

Перспективы использования больших языковых моделей (LLM), внешних источников знаний и предобученных нейронных сетей в медицинской практике

Докладчик: Ваганов Дмитрий Валерьевич,

инженер компании «Сириус-Софт», преподаватель технических дисциплин

(ТОГУ, Хабаровский институт инфокоммуникаций)

План

1. Общие сведения о больших языковых моделях
2. Проблема галлюцинаций в больших языковых моделях и качество представления знаний
3. Архитектура системы поддержки принятия медицинских решений на основе LLM
4. Перспективы использования новых классов нейросетей в области медицинской диагностики
5. Программные разработки образовательного проекта Vertexprize в области кардиологии.

Общие сведения о больших языковых моделях (Large Language Model - LLM)

Большие языковые модели — это искусственная нейронная сеть, построенная на специализированной архитектуре — трансформерах

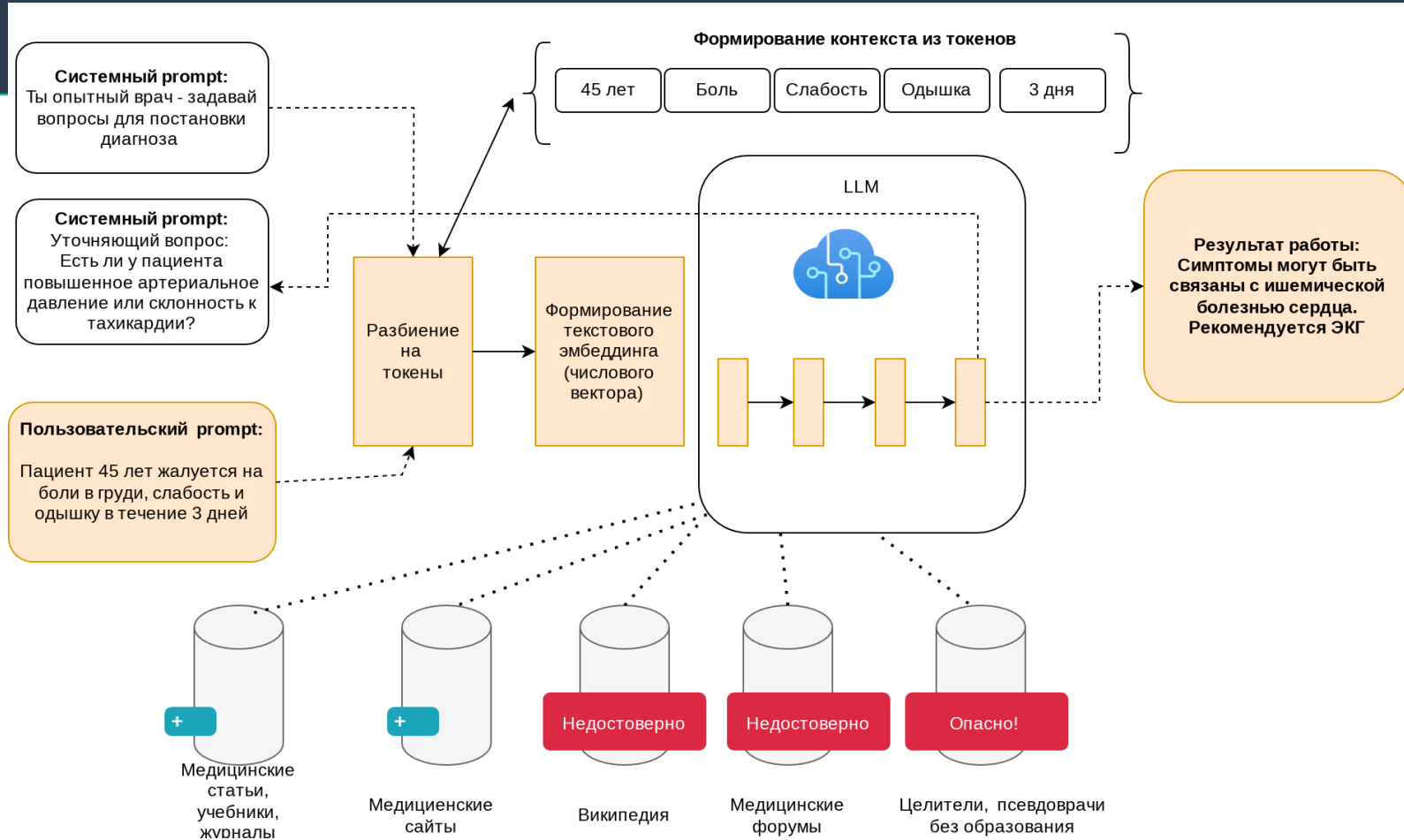
Основная задача трансформера:

преобразовании текстовых данных в контекстно-зависимое представление с использованием числового представления слов в многомерном пространстве.

