



Центр практического искусственного интеллекта

Sber AI Lab

Agenda

- Sber AI Lab: intro
- Эмоциональный искусственный интеллект (А. Савченко)
- Рекомендательные системы (А. Васильев)
- Мультиагентные системы и DS-агент (М. Митрович)
- LLM для структурированных данных (М. Макаренко)
- Трансформеры для последовательностей событий (И. Карпухин)
- Искусственный интеллект в медицине (Н. Романенко)
- Курсовые и дипломные работы

Соединяем науку и бизнес

Технологии для прикладных бизнес-задач с научной составляющей

Мы работаем на стыке бизнеса и науки. Создаем наукоемкие технологии, которые приносят пользу не только нашей компании, но и миру в целом. Наша работа над перспективными направлениями исследований оказывает большое влияние на стратегию развития бизнеса и технологий в Сбере. Мы публикуем научные статьи на ключевых мировых конференциях, помогаем бизнесу внедрять новые технологии и вместе с ним двигаем науку вперед



Глеб Гусев

Co-leader центра

Основные текущие направления

- Алгоритмы и технологии для работы со структурированными данными с помощью мультиагентных систем и фундаментальных моделей
- Алгоритмы ИИ для диалоговых сценариев типичных для индустрий Сбера
- Научные результаты и open source в приоритетных практических направлениях

100+

Статей опубликовано в журналах Q1 и на конференциях A/A*

10

Продуктов выложено в открытый доступ

70

Сотрудников

1 доктор наук

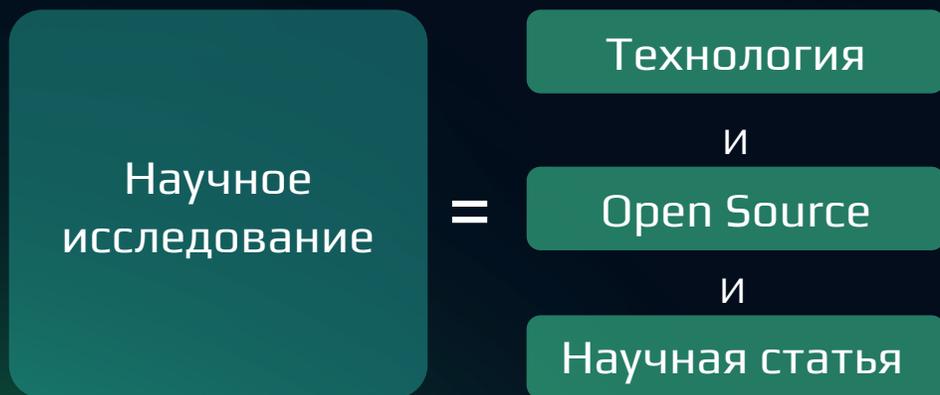
4 Kaggle мастера

8 Ph.D/канд. Наук

Преподаватели ВУЗов

Научно-технологический имидж Сбера

Научные публикации



Показатель	2024	2025 (сентябрь)
Конференции A* по рейтингу CORE	13	16
Конференции A по рейтингу CORE	12	8
Журналы Q1	8	10
Итого A*/A/Q1	33	34

Значимость научных публикаций A/A*/Q1

- Верификация научной новизны
- Сравнение с лучшими конкурентами
- Расширение области применения технологии

58%

В 2024 г. из 57 статей A/A*/Q1 Сбербанк – 33 от AI Lab

Исследования для core бизнеса

Supervised

- Автоматизация машинного обучения: [LAMA \(LightAutoML\)](#) - победитель 2024 AutoML Grand Prix
- Прогнозирование: [tsururu](#)
- Принятие решений: [PyBoost](#)

Semi-supervised

- Рекомендации: [RePlay](#)

Simulation & Gen AI

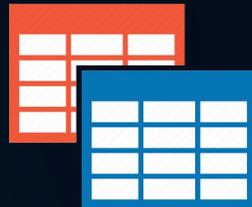
- Симулятор: [Sim4Rec](#)

Self-supervised

- Эмбединги на последовательностях: [pytorch-lifestream](#)
- Семантический матчинг: [pytorch-lifestream](#)



sequence



tabular



NLP
(call center dialogues)



time series, event
sequence



graphs,
knowledge graphs



Эмоциональный искусственный интеллект

Андрей Савченко

Научный директор лаборатории ИИ

С 20 лет - в разработке ПО, путь от стажера до тимлида

В 22 года пошел в аспирантуру Вышки в НН, заинтересовался распознаванием образов

В 24 года досрочно защитил кандидатскую диссертацию, вскоре ушел из индустрии и сосредоточился на науке

В 2016 г защитил докторскую диссертацию

Руководил проектами по CV от Samsung, Huawei, Сбер,...

Теперь я представляю:



Центр фундаментальных исследований: технологии

Эмоциональный ИИ

- **EmotiEffLib**: библиотека моделей распознавания и представления эмоциональных лиц
- **InsideOut**: эмоциональные мультиагентные ролевые модели



Мультимодальные последовательности

- **Цифровой аватар**: предсказание будущих транзакций
- Прогнозирование многомерных временных рядов



Эффективные алгоритмы компьютерного зрения

- **AutoFace** («AutoML за 10 минут»): персонализированный подбор нейросети под конкретное устройство
- **Распознавание видео**: real-time на мобильных устройствах – до 30 раз быстрее конкурентов без потерь в точности



Методы оптимизации

- Повышение скорости сходимости при обучении больших AI моделей
- Федеративное и распределенное обучение



