

Применение технологий LLM для анализа электронных медицинских карт: опыт Webiomed

Александр Гусев

к.т.н., директор по развитию Webiomed,
эксперт по искусственному интеллекту
ЦНИИОИЗ Минздрава России

Елена Макарова

к.т.н., руководитель ИИ-направления
Webiomed

IX Всероссийская конференция по искусственному интеллекту в здравоохранении
и системам поддержки принятия врачебных решений
6-7 февраля 2025 года, Москва

Платформа прогнозной аналитики Webiomed

Предназначена для обработки обезличенных медицинских данных с помощью технологий ИИ: автоматически формирует цифровые профили пациентов и предоставляет клиническую и управленческую аналитику для врачей, пациентов и руководителей в сфере здравоохранения

41

субъект РФ
подключен
к платформе

50

тыс. врачей
используют
СППВР Webiomed

51

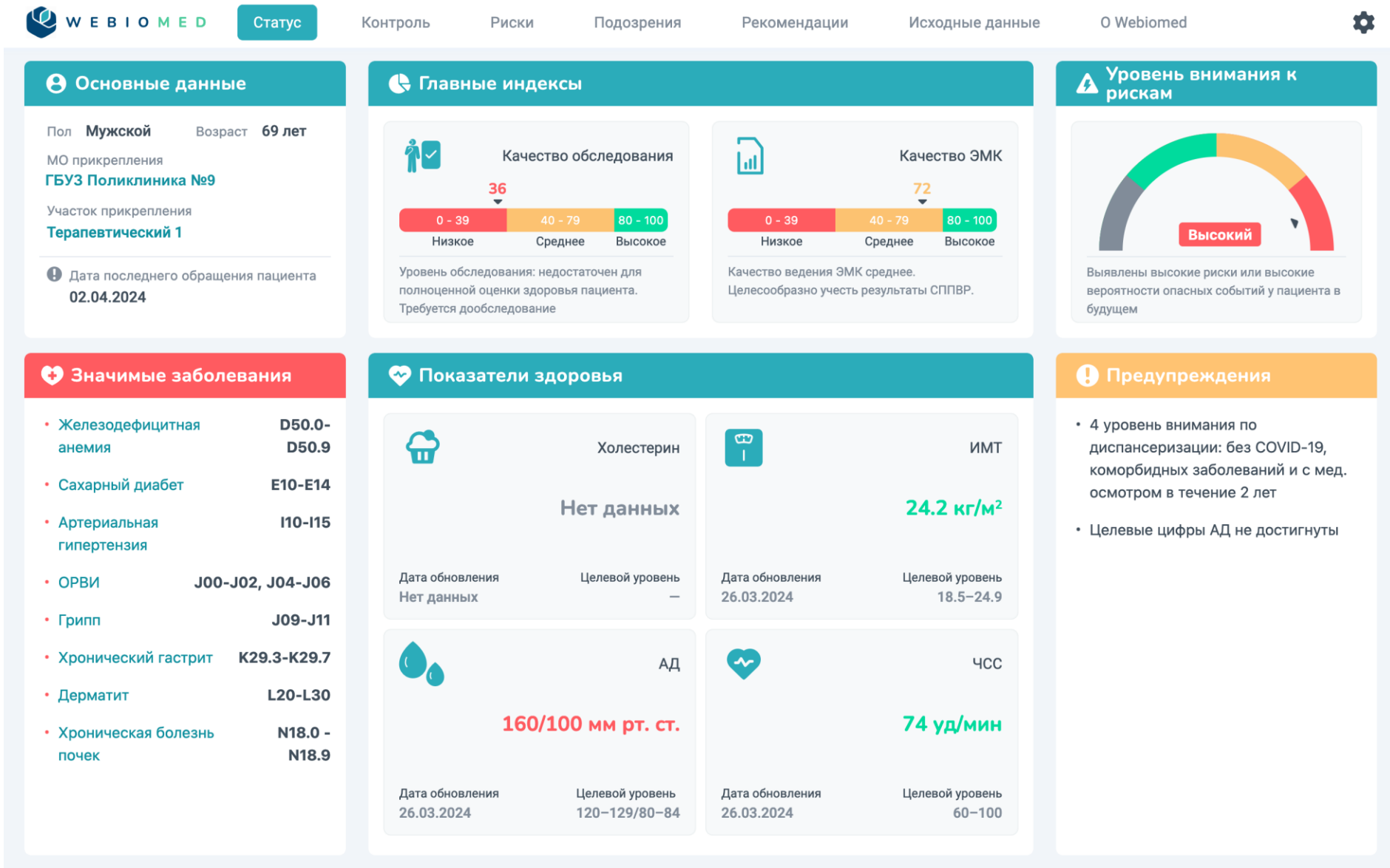
млн. пациентов
накоплено в
платформе

3тыс

признаков
извлекает NLP-
сервис из ЭМК

Ключевые функциональные возможности

- ✓ Выявление пациентов высокого риска на основе прогнозной аналитики
- ✓ Выявление пациентов с подозрениями на различные заболевания
- ✓ Формирование персональных рекомендаций врачу по сохранению здоровья пациента
- ✓ Контроль соблюдения требований Минздрава и клинических рекомендаций по профилактике заболеваний
- ✓ Контроль качества обследования пациентов
- ✓ Контроль качества ведения врачами электронных медицинских карт
- ✓ Поддержка принятия управленческих решений по эпидемиологии и контролю работы врачей



Проектная задача: улучшить выявление пациентов с высоким риском развития наркологических заболеваний

Текущий алгоритм выявляет пациентов высокого наркологического риска у 1-5% взрослого населения

W E B I O M E D

Статус

Контроль

Риски

Подозрения

Рекомендации

Исходные данные