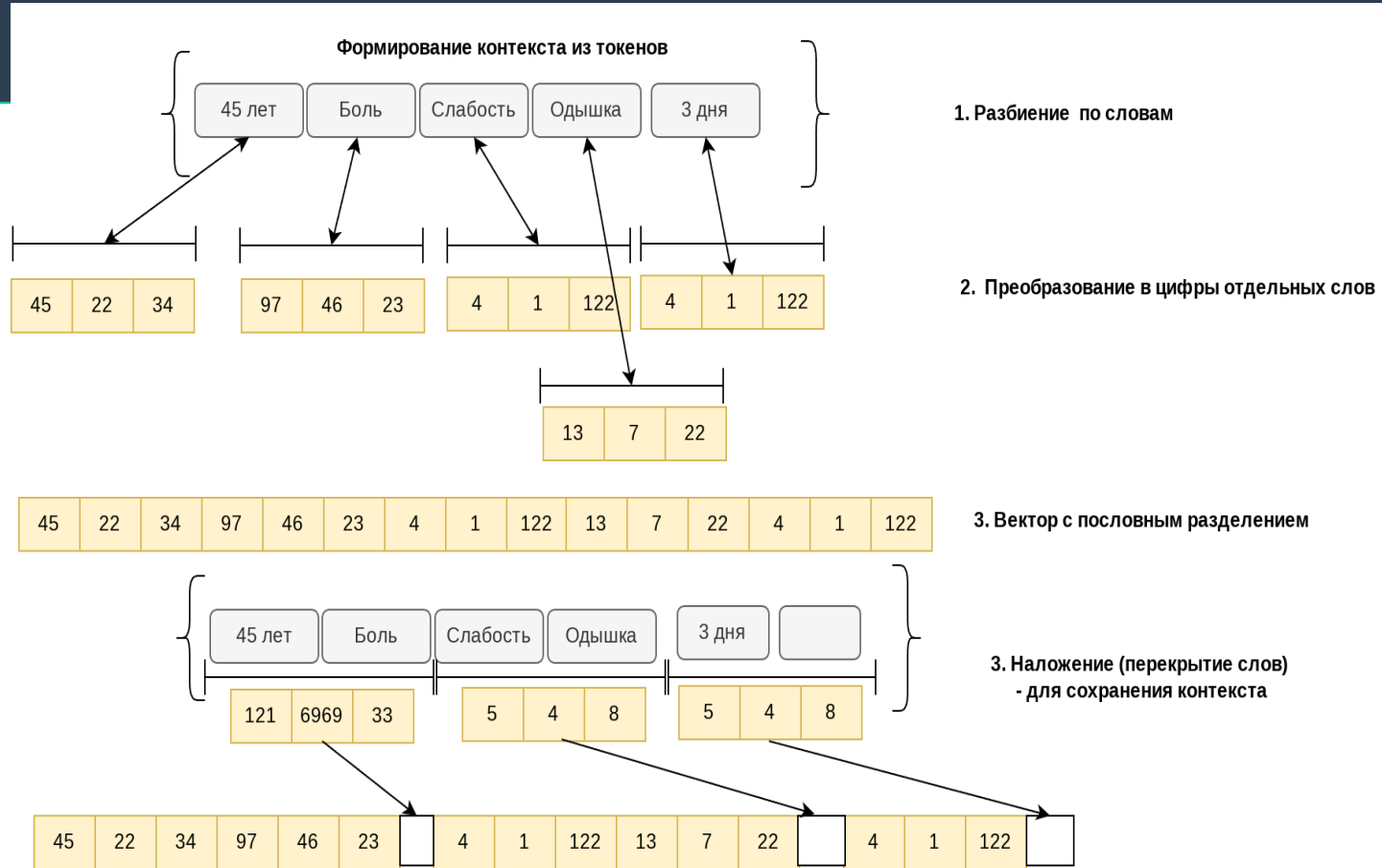
Функции трансформеров:

Поиск глобального контекста
Моделирование взаимосвязей между любыми частями входной последовательности символов

Общая схема работы LLM (на примере медицинской консультации)