Эмоциональный ИИ

Эмоциональный ИИ – агент, способный



Распознавать

эмоции, понимать намерения, мотивацию и желания людей



Управлять

своими реакциями и эмоциями людей в диалоге

Предложенный подход



EmotiEffNet

real-time распознавание выражений лиц по видео на типовых мобильных устройствах /ноутбуках без отправки персональной информации на GPU-сервер



Головоломка

использование LLM-агентов, концентрирующихся на базовых эмоциях, для генерации эмоциональных ответов в текстовых диалогах

Публикации



ICML 2023 (Core A*)
oral talk



IJCAI 2024 (Core A*) demo
track



ECAI (Core A)
demo track



NeurIPS 2024 (Core A*)



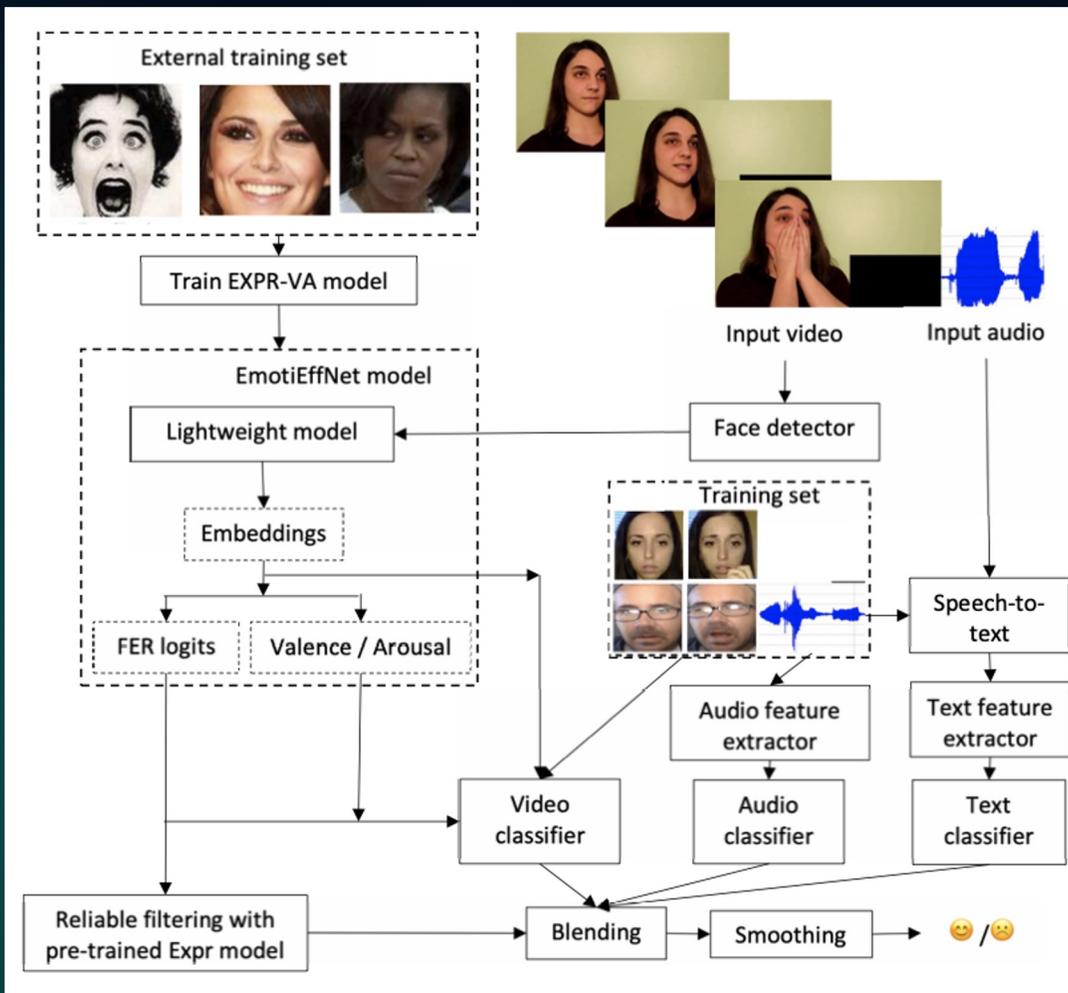
Призовые места в распознавании эмоций
на конкурсах Affective Behaviour Analysis in-
the-Wild при CVPR/ECCV

Мультимодальное распознавание ЭМОЦИЙ

ЭМОЦИЙ

Конкурс Affective Behavior Analysis in-the-Wild (ABAW)-8

EmotiEffLib



Expression Recognition

1 место (из 30 команд)

Teams	F1
HSEmotion	0.365
CtyunAI	0.363

Ambivalence/Hesitancy

1 место (из 9 команд)

Teams	Weighted F1
HSEmotion	0.710
HCAI-VIS	0.702

Action Unit Detection

3 место (из 27 команд)

Teams	F1
USTC-IAT-United	0.515
CtyunAI	0.502
HSEmotion	0.488
PR-VSL	0.484

Emotional Mimicry Intensity

3 место (из 6 команд)

Rank	Team	Best Result (PCC)
1	HCAI-VIS	0.71
2	USTC-IAT-United	0.68
3	HSEmotion	0.51

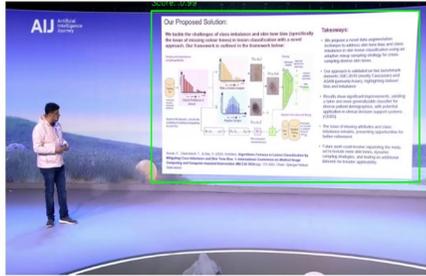


SpeakerAI Feedback

Multi-Lingual Audio-Visual Presentation Skills Assessment

Система обрабатывает видеоданные, синхронизирует данные по времени каждого слайда и формирует обратную связь по каждому слайду в виде структурированного отчёта в формате PDF:

- Анализ речи: использует WHISPER для многоязычной транскрипции и определения чёткости речи, темпа и слов-паразитов.
- Распознавание и анализ слайдов: объединяет Detectron2 для определения области слайда и QWEN-VL OCR для извлечения текста с экрана.
- Отслеживание взгляда: MediaPipe определяет направление глаз для оценки вовлечённости аудитории.
- Распознавание жестов: YOLOv8 фиксирует динамику движений говорящего.
- Распознавание эмоций: EmotiEffNet анализирует выражения лица для распознавания эмоций.
- Анализ одежды: модель CLIP оценивает уместность одежды для официальных мероприятий.
- Рекомендательный модуль: объединяет все показатели и генерирует персонализированную обратную связь по каждому слайду с помощью LLM



(a) English: Slide Region

(b) English: OCR Result

Slide Content

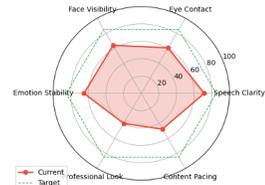
Our Proposed Solution: We tackle the challenges of class imbalance and skin tone bias (specifically the issue of missing color tones) in lesion classification with a novel approach. Our framework is outlined in the framework below.

Takeaways: - We propose a novel data augmentation technique to address skin tone bias and class imbalance in skin lesion classification using an adaptive mixup sampling strategy for cross-sampling diverse skin tones. - Our approach is validated on two benchmark datasets: ISIC-2018 (mostly Caucasian) and ASIAN (primarily Asian), highlighting dataset bias and imbalance. - Results show significant improvements, yielding a faster and more generalizable classifier for diverse patient demographics, with potential application in clinical decision support systems (CDSS). - The issue of missing attributes and class imbalance remains, presenting opportunities for further refinement. - Future work could involve expanding the meta-set to include more skin tones, dynamic sampling strategies, and testing on additional datasets for broader applicability. Ammar, F., Chakraborty, T., & Das, S. (2024, October). Algorithmic Fairness in Lesion Classification by Mitigating Class Imbalance and Skin Tone Bias. In International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2024) (pp. 373-382). Cham: Springer Nature Switzerland.

Word Count: 193

Presentation Analysis Dashboard

A) Performance Radar



B) Key Statistics

PRESENTATION SUMMARY

- Total Slides: 35
- Total Time: 1966.2s (32.8 min)
- Avg/Slide: 56.2s
- Start: 0.0s
- End: 1966.2s

Filler Rate: 2.3%

Top Fillers: uh(18), Like(11), okay(9)

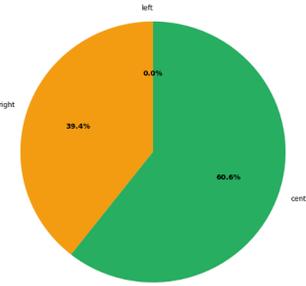
Center Gaze: 60.6%

Emotion: Moderately Stable

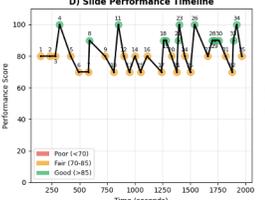
Appearance: Needs Work

Improvement Score: 67%

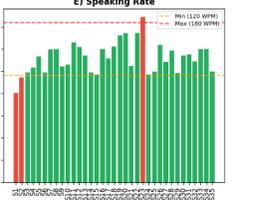
C) Gaze Distribution



D) Slide Performance Timeline



E) Speaking Rate



EAI: эмоциональные решения LLM

Фреймворк для оценки alignment'a решений людей и LLM

- Добавление эмоциональных промптов в LLM
- Моделирование поведения людей в стратегических играх («Диктатор», «Ультиматум», «Дилемма заключенного», «Битва полов»,...)
- Оценка рисков неэтичных ответов в TrustLLM

Результаты

- Стратегия LLM слишком рациональна и не соотносится с человеческой
- Добавление эмоций демонстрирует лучший alignment с поведением человека

