Федеральное государственное автономное образовательное учреждение
высшего образования
**«Московский физико-технический институт
(национальный исследовательский университет)»**

Физтех-школа прикладной математики и информатики
Кафедра машинного обучения и цифровой гуманитаристики

Направление: 09.04.01 Информатика и вычислительная техника
Профиль: Анализ данных и разработка информационных систем

МОДЕЛЬ МАШИННОГО ПЕРЕВОДА ХУДОЖЕСТВЕННЫХ ТЕКСТОВ

(магистерская работа)

Студент: Кулиш Дмитрий Александрович

_____
(подпись студента)

Научный руководитель: Воронцов Константин Вячеславович, докт. физ.-мат. наук

_____
(подпись научного руководителя)

Консультант (при наличии): _____
(подпись консультанта)

Москва 2025

Аннотация

В работе создается и анализируется модель машинного перевода художественных текстов. Модель создается с помощью дообучения большой языковой модели Qwen2.5 на специально подготовленном параллельном англйско-русском корпусе художественных текстов (оригинальные тексты на английском языке и их переводы на русский). В результате работы получилось проанализировать влияние дообучения на качество перевода по сравнению с базовой моделью и убедиться в улучшении качества перевода дообученной модели.

Постановка задачи

Данная работа посвящена построению модели машинного обучения, предназначенной для перевода художественных текстов с английского языка на русский, на основе большой языковой модели Qwen2.5[1], дообученной на специально подготовленном корпусе примеров. Художественный текст — это особый тип вербального высказывания, обладающий многоуровневой смысловой структурой и направленный на эстетическое восприятие. Он передаёт содержание как в явной форме, так и через символы, метафоры, ассоциации и культурные отсылки. Для усиления эмоционального и художественного воздействия в нём широко применяются выразительные средства языка — ритм, образность, особая лексика, риторические приёмы и пунктуационные особенности. Художественный текст организован как цельное произведение с внутренней сюжетной логикой (завязка, кульминация, развязка) и построен вокруг действующих персонажей, играющих ключевую роль в раскрытии темы. Он предполагает множественность интерпретаций в зависимости от читательского опыта, культурного контекста и личного восприятия, обеспечивая не только передачу информации, но и создание эстетического впечатления. Критериями выполнения задачи будем считать грамматическую правильность перевода, художественность повествования и сохранение смысла исходного текста[2].

Актуальность темы

Машинный перевод художественных текстов остаётся одной из наиболее сложных задач обработки естественного языка (NLP). Литературные произведения характери-

зуются высокой семантической сложностью, культурными и эмоциональными контекстами, а также лексической и синтаксической нестандартностью, что усложняет как создание таких моделей, так и оценку качества перевода, в процессе перевода могут объединяться предложения, а в некоторых случаях и абзацы, что дополнительно усложняет задачу.

Основные проблемы в области систем перевода художественных текстов:

- Нехватка специализированных корпусов. Долгое время для задачи не существовало крупных document-level датасетов. Появление корпусов GuoFeng Webnovel v1 (WMT-2023) и GuoFeng Webnovel v2 (WMT-2024) частично решает эту проблему, однако корпус содержит только 3 языковые пары (китайский - английский, китайский - русский, китайский - немецкий).

- Длинный контекст. Романы и повести требуют учёта десятков тысяч токенов, поэтому для document-level перевода стандартного "окна"классических трансформеров уже не достаточно, требуются большие языковые модели.

- Оценка качества. Классические автоматические метрики, такие как BLEU, d-BLEU и другие, не всегда релевантны для оценки художественных текстов, так как основаны на простом совпадении n-грамм (и других подобных решениях) и не могут учитывать сложные стилистические особенности текстов. Поэтому, как правило, классические метрики дополняются другими, более адаптированными для оценки художественных переводов.

- Надёжность метрик, основанных на человеческой аннотации. Например, стандартная метрика MQM, зарекомендовавшая себя в других задачах перевода, при оценке студентами-аннотаторами отдаёт предпочтение переводам художественных текстов, выполненным человеком, только в 60 % примеров при сравнении с лучшими моделями. Кроме того, в отличие от автоматических метрик, MQM все еще может быть субъективной, так, аннотации двух оценщиков могут отличаться.

- Развитие больших языковых моделей. Крупные языковые модели (GPT-4o, Llama, Qwen и т.д.) постепенно приближаются к уровню профессионального перевода, однако системы, основанные на этих моделях, все еще имеют ряд нерешенных проблем, связанных с качеством перевода.

- Активизация исследовательского сообщества. Благодаря секции WMT[4] Literary

Translation в 2023 и 2024 годах возникло множество новых работ, посвященных переводу художественных текстов.

- Несмотря на то, что в направлении машинного перевода художественных текстов написано немало работ, а языковая пара английский - русский достаточно хорошо изучена, в контексте перевода художественных текстов этой паре уделено мало внимания.

Таким образом, повышение качества машинного перевода художественных текстов остаётся актуальной и многоплановой задачей, объединяющей проблемы работы с длинным контекстом, небольшого количества специализированных параллельных корпусов и недостатки существующих метрик в условиях достаточно чатого появления новых, все более совершенных больших языковых моделей (с начала 2025 года были опубликованы модели Qwen3, ChatGPT o4-mini-high, o3, GPT-4.1, GPT-4.5 и Llama-4).

1. Литературный обзор

Машинный перевод художественных текстов представляет собой сложную задачу, требующую не только лингвистической точности, но и сохранения стилистической целостности, авторского голоса и культурных особенностей оригинала. В последние годы (2023, 2024 гг.) у этой научной области появилась специализированная секция на конференции WMT - Literary Translation, где публикуются исследования разных авторов по данной тематике, а также происходит сравнение моделей машинного перевода. Чтобы полученная оценка была наиоблее объективной, исследователи используют разные метрики, как базовые, такие как BLEU, так и более продвинутые, наподобие MQM, а также используют специализированные датасеты (например, Guofeng Webnovel) для сравнения результатов обучения.

1.1Секция WMT Literary Translation 2023

Приведем краткий обзор основных моделей, представленных на конференции WMT 2023 в секции Literary Translation[3]. Все модели обучены для перевода на языковой паре китайский - английский и поделены на "ограниченные"и "неограниченнные"в зависимости от того, происходило обучение только на датасете Guofeng Webnovel, предоставленном организиторами секции, или же авторы использовали для обучения дополнительные данные.

1.1.1 MaxLab

Работа[5] команды MAX-ISI на конференции WMT 2023 основана на двух современных архитектурах больших языковых моделей: классический Transformer[6] и модель MEGA[7] (Moving Average Equipped Gated Attention), оптимизированная для более эффективной работы с длинными последовательностями.

Для обучения моделей использовались два датасета: выровненный на уровне предложений и на уровне параграфов. Исходный литературный корпус был предварительно обработан — удалялись примечания переводчиков, объединялись диалоги, устранялись пустые строки. На основе этих данных сформированы две версии датасета: с выравниванием по предложениям и с агрегацией предложений в параграфы, что позволяет моделям использовать более широкий контекст.

В ходе экспериментов сравнивались четыре системы: Transformer и MEGA, обученные на датасетах, выравненных по предложениям и по параграфам. Для оценки качества перевода применялись стандартные метрики BLEU, d-BLEU и BlonDe. Анализ результатов показал, что классический Transformer немного превосходит MEGA по всем основным метрикам, несмотря на то, что теоретически MEGA должна лучше работать с длинным контекстом. sentence-level модели набрали больше баллов по метрикам BLEU и d-BLEU, а paragraph-level — по BlonDe. Авторы объясняют это тем, что обучение на параграфах текста позволяет лучше переводить сложные литературные конструкции.

1.1.2 MAKE-NMT-VIZ

В данной работе[8] для задачи перевода художественных текстов с китайского языка на английский в качестве основной выбрана модель mBART50[9]. Для обучения использовался корпус GuoFeng, предоставленный организаторами секции.

Процесс обучения основной модели заключается в fine-tuning mBART50 на датасете Guofeng. Кроме основной модели, обучены еще две, основанные на работе[10], основная идея которой заключается в обучении модели архитектуры Transformer с учетом относительных позиций предложений в тексте, а также другими небольшими изменениями. Обучение проходит в два этапа: (1) на General Data WMT 2023 с тестированием и валидацией на датасете Guofeng, и (2) дообучение уже полностью на датасете Guofeng на примерах, состоящих из конкатенации трех предложений. Вторая модель аналогична первой, но для нее применяется только (1) этап обучения.

Для оценки качества перевода использовался BLEU и человеческие аннотации.

Анализ результатов показал, что баллы BLEU основной модели и второй модели практически совпадают, в то время как у первой модели результат немного хуже. Авторы предполагают, что это связано с небольшим числом эпох обучения из-за ограниченных ресурсов. Аннотации использовались только для основной модели, тексты оценивались по шкале от 1 до 10, в средней переводы получили 5 баллов из 10, 31% переведенных текстов был без ошибок, 47% сегментов с семантическими ошибками, 11% имеют стилистические ошибки, 5% содержат логические ошибки и 3% грамматические.
Авторы показывают, что, несмотря на то, что им удалось улучшить качество перевода и сократить количество ошибок, перевод художественных текстов все еще сталкивается с рядом нерешённых проблем.

1.1.3 TJUNLP

В статье[11] в качестве базовой архитектуры была выбрана модель Transformer с применением подхода Mixture of Experts (MOE).
Для обучения использовался предоставленный организаторами конкурсный корпус Guofeng.
Процесс построения модели включал два этапа. Сначала была обучена стандартная модель Transformer. После этого параметры обученной модели использовались для инициализации MOE-модели, в которой часть feed-forward-слоев была заменена на MOE-слои. Для повышения разнообразия к инициализируемым параметрам feed-forward-слоев добавлялся шум. Routing module инициализировался случайным образом.
Далее MOE-модель проходила этап обучения без учителя, при котором модель сама предсказывает результат перевода. 35% входных данных маскировалось. Затем модель сравнивала предсказанный текст с оригиналом, вычисляла функцию потерь и обновляла веса.
Для оценки качества модели применялась sacreBLEU.
Результаты экспериментов показали, что MOE-модель превосходит по качеству обычную dense-модель. MOE-модель набрала на тестовых подвыборках 21.59 и 17.89 баллов sacreBLEU, в то время как обычная dense-модель 19.08 и 17.89 баллов соответственно. Обученная dense-модель уже захватывает основные паттерны в данных, и инициализация MOE-модели с её весами позволяет начать обучение с более оптимального состояния, обеспечивая быструю сходимость. MOE-модель, за счёт своей структуры, может выявлять более сложные закономерности в данных.

1.1.4 NTU

В данной работе модель Opus-MT, обученная на датасете OPUS[12], дообучается на корпусе Guofeng.

1.1.5 DLUT

В статье[13] основной идея заключается в использовании большой языковой модели gpt-3.5-turbo без дообучения: вместо fine-tuning авторы разрабатывают и тестируют различные стратегии промптинга и in-context learning для повышения качества перевода на уровне документа.

Для экспериментов использовались данные из предоставленного оргазинаторами датасета Guofeng. Особое внимание уделялось дополнительной обработке данных: удалялись дубликаты, предложения с некорректными или невидимыми символами, пунктуация нормализовалась (скрипты Moses для английского и китайского языков), а китайский текст дополнительно сегментировался с помощью Jieba. Также проводилось преобразование символов из full-width в half-width и перевод традиционных иероглифов в упрощённые.

В ходе исследования разработаны и сопоставлены три prompt для gpt-3.5-turbo. Лучший prompt позволил достичь 28.16 баллов BLEU на тестовой выборке. Кроме того, авторы изучили зависимость качества перевода от длины текстовых сегментов и способа сегментации: оптимальными оказались длина сегмента в 20 предложений (если сегмент короткий, то модель получает слишком мало контекста, если сегмент слишком длинный, то модель не справляется с объемом контекста) и сегментация китайского текста с помощью Jieba.

Для сравнения авторы обучили модель G-transformer с различными стратегиями (обучение с нуля, fine-tuning, дообучение на весах mBART). Максимальный результат для этой архитектуры составил 25.26 BLEU, что ниже, чем у gpt-3.5-turbo с использованием оптимальных промптов.

Таким образом, авторы показали, что грамотная стратегия применения prompt в сочетании с качественной предобработкой текста позволяют большим языковым моделям значительно превосходить специализированные системы перевода без необходимости дополнительного обучения на целевых данных.

1.1.6 HITer-WMT

В данной работе представлены две системы перевода. Основная система основана на instruction fine-tuning и реализована на основе модели Llama-7b. Авторы создали instruction-датасет на основе двух полноценных глав из предоставленного организаторами корпуса.
Вторая система предполагает дообучение модели GuoFeng mBART, предоставленной организаторами секции.

1.1.7 HW-TSC

В работе[14] HW-TSC в качестве базовой модели используется стандартная sentence-level архитектура Transformer. Эта модель ранее была обучена для общей задачи машинного перевода на корпусе General MT.
Для дообучения авторы использовали несколько корпусов. В качестве основного набора данных использовались те же данные, что и для общей задачи машинного перевода (25 млн пар предложений). Для адаптации под задачу и улучшения возможностей перевода длинных текстов используются дополнительные датасеты: GuoFeng Webnovel Corpus, предоставленный организаторами секции, а также корпус, который авторы собрали сами. Он состоит из 100 миллионов предложений на китайском и 400 миллионов предложений на английском, а также 10 миллионов предложений параллельного корпуса. Все данные прошли тщательную очистку и обработку по методике, описанной в предыдущих работах HW-TSC.
Сначала авторы дообучают baseline, в процессе обучения авторы применяют Regularized Dropout, Data Diversification (предсказание переводов в "прямом"и "обратном"направлениях для увеличения размеров датасета), Forward Translation (на монолингвальных данных генерируется перевод, получившийся параллельный корпус включается в датасет), Back Translation (аналогично Forward, но генерируется перевод монолингвальных данных в "противоположном"направлении), Alternated Training (для компенсации шумов в синтетических данных и деградации качества модели реальные данные чередуются со сгенерированными), Curriculum Learning (ранжирование обучающих примеров по сложности).
На следующем этапе модель дообучается на узконаправленном датасете, чтобы адаптировать модель к переводу художественных текстов: для дообучения используются описанные выше данные Guofeng и собранные авторами данные.

После этого авторы используют специальные техники для document-level обучения:

- MR-doc2doc: Модель обучают переводить уже не отдельные предложения, но и фрагменты разной длины — от одного предложения до целого документа. Текст разбивается на куски по несколько предложений, и модель обучают переводить такие сегменты. Это помогает модели лучше улавливать контекст и обеспечивать связность перевода на уровне целых глав, а не только предложений.
- TADA: В исходный текст добавляются специальные теги (метки) на слова/фразы, которые должны быть переведены определённым образом (например, имена или названия). Модель учится копировать эти элементы строго по тегу, обеспечивая согласованность терминологии и именованных сущностей на всём протяжении документа.

Анализ результатов показал, что каждый этап обучения добавляет баллы по метрике sacreBLEU. Внедрение MR-doc2doc и TADA позволило дополнительно повысить согласованность перевода имен собственных: точность по этому критерию выросла с 43.3% для базовой модели до 71.8% при использовании TADA. Таким образом, сочетание адаптации к конкретной задаче и document-level стратегии обучения позволило достичь существенного прогресса в качестве машинного перевода художественных текстов.

1.2 Анализ результатов WMT Literary Translation 2023

В таблице 1 сравниваются все описанные выше системы с помощью автоматических метрик. Среди основных систем, обучение которых было ограничено только корпусом Guofeng Webnovel, система MAKE-NMT-VIZ показывает лучшие значения по всем метрикам. Аналогично, основная система HW-TSC также достигает наилучших результатов среди систем, обучение которых не ограничивалось корпусом, предоставленным авторами.

В большинстве работ основная система показывает лучшие результаты по сравнению с дополнительными. Исключением являются работы HITer-WMT и HW-TSC, где такая закономерность не соблюдается.

