ЛАБОРАТОРИЯ № 80
«КИБЕРФИЗИЧЕСКИХ СИСТЕМ»

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ УЧРЕЖДЕНИЕ НАУКИ
ИНСТИТУТ ПРОБЛЕМ УПРАВЛЕНИЯ
ИМ. В.А. ТРАПЕЗНИКОВА
РОССИЙСКОЙ АКАДЕМИИ НАУК

О методах обработки слабоструктурированной информации на естественных языках в условиях сильных шумов для решения задач безопасности

Мещеряков Р.В.
д.т.н., проф.

Мельников С.Ю.
д.ф.-м.н.

meshcheryakov.roman@gmail.com

Комплексная защита информации
6 октября 2026 г.

Постановка задачи. Примеры

Эталонный текст

КИШИНЕВ, 28 ф̌ев — РИА Новости.
Прокуроры генеральной прокуратуры Румынии задержали 18 подозреваемых в рамках расследования дела, в котором обвинения предъявлены бывшему кандидату в президенты страны Кэлину Джеорджеску, сообщают в пятницу СМИ страны, включая телеканал Realitatea.
СМИ отмечают, что задержания были проведены в ходе масштабной операции с обысками по 47 адресам в пяти регионах Румынии.
Подозреваемые, среди которых сторонники Джеорджеску, обвиняются в действиях против конституционного строя, создании организации ф̌ашистского характера и нарушениях, связанных с ф̌инансированием его избирательной кампании.
"Из 21 сторонника Джеорджеску 18 находятся под стражей»

Результат OCR распознавания

К III II III II В 128 ф̌ев РИА Нового и Прокуроры генеральной прокура у рыб умер ни задержали во дозревае мы х в рамках расследования цела в котором обвинения предъявлены бывшему канди и президенты страны Кэлину Джо орджеску сообщают в пятницу СМИт раны включая пол с канал Realitatea СМИ отмечают в Бо за сржания были проведены в ходе за outi зоной по рации с обысками по 47 а ре сам в пяти
во гионах Румынии
Подозреваемые сро и которых сторонники Л коорjжер ку обвиняюсся в депо гиях против оно и туционного строя создании организации ф̌ашистско она рак трате веру uclinях связанных ф̌инансировани см его избирательной кампании
На 21 сторонника Джеордже стул 8 находятся под стражей

Основные подходы к коррекции искаженного текста

Задачи коррекции искаженных текстов, полученных при машинном распознавании речи или изображений текстов, называются **post-ASR** и **post-OCR коррекцией** соответственно.

Подходы:

- использование **вероятностно-лингвистических моделей** языка

  минусы: недостаточная точность,
  плюсы: высокая скорость и интерпретируемость результатов

- использование **Seq2Seq** подходов

  минусы: нужно много данных для обучения, низкая скорость,
  плюсы: выше точность

- использование **LLM**

  минусы: галлюцинации, выч. затраты, неинтерпретируемость
  плюсы: высокая точность

С 2023 года наиболее активно развивается именно последнее направление.

Степень проработанности темы в мировой науке. Коррекция с использованием LLM

Позитивные результаты использования LLM

Hajiali, 2023. Корпус распознанных OCR Tesseract ver. 3.0.2 современных англ. книг по биологии, в котором 3000 ошибок.
Точность коррекции BERT – 68%, GPT-4 – 75%.
Thomas, 2024. Корпус английских газет XIX в.
Модель BART уменьшает CER на 23%, модель Llama2 на 55%.
Löfgren, 2024. Высокая точность LLM при post-OCR коррекции шведских газет XIX в.
Jasonarson, 2023. Высокая точность при коррекции исландских текстов.
de Araújo, 2024. Высокая точность для португальского.
Evaggelatos, 2025. Высокая точность для греческого.