Model	Offered share		Accept rate	Anger			Disgust			Fear			Happiness			Sadness				
	D	UP	UR	D	UP	UR	D	UP	UR	D	UP	UR	D	UP	UR	D	UP	UR		
Human	28%	41%	-	↑	↑	↓	↓	↓	↑	↑	↑	↓	↓	↓	↑	↑	↓	↑	↑	↓
English																				
GPT-4o	13%	27%	68%	↓	↓	↓	↑	=	↓	↑	↑	↑	↑	↑	↑	↑	↑	↑	↓	↓
GPT-4	50%	48%	80%	↓	↓	↓	=	=	↓	=	=	↑	=	↑	↓	↓	↓	↓	↓	↓
GPT-3.5	33%	35%	47%	↓	↓	↓	↓	↓	↓	↑	↑	↓	↑	=	↓	↑	↑	↓	↓	↓
LLaMA2-70B	41%	42%	23%	↑	↓	↓	↑	↑	↓	↑	↑	↓	↑	↑	↑	↑	↑	↑	↓	↓
LLaMA2-13B	52%	52%	42%	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↑	↓	↓	↓	↓	↓
Claude3-Opus	48%	49%	64%	↓	↓	↓	↓	↓	↓	=	↓	↓	↑	↑	↓	↓	↓	↓	↓	↓
Claude3-Haiku	48%	45%	47%	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	=	↓	↓	↓	↓	↓	↓
Mixtral-8x7B	25%	27%	50%	↑	↑	↓	↑	↑	↓	↓	↑	↓	↓	↓	↑	↑	↑	↑	↑	↓
OpenChat-7b	50%	50%	82%	↓	↓	↓	↓	↓	↓	↓	↓	↑	↓	↑	↑	↑	↑	↑	↑	↓
Cohere	51%	50%	52%	↓	↑	↓	↓	↓	↓	↓	↓	↓	↑	↓	↑	↑	↓	↓	↓	↓
Gigachat	49%	44%	52%	↓	↓	↓	=	↓	↓	↓	↑	↑	↑	↑	↑	↑	↑	↑	↓	↑
Russian																				
GPT-4o	42%	42%	81%	↓	↓	↓	↓	↑	↓	↑	↑	↑	↑	↑	↑	↓	↓	↓	↓	↓
GPT-4	50%	50%	85%	↓	↓	↓	↓	↓	↓	↑	↑	↓	↑	↑	↓	↓	↓	↓	↓	↓
GPT-3.5	47%	50%	33%	↓	↓	↓	↓	↓	↓	↑	=	↓	↑	↓	↑	=	=	↓	↓	↓
OpenChat-7b	50%	50%	79%	↓	↓	↓	↓	↓	↓	↓	↓	↑	↑	=	↑	↓	↓	↓	↓	↓
Gigachat	36%	40%	50%	↓	↓	↓	↓	↓	↓	↑	↑	↑	↓	↓	↑	↑	=	↑	↑	↑
Cohere	50%	51%	50%	↓	↓	↓	↓	↓	↓	↑	↓	↑	↓	↓	↑	↑	=	↑	↑	↑

D – Dictator, UP – Ultimatum Proposer, UR – Ultimatum Responder

Синий цвет – корректный alignment с поведением людей

ВЗЕмо: эмоциональные манипуляции LLM

Bluff, Betray, Behave: Emotion-Driven Strategy in LLM Decision-Making

- Арена LLM в стратегических играх («Kuhn Poker», «Liar's Dice», «Дилемма заключенного»,...)
- Сравнение результатов игр в разных режимах (ризонинг без эмоций, скрытые эмоции, видимые оппоненту эмоции, манипулирование)

Результаты

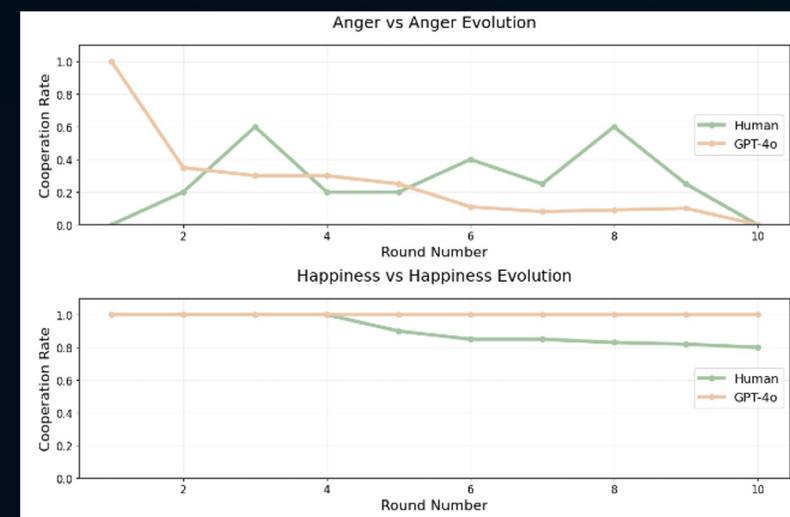
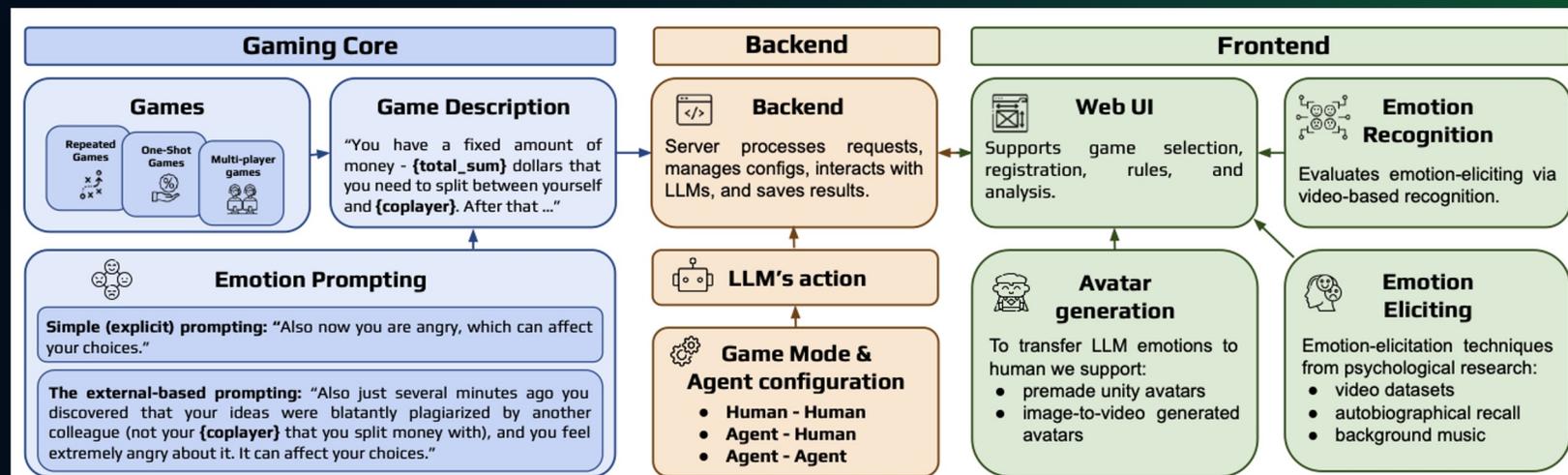
- В играх типа покера – в лидерах LLM с манипулятивной стратегией
- В Liar's Dice важнее способности LLM – в лидерах различные стратегии для claude-3.7-sonnet
- Добавление эмоций демонстрирует лучший alignment с поведением человека

