источников. В отличие от GLIP, где классы объединяются в одно предложение, DetCLIP обрабатывает каждый концепт отдельно. В своей работе авторы объединяют масштабные наборы изображений с текстовыми описаниями и использует сгенерированные псевдо-метки для расширения базы знаний детектора, что значительно повышает его способность к обобщению на новые классы. Это позволяет эффективно использовать разнородные данные и обучать модель на более широкой семантической базе. На LVIS модель DetCLIP-T превзошла GLIP-T, особенно выигрывая на редко встречающихся и уникальных классах. Несмотря на высокое качество распознавания, DetCLIP использует глубокое слияние визуальных и текстовых признаков, что увеличивает вычислительную нагрузку и может замедлить инференс, особенно при обработке изображений с большим количеством объектов.

GroundingDINO [49] расширяет архитектуру классической трансформерной

модели детекции DINO [32], внедряя механизм слияния визуальных и текстовых признаков на различных этапах обработки. Идейно он также продолжает идеи GLIP с использованием привязки текста к регионам изображений. Ключевым нововведением GroundingDINO является интеграция языковых и визуальных модальностей. Также важной особенностью является использование текстового запроса для инициализации детекторных запросов и перекрёстного внимания в декодере. Архитектура модели включает в себя текстовый и визуальный бэкбоны, модуль объединения признаков, механизм выбора запросов, «управляемый языком», и декодер для кроссмодального взаимодействия. Это позволяет модели эффективно обрабатывать как визуальные, так и текстовые данные, обеспечивая точное обнаружение объектов на основе текстовых описаний. Реальные эксперименты показывают высокое качество распознавания при высоком быстродействии, что делает эту модель одной из самых перспективных и полезных в задачах реального применения.

Серия моделей YOLO (You Only Look Once) [11], известная своей вычислительной эффективностью, изначально ограничена фиксированным набором классов, что снижает её применимость в условиях реального мира. Решением данной проблемы выступает **YOLO-World** [67], в которой используется моделирование вида: «визуальное представление + язык» вместе с масштабным предварительным обучением. Архитектура включает в себя Reparameterizable Vision-Language Pathway Aggregation Network (RepVL-PAN), предназначенному для агрегации визуальных и языковых признаков. Это позволяет существенно повысить контекстную релевантность и точность обнаружения объектов. Также особенностью модели является использование контрастивной функции потерь между региональными признаками и текстом, что способствует более глубокой интеграции модальностей. Архитектура YOLO-World включает три основных компонента. Первый компонент — это визуальный экстрактор на основе YOLOv8, отвечающий за извлечение многоуровневых признаков с использованием сверточной нейросети. Второй компонент — текстовый энкодер, например CLIP, который преобразует текстовые подсказки в эмбеддинги, применяемые в процессе обнаружения. Третий компонент — RepVL-PAN, агрегирующий визуальные и языковые представления с применением механизмов перекрёстного внимания. Модель демонстрирует высокую эффективность в прикладных задачах, включая автоматическую разметку данных, анализ видеопотоков в реальном времени, а также использование в мобильных и встроенных устройствах, благодаря оптимизированному использованию вычислительных ресурсов.

2.1.2 Модели сегментации

Ярким примером современных методов сегментации с открытым словарём является модель **YOLO-E** [70]. Эта модель также является членом серии YOLO, однако помимо детекции поддерживает задачу сегментациии по текстовым запросам. Авторы представили модель поддерживающую текстовые запросы, визуальные промпты, а также режим работы без явных запросов, что реализуется в рамках единой и эффективной архитектуры. Построенная на базе YOLOv8/YOLOv10, модель демонстрирует передовые результаты в zero-shot сегментации при существенном снижении вычислительных затрат на обучение и ускорении инференса по сравнению с ранее предложенными подходами.

Аналогичную универсальность демонстрирует модель **SEEM** [55]. предназначенная для выполнения широкого спектра задач сегментации. Данная модель является универсальной мультимодальной моделью сегментации. Она поддерживает различные типы входных запросов — в том числе текстовые описания, ключевые точки, ограничивающие рамки, штриховые пометки и маски, — что позволяет использовать её в интерактивных сценариях и задачах с произвольным вводом. Однако, как показывают реальные эксперименты, качество работы данной модели не одинаково при каждом типе запросов или их комбинации, и наилучшие результаты распознавания достигаются при использовании только текстовых запросов.

Модель **OpenSeeD** [45] представляет собой единый фреймворк для распознавания и сегментации с открытым словарём. Обучение осуществляется одновременно на данных для задач детекции и сегментации, с использованием общего семантического пространства и раздельного механизма декодирования, что позволяет эффективно учитывать различия между типами задач и достигать конкурентоспособных результатов в zero-shot режиме. Хотя модель вводит декуплированные запросы для разделения переднего и заднего планов, она всё ещё может испытывать трудности при обнаружении мелких или перекрывающихся объектов. Это связано с ограничениями в точности локализации и сегментации в сложных визуальных сценах .

Несмотря на различия в архитектурах и методах обучения, эффективность всех перечисленных систем существенно зависит от корректного подбора текстовых описаний. Даже при наличии мощных моделей, некорректно сформулированный запрос может значительно снизить качество распознавания.

2.2 Алгоритмы подбора текстовых промтов для моделей распознавания изображений с открытым словарем

Как было показано выше, в задачах распознавания и классификации объектов с произвольным набором классов одной из ключевых проблем является выбор и подстройка текстовых запросов, описывающих категории интереса. Качество текстовых промптов напрямую влияет на эффективность связывания визуальных признаков с нужными классами. Далее будут рассмотрены различные подходы к данной задаче, разные способы их реализации, проанализированы их плюсы и минусы

Метод **Context Optimization (CoOP)** [37] предназначен для few-shot классификации и оптимизирует контекстные представления на основе непрерывных эмбеддингов. Этот метод вводит в шаблон промпта специальные контекстные вектора, обучаемые на небольшом относительно размера общей обучающей выборки числе размеченных примеров, при замороженной визуально-языковой модели (например, CLIP). Такой подход позволяет отказаться от ручного инжиниринга шаблонов и добиться высокой точности в few-shot классификации. CoOp демонстрирует высокую точность на классах, присутствующих в обучающем наборе, но значительно хуже справляется с новыми, ранее не встречавшимися классами. Это связано с тем, что обученные контекстные векторы фиксированы и не адаптируются к новым данным, что приводит к переобучению на базовые классы.

Метод **DetPro** [36], созданный на основе ViLD, расширяет идею CoOP для задач детекции. Он включает специальные механизмы: интерпретацию фоновых регионов и градацию контекста. Это позволяет автоматически адаптировать шаблоны и повысить качество обнаружения новых классов. Однако важной особенностью всех этих методов остаётся зависимость от заранее определённого набора классов, что ограничивает универсальность применяемых шаблонов. Кроме того, несмотря на введение схемы градации контекста, DetPro может испытывать трудности при обнаружении мелких или перекрывающихся объектов, особенно в сложных визуальных сценах. Это связано с ограничениями в точности локализации и сегментации в таких условиях.

Отдельный класс решений включает адаптацию запросов при помощи языковых моделей. Так, метод **OPT2I** [61] применяет итеративную оптимизацию текстовых запросов на основе отклика модели генерации изображений по описанию. Фреймворк OPT2I состоит из трёх основных компонентов: предобученной модели генера-

ции изображений, LLM и метрики согласованности между текстом и изображением. Процесс начинается с пользовательского текстового запроса, который передаётся в T2I-модель для генерации изображения. Затем сгенерированное изображение оценивается с помощью метрики согласованности, например, CLIPScore или Davidsonian Scene Graph. Полученная оценка используется для обновления истории запросов и их соответствующих оценок.

На следующем этапе LLM, используя историю предыдущих запросов и их оценок, генерирует новые формулировки запроса, стремясь улучшить метрику согласованности. Эти новые запросы снова передаются в T2I-модель, и процесс повторяется итеративно до достижения заданного уровня согласованности или достижения максимального количества итераций.

OPT2I не требует дообучения моделей и может быть применён к различным комбинациям T2I-моделей, LLM и метрик согласованности. Эксперименты, проведённые на датасетах COCO и PartiPrompts, показали, что OPT2I может повысить начальную оценку согласованности по метрике DSG, при этом сохраняя показатели FID и увеличивая recall между сгенерированными и реальными данными. Несмотря на продемонстрированную эффективность с точки зрения качества, OPT2I требует многократных обращений к LLM и генерации изображений на каждом шаге итерации. Это приводит к значительным вычислительным затратам, особенно при использовании мощных моделей генерации изображений и языковых моделей. Например, для достижения улучшения согласованности может потребоваться до 30 итераций, что делает метод ресурсоемким и неэффективным для задач, требующих быстрого отклика. Также, OPT2I был разработан и протестирован на определенных моделях генерации изображений и может не обеспечивать аналогичное улучшение согласованности на других моделях без дополнительной настройки. Это ограничивает универсальность метода и требует индивидуальной адаптации для каждой новой модели.

Promptist [64] использует алгоритм Proximal Policy Optimization (PPO) для обучения языковой модели в задаче генерации промтов. Основная цель — преобразовать исходные пользовательские запросы в формулировки, предпочтительные для модели генерации изображений, таких как Stable Diffusion [34], с целью улучшения качества и соответствия сгенерированных изображений без необходимости дообучения самой модели генерации.

Фреймворк Promptist состоит из двух основных этапов обучения. На первом этапе проводится supervised fine-tuning предобученной языковой модели на паре дан-

ных, состоящей из исходных пользовательских запросов и соответствующих им вручную созданных эффективных формулировок. Эта донастройка позволяет модели научиться преобразовывать пользовательские запросы в более эффективные для генерации изображения формулировки.

На втором этапе применяется обучение с подкреплением для дальнейшего улучшения способности модели генерировать оптимальные запросы. В процессе обучения используется функция вознаграждения, которая учитывает как релевантность сгенерированных изображений исходному запросу, так и их эстетическое качество. Таким образом, модель обучается генерировать запросы, которые приводят к созданию изображений, максимально соответствующих намерениям пользователя и обладающих высоким визуальным качеством.

Эксперименты, проведённые с использованием модели Stable Diffusion, показали, что Promptist превосходит ручную настройку запросов как по автоматическим метрикам, так и по оценкам предпочтений пользователей. Особенно значительные улучшения были достигнуты при обработке запросов, выходящих за рамки обучающего домена, что свидетельствует о высокой способности модели к обобщению.

Несмотря на продемонстрированную эффективность данного метода, его применение сопряжено с определёнными ограничениями: функции вознаграждения, используемые на этапе обучения с подкреплением, основаны на субъективных оценках эстетической привлекательности и релевантности, что затрудняет достижение стабильности и универсальности моделей, обученных при помощи данного метода.

2.3 Алгоритмы дообучения языковых моделей

Современные подходы к дообучению языковых моделей всё чаще используют методы обучения с подкреплением. Такие методы позволяют эффективно оптимизировать поведение модели на основе специфичных ограничений и требований. Тем не менее, существуют альтернативные подходы, не использующие RL, включая обучение с помощью имитации и оптимизацию с использованием прямой оценки предпочтений. Дальнейший анализ будет разделен на две части: в первой будут рассмотрены RL-подходы, во второй альтернативные.

2.3.1 Подходы, основанные на обучении с подкреплением

Reinforcement Learning from Human Feedback(RLHF) [19, 24] это один из самых популярных на сегодняшний день методов дообучения языковых моделей, заключающийся в использовании человека для оценки качества работы модели. Более глобально этот метод можно назвать *RL с KL штрафами*. Общая идея использования KL-расхождений в RL [5] была перенесена в область дообучения языковых моделей [15]. Существует множество работ в этом направлении, но основным является именно подход, использующий человеческий отклик в качестве вознаграждения. Главной мотивацией его появления являлось то, что различные автоматизированные метрики, такие как, например, BLEU для машинного перевода [2] или ROGUE для саммаризации [4], не могут полноценно оценить качество выполнения задачи, так как не способны ни учесть смысловое наполнение сгенерированного текста, ни гарантировать безопасность, полезность и правдивость своих ответов.

Описываемый подход глобально состоит из 3 шагов:

1. Сбор данных о человеческих предпочтениях;
2. Обучения модели вознаграждения(RM) при помощи собранного датасета;
3. Дообучение языковой модели при помощи обучения с подкреплением, максимизируя награду от RM;

Рассмотрим каждый из этих шагов по отдельности:

Сбор данных

Это первый этап, включающий в себя большую часть работы непосредственно с людьми, и, соответственно, генерирующий наибольшее количество трудностей. Здесь собираются данные о человеческих предпочтениях в условиях конкретной задачи. Это может быть выбор одного из четырех ответов модели в качестве лучшего [19], выбор одного из двух как в [24], прямые демонстрации, то есть написанные человеком «правильные» ответы, удовлетворяющие предпочтениям, например, как в [43], ответы на закрытые вопросы(с вариантами ответа да/нет) [35] или другим способом. В результате получается набор данных S.

Обучение модели вознаграждения

На полученном датасете обучается модель вознаграждения $r : X \times Y \to \mathbb{R}$, минимизирующая функцию потерь. Предложенная в работе [19] функция потерь выглядит

следующим образом:

$$loss(r_\theta) = \mathbb{E}_{(x,\{y_i\}_i,b)\sim S}[log\frac{e^{r_\theta(x,y_b)}}{\sum\limits_i e^{r_\theta(x,y_i)}}], \tag{2.1}$$

где b - номер выбранного человеком ответа. Эта функция легко обобщается на любое количество примеров, однако, на практике, для двух примеров сейчас больше используется функция потерь, предложенная в [24]:

$$loss(r_\theta) = \mathbb{E}_{(x,0,y_1,b)\sim S}[log(\sigma(r_\theta(x,y_b) - r_\theta(x,y_{1-b})))]. \tag{2.2}$$

Обе эти функции стремятся к максимизации расстояния между вознаграждением за выбранный человеком ответ и отвергнутым(-и). Предложенная принципиальная схема обучения модели вознаграждения из оригинальной статьи [19] приведена на Рис. 2

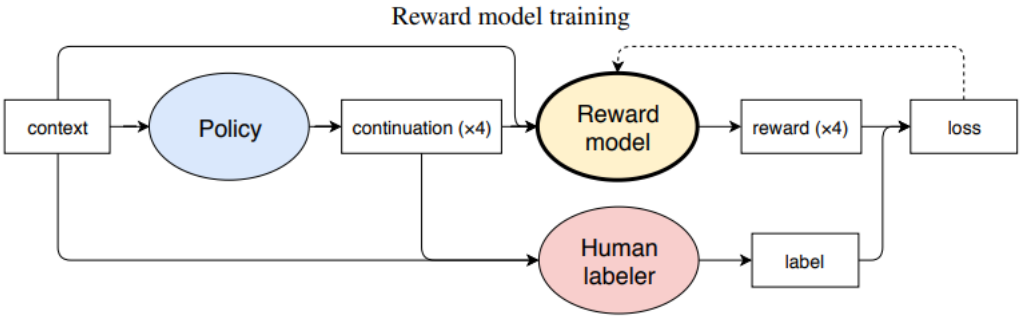


Рис. 2: Схема обучения модели вознаграждения [19]

Важно отметить, что в различных исследованиях левая часть схемы может меняться, как и виды выводов самой модели вознаграждения.