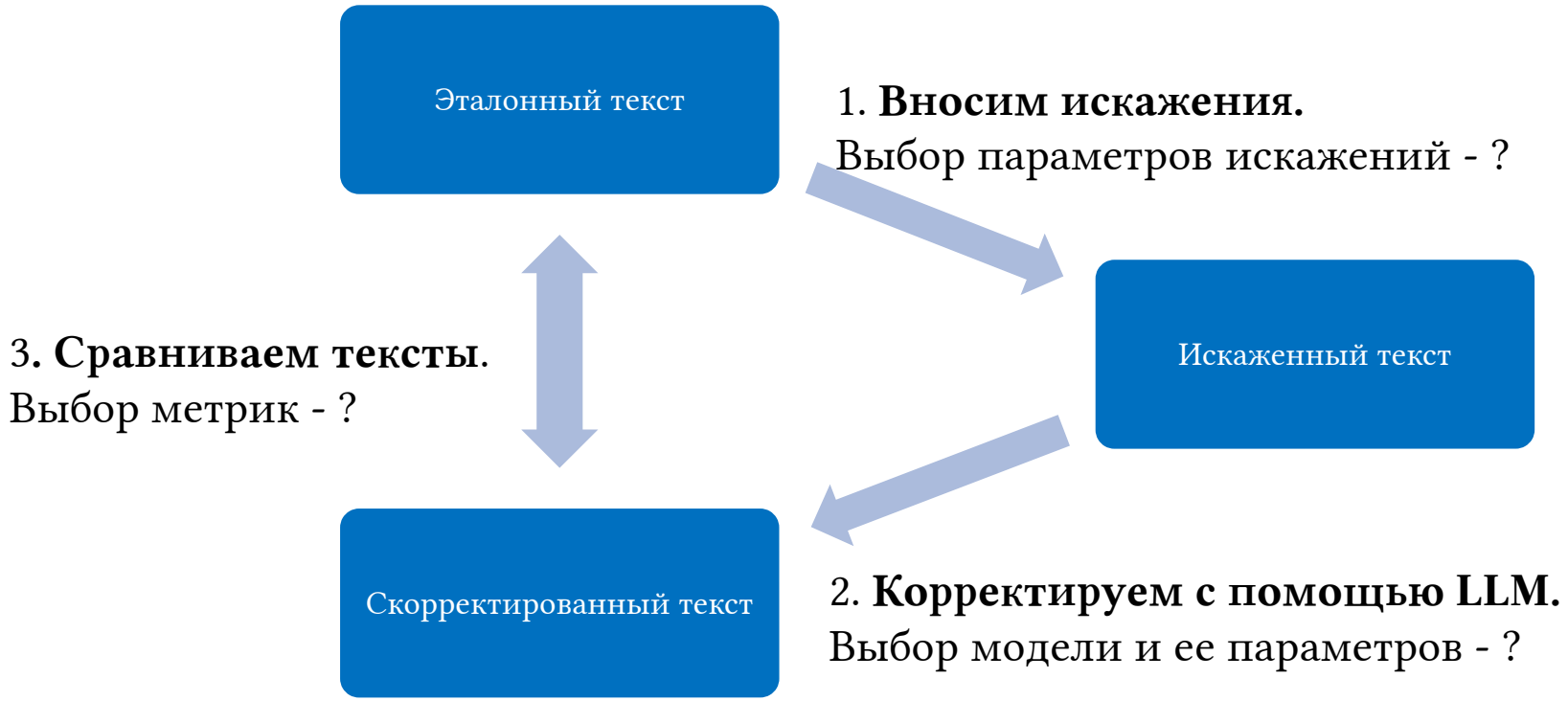
Негативные результаты использования LLM

Kanerva, 2025. «OCR error post-correction with LLMs in historical documents: No free lunches». Исторические тексты на финском языке.
Boroş, 2024. Мультиязычные материалы XVII-XX вв.
Biriuchinskii, 2025. Исторические французские тексты.

Коррекция. Схема исследования

Вопрос 1: Искажения какого уровня могут корректироваться LLM?
Вопрос 2: Какие LLM корректируют лучше?

Схема:



Выбор программной среды и параметров модели

Word Error Rate: WER= (S+D+I)/N где S – количество замен, D – количество удалений, I – количество вставок.

Синтетические искажения текстов.

Текст просматривается посимвольно слева направо.
С заданной вероятностью P, 0.1<P<0.45, принимается решение об искажении текущего символа.

- с вероятностью 1/3 символ удаляется,
- с вероятностью 1/3 вставка символа (равновероятно на алфавите),
- с вероятностью 1/3 замена символа на другой (равновероятно на алфавите).

Выбор программной среды для загрузки и локального подключения offline LLM.