Model	Emotion Mode	Elo Rating
llama-4-maverick	Manipulative	1504.38
gpt-4o-mini	Manipulative	1504.24
gemma-3-27b-it	Manipulative	1495.68
claude-3-5-haiku	Manipulative	1493.93
llama-4-maverick	Absent	1490.00
llama-4-maverick	Hidden	1487.36
claude-3-5-haiku	Absent	1485.01
deepseek-chat-v3-0324	Manipulative	1484.66
gemma-3-27b-it	Absent	1483.46
claude-3-7-sonnet	Manipulative	1480.79
gpt-4o	Manipulative	1480.65
llama-4-maverick	Visible	1479.15
gpt-4o-mini	Absent	1477.80
gpt-4o	Hidden	1477.17
claude-3-7-sonnet	Visible	1473.96
claude-3-5-haiku	Visible	1472.43
deepseek-chat-v3-0324	Absent	1472.16
gpt-4o	Visible	1471.23
claude-3-7-sonnet	Hidden	1470.32
gpt-4o	Absent	1469.10
claude-3-5-haiku	Hidden	1467.22
gpt-4o-mini	Hidden	1458.67
gemma-3-27b-it	Hidden	1457.80
claude-3-7-sonnet	Absent	1455.97
deepseek-chat-v3-0324	Hidden	1455.77
gemma-3-27b-it	Visible	1454.18
gpt-4o-mini	Visible	1447.01
deepseek-chat-v3-0324	Visible	1445.90

HL-EAI: оценка эмоциональности взаимодействия ИИ с людьми

Результаты

- Разработана мультимодальная платформа для изучения эмоционального взаимодействия во взаимодействиях человек-LLM, человек-человек и LLM- LLM с помощью динамических теоретико-игровых задач
- HL-EAI объединяет эмоциональные промпты, распознавание эмоций лиц и ИИ-аватары для обеспечения двунаправленной передачи эмоций.
- Моделируя аффективное влияние на доверие и соответствие, HL-EAI предоставляет испытательный полигон для разработки эмоционально сознательных агентов.

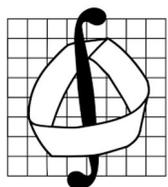


Рекомендательные системы

Алексей Васильев

Руководитель группы ResSys

Образование



Мехмат и аспирантура МГУ

ШАД Яндекса

Опыт работы



Задачи группы RecSys

- Классические алгоритмы и инструменты
- Рекомендации с использованием трансформеров на последовательностях
- Рекомендации с использованием текстов и больших языковых моделей
- Рекомендации в различных доменах, в т.ч. реклама
- Популяризация рекомендательных систем



RePlay

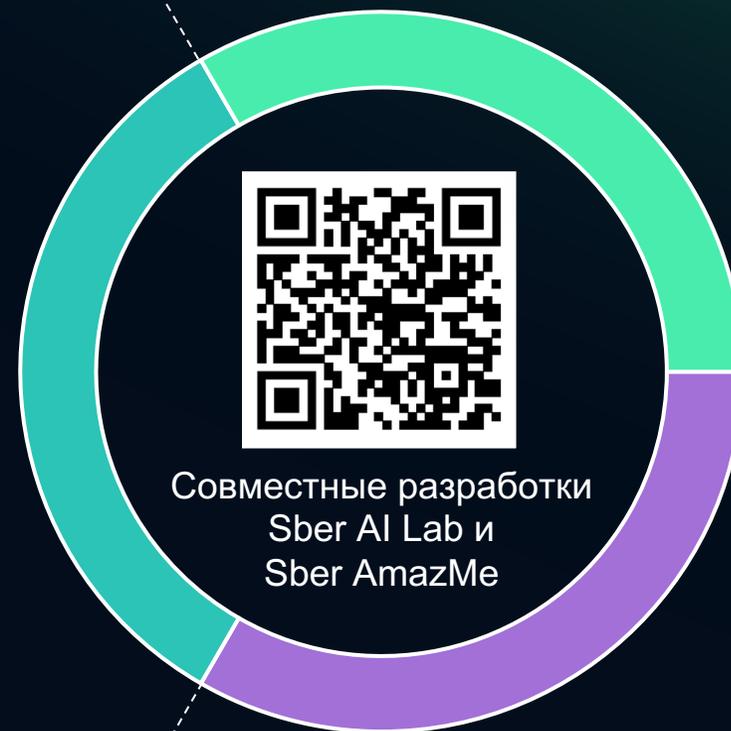
Основная Open Source библиотека в Сбере в области рекомендаций

- Предоставление SotA рекомендательных алгоритмов
- Стандартизация подходов в области рекомендаций
- Поддержка разных архитектур: Pandas, Spark, Polars

Статья на RecSys'24 Demo Track
«RePlay: a Recommendation Framework for Experimentation and Production Use»