О Webiomed

Сердечно-сосудистые заболевания

Высокий уровень внимания

Сахарный диабет

Обычный уровень внимания

Заболевания органов дыхания

Высокий уровень внимания

COVID-19 и воспалительный синдром

Высокий уровень внимания

Наркологические заболевания

Обычный уровень внимания

МЕТОДИКИ ОЦЕНКИ

Риск пагубного употребления алкоголя:

Низкий

Методика оценки риска опасного употребления алкоголя на основании Методических рекомендаций «Организация и поведение профилактического медицинского осмотра и диспансеризации определенных групп взрослого населения 2019 г.»

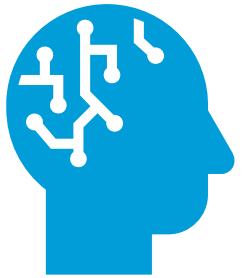
ФАКТОРЫ РИСКА

Факторы риска не выявлены

Факторы риска

10

Для этого нужно извлечь сложный признак из неструктурированной медицинской записи в ЭМК, с учетом контекста



В ходе добавления алгоритма оценки рисков развития наркологических заболеваний потребовалось извлечь из врачебных записей указания на злоупотребления пациентом алкоголя и на фиксацию фактов алкогольного объяснения

Сложности выполнения этой задачи связана с высокой вариативностью описания признаков:

«Злоупотребление алкоголем» (в анамнезе)

- 1) прямое указание врача-нарколога на злоупотребление
- 2) частое, чрезмерное употребление
- 3) повышенная толерантность к алкоголю
- 4) запои, амнестические формы опьянения
- 5) И т.д.

«Алкогольное опьянение» (в момент осмотра)

- 1) нахождение в алкогольном опьянении
- 2) результаты медицинского освидетельствования
- 3) алкогольная интоксикация
- 4) запах перегара, алкоголя
- 5) и т.д.

Необходимость оценки контекста употребления терминов:

- 1) Описание семейного анамнеза, алкогольного опьянения у родственников (например, детей, изъятых опекой)
- 2) Описание факта злоупотребления в прошлом у пациента, вышедшего в ремиссию и т.д.