Среди baseline систем сервис Google Translate превосходит GPT-4 и Llama-MT по d-BLEU. Примечательно, что и лучшая среди моделей, обученных на датасете организаторов, и две лучшие модели, обученные на неограниченных данных, по d-BLEU превос-

ходят коммерческую систему машинного перевода. В таблице 2 представлено сравнение

Таблица 1: Сравнение результатов моделей с помощью автоматических метрик

Тип	Система	BLEU	chrF	COMET	TER	d-BLEU
Baseline	Llama-MT	n/a	n/a	n/a	n/a	43.1
	GPT-4	n/a	n/a	n/a	n/a	43.7
	Google	37.4	57.0	80.50	57.4	47.3
Основные модели (ограниченный датасет)	MaxLab	34.1	53.3	78.24	62.4	45.0
	MAKE-NMT-VIZ	37.9	56.6	81.50	58.7	48.0
	TJUNLP	32.1	51.9	77.93	64.1	43.3
Основные модели (неограниченный датасет)	DLUT	40.5	58.5	82.58	54.6	50.2
	NTU	32.3	52.5	78.07	64.3	43.4
	HITer-WMT	16.1	37.1	69.84	80.1	28.0
	HW-TSC	44.3	61.1	82.69	51.8	52.2
Дополнительные модели	MaxLab1	34.5	54.7	79.14	62.7	44.9
	MaxLab2	33.1	52.4	77.84	63.6	44.4
	MAKE-NMT-VIZ1	33.8	51.2	76.91	63.5	45.5
	MAKE-NMT-VIZ2	35.0	52.7	77.26	61.5	46.2
	HITer-WMT 1	30.8	49.2	76.41	67.2	40.6
	HW-TSC 1	44.6	61.0	82.67	51.8	52.6
	HW-TSC 2	44.4	61.5	82.63	52.1 5	2.2

baseline и основных моделей с помощью человеческой аннотации с использованием метрики MQM. Модель MAKE-NMT-VIZ превосходит остальные ограниченные по датасету системы, в то время как DLUT занимает первое место среди четырёх систем с неограниченным датасетом. Это не полностью совпадает с результатами автоматической оценки, приведёнными в таблице 1. Две лучшие неограниченные системы превосходят лучшую ограниченную систему, что, как и следовало ожидать, подчеркивает преимущества использования внешних знаний. Это наблюдение согласуется с результатами автоматической оценки.

Среди baseline систем наилучший результат показывает система на основе большой языковой модели, в то время как система машинного перевода Google Translate демон-

стрирует худшие показатели, это отличается от результатов автоматической оценки. При этом модели, дополнительно обученные на литературных данных, показывают результаты, сопоставимые с такими системами, как MaxLab и Google Translate.

При анализе baseline систем GPT-4 показал высокую частоту незначительных ошибок, особенно в категориях "Грамотность"и "Стиль". Llama-MT допускает большое количество серьёзных и критических ошибок в категориях "Точность"и "Терминология". У Google заметно выделяются ошибки категории "Грамотность что может указывать на проблемы с сохранением связности и естественности текста по сравнению с большими языковыми моделями.

В категории систем с ограниченным датасетом анализ MAKE-NMT-VIZ показывает равномерное распределение ошибок с относительно меньшим числом случаев в каждой категории, что свидетельствует о сбалансированности работы системы по разным аспектам перевода. В то же время у MaxLab и TJUNLP наблюдается увеличение числа ошибок в категориях "Точность"и "Грамотность что говорит о сложностях в обеспечении одновременно точности передачи исходного текста и естественности перевода на целевом языке.

При анализе систем с неограниченным датасетом эти системы, а в особенности HW-TSC и DLUT, показывают заметное снижение количества ошибок по категоряим "Точность"и "Грамотность"перевода по сравнению с ограниченными системами. Авторы предполагают,что отсутствие ограничений даёт системам большую гибкость, что приводит к более точным и плавным переводам. Тем не менее, общее распределение ошибок среди разных систем подчёркивает сложные компромиссы и вызовы, присущие машинному переводу, и необходимость дальнейших исследований и оптимизации в этой области.

Основные ограничения работ в секции WMT Literary Tranlsation 2023:

- Языковая пара. Задача была сфокусирована только на направлении китайский - английский.
- Литературный жанр. Корпус в основном состоит из веб-романов, то есть нет разнообразия жанров художественных произведений. Жанр был выбран по двум причинам: во-первых, по мнению авторов, такие художетсвенные произведения проще написаны по сравнению с другими жанрами, следовательно, модели будет проще обучить на перевод веб-романов; во-вторых объём параллельных корпусов для

этого жанра достаточно большой и постоянно увеличивается.

Таблица 2: Сравнение результатов моделей с помощью человеческой аннотации

Тип	Система	MQM	Позиция
Baselines	GPT-4	54.81	1
	Llama-MT	28.40	2
	Google	22.66	3
Основные системы (ограниченный датасет)	MAKE-NMT-VIZ	42.36	1
	MaxLab	28.58	2
	TJUNLP	18.34	3
Основные системы (неограниченный датасет)	DLUT	63.35	1
	HW-TSC	53.01	2
	NTU	31.66	3
	HITer-WMT	5.56	4

1.3 Секция WMT Literary Translation 2024

Перейдем к секции Literary Translation конференции WMT 2024. В отличие от предыдущего года, в датасете появились языковые пары китайский - немецкий и китайский русский, а также авторы убрали разделение моделей на "ограниченные"и "неограниченные". Кроме того, вместо MQM теперь авторы используют для человеческой аннотации другую систему оценивания. Рассмотрим представленные модели:

1.3.1 Cloudsheep

Модель[15] построена на основе Jeiba, CC-CEDICT, GPT-4 и Google Translate. В данном подходе при переводе используется многоступенчатая обработка текста: создается словарь собственных имен, чтобы персонажи переводились одинаково в течение всего текста (для улучшения читаемости), используются Jeiba и GPT-3.5, затем с помощью GPT-3.5 переводятся гоноративы (‘brother’, ‘grandmother’, и т.д.), они также стандартизируются, чтобы во всем тексте персонаж переводился одинаково. В следующем блоке с помощью CC-CEDICT происходит выделение идиом, крылатых выражений в тексте,

они переводятся отдельно, весь остальной текст остается нетронутым, идиомы берутся в кавычки, после них ставится обозначение (idiom), это нужно для дальнейшей обработки. После этого оставшийся текст переводится в Google Translate / DeepL. Далее следует обработка идиом с помощью GPT-4 и специального prompt'а они выстраиваются в текст подходящим образом.

1.3.2 HW-TSC (2024)

В данном эксперименте[16] использовались предложенный датасет и general MT shared task датасет. Основная используемая модель: Chinese-LLaMA2. Обучение состоит из трех этапов:

1. Предобучение на китайских и английских монолингвальных литературных данных. На данном этапе осуществляется адаптация языковой модели общего назначения (LLM) к специализированной литературной модели с использованием монолингвальных литературных данных на китайском и английском языках.

2. Предобучение на выровненных китайско-английских литературных текстах в интерлинейном формате. На этом этапе улучшаются способности модели к перекрестному переводу, для чего используются выровненные литературные тексты в интерлинейном формате, основанные на базе, заложеннйм в первом этапе. Интерлинейный формат текста, где каждое предложение исходного текста напрямую связано с его переводом на уровне слов или фраз, играет ключевую роль в том, чтобы модель могла понять и сопоставить синтаксические и семантические структуры между китайским и английским языками. Это особенно важно для достижения высокого качества перевода. В работе применяется метод постоянного предобучения с использованием LoRA для эффективной адаптации модели к этим интерлинейным текстам.

3. Обучение с учителем с использованием контекстно-осмысленных и стилево-ориентированн инструкций. На заключительном этапе проводится обучение модели с учителем (supervised fine-tuning), для чего используются контекстно-осмысленные и стилево-ориентированные инструкции. Этот этап специально разработан для решения задач, связанных с сохранением семантической согласованности и стилистической целостности при переводе литературных текстов.

В отличие от подхода, основанного на инструкциях, согласованных с исходным язы-

ком (source-language consistent instruction), предлагаемый метод фокусируется на том, чтобы переведенный текст сохранял связность повествования и соответствовал стилистическим нюансам оригинального текста. Это особенно важно для литературного перевода, где сохранение авторского голоса и общего тона произведения так же важно, как и точность перевода.

1.3.3 NLP2CT-UM

Данная модель[17] основана на GPT-4o. Первоначально для каждого документа, поступающего на вход, создается терминологическая таблица, и в исходном документе вся терминология заменяется на терминологию из таблицы, чтобы обеспечить последовательный и согласованный перевод. После этого текст разделяется на части, и переводится с использованием MAPS (три перевода от GPT-4o: один с объяснениями разговорных фраз, один с использованием summary и один просто, выбираем лучший по xComet) и R-BM25 (подбор предложений для контекстуального перевода). После этого GPT-4o корректирует пунктуационные ошибки в переводе.

1.3.4 NTU (2024)

В данной работе[18] фактически можно выделить 4 основных эксперимента:

1. Эксперимент по полному дообучению (SFT) Llama-3-Chinese-8B-Instruct не удался из-за нехватки мощностей GPU;

2. Эксперимент по дообучению Llama-3B с применением метода LoRA на датасете из 10 000 примеров. В результате дообучения получены высокие баллы BLEU и ROUGE;

3. В эксперименте по полному дообучению Phi Chinese на корпусе из 2 000 примеров получены баллы BLEU и ROUGE ниже, чем при дообучении Llama-3B;

4. Эксперимент по полному дообучению Phi3 Chinese 3.5B на 1 500 000 примерах: в результате эксперимента получены баллы BLEU ниже, чем в других экспериментах. В заключении авторы сравнивают дообученные модели и приходят к выводу, что размер модели не всегда обеспечивает лучшую производительность (баллы Phi и Phi3 почти одинаковы, хотя у модели Phi Chinese 3,5 миллиарда параметров, в то время как у Phi3 только 128 тысяч), важно правильно выбирать метод

дообучения, контролировать объем данных для получения качественного перевода, а также учитывать энергоэффективность методов дообучения.

1.3.5 SJTU-LoveFiction

Данная модель[19] основана на GPT-4o, Claude-3.5 и Qwen2-72B. Модель предполагает препроцессинг текстов (убрать предложения без перевода, убрать символы, кроме языка оригинала и языка перевода, стандартизировать пунктуацию). Далее текст разбивается на chunks, каждый из которых представляет из себя несколько предложений, при этом выбирается оптимальный размер chunk. После этого проводится supervised fine tuning для Qwen2-72B на корпусе из конференции. Каждая модель генерирует свой перевод, они объединяются (translation merging) с помощью GPT-4o, которая собирает из трех переводов лучшую версию. В конце выполняется стандартизация терминологии: создание словаря, подбор для него оптимальных терминов, в результате чего и получается итоговый перевод.

1.4 Анализ результатов WMT Literary Translation 2024

В таблице 3 с помощью автоматических метрик сравниваются системы с языковой парой китайский - английский. На уровне предложений система NLP2CT-UM показала наивысшие значения BLEU (41.6) и COMET (83.56), что свидетельствует о высоком качестве перевода, chrF2, отражающая точность и полноту по символным n-граммам, также была высокой — 58.7. Кроме того, NLP2CT-UM достигла наименьшего значения TER (52.7), то есть количество ошибок перевода относительно длины вывода было минимальным.

На уровне документов метрика d-BLEU выделяет NLP2CT-UM с максимальным результатом 50.9 — немного выше, чем у HW-TSC (50.2) и существенно опережая остальные системы.

Основные системы превосходят baseline системы по d-BLEU. Среди базовых систем Google набрал 47.3 балла, но результат системы NLP2CT-UM оказался выше (50.9). Аналогично, HW-TSC также показывает высокий d-BLEU (50.2), демонстрируя больший результат по сравнению с baseline. Таким образом, основные системы лучше справляются с задачей генерации точных и связных переводов на уровне всего документа по сравнению с baseline системами.

В таблице 4 приведены результаты обучения моделей для направлений китайский–немецкий

Таблица 3: Сравнение результатов моделей с помощью автоматических метрик (китайский - английский)

Тип	Система	BLEU	chrF2	COMET	TER	d-BLEU
Baseline	Google	37.4	57.0	80.50	57.4	47.3
	Llama-MT	n/a	n/a	n/a	n/a	43.1
	GPT-4	n/a	n/a	n/a	n/a	43.7
Основные системы	Cloudsheep	39.5	57.5	81.22	55.5	48.5
	HW-TSC	40.5	58.5	82.61	56.0	50.2
	NLP2CT-UM	41.6	58.7	83.56	52.7	50.9
	NTU	20.9	41.9	74.53	73.9	34.6
	SJTU-LoveFiction	35.1	54.7	80.79	62.1	47.2
Дополнительные системы	HW-TSC	40.6	58.6	82.59	55.9	50.3
	NLP2CT-UM1	41.6	58.7	83.54	52.8	50.8
	NLP2CT-UM2	41.5	58.6	83.38	52.8	50.7
	SJTU-LoveFiction1	35.7	56.0	82.67	59.7	46.3
	SJTU-LoveFiction2	38.6	56.5	82.49	57.1	49.6

и китайский–русский. В обоих направлениях baseline-система Google показывает наивысшие значения d-BLEU: 31.3 для китайско-немецкого и 25.2 для китайско-русского, демонстрируя высокое общее качество перевода по этим парам. Среди основных систем NLP2CT-UM набирает 26.7 на паре Zh-De и 22.7 на Zh-Ru, а SJTU-LoveFiction набирает баллы немного ниже: 25.4 и 21.5 соответственно. Эти результаты показывают, что хотя основные системы работают конкурентоспособно, они всё же уступают baseline Google, особенно по согласованности перевода на уровне документа для этих языковых пар. В таблице 5 приведены результаты сравнения моделей с помощью че-

Таблица 4: Сравнение результатов моделей с помощью автоматических метрик (китайский - немецкий и китайский - русский)

Система	Zh-De	Zh-Ru
Google	31.3	25.2
GPT-4	26.7	20.5
NLP2CT-UM	26.7	22.7
SJTU-LoveFiction	25.4	21.5

ловеческой аннотации для направлений китайский - английский. Системы оценивались по двум критериям: общее качество перевода (General) и контекстная согласованность (Discourse), оба параметра оценивались по шкале от 0 до 5.

- NLP2CT-UM показала наивысшие результаты по обоим критериям: 3.96 по общему качеству (1-е место) и 4.00 по контекстной согласованности (разделённое 2-е место), что свидетельствует о высоком общем качестве перевода и хорошей связности.
- SJTU-LoveFiction особенно хорошо проявила себя по контекстной согласованности (4.13 — наивысший результат), а также набрала 3.92 по общему качеству, заняв 2-е место по общему качеству и 1-е место по дискурсу.
- Cloudsheep получила сбалансированные оценки: 3.67 (общее качество) и 3.71 (контекстная согласованность), заняв общее 3-е место.
- HW-TSC набрала 3.54 (общее качество) и 3.46 (контекстная согласованность), заняв 4-е места в обеих категориях.
- NTU показала наименьшие результаты: 2.58 (общее качество) и 2.42 (контекстная согласованность), заняв 5-е места.

Эти результаты выделяют системы NLP2CT-UM и SJTU-LoveFiction как модели, набравшие наибольшие баллы и продемонстрировавшие наилучшее качество перевода и хорошую контекстную согласованность.

Таблица 5: Сравнение результатов моделей с помощью человеческой аннотации

Система	General	Discourse	Место
Cloudsheep	3.67	3.71	3
HW-TSC	3.54	3.46	4
NLP2CT-UM	3.96	4.00	1 / 2
NTU	2.58	2.42	5
SJTU-LoveFiction	3.92	4.13	2 / 1

1.5 Transagents

Отдельного внимания заслуживает Transagents[21]: статья представляет виртуальную multi-agent компанию для перевода длинных литературных текстов. В основе под-

хода лежит имитация традиционного процесса перевода, где различные агенты (CEO, старшие и младшие редакторы, переводчики, специалисты по локализации и корректоры) выполняют специализированные роли. Каждый агент обладает детально описанным профилем, что позволяет смоделировать реальные условия работы переводческой команды. Для координации работы авторы предлагают две ключевые стратегии взаимодействия между агентами:

- Addition-by-Subtraction Collaboration: два агента работают итеративно. Один генерирует максимально подробный перевод (или выполняет другую задачу с максимально подробным response), а второй устраняет избыточную информацию, корректируя и уточняя результат до достижения стабильного варианта.