Дообучение языковой модели при помощи RL-алгоритма

Итак, есть модель вознаграждений, которая каждой паре (x, y) сопоставляет скалярное значение $r(x, y)$. Для того, чтобы ограничить обновление модели так, чтобы каждое новое распределение $\pi(s, \theta)$ не сильно отличалось от начального, вознаграждение от модели дополняется штрафом вида $\beta KL(\pi, a)$, где за a обозначено начальное распределение. Это позволяет использовать положительные лингвистические качества оригинальной предобученной модели a, а также гарантирует, отсутствие критических последствий после совершения очередного шага обновления. Итоговое вознаграждение имеет вид:

$$R(x, y) = r(x, y) - \beta log\frac{\pi_\theta(y, x)}{a(y, x)}. \tag{2.3}$$

Обновление параметров модели проводится при помощи алгоритмов обучения с подкреплением. В частности, широко применяется алгоритм Proximal Polixy Optimization [14]. Предложенная схема дообучения языковой модели из работы [19] приведена на Рис. 3

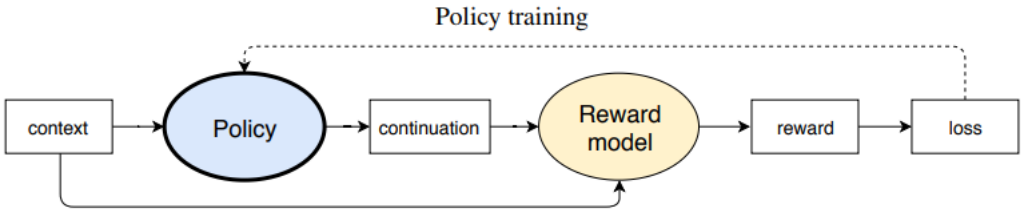


Рис. 3: Схема донастройки языковой модели [19]

Как уже говорилось выше, RLHF явялется крайне популярным подходом и применялся для обновления языковых моделей на широком спектре задач, таких как суммаризация текстов [19, 24, 30], диалог [35], задача соответствия истинности, полезности и безопасности, представляющая собой задачу дообучения языковой модели таким образом, что её ответы будут истинными, соответствовать запросу пользователя и не будут предоставлять никакой опасной информации [41, 44], а также многих других.

Однако, несмотря на свою популярность и полезность, данный подход сталкивается с рядом проблем и ограничений. Самым главным недостатком данного подхода является большое количество моделей, необходимых для обучения. Это две копии оригинальной модели(обучаемая и референсная), модель-критик из алгоритма PPO и модель вознаграждения. С ростом размерности моделей это становится очень серьезным препятствием. Есть способы снизить количество используемых ресурсов, например, Low Rank Adaptation(LORA) [28], однако это приводит к замедлению обучения ввиду увеличения количества вычислений. Также необходимо упомянуть о сложностях работы с абстрактными человеческими ценностями. Во многих работах, использующих RLHF [19, 41, 16], говорится о проблеме субъективности человеческих оценок, заключающуюся в том, что зачастую два оценщика могут присвоить кардинально разные значения вознаграждений одному и тому же ответу модели. Более того, часто даже сами авторы не согласны с оценщиками.

Метод **Reinforced Self-Training(ReST)** [53] представляет собой подход, позволяющий использовать оффлайн-RL алгоритмы для дообучения языковых моделей. Принципиально представляет собой повторение двух шагов: увеличения размера на-

бора данных и обновления параметров языковой модели. Пусть D - набор данных, π_θ - оптимизируемая стратегия, в начале обучения совпадающая с начальной моделью a, N_g - параметр количества шагов увеличения набора данных, R(x,y) - функция вознаграждения. Метод состоит из двух шагов:

1. Grow На этом шаге создается новый набор данных:
   $D_g = \{(x^i, y^i)|_{i=1}^{N_g}, x^i \sim D, y^i \sim \pi_\theta(y|x^i)\} \cup D.$
   Он аннотируется при помощи модели $R(x, y)$

2. Improve На этом шаге производится фильтрация полученного датасета:
   $F_{filter}(x, y; \tau) = 1_{R(x,y)<\tau},$
   где τ - некоторое пороговое значение вознаграждения. Далее проводится обновление параметров модели при помощи функций потерь оффлайн-RL L_{RL} или функций потерь обучения с учителем $L_{supervised}$
   Общая формула для этого шага дается формулой:

$$J(\theta) = \mathbb{E}_{(x,y)\sim D_g}[F(x, y; \tau)L_{RL}]. \tag{2.4}$$

Принципиальная схема алгоритма изображена на Рис. 4

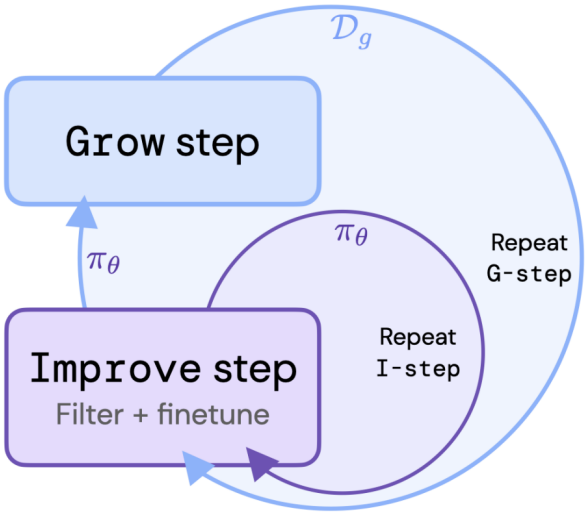


Рис. 4: Принципиальная схема ReST [53]

При обновлении весов алгоритма не используется KL-расхождения, так как в наборе данных все еще остаются старые выборки, что гарантирует, что на каждом шаге модель ответы модели не будут слишком сильно отличаться от ответов на предыдущем шаге.

Важно отметить, что так как шаг с увеличением датасета достаточно затратен с точки зрения вычислений, на один шаг типа Grow приходится несколько шагов типа Improve.

Для обеспечения улучшения результатов, порог τ постоянно увеличивается. Но с увеличением порога уменьшается размер набора данных, поэтому, чтобы не допустить переобучения, необходимо с увеличением порога снижать learning rate.

Основным практическим результатом данной работы является то, что они улучшают результаты в сравнении с PPO, который показал результаты, сравнимые с работой ReST с единственным шагом увеличения датасета.

Стоит также упомянуть работу о подходе **Reward rAnked FineTuning(RAFT)** [52], которая сделана независимо от ReST и исследует то же направление. Авторы добились улучшения по сравнению с PPO в различных задачах языкового моделирования. Они также предложили свой, более оптимальный вариант генерации порогового значения для фильтрации сгенерированного набора данных. Однако их подход эквивалентен частному случаю ReST с одним шагом типа Improve на каждый шаг типа Grow.

Group Relative Policy Optimization (GRPO) [56] Недавним дополнением к данной области стал алгоритм GRPO [56], разработанный для устранения необходимости в value-функции в алгоритме PPO. Вместо этого GRPO использует множество сгенерированных ответов на один и тот же запрос и нормализует их качества внутри группы, что упрощает обучение и повышает его стабильность за счёт перехода к относительной, а не абсолютной, шкале оценки. Пусть q — запрос, для которого стратегии $\pi_{\theta_{\text{old}}}$ генерирует G ответов $\{o_i\}_{i=1}^{G}$ с соответствующими вознаграждениями $\{r_i\}_{i=1}^{G}$.

Среднее и стандартное отклонение вознаграждений в группе:

$$\mu = \frac{1}{G}\sum_{i=1}^{G} r_i, \quad \sigma = \sqrt{\frac{1}{G}\sum_{i=1}^{G}(r_i - \mu)^2}. \tag{2.5}$$

Нормализованное преимущество:

$$\hat{A}_i = \frac{r_i - \mu}{\sigma}. \tag{2.6}$$

Отношение вероятностей токена t:

$$r_{i,t} = \frac{\pi_\theta(o_{i,t}|q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q, o_{i,<t})}. \tag{2.7}$$

Функция потерь GRPO:

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\left[\frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{[\pi_\theta(o_{i,t} \mid q, o_{i,<t})]_{\text{no grad}}}\hat{A}_{i,t} - \beta\mathbb{D}_{\text{KL}}\left[\pi_\theta \| \pi_{\text{ref}}\right]\right]. \quad (2.8)$$

GRPO имеет несколько важных преимуществ. Во-первых, он не требует использования value-функции, что упрощает обучение. Во-вторых, он использует относительное преимущество внутри группы ответов, что делает обучение более стабильным. Данный алгоритм получил широкое распространение в области дообучения языковых моделей под раличные задачи [68, 69]

2.3.2 Альтернативные подходы

Rank Responses to align Human Feedback(RRHF) [54] это подход, призванный решить проблему большого количества моделей, необходимых для донастройки. Он представляет собой расширение обучения с учителем, использующим более одного примера на каждом шаге обучения. Главная идея заключается в использовании большого количества заранее подготовленных выборок из разных моделей и стратегий, в том числе можно использовать человеческие демонстрации. Принципиально подход похож на стандартный подход к RLHF, все также есть модель вознаграждений, которая каждому ответу модели присваивает скалярное значение r, однако для ранжирования вознаграждений используются условные логарифмические вероятности:

$$p_i = \frac{\sum_t logP_\pi(y_{i,t}|x, y_{i,<t})}{\|y_i\|}. \quad (2.9)$$

Здесь $y_i, i \in \{1 \dots k\}$ - один из k ответов на промпт-запрос x, полученных из разных стратегий до начала обучения Функция потерь составляется из двух функций, первая из которых гарантирует, что большие награды даются лучшим ответам модели, а меньше - худшим:

$$L_{rank} = \sum_{r_i<r_j} max(0, p_i - p_j). \quad (2.10)$$

Вторая функция потерь является кросс-энтропией:

$$i' = argmax_i(r_i); \quad (2.11)$$

$$L_{ft} = -\sum_t logP_\pi(y_{i',t}|x, y_{i',<t}). \quad (2.12)$$

Общая функция потерь дается суммой двух предыдущих:

$$L = L_{rank} + L_{ft}. \quad (2.13)$$

Схема работы RRHF представлена на Рис. 5

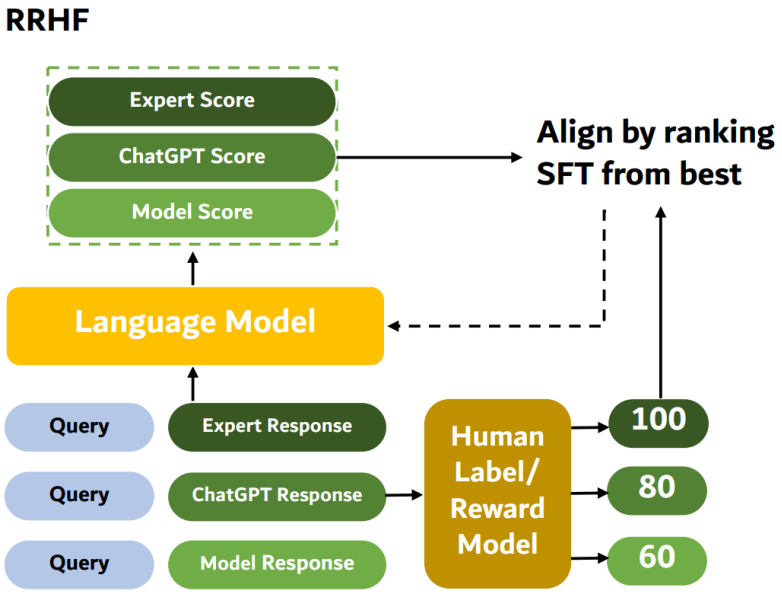


Рис. 5: Принципиальная схема RRHF [54]

В качестве основного практического результата можно выделить то, что авторам удалось достичь результатов, сравнимых с результатами RLHF, используя меньшее количество моделей.

Большим преимуществом данного подхода является снижение количества используемых моделей, что может оказаться критически важным в случае использования больших моделей, количество параметров которых исчисляется сотнями миллиардов.

Также стоит отметить, что ранжирование лучше отражает человеческие предпочтения, чем сопоставление скалярного значения.

Однако, преимущество в уменьшении ресурсов для обучения приводит к значительному увеличению использования ресурсов на этапе подготовки к обучению, так как необходимо большое количество выборок.

Generation with Distributional Control(GDC) [23] формулирует задачу дообучения языковой модели, как задачу удовлетворения ограничениям на целевом вероятностном распределении p, представляющем собой желаемую языковую модель. В частности, формулируется два вида ограничений: точечные(например, говорить только о спорте) и дистрибутивные(например, в 50% ответов упоминать персонажей женского пола). В качестве дополнительного ограничения авторы требуют обеспечения минимального KL-расхождения $D_{KL}(p, a)$ между распределением p и предварительно обученной языковой моделью a.

Итак, пусть X - набор всех последовательностей x длины L, Вероятности x равны $a(x)$ и $p(x)$ для начального и целевого распределения, соответственно. Ограничения выражаются μ_i через заранее определенные функции, $\mu_i = \mathbb{E}_{x\sim p}\varphi_i(x), i \in 1, \ldots, k..$ Так, например, ограничения, описанные в примере, могут быть выражены через бинарные функции: $\varphi_1(x) = 1$, если в x говорится о спорте, $\varphi_2(x) = 1$, если упомянут женский персонаж. Тогда ограничения примут вид: $\mu_1 = \mathbb{E}_{x\sim p}\varphi_1(x) = 1$, $\mu_1 = \mathbb{E}_{x\sim p}\varphi_1(x) = 0.5$.

Обозначим за C все распределение c над X, удовлетворяющие ограничениям. Тогда целевое распределение p, удовлетворяющее ограничениям, при этом минимизирующее KL-расхождение с начальным распределением a:

$$p \doteq \underset{c\in C}{argmin}\, D_{KL}(c, a). \tag{2.14}$$

Стоит отметить, что минимизация KL-расхождения идентична максимизации энтропии. Поэтому, говоря неформально, это задача поиска наиболее разнообразного распределения, такого, что оно будет удовлетворять всем ограничениям.

Предлагаемый авторами статьи подход основывается на двух утверждениях, которые были доказаны в информационной геометрии [3]:

1. Существует единственное решение p такого уравнения: $p(x) \propto P(x)$, где:

$$P(x) = a(x)1[x \in X_C]e^{\sum_i \lambda_i \varphi_i(x)}. \tag{2.15}$$

   Другими словами, $p(x) = 1/ZP(x)$, $Z = \sum_{x\in X} P(x)$; P - ненормализованное распределение, то есть *энергетическая модель*. Здесь $X_C = \{x \in X | \exists c \in Cm.чc(x) > 0\}$. λ_i - натуральные числа.