Признаки в неверном контексте не должны извлекаться как актуальные для пациента.

Для проверки точности работы модели извлечения нарколологических признаков был создан валидационный набор

- ✓ В наборе данных присутствовали записи врачей следующих специальностей: психиатров-наркологов, терапевтов, травматологов, фельдшеров.
- ✓ Тексты были предварительно отобраны, исходя из разнообразия медицинских учреждений и используемой в текстах лексики.
- ✓ Разметка выполнялась медицинскими экспертами Webiomed

Признак	Количество записей	Из них с положительным классом
Злоупотребление алкоголем	470	204
Алкогольное опьянение	470	74

Попытка решения задачи привычными NLP-методами (1я версия NLP-модели) для «Злоупотребления алкоголем»

Пример (1)

Впервые спиртное попробовал в 18 лет. Употребление алкоголя эпизодическое, "по праздникам", частотой 1 раз в месяц.

Употребляет преимущественно вино, пиво, толерантность до 2.5 литров в сутки. В АО спокоен, амнезии отрицает. ААС отрицает, не опохмеляется. Ситуационный контроль снижен. Последняя алкоголизация – новый год 2020г.

Ответ модели: True (ошибка)

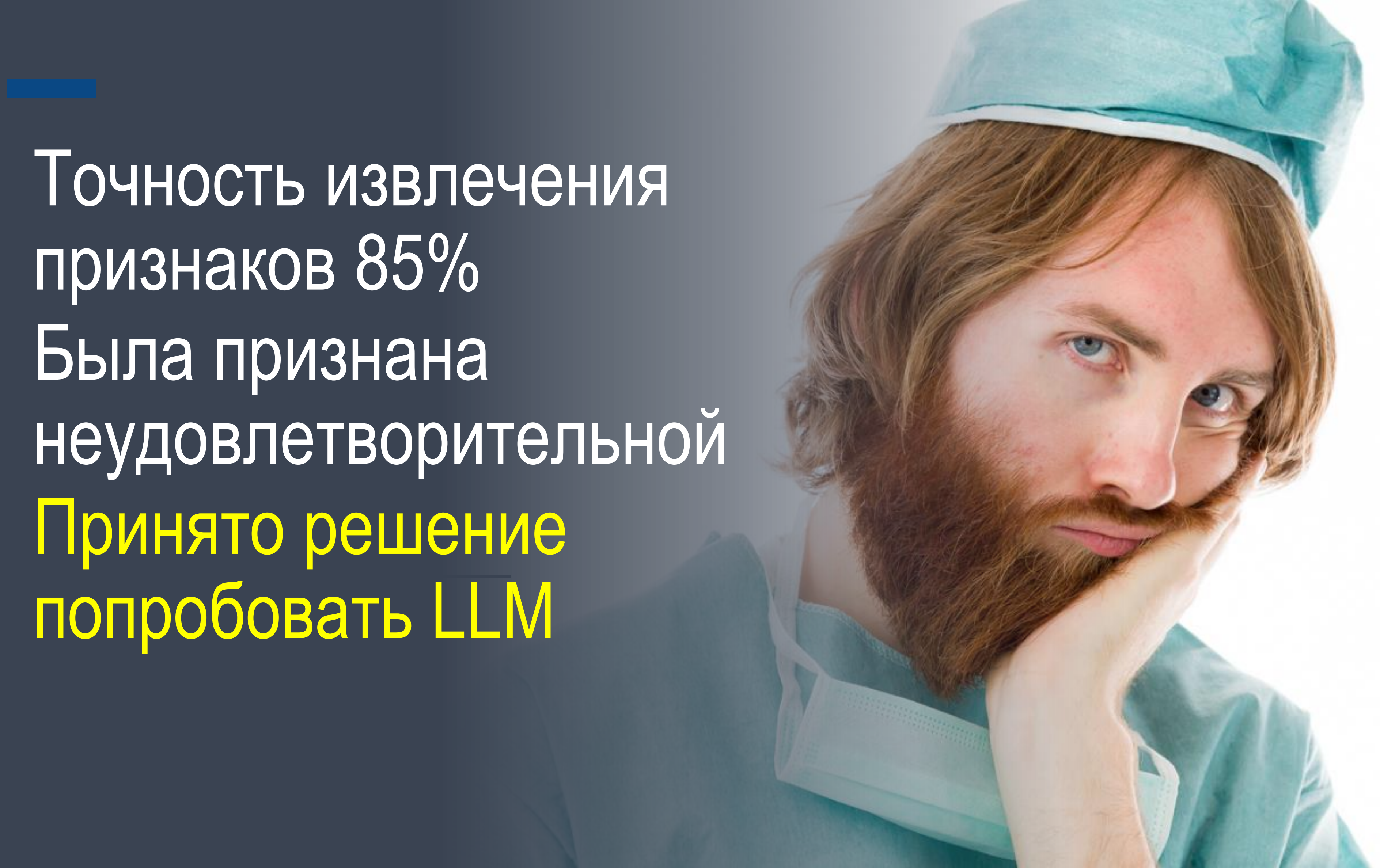
Пример (2)