- Trilateral Collaboration: процесс включает трёх агентов — Action (исполнитель), Critique (критик) и Judgment (судья). Первый создаёт первоначальный вариант перевода, второй высказывает замечания, они работают итеративно, затем третий принимает окончательное решение о качестве перевода. Такой подход позволяет не только генерировать текст, но и систематически улучшать его. Роль «агентов» выполняет GPT-4, у каждого агента есть профиль (вымышленная биография, ключевые навыки и характеристики) и детально составленная инструкция, которой агент должен следовать. Для оценки качества перевода предложены две стратетии:

    1. Monolingual Human Preference (MHP), ориентированная на восприятие носителей целевого языка;
    2. Bilingual LLM Preference (BLP), где LLM сравнивают перевод с оригиналом. Результаты демонстрируют, что, несмотря на более низкие оценки по классическим метрикам (например, d-BLEU), переводы системы TRANSAGENTS предпочитаются как экспертами, так и LLM, особенно в жанрах, требующих глубокого культурного и исторического контекста, а также позволяют существенно снизить затраты по сравнению с традиционными услугами профессиональных персводчиков.

Несмотря на то, что каждое решение задачи машинного перевода художественных текстов, которое мы рассмотрели выше, уникально, можно выделить две основные архитектуры для построения моделей: дообучение LLM на параллельных корпусах текстов

для улучшениях их возможносетй к переводу, а также (в некоторых случаях) на монолигнвальных данных, для того, чтобы переведенный моделью текст как можно больше походил на тексты, написанные человеком, а также использование моделей ChatGPT без дообучения, но с различными улучшениями: prompting, создание словарей имен собственных и других приемов.

Также существуют примеры коммерческих систем, таких как Omni и AINovelTranslation, которые предлагают услуги перевода художественных текстов с возможностью создания своего prompt для перевода, а также редактирования словаря сущностей: пользователь может самостоятельно выбрать, какие имена собственные будут в словаре, а также как они будут переводиться.

2. Анализ используемых метрик

Корректность перевода и сохранение художественного стиля на тестовом датасете будем определять с помощью метрик BLEU (в реализации sacreBLEU) и GEMBA-MQM, а также анализировать другие метрики. Рассмотрим основные метрики, которые используются для оценки перевода в целом и в нашей задаче.

2.1 BLEU

Метрика BLEU[20] была представлена в 2002 году как автоматический способ оценки качества машинного перевода, коррелирующий с оценками людей-переводчиков, но при этом значительно проще в вычислениях и дешевле (нет необходимости нанимать переводчиков). Метод не зависит от языковых пар и позволяет быcро сравнивать системы машинного перевода, а также переводы, выполненные людьми.

Идея заключатеся в том, что для оценки качества машинного перевода есть какой-то перевод (или несколько переводов), который считается эталонным. Предполагается, что хороший перевод должен походить на эталонный текст (или эталонные тексты), поэтому выделяются следующие основные шаги:

- Для каждого n от 1 до N определяется набор n-грамм в эталонных текстах и в машинном переводе;
- Для каждой такой n-граммы определяется максимальное число вхождений в эталонные переводы;

- Если в сгенерированном тексте больше вхождений n-граммы, чем в эталоне, дополнительные вхождения не учитываются. (Логика в том, чтобы штрафовать не только за недостаточные, но и за избыточные вхождения n-грамм).

- Модифицированный precision p_{n} определяется как минимум между отношением n-грамм в эталоне ко всем n-граммам сгенерированного текста и единицей:

$$p_{\text{n}} = min(n_{\text{эталон}}/n_{\text{MT}}, 1)$$

Совпадения n-грамм вычисляются для каждого предложения, но числитель и знаменатель для p_{n} суммируются по всему тестовому корпусу. Это даёт более устойчивую оценку качества.

Также предлагается ввести штраф за краткость: если машинный перевод получился короче эталонного. При этом нужно учитывать, что в коротких предложениях штраф будет сильнее штрафовать за отклонения от длины эталона. Чтобы нивелировать этот эффект, предполагается использовать экспоненциальное сглаживание. Таким образом, штраф за краткость brevity penalty (BP) имеет вид:

$$\text{BP} = \begin{cases} 1, & c > r, \\ e^{1-\frac{r}{c}}, & c \leq r, \end{cases}$$

где c - длина сгенерированного текста, r - длина эталонного текста.

Таким образом, итоговый балл BLEU будет определяться как:

$$BLEU = BP * exp(\sum_{n=1}^{N} w_n \log p_n),$$

где w_n - нормировочный коэффициент, в оригинальной работе $w_n = 1/N$, при $N = 4$ достигнута максимальная корреляция с экспертными (человеческими) оценками.

Преимущество BLEU заключается в простоте, скорости и дешевизне вычислений, воспроизводимости, а также объективности: в то время как более сложные метрики, для которых требуется человеческая оценка, могут зависеть от переводчика-оценщика, BLEU выдает предсказуемый результат.

Недостатки метрики заключается в ее зависимости от эталонных переводов: даже хороший перевод с точки зрения человека-переводчика может получить низкие баллы BLEU, если он сильно отличается от эталонного. Метрика не учитывает напрямую ошибки, связанные с грамотностью, стилем и терминологией переведенного текста. Именно для решения этих проблем в работе используется MQM.

2.2 sacreBLEU

Рассмотрим стандартизированный вариант реализации метрики BLEU, sacreBLEU[23]. Метрика BLEU определяется через подсчет совпадающих n-грамм между машинным переводом и эталонным текстом, при этом существует проблема из-за зависимости реальных реализаций метрики от дополнительных параметров, из-за чего значения BLEU, полученные в разных статьях, не всегда возможно сравнить напрямую. В некоторых случаях различие между баллами BLEU, вычисленными с разными параметриами, превышает заявленный рост баллов. Главная причина такой разницы заключается в различных схемах предобработки машинного перевода и эталонного текста.

Предобработка включает в себя различные модификации текста: нормализацию (объединение знаков препинания, удаление специальных символов), токенизацию (разделение знаков препинания), разбиение составных слов, приведение к нижнему регистру и другие операции. Главная цель предобработки — подготовить для системы осмысленные токены, разделённые пробелами. Наиболее критичной является токенизация, поскольку BLEU — это метрика, основанная на подсчёте точности n-грамм, и изменение токенизации эталона напрямую влияет на набор n-грамм, с которыми сравнивается выход модели. На практике и сгенерированный текст, и эталон всегда токенизируются, различие лишь в том, выполняется ли предобработка эталона самим пользователем или это определено в коде метрики. Сравнивать BLEU-оценки можно только в том случае, если предобработка была идентичной. Пользовательская предобработка приводит к ошибкам и делает сравнение между статьями невозможным.

Для решения этой проблемы автор статьи[23] предлагает свой, стандартизированный вариант метрики BLEU, sacreBLEU. Метрика ожидает на входе детокенизированные данные, применяет внутреннюю предобработку (аналогичную предобработке на конференции WMT), и выдаёт такие же значения, как и метрика BLEU организаторов конференции WMT.

2.3 chrF

Рассмотрим метрику chrF[24]. Идея chrF также заключается в подсчете совпадений n-грамм между сгенерированным переводом и эталонным текстом, однако, в отличие от BLEU, единицей n-граммы в этой метрике является не слово, а один символ. Приведем

формулу для вычисления chrF:

$$chrF\beta = (1 + \beta^2)\frac{chrP * chrR}{\beta^2 chrP + chrR}, \quad (1)$$

где $chrP$ — это доля n-грамм из перевода модели, которые также встречаются в эталонном переводе (character n-gram precision), а $chrR$ — это доля n-грамм из эталонного перевода, которые присутствуют в гипотезе (character n-gram recall).

В оригинальной статье максимальная корреляция с человеческой оценкой достигается при параметрах $n = 6$ и $\beta = 3$, при такой конфигурации модель лучше коррелирует с человеческими суждениями, чем BLEU, в 70-80% примеров перевода.

К плюсам метрики, в сравнении с BLEU, можно отнести независимость от токеназции: в BLEU, для сравнения слов, необходимо сначала токенизировать слова, в то время как в chrF для символов токенизация не нужна. Также плюсом является независимость от языка, метрику можно использовать для любых языковых пар.

С другой стороны, из-за меньшего масштаба сравнения (по символам), метрика может быть менее чувствительной к ошибкам: например, если слово написано с одной ошибкой, BLEU оштрафует сильнее, так как не засчитает совпадение n-грамм. chrF засчитает совпадение n-грамм на правильно написанной части слова и, следовательно, ее штраф будет меньше. Также, аналогично другим автоматическим метрикам, chrF не штрафует напрямую за ошибки, связанные с грамотностью, стилем и за другие ошибки более высокого уровня, чем совпадение n-грамм слов или символов.

2.4 TER

В статье[25] авторы предлагают метрику для оценки качества машинного перевода, основанную на количестве изменений, которые нужно сделать в переведенном тексте, чтобы он был идентичен эталонному тексту (либо одному из нескольких эталонных текстов). Авторы отмечают, что метрика TER более наглядна, чем классический BLEU: для человеческого восприятия интуитивно более понятна метрика, основанная на правках переведенного текста, чем метрика, основанная на совпадении n-грамм.

Метрика TER определяется через минимальное количество правок, необходимых для преобразования сгенерированного перевода в эталонный текст. Количество правок делится на среднюю длину эталонных текстов. Поскольку считается минимальное число правок, переведенный текст сравнивается только с тем эталонным примером, для соответствия которому нужно внести наименьшее количество правок. Формула для под-

счета балла TER:

$$TER = \frac{\text{количество исправлений}}{\text{средняя длина эталонного текста}}. \quad (2)$$

Возможные правки подразумевают вставку, удаление и замену отдельных слов, а также перемещение последовательностей слов. Под таким перемещением подразумевается перенос последовательно стоящих слов внутри сгенерированного текста на другую позицию в рамках одного предложения. При этом все правки, включая перемещения последовательностей любой длины и на любое расстояние, одинаково штрафуют перевод. Знаки препинания учитываются как отдельные слова, ошибки в регистре букв также учитываются, модель штрафуется за правки регистра.

Определение оптимального расстояния перемещения представляет собой NP-полную задачу, в практической реализации авторы предлагают следующий приближенный алгоритм:

- Сначала оценивается минимальное количество вставок, удалений и замен;
- Затем применяется алогоритм жадного поиска: на каждом шаге выбирается такое перемещение, которое максимально сокращает число остальных правок. Перемещения применяются до тех пор, пока они дают выгоду. Если перемещение сокращает число правок только на одну, итоговая штраф не меняется, но перемещение всё равно применяется, так как обычно это улучшает выравнивание с эталонным текстом и может уменьшить число правок на следующих шагах;
- После выполнения всех возможных сдвигов, остаются несоответствия между сгенерированным переводом и эталоном, которые не удалось устранить сдвигами. Оставшиеся отличия устраняются с помощью операций вставок, удалений и замен.

Ограничения на операции перемещения:

- Перемещаемые слова должны полностью совпадать со словами в эталонном тексте;
- Последовательность слов в сгенерированном тексте и соответствующая последовательность в эталоне не должны уже совпадать в своих позициях (иначе перемещение бессмысленно).
- Последовательность слов в эталонном переводе на предполагаемой позиции перемещения тоже должна быть изначально несовпадающей.

После перечисленных выше операций вычисляется балл TER для всех эталонных переводов и выбирается минимальный. TER вычисляет количество правок между машинным переводом и тем эталонным текстом, с которым совпадение максимальное. Наиболее точно этот показатель отражает ошибку перевода, если выбранный эталон действительно максимально близок к сгенерированному переводу. Для самой точной оценки авторы предлагают использовать специально созданные эталоны, которые готовит аннотатор (человек).

В качестве исходных данных используется результат машинного перевода и один или несколько эталонных переводов. Аннотатор может отредактировать как саму гипотезу, так и один из исходных эталонов. Затем для измерения HTER (Human-targeted Translation Edit Rate) используется тот же алгоритм TER, но только по отношению к новому эталону, созданному человеком.

Авторы показали, что TER хорошо коррелирует с человеческими суждениями и автоматическими метриками, такими как BLEU.

К плюсам метрики можно отнести хорошую интерпретируемость в сочетании с корреляцией с другими автоматическими метриками.

Недостатком, аналогично BLEU и другим автоматическим метрикам, является зависимость от эталонного текста, метрика не может учитывать напрямую ошибки в стиле, терминологии и т.д.

2.5 Comet

В статье[26] авторы представляю COMET, метрику для автоматической оценки качества машинного перевода. В отличие от классических метрик, основанных на подсчёте совпадений n-грамм (таких как BLEU или chrF), COMET использует современные нейросетевые модели для анализа перевода на семантическом уровне. Модель строится на основе больших языковых моделей архитектуры Transformer (авторы используют XLM-RoBERTa) и обучается на корпусах данных, основанных на человеческих аннотациях.

Модель поддерживает две разные архитектуры: модель-оцениватель и модель ранжирования переводов. Ключевое отличие между ними — в целевой функции обучения:

- Модель-оцениватель обучается непосредственно предсказывать оценку качества;
- Модель ранжирования минимизирует расстояние между лучшим переводом и соответствующими ей эталоном и исходным текстом.

В процессе оценки COMET учитывает не только сгенерированный перевод и эталон, но и исходное предложение. Благодаря этому подходу метрика способна улавливать более сложные ошибки, связанные с передачей смысла, стилем, терминологией и структурой предложения, а не только формальные совпадения с эталоном.
В статье авторы обучают три модели на трех разных корпусах данных:

1. Корпус, содержащий пост-редактирование для подхода HTER, модель Comet-HTER обучается вносить минимальные правки в машинный перевод.

2. Корпус с ранжированием текстов, модель Comet-RANK учится проставлять баллы для машинных переводов от лучшего к худшему.

3. Корпус, содержащий исходный текст, машинный перевод, эталонный перевод, MQM, модель Comet-MQM учится предсказывать значение MQM.

В оригинальной статье модель COMET-RANK, обученная на данных Direct Assessment/Ranking (DARR), показала наивысшую корреляцию с человеческими оценками по сравнению с остальными вариантами и классическими метриками.
Модель, обученная на MQM, также продемонстрировала высокую корреляцию, но чуть ниже, чем COMET-RANK.
COMET-HTER продемонстрировала хорошую, но более низкую корреляцию по сравнению с двумя предыдущими.
Основные преимущества метрики COMET:

- Высокая корреляция с человеческими оценками: авторы статьи подчеркивают, что COMET показывает более высокое соответствие с экспертной оценкой по сравнению с BLEU, chrF и другими классическими метриками.

- Учет контекста и семантики: благодаря использованию языковых моделей, COMET оценивает качество перевода с учётом смысла, а не только совпадения на уровне символов или слов.

- Универсальность: модель можно дообучить под конкретную задачу, например, для оценки стиля, точности или грамотности перевода.

К недостаткам COMET относятся:

- Необходимость обучения на качественно размеченных данных: работа модели в значительной степени зависит от того, насколько хорошо подготовлены человеческие оценки в обучающем корпусе.

- Зависимость от эталонного перевода: хотя COMET учитывает исходное предложение, модели всё еще нужен эталонный текст, поэтому нестандартные, но адекватные переводы могут получать заниженные оценки — как и в случае с классическими BLEU или TER.

- Как и у большинства нейросетевых подходов, проблемой является эффект “чёрного ящика”. В отличие от классических метрик, при использовании которых результат вычисляется по прозрачной формуле и легко поддаётся ручной проверке, значение балла COMET представляет собой результат работы нейронной сети, и его невозможно воспроизвести вручную. Это затрудняет анализ ошибок и понимание причин получения тех или иных оценок.

Таким образом, COMET позволяет получать оценки, которые лучше отражают восприятие качества перевода человеком, особенно при сравнении современных моделей машинного перевода.

2.6 Multidimensional Quality Metrics

Метрика Multidimensional Quality Metrics (MQM) представляет собой открытую и расширяемую систему для оценки качества перевода, основанную на структурированном словаре типов ошибок. Основное достоинство MQM заключается в её гибкости: она позволяет оценивать качество перевода с различной степенью детализации и адаптировать их под конкретные задачи.