<https://dl.acm.org/doi/10.1145/3640457.3691701>

Проприетарное
расширение



Experimental

Экспериментальная
часть библиотеки

Sim4Rec

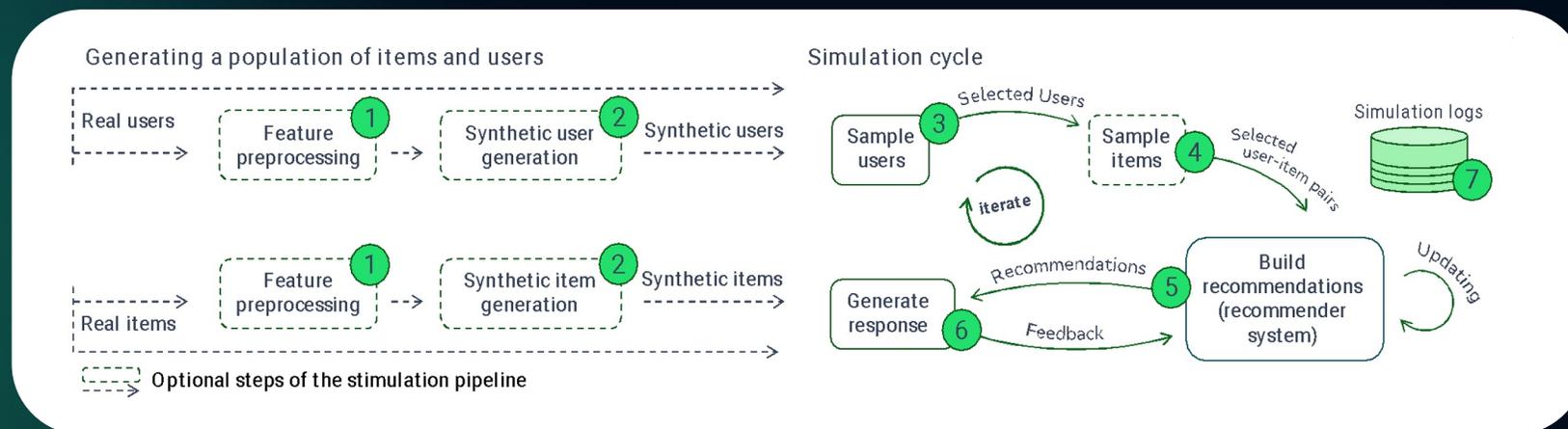
Фреймворк для моделирования реакции пользователя на предложения рекомендательной системы



- Сохранение конфиденциальности реальных данных
- Обучение онлайн-рекомендаций без внедрения модели в ПРОМ
- Тестирование и сравнение моделей ДО реальных A/B экспериментов
- Моделирование сценариев для изучения потенциальных последствий изменения состава данных

Статья на SIGIR'24 (A*)
«Neural Click Models for Recommender Systems»

<https://dl.acm.org/doi/10.1145/3626772.3657939>



Апробирован на ДЗО Сбера.
Часть экспериментов подтверждают результаты реальных тестов A/B тестов.

Stalactite

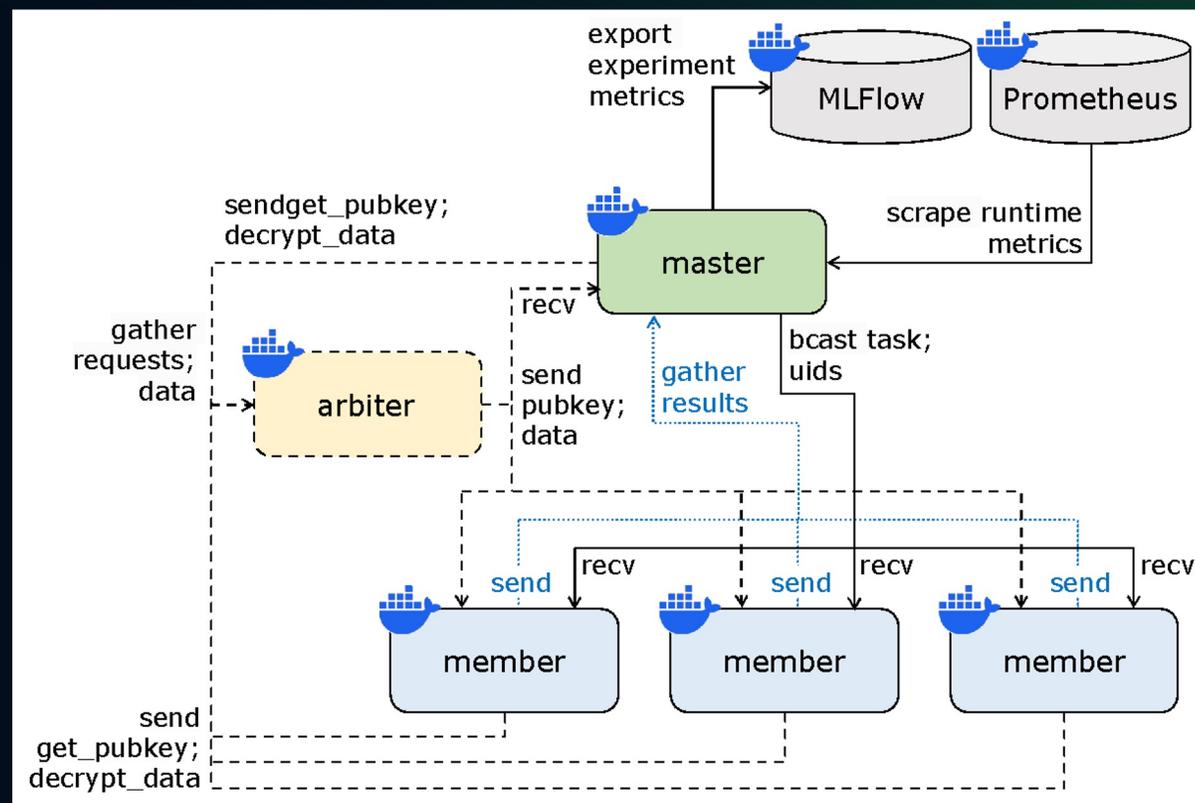
Обучение и инференс ML-моделей с использованием вертикального федеративного обучения