Вычислитель: процессор - AMD Ryzen 9 5950X 16-Core, 3.40 GHz; оперативная память – 128 Gb; видеопамять - NVIDIA GeForce RTX 4070 Ti, 12Gb; ОС – Windows 11.

Рассматривались варианты программ запуска LLM: **LM Studio,** GPT4All, Ollama, LLaMa.cpp.

Выбраны следующие параметры моделей*:* **temperature=0, Top-p=0, Top-k=5, frequency_penalty=0, presence_penalty=0, max_tokens=-1.**

Выбор таких параметров обеспечивает низкую вариативность генерируемого текста.

Инструментарий для сравнения моделей по качеству коррекции

Выбор offline LLM

Проанализированы более 40 offline LLM (универсальные и специализированные для грамматической коррекции текста), с разным количеством параметров и степенями квантования.

Таблица 1. Снижение WER при коррекции **offline LLM**, 10 лучших моделей.
(**жирным шрифтом** выделены предпочтительные значения параметров).

Модель	Разработчик	Архитектура	Параметры	Квантование	Объем Гб	Токены	*WER* снижение %
airoboros-l2-gpt4-1.4.1	Meta	Llama	13B	**Q5_0** Q6_K	8.98 10.7	4096	23
LLaMA2-Psyfighter2	Meta	Llama	13B	**Q4_0** Q6_K	7.36 10.7	4096	23
Meta-Llama-3-Instruct	Meta	Llama	8B	F16 **Q8_0** Q6_K Q5_K_M Q4_K_M	16.07 8.54 6.6 5.73 4.92	8192	33
Mistral-Instruct-v0.1, **v0.2**	Mistral AI	Llama	7B	**Q8_0** Q4_K_S	7.7 4.14	32768	26
MLewd-L2-Chat	Undi	Llama	13B	**Q4_0** Q6_K	7.37 10.7	4096	26
Orca-2	Microsoft	Llama	13B	**Q4_0** Q6_K	7.37 10.7	4096	31
Starling-LM-beta	Stability AI	Llama	7B	Q8_0	7.7	4096	23
StableBeluga	Stability AI	Llama	**7B** **13B**	Q8_0 **Q4_K_M** Q6_K	7.16 7.87 10.7	4096	28
Wizard-Vicuna	LMSYS Org	Llama	7B **13B**	Q6_K **Q4_0**	5.53 7.36	16384	25
xwin-mlewd-v0.2	Undi	Llama	13B	Q4_0 **Q5_K_M**	7.36 9.23	4096	26

Выбор online llm

Таблица 2. Снижение WER при коррекции online LLM.

Модель	Разработчик	Архитектура	Параметры	Длина контекста	*WER* снижение %
Claude 3 Haiku	Anthropic	Claude	52B	200к	45
Gemini-1 Pro	Google	MultiTransformer	неизвестно	32к	38
GigaChat	СБЕР	Giga	неизвестно	8k	32
GPT - 3.5	OpenAI	GPT	175B	2к	38
GPT - 4	OpenAI	GPT	175B	128к	31
GPT – 4o	OpenAI	GPT	80B	8к	32
GPT – 4o mini	OpenAI	GPT	30B	4к	41
Grammer GPT	Grammarly	GPT	неизвестно	неизвестно	40
Llama 3.1	Meta	Llama	70B	16к	35
Mistral large 2	Mistral AI	Mistral	70B	32к	31
Pixtral	Mistral AI	Mixtral	124B	128k	29
Qwen 2.5	Qwen AI	Qwen	70B	32k	38
Zero GPT	Zero GPT	GPT	неизвестно	неизвестно	35

Промпты для коррекции

Примеры промптов, тестировавшихся для задачи коррекции текстов:

№1 – «I want you act as a professional corrector. I will provide you texts and I would like you to review them for any spelling, grammar, or punctuation errors. The length of the words and backspaces in the source and correct text must be the same. Give me a correct text with all corrections. It is very important for me!»

№3 – «For the duration of this conversation, I want you to act as a meticulous proofreader. I will provide you with various texts and I expect you to thoroughly review them for any spelling, grammar or punctuation errors. Your attention to detail and accuracy in this task is of utmost importance.»

№6 – «I want you act a proofreader. Repair this text!»