Одной из ключевых особенностей MQM является иерархическая структура типов ошибок. Система делит ошибки на несколько высокоуровневых категорий, таких как точность, грамотность, стиль, терминология, локальные адаптации и прочие. Каждая из этих категорий может быть дополнительно разбита на подкатегории. Например, категория «Точность» включает подтипы, такие как «Добавление», «Неправильный перевод» и «Пропуск», а последний из них может детализироваться до специфичных случаев, таких как «False friend» (использование слова, внешне похожего на ожидаемое) или «Галлюцинация машинного перевода». Такая структурированность позволяет оценивать перевод как на общем уровне, так и на достаточно высоком уровне детализации в зависимости от поставленных целей.

Гибкость MQM проявляется в возможности настройки метрики под различные задачи перевода. При оценке переводов для общего контроля качества может быть ис-

пользована достаточно обобщённая версия метрики, в то время как для диагностики проблем в системах машинного перевода возможно применение более детализированных вариантов. MQM не навязывает единый стандарт, а предоставляет инструментарий для создания метрик, оптимально соответствующих требованиям конкретной ситуации.

Еще одно преимущество системы заключается в стандартизации используемой терминологии. Благодаря единому словарю типов ошибок становится возможным сравнение результатов оценок, выполненных разными экспертами или с использованием различных инструментов. Это способствует повышению объективности и согласованности оценивания, а также упрощает обмен информацией между участниками процесса перевода.

MQM поддерживает два подхода к оценке качества перевода: аналитический и целостный. Аналитический метод предполагает детальное выявление и подсчёт конкретных ошибок в переводе, что позволяет точно определить проблемные аспекты работы переводчика или системы машинного перевода. При этом каждая ошибка ранжируется по степени критичности, что даёт возможность количественно оценить соответствие перевода заданным требованиям. Целостный подход, напротив, основан на более общем оценивании качества перевода как единого целого – например, по таким вопросам, как «Насколько перевод соответствует ожиданиям по точности?» с использованием шкальных значений. Возможность перехода от аналитической к целостной оценке позволяет адаптировать систему под различные сценарии и цели исследования.

Языковая нейтральность MQM – еще один значимый аспект: система разработана таким образом, что её можно применять для оценки качества перевода между любыми языковыми парами. Это делает MQM универсальным инструментом для международных проектов, независимо от специфики языков.

Таким образом, MQM представляет собой мощный инструмент для оценки качества перевода, сочетающий гибкость, масштабируемость и стандартизированный подход. Благодаря иерархической структуре типов ошибок, возможности настройки метрик под конкретные задачи, а также поддержке как аналитического, так и целостного подходов, MQM позволяет эффективно решать задачи контроля качества перевода в самых различных условиях. Эти особенности делают систему MQM востребованной как в академических исследованиях, так и в практических приложениях в сфере перевода и локализации текстов.

В работе предполагается использовать версию метрики MQM, аналогичную представ-

ленной на конференции WMT 2023[29], использовавшейся для оценки качества перевода художественных текстов. Критерии оценки художественных переводов:

- Точность
  - Добавление: перевод включает текст, отсутствующий в исходном тексте.
  - Пропуски: в переводе пропущена часть оригинального текста
  - Неправильный перевод: переведенный текст не соответствует исходному.
  - Неправильное обозначение: перевод более/менее специфичный, чем исходный текст, какие-то детали упускаются, либо наоборот присутствуют детали, которых не было в оригинале.
  - Отсутствие перевода: часть исходного текста осталась непереведенной.
- Грамотность
  - Пунктуация: нет знаков препинания, либо они используются неправильно.
  - Написание: проблемы с правописанием слов (в том числе заглвные буквы, дефисы, звездочки и т.д.)
  - Грамматика: проблемы, связанные с грамматикой или синтаксисом текста, кроме правописания и орфографии. (особенно несоответствие времен и условных предложений)
  - Непоследовательность: текст переведен непоследовательно
- Стиль
  - Несуразность: текст написан неестественным, громоздким стилем, текст трудно воспринимать и понимать, хотя основной смысл понятен.
  - Непоследовательность: стиль непоследователен на протяжении текста.
  - Неидеоматический перевод: текст грамматически правильный, но идиомы неправильно переводятся.
- Терминология
  - Неточный перевод: неправильно переведена жанрово-специфическая или культурно-специфическая терминология.

  - Непоследовательность: терминология в тексте используется непоследовательно.

- Локальные адаптации
  - Локальные форматы: использование неправильных форматов для адреса, имени и т.д.
  - Числовой формат: переведённые дата, время, валюта, телефон оформлена в формате, не соответствующем целевым нормам.

Для количественной оценки и контроля качества будем присваивать числовые значения обнаруженным ошибкам перевода, исходя из типа ошибки, её серьёзности и других факторов, чтобы сделать результаты оценки более наглядными. Общая оценка качества рассчитывается на основе точности перевода по отношению к числу слов:

$$S = 1 - \frac{5 \times C_{\text{Min.}} + 10 \times C_{\text{Maj.}} + 25 \times C_{\text{Cri.}}}{N},$$

определим четыре уровня ошибок: нейтральный $\mathrm{C}_{\text{Neu.}}$, незначительный $\mathrm{C}_{\text{Min.}}$, значительный $\mathrm{C}_{\text{Maj.}}$ и критический $\mathrm{C}_{\text{Cri.}}$, с коэффициентами 0, 5, 10 и 25 соответственно. Общее число слов рассчитывается на основе оригинального текста (на английском языке).

2.7 Gemba-MQM

Рассмотрим метрику Gemba-MQM[27]. Метрика GEMBA-MQM представляет собой современный подход к автоматической оценке качества машинного перевода, основанный на использовании крупных языковых моделей, в частности GPT-4. Основная идея заключается в имитации работы профессиональных аннотаторов, применяющих фреймворк MQM (Multidimensional Quality Metrics) для идентификации и классификации ошибок перевода.

В основе GEMBA-MQM лежит методика, использующая фиксированную схему трёхшотового (three-shot) обучения. Эта схема предусматривает предоставление модели заранее отобранных примеров, где ошибки перевода размечены по категориям (например, точность, беглость, стиль, терминология и т.д.) с учетом их критичности (критические, значимые, незначительные). Такой подход позволяет модели достаточно точно и быстро выявлять проблемные фрагменты в переводе без необходимости использования кон-

трольных (эталонных) переводов, что является существенным преимуществом, особенно для сценариев, где контрольные данные отсутствуют или их получение затруднено.

Ключевым достоинством GEMBA-MQM является его универсальность. Благодаря использованию универсальных, языконезависимых подсказок, методика успешно применяется для широкого спектра языковых пар без дополнительной ручной настройки. Это позволяет избежать традиционных ограничений, когда для каждого нового языкового направления требуется разработка специализированных подсказок или адаптация существующих методик. В экспериментальных исследованиях, проведённых на данных WMT и внутренних тестах Microsoft, GEMBA-MQM показала высокую точность в ранжировании систем машинного перевода, демонстрируя конкурентоспособные, а зачастую и превосходящие показатели по сравнению с классическими метриками, такими как BLEU или COMET.

Однако, несмотря на значительные преимущества, у GEMBA-MQM есть и определённые ограничения. Одним из главных недостатков является зависимость от проприетарной модели GPT-4, что может приводить к проблемам с воспроизводимостью результатов, особенно в условиях возможных обновлений или изменений в работе модели. Кроме того, проведение экспериментов преимущественно на высокоресурсных языках вызывает вопросы о применимости данной методики для языков с ограниченными ресурсами.

Еще одной особенностью, на которую стоит обратить внимание, является вопрос корректного распределения ошибок. В ходе исследований было выявлено, что определённые категории ошибок, например, связанные с локальными адаптациями (locale convention), могут некорректно интерпретироваться моделью. В результате авторы приняли решение исключить этот класс ошибок из финальной версии метрики, что позволило перераспределить ошибки в более релевантные категории и повысить общее качество оценки.

Таким образом, GEMBA-MQM представляет собой инновационное решение для автоматизированной оценки качества перевода, объединяющее возможности современных языковых моделей с проверенной методологией MQM. Его универсальность, способность обходиться без эталонных переводов и высокая точность в системном ранжировании делают его привлекательным инструментом для исследования и практического применения в области машинного перевода, несмотря на некоторые ограничения, связанные с зависимостью от «черного ящика» GPT-4.

К преимуществам метрики также можно отнести наглядность (в отличие и от Comet, и от классических метрик, метрика выводит не просто значение балла, а сами ошибки, которые можно проанализировать и проверить на галлюцинации в работе модели), а также выставление баллов за более высокоуровневые ошибки (нарушения стиля, грамматически ошибки и т.д.), чем просто совпадения n-грамм.

В данной работе предполагается испоользование незначительно измененного prompt: изменятся виды ошибок, чтобы привести их в соответствие с MQM, а также вместо устаревшей GPT-4 будем использовать более релевантные модели от OpenAI: o4-mini-high, o3 и GPT-4o.

3. Датасет, используемый для обучения и тестирования модели.

Проанализируем датасеты, которые используются для обучения и оценки качества перевода художественных текстов в похожих работах: LitEval[28] и GuofengWebnovel[29].

LitEval

Датасет представляет из себя параллельный корпус на уровне абзацев, разработанный для оценки литературного машинного перевода (MT). Он включает в себя проверенные человеческие переводы и машинные переводы 9 различных моделей, что позволяет проводить сравнительный анализ их качества. Содержит классические и современные произведения. Классическая литература часто представляет большие трудности из-за сложного синтаксиса, изменения языка со временем и культурных отсылок, которые могут быть непонятны современным читателям. При этом особую ценность представляют переводы, которые появились до широкого распространения машинного перевода (MT) и, следовательно, были созданы без участия этих систем. В то же время, включение современных произведений помогаст снизить риск контаминации данных в LLM, так как эти модели с большей вероятностью были обучены на широко доступных текстах, включая классическую литературу, и современные произведения, скорее всего, просто не попали в их обучающие датасеты. Перечислим основные характеристики датасета:

- Количество языков: 3 (английский, китайский, немецкий)
- Количество языковых пар: 4 (En-Zh, De-En, En-De, De-Zh)
- Количество абзацев: 2 184

- Количество предложений: 13 346
- Включаст как классические произведения, так и современные
- Модели:
  - Коммерческие модели: Google Translate и DeppL
  - State-of-the art модели: NLLB-3.3b, M2M_100-1.3b
  - Closed-source модель GPT-4o от OpenAI
  - Open-source модели: Llama 3, Qwen 2, Gemma 1.1, Towerinstruct

GuoFeng Webnovel Corpus Датасет GuoFeng Webnovel Corpus представляет из себя общедоступный, высококачественный, многоязычный корпус художественных произведений (web fiction).

Особенности:

- Наличие лингвистических и культурных феномснов: литературные тексты содержат более сложные лингвистические и культурные феномены, чем другие тексты.
- Длинный контекст: литературные произведения, такие как романы, обладают значительно более длинным контекстом по сравнению с текстами из других областей.
- Количество языков: 4 (китайский, русский, английский, немецкий)
- Количество книг в корпусе: 120-180 (в зависимости от языковой пары)
- Количество глав: 19000

Так как в одном из указанных выше датасетов нет пары русский-английский, а в другом тексты и на русском, и на английском языках переведены с китайского (причем на русский тексты были переведны с помощью gpt4o с постредактированием), принято решение собрать свой датасет для обучения модели, который будет состоит из оригинальных текстов на английском языке и их переводов на русский язык, переведенных профессиональными переводчиками, взятых из истоничков в открытом доступе.

Характеристики датасета:

- Количество книг: 200
- Количество глав: 5681

- Языковая пара: английский - русский
- Временной диапазон: 1742 - 1932
- Источники: Project Gutenberg, "Библиотека Максима Мошкова"(lib.ru), Project Gutenberg Canada, Project Gutenberg Australia, Internet Archive
- Количество токенов в датасете (английский): 23 633 131
- Количество токенов в датасете (русский): 35 695 332
- Средняя длина главы (английский): 4 161 токенов
- Средняя длина главы (русский): 6 285 токенов
- Фактическая максимальная длина главы (после обработки): 4 096 токенов

Для создания датасета выбирались произведения, написанные в оригинале на английском языке и переведенные на русский и находящиеся в открытом доступе на перечисленных выше ресурсах.

Перед формированием датасета для обучения произведения предобрабатываются: книги сегментируются на главы, невалидные главы (главы, которые есть только на одном из языков, либо разбиение на главы в оригинале и переводе не совпадает, либо другие недочеты) удаляются. Также из текстов удаляются все символы, кроме символов латиницы, кириллицы, знаков препинания и цифр, нумерация страниц, комментарии автора, примечания, ссылки и т.п. При выравнивании глав применяется ручная проверка соответствия текста и перевода.

На рисунке 1 изображено распределение длин текстов обучающего датасета до применения обрезки с ограничением по количеству токенов.

Если токенизированный текст длиннее, чем 4096 токенов токенизатора Qwen2.5, то он обрезается до этой длины. Текст разбивается на предложения с помощью библиотеки nltk[31] и подбирается количество предложений, при котором выполняется условие на ограничение по токенам. Алгоритм применяется и к русскому, и а английскому текстам. После этого применяется метод vecalign[30], который позволяет выровнять тексты так, чтобы последнее предложение на английском соответствовало последнему на русском.

Разбиение одной главы на несколько сегментов по 4096 токенов не применяется, чтобы избежать генерации некачественных примеров: хотя vecalign позволяет выравнивать

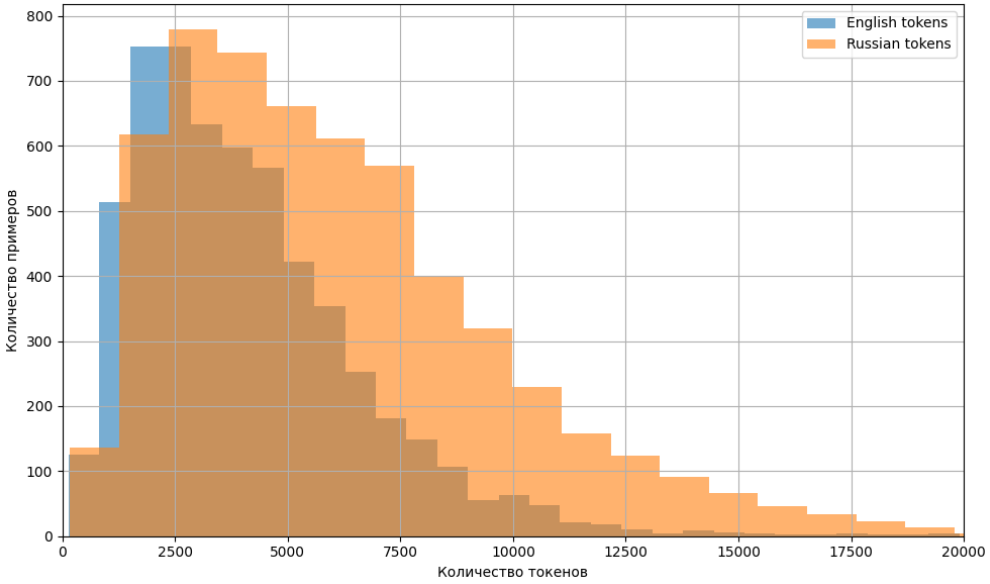

Рисунок 1: График распределения длин текстов обучающего датасета

тексты по предложениям, для того, чтобы гарантировать корректную генерацию примеров необходима достаточно трудозатратная ручная проверка, так как в этом случае метод будет определять и начало, и конец примера, в случае длинных текстов несколько раз, что может привести к накоплению ошибок. При этом нужно отметить, что ограничение по токенам зависит от конкретной модели Qwen, у других моделей из-за разницы работы токенизаторов может получаться больше или меньше токенов, поэтому ручная проверка для конкретного токенизатора иррациональна. Как можно увидеть из рисунка 1, примерно 50% текстов на английском и 35% текстов на русском короче 4096 токенов до применения обрезки (то есть обрезка не требуется). Разница в длине токенизированных текстов объясняется особенностями работы токенизатора Qwen: всего в датасете 44 392 уникальных токена для английских текстов и только 10 082 уникальных токена для русских текстов, так как модель первоначально более подготовлена для работы с английским языком, чем с русским.