История заболевания: беспокоит не помнит сколько , после прививки какой не помнит , когда сделали прививку не помнит , травмы не было. Со слов пациента находится в алкогольном опьянении (сколько употребила вопрос 2 банки пиво), предложена помощь психиатра нарколога отказалась и ушла

Ответ модели: False (ошибка)

* Для выполнения задачи использовались подходы на основе эмбедингов.

Точность извлечения
признаков 85%
Была признана
неудовлетворительной
Принято решение
попробовать LLM

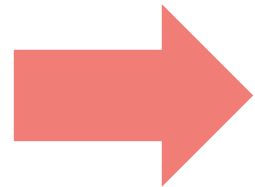


Эксперимент: изучить эффективность больших языковых моделей 8

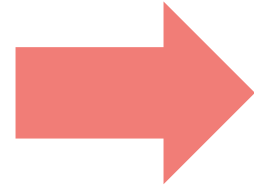
для извлечения сложных смыслов из текста (1)

Анализ текста с помощью NLP-моделей

Текстовая
запись



NLP-
модель



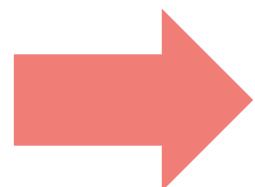
Список
признаков



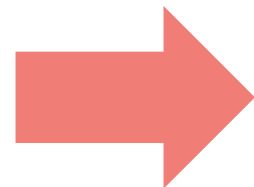
Извлекает поименные сущности
(например, значение роста, веса или АД) или **факты** (например, симптомы из жалоб)

Анализ текста с помощью LLM-модели

Запрос
(prompt)



Большая языковая
модель (LLM)



Ответ +
объяснение



Генерирует ответ на основе заданного вопроса и включенной в него текстовой записи. Ответ LLM-модели можно конвертировать в значение извлеченного признака

Эксперимент: изучить эффективность больших языковых моделей для извлечения сложных смыслов из текста (2)

9



Исходный текст

*На приеме к\я по приглашению для сдачи анализов и прохождения Флг. **Не отрицает** , что 2 дня назад **употреблял алкоголь** сожительницей. Вне абстиненции, **трезв** .*

Вопрос к Gema2

prompt = f'Можно ли сказать по этому тексту "{text}" что человек в настоящий момент пьян? Поясни ответ'

Ответ модели

*Нет, по данному тексту нельзя сказать, что человек в настоящий момент **пьян***

Объяснение

**"Вне абстиненции, трезв": Этот фрагмент указывает на то, что человек в данный момент не испытывает симптомов отмены алкоголя (абстиненции) и чувствует себя трезвым.*

"2 дня назад употреблял алкоголь сожительницей": Это говорит о том, что последний раз он пил два дня назад. В большинстве случаев, спустя два дня после употребления алкоголя, его влияние на организм уже исчезает.