2. p также может быть аппроксимировано с произвольной точностью распределениями p_ε следующего вида:

$$p_\varepsilon \propto a(x)e^{\sum_i \lambda_{\varepsilon,i}\varphi_i(x)}, \tag{2.16}$$

   для соответствующих натуральных значений $\lambda_{\varepsilon,i}$

Оба этих утверждения используются для получения распределения p. Первое в случае, когда присутствуют только точечные ограничения, второе, когда ограничения строго дистибутивные или же представляют некоторое сочетание точечных и дистрибутивных.

Сам подход заключается в декомпозиции глобальной задачи на две подзадачи.

Задача 1: вычисление λ

Учитывая указанные выше утверждения, для нахождения такого оптимального распределения p, что $\mathbb{E}_{x\sim p}\varphi = \mu$, где μ это набор ограничений, из условий не могут быть определены только значения λ_i. Для вычисления этих значений применяется алгоритм, псевдокод которого изображен на Рис. 6, определяются «веса важности» $w_j(\lambda) = \frac{P(X)}{a(x)} = e^{\lambda\varphi(x_j)}$. Далее применяется Self Normalized Importance Sampling (SNIS) [9], в результате работы которого находится оценка $\hat{\mu}(\lambda)$, представляющая собой параметрическую функцию переменной λ. Для нахождения значений λ, т.ч. $\hat{\mu}(\lambda) = \mu$ применяется стохастический градиентный спуск.

Задача 2: аппроксимация целевого распределения

Algorithm 1 Computing $\boldsymbol{\lambda}$

Input: a, features $\boldsymbol{\phi}$, imposed moments $\bar{\boldsymbol{\mu}}$

1: sample a batch $x_1, \ldots, x_N$ from a

2: for each $j \in [1, N]$: $w_j(\boldsymbol{\lambda}) \leftarrow e^{\boldsymbol{\lambda}\cdot\boldsymbol{\phi}(x_j)}$

3: $\hat{\boldsymbol{\mu}}(\boldsymbol{\lambda}) \leftarrow \frac{\sum_{j=1}^{N} w_j(\boldsymbol{\lambda})\ \boldsymbol{\phi}(x_j)}{\sum_{j=1}^{N} w_j(\boldsymbol{\lambda})}$

4: solve by SGD: $\arg\min_{\boldsymbol{\lambda}} ||\bar{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}(\boldsymbol{\lambda})||_2^2$

Output: parameter vector $\boldsymbol{\lambda}$

Рис. 6: Псевдокод алгоритма для вычисления значений λ [23]

Представление в виде энергетической модели однозначно определяет оптимальную $p = Z^{-1}P$. Из этого мы можем сразу вычислить отношения вида $p(x)/p(x')$ двух последовательностей x и x', но не зная Z, нельзя вычислить p(x), а даже зная, нельзя делать выборки из него. Для решения этой типичной для энергетических моделей проблемы авторы предлагают использовать алгоритм, называемый «Distributional Policy Gradient» (DPG). DPG находит авторегрессионную стратегию π_θ, аппроксимирующую p в терминах минимизации кросс-энтропии $CE(p, \pi_\theta) = \sum_x p(x) log \pi_\theta(x)$, используя тот факт, что для любого распределения q, чей носитель содержит носитель p, верно равенство:

$$\nabla_\theta CE(p, \pi_\theta) = -\mathbb{E}_{x\sim q}\frac{p(x)}{q(x)}\nabla_\theta log \pi_\theta(x). \tag{2.17}$$

Авторы предлагают улучшить сходимость DPG путем периодического изменения распределения q в случае, если KL-расхождение между p и π_θ меньше, чем расхожде-

ние между p и q: $D_{KL}(p||\pi_\theta) < D_{KL}(p||q)$. Авторы назвали свой алгоритм KL-adaptive DPG, где адаптивной выражается в терминах изменения распределения q. Псевдокод алгоритма приведен на Рис. 7.

Algorithm 2 KL-Adaptive DPG

Input: P, initial policy q

$\pi_\theta \leftarrow q$
for each iteration **do**
  for each episode **do**
    sample x from $q(\cdot)$
    $\theta \leftarrow \theta + \alpha^{(\theta)} \frac{P(x)}{q(x)} \nabla_\theta \log \pi_\theta(x)$
  if $D_{\mathrm{KL}}(p||\pi_\theta) < D_{\mathrm{KL}}(p||q)$ **then**
    $q \leftarrow \pi_\theta$

Output: π_θ

Рис. 7: Псевдокод алгоритма KL-adaptve DPG [23]

Предлагаемый метод был применен к языковой модели GPT-2 [21] и сравнивался с тремя другими алгоритмами: двумя версиями REINFORCE с разными вознаграждениями, и классическим RLHF, который авторы называют Ziegler. Для оценки использовались пять метрик: $\mathbb{E}_{x\sim\pi_\theta}\varphi_i(x)$ в качестве оценки достижения цели по конкретному ограничению, $D_{KL}(p||\pi_\theta)$ и $D_{KL}(\pi_\theta||a)$, а также Dist и Self-BLEU для оценки количества повторений в одной последовательности и по всему распределению. Предлагаемый метод показывает значительные улучшения в сравнении со всеми тремя алгоритмами. Хотя Ziegler показывает схожие результаты, он значительно уступает в стабильности. Также применение GDC к GPT-2, дообученному на наборе биографий, позволило значительно улучшить качество работы модели как в условиях одного или нескольких дистрибутивных ограничений, так и в случае гибрида точечных и дистрибутивных.

Несомненным преимуществом предложенного GDC является возможность использовать как точеные, так и дистрибутивные ограничения, а также их комбинации. Это первый подход, который позволяет использовать настолько широкий спектр ограничений.

К проблемам же данного метода стоит отнести необходимость задавать ограничения явно. Это значительно снижает возможности применения, например, когда необходимо соответствовать абстрактным человеческим предпочтениям, таким как

безопасность, правдивость и полезность.

f-DPG [47] является принципиальным обобщением всех методов, использующих меры из класса f-дивергенций. Авторы вводят новый подход, который называют f-DPG, позволяющий любой мере из этого класса аппроксимировать любое целевое распределение.

Одной из предпосылок данной работы можно назвать исследование, показавшее, что несмотря на то, что RL с KL штрафами основано только на максимизации вознаграждения, оно эквивалентно минимизации *обратной KL-дивергенции* $D_{KL}(\pi_\theta||p)$ [38].

Авторы разделяют задачу дообучения языковых моделей на две подзадачи: то, как конструируется целевое распределение и то, как оно аппроксимируется и рассматривают их по отдельности.

Целевые распределения

Авторы следуют идее, озвученной в GDC, и моделируют распределения с помощью энергетической модели(или, что эквивалентно, ненормализованного распределения) P(x). В данной работе рассматриваются три вида целевых распределений: бинарные, скалярные и дистрибутивные. Бинарные и дистрибутивные распределения полностью соответствуют тем, что были рассмотрены в предыдущей описанной мной работе, и моделируются как:

$$p_{GDC_bin} \propto a(x)b(x), \tag{2.18}$$

для бинарного распределения. Для дистрибутивного:

$$p_{GDC_dist} \propto a(x)e^{\sum_i \lambda_i \varphi_i(x)}, \tag{2.19}$$

где φ_i - признаки, λ_i - коэффициенты, т.ч. $\mathbb{E}_{x\sim p}[\varphi_i(x)] = \mu_i$, μ_i - ограничения.

Авторы также вводят распределение, основанное на RL с KL-штрафами, которое определяется, как:

$$p_{RLKL} \propto a(x)e^{\frac{r(x)}{\beta}}, \tag{2.20}$$

где β представляет собой гиперпараметр.

Аппроксимация

Для начала, необходимо ввести определение f-дивергенции. Пусть $f : (0, \infty) \to \mathbb{R}, f(1) = 0, f(0) = lim_{t\to 0} f(t), f'(\infty) = lim_{t\to 0} tf(\frac{1}{t})$. Пусть p_1 и p_2 - какие-то распределения на множестве X. Тогда f-дивергенция между p_1 и p_2:

$$D_f(p1||p2) = \mathbb{E}_{x\sim p_2}[f(\frac{p_1(x)}{p_2(x)})] + f'(\infty)p1(p2 = 0). \tag{2.21}$$

Функция f называется *генератором* D_f. Итак, задача состоит в нахождении π_θ, аппроксимирующего p. Тогда естественным образом возникает целевая функция:

$$min_\theta D_f(\pi_\theta||p). \tag{2.22}$$

Как и в предыдущей описанной работе, возникает проблема генерации выборки из p. Техника аппроксимации, описанная в данной работе, основывается на следующем утверждении:
В случае, когда носитель π_θ не зависит от параметров θ, верно, что:

$$\nabla_\theta D_f(\pi_\theta||p) = \mathbb{E}_{x\sim\pi_\theta}[f'(\frac{\pi_\theta(x)}{p(x)})\nabla_\theta log\pi_\theta(x)]. \tag{2.23}$$

Разумеется, это утверждение будет верным и в случае языковых моделей.

Авторы провели большое количество экспериментов на различных задачах обработки естественного языка, однако в данном обзоре будут приведены только самые интересные с точки зрения результатов. Для экспериментов авторы применяли 4 разных версии f-DPG, а именно KL-DPG, RKL-DPG, TV-DPG and JS-DPG, соответствующие минимизации прямого KL-расхождения, обратного, Total Variation, и дивергенции Дженсена-Шеннона. В качестве метрик выступали: Сами KL-расхождения, $D_{KL}(\pi_\theta||a)$, $\mathbb{E}_{x\sim\pi_\theta}\varphi_i(x)$, нормализованная энтропия и псевдонаграда, представляющая собой величину $r_\theta(x) = -f'(\frac{\pi_\theta(x)}{p(x)})$. В качестве языковых моделей выступали GPT-2 «small» и T5

В экспериментах по дообучению для соответствия скалярным предпочтениям, то есть по аппроксимации целевого распределения p_{RLKL}, в качестве вознаграждения выступала величина $r(x) = log\varphi(x)$, где величина φ - вероятность, возвращенная классификатором сентиментов. Интересным в результатах следующее: хоть RKL-DPG и показал наилучшие результаты, по минимизации RKL, он полностью проваливается в минимизации остальных расхождений. Иными словами, минимизация одного расхождения не влечет за собой минимизацию остальных.

Также необходимо упомянуть, что согласно приведенным исследованиям, увеличение размера модели никак не влияет на общее качество работы разных вариантов f-DPG в условиях конкретной задачи. Имеется в виду, что, например, KL-DPG все также останется наихудшим при решении задачи с бинарными ограничениями, вне зависимости от размера модели.

Говоря о проблемах, стоит отметить, что так как данный подход использует явные формулировки целевого распределения, он все так же сталкивается с проблемой необходимости явно задавать каждое ограничение. Также подход с явным заданием целевых распределений подвержен проблеме неправильного определения(misspecification). Она возникает, когда параметрическое семейство π_θ не включает в себя p. Это может крайне негативно повлиять на итоговый результат.

Direct Preference Optimization(DPO) [48] Данный метод направлен на решение таких типичных проблем RLHF, как сложность и нестабильность. Авторы предлагают простой подход к оптимизации стратегии, используя предпочтения напрямую. В отличие от методов RLHF, которые изучают вознаграждение, а затем оптимизируют его с помощью RL, их подход использует конкретный выбор параметризации модели вознаграждения, который позволяет изучить оптимальную стратегию без цикла обучения RL. Ключевая идея заключается в использовании аналитического сопоставления функций вознаграждения с оптимальными стратегиями, что позволяет преобразовать функцию потерь в терминах функций вознаграждения в функцию потерь в терминах стратегий.

Авторы используют модель Breadly-Terry(BT) [1], предполагающую, что распределение человеческих предпочтений может быть представлено в виде:

$$p^*(y_1 \succ y_2|x) = \frac{e^{r^*(x,y_1)}}{e^{r^*(x,y_1)} + e^{r^*(x,y_2)}}. \tag{2.24}$$

Здесь x, y_1y_2 - запрос и ответы языковой модели соответственно, r^* - некоторая скрытая функция вознаграждения, которая генерирует предпочтения.

Итак, предлагаемый метод основывается на следующих размышлениях: Берется явное целевое распределение $p_{RLKL} = \frac{1}{Z(x)}a(x,y)e^{\frac{r(x,y)}{\beta}}$, $Z(x) = \sum_y a(x,y)e^{\frac{r(x,y)}{\beta}}$. Очевидно, что любое оптимальное решение π_rзадачи RLHF будет представимо в таком в виде. Далее, при помощи относительно простых алгебраических преобразований получается выражение для функции вознаграждения, зависящее только от целевого распределения, начального и функции $Z(x)$:

$$r(x,y) = \beta log\frac{\pi_r(y|x)}{a(y|x)} + \beta log Z(x). \tag{2.25}$$

Далее используется наблюдение, что модель BT зависит только от разницв в вознаграждениях между двумя ответами языковой модели: $p^*(y_1 \succ y_2|x) = \sigma(r^*(x,y_1) - r^*(x,y_2))$. Поэтому член с $Z(x)$ в (2.25) может быть опущен. Тогда выражение для вероятности, что ответ y_1 будет предпочтительнее ответа y_2 может быть выражен в

терминах оптимального и начального распределений:

$$p^*(y_1 \succ y_2|x) = \frac{1}{1 + exp(\beta log \frac{\pi^*(y_1|x)}{a(y_1|x)} - \beta log \frac{\pi^*(y_2|x)}{a(y_2|x)})}. \tag{2.26}$$

Тогда функция потерь для стратегии тогда принимает вид, аналогичный уравнению (2.2) для модели вознаграждения:

$$L_{DPO}(\pi_\theta, a) = -\mathbb{E}_{(x,0,y_1,b)\sim S}[log(\sigma(\beta log \frac{\pi_\theta(y_1|x)}{a(y_1|x)} - \beta log \frac{\pi_\theta(y_{1-b}|x)}{a(y_{1-b}|x)}))]. \tag{2.27}$$

Простыми словами, эта функция потерь вычисляет вероятность того, что предпочтительный ответ модели (согласно данным о предпочтениях человека) более вероятен, чем менее предпочтительный ответ, учитывая контекст x. Соответственно, алгоритм, предлагаемый авторами для дообучения языковой модели, состоит из 2 шагов:

1. Сбор и оценка данных, аналогичный RLHF. На выходе также получается датасет S

2. Дообучение языковой модели π_θ путем минимизации L_{DPO}, при данных S и β

В оригинальной работе DPO был применен к 3 задачам NLP: стилистическое продолжение, суммаризация и диалог. Для оценки использовались значения вознаграждения и KL-дивергенции, а также процент побед в сравнении с моделью-соперником. На всех задачах DPO значительно превосходит PPO.

Главным преимуществом DPO является его простота. Этот подход использует всего один этап для дообучения, в то время как RLHF использует два: обучение модели вознаграждения и дообучение языковой модели. Помимо очевидного снижения вычислительных затрат, это также потенциально позволит повысить стабильность и надежность обучения.