На рисунке 2 приведено распределение книг по количеству глав. В среднем в одной книге содержится 28,5 глав, однако большинство книг укладывается в диапазон 6–40 глав. Наиболее часто встречаются книги с 6–15 главами. Длинный "хвост"распределения свидетельствует о наличии небольшого числа книг с необычно большим количеством глав (в том числе свыше 100), однако основная масса произведений существенно короче. Такой разброс говорит о разнообразии структуры литературных произведений в корпусе, так как в датасет для разнообразия примеров включены как длинные романы, так и небольшие рассказы.

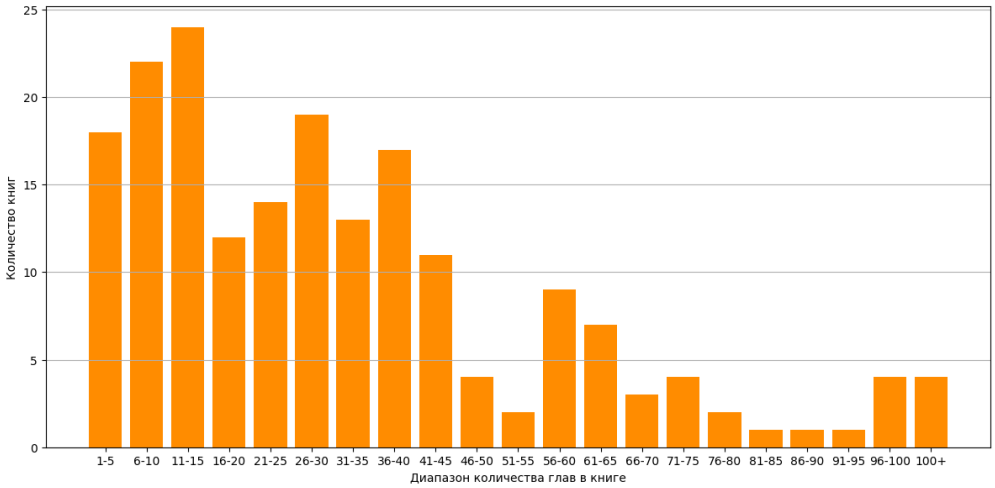

Рисунок 2: График распределения книг по количеству глав

Поскольку данные взяты из открытых источников, в процессе формирования датасета возникли некоторые ограничения: так, в датасете недостаточно представлены произведения второй половины XX века и полностью отсутствуют произведения начала XXI века: как правило, они реже находятся в открытом доступе, чем более ранние произведения. Информацию о распределении книг по десятилетиям можно увидеть на рисунке 3: большая часть книг находится во временном промежутке от начала XIX века до 30-х годов XX века. На этом промежутке книги распределены практически равномерно, за исключенем небольшого спада в 30-х годах XIX века и увеличения в начале XX века. Алгоритм поиска книг состоит из проверки наличия перевода книги на русский язык, и поиска соответствующего оригинала (как правило, если перевод книги есть в открытом доступе, за редким исключением оригинал тоже доступен, в то же время очевидно, что далеко не все произведения, написанные на английском, переведены на русский язык). Небольшое количество произведений в начале XVIII века и их полное отсутствие во второй его половине объясняется небольшим количеством переведенных книг этого периода в открытом доступе.

Также можно отметить другой недостаток подготовленного датасета, дисбаланс в количестве произведений между разными авторами: число произведений отдельных авторов может доходить до 10-15, в то время как часть из них представлена только одним-двумя произведениями. Однако этот дисбаланс вынужденный: если исключить часть авторов, датасет станет еще менее разнообразным, а если у каждого автора оставить по 1-2 произведения, значительно сократится объем датасета для дообучения.

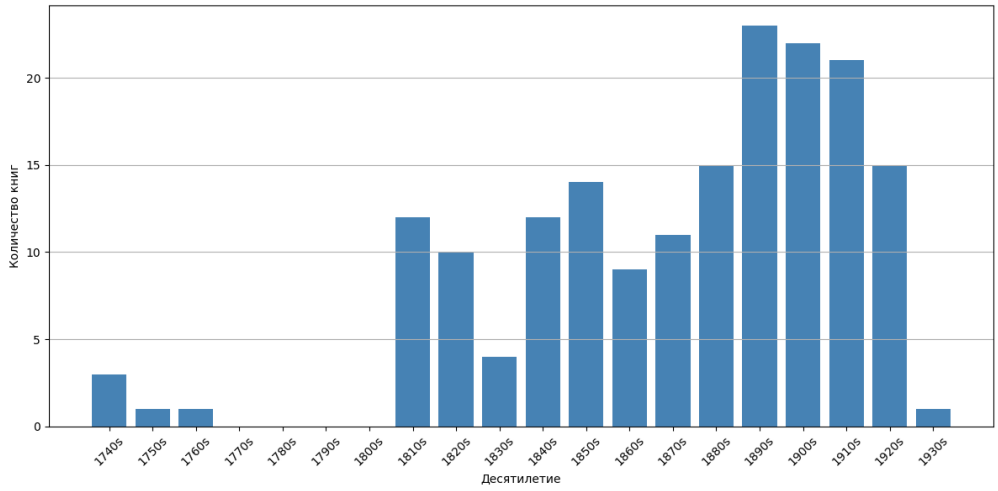

Рисунок 3: График распределения книг по десятилетиям

4. Обзор модели

Перейдем к обзору модели, которая будет использоваться как основа для дообучения в этой работе, начнем с рассмотрения классической архитектуры Transformer.

4.1 Архитектура Transformer

Большинство моделей преобразования последовательностей используют архитектуру «Encoder–Decoder». Encoder сопоставляет входной последовательности символов последовательность векторов. Decoder, на основе этой последовательности, генерирует выходную последовательность. На каждом шаге модель использует ранее сгенерированные символы в качестве дополнительного входа при создании следующего.
Архитектура Transformer[6] следует этой общей схеме, но как в блоке Encoder, так и в Decoder вместо рекуррентных или сверточных сетей применяет механизмы Multi-Head Self-Attention и покомпонентные полносвязные слои Feed-Forward Network.
Encoder.
В оригинальной статье Encoder состоит из $N = 6$ одинаковых слоёв. Каждый слой включает два подслоя:

1. Механизм Multi-Head Self-Attention;
2. Feed-Forward Network.

К выходу из каждого подслоя применяется механизм остаточной связи и нормализация по слоям. Таким образом, output каждого подслоя имеет вид $LayerNorm(x +$

$Sublayer(x)$, где $Sublayer(x)$ - output самого подслоя. Чтобы механизм остаточной связи работал корректно, output всех подслоев в оригинальной статье имеют размерность $d = 512$.

Decoder.

Decoder в оригинальной статье также состоит из $N = 6$ одинаковых слоёв, но каждый слой содержит уже три подслоя:

1. Механизм Masked Multi-Head Self-Attention. В Decoder нужно генерировать последовательность по одному токену за раз и не "смотреть"на еще не сгенерированные токены. Для этого применяется сдвиг эмбеддингов на одну позицию, а также все "запрещенные"элементы затираются (устанавливаются в значение $-\infty$;

2. Дополнительный подслой с механизмом Multi-Head Attention, но на вход подается output Encoder;

3. Feed-Forward Network.

Аналогично Encoder, к выходу каждого подслоя Decoder применяется механизм остаточной связи и нормализация по слоям.

Attention

Функция Attention описывается как отображение запроса и множества пар «ключ–значение» в выходной вектор: при этом запрос, ключи, значения и сам выход - вектора. Output вычисляется как взвешенная сумма значений, где вес каждого значения определяется функцией совместимости (compatibility) между запросом и соответствующим ключом.

Scaled Dot-Product Attention

В качестве input поступают запросы и ключи размерности d_k и значения размерности d_v. Вычисляются скалярные произведения между запросом и всеми ключами, нормируются на $\sqrt{d_k}$, применяется softmax и получаем весы для всех значений.

На практике функция внимания вычисляется на множестве запросов, сгруппированных в матрицу Q, ключи и значения также формируют матрицы K и V соответственно. Матрица attention принимает вид: $Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$. Авторы применили этот вид внимания, а не additive attention[32], так как, несмотря на одинаковую теоретическую сложность, этот вариант быстрее и более эффективнее на практике. При малых значениях d_k механизмы работают схоже, чтобы нивелировать разницу (алгоритм уступает additive attention) при больших значениях d_k, применяется нормирование на $\frac{1}{d_k}$.

Механизм Multi-head attention обрабатывает информацию через несколько «голов», каждая из которых работает в собственном подпространстве представлений, и выполняет над ними функцию внимания параллельно, после чего объединяет результаты и проецирует обратно в исходное пространство. Это позволяет модели одновременно улавливать разные типы взаимосвязей между элементами последовательности, в то время как одна голова, усредняя всё в одном представлении, не позволяет извлекать такое количество информации. $MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^o$, где $head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$, где проекции это матрицы параметров $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$ и $W^O \in \mathbb{R}^{hd_v \times d_{model}}$.

Вместо вычисления функции внимания единожды с запросами, ключами и значениями размерности d_{model}, авторы предлагают линейно проецировать запросы, ключи и значения h раз с разными обучаемыми линейными проекциями с размерностями d_k, d_k, d_v соответственно. К каждой из этих проекций параллльельно применяется функция внимания с output размерности d_v. После этого они объединяются и проецируются в финальные значения.

В оригинальной статье авторы использовали $h = 8$ параллельных слоев, для каждого из них используется $d_k = d_v = d_{model}/h = 64$. Благодаря снижению размерности в каждой голове, авторы добились стоимости вычислений, сопоставимой со стоимостью вычислений для одной головы, но с полной размерностью.

Применение механизма внимания в Transformer:

1. Encoder–Decoder attention: запросы берутся из предыдущего слоя Decoder, а ключи и значения — из output Encoder. Это позволяет Decoder «смотреть» на все позиции входной последовательности, как в классических Seq2Seq-моделях.

2. Self-attention в Encoder: все запросы, ключи и значения поступают из output предыдущего слоя Encoder, что позволяет каждой позиции "смотреть"на все позиции в предыдущем слое Encoder.

3. Self-attention в Decoder: аналогично, но с маскированием будущих позиций (веса соответствующих связей перед вычислением softmax устанавливаются в $-\infty$), чтобы сохранить авторегрессию и не допускать утечки информации о будущих токенах.

В дополнение к подслоям внимания, каждый слой Encoder и Decoder содержит полносвязную FFN, которая применяется к каждой позиции независимо и идентично. Она

состоит из двух линейных преобразований и функции ReLU между ними: $FFN(x) = max(0, xW_1 + b_1)W_2 + b_2$. Хотя линейные преобразования идентичны для разных позиций, их параметры отличаются от слоя к слою.

Эмбеддинги и softmax

Аналогично другим моделям преобразования последовательностей, авторы используют обучаемые эмбеддинги, чтобы преобразовать входные и выходные токены в векторы размерности d_{model}. Для преобразования выхода Decoder в вероятности следующих токенов также используется стандартное обучаемое линейное преобразование и функция softmax. В модели одна и та же матрицу весов используется двумя слоями эмбеддингов и пред-softmax линейным преобразованием. В слоях эмбеддинга эти веса умножаются на $\sqrt{d_{\text{model}}}$.

Поскольку в архитектуре Transformer нет рекуррентных и свёрточных слоёв, чтобы включить в неё информацию о порядке токенов, следует добавить информацию об абсолютных или относительных позициях в последовательности. Для этого в нижней части стеков Encoder и Decoder авторы добавили «позиционные кодировки» к эмбеддингам с той же размерностью d_{model}, что и эмбеддинги, чтобы их можно было просто сложить. В оригинальной статье используются синусоидальные функции разной частоты: $PE_{pos,2i} = sin(pos/10000^{2i/d_{model}})$, $PE_{pos,2i+1} = cos(pos/10000^{2i/d_{model}})$ где pos — индекс позиции, i — номер измерения. То есть каждое измерение соответствует одной синусоиде, а длины волн образуют геометрическую прогрессию от 2π до $10000 \cdot 2\pi$. Авторы выбрали этот вариант исходя из предположения, что модель сможет легко учиться учитывать относительный сдвиг позиций: для любого фиксированного смещения k PE_{pos+k} выражается как линейная функция от PE_{pos}.

4.2 Grouped-query attention

Рассмотрим механизм Grouped-Query Attention (GQA)[34], применяемый в Qwen2.5. GQA делит все головы запросов (query heads) на G групп, в каждой из которых используется по одной общей голове ключей (key head) и одной общей голове значений (value head). Обозначение GQA-G означает использование G групп.

- GQA-1 (одна группа) эквивалентна классическому MQA (одна пара Key/Value-голов на все запросы);
- GQA-H (число групп равно общему числу голов H) точно соответствует MHA (каждая Query-голова имеет свою пару Key/Value).

При конвертации MHA-чекпоинта в GQA-чекпоинт для каждой группы авторы усредняют (mean pooling) проекционные матрицы всех оригинальных голов key и value, входящих в эту группу, чтобы получить по одной паре key/value на каждую группу.

Промежуточное число групп ($1 < G < H$) даёт модель, качество которой выше, чем у MQA, но скорость работы почти не уступает MQA. Переход от MHA к MQA уменьшает количество голов key/value с H до одной, сокращая размер кэша ключей и значений и объём загружаемых данных в H раз. При этом у крупных моделей число голов обычно растёт, и MQA обеспечивает всё более резкое снижение нагрузки на память и ёмкость кэша. GQA позволяет сохранять ту же пропорциональную экономию пропускной способности и памяти по мере увеличения числа голов.

Кроме того, у крупных моделей относительная доля пропускной способности, занятой attention-кэшем, снижается: размер KV-кэша растёт линейно с размерностью модели, тогда как FLOPs и количество параметров — квадратично. И наконец, при стандартном шардинге (разбиении модели на P партиций) одна голова key/value реплицируется P раз, что создаёт избыточность; GQA устраняет эту избыточность. Поэтому GQA особенно выгоден для больших моделей.

GQA не применяется к слоям self-attention в Encoder: представления Encoder вычисляются параллельно, и пропускная способность памяти там обычно не является узким местом.

4.3 Rotary Position Embedding

В Qwen2.5 авторы используют Rotary Position Embedding[35]. Модели на базе Transformer обычно используют позиционную информацию отдельных токенов через механизм self-attention. Скалярное произведение $q_m^T k_n$ обычно позволяет передавать информацию между токенами на разных позициях. Чтобы встроить относительную информацию о позициях, скалярное произведение запросов q_m и ключей k_n должно задаваться некоторой функцией g, принимающей на вход только эмбеддинги слов x_m, x_n и их относительную разницу $m - n$. Иными словами, позиционная информация должна кодироваться только в виде разности индексов, а не абсолютных положений:

$$\langle f_q(x_m, m), f_k(x_n, n) \rangle = g(x_m, x_n, m - n) \tag{3}$$

Нужно найти такие функции $f_q(x_m, m)$ и $f_k(x_n, n)$, которые удовлетворяют этому условию.

Как показано в статье[35], в простейшем, двумерном случае, такие функции сводятся к

повороту на угол, соответствующий их индексу. В общем случае такое решение имеет вид:

$$f_{\{q,k\}}(x_m, m) = R^d_{\Theta,m} W_{\{q,k\}} x_m, \tag{4}$$

где

$$R^d_{\Theta,m} = \begin{pmatrix} \cos(m\theta_1) & -\sin(m\theta_1) & 0 & \cdots & 0 \\ \sin(m\theta_1) & \cos(m\theta_1) & 0 & \cdots & 0 \\ 0 & 0 & \cos(m\theta_2) & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \cos(m\theta_{d/2}) \end{pmatrix} \tag{5}$$

матрица вращения с предопределенными параметрами $\Theta = \{\theta_i = 10000^{-2(i-1)/d}, i \in [1, 2, ..., d/2]\}$. После применения RoPE к функции внимания:

$$q_m^T k_n = (R^d_{\Theta,m} W_q x_m)^T (R^d_{\Theta,n} W_k x_n) = x^T W_q R^d_{\Theta,n-m} W_k x_n, \tag{6}$$

где $R^d_{\Theta,n-m} = (R^d_{\Theta,m})^T R^d_{\Theta,n}$. Итоговая формула внимания с RoPE принимает вид:

$$Attention(Q, K, V)_m = \frac{\sum_{n=1}^N (R^d_{\theta,m} \phi(q_m))^T (R^d_{\theta,n} \varphi(k_n)) v_n}{\sum_{n=1}^N \phi(q_m)^T \varphi(k_n)}, \tag{7}$$

где $\varphi(*)$, $\phi(*)$ - неотрицательные функции.