Важно:

** Текст не дает информации о количестве выпитого алкоголя и индивидуальной реакции человека.*

** Возможность того, что человек все еще испытывает некоторое влияние алкоголя через два дня после приема маловероятна, но исключать ее полностью нельзя.*

Чтобы достоверно определить состояние человека в настоящий момент, необходимо использовать более точные показатели, такие как анализ крови на содержание алкоголя.

Технический способ решения задачи	Затраты человеко-часов на создание модели
NLP: 1 этап (базовый техпроцесс)	≈ 40 часов
NLP: 2 этап (тонкий тюнинг)	≈ 40 часов
LLM: Gemma2	≈ 10 минут
LLM: saiga_llama3	≈ 10 минут



Решение задачи с помощью LLM выигрывает с огромным отрывом по сравнению с обычными NLP. Фактически, даже LLM общего назначения может хорошо анализировать ЭМК

Экспериментальные данные. Время извлечения 1 признака

Технический способ решения задачи	Признак	Средняя скорость обработки, сек/запись
NLP	Злоупотребление алкоголем	0.053
NLP	Алкогольное опьянение	0.051
LLM: Gemma2	Злоупотребление алкоголем	96.156
LLM: Gemma2	Алкогольное опьянение	94.658
LLM: saiga_llama3	Злоупотребление алкоголем	96.156
LLM: saiga_llama3	Алкогольное опьянение	94.658



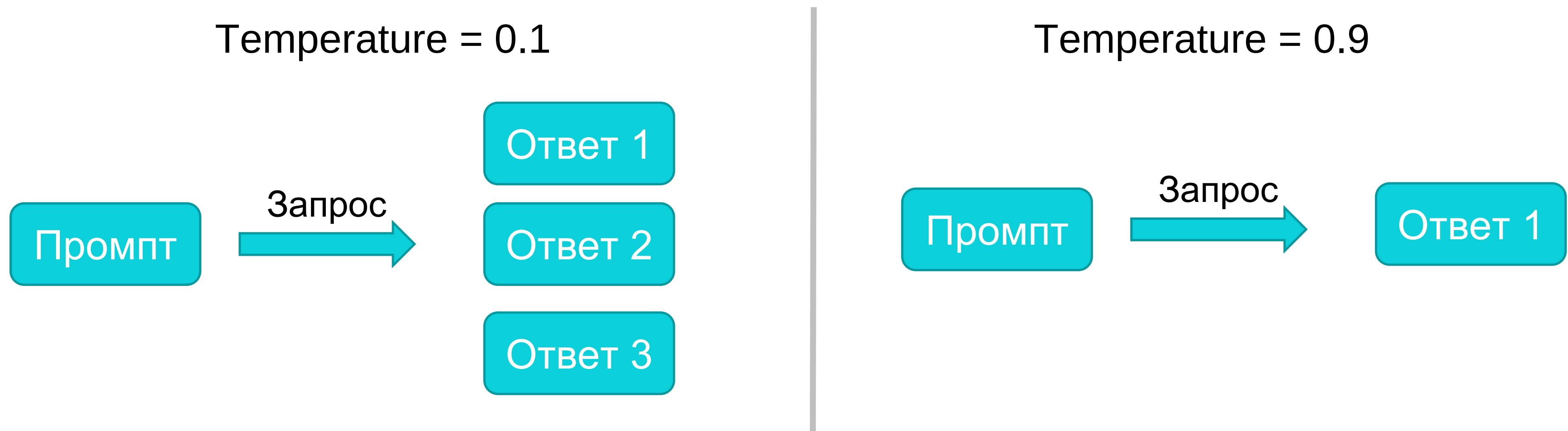
Решение задачи с помощью LLM существенно проигрывает по производительности по сравнению с NLP. Для устранения этого недостатка нужны мощные и дорогие GPU

Экспериментальные данные. Влияние «температуры» в LLM

Температура — важный параметр, который можно настроить в LLM.