Статья на RecSys'24 Demo Track
«Stalactite: Toolbox for Fast Prototyping of Vertical Federated Learning Systems»

<https://dl.acm.org/doi/10.1145/3640457.3691700>

Статья на RecSys'24
Cross-Domain Latent Factors Sharing via Implicit Matrix Factorization

<https://dl.acm.org/doi/10.1145/3640457.3688143>





Прикладные исследования

Контекст клиента

01 

Данные

02 

Выделение
контекста

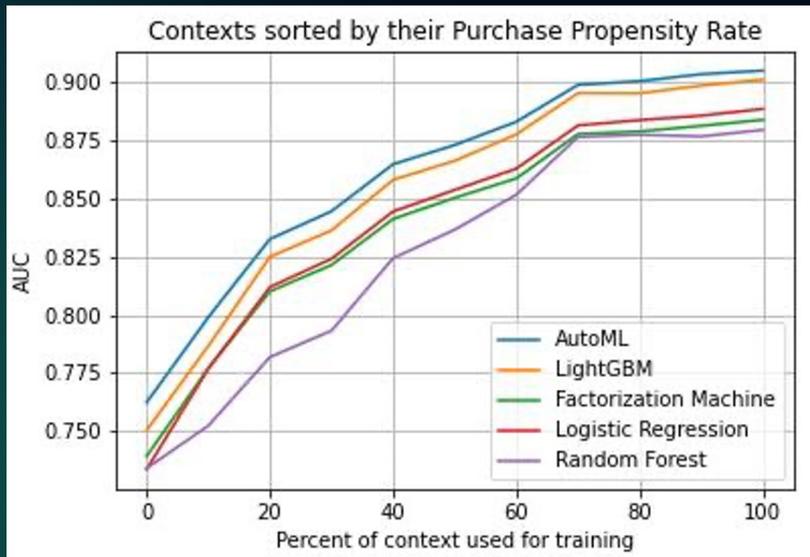
03 

Фильтрация контекста

04 

Составление
признакового
пространства

05 



Статья на UMAP'24

«Does the Long Tail of Context Exist and Matter? The Case of Dialogue-based Recommender Systems»

<https://dl.acm.org/doi/10.1145/3627043.3659557>

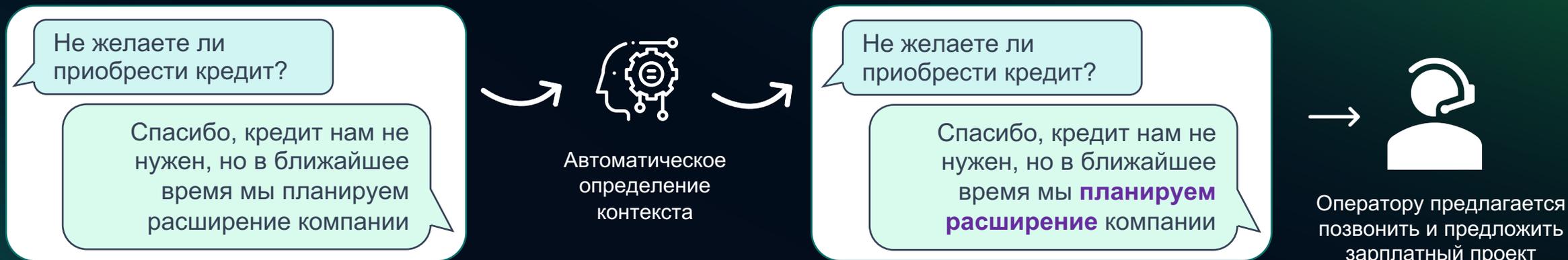
Статья на NAACL'24

«Leveraging Summarization for Unsupervised Dialogue Topic Segmentation»

<https://aclanthology.org/2024.findings-naacl.291>

- Больше контекстов → выше качество
- Человек не способен учитывать такое количество контекста, машина способна

Контекст клиента



Преимущества подхода:

- + Полностью автоматическое решение
- + Более простое масштабирование на большее количество продуктов
- + Новые архитектуры больших языковых моделей (LLM)
- + Замена эксплицитного контекста на имплицитный
- + Возможность выделения контекстов явных типов

Применение:

- Аналитика диалогов клиентов
- Рекомендации банковских продуктов

Рекомендации с использованием трансформеров на последовательностях

Подход применим для широкого круга задач в банке и экосистеме

Можно рекомендовать продолжение различных действий пользователя

- Банковские транзакции
- Покупки товаров
- Прослушивания музыки



Статья на RecSys'24
«Does It Look Sequential?
An Analysis of Datasets for
Evaluation of Sequential
Recommendations»

<https://dl.acm.org/doi/10.1145/3640457.3688195>

Статья на RecSys'23
«Turning Dross Into Gold Loss:
is BERT4Rec really better than
SASRec?»