Говоря о проблемах DPO, нужно упомянуть, что он в значительной степени подвержен особенностям датасета S, который генерируется человеком, а значит, сталкивается с теми же проблемами, что и RLHF, такими как субъективность, абстрактность и многими другими. Также сами авторы отмечают необходимость дальнейших исследований возможностей обобщения стратегий DPO, так как приведенные эксперименты не до конца исследовали эту область.

2.3.3 Сравнение подходов к дообучению языковых моделей

Для сравнения их друг с другом каждый из подходов может быть охарактеризован по трем признакам: минимальное количество используемых моделей, исполь-

зование какой-либо дивергенции, а также то, явным или неявным является целевое распределение в данном подходе. Характеристики каждой из моделей приведены в таблице 1

Номер	Название подхода	Минимальное количество моделей	Использование дивергенции	Целевое распределение
1	RLHF(PPO)	4	Да	Неявное
2	RLHF(GRPO)	3	Да	Неявное
3	ReST	2	Нет	Неявное
4	RRHF	1	Нет	Неявное
5	GDC	2	Да	Явное
6	f-DPG	2	Да	Явное
7	DPO	2	Да	Неявное

Таблица 1: Сравнение различных подходов к дообучению языковых моделей

Снижение количества используемых моделей, разумеется, положительный аспект, однако, этого редко получается достичь без последствий. Так в [54] получилось снизить количество моделей при обучении до одной, однако, ценой такого снижения является кратное увеличение количества ресурсов, используемых для генерации до обучения. Отказ от использования дивергенций в этом смысле немного более перспективен. Так, например, в [53] оптимизируемая стратегия не сравнивается ни с начальным распределением, ни с каким-либо целевым, оптимизация происходит с опорой только на модель вознаграждения. Это также имеет ряд значительных недостатков, о чем будет сказано далее, однако в терминах снижения количества используемых моделей это улучшение. С другой стороны, использование дивергенций хоть и увеличивает количество моделей, но также гарантирует, что в результате очередного шага обновления параметров θ не произойдет никаких критических ошибок, так как требует постоянного сравнения либо с начальным распределением [19, 24], либо с целевым[23, 47].

Использование явного распределения[23, 47] позволяет четко понимать качество выполнения задачи, так как ставит явную же цель: достичь целевого распределения. Однако, такие подходы сталкиваются с проблемой неправильного определения, что

может привести к тому, что эта цель станет недостижимой. К тому же, применение такого подхода к сложным задачам, таким как соответствие человеческим предпочтениям в целом, становится крайне непростым, так как вынуждает явно определять каждое ограничение. С другой стороны, использование неявного распределения [19, 24, 54, 53, 48] требует дополнительной модели или более того, человека, для понимания, какие изменения в стратегии хорошие, а какие плохие. Эта проблема не актуальна для явного распределения, потому что полезность изменений можно оценить по приближению к целевому распределению. К тому же, опора на модель вознаграждений сама по себе является опасной, потому что эти модели часто подвержены проблемам ошибок и высокой дисперсии, так как зачастую обучаются на основе обратной связи от человека, качество которой зависит от количества респондентов, предложенного метода оценки, а также непосредственно от конкретных людей, привлеченных для оценки. Именно поэтому предложенный в [53] подход с использованием исключительно модели вознаграждения является сомнительным с точки зрения универсальности и надежности. К тому же, нельзя не упомянуть субъективность оценки. Так, например, легко представить ситуацию, когда модели, обученные по одинаковому алгоритму, с одинаковым размером датасетов, и даже одинаковыми промптами и ответами в них, но оцененными разными людьми, в итоге будут давать совершенно разные вознаграждения. Этот вопрос был исследован в работе [24], и даже удалось добиться высокого процента согласия между оценщиками и авторами, что может говорить о некотором повышении общей объективности, однако это потребовало огромного количества человеко-часов со стороны оценщиков и крайне четкой формулировки задачи со стороны авторов.

3 Формальная постановка задачи

В данном разделе представлена формальная постановка исследуемой задачи. В разделе 3.1 даётся её обобщённое математическое описание, позволяющее выделить ключевые переменные и ограничения. Затем, в разделе 3.2 задача формулируется в терминах марковского процесса принятия решений, что обеспечивает основу для последующего применения методов обучения с подкреплением.

3.1 Формальная постановка задачи в общем виде

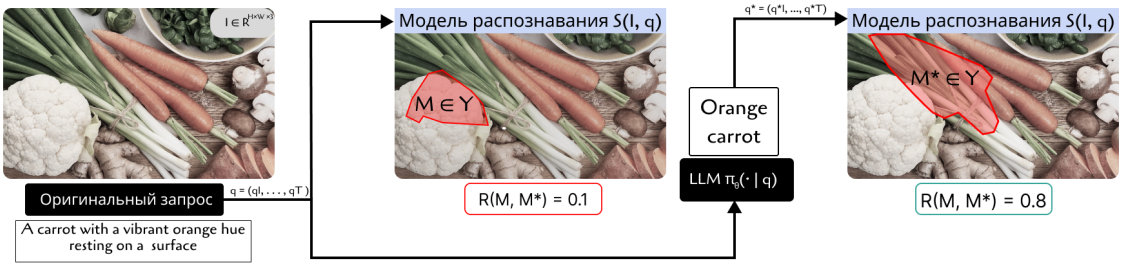


Рис. 8: Формальная предлагаемого подхода

Пусть $I \in \mathbb{R}^{H\times W\times 3}$ — входное изображение, $q = (q_1, \ldots, q_T) \in \mathcal{V}^T$ — исходный текстовый запрос в виде последовательности токенов, а $q^* = (q_1^*, \ldots, q_{\tilde{T}}^*) \sim \pi_\theta(\cdot \mid q)$ — переформулированный запрос, сгенерированный языковой моделью $\pi_\theta(\cdot \mid q)$ с параметрами θ, действующей как стохастическая автогрегрессионная стратегия.

Модель распознавания $\mathcal{S}(I, q^*) = \hat{M} \in \mathcal{Y}$ принимает изображение и запрос, и возвращает предсказание:

- В задаче сегментации $\mathcal{Y} = \{0, 1\}^{H\times W}$ и $\hat{M}$ — бинарная маска;
- В задаче детекции $\mathcal{Y} = \mathbb{R}^4$ и $\hat{M}$ — bounding box.

Пусть $M^* \in \mathcal{Y}$ — эталонная маска соответствующего формата. Функция вознаграждения определяется как —

$$R(\hat{M}, M^*) = \mathrm{IoU}(\hat{M}, M^*) = \frac{|\hat{M} \cap M^*|}{|\hat{M} \cup M^*|},$$

и представляет собой метрику качества предсказанной маски относительно эталонной.

Целью является оптимизация параметров θ переформулирующей модели π_θ таким образом, чтобы максимизировать ожидаемое значение функции вознаграждения:

$$\max_\theta \ \mathbb{E}_{(I,q)\sim\mathcal{D},\ q^*\sim\pi_\theta(\cdot|q)} \left[\mathrm{IoU}(\mathcal{S}(I,q^*),M^*)\right].$$

Оптимизация параметров языковой модели π_θ формулируется как минимизация функции потерь $\mathcal{L}(\theta)$, определённой как отрицательное ожидаемое значение функции вознаграждения:

$$\mathcal{L}(\theta) = -\mathbb{E}_{(I,q)\sim\mathcal{D},\ q^*\sim\pi_\theta(\cdot|q)} \left[\mathrm{IoU}(\mathcal{S}(I,q^*),M^*)\right].$$

В данной работе предлагается использовать алгоритмы обучения с подкреплением для донастройки языковых моделей. Следовательно, задачу дообучения языковой модели следует формализовать в терминах марковского процесса принятия решений. Это позволит верно формализовать задачу адаптации модели под сложные критерии качества.

3.2 Дообучение языковой модели как марковский процесс

Марковский процесс принятия решений задаётся кортежем $(\mathcal{S},\mathcal{A},P,R,\gamma)$, где $\mathcal{S}$ — множество состояний, $\mathcal{A}$ — множество действий агента, $P(s'|s,a)$ — вероятности перехода в состояние s' из состояния s при выполнении действия a, $R(s,a)$ — функция вознаграждения, возвращающая числовой сигнал качества за действие a в состоянии s, и $\gamma\in[0,1]$ — коэффициент дисконтирования будущих вознаграждений. Задача обучения с подкреплением заключается в поиске оптимальной стратегии $\pi(a|s)$, которая максимизирует ожидаемое суммарное дисконтированное вознаграждение агента за эпизод взаимодействия со средой. Далее будет показано, как каждый из этих элементов определяется для задачи генерации текста языковой моделью при её дообучении.

3.2.1 Состояния и переходы

В контексте языковой модели естественным образом определить состояние $s\in\mathcal{S}$ как текущее текстовое состояние: например, первоначально s_0 содержит заданный контекст или запрос, а в процессе генерации состояния будут представлять собой

частично сгенерированные тексты. Формально состояние на шаге t может быть определено как последовательность уже предсказанных моделью токенов $y_{<t} = (y_1, y_2, \ldots, y_{t-1})$ с учётом исходного промпта, если он есть. Марковское свойство при таком выборе выполняется: вероятностное распределение следующего шага зависит только от текущего состояния (текущих сгенерированных токенов), а не от более ранней истории. Правило перехода $P(s'|s, a)$ описывает динамику изменения состояния. В задачах генерации текста этот переход обычно детерминирован: если агент (модель) выбирает в состоянии s некоторый токен a в качестве следующего слова, то новое состояние s' будет содержать расширенную последовательность, включающую выбранный токен. Иначе говоря, каждый шаг генерации добавляет очередной токен к выходному тексту, формируя новое состояние. Эпизод (полная последовательность взаимодействий) завершается, когда модель предсказывает специальный токен конца последовательности либо достигнута максимальная длина ответа. Таким образом, единичный эпизод MDP соответствует генерации моделью одного завершённого фрагмента текста (например, ответа на запрос).

3.2.2 Действия агента

Действия $a \in \mathcal{A}$ в описываемом MDP соответствуют выбору языковой моделью очередного токена (слова или символа) в тексте. Пространство действий $\mathcal{A}$ совпадает с словарём модели, то есть всеми токенами, которые она может предсказать. На каждом шаге t агент наблюдает текущее состояние (уже сгенерированный текст) и согласно своей стратегии $\pi_\theta(a|s_t)$ выбирает следующий токен. Стратегия π_θ параметризуется весами модели θ; фактически это и есть вероятностное распределение, которое языковая модель задаёт на пространствах слов в зависимости от контекста. При дообучении параметры θ обновляются для оптимизации некоторого критерия качества, задаваемого моделью вознаграждения, что в MDP-формулировке эквивалентно улучшению стратегии с точки зрения получаемых наград.

3.2.3 Функция вознаграждения

Ключевым компонентом MDP является функция вознаграждения $R(s, a)$, определяющая, насколько ценным является действие a, совершённое в состоянии s. В контексте RLHF награда формируется на основе предпочтений человека. В частности, сначала обучается отдельная модель вознаграждения, обученная на определен-

ным образом размеченных и оценённых людьми ответах языковой модели, и возвращающая скалярное значение, отражающее степень удовлетворённости пользователя или соответствия некоторым заранее определенным критериям качества. Обратная связь от человеческой оценки может быть представлена различными способами, например в виде пар сравнений «какой из двух ответов лучше» [24, 12]. После обучения модели вознаграждения она используется как R: то есть, для каждого законченного ответа, сгенерированного агентом, вычисляется оценка $R = R(s_{\text{final}}, a_{\text{final}})$, которая служит сигналом вознаграждения для алгоритма обучения с подкреплением. В случае же некачественного или нежелательного вывода награда может быть отрицательной, что даёт возможность явно «наказывать» модель за нежелательное поведение. В случае задачи обучения языковой модели с учетом особенностей модели распознавания, нет необходимости в каком-либо дополнительном обучении. Сама модель распознавания может выступать в качестве модели вознаграждения. В качестве скалярного значения вознаграждения может выступать некая функция от выходных значений распознавания(масок, мер уверенности распознавания), отражающая качество распознавания объектов на основе текстового запроса, сгенерированного языковой моделью. Функция потерь Для оптимизации стратегии $\pi_\theta(a|s)$ в рамках сформулированной задачи используется градиентная оптимизация параметров θ на основе функции потерь $\mathcal{L}(\pi, R)$, которая зависит от вознаграждения R, полученного в ходе взаимодействия агента с окружением. В процессе обучения функция потерь служит связующим звеном между вознаграждением R и обновлением параметров стратегии $\pi(s, \theta)$, позволяя корректировать поведение модели таким образом, чтобы максимизировать ожидаемое суммарное вознаграждение:

$$\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}(\pi(s, \theta), R), \tag{3.1}$$

где α — параметр learning rate, отвечающий за скорость обучения.

Таким образом, функция потерь направляет обучение агента в пространстве состояний и действий, формируя стратегию, способную генерировать последовательности токенов, максимизирующие качество предсказаний с точки зрения модели распознавания.

3.3 Формальная постановка задачи в виде марковского процесса

Формализовав задачу дообучения языковой модели с учетом отклика от модели распознавания $\mathcal{S}$ изображений I как MDP, получим:

- *Пространство состояний* S: область всех возможных предсказанных токенов, состояние $s \in S$ - текущая текстовая последовательность;
- *Пространство действий* A: ответы модели на запросы, действие $a \in A$ - предсказание следующего токена;
- *Функция вознаграждения* R: $R = \mathrm{IoU}(\mathcal{S}(I, q^*), M^*)$, $q_i = \pi(s_i)$;
- *Стратегия* $\pi(s, \theta)$: сама языковая модель, которая каждому состоянию(текущей последовательности токенов) сопоставляет вероятностное распределение действий(возможных следующих токенов);
- *Функция потерь*

$$\mathcal{L}(\theta) = \mathcal{L}(\pi(s, \theta), R) = -\mathbb{E}_{(I,q)\sim\mathcal{D},\ q^*\sim\pi_\theta(\cdot|q)} \left[\mathrm{IoU}(\mathcal{S}(I, q^*), M^*)\right];$$

  - функция от текущей стратегии и полученного вознаграждения, используемая для обновления параметров обучаемой языковой модели

4 Разработка метода улучшения качества работы нейросетевых трансформерных моделей распознавания изображений

4.1 Общая структура разрабатываемого подхода

В данной работе предложен новый подход к адаптации текстовых запросов для моделей распознавания изображений с открытым словарём, ориентированных на задачи детекции и сегментации. Разрабатываемый метод направлен на повышение качества распознавания за счёт автоматизированной переформулировки пользовательских текстов с использованием большой языковой модели. Схематическое представление общей архитектуры метода приведено на рис. 9.