Чем **ниже** значение температуры, тем более **детерминированными** будут результаты.

Детерминируемость результатов состоит в том, что при генерации будет выбран самый вероятный следующий токен. Увеличение температуры приводит к большей случайности, что способствует более разнообразным или творческим результатам.



Для повышения точности работы LLM до уровня, пригодного для реальной клинической практики, требуется аккуратное использование и такая же проверка качества, как и в случае NLP

Было предложено 2
сценария:

1. Использовать
«маленькие» LLM
2. Вернутся к более
глубокой проработки
привычной NLP-
модели





«Большие» LLM
(20 и более миллиардов параметров)

- ✓ Требуют значительных вычислительных мощностей, превышающих возможности обычного ПК
- ✓ Более «умные» - понятнее отвечают на вопросы, генерируют текст и т.д.
- ✓ Обучены на триллионах токенах

Параметр	Gemma 2 27B	Gemma 2 9B
Рекомендуемая видеопамять	40 Гб	24 Гб
Пример видеокарты	Tesla A100	NVIDIA RTX 4090
Цена видеокарты	≈ 1 млн. руб.	≈ 240 тыс. руб.

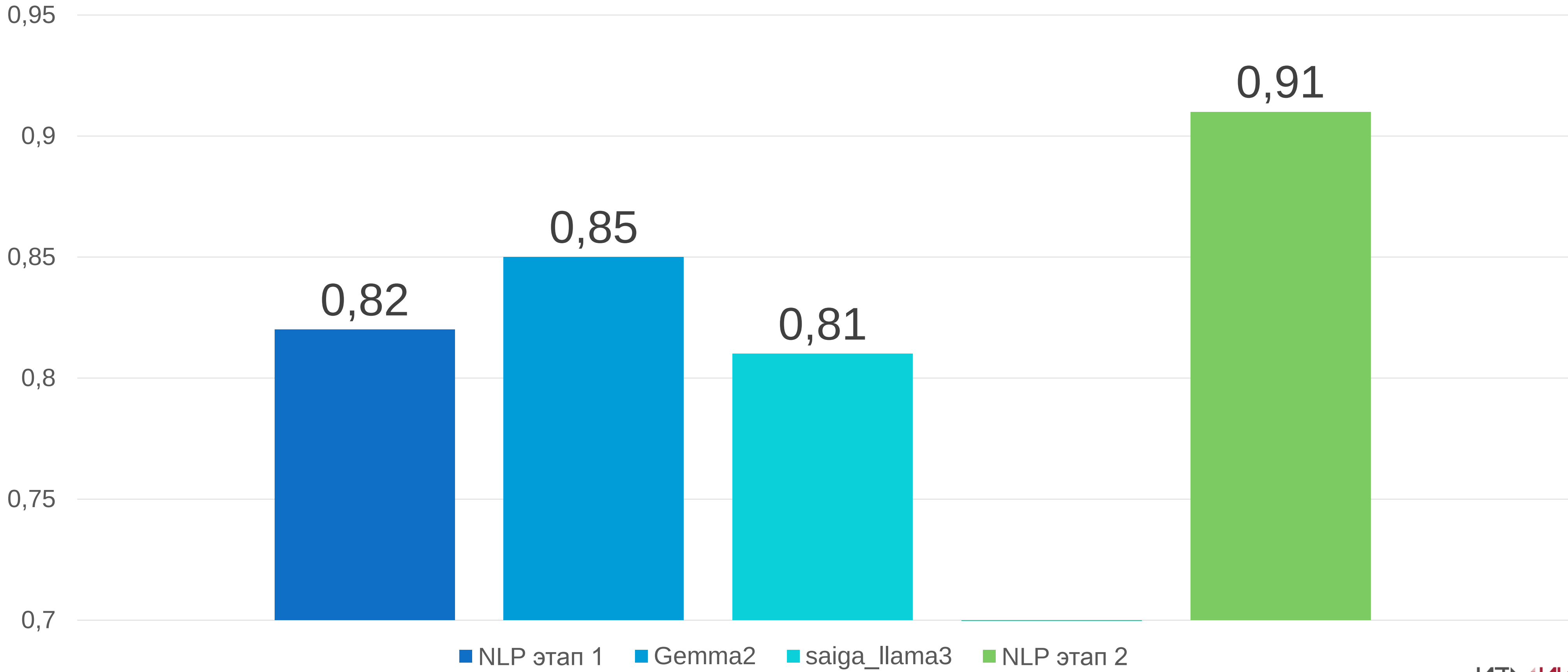
«Маленькие» LLM
(до 12-20 миллиардов параметров)