Разбиение на токены и формирование эмбединга



Векторные представления слов и эмбединги LLM

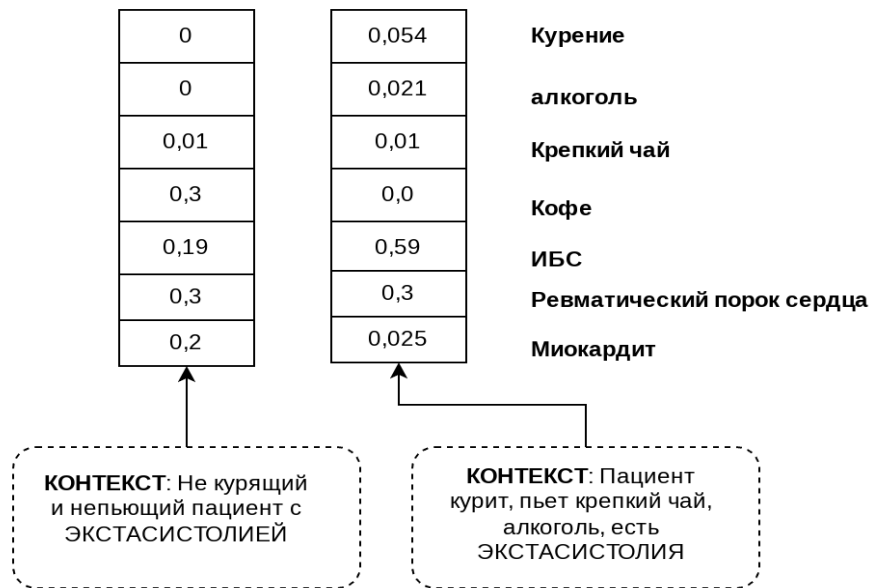
Пример двоичного вектора ЭКСТРАСИСТОЛИЯ

9 мерный вектор двоичный вектор
(не устанавливает связи между компонентами)

0	Синусовая тахикардия
0	Синусовая брадикардия
0	Синусовая аритмия
1	Экстрасистолия
0	Пароксизмальная тахикардия
0	Трепетание предсердий
0	Мерцательная аритмия
0	Острый инфаркт
0	Подострая стадия инфаркта миокарда

Пример эмбединга ЭКСТРАСИСТОЛИЯ

Смена эмбединга в зависимости от контекста
(присутствует связь между компонентами)



Алгоритм работы LLM: 1. Формирование входных данных

Входные данные:

Системный prompt: задаёт общие правила работы модели: контекст, который описывает её роль, задачу и формат вывода (например: Вы медицинский консультант, оценивающий симптомы для постановки диагноза. Запрашивайте недостающую информацию у пользователя).

Пользовательский prompt: описывает конкретный случай, включая симптомы, анамнез и жалобы пациента (например: Пациент 45 лет жалуется на боли в груди, слабость и одышку в течение 3 дней).

Вход модели = **Системный prompt + Пользовательский prompt:**

Функции трансформера:

1. Разбиение на токены: (объединённый prompt) сначала токенизируется — текст разбивается на токены.

2. Начальное формирование цифровых векторов: Каждый токен преобразуется в эмбединг через заранее обученную эмбединг-матрицу. Это создаёт начальное представление текста в виде числовых векторов.

Алгоритм работы LLM: 2. Обработка входных данных в слоях трансформера

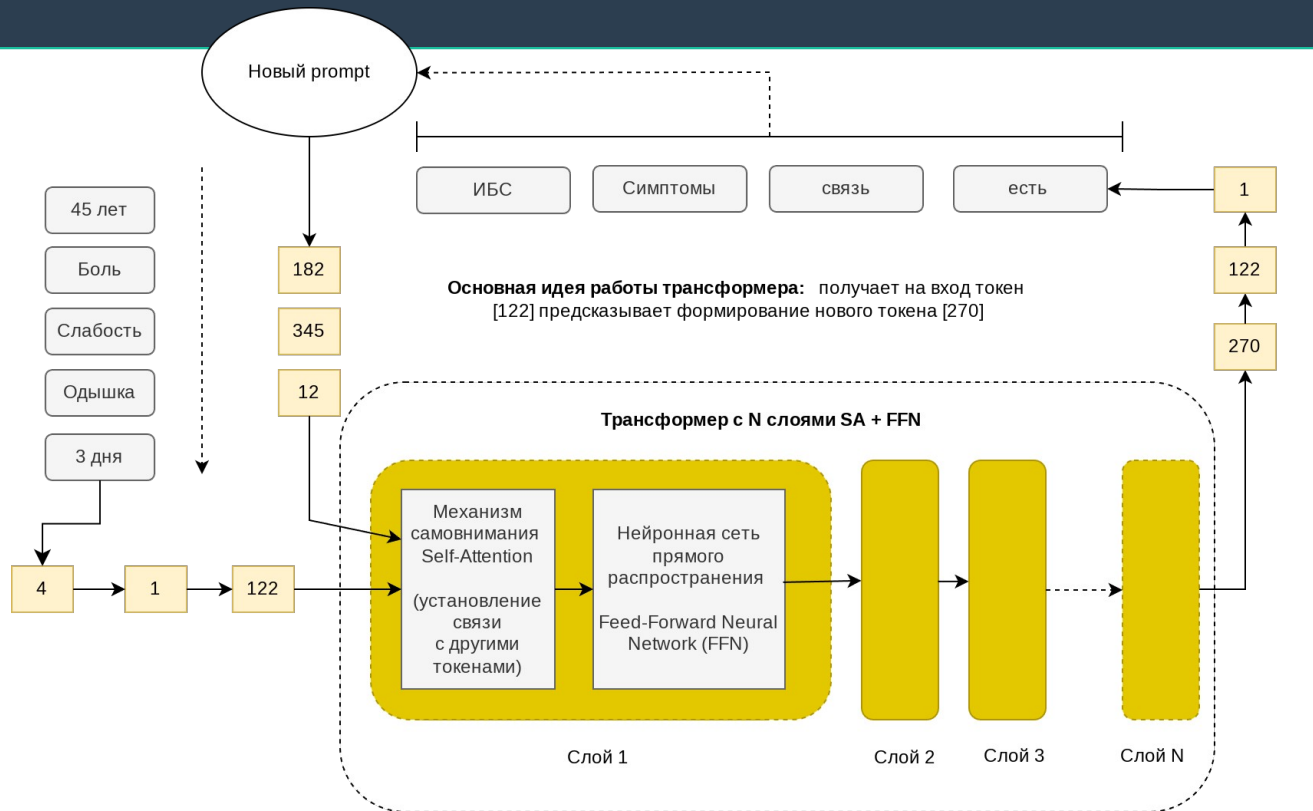
Функции трансформера:

1. Модель последовательно обрабатывает токены с помощью нескольких слоёв. Каждый слой трансформера включает:

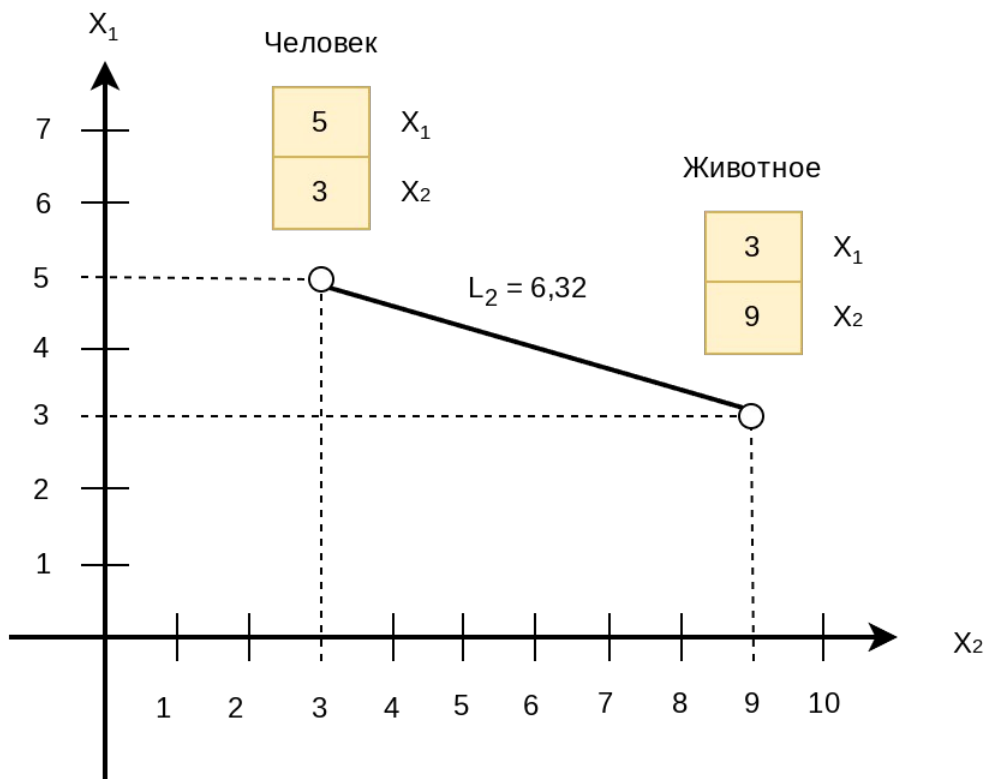
Self-Attention механизм : вычисляет, насколько каждый токен связан с другими токенами в тексте. Это позволяет учитывать контекст (например, при анализе фразы «боль в груди» важно понимать, связана ли она с одышкой или другими симптомами). Механизм Self-Attention преобразует каждое токен-эмбеддинг в новое, "контекстно-зависимое" представление, беря во внимание все токены в последовательности.