4.4 Pre-RMSNorm

В отличие от оригинальной статьи Transformers, в Qwen2.5 используется не $LayerNorm$ нормализация, а Pre-RMSNorm[36]: $RMSNorm(x) = \frac{x}{\sqrt{\|x\|_2^2/d + \epsilon}}$, где $\epsilon > 0$, приставка "Pre"означает, что вместо формулы $LayerNorm(x + Sublayer(x))$ применяется $x + Sublayer(RSMNorm(x))$. Нормализация $RMSNorm$ демонстрирует лучшую вычислительную эффективность по сравнению с $LayerNorm$, но при этом может снижать выразительные возможности модели. В статье[36] авторы формально доказывают эквивалентность нормализаций Pre-LayerNorm и Pre-RMSNorm для обучения и inference. В экспериментах время обучения и вывода с нормализацией Pre-RMSNorm сокращается по сравнению с Pre-LayerNorm на 1-10%.

4.5 Qwen2.5

В качестве основы для fine-tuning в работе используется Qwen2.5[1] - серия больших языковых моделей для различных задач. В открытом доступе находятся базовые

и инструкционно-дообученные модели размерами 0.5, 1.5, 3, 7, 14, 32 и 72 миллиардов параметров. Модели семейства Qwen2.5 показывают хорошие результаты на большом наборе бенчмарков по пониманию языка, рассуждению, математике, программированию и выравниванию с человеческими предпочтениями. Флагманская модель Qwen2.5-72B-Instruct превосходит множество открытых и закрытых моделей и сопоставима по качеству с Llama-3-405B-Instruct, которая в пять раз больше по размеру.
Модель доработана по сравнению с предыдущей версией Qwen2, а именно:

- Измененные размеры моделей. По сравнению с Qwen2, помимо моделей 0,5 B, 1,5 B, 7 B и 72 B, в Qwen2.5 возвращены модели 3 B, 14 B и 32 B — более экономичные варианты для ограниченных по ресурсам сценариев.
- Увеличенный объем данных. Объём и качество данных для предобучения и дообучения улучшены: набор данных предобучения увеличен с 7 триллионов до 18 триллионов токенов.
- Устранены ключевые ограничения Qwen2: увеличена длина генерации (с 2 000 до 8 000 токенов), улучшена поддержка структурированного ввода и вывода (таблицы, JSON), упростилось использование инструментов.

Для dense-моделей сохранена классическая архитектура декодера Transformer, аналогичная Qwen2. Отличия от архитектуры Transformer:

- Grouped Query Attention (GQA) для более эффективного использования кеша ключей/значений;
- SwiGLU в качестве функции нелинейной активации;
- Rotary Positional Embeddings (RoPE) для кодирования информации о позициях;
- QKV-смещение внутри механизма внимания;
- RMSNorm с преднормализацией для стабильности обучения.

Для токенизации используется собственный токенизатор Qwen, основанный на byte-level byte-pair encoding (BBPE) со словарём из 151 643 обычных токенов. По сравнению с предыдущими версиями расширено число служебных (control) токенов с 3 до 22: два новых выделены на работу с внешними инструментами, остальные — под другие возможности модели. Это позволяет унифицировать словарь для всех моделей Qwen2.5,

повысив совместимость и согласованность. Некоторые преимущества Qwen2.5 для нашей работы:

- В процессе предобучения применяется этап предобучения для работы с длинным контекстом, что повышает способность модели обрабатывать и понимать длинные последовательности. Это улучшение особенно актуально для нашей задачи.

- Кроме того, благодаря дообучению с помощью Reinforcment Learning, ориентированному на правдоподобность и лаконичность ответа, переводы еще до нашего дообучения приближены к человеческому стилю и требуют меньше изменений.

- В процессе дообучения с учителем используется большой датасет, включающий миллионы высококачественных примеров. Расширение датасета по сравнению с Qwen2 нацелено на устранение проблем предыдущей модели в ключевых областях, таких как генерация длинных последовательностей, решение математических задач, программирование, следование инструкциям, понимание структурированных данных, логическое рассуждение, перенос знаний между языками и надёжное выполнение системных команд. Для нашей работы, очевидно, плюсами являются улучшения, связанные с работой с разными языками и с генерацией длинных последовательностей.

- Qwen 2.5-3B-Instruct распространяется под специализированной исследовательской лицензией Qwen Research License, которая разрешает ее применение, модификацию и дообучение в рамках любых научно-исследовательских и учебных проектов без каких-либо дополнительных разрешений или уплаты лицензионных отчислений.

5. Методы, используемые в работе

5.1 LoRA

Так как в работе проводится дообучение большой языковой модели (на 3 миллиарда параметров), полный fine-tuning в условиях ограниченных ресурсов неэффективен с точки зрения времени работы GPU. Для оптимизации времени дообучения в работе используется метод LoRA[37] (LOW-RANK ADAPTATION OF LARGE LANGUAGE MODELS).

В методе LoRA веса предобученной модели замораживаются, и в каждый слой архитектуры Transformer внедряются обучаемые матрицы низкорангового разложения. Это значительно сокращает число параметров для дообучения на конкретные задачи. Рассмотрим метод подробнее.

Нейросети содержат множество dense-слоёв, в которых происходят матричные умножения. Матрицы весов обычно имеют полный ранг. Однако, как показано в работе[38], предобученные языковые модели обладают низкой intrinsic dimension и могут эффективно учиться даже при проекции в гораздо меньшее подпространство. Основываясь на этом авторы предполагают, что и обновления весов при адаптации тоже имеют низкий «внутренний ранг».

Пусть есть предобученная матрица весов $W_0 \in \mathbb{R}^{d\times k}$. Ее обновление ограничивается низкоранговым разложением: $W_0 + \Delta W = W_0 + BA$, где $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$, а ранг $r \ll \min(d, k)$.

Во время обучения W_0 заморожена и не получает обновлений весов, а веса A и B обучаются. При этом и W_0, и $\Delta W = BA$ умножаются на один и тот же input, а их выходы складываются поэлементно. То есть, если $h = W_0 x$, то $h = W_0 x + \Delta W x = W_0 x + BAx$.

При инициализации матрица A заполняется случайными значениями по Гауссовому распределению, матрица B заполняется нулями, поэтому изначально $\Delta W = 0$. Затем $\Delta W x$ масштабируется на α/r, где α — константа. В процессе оптимизации Adam это аналогично подстройке learning rate, поэтому α фиксируется при первом выборе ранга r. Это избавляет от необходимости постоянной настройки гиперпараметров при изменении r.

Более общий вариант fine-tuning подразумевает обучение только части параметров предобученной модели. В методе LoRA обновление матриц весов при адаптации не обязательно должно быть полного ранга. То есть, если применить LoRA ко всем матрицам весов и обучать все смещения, можно почти полностью восстановить выразительную мощность полного fine-tuning, если выставить ранг LoRA равным рангу матриц весов. Таким образом, с увеличением числа обучаемых параметров LoRA стремится к обычному обучению всей модели.

В production можно явно вычислить и сохранить $W = W_0 + BA$, и выполнять inference как обычно. Если необходимо сменить задачу, достаточно вычесть BA и добавить новые $B'A'$, такая операция требует минимальных затрат памяти и выполняется очень быстро. Это гарантирует отсутствие какой-либо дополнительной задержки на inference по сравнению с полностью дообученной моделью.

Теоретически LoRA можно применить к любому подмножеству весов в нейросети для сокращения числа обучаемых параметров. В архитектуре Transformer есть четыре матрицы весов в модуле self-attention (Wq, Wk, Wv, Wo) и две — в MLP-модуле. Обычно Wq (или Wk, Wv) — это матрица размера $d_{model} \times d_{model}$, хотя на практике выход разбивается по attention heads.

В оригинальной работе авторы ограничиваются адаптацией только весов attention для конкретных задач и оставляют MLP-модули "замороженными это оптимальнее с точки зрения количества параметров.

Главное преимущество метода заключается в резком сокращении потребления памяти GPU и оптимизации хранения весов моделей. Например, для большого Transformer с Adam расход VRAM снижается до 1/3, если $r \ll d_{model}$, так как нет необходимости хранить optimizer state для замороженных параметров.

Другое преимущество LoRA состоит в том, что при развертывании модели можно быстро переключаться между задачами, меняя только веса LoRA, а не всю модель. Это удобно для хранения большого числа "custom"моделей, которые можно быстро подгружать, если основные веса модели находятся в VRAM.

В обучении на GPT-3 175B авторы наблюдают ускорение до 25% по сравнению с полным fine-tuning, так как не нужно считать градиенты для основной массы параметров.

Ограничение метода может заключаться в том, что для максимальной скорости inference обычно объединяют веса заранее, но тогда можно обрабатывать только одну задачу за раз. Если нужна многозадачность, то нужно хранить веса LoRA отдельно и добавлять разные A и B для разных задач, но тогда inference будет немного медленнее.

5.2 Prompt для Gemba-MQM

Так как в первоначальном варианте prompt не предполагает использование ChatGPT для оценки художественных текстов, в него были внедрены изменения. Модифицированы сами типы ошибок и приведены в соответсвие с теми, по которым оценивались модели на секции Literary Translation WMT 2023.

При тестировании применялись разные технологии, от zero-shot prompting (отсутствие примеров в запросе) до few-shot prompting (несколько примеров, состоящих из входных данных в определенном формате и ожидаемого формата ответа модели).

Для определения оптимального prompt результат вручную валидировался, чтобы убедиться, что модель не галлюцинирует. Также различные варианты prompt сравни-

вались на устойчивость генерации: на вход подавался один и тот же пример 5 раз, для оценки воспроизводимости результатов: если разброс оценок от итерации к итерации сравним с различиями между системами, то сама метрика теряет объективность как инструмент точного сравнения, поэтому нужен prompt, который генерирует оценку с максимальной воспроизводимостью. Результаты сравнения находятся в таблице 6. В варианте two-shot prompt № 4 демонстрирует лучшие результаты: в 3 из 5 тестов ошибки, на основе которых вычисляется балл MQM (а следовательно и сам балл MQM), не меняются, а в четвертом меняется незначительно, в то время как у других запросов количество ошибок может меняться от запуска к запуску. Prompt № 2 самый длинный, с дополнительными инструкциями ("You are a professional literary translation editor specializing in English–Russian translation and MQM annotation. Be extremely strict and precise in your evaluation. "Double-check your output for missed critical errors or mismatches between source and translation before answering."), однако результаты все еще хуже prompt 4. Поскольку этот prompt самый длинный, для проверки гипотезы, что этот prompt вместе с двумя примерами не помещается в контекстное окно ChatGPT, он был протестирован в режиме one-shot, однако это только понизило стабильность работы prompt и показало, что контекстного окна у модели достаточно. Также prompt 4 был протестирован в режимах zero-shot и one-shot, результаты показали, что лучшим по воспроизводимости остается вариант two-shot. Поэтому принято решение использовать для оценки MQM именно prompt 4. Этот запрос (prompt) находится в Приложении 2. Также можно отметить, что это самый короткий prompt из всех тестируемых: как показали эксперименты, чем prompt проще, тем его результаты более предсказуемы.

6. Эксперимент

Основной эксперимент состоит в дообучении большой языковой модели Qwen2.5-3B-Instruct на собранном датасете с помощью метода LoRa, на одном GPU Nvidia A100, обучение заняло 5 часов, один раз в эпоху фиксировались баллы BLEU на тестовом датасете.

Параметры эксперимента:

- Число эпох: 3
- batch size: 7
- cutoff len: 4096

Таблица 6: Сравнение разных prompt в тесте на воспроизводимость результатов

№	Примеры	Тест 1			Тест 2			Тест 3			Тест 4			Тест 5		
		Cr	Mj	Min	Cr	Mj	Min	Cr	Mj	Min	Cr	Mj	Min	Cr	Mj	Min
1	two-shot	0	9	8	0	5	6	0	4	5	0	7	11	0	11	4
2	one-shot	0	10	5	0	9	6	0	5	10	0	7	4	0	8	3
2	two-shot	0	7	4	0	6	6	0	7	3	0	9	4	0	7	7
3	two-shot	0	7	7	0	9	7	0	5	6	0	8	5	0	9	4
4	zero-shot	0	4	8	0	6	5	0	7	3	0	4	3	0	4	6
4	one-shot	0	5	6	0	3	8	0	5	6	0	6	5	0	4	8
4	two-shot	0	8	3	0	8	4	0	5	7	0	8	4	0	8	4
5	two-shot	0	7	4	0	4	7	0	7	6	0	10	5	0	7	4

- метод fine-tuning: LoRA
- lora rank: 32
- lora alpha: 16
- Lora dropout: 0.05

В ходе эксперимента наблюдается снижение loss на train датасете и стабильное возрастание BLEU на тестовом датасете. На графике 4 демонстрируется снижение train loss при обучении. На графике 5 представлена зависимость усредненного значения BLEU на тестовом датасете от эпохи. Обучение остановлено на 3 эпохе, так как далее начинается переобучение и результат BLEU на тестовом датасете снижается. Также из графика можно увидеть, что основной прирост баллов происходит на 2 эпохе, дальнейшее обучение приносит меньший эффект.

7. Анализ результатов

В качестве Baseline для сравнения были выбраны:

- Модели ChatGPT o3-mini-high и ChatGPT GPT4-o, как одни из лучших моделей на сегодняшний день;
- Яндекс.Перевдочик, как система, при создании которой большое внимание уделялось переводу на языковой паре английский-русский;

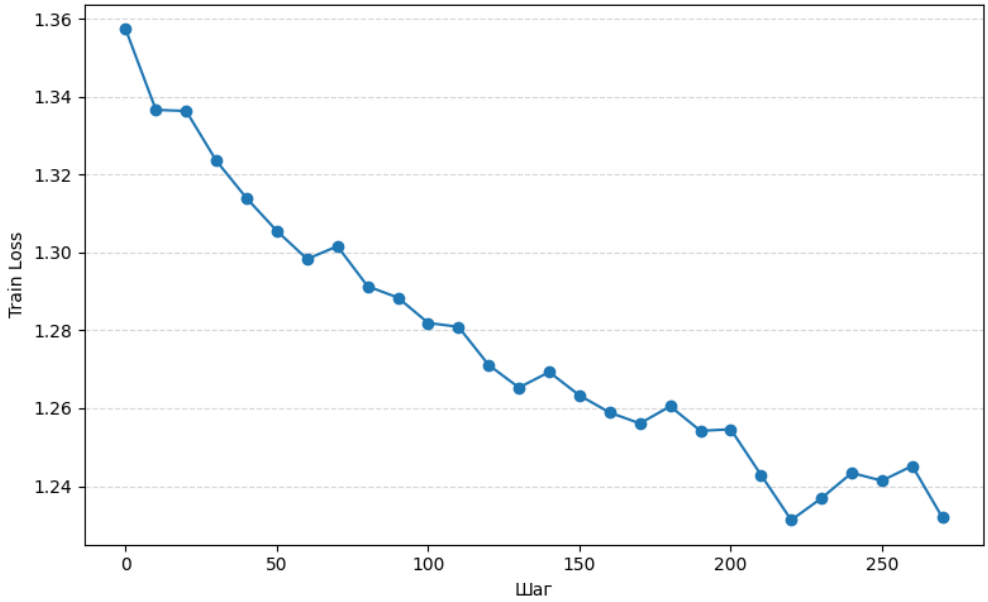

Рисунок 4: График зависимости train loss от числа шагов обучения

- Qwen2.5-3B-Instruct, так как в качестве основы мы используем именно эту модель.