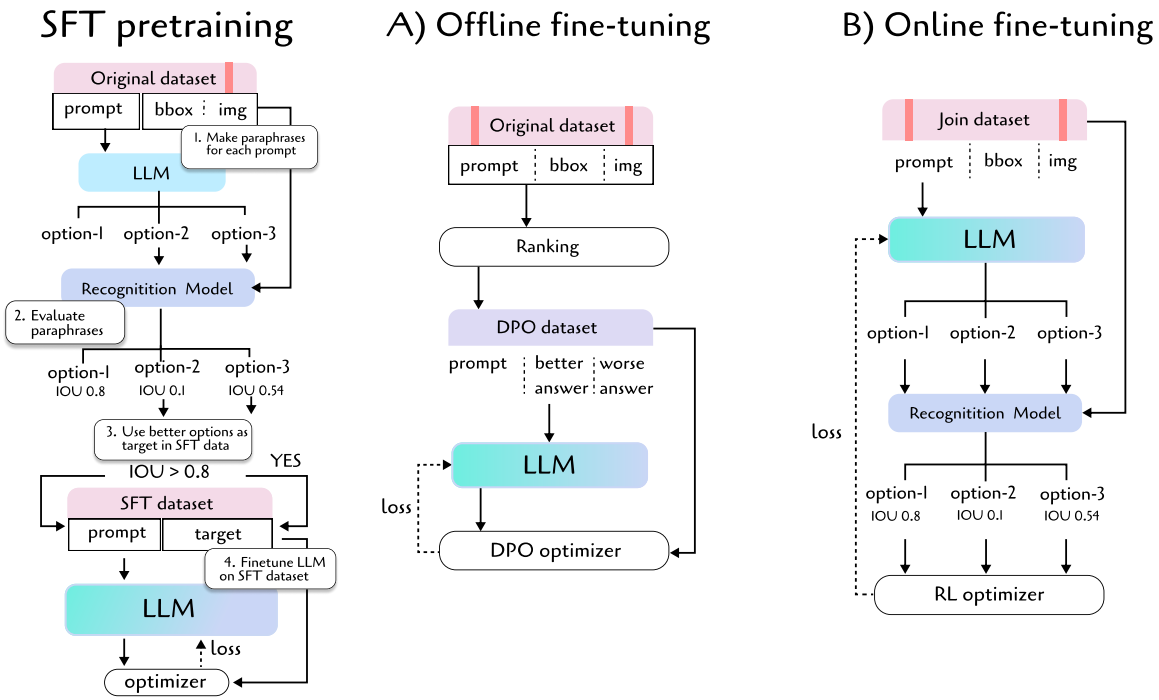


Рис. 9: Схема предлагаемого подхода

Предлагаемый подход реализуется в два этапа.

На первом этапе (см. раздел 4.2) применяется методика обучения с учителем, в рамках которой языковая модель обучается преобразовывать менее эффективные запросы в альтернативные, обеспечивающие более высокие показатели распознавания.

На втором этапе разработанный подход поддерживает два режима обучения: **offline** и **online**.
В offline-режиме обучение осуществляется исключительно на заранее размеченных данных, содержащих информацию о качестве распознавания объектов при использовании различных текстовых запросов для конкретной модели распознавания. Такой подход позволяет моделировать зависимости между формулировкой запроса и эффективностью ответа модели на основе накопленного эмпирического опыта.

В online-режиме используется онлайн-обучение с подкреплением, при котором языковая модель генерирует несколько перефразированных вариантов для каждого исходного запроса, а обратная связь формируется на основе оценки качества каждого из вариантов. В качестве основной метрики оценки используется значение Intersection over Union между предсказанным и эталонным результатом — в зависимости от задачи это может быть маска сегментации или bounding box. Полученное значение IoU служит сигналом вознаграждения для процедуры оптимизации с подкреплением.

Оба режима требуют наличия специализированного датасета определённого формата. Такой датасет должен включать исходные пользовательские запросы, их перефразированные версии, соответствующие целевые объекты, а также результаты оценки качества работы модели распознавания на различных формулировках (например, значения IoU или confidence score).

Для реализации обоих этапов обучения был сформирован **специализированный датасет** (см. раздел 4.5), включающий переформулированные версии исходных запросов и модифицированные визуальные представления целевых объектов. Для каждого объекта в датасете хранится эталонная маска, а также предсказанные маски, соответствующие различным вариантам запросов, содержащим описания данного объекта. Такая структура позволяет использовать малоэффективные и некачественные запросы в качестве входных данных, а наиболее результативные — в качестве целевых при обучении на этапе SFT. На этапе RL такая организация обогащает вариативность входных запросов и способствует повышению устойчивости вырабатываемых стратегий переформулировки.

4.2 Первый этап обучения: SFT

На первом этапе применяется обучение с учителем, целью которого является смещение вероятностного распределения выходов языковой модели в сторону таких формулировок, которые демонстрируют более высокую эффективность в задаче распознавания объектов. Для этого используется специально отобранное подмножество D_t исходного датасета D.

Для отбора обучающих примеров вычисляется 80-й перцентиль значений метрики вознаграждения в D_t, после чего выбираются только те строки, в которых значение метрики превышает указанный порог. Это позволяет отфильтровать менее информативные или неэффективные примеры.

Каждый пример в датасете содержит значение метрики IoU, отражающее качество распознавания объекта на основе конкретной формулировки. На основании этого:

- В качестве запросов используются описания с IoU < 0.8, что указывает на неудовлетворительное распознавание;
- В качестве ответов — описания с IoU $\geqslant 0.8$, соответствующие успешным результатам.

При этом гарантируется, что соответствующие описания соответствуют одному и тому же объекту.

Далее веса обновляются с использованием кросс-энтропийной функции потерь:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\sum_{t=1}^{T} \log \pi_\theta(y_t \mid y_{<t}, x) \right], \tag{4.1}$$

где θ — параметры языковой модели, x — входные даннные(запрос), $y = (y_1, \ldots, y_T)$ — целевая последовательность токенов(ответ), $y_{<t}$ — префикс целевой последовательности до позиции t, $\pi_\theta(y_t \mid y_{<t}, x)$ — вероятность, присвоенная моделью токену y_t на шаге t, $\mathcal{D}$ — обучающая выборка, состоящая из пар (x, y).

Таким образом, на этапе SFT обучение модели осуществляется на примерах, эмпирически демонстрирующих улучшение качества распознавания. Это позволяет сместить вероятностное распределение выходов языковой модели в сторону таких генераций, которые соответствуют требованиям целевой задачи.

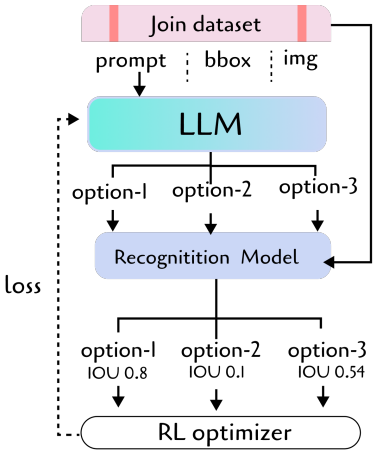


Рис. 10: Принципиальная схема предлагаемого онлайн-подхода

4.3 Второй этап обучения: онлайн-режим

Данный этап позволяет языковой модели формировать новые, ранее не встречавшиеся ответы, способствующие повышению качества работы модели. Обучение происходит в виде циклического процесса: языковая модель поочерёдно генерирует текстовые ответы, оценивает их с помощью модели распознавания и обновляет свои параметры на основе полученной обратной связи. Принципиальная схема предлагаемого онлайн подхода приведена на Рис. 10.

В качестве метода обучения предлагается два варианта функции потерь. Первая из них это Proximal Policy Optimization [13], одна из самых популярных функций потерь в данной задаче. Градиент функции потерь PPO в используемой конфигурации имеет следующий вид:

$$\nabla_\theta L^{\text{PPO}}(\theta) = \mathbb{E}_t \left[\min \left(r_t(\theta)\hat{A}_t,\, \text{clip}\left(r_t(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_t \right) \right], \tag{4.2}$$

где:

- $r_t(\theta)$ — отношение вероятностей между новой и предыдущей стратегиями;
- $\hat{A}_t$ — оценка функции преимущества;
- ϵ — гиперпараметр, задающий границы обрезки (clipping).

Такой вид функции потерь ограничивает величину обновлений стратегии, обеспечивая устойчивость процесса обучения.

Второй предлагаемой функцией потерь является метод Group Relative Policy Optimization (GRPO), представляющий собой современную модификацию PPO. Данный метод устраняет необходимость в явной функции ценности, используя групповые оценки преимущества и нормализацию внутри выборки. Это упрощает процесс обучения и демонстрирует улучшенные результаты в задачах, требующих сложных логических выводов.

Формула функции потерь GRPO имеет следующий вид:

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\left[\frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{[\pi_\theta(o_{i,t} \mid q, o_{i,<t})]_{\text{no grad}}}\hat{A}_{i,t} - \beta\mathbb{D}_{\text{KL}}\left[\pi_\theta \| \pi_{\text{ref}}\right]\right], \tag{4.3}$$

где:

- $r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\text{old}}(o_{i,t}|q,o_{i,<t})}$ — отношение вероятностей новой и старой стратегии для токена t в ответе o_i;

- $\hat{A}_{i,t} = \frac{r_i - \mu}{\sigma}$ — нормализованная оценка преимущества, где r_i — вознаграждение за ответ o_i, а μ, σ — среднее и стандартное отклонение вознаграждений внутри группы;

- β — коэффициент регуляризации для слагаемого KL-дивергенции;

- $\mathbb{D}_{\text{KL}}[\pi_\theta \| \pi_{\text{ref}}]$ — KL-дивергенция между текущей и эталонной стратегиями.

Такая формулировка позволяет проводить умеренные, но направленные обновления модели, при этом учитывая эффективность ответов по сравнению с другими примерами внутри группы.

Псевдокод описываемого метода с использованием PPO представлен в алгоритме 4.1.

Алгоритм 4.1 Обучение в онлайн-режиме

Вход: Датасет $\mathcal{D}$, число итераций N, число итераций N_1 для SFT, предварительно обученная модель π_θ, число сэмплируемых ответов для SFT Q

Сгенерировать датасет в формате SFT: $\mathcal{D}_{SFT}$

// Supervised Fine-Tuning (SFT)

для $k = 1$ до N_1

  $(p, \hat{y}) \leftarrow$ выборка из $\mathcal{D}_{SFT}$

  $y \leftarrow \pi_\theta(p)$

  $\pi_\theta \leftarrow \pi_\theta - \alpha\nabla(-\hat{y}\log(y))$

// Proximal Policy Optimization (PPO)

для $j = 1$ до N

  $(d, I, bbox) \leftarrow$ выборка из $\mathcal{D}$

  $p \leftarrow$ «Paraphrase sentence:: {d}»

  $y \leftarrow \pi_\theta(p)$

  $(bboxes/masks) \leftarrow$ выход модели для y

  $r \leftarrow$ вычислить вознаграждение $R(\text{IoU, score} \mid y)$

  $\mathcal{L} \leftarrow \min(r(\theta)\hat{A}, \text{clip}(r(\theta), 1-\epsilon, 1+\epsilon)\hat{A})$

  $\pi_\theta \leftarrow \pi_\theta + \beta\nabla\mathcal{L}$

4.4 Второй этап обучения: оффлайн-режим

В оффлайн-режиме используется заранее подготовленный датасет, а обучение проводится с применением метода Direct Preference Optimization (DPO) [57]. Особенностью данного метода является отсутствие прямого использования сигнала вознаграждения; вместо этого данные представляются в виде предпочтений. Для этого исходный датасет преобразуется в формат, содержащий пары альтернативных формулировок для одного и того же запроса. Принципиальная схема данного подхода приведена на Рис. 11 На этапе генерации датасета предпочтений для каждого объекта выбираются три описания: одно выступает в роли исходного запроса, а два других сравниваются между собой для оценки их релевантности визуальной модели. Описания классифицируются как «предпочтительное» (chosen) и «отклонённое» (rejected). Ранжирование осуществляется внутри каждой тройки описаний, где одно используется в качестве контекста, а два других — для сравнения. При этом предполагается, что описание в первой колонке обладает большей полезностью по сравнению с вари-

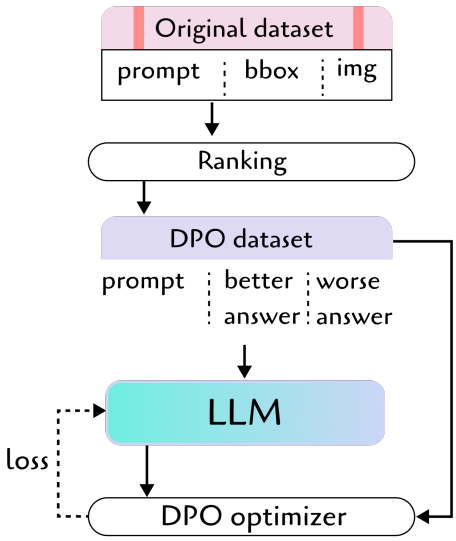


Рис. 11: Принципиальная схема предлагаемого оффлайн-подхода

антом во второй колонке. Основная идея метода DPO заключается в минимизации функции потерь L_{DPO}, для которой градиент имеет следующий вид:

$$\nabla_\theta \mathcal{L}_{DPO}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim D}\left[\sigma(\hat{r}_\theta(x, y_l) - \hat{r}_\theta(x, y_w))\left(\nabla_\theta \log \pi_\theta(y_w|x) - \nabla_\theta \log \pi_\theta(y_l|x)\right)\right], \tag{4.4}$$

где $\hat{r}_\theta(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$ — это косвенно определяемая функция вознаграждения, зависящая от текущей языковой модели π_θ и эталонной модели π_{ref}. Функция σ представляет собой логистическую сигмоиду, усиливающую вклад в градиент в тех случаях, когда оценка предпочтения является ошибочной. Таким образом, градиент функции потерь направлен на увеличение вероятности генерации предпочтительных завершений y_w и снижение вероятности генерации нежелательных y_l, с учётом силы ограничения Кульбака — Лейблера (KL-дивергенции).

Псевдокод, иллюстрирующий процесс обучения в offline-режиме, приведён в алгоритме 4.2.

4.5 Формирование датасета

Формат данных. В данной работе для задачи детекции используется модифицированная версия датасета GOLD [26], дополненная обновлёнными изображениями

Алгоритм 4.2 Обучение в оффлайн-режиме

Вход: Датасет $\mathcal{D}$, число итераций N для DPO, число итераций N_1 для SFT, предварительно обученная языковая модель π_θ

Сгенерировать датасет в формате SFT: $\mathcal{D}_{SFT}$

// Supervised Fine-Tuning (SFT)

для $k = 1$ до N_1

  $(p, \hat{y}) \leftarrow$ выборка из $\mathcal{D}_{SFT}$

  $y \leftarrow \pi_\theta(p)$

  $\pi_\theta \leftarrow \pi_\theta - \alpha\nabla(-\hat{y}\log(y))$

Сгенерировать датасет в формате DPO: $\mathcal{D}_{DPO}$

$\pi_{ref} \leftarrow \pi_\theta$

для $i = 1$ to N

  для всех кортежей (x, y_w, y_l) в $\mathcal{D}_{DPO}$

    $\nabla_\theta \mathcal{L}_{DPO}(\pi_\theta; \pi_{ref}) \leftarrow$

      $\sigma(\hat{r}_\theta(x, y_l) - \hat{r}_\theta(x, y_w)) \cdot (\nabla_\theta \log \pi_\theta(y_w \mid x) - \nabla_\theta \log \pi_\theta(y_l \mid x))$

и описаниями с целью повышения вариативности. Данный датасет был выбран в качестве основы, поскольку он использовался при обучении моделей детекции, применяемых в экспериментах. Предполагается, что отклик моделей, обученных на этих данных, должен быть максимальным, а значит, именно эти данные отражают целевые текстовые смещения конкретных моделей.