Feed-Forward Neural Network (FFN): применяется к каждому токenu индивидуально для их нелинейной трансформации. FFN работает как фильтр: каждый токен индивидуально дорабатывается, чтобы содержать более качественное, нелинейно трансформированное представление. После FFN модель становится способной более точно решать задачи, такие как генерация текста, классификация или перевод.

Алгоритм работы LLM: 2. Алгоритма работы трансформера



Выражение семантического сходства через векторное расстояние (L2) для слова «боль»



Формула вычисления L_2 :

$$L_2 = \sqrt{(x_{1ч} - x_{1ж})^2 + (x_{2ч} - x_{2ж})^2} = \sqrt{(5 - 3)^2 + (3 - 9)^2} = 6,32$$

Выражение семантического сходства через векторное расстояние (L2) для слова «боль»

Как LLM выражает семантическое сходство?

LLM на основе трансформеров обрабатывают текст, они представляют слова, фразы и предложения в виде числовых векторов в многомерном пространстве (например, в 768-мерном для стандартного BERT или 1024-мерном для GPT-3). Эти векторы кодируют семантическую информацию так, что близость между векторами отражает степень семантической или контекстуальной похожести между словами или предложениями.

Пример 1: "боль" в двух семантически схожих контекстах:

"Пациент жалуется на сильную боль в спине."

"Человек испытывает боль после травмы."

В этом случае слово "боль" окружено схожим контекстом, относящимся к медицинской или телесной боли.

Нейросеть обнаружит высокую семантическую близость между этим употреблением слова в обоих предложениях.

Векторы для слова "боль" будут находиться близко друг к другу в пространстве. $L2 = 0.8$ (где меньшие значения означают большую схожесть).

Пример 2: "боль" в кардинально разных контекстах:

"Боль не покидала его сознание после аварии."

(физическое состояние или медицинский термин).

"Боль потери навсегда осталась в её сердце."

(эмоциональное состояние).

Контекст вокруг слова влияет на его векторное представление, из-за чего векторы этих двух слов "боль" будут гораздо дальше друг от друга в пространстве. $L2=4,5$

Причины галлюцинаций при использовании LLM в качестве медицинского ассистента

1. Неадекватное обучение (недостаточное качество данных):

LLM обучаются на огромных наборах данных, которые могут включать ошибки, устаревшую информацию, персональные мнения или неточные медицинские факты. Это приводит к унаследованию систематических ошибок.

2. Смещение источников знаний:

Модели могут комбинировать разрозненные части данных из разных источников, приводя к появлению несоответствующей или ложной информации.

3. Обучение на общих текстах (например, интернет-страницы, статьи, книги) может привести к недостаточному пониманию медицинских концепций, что снижает точность ответов.

4. Контекстуальные ошибки: модель может запутаться в сложных медицинских контекстах, особенно если запрос плохо сформулирован или требует сложного анализа.

5. Стремление модели "угодить": LLM склонны генерировать ответы даже тогда, когда данных недостаточно или модель "не знает". Это связано с обучением на задачах предсказания текста, где правдоподобие ответа имеет приоритет.

Архитектура Дополненная генерация ответа Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) — это современная архитектура в области обработки естественного языка, которая сочетает в себе две ключевые технологии: извлечение информации и генерацию текста. Основная идея RAG заключается в том, чтобы улучшить качество текстовой генерации за счёт того, что модель имеет доступ к внешнему источнику информации, который помогает ей генерировать более точные и контекстуально релевантные ответы

Дополненная генерация ответа Retrieval-Augmented Generation (RAG) и системный prompt

Роль системного prompt в модели:

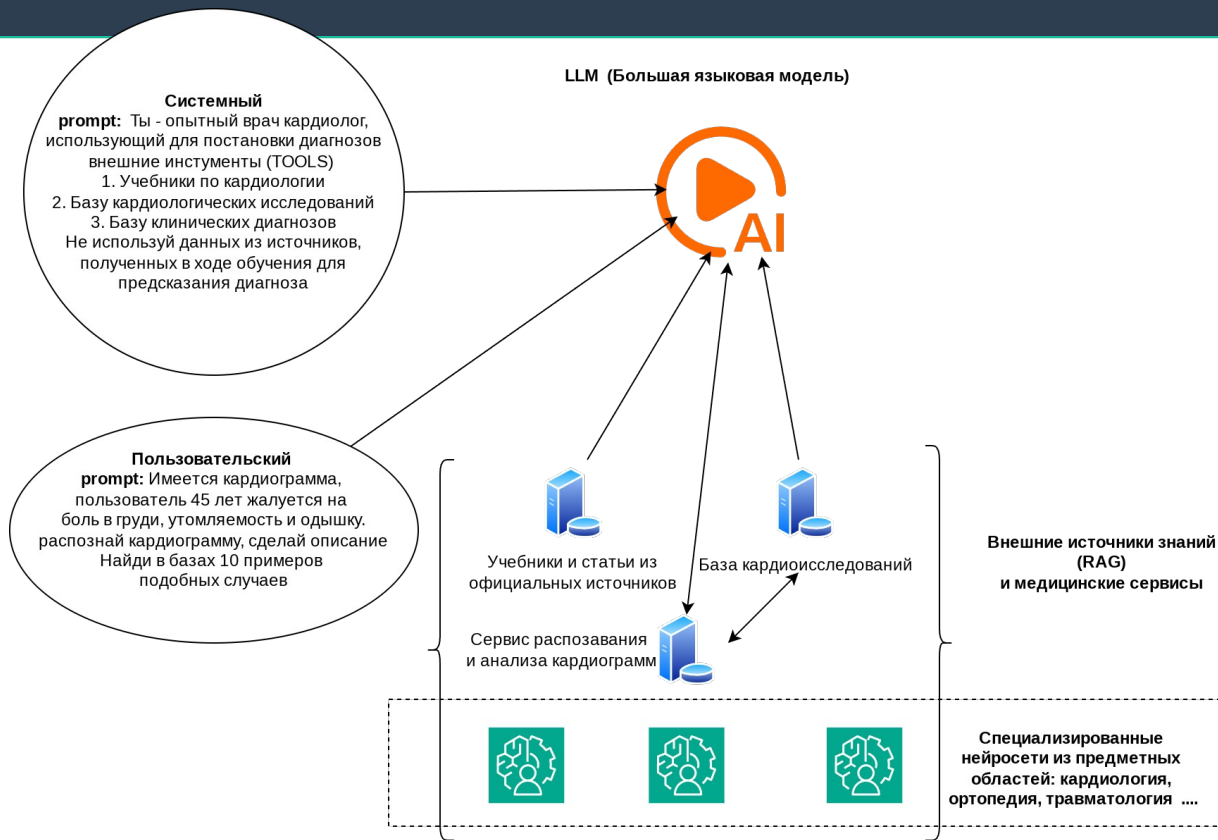
Системный prompt задаёт общий тон и поведение медицинского ассистента, определяя его границы, рамки и цели. С его помощью можно:

1. Прописать чёткие инструкции о том, как модель должна формулировать ответы (например, на основе медицинских протоколов);
2. Напомнить модели о необходимости быть осторожной в трактовке данных, ограничивать выводы, если информация недостаточна
3. Выдать результат работы: НЕТ ОТВЕТА НА ВОПРОС — НЕДОСТАТОЧНО КОНТЕКСТНОЙ ИНФОРМАЦИИ.

Пример системного prompt для медицинского ассистента:

> "Вы являетесь медицинским помощником, обученным предоставлять точные и обоснованные медицинские ответы. Всегда опирайтесь только на данные, извлечённые из аутентичных медицинских источников. Если информация отсутствует или неоднозначна, сообщите об этом и рекомендуйте обратиться к врачу. Всегда соблюдайте формальный стиль и избегайте личных интерпретаций."

Архитектура медицинского ассистента с использованием RAG



Как и почему использование RAG улучшает качество работы модели

RAG (Retrieval-Augmented Generation) дополняет работу языка модели, извлекая из внешних баз данных достоверную информацию перед генерацией ответа.

Пример в медицинской области: извлечение данных из таких источников, как PubMed (США) или RusMed (Россия) для ответа на вопросы пользователей.

Ключевые преимущества интеграции RAG:

Актуальность: Модель получает доступ к информации о новейших исследованиях и рекомендациях, в отличие от завершённого набора обучающих данных.

Достоверность: Система ссылается на извлечённые фрагменты текста, что снижает шанс на галлюцинации модели.

Преимущества RAG и корректно составленного system prompt : управление внешними запросами

Системный prompt управляет тем, как модель формирует запросы к источникам RAG (то есть поисковые запросы, исходя из вопроса пользователя).

Пример:

Если пользователь спрашивает о лекарстве: Системный prompt может направить модель к запросу информации по базе FDA или актуальных рекомендаций ВОЗ.

Если вопрос общеклинический: Модель может базироваться на официальных медицинских рекомендациях.

Пример работы с RAG:

> Пользователь спрашивает: "Какие побочные эффекты имеют препараты для лечения гипертонии?"

Системный prompt настраивает формат запроса: "Извлечь из базы данных точные побочные эффекты, упомянутые в клинических рекомендациях для лечения гипертонии."

RAG возвращает релевантные данные из базы.

Генеративная (LLM) модель **использует эти данные и структурирует ответ.**

Преимущества RAG и корректно составленного system prompt : управление внешними запросами

Системный prompt управляет тем, как модель формирует запросы к источникам RAG (то есть поисковые запросы, исходя из вопроса пользователя).