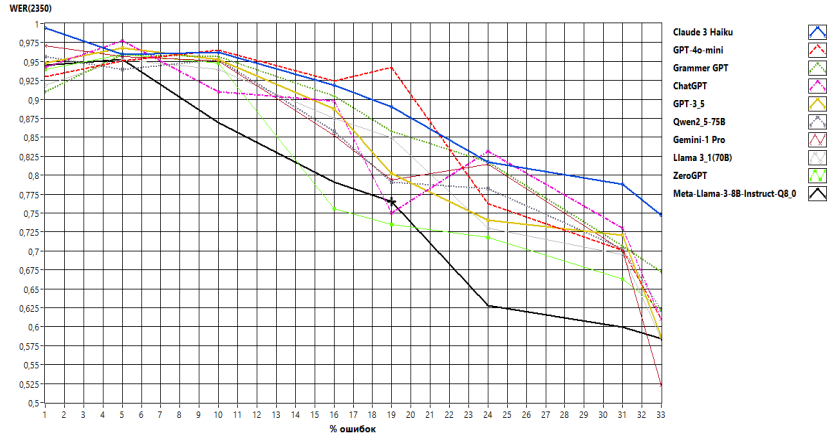
№7 – «Give me the best repaired text without any errors! It is very important for my business!»

Промпт №3 был искусственно сгенерирован моделью GPT-4o

Лучший промпт для большинства моделей при коррекции текстов - промпт №1, худший - №6.

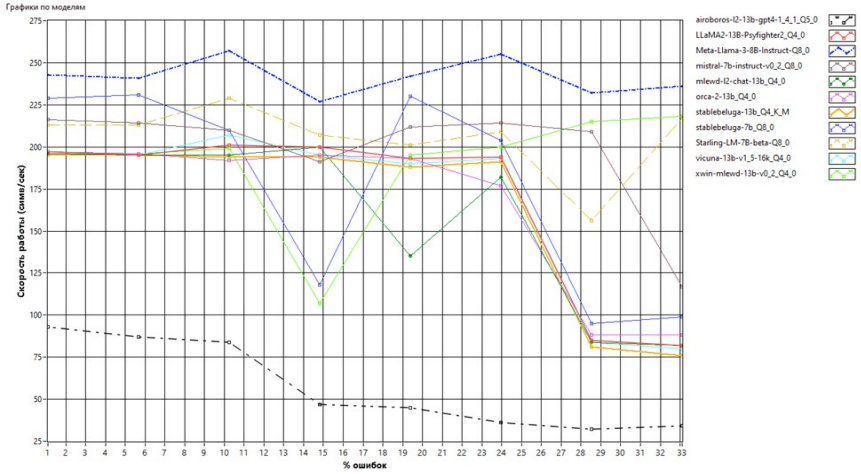
Точность (WER) и скорость коррекции в зависимости от степени искажения

Зависимость точности от уровня искажений



Точность коррекции зависит от размера offline моделей – 13B практически всегда лучше 7B (есть исключения в виде instruct моделей, дообучающихся на корпусе инструкций).
Для online моделей качество коррекции слабо зависит от объема (30B достаточно), но зависит от модели и оптимизации её параметров.
С ростом искажений точность коррекции падает.