Созданный датасет включает набор изображений, ограничивающие рамки (bounding boxes) и текстовые описания объектов, представленных на этих изображениях. Процесс генерации нового датасета (далее — GOLD-A) включает серию этапов, направленных на сохранение высокого качества и релевантности по отношению к исходной выборке. Структура датасета приведена в таблице 2. Данный набор используется как в offline-, так и в online-режиме обучения в задаче детекции.

Генерация описаний. Для генерации переформулированных описаний используется предварительно обученная языковая модель в zero-shot режиме — без дополнительной адаптации под конкретный датасет. Формат текстового запроса к модели имеет вид: `«Paraphrase sentence: {sentence} [/INST]»`, что позволяет получить перефразированную версию исходного текста при сохранении его семантического содержания.

Дополнительно применяется механизм контроля качества: исключаются предло-

Таблица 2: Структура записей в датасете.

Название поля	**Описание**	**Пример**
item id	Уникальное обозначение объекта и ссылка на соответствующее изображение	*fork_1_2*
true bbox	Истинные координаты bbox для объекта	[236, 286, 231, 328]
description	Текстовое описание объекта: либо из базового набора, либо сгенерированное моделью	*A fork designed for picking up and eating food from a container is used to insert food into the mouth.*
score	Оценка от модели распознавания, используемая для фильтрации и ранжирования	0.49

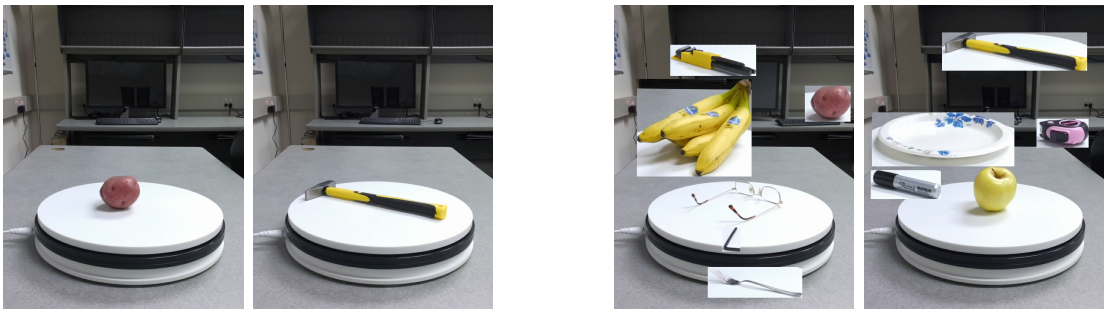
жения, содержащие вопросительные и восклицательные знаки, двоеточия и точки с запятыми; проверяется наличие финального токена и минимальная длина текста (не менее пяти символов). Это позволяет обеспечить стилистическую и синтаксическую согласованность полученных описаний.

Процедура перефразирования применяется к базовому набору описаний, привязанных к уникальным объектам с определёнными ограничивающими рамками. Для каждого объекта генерируется несколько вариантов описания, каждому из которых присваивается оценка качества. Обогащённый датасет сохраняется поэтапно с целью фиксации прогресса.

Аугментация изображений. Изначально изображения в датасете GOLD характеризуются высокой однородностью: целевой объект располагается по центру и визуально контрастирует с фоном. Это приводит к тому, что модели детекции с открытым словарём обученные на нем демонстрируют высокую точность вне зависимости от формулировки запроса, что не отражает реалий практического применения.

Для повышения сложности и реалистичности задач была проведена аугментация изображений. Целевые объекты вырезаются из исходных изображений и вставляются в случайные позиции на фоне, при этом добавляются до пяти дополнительных объектов из того же датасета. Условием является отсутствие пересечений объектов и различие классов между ними. Такой подход повышает разнообразие визуальных сцен и обеспечивает необходимость в точной формулировке запросов. Примеры изображений из GOLD и GOLD-A приведены на рисунке 12.

Оценка датасета. Расширенный датасет дополнительно обогащается данными, полученными на основе отклика модели детекции. В каждую запись добавляются предсказанные координаты ограничивающей рамки, значение IoU, отражающее



(a) Изображения из датасета GOLD

(b) Изображения из датасета GOLD-A

Рис. 12: Примеры изображений из датасетов GOLD и GOLD-A

соответствие между предсказанием и эталоном, а также confidence score, характеризующий степень уверенности модели в корректности предсказания. Эти данные используются как при отборе обучающих примеров, так и при формировании сигнала обратной связи в процессе обучения языковой модели.

5 Вычислительные эксперименты

В данном разделе представлена экспериментальная оценка предложенного метода адаптации текстовых запросов. Описаны используемые модели и датасеты, приведены технические детали реализации, а также проведены описания дополнительных экспериментов, направленных на изучение обобщаемости метода и чувствительности к выбору функции вознаграждения и полученные результаты.

5.1 Экспериментальная конфигурация

Языковые модели. В экспериментальной части рассматривались различные архитектуры языковых моделей, обученные на инструктивных корпусах. В частности, использовались модели Llama2 (7B) [50], Gemma (7B) [59] и Qwen2.5 (7B) [65]. Основным критерием выбора указанных моделей выступала их способность к zero-shot генерации, в частности в задаче перефразирования текстов. Теоретически, допускается применение более компактных моделей при условии их дообучения под конкретную задачу.

Модели распознавания изображений. Реализация предлагаемой методики производилась на двух моделях для детекции с открытым словарём: GroundingDINO и YOLO-World. Эти модели обеспечивают корректную обработку сложных естественно-языковых описаний, включая контекстуальные зависимости, такие как: «мягкая игрушка в виде медведя, лежащая на правом столике».

Для оценки генерализации разработанного подхода также была проведена серия экспериментов в задаче сегментации с открытым словарём. В данной части использовалась модель YOLO-E [70], которая на момент проведения экспериментов демонстрировала одни из лучших результатов по метрикам качества и скорости инференса. Методика обучения была адаптирована к задаче сегментации без изменения принципов взаимодействия между языковой и визуальной компонентами.

Датасеты. Для обучения и тестирования использовались три различных датасета:

- **RefCOCOg** [8] — один из стандартных наборов данных, широко применяемый в задачах языковой привязки (language grounding). Данный набор отличается высокой лингвистической вариативностью и служит устойчивой тестовой средой для оценки качества интерпретации естественного языка.

- **GOLD-A** [66] — специализированный датасет, разработанный в рамках данной работы. Он учитывает распределения, близкие к тем, что использовались при обучении целевых моделей детекции и сегментации, что делает его релевантным для оценки адаптивных стратегий.

- **External** — дополнительный тестовый набор, составленный из реальных размеченных фотографий. Для каждого изображения с помощью модели ChatGPT-4 были сгенерированы подробные описания объектов, представленных на изображении. В общей сложности датасет включает 50 уникальных описаний, охватывающих 15 различных классов. Формат описаний соответствует структуре, приведённой в разделе 4.5.

5.2 Технические детали реализации

Реализация модели проводилась с использованием библиотеки Transformer Reinforcement Learning (TRL) [25]. Все эксперименты выполнялись на одной графической карте NVIDIA A100. Размер батча при обучении составлял 16.

В каждом из предлагаемых подходов использовалось: 20 000 шагов на этапе обучения с учителем и 30 000 шагов на этапе обучения с подкреплением.

5.3 Результаты экспериментов над предлагаемым подходом в задаче детекции

В рамках проведённой работы были обучены несколько версий архитектур языковых моделей, после чего их эффективность была оценена с использованием различных моделей распознавания изображений. В качестве основной метрики применялась величина IoU, рассчитываемая для следующих категорий запросов: (1) исходные промты, из наборов данных (2) запросы, сформированные LLM без дообучения, и (3) запросы, сгенерированные LLM после обучения в offline- и online-режимах с использованием алгоритмов DPO, PPO и GRPO соответственно, на основе предлагаемого подхода.

5.3.1 Количественный анализ результатов

Проведен ряд экспериментов с целью оценки того, позволяет ли описываемый метод получать запросы, которые лучше соответствуют внутренним предпочтениям

Таблица 3: Сравнение результатов обучения различных языковых моделей для задачи детекции с открытым словарём.

Модель детекции	**Языковая модель**	**Алгоритм обучения**	**External, IoU**	**GOLD-A, IoU**
	Оригинальный промпт	-	*0.561*	*0.572*
	Gemma-7b	Предобученная	0.379	0.436
	Gemma-7b	DPO	0.398	0.523
GroundingDINO	Gemma-7b	PPO	0.542	0.564
	Llama2-7b	Предобученная	0.428	0.550
	Llama2-7b	DPO	0.448	0.562
	Llama2-7b	PPO	0.572	0.585
	Llama2-7b	GRPO	0.579	0.593
	Qwen2.5-7b	Предобученная	0.445	0.480
	Qwen2.5-7b	GRPO	**0.593**	**0.620**
	Оригинальный промпт	-	*0.501*	*0.535*
	Gemma-7b	Предобученная	0.270	0.433
	Gemma-7b	DPO	0.302	0.508
YOLO-World	Gemma-7b	PPO	0.442	0.521
	Llama2-7b	Предобученная	0.411	0.511
	Llama2-7b	DPO	0.453	0.528
	Llama2-7b	PPO	0.504	0.545
	Llama2-7b	GRPO	**0.511**	0.550
	Qwen2.5-7b	Предобученная	0.464	0.531
	Qwen2.5-7b	GRPO	0.508	**0.584**

визуальной модели. Результаты, приведённые в таблице 4, демонстрируют, что все предложенные методы, как offline-, так online-варианты обучения обеспечивают значительный прирост качества по сравнению с предобученными моделями. Кроме того, online-обучение приводит к результатам, превосходящим изначальные формулировки: так, например, модель Llama2-7b, обученная в online-режиме с использованием PPO, достигает среднего значения IoU = 0.585, тогда как для исходных пользовательских запросов этот показатель составляет 0.572. В свою очередь, модель Qwen2.5-7b обученная с использованием GRPO в предложенном online-режиме, достигает результата в 0,620, значительно превосходя качество изначальных промптов.

5.3.2 Переформулирование текстовых запросов без обучения.

Был проведен ряд экспериментов с целью изучения возможности языковой модели к переформулированию запросов без дообучения. Без специализированного обучения языковая модель действительно способна улучшать отдельные запросы за счёт добавления описательных деталей. Однако, как видно из таблицы 4, такие улучшения

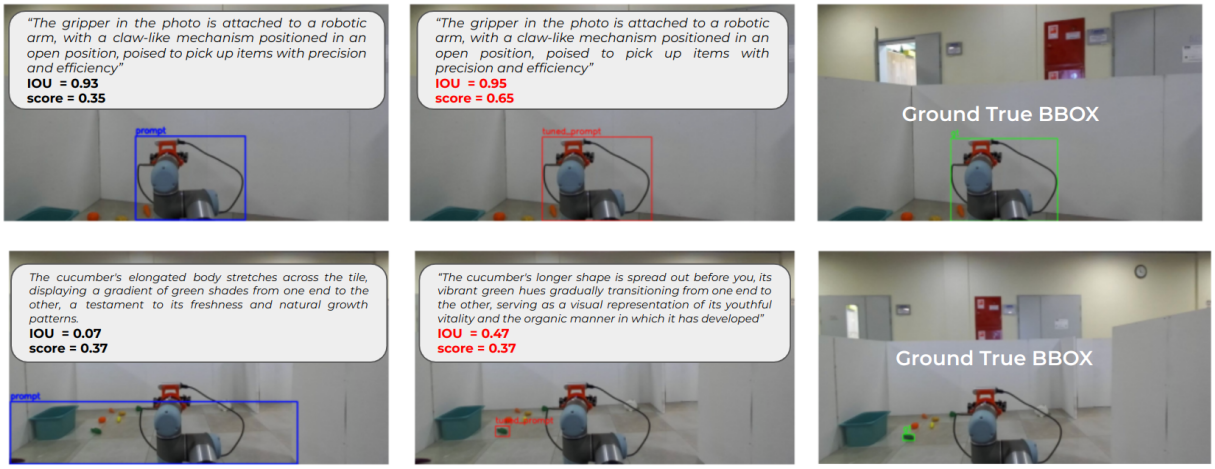


Рис. 13: Результаты обнаружения: оригинал (слева), после перефразирования (в центре), целевой bbox (справа). Перефразирование улучшает метрики и точность распознавания

Таблица 4: Сравнение запросов до и после переформулирования языковой моделью

Изначальный промпт	**IoU**	**Преобразованный промпт**	**IoU**
A sleek, black **photo frame** with squared-off edges is standing upright on a flat surface, its imposing presence drawing attention.	0.00	This is a black **picture frame**. The frame has a sleek, modern design that looks great.	0.97
The **plate** being referred to **is red in color**.	0.80	The **item is a red dinner plate** that exudes a vintage aesthetic, characterized by its classic design and timeless appeal.	0.82
A *set* of **blue scissors** with handles.	0.91	A *pair* of scissors with **blue handles**.	0.93

не носят системного характера и не учитывают специфику конкретной архитектуры модели распознавания, что ограничивает результативность метода.

5.3.3 Влияние качества исходного датасета на качество результатов.

В процессе экспериментов было проведено следующее наблюдение: среднее значение IoU для описаний из изначального набора данных обратно пропорционально максимальному достигаемому приросту качества в результате обучения. Для формализации и подтверждения этой гипотезы был проведен ряд дополнительных экспериментов. Результаты демонстрируют чёткую зависимость между качеством исходных запросов и итоговой эффективностью применения предлагаемого подхода. Как показано на рисунке 14, относительное улучшение снижается по мере изначального качества описаний. В частности, при среднем значении IoU в диапазоне от 0.1 до 0.15 достигается прирост около 30%. В случаях, когда начальное значение IoU находится в пределах 0.25–0.3, улучшение составляет 15–20%, тогда как при значениях выше

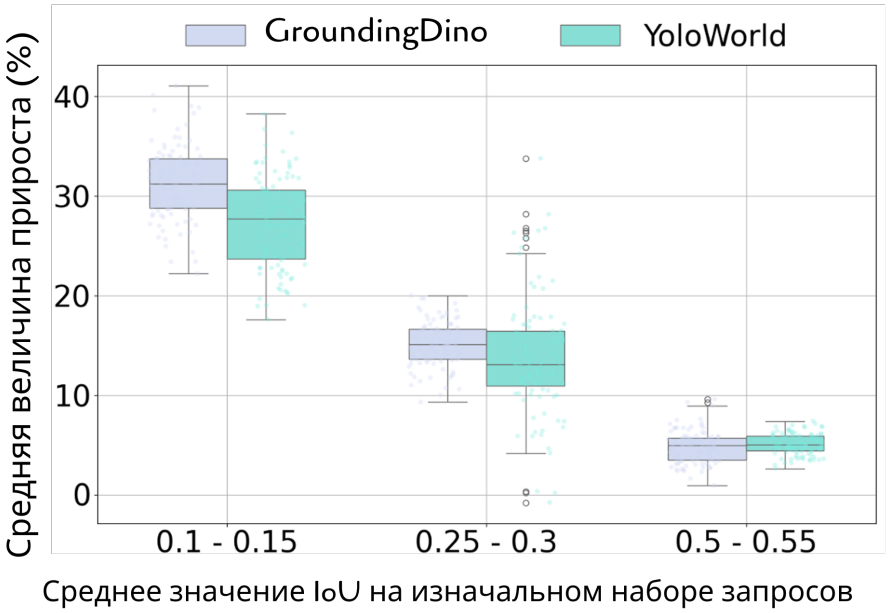


Рис. 14: Зависимость прироста качества распознавания от исходного значения IoU

0.5 эффект ограничивается 5–7%. Эти наблюдения указывают на то, что предлагаемая стратегия оптимизации текстовых запросов особенно эффективна в сценариях с низкой исходной точностью.