Пример:

Если пользователь спрашивает о лекарстве: Системный prompt может направить модель к запросу информации по базе FDA или актуальных рекомендаций ВОЗ.

Если вопрос общеклинический: Модель может базироваться на официальных медицинских рекомендациях.

Пример работы с RAG:

> Пользователь спрашивает: "Какие побочные эффекты имеют препараты для лечения гипертонии?"

Системный prompt настраивает формат запроса: "Извлечь из базы данных точные побочные эффекты, упомянутые в клинических рекомендациях для лечения гипертонии."

RAG возвращает релевантные данные из базы.

Генеративная (LLM) модель **использует эти данные и структурирует ответ.**

Преимущества использования RAG и корректно составленного system prompt при построении медицинских рекомендательных систем

- 1. Предупреждение "галлюцинаций"** : Наличие системного prompt заставляет модель генерировать ответы только на основе данных, которые были извлечены из доверенного источника. Если нужная информация отсутствует, модель будет указана на это, вместо попытки "угадать" ответ.
- 2. Стандартизация ответов:** Системный промпт помогает задавать согласованный стиль и формат ответа, что важно при общении в медицинской практике (например, использование формата "причина-признаки-лечение"; ссылка на безопасные источники) в формате, принятом в лечебных учреждениях.
- 3. Уверенность в ограничениях.** Системный промпт ставит акцент на важности консультации врача в случае сложных медицинских вопросов. Например:

> "На основании представленных данных XYZ побочные эффекты связанные с препаратом А включают головокружение и тошноту. Однако перед началом лечения проконсультируйтесь с врачом, чтобы учесть вашу индивидуальную медицинскую историю."

Выводы:

1. Современные **LLM**, изначально придуманные как средство генерации текста, ответа на вопросы в областях знаний, заложенных в LLM **могут использоваться при построении медицинских рекомендательных систем при соблюдении следующих условий:**
 - непроверенные и недостоверные сведения, полученные на этапе обучения LLM блокируются правилами, заложенными в system prompt
 - непроверенные и потенциально недостоверные источники информации заменяются проверенными данными с использованием технологии RAG.
2. Медицинские рекомендательные системы должны быть дополнены внешними узкоспециализированными системами в числе которых: медицинские базы данных, включая истории болезни, специализированные нейронные системы (предобученные нейронные сети) для автоматизации процессов постановки диагнозов в специальных медицинских областях.
3. Возможности LLM не ограничиваются обработкой текстов: эмбединги позволяют **ОДИНАКОВО (ОДНОТИПНО)** работать с кардиограммами, графическими изображениями МРТ и другими источниками, поскольку LLM внутри трансформера LLM не различает эти виды данных.

2024 год — Прорыв в развитии нейростей (тихая, никем не замеченная революция) - появление нового класса сетей, основанных на работах Kolmogorov-Arnold Networks - KAN — сети названы в честь русских ученых Колмогорова Андрея Николаевича (1903-1987), Арнольда Владимира Игоревича (1937 -2010), доказавших великую теорему, которая утверждает:

сложные многомерные функции могут быть разложены на более простые одномерные функции, полагая основу для уникальной структуры нейросети KAN

Доказанная теорема решает проблему, которая носит название **«ПРОКЛЯТИЕ РАЗМЕРНОСТИ»** Проклятие размерности в нейронных сетях — это проблема, возникающая при работе с данными, у которых очень много признаков или параметров.

В низкоразмерных пространствах (например, с двумя или тремя признаками) данные располагаются близко к друг другу, поэтому их легко проанализировать. С увеличением числа измерений (размерностей) объём пространства данных растёт экспоненциально.

Суть проклятия: с ростом числа измерений анализировать данные становится **НЕВОЗМОЖНО!**

Перспективы использования новых классов нейросетей в области медицинской диагностики

2024 год — Прорыв в развитии нейростей (тихая, никем не замеченная революция) - появление нового класса сетей, основанных на работах Kolmogorov-Arnold Networks - KAN — сети названы в честь русских ученых Колмогорова Андрея Николаевича (1903-1987), Арнольда Владимира Игоревича (1937 -2010), доказавших великую теорему, которая утверждает:

сложные многомерные функции могут быть разложены на более простые одномерные функции, полагая основу для уникальной структуры нейросети KAN

Доказанная теорема решает проблему, которая носит название **«ПРОКЛЯТИЕ РАЗМЕРНОСТИ»** Проклятие размерности в нейронных сетях — это проблема, возникающая при работе с данными, у которых очень много признаков или параметров.

В низкоразмерных пространствах (например, с двумя или тремя признаками) данные располагаются близко к друг другу, поэтому их легко проанализировать. С увеличением числа измерений (размерностей) объём пространства данных растёт экспоненциально.