- ✓ Требуют меньше вычислительных мощностей, можно запускать и использовать на мощном ПК
- ✓ Как правило используются для узких задач
- ✓ Могут создаваться на базе «больших» LLM путем их дообучения

Бенчмарк	Метрика	Gemma 2 27B	Gemma 2 9B
MMLU	5-shot, top-1	75,2	71,3
BBH	3-shot, CoT	74,9	68,2
HellaSwag	10-shot	86,4	81,9

Сценарий №2 – развитие привычного подхода на основе NLP-модели и более глубокого тюнинга

F1-score по признаку «Злоупотребление алкоголем»



Выводы из опыта команды Webiomed

✓ Главные плюсы LLM в анализе медицинских текстов:

- Имитация что LLM «понимает» текст
- Возможность получить объяснение ответа, что в классических NLP практически невозможно

✓ Главные минусы LLM в анализе медицинских текстов:

- Необходимость дорогостоящих GPU для дообучения базовых LLM и их последующей эксплуатации, что в несколько раз увеличивает стоимость инфраструктуры и снижает экономическую эффективность
- Гораздо медленная работа (в несколько раз)

✓ Маленькие LLM вполне хорошо справляются с пониманием сырых медицинских документов «из коробки» со средней метрикой $f1 = 0.825$

✓ Не стреляйте из пушки по воробьям: LLM – хороший инструмент для работы с текстами, но затратный. Не стоит использовать его для задач, которые можно решить классическими методами NLP, потому что это не оправдано как по скорости / стоимости работы, так и по риску галлюцинаций. Для применения в SaMD нужно дообучение LLM на релевантном тексте

✓ LLM перспективны для работы с объёмными, сложными и противоречивыми текстами: аннотировать, классифицировать, извлекать ключевые моменты, строить логические выводы в полуавтоматическом режиме, а также для диалогового ИИ (чат-боты, диалоговый интерфейс)

Заключение команды Webiomed

Мы не видим целесообразность применения LLM в задачах извлечения признаков из ЭМК. Это слишком рискованно (галлюцинации, точность), затратно и медленно на потоке данных

Более интересным выглядит задача поддержки и ускорения процесса разработки NLP-моделей на основе новых подходов: аннотирование текстов, поддержка разработчика, обобщение накопленных данных и т.д.



Спасибо за внимание!



WEBIOMED

**Есть вопросы?
Пожалуйста, обращайтесь
по контактам ниже:**

Мои контакты

ГУСЕВ АЛЕКСАНДР

к.т.н., директор по развитию ООО «К-Скай»

 agusev@webiomed.ru

 +7 (911) 402-35-00



Контакты компании



Сайт

<https://webiomed.ru>

Мы в социальных сетях

ВКонтакте



<https://vk.com/webiomed>

Telegram



<https://t.me/webiomed>