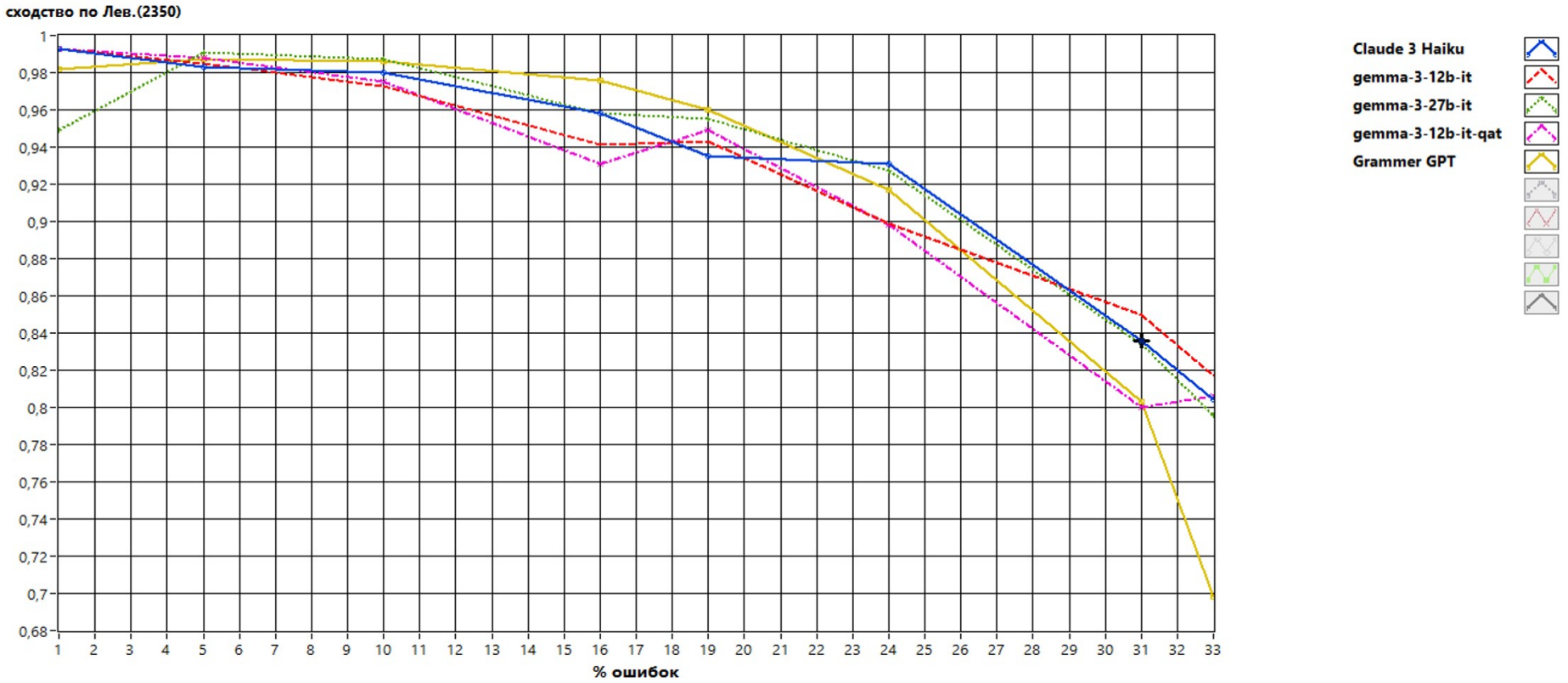
Зависимость скорости от уровня искажений



Скорость коррекции текстов для разных моделей варьируется в диапазоне 50 до 250 симв/сек.
Зависит от размера модели, большие модели работают медленнее.
При сильных искажениях (более 24% текста) скорость резко уменьшается, и модели часто галлюцинируют.

Точность (Левенштейн) коррекции в зависимости от степени искажения

Точность коррекции (сходство по расстоянию Левенштейна между эталонным и скорректированным текстом) в зависимости от процента ошибок в искаженном тексте.



Заключение

Основные результаты

1. Протестированы более 50 LLM с различными параметрами. По результатам тестов, проведенных летом 2025 года, лучшая offline модель – Meta-Llama-3-8b-instruct, лучшая online LLM – Claude 3 Haiku.
2. Online LLM модели лучше корректируют текст, чем offline модели, при искажениях < 30% количество ошибочных слов после коррекции уменьшается на 35-40%. Лучшие offline LLM модели эффективны при искажениях < 20%, количество слов с ошибками уменьшается на 25-30%.
3. Offline модели бОльшего объема обеспечивают в среднем лучшую точность коррекции.

Выявленные ограничения

4. Для устойчивой работы размер искаженного текста плюс запрос к модели (параметры модели + промпт) должны быть меньше максимального количества токенов для конкретной модели. Модель должна полностью загружаться в видеопамять.
5. При большом количестве ошибок скорость может упасть на порядок и модель может “подвиснуть”.
   Точность коррекции зависит от объема offline моделей и мало зависит от уровня квантования.
   Минимальный уровень квантования должен быть не меньше Q=4 и количество параметров при обучении не меньше 7B. Изменение параметров temperature, Top_p, и др. приводят к незначительным изменениям точности коррекции.
   Для online моделей точность коррекции слабо зависит от объема модели.

Работа поддержана грантом РНФ 24-11-00340 Исследование и разработка методов обработки слабоструктурированной информации на естественных языках в условиях сильных шумов для решения задач безопасности