Подобное поведение может быть интерпретировано следующим образом: в процессе обучения языковая модель оптимизируется исключительно под задачу преобразования запросов и не получает примеров, в которых начальная формулировка должна оставаться неизменной. В результате, при применении модели к датасетам с низким качеством исходных описаний она способна преобразовать большую долю промптов, достигая более высокого прироста целевой метрики.

Напротив, при использовании датасетов, в которых большая часть запросов уже изначально сформулирована корректно, потенциал модели для улучшения снижается, поскольку число случаев, действительно нуждающихся в переформулировке, невелико. Соответственно, максимальный возможный прирост целевой метрики значительно уменьшается.

5.3.4 Качественный анализ работы предложенного метода

На Рис. 13 представлен качественный анализ предлагаемого подхода в задаче детекции. Визуализация включает эталонные маски, а также результаты детекции до и после применения механизма адаптации текстовых запросов. Как можно видеть, переформулирование запросов позволяет не только улучшить распознавание

конкретного объекта, но и находить те объекты, которые при изначальном запросе не распознаются, что работает также и для мелких объектов Семантический и синтаксический анализ описаний до и после преобразования позволит наглядно оценить влияние предлагаемого подхода. Как видно из таблицы 4, начальные формулировки зачастую оказываются недостаточно точными или перегруженными лишними деталями, что приводит к ошибкам или низкой уверенности со стороны модели распознавания. Переформулированные запросы, напротив, демонстрируют более чёткую и целенаправленную семантику, способствуя повышению точности предсказания и уверенности модели. Такие наблюдения подчёркивают критическую важность структуры текстового запроса для надёжной работы систем визуального распознавания. В частности, в одном из случаев *переформулировка описания целевого объекта в запросе* приводит к существенному повышению эффективности модели в аспекте качества детекции. В другом случае *уточнение формулировки и более точная детализация искомого объекта* обеспечивает более точное семантическое описание, что, в свою очередь, позволяет модели детекции более корректно локализовать целевой объект. В третьем случае детализация и *изменение описания объекта от более общего к более частному* позволяет незначительно улучшить качество детекции, немного подстроив целевую рамку. Эти эмпирические наблюдения подчёркивают критическую значимость тщательной и контекстуально релевантной конструкции запросов для повышения точности и достоверности предсказаний.

5.3.5 Влияние выбора функции вознаграждения на качество результата

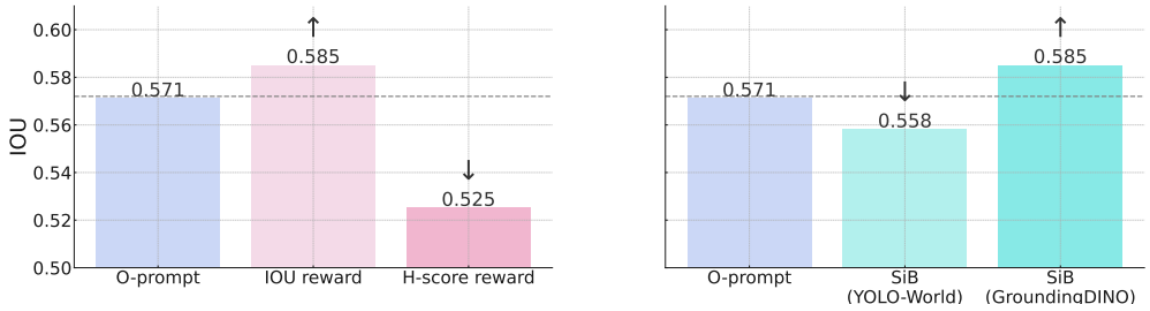
В рамках данного эксперимента была проанализирована зависимость качества обработки текстовых запросов от выбора функции вознаграждения в online-режиме для модели Llama2. Рассматривались два варианта: использование значения IoU в качестве единственного критерия и применение гармонического среднего между IoU и confidence score модели визуального распознавания. Полные результаты эксперимента представлены на рисунке 15а.

Проведённый анализ показал, что наилучшее качество достигается при использовании чистого значения IoU в качестве функции вознаграждения без дополнительных модификаций. Такой подход позволяет напрямую сопоставлять эффективность текстового запроса с результатами распознавания: выдавая высокую оценку в случаях, когда описание корректно соответствует целевому объекту, и низкую — когда

модель не может установить соответствие между запросом и изображением.

5.3.6 Возможность переноса обученных языковых моделей между различными моделями распознавания

Дополнительно была изучена возможность использования языковых моделей, обученных на одной архитектуре визуального распознавания, для повышения качества работы другой. Результаты, представленные на рисунке 15b, демонстрируют, что такие модели действительно обеспечивают прирост по сравнению с вариантом без дообучения. Однако наибольший эффект достигается в случае, когда языковая модель обучается непосредственно с учётом отклика той архитектуры, на которой впоследствии планируется её применение. Это подтверждает важность согласования между структурой обучающей выборки и характеристиками целевой модели визуального восприятия.



(a) (b)

Рис. 15: a) Сравнение результатов обучения Llama2 для детектора GroundingDINO с использованием различных функций вознаграждения. O-prompt обозначает использование модели распознавания с оригинальным промптом из датасета. b) Сравнение результатов Llama2-7b, обученной с различными моделями распознавания и протестированной на GroundingDINO.

5.4 Результаты экспериментов над предлагаемым подходом в задаче сегментации

5.4.1 Количественный анализ результатов

Проведен ряд экспериментов с целью оценки возможностей генерализации предложенного метода к другим задачам распознавания с открытым словарём, работающим с более сложными типами и формами масок. В экспериментах применялся предложенный в данной работе online-подход, как показавший наилучшие результа-

ты в задаче детекции. Для экспериментов также использовалась языковая модель Qwen2.5, показавшая наилучшее качество обучения в предыдущих экспериментах. Результаты, представленные в таблице 5, демонстрируют значительное улучшение

Таблица 5: Сравнение результатов обучения для задачи сегментации с открытым словарём.

Модель сегментации	**Языковая модель**	**Алгоритм обучения**	**RefCOCOg, IoU**
YOLO-E	*Оригинальный промпт*	-	*0.276*
	Qwen2.5-7b	Предобученная	0.232
	Qwen2.5-7	GRPO	**0.498**
SEEM	*Оригинальный промпт*	-	*0.256*
	Qwen2.5-7b	Предобученная	0.202
	Qwen2.5-7	GRPO	**0.362**

качества сегментации при использовании предложенного в работе online-подхода с языковой моделью Qwen2.5-7b и алгоритмом обучения GRPO. Для обеих моделей сегментации — YOLO-E и SEEM — наблюдается уверенный прирост в значениях целевой метрики в сравнении как с исходным вариантом (использующим оригинальные промпты), так и с вариантом, в котором применялась предобученная языковая модель Qwen2.5-7b без дообучения.

В частности, модель YOLO-E при использовании GRPO и Qwen2.5-7 достигает максимального значения IoU, равного 0.498, что почти в два раза превышает результат, полученный с оригинальным промптом (0.276), и более чем вдвое превосходит результат с предобученной Qwen2.5-7b (0.232). Это указывает на высокую эффективность совместного использования GRPO и более мощной языковой модели.

Аналогичная тенденция наблюдается и для модели SEEM. При обучении достигается значение IoU 0.362, что существенно превышает как результат с оригинальным промптом (0.256), так и результат с предобученной моделью (0.202). Несмотря на то что абсолютные значения IoU у SEEM ниже, чем у YOLO-E, прирост качества при использовании предложенного метода также весьма значительный.

Таким образом, можно заключить, что предложенный подход демонстрирует хорошую способность к обобщению на новые архитектуры и задачи, а сочетание GRPO с языковой моделью Qwen оказывается наилучшим вариантом для достижения высокой точности сегментации.

5.4.2 Качественный анализ работы предложенного метода

На Рис. 16 представлен качественный анализ предлагаемого подхода в задаче сегментации. Визуализация включает оригинальные изображения, соответствующие эталонные маски, а также результаты сегментации до и после применения механизма адаптации текстовых запросов. До перефразирования модель нередко демонстрирует ошибки в интерпретации запроса, что приводит к частичному или некорректному выделению целевых областей или полному отсутствию распознавания(пример 3). В некоторых случаях наблюдается неправильная локализация объектов, смещение границ или появление нерелевантных сегментированных областей. Это связано с несоответствием формулировок пользовательского запроса лингвистическим особенностям, которым была обучена визуальная модель. После применения перефразирования с использованием дополнительной языковой модели, сегментация становится значительно более точной. Маски демонстрируют лучшую согласованность с эталоном, объекты выделяются полнее и с более чёткими границами, снижается количество ложных срабатываний. Таким образом, адаптация текстовых описаний позволяет эффективнее согласовывать входной запрос с внутренним пространством признаков визуальной модели, что приводит к повышению качества сегментации.

Семантический анализ работы модели.

Рассмотрим приведенные на Рис. 16 запросы и проанализируем работу языковой модели, изучив как именно переформулирование помогает улучшить качество распознавания как с точки зрения модели сегментации, так и с точки зрения естественного языка. В первом примере *перестройка порядка компонентов предложения* и *смещение фокуса с пространственного положения на активное действие* позволяют улучшить качество распознавания. Первоначальный запрос — «One dog is behind the other, climbing on the window» — представляет собой неоднозначную структуру: неясно, кто именно совершает действие climbing on the window, и основной акцент сделан на расположении объектов друг относительно друга. Такая формулировка может затруднить интерпретацию для модели, особенно в определении, какой из объектов является действующим лицом. Переформулированный запрос — «The dog is climbing on the window behind the other dog» — устраняет эту неоднозначность, чётко указывая, **кто** выполняет действие (the dog is climbing) и **где** это происходит (on the window behind the other dog). Таким образом, уточнение субъектно-предикатной структуры и более точное пространственное описание обеспечивают более коррект-



Рис. 16: Примеры улучшения качества сегментации по текстовым запросам

ную локализацию целевого объекта моделью распознавания.

Во втором случае *сокращение и упрощение описания целевого объекта* приводит к улучшению результата сегментации. Изначальный запрос — «Larger of two sheep grazing with a fur on its back» — содержит избыточную и потенциально сбивающую с толку информацию: конструкция «with a fur on its back» может быть интерпретирована неоднозначно, а дополнительное описание действия «grazing» усложняет семантику запроса. В результате модель не смогла обнаружить целевой объект. В переформулированной версии — «The larger of two sheep» — происходит *фокусировка на ключевом различительном признаке* (размере), важном для идентификации нужного объекта. Это упрощает семантическую задачу и позволяет модели точно локализовать целевой объект на изображении.

В третьем примере первоначальный запрос — «left side monitor» — является краткой и синтаксически неполной фразой, в которой отсутствует чёткая структура и контекст. Такая формулировка может быть интерпретирована моделью неоднозначно: неясно, идёт ли речь об экране, расположенном слева, или о части самого монитора. Переформулированный запрос — «A monitor on the left side of the screen» — содержит *полноценную синтаксическую структуру и уточнённое пространственное*

положение объекта, что позволяет модели более точно интерпретировать намерение пользователя и корректно локализовать нужный монитор на изображении.

В четвертом примере *замена специфичного обозначения класса на более общий и типичный для модели термин* приводит к улучшению качества сегментации. Первоначальный запрос — «small girl starring at something, along with her elder sister» — включает в себя уточнение пола и возраста («small girl»), а также дополнительную, нецелевую информацию (о присутствии старшей сестры), что может затруднить фокусировку модели на нужном объекте. Более того, термин «girl» может не совпадать с классами, которые многократно встречались при обучении модели, что негативно сказывается на качестве сегментации. В переформулированном запросе — «A young person is looking at something» — используется *термин, соответствующий более распространённой категории объекта* («person» вместо «girl»), что делает описание более совместимым с внутренними представлениями модели. Это обеспечивает более точную идентификацию и сегментацию искомого объекта.

6 Заключение

Модели распознавания изображений с открытым словарём представляют собой важное направление в области компьютерного зрения, поскольку позволяют идентифицировать объекты, не входящие в исходный обучающий набор классов, что крайне актуально для реальных задач, где исчерпывающий перечень категорий заранее определить невозможно. Однако высокая чувствительность этих моделей к лексическим и синтаксическим вариациям входных описаний, вызванная языковыми смещениями обучающих данных, снижает устойчивость результатов и ограничивает надёжность их применения в критически значимых сценариях. В данной работе был предложен подход к повышению качества работы моделей распознавания изображений с открытым словарём путём адаптации текстовых запросов с учётом специфики используемой модели. Для реализации данного подхода используется дополнительная языковая модель, обученная методами обучения с подкреплением, задача которой — согласование пользовательских запросов с лингвистическими смещениями, характерными для конкретных моделей распознавания. Такая адаптация позволяет повысить точность предсказаний за счёт более точного соответствия между входными запросами и внутренними представлениями модели.

Основные результаты работы:

1. Был предложен подход к адаптации текстовых запросов для моделей распознавания с открытым словарём. Предложенный подход решает проблемы существующих методов, такие как необходимость шаблона и большого количества мультимодальных данных для обучения.

2. В рамках работы был предложен новый датасет для задачи детекции с открытым словарём: GOLD-A. Предложенный набор данных решает проблему однообразия существующего набора данных GOLD и расширяет его большим количеством текстовых описаний. Он содержит более 40000 текстовых описаний и более 800 разнообразных изображений.

3. Предложенный метод был применен к моделям детекции и сегментации с открытым словарём

4. Показано, что предложенный подход обеспечивает стабильный прирост качества на различных архитектурах моделей, охватывающих как задачи детекции, так и сегментации, что подтверждает его универсальность и эффективность.

Полученные результаты показывают, что предложенный подход решает поставленную задачу улучшения качества распознавания объектов по текстовым запросам. При этом сохраняются определённые ограничения. Во-первых, используемая стратегия аугментации данных согласует поведение языковой модели с моделью визуального распознавания с помощью механизмов глубокого обучения. Однако возможно, что применение той же модели, которая впоследствии используется для дообучения, могло бы обеспечить более эффективную интеграцию данных. Во-вторых, применяемые алгоритмы обучения с подкреплением отличаются высокой вычислительной затратностью и требуют значительных объёмов данных и ресурсов для достижения стабильных и качественных результатов.