Суть проклятия: с ростом числа измерений анализировать данные становится **НЕВОЗМОЖНО!**

Решение проблемы проклятья размерности с использованием KAN сетей

Преимущества сетей KAN:

1. Более высокая точность
2. Меньший размер сети (требует меньше вычислительных ресурсов)

Недостатки сетей KAN:

Очень большое время обучения

Kolmogorov



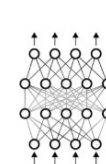
+

Arnold



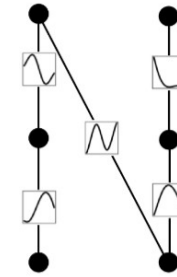
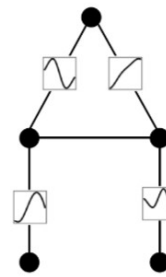
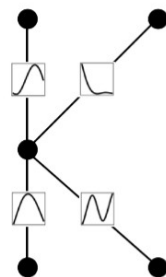
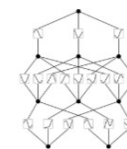
+

Network



=

KAN

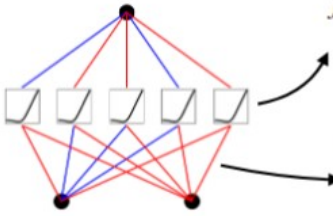
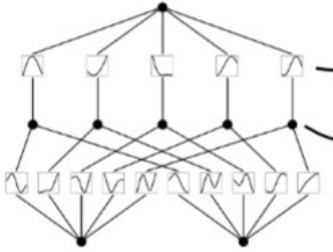


Mathematical

Accurate

Interpretable

Отличие традиционных ИНС от сетей KAN

Multi-Layer Perceptron (MLP)	Kolmogorov-Arnold Network (KAN)
<p>Universal Approximation Theorem</p>	<p>Kolmogorov-Arnold Representation Theorem</p>
$f(\mathbf{x}) \approx \sum_{i=1}^{N(\epsilon)} a_i \sigma(\mathbf{w}_i \cdot \mathbf{x} + b_i)$	$f(\mathbf{x}) = \sum_{q=1}^{2n+1} \Phi_q \left(\sum_{p=1}^n \phi_{q,p}(x_p) \right)$
<p>(a)</p>  <p><i>fixed activation functions on nodes</i></p> <p><i>learnable weights on edges</i></p>	<p>(b)</p>  <p><i>learnable activation functions on edges</i></p> <p><i>sum operation on nodes</i></p>
$\text{MLP}(\mathbf{x}) = (\mathbf{W}_3 \circ \sigma_2 \circ \mathbf{W}_2 \circ \sigma_1 \circ \mathbf{W}_1)(\mathbf{x})$	$\text{KAN}(\mathbf{x}) = (\Phi_3 \circ \Phi_2 \circ \Phi_1)(\mathbf{x})$

Суть обучения: изменение весов в матрицах связей

Весы связей (числа) заменяются обучаемыми функциями, расположенными на ребрах

Применение KAN сетей в медицине:

Высокая точность положительно влияет на точность постановки диагнозов в медицинских рекомендательных системах

Меньшие требования к вычислительным ресурсам открывают перспективы встраивания диагностических непосредственно в носимое и (или) малогабаритное медицинское оборудование

Перспективы: Нейросети как универсальный инструмент саморегулирования в сложных системах, включая медицинские и технические системы



AIP Publishing AIP Conference Proceedings

HOME BROWSE ▾ FOR AUTHORS ▾ FOR ORGANIZERS ▾ ABOUT ▾

Volume 2402, Issue 1
15 November 2021

RESEARCH ARTICLE | NOVEMBER 15 2021

Neural networks with radial-basis functions for forming of soil profile in pipeline's cathodic protection system

Sergei Nikulin ; Dmitriy Vaganov

+ Author & Article Information

AIP Conf. Proc. 2402, 0700339 (2021)

<https://doi.org/10.1063/5.0071832>

Share ▾

Tools ▾

Anti-corrosion protection of steel buried pipelines is effected by means of cathodic protection system and special protective insulation coatings. It has been detected that the best control option for the whole assembly of cathodic protection stations on a certain pipeline section is based on development

AIP Publishing — это полностью некоммерческая дочерняя компания Американского института физики (AIP). Портфолио включает в себя высоко оцененные, рецензируемые журналы, включая растущий портфель изданий открытого доступа, которые охватывают все области физических наук.

Нейронные сети с радиально-базисными функциями для формирования грунтового профиля систем катодной защиты трубопроводов
Авторы: Сергей Никулин, Дмитрий Ваганов, ноябрь 2021 г.

<https://doi.org/10.1063/5.0071832>