В таблице 7 представлено сравнение систем. Дообученная модель демонстрирует результаты лучше, чем базовая, но все еще уступает другим, более крупным моделям. Это объясняется ограниченным объемом датасета и небольшим размером модели (3 миллиарда параметров). В таблице 8 анализируются баллы модифицированной метри-

Таблица 7: Сравнение дообученой модели и baseline на тестовом датасете

Система	BLEU	Gemba-MQM
Qwen2.5-3B-Instruct (fine-tuned)	11,6	67
Qwen2.5-3B-Instruct	9,0	62
Яндекс.Переводчик	17,8	70
ChatGPT GPT-4o	14,8	82
ChatGPT o3-mini-high	13,0	75

ки Gemba-MQM. Значения метрики считались на пяти текстах из тестового датасета, значения BLEU на десяти текстах. Обе метрики ранжируют модели практически одинаково, за исключением Яндекс.Переводчика: по BLEU данная система занимает первое место, в то время как по баллам MQM она уступает GPT-4o и o3-mini-high. Это объясняется разницей в работе метрик: при большем количестве совпавших n-грамм модель может совершать больше ошибок. Также у Gemba-MQM закономерное распределение

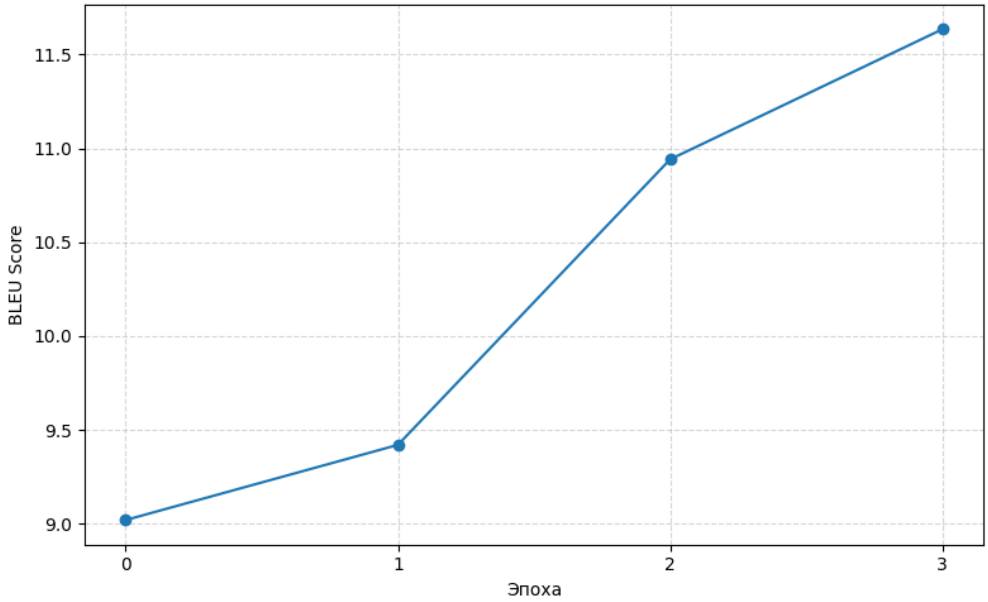

Рисунок 5: График зависимости BLEU(n) на тестовом датасете

ошибок в зависимости от длины текста: как и ожидалось, в текстах 4 и 5, более длинных, модели в среднем совершают больше ошибок. Как видно из таблицы, в некоторых примерах модели, близкие по набранным баллам модели, например Qwen и Qwen-finetuned, могут то отставать, то опережать друг друга по полученным значениям Gemba-MQM. Это можно объяснить как неопределенностью работы самих моделей, которые могут совершать разное количество ошибок от итерации к итерации, так и неполной воспроизводимостью Gemba-MQM, результат работы ChatGPT в качестве аннотатора не является константой для фиксированного текста.

Таблица 8: Подробная таблица с подсчетом Gemba-MQM

Номер текста	Система	Critical	Major	Minor	Количество слов	Итог
1	Qwen-finetuned	0	8	4	207	0,52
1	Qwen	0	7	5	207	0,54
1	GPT-4o	0	1	8	207	0,76
1	Яндекс.Переводчик	0	3	4	207	0,76
1	o3-mini-high	0	3	9	207	0,64
2	Qwen-finetuned	0	4	8	289	0,72
2	Qwen	1	3	13	289	0,58
2	GPT-4o	0	1	7	289	0,85

2	Яндекс.Переводчик	0	4	7	289	0,74
2	o3-mini-high	0	1	10	289	0,79
3	Qwen-finetuned	0	5	6	266	0,70
3	Qwen	0	3	7	266	0,76
3	GPT-4o	0	1	8	266	0,81
3	Яндекс.Переводчик	0	6	6	266	0,66
3	o3-mini-high	0	3	7	266	0,76
4	Qwen-finetuned	0	5	9	381	0,75
4	Qwen	0	11	7	381	0,62
4	GPT-4o	0	0	12	381	0,84
4	Яндекс.Переводчик	0	9	5	381	0,70
4	o3-mini-high	0	2	16	381	0,74
5	Qwen-finetuned	0	8	12	407	0,66
5	Qwen	0	10	12	407	0,61
5	GPT-4o	0	0	11	407	0,86
5	Яндекс.Переводчик	0	7	14	407	0,66
5	o3-mini-high	0	0	16	407	0,80

В таблице 9 анализируются ошибки MQM по типам, для каждой модели вычислен суммарный штраф за ошибки с учетом нормировки (количество Critical ошибок умножается на 25, Major на 5, Minor суммируются без домножения).

Наибольший вклад в штраф MQM вносят ошибки категории "Точность". Такой результат объясняется спецификой метрики MQM, в рамках которой под ‘Точностью’ понимается не только буквальное соответствие исходному тексту, но и правильная передача всех смысловых, стилистических и культурных оттенков оригинала. Категория ‘Точность’ включает наиболее критичные ошибки — такие как искажение смысла, пропуски, излишние дополнения, неправильный перевод ключевых терминов и т.д. В художественных текстах, насыщенных сложными оборотами, культурными реалиями и авторским стилем, именно смысловые расхождения встречаются чаще всего и оказывают наибольшее влияние на восприятие читателем перевода. Следует отметить, что даже современные большие языковые модели, несмотря на значительный прогресс в генерации грамотного и стилистически корректного текста, пока не достигают столь же высокого качества именно при передаче смысла исходного текста. В этой категории

число ошибок у дообученной модели снизилось по сравнению с базовой Qwen.

На втором месте по суммарному штрафу находятся ошибки в категории "Грамотность". Большие языковые модели, демонстрируют схожий — и в целом высокий — уровень владения орфографией, пунктуацией и грамматикой русского языка. Это объясняется тем, что современные системы перевода обучаются на огромных корпусах текстов, что позволяет им эффективно воспроизводить корректные грамматические и орфографические конструкции, даже при переводе сложных литературных фрагментов. Дообученная в работе модель (в таблице Qwen-finetuned) существенно превзошла базовую версию в этой категории ошибок. Это объясняется тем, что базовая модель Qwen изначально обучалась на универсальных корпусах, в которых русскоязычные тексты в целом и художественные тексты в частности были недостаточно представлены. В результате модель часто допускала ошибки при переводе сложных художественных фрагментов. Дополнительное дообучение на корпусе художественных текстов позволило нашей модели лучше усвоить языковые особенности литературного стиля и снизить количество ошибок в категории "Грамотность".

Анализ ошибок по категории "Стиль"показал, что даже современные большие языковые модели могут иногда допускать неестественные, громоздкие или нехарактерные для русского языка выражения, особенно в случаях сложных синтаксических конструкций, игры слов или стилистически окрашенных фрагментов оригинального текста. Однако в целом эти модели демонстрируют достаточно высокий уровень стилистической адаптации.

В категории "Терминология"ошибки встречаются относительно редко. Тем не менее, даже у лидирующих моделей отдельные ошибки стиля и терминологии сохраняются, что указывает на перспективные направления дальнейшей доработки систем машинного перевода художественных текстов.

Интересно отметить, что в распределении ошибок базовой модели Qwen наблюдается относительно небольшое количество ошибок по категориям "Стиль"и "Терминология". Однако это не означает, что модель превосходит другие системы по данным аспектам. Данное явление связано с тем, что основная масса ошибок этой модели приходится на грубые смысловые ("Точность") и грамматические ("Грамотность") нарушения, из-за чего более тонкие стилистические и терминологические неточности либо не проявляются, либо не аннотируются отдельно. В случае сильных моделей (например, GPT-4o и o3-mini-high), напротив, базовые смысловые и грамматические ошибки встречаются

редко, и на первый план выходят вопросы стилистики и терминологии.

Меньше всего модели были оштрафованы за "Локальные адаптации". Такое распределение можно объяснить с одной стороны тем, что современные большие языковые модели легко справляются с этой задачей, а с другой стороны, в художественных текстах редко встречаются специфические элементы форматирования дат, чисел, валют и т.д.

Также можно отметить корреляцию результатов с конференцией WMT 2023, где у многих моделей преобладают ошибки в категории "Точность на втором месте находятся ошибки типа "Грамотность ошибок по категориям "Стиль"и "Терминология"значительно меньше, а штрафов за категорию "Локальные адаптации"почти не наблюдается.

Таблица 9: Анализ ошибок MQM по типам

Система	Точность	Грамотность	Стиль	Терминология	Локальные адаптации
Qwen-finetuned	163	12	8	6	0
Qwen	197	38	2	1	1
Яндекс.Переводчик	137	11	10	8	0
o3-mini-high	69	11	13	0	0
gpt-4o	40	9	10	1	1

8. Выводы

В данной работе подготовлен параллельный корпус художественных текстов, внесены и исследованы изменения в метрику Gemba-MQM для адаптации к задаче, построена и проанализирована модель перевода художественных текстов на основе Qwen2.5-3B-Instruct. В результате экспериментов получены результаты:

- 11,63 баллов BLEU (+2 балла к базовой модели, + 22%);
- 67 баллов Gemba-MQM.

Возможные перспективы развития:

- Дообучение больших и/или более современных моделей (например, Qwen2.5-14B);
- Увеличение датасета для обучения;

- Применение дополнительных стадий обработки для повышения качества основного датасета;
- Улучшение метрики Gemba-MQM;
- Применение других метрик.

Список используемой литературы

[1] - Qwen Team, "Qwen2.5 Technical Report arXiv:2412.15115v2

[2] - Лотман Ю.М, "Структура художественного текста Лотман Ю.М. Об искусстве. – СПб.: «Искусство – СПБ», 1998. – С. 285

[3] - Longyue Wang, Siyou Liu, Chenyang Lyu, Wenxiang Jiao, Xing Wang, Jiahao Xu, Zhaopeng Tu, Yan Gu, Weiyu Chen, Minghao Wu, Liting Zhou, Philipp Koehn, Andy Way, Yulin Yuan, "Findings of the WMT 2024 Shared Task on Discourse-Level Literary Translation WMT 2024 Conference.

[4] - https://www2.statmt.org/wmt25/

[5] - Li An, Linghao Jin, Xuezhe Ma, "MAX-ISI System at WMT23 Discourse-Level Literary Translation Task WMT 2023

[6] - Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin, "Attention Is All You Need arXiv:1706.03762v7

[7] - Xuezhe Ma, Chunting Zhou, Xiang Kong, Junxian He, Liangke Gui, Graham Neubig, Jonathan May, Luke Zettlemoyer, "Mega: Moving Average Equipped Gated Attention arXiv:2209.10655v3

[8] - Fabien Lopez, Gabriela Gonzalez-Saez, Damien Hansen, Mariam Nakhle, Behnoosh Namdarzadeh, Marco Dinarelli, Emmanuelle Esperança-Rodier, Sui He, Sadaf Mohseni, Caroline Rossi, Didier Schwab, Jun Yang, Jean-Baptiste Yunès, Lichao Zhu, Nicolas Ballier,"The MAKE-NMTViz System Description for the WMT23 Literary Task"

[9] - Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, Luke Zettlemoyer, "Multilingual Denoising Pre-training for Neural Machine Translation arXiv:2001.08210v2

[10] - "Encoding Sentence Position in Context-Aware Neural Machine Translation with Concatenation Lorenzo Lupo, Marco Dinarelli, Laurent Besacier, arXiv:2302.06459v2

[11] - Shaolin Zhu, Deyi xiong, "TJUNLP:System Description for the WMT23 Literary Task in Chinese to English Translation Direction WMT 2023

[12] - J¨org Tiedemann, Santhosh Thottingal, "OPUS-MT – Building open translation services for the World"

[13] - Anqi Zhao, Kaiyu Huang, Hao Yu, Degen Huang, "DUTNLP System for WMT23 Discourse-Level Literary Translation"

[14] - Yuhao Xie, Zongyao Li, Zhanglin Wu, Daimeng Wei, Xiaoyu Chen, Zhiqiang Rao, Shaojun Li, Hengchao Shang, Jiaxin Guo, Lizhi Lei, Hao Yang, Yanfei Jiang"HW-TSC's Submissions to the WMT23 Discourse-Level Literary Translation Shared Task"

[15] - Lisa Liu, Ryan Liu, Angela Tsai, Jingbo Shang, "CloudSheep System for WMT24 Discourse-Level Literary Translation WMT 2024

[16] - Yuanchang Luo, Jiaxin Guo, Daimeng Wei, Hengchao Shang, Zongyao Li, Zhanglin Wu , Zhiqiang Rao, Shaojun Li, Jinlong Yang, Hao Yang, "Context-aware and Style-related Incremental Decoding framework for Discourse-Level Literary Translation WMT 2024

[17] - Yuchen Liu1, Yutong Yao1, Runzhe Zhan1, Yuchu Lin2, Derek F. Wong, "NovelTrans: System for WMT24 Discourse-Level Literary Translation WMT 2024

[18] - Kechen Li, Yaotian Tao1, Hongyi Huang, Tianbo Ji, "LinChance×NTU for Unconstrained WMT2024 Literary Translation WMT 2024

[19] - Haoxiang Sun1, Tianxiang Hu, Ruize Gao, Jialong Tang, Pei Zhang, Baosong Yang, Rui Wang, "SJTU LoveFiction's System for WMT24 Discourse-Level Literary Translation WMT 2024

[20] - Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002

[21] - Minghao Wu, Jiahao Xu, Yulin Yuan, Gholamreza Haffari, Longyue Wang, Weihua Luo, Kaifu Zhang, "(Perhaps) Beyond Human Translation: Harnessing Multi-Agent Collaboration for Translating Ultra-Long Literary Texts arXiv:2405.11804v2

[22] - Arle Lommel, Hans Uszkoreit, Aljoscha Burchardt, "Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics Tradumàtica tecnologies de la traducció, 10.5565/rev/tradumatica.77

[23] - Matt Post, "A Call for Clarity in Reporting BLEU Scores"

[24] - Maja Popovi´c, "CHRF: character n-gram F-score for automatic MT evaluation"

[25] - Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, John Makhoul, "A Study of Translation Edit Rate with Targeted Human Annotation"

[26] - Ricardo Rei, Craig Stewart, Ana C Farinha, Alon Lavie, "COMET: A Neural Framework for MT Evaluation"

[27] - Tom Kocmi, Christian Federmann, "GEMBA-MQM: Detecting Translation Quality Error Spans with GPT-4 arXiv:2310.13988

[28] - Ran Zhang, Wei Zhao Steffen Eger, "How Good Are LLMs for Literary Translation, Really? Literary Translation Evaluation with Humans and LLMs arXiv:2410.18697

[29] - Longyue Wang, Zhaopeng Tu, Yan Gu, Siyou Liu, Dian Yu, Qingsong Ma, Chenyang Lyu, Liting Zhou, Chao-Hong Liu, Yufeng Ma, Weiyu Chen, Yvette Graham, Bonnie Webber, Philipp Koehn, Andy Way, Yulin Yuan, Shuming Shi, "Findings of the WMT 2023 Shared Task on Discourse-Level Literary Translation: A Fresh Orb in the Cosmos of LLMs WMT 2023 Conference.