Результаты проведенных исследований были опубликованы в следующих работах:

- Reframing: Detector-Specific Prompt Tuning for Enhancing Open-Vocabulary Object Detection / M. Avshalumov, Z. Volovikova, D. Yudin, A. Panov // International Conference on Hybrid Artificial Intelligence Systems. — Springer. 2024. — С. 128—140

- Say It Better: RL-Based Prompt Tuning for Enhancing Open-Vocabulary Recognition / Mikhail Avshalumov, Zoya Volovikova, Dmitry Yudin, Aleksandr Panov / Neurocomputing. 2025.(На ревью)

Список литературы

1. *Bradley R. A.*, *Terry M. E.* Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons // Biometrika. — 1952. — Т. 39. — С. 324. — URL: https://api.semanticscholar.org/CorpusID:125209808.
2. Bleu: a method for automatic evaluation of machine translation / K. Papineni, S. Roukos, T. Ward, W.-J. Zhu // Proceedings of the 40th annual meeting of the Association for Computational Linguistics. — 2002. — С. 311—318.
3. *Csiszár I.*, *Shields P.* Information Theory and Statistics: A Tutorial // Foundations and Trends® in Communications and Information Theory. — 2004. — Т. 1, № 4. — С. 417—528. — ISSN 1567-2190. — DOI: 10.1561/0100000004. — URL: http://dx.doi.org/10.1561/0100000004.
4. *Lin C.-Y.* Rouge: A package for automatic evaluation of summaries // Text summarization branches out. — 2004. — С. 74—81.
5. *Todorov E.* Linearly-solvable Markov decision problems // Advances in neural information processing systems. — 2006. — Т. 19.
6. Microsoft COCO: Common Objects in Context / T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick // Computer Vision – ECCV 2014. — Cham : Springer International Publishing, 2014. — С. 740—755. — DOI: 10.1007/978-3-319-10602-1_48. — URL: https://arxiv.org/abs/1405.0312.
7. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks / S. Ren, K. He, R. Girshick, J. Sun // Advances in Neural Information Processing Systems. Т. 28 / под ред. С. Cortes, N. Lawrence, D. Lee, M. Sugiyama, R. Garnett. — Curran Associates, Inc., 2015. — URL: https://proceedings.neurips.cc/paper_files/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf.
8. Generation and comprehension of unambiguous object descriptions / J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, K. Murphy // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2016. — С. 11—20.
9. *Kim T.*, *Bengio Y.* Deep directed generative models with energy-based probability estimation // arXiv preprint arXiv:1606.03439. — 2016.

10. Sequence Level Training with Recurrent Neural Networks / M. Ranzato, S. Chopra, M. Auli, W. Zaremba // International Conference on Learning Representations (ICLR). — 2016.

11. You Only Look Once: Unified, Real-Time Object Detection / J. Redmon, S. K. Divvala, R. B. Girshick, A. Farhadi // 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. — IEEE Computer Society, 2016. — C. 779—788.

12. Deep reinforcement learning from human preferences / P. F. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, D. Amodei // Advances in Neural Information Processing Systems. T. 30. — 2017.

13. Proximal policy optimization algorithms / J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov // arXiv preprint arXiv:1707.06347. — 2017.

14. Proximal policy optimization algorithms / J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov // arXiv preprint arXiv:1707.06347. — 2017.

15. Sequence tutor: Conservative fine-tuning of sequence generation models with kl-control / N. Jaques, S. Gu, D. Bahdanau, J. M. Hernández-Lobato, R. E. Turner, D. Eck // International Conference on Machine Learning. — PMLR. 2017. — C. 1645—1654.

16. Can neural machine translation be improved with user feedback? / J. Kreutzer, S. Khadivi, E. Matusov, S. Riezler // arXiv preprint arXiv:1804.05958. — 2018.

17. Texygen: A Benchmarking Platform for Text Generation Models / Y. Zhu, S. Lu, L. Zheng, J. Guo, W. Zhang, J. Wang, Y. Yu // The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. — 2018. — URL: https://api.semanticscholar.org/CorpusID:3636178.

18. Zero-Shot Object Detection / A. Bansal, K. Sikka, G. Sharma, R. Chellappa, A. Divakaran // Proceedings of the European Conference on Computer Vision (ECCV). — 09.2018.

19. Fine-tuning language models from human preferences / D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, G. Irving // arXiv preprint arXiv:1909.08593. — 2019.

20. *Gupta A.*, *Dollár P.*, *Girshick R.* LVIS: A Dataset for Large Vocabulary Instance Segmentation // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). — 2019.

21. Language models are unsupervised multitask learners / A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever [и др.] // OpenAI blog. — 2019. — Т. 1, № 8. — С. 9.

22. End-to-end object detection with transformers / N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko // European conference on computer vision. — Springer. 2020. — С. 213—229.

23. *Khalifa M.*, *Elsahar H.*, *Dymetman M.* A distributional approach to controlled text generation // arXiv preprint arXiv:2012.11635. — 2020.

24. Learning to summarize with human feedback / N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, P. F. Christiano // Advances in Neural Information Processing Systems. Т. 33 / под ред. H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin. — Curran Associates, Inc., 2020. — С. 3008—3021. — URL: `https://proceedings.neurips.cc/paper_files/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf`.

25. TRL: Transformer Reinforcement Learning / L. von Werra, Y. Belkada, L. Tunstall, E. Beeching, T. Thrush, N. Lambert, S. Huang. — 2020.

26. A Spoken Language Dataset of Descriptions for Speech-Based Grounded Language Learning / G. Y. Kebe, P. Higgins, P. Jenkins, K. Darvish, R. Sachdeva, R. Barron, J. Winder, D. Engel, E. Raff, F. Ferraro, C. Matuszek // Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1). — 2021.

27. Learning transferable visual models from natural language supervision / A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark [и др.] // International conference on machine learning. — PMLR. 2021. — С. 8748—8763.

28. Lora: Low-rank adaptation of large language models / E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen // arXiv preprint arXiv:2106.09685. — 2021.

29. Open-vocabulary object detection via vision and language knowledge distillation / X. Gu, T.-Y. Lin, W. Kuo, Y. Cui // arXiv preprint arXiv:2104.13921. — 2021.

30. Recursively summarizing books with human feedback / J. Wu, L. Ouyang, D. M. Ziegler, N. Stiennon, R. Lowe, J. Leike, P. Christiano // arXiv preprint arXiv:2109.10862. — 2021.

31. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection / L. Yao, J. Han, Y. Wen, X. Liang, D. Xu, W. Zhang, Z. Li, C. Xu, H. Xu // Advances in Neural Information Processing Systems. — 2022. — Т. 35. — С. 9125—9138.

32. Dino: Detr with improved denoising anchor boxes for end-to-end object detection / H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, H.-Y. Shum // arXiv preprint arXiv:2203.03605. — 2022.

33. Grounded language-image pre-training / L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang [и др.] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. — 2022. — С. 10965—10975.

34. High-Resolution Image Synthesis with Latent Diffusion Models / R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). — 2022. — С. 10684—10695. — DOI: 10.48550/arXiv.2112.10752. — URL: https://arxiv.org/abs/2112.10752.

35. Lamda: Language models for dialog applications / R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du [и др.] // arXiv preprint arXiv:2201.08239. — 2022.

36. Learning to prompt for open-vocabulary object detection with vision-language model / Y. Du, F. Wei, Z. Zhang, M. Shi, Y. Gao, G. Li // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. — 2022. — С. 14084—14093.

37. Learning to prompt for vision-language models / K. Zhou, J. Yang, C. C. Loy, Z. Liu // International Journal of Computer Vision. — 2022. — Т. 130, № 9. — С. 2337—2348.

38. On reinforcement learning and distribution matching for fine-tuning language models with no catastrophic forgetting / T. Korbak, H. Elsahar, G. Kruszewski, M. Dymetman // Advances in Neural Information Processing Systems. — 2022. — T. 35. — C. 16203—16220.

39. Open-vocabulary detr with conditional matching / Y. Zang, W. Li, K. Zhou, C. Huang, C. C. Loy // European Conference on Computer Vision. — Springer. 2022. — C. 106—122.

40. Simple Open-Vocabulary Object Detection with Vision Transformers / M. Minderer, A. A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen, X. Wang, X. Zhai, T. Kipf, N. Houlsby // ArXiv. — 2022. — T. abs/2205.06230. — URL: `https://api.semanticscholar.org/CorpusID:248721818`.

41. Training a helpful and harmless assistant with reinforcement learning from human feedback / Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan [и др.] // arXiv preprint arXiv:2204.05862. — 2022.

42. Training language models to follow instructions with human feedback / L. Ouyang [и др.] // Advances in Neural Information Processing Systems. T. 35. — 2022. — C. 27730—27744.

43. Training language models to follow instructions with human feedback / L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray [и др.] // Advances in Neural Information Processing Systems. — 2022. — T. 35. — C. 27730—27744.

44. Training language models to follow instructions with human feedback / L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray [и др.] // Advances in Neural Information Processing Systems. — 2022. — T. 35. — C. 27730—27744.

45. A Simple Framework for Open-Vocabulary Segmentation and Detection / H. Zhang, F. Li, X. Zou, S. Liu, C. Li, J. Gao, J. Yang, L. Zhang // arXiv preprint arXiv:2303.08131. — 2023.

46. A Simple Framework for Open-Vocabulary Segmentation and Detection / H. Zhang, F. Li, X. Zou, S. Liu, C. Li, J. Yang, L. Zhang // Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). — 10.2023. — C. 1020—1031.

47. Aligning Language Models with Preferences through f-divergence Minimization / D. Go, T. Korbak, G. Kruszewski, J. Rozen, N. Ryu, M. Dymetman // International Conference on Machine Learning. — 2023. — URL: https://api.semanticscholar.org/CorpusID:256900692.

48. Direct preference optimization: Your language model is secretly a reward model / R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, C. Finn // arXiv preprint arXiv:2305.18290. — 2023.

49. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection / S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, L. Zhang. — 2023. — arXiv: 2303.05499 [cs.CV].

50. Llama 2: Open foundation and fine-tuned chat models / H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale [и др.] // arXiv preprint arXiv:2307.09288. — 2023.

51. ONE-PEACE: Exploring One General Representation Model Toward Unlimited Modalities / P. Wang, S. Wang, J. Lin, S. Bai, X. Zhou, J. Zhou, X. Wang, C. Zhou // arXiv preprint arXiv:2305.11172. — 2023.

52. Raft: Reward ranked finetuning for generative foundation model alignment / H. Dong, W. Xiong, D. Goyal, R. Pan, S. Diao, J. Zhang, K. Shum, T. Zhang // arXiv preprint arXiv:2304.06767. — 2023.

53. Reinforced self-training (rest) for language modeling / C. Gulcehre, T. L. Paine, S. Srinivasan, K. Konyushkova, L. Weerts, A. Sharma, A. Siddhant, A. Ahern, M. Wang, C. Gu [и др.] // arXiv preprint arXiv:2308.08998. — 2023.

54. Rrhf: Rank responses to align language models with human feedback without tears / Z. Yuan, H. Yuan, C. Tan, W. Wang, S. Huang, F. Huang // arXiv preprint arXiv:2304.05302. — 2023.

55. Segment everything everywhere all at once / X. Zou, J. Yang, H. Zhang, F. Li, L. Li, J. Wang, L. Wang, J. Gao, Y. J. Lee // Advances in neural information processing systems. — 2023. — Т. 36. — С. 19769—19782.

56. Deepseekmath: Pushing the limits of mathematical reasoning in open language models / Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu [и др.] // arXiv preprint arXiv:2402.03300. — 2024.

57. Direct preference optimization: Your language model is secretly a reward model / R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, C. Finn // Advances in Neural Information Processing Systems. — 2024. — T. 36.

58. Enhancing Interactive Image Retrieval With Query Rewriting Using Large Language Models and Vision Language Models / H. Zhu, J.-H. Huang, S. Rudinac, E. Kanoulas // arXiv preprint arXiv:2404.18746. — 2024.

59. Gemma / T. M. Gemma Team, C. Hardin, R. Dadashi, S. Bhupatiraju, L. Sifre, M. Rivière, M. S. Kale, J. Love, P. Tafti, L. Hussenot, et al. — 2024. — DOI: 10.34740/KAGGLE/M/3301.

60. HyperSeg: Towards Universal Visual Segmentation with Large Language Model / C. Wei, Y. Zhong, H. Tan, Y. Liu, Z. Zhao, J. Hu, Y. Yang // arXiv preprint arXiv:2411.17606. — 2024.

61. Improving Text-to-Image Consistency via Automatic Prompt Optimization / O. Mañas, P. Astolfi, M. Hall, C. Ross, J. Urbanek, A. Williams, A. Agrawal, A. Romero-Soriano, M. Drozdzal // ArXiv. — 2024. — T. abs/2403.17804.

62. Instruction Following with Goal-Conditioned Reinforcement Learning in Virtual Environments / Z. Volovikova, A. Skrynnik, P. Kuderov, A. I. Panov. — 2024. — arXiv: 2407.09287 [cs.AI].

63. *Meta AI*. Introducing Meta Llama 3. — 2024. — Accessed: 2025-04-21.

64. Optimizing prompts for text-to-image generation / Y. Hao, Z. Chi, L. Dong, F. Wei // Advances in Neural Information Processing Systems. — 2024. — T. 36.

65. Qwen2.5 technical report / A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei [и др.] // arXiv preprint arXiv:2412.15115. — 2024.

66. Reframing: Detector-Specific Prompt Tuning for Enhancing Open-Vocabulary Object Detection / M. Avshalumov, Z. Volovikova, D. Yudin, A. Panov // International Conference on Hybrid Artificial Intelligence Systems. — Springer. 2024. — С. 128—140.

67. YOLO-World: Real-Time Open-Vocabulary Object Detection / T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, Y. Shan // Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR). — 2024.

68. *Dao A.*, *Vu D. B.* AlphaMaze: Enhancing Large Language Models' Spatial Intelligence via GRPO // arXiv preprint arXiv:2502.14669. — 2025. — URL: https://arxiv.org/abs/2502.14669.

69. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning / D. Guo [и др.] // arXiv preprint arXiv:2501.12948. — 2025. — URL: https://arxiv.org/abs/2501.12948.

70. Yoloe: Real-time seeing anything / A. Wang, L. Liu, H. Chen, Z. Lin, J. Han, G. Ding // arXiv preprint arXiv:2503.07465. — 2025.

71. *Tutor S.* Conservative Fine-Tuning of Sequence Generation Models with KL-control Natasha Jaques, Shixiang Gu, Dzmitry Bahdanau, José Miguel Hernández-Lobato, Richard E // Turner, Douglas Eck arXiv (2016-11-09) https://arxiv.org/abs/1611.02796 v9. —.