[30] - Brian Thompson, Philipp Koehn, "Vecalign: Improved Sentence Alignment in Linear Time and Space"

[31] - Python nltk library

[32] - Dzmitry Bahdanau, KyungHyun Cho, Yoshua Bengio, "NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE arXiv:1409.0473v7

[33] - Noam Shazeer, "Fast Transformer Decoding: One Write-Head is All You Need arXiv:1911.02150v1

[34] - Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, Sumit Sanghai, "GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints arXiv:2305.13245v3

[35] - Jianlin Su Shenzhen, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, Yunfeng Liu, "ROFORMER: ENHANCED TRANSFORMER WITH ROTARY POSITION EMBEDDING arXiv:2104.09864v5

[36] - Zixuan Jiang, Jiaqi Gu, Hanqing Zhu, David Z. Pan, "Pre-RMSNorm and Pre-CRMSNorm Transformers: Equivalent and Efficient Pre-LN Transformers arXiv:2305.14858v2

[37] - Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, "LORA: LOW-RANK ADAPTATION OF LARGE LANGUAGE MODELS arXiv:2106.09685v2

[38] - Armen Aghajanyan, Luke Zettlemoyer, Sonal Gupta, "INTRINSIC DIMENSIONALITY EXPLAINS THE EFFECTIVENESS OF LANGUAGE MODEL FINE-TUNING arXiv:2012.13255v1

Приложения

Приложение 1.

Таблица 10: Список книг, используемых в датасете.

№	Название книги	Автор
1	Рассказы	Агата Кристи
2	Белый отряд	Артур Конан Дойл
3	Долина ужаса	Артур Конан Дойл
4	Дядя Бернак	Артур Конан Дойл
5	Затерянный мир	Артур Конан Дойл
6	Знак четырех	Артур Конан Дойл
7	Маракотова бездна	Артур Конан Дойл
8	Отравленный пояс	Артур Конан Дойл
9	Приключения Михея Кларка	Артур Конан Дойл
10	Собака Баскервилей	Артур Конан Дойл
11	Сэр Найджел Лоринг	Артур Конан Дойл
12	Торговый дом Гердлстон	Артур Конан Дойл
13	Хирург с Гастеровских болот	Артур Конан Дойл
14	Этюд в багровых тонах	Артур Конан Дойл
15	Гэбриель Конрой	Брет Гарт
16	Аббат	Вальтер Скотт
17	Айвенго	Вальтер Скотт
18	Антикварий	Вальтер Скотт
19	Гай Мэннеринг, или Астролог	Вальтер Скотт
20	Квентин Дорвард	Вальтер Скотт
21	Кенилворт	Вальтер Скотт
22	Ламмермурская невеста	Вальтер Скотт
23	Легенда о Монтрозе	Вальтер Скотт
24	Монастырь	Вальтер Скотт
25	Певерил Пик	Вальтер Скотт

№	Название книги	Автор
26	Пертская красавица, или Валентинов день	Вальтер Скотт
27	Пират	Вальтер Скотт
28	Приключения Найджела	Вальтер Скотт
29	Пуритане	Вальтер Скотт
30	Сент-Ронанские воды	Вальтер Скотт
31	Талисман, или Ричард Львиное-Сердце в Палестине	Вальтер Скотт
32	Уэверли, или шестьдесят лет назад	Вальтер Скотт
33	Черный карлик	Вальтер Скотт
34	История Нью-Йорка	Вашингтон Ирвинг
35	Рип ван Винкль	Вашингтон Ирвинг
36	Воспитание Генри Адамса	Генри Адамс
37	В клетке	Генри Джеймс
38	Вашингтонская площадь	Генри Джеймс
39	Веселый уголок	Генри Джеймс
40	Дэзи Миллер	Генри Джеймс
41	Европейцы	Генри Джеймс
42	Письма Асперна	Генри Джеймс
43	Поворот винта	Генри Джеймс
44	Подлинные образцы	Генри Джеймс
45	Связка писем	Генри Джеймс
46	Урок мастера	Генри Джеймс
47	Ученик	Генри Джеймс
48	Амелия	Генри Филдинг
49	История жизни Джонатана Уайлда Великого	Генри Филдинг
50	История приключений Джозефа Эндруса и его друга Абраама Адамса	Генри Филдинг
51	История Тома Джонса, найденыша	Генри Филдинг
52	Анна-Вероника	Герберт Уэллс
53	В дни кометы	Герберт Уэллс

№	Название книги	Автор
54	Война в воздухе	Герберт Уэллс
55	Война миров	Герберт Уэллс
56	Жена сэра Айзека Хармана	Герберт Уэллс
57	История мистера Полли	Герберт Уэллс
58	Когда спящий проснется	Герберт Уэллс
59	Колеса фортуны	Герберт Уэллс
60	Машина времени	Герберт Уэллс
61	Освобожденный мир	Герберт Уэллс
62	Остров доктора Моро	Герберт Уэллс
63	Первые люди на Луне	Герберт Уэллс
64	Человек невидимка	Герберт Уэллс
65	Билли Бадд, фор-марсовый матрос	Герман Мелвилл
66	Моби Дик, или Белый кит	Герман Мелвилл
67	Гордость и предубеждение	Джейн Остин
68	Доводы рассудка	Джейн Остин
69	Белый Клык	Джек Лондон
70	Дочь снегов	Джек Лондон
71	Зов предков	Джек Лондон
72	Лунная долина	Джек Лондон
73	Маленькая хозяйка Большого дома	Джек Лондон
74	Межзвездный скиталец	Джек Лондон
75	Морской Волк	Джек Лондон
76	Прежде Адама	Джек Лондон
77	Смок Беллью	Джек Лондон
78	Как мы писали роман	Джером Клапка Джером
79	Памяти Джона Ингерфилда и жены его Анны	Джером Клапка Джером
80	Трое в одной лодке, не считая собаки	Джером Клапка Джером
81	Трое на велосипедах	Джером Клапка Джером
82	Дуэль	Джозеф Конрад
83	Зеркало морей	Джозеф Конрад
84	Лорд Джим	Джозеф Конрад

№	Название книги	Автор
85	Сердце тьмы	Джозеф Конрад
86	Тайный сообщник	Джозеф Конрад
87	Тайфун	Джозеф Конрад
88	Фрейя Семи Островов	Джозеф Конрад
89	Вилла Рубейн	Джон Голсуорси
90	Пылающее копье	Джон Голсуорси
91	Фриленды	Джон Голсуорси
92	Миддлмарч	Джордж Элиот
93	Испытание Гилберта Пинфолда	Ивлин Во
94	Не жалейте флагов	Ивлин Во
95	Незабвенная	Ивлин Во
96	Офицеры и джентльмены	Ивлин Во
97	Пригоршня праха	Ивлин Во
98	Упадок и разрушение	Ивлин Во
99	Ветер в ивах	Кеннет Грэм
100	Алиса в Зазеркалье	Льюис Кэролл
101	Алиса в стране чудес	Льюис Кэролл
102	Белый вождь	Майн Рид
103	Всадник без головы	Майн Рид
104	Затерянные в океане	Майн Рид
105	Квартеронка	Майн Рид
106	Морской волчонок	Майн Рид
107	Оцеола, вождь семинолов	Майн Рид
108	Детектив с двойным прицелом	Марк Твен
109	Жанна д'Арк	Марк Твен
110	Запоздавший русский паспорт	Марк Твен
111	Наследство в тридцать тысяч долларов	Марк Твен
112	Позолоченный век	Марк Твен
113	Приключения Тома Сойера	Марк Твен
114	Простофиля Вильсон	Марк Твен

№	Название книги	Автор
115	Путешествие капитана Стормфтлда в рай	Марк Твен
116	Рассказы	Марк Твен
117	Таинственный незнакомец	Марк Твен
118	Том Сойер, сыщик	Марк Твен
119	Алая буква	Натаниель Готорн
120	Векфильдский Священник. История его жизни, написанная, как полагают, им самим	Оливер Голдсмит
121	Дживс и Вустер	Пэлем Грэнвил Вудхауз
122	Вечера с историком	Рафаэль Сабатини
123	Одиссея Капитана Блада	Рафаэль Сабатини
124	Псы Господни	Рафаэль Сабатини
125	Скарамуш	Рафаэль Сабатини
126	Ким	Редьярд Киплинг
127	От моря до моря	Редьярд Киплинг
128	Отважные мореплаватели	Редьярд Киплинг
129	Свет погас	Редьярд Киплинг
130	Алмаз Раджи	Роберт Луис Стивенсон
131	Владетель Баллантрэ	Роберт Луис Стивенсон
132	Катриона	Роберт Луис Стивенсон
133	Остров сокровищ	Роберт Луис Стивенсон
134	Потерпевшие кораблекрушение	Роберт Луис Стивенсон
135	Похищенный, или Приключения Дэвида Бэлфура	Роберт Луис Стивенсон
136	Сент Ив	Роберт Луис Стивенсон
137	Странная история доктора Джекила и мистера Хайда	Роберт Луис Стивенсон
138	Черная стрела	Роберт Луис Стивенсон
139	Бэббит	Синклер Льюис
140	Кингсблад, потомок королей	Синклер Льюис
141	Эроусмит	Синклер Льюис

№	Название книги	Автор
142	Каталина	Сомерсет Моэм
143	Луна и грош	Сомерсет Моэм
144	Острие бритвы	Сомерсет Моэм
145	Пироги и пиво, или Скелет в шкафу	Сомерсет Моэм
146	Рассказы	Сомерсет Моэм
147	Рождественские каникулы	Сомерсет Моэм
148	Театр	Сомерсет Моэм
149	Тогда и теперь	Сомерсет Моэм
150	Узорный покров	Сомерсет Моэм
151	Американская трагедия	Теодор Драйзер
152	Гений	Теодор Драйзер
153	Домой возврата нет	Томас Вулф
154	Вдали от обезумевшей толпы	Томас Гарди
155	Возвращение на родину	Томас Гарди
156	Мэр Кэстербриджа	Томас Гарди
157	Под деревом зеленым или Меллстокский хор	Томас Гарди
158	Тэсс из рода д'Эрбервиллей	Томас Гарди
159	Аббатство Кошмаров	Томас Лав Пикок
160	Усадьба Грилла	Томас Лав Пикок
161	Лунный камень	Уилки Коллинз
162	Виргинцы	Уильям Мейкпис Теккерей
163	Записки Барри Линдона, эсквайра, писанные им самим	Уильям Мейкпис Теккерей
164	История Пенденниса, его удач и злоключений, его друзей и его злейшего врага	Уильям Мейкпис Теккерей
165	Кольцо и роза, или история принца Обалду и принца Перекориля	Уильям Мейкпис Теккерей
166	Кэтрин	Уильям Мейкпис Теккерей

№	Название книги	Автор
167	Ньюкомы	Уильям Мейкпис Теккерей
168	Роковые сапоги	Уильям Мейкпис Теккерей
169	Ярмарка тщеславия	Уильям Мейкпис Теккерей
170	Маленький лорд Фаунтлерой	Френсис Ходгсон Бернет
171	Алмаз величиной с отель "Риц"	Фрэнсис Скотт Фицджеральд
172	Последний магнат	Фрэнсис Скотт Фицджеральд
173	Барнеби Радж	Чарльз Диккенс
174	Большие надежды	Чарльз Диккенс
175	Жизнь Дэвида Копперфилда	Чарльз Диккенс
176	Жизнь и приключения Николаса Кильби	Чарльз Диккенс
177	Земля Тома Тиддлера	Чарльз Диккенс
178	Картины Италии	Чарльз Диккенс
179	Лавка древностей	Чарльз Диккенс
180	Оливер Твист	Чарльз Диккенс
181	Посмертные записки Пиквикского клуба	Чарльз Диккенс
182	Путешественник не по торговым делам	Чарльз Диккенс
183	Рождественская песнь	Чарльз Диккенс
184	Тайна Эдвина Друда	Чарльз Диккенс
185	Торговый дом Домби и сын	Чарльз Диккенс
186	Тяжелые времена	Чарльз Диккенс
187	Холодный дом	Чарльз Диккенс
188	Городок	Шарлотта Бронте
189	Джен Эйр	Шарлотта Бронте
190	Шерли	Шарлотта Бронте

№	Название книги	Автор
191	Уайнсбург, Огайо	Шервуд Андерсон
192	Повесть о приключениях Артура Гордона Пима	Эдгар Алан По
193	Боксер Билли	Эдгар Райс Берроуз
194	Крэнфорд	Элизабет Гаскелл
195	Грозовой перевал	Эмилия Бронте
195	Овод	Этель Лилиан Войнич

Приложение 2.

Используемый в работе prompt:

You are an annotator for the quality of machine translation of literary texts. Your task is to identify and classify all errors in the translation below using the provided categories and severity levels.

Instructions:

- Output only a markdown table as shown below.
- Do not add any explanations, summaries, or extra text.
- If there are no errors, output a single row with "No errors detected" in the comment column.

Categories of errors:

- Accuracy: addition, omission, mistranslation, misnomer, untranslated
- Fluency: punctuation, spelling, grammar, inconsistency
- Style: awkwardness, inconsistent, unidiomatic
- Terminology: mistranslation, inconsistent
- Locale convention: location format, number format
- Other

Severity:

- Critical: Inhibits comprehension.
- Major: Disrupts flow but meaning is still clear.
- Minor: Error present but does not disrupt comprehension.

Markdown Table Format:

№|Category/Subtype|Severity|Source fragment|Target fragment|Comment

source:

«. . . »

translation:

«. . . »

example_one:

«. . . »

example_two:

«. . . »

Приложение 3.

Пример перевода текста до и после дообучения.

Текст, переведенный с использованием исходной модели:

Во всех общественных учреждениях Америки преобладает исключительная вежливость. Многие из наших министерств могли бы существенно улучшить себя в этом отношении, но таможня в особенности должна поучиться у Соединенных Штатов и сделать себя менее неприятным и отталкивающим для иностранных граждан. Слуги французского чиновничества достаточно презираемы из-за своей подхалимской жадности; но у нас есть грубая неуважительность к нашим людям, которая отвратительна для всех, кто попадает в их руки, и порочит страну, которая держит таких неприветливых псарей рычавшими возле своих ворот.

Когда я приземлился в Америку, я не мог не ощутить сильного впечатления от контраста, который предстает передо мной таможнями, и вежливости, внимания и добродушия, с которыми их чиновники выполняют свои обязанности.

Так как мы не приземлились в Бостоне из-за некоторой задержки на причале до вечера, я получил свои первые впечатления от города, прогуливаясь к таможне после нашего прибытия, что было воскресенье. Я боюсь сказать, как много приглашений на места в церкви для этого утра было предложено нам, путем официального приглашения, прежде чем мы даже завершили первый ужин в Америке, но если я могу позволить себе сделать умеренное предположение, без более тщательного расчета, я думаю, что предложено нам хотя бы столько мест, сколько могло бы вместить несколько десятков взрослых семей. Количество верований и форм религии, к которым было приглашено удовольствие нашего присутствия, было в очень справедливом пропорциональном соотношении.

Тот же текст, но переведенный с помощью итоговой модели:

В общественных учреждениях Соединенных Штатов преобладает величайшая вежливость. Большинство наших департаментов могло бы существенно улучшить себя в этом отношении, но тамошняя таможня особенно заслуживает внимания и примера: она должна сделать себя менее неприятной и отталкивающей для иностранных граждан. Служители французского таможенного ведомства известны своей подхалимской жадностью, но нашим таможенным служащим свойственна грубая неуважительность,

вызывающая отвращение у всех, кто попадает в их руки, и оскорбляющая достоинство страны, которая держит в своих воротах таких негодяев, как они.

Когда я прибыл в Америку, я не мог не поразиться контрасту между таможенным ведомством Соединенных Штатов и той вежливостью, с которой его служители исполняли свои обязанности.

Поскольку мы прибыли в Бостоне только вечером, после того как нас задержали на причале, я получил свои первые впечатления о городе, прогуливаясь по улицам в воскресенье утром, после того как завтракали впервые в Соединенных Штатах. Мне страшно признаться, сколько было предложено нам мест в церкви, чтобы мы могли присутствовать на богослужении, когда мы только что закончили завтрак в Америке, но если я могу позволить себе приблизительный расчет, не вдаваясь в подробности, я скажу, что предлагалось нам столько мест, сколько было бы достаточно для двух-трех десятков взрослых семей. Число вероисповеданий и религиозных обрядов, с которыми нам предлагали присутствовать на богослужении, было весьма разнообразным.