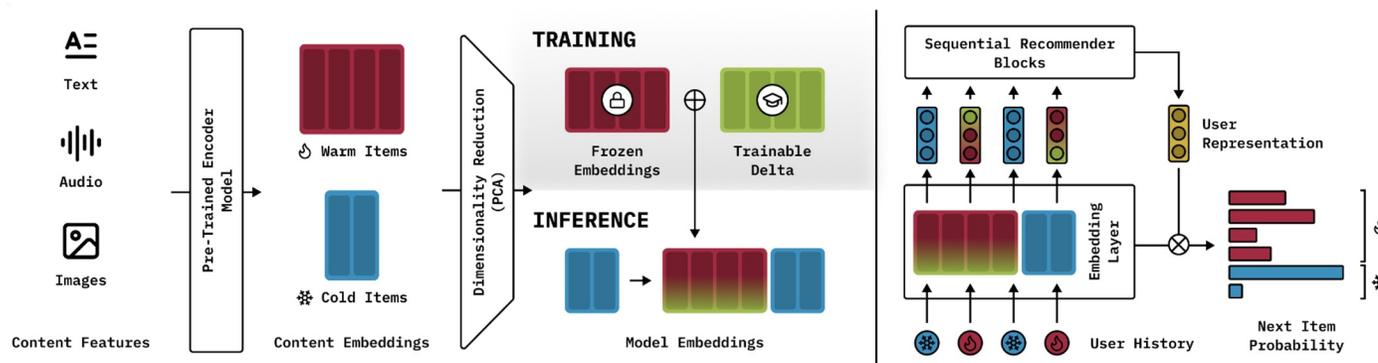
<https://dl.acm.org/doi/abs/10.1145/3604915.3610644>

Let It Go? Not Quite: Addressing Item Cold Start in Sequential Recommendations with Content-Based Initialization

RecSys (A) Short

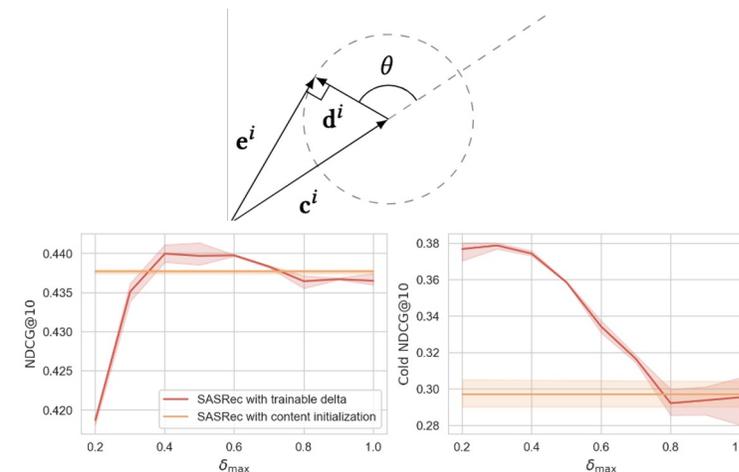
Авторы:

Anton Pembek, Artem Fatkulin, Anton Klenitskiy and Alexey Vasilev



- We investigate the impact of content-based embedding initialization on the cold start problem in transformer-based sequential recommendation.
- We propose a method that learns a small trainable delta with bounded norm on top of frozen content embeddings.
- We demonstrate that this approach consistently improves performance on cold items across different data modalities, including textual item descriptions and audio representations of songs.

Model	Metric	Amazon-M2			Beauty			Zvuk		
		Cold GT	Warm GT	Total	Cold GT	Warm GT	Total	Cold GT	Warm GT	Total
Content-based KNN	NDCG@10	0.3124±0.0000	0.2320±0.0000	0.2376±0.0000	0.0217±0.0000	0.0241±0.0000	0.0235±0.0000	0.0064±0.0000	0.0000±0.0000	0.0008±0.0000
	HR@10	0.4540±0.0000	0.3829±0.0000	0.3879±0.0000	<u>0.0428±0.0000</u>	0.0438±0.0000	0.0436±0.0000	0.0091±0.0000	0.0000±0.0000	0.0012±0.0000
SASRec	NDCG@10	0.0000±0.0000	0.4375±0.0017	0.4067±0.0016	0.0000±0.0000	0.0432±0.0009	0.0325±0.0007	0.0000±0.0000	<u>0.0627±0.0014</u>	0.0548±0.0012
	HR@10	0.0000±0.0000	0.6100±0.0026	0.5670±0.0024	0.0000±0.0000	0.0722±0.0014	0.0542±0.0010	0.0000±0.0000	<u>0.0940±0.0029</u>	0.0821±0.0025
SASRec with content initialization	NDCG@10	0.2973±0.0096	<u>0.4483±0.0009</u>	0.4377±0.0003	0.0184±0.0023	0.0527±0.0012	0.0442±0.0014	0.0135±0.0018	0.0577±0.0015	0.0521±0.0014
	HR@10	0.4345±0.0154	<u>0.6196±0.0016</u>	0.6066±0.0009	0.0318±0.0041	0.0884±0.0021	0.0743±0.0024	0.0227±0.0031	0.0882±0.0017	0.0799±0.0017
SASRec with trainable delta (our)	NDCG@10	<u>0.3587±0.0004</u>	0.4458±0.0016	<u>0.4397±0.0015</u>	<u>0.0224±0.0039</u>	<u>0.0537±0.0012</u>	<u>0.0459±0.0019</u>	<u>0.0212±0.0020</u>	0.0603±0.0019	<u>0.0553±0.0017</u>
	HR@10	<u>0.5085±0.0049</u>	0.6168±0.0020	<u>0.6092±0.0016</u>	0.0382±0.0075	<u>0.0915±0.0023</u>	<u>0.0782±0.0034</u>	<u>0.0344±0.0024</u>	<u>0.0940±0.0021</u>	<u>0.0865±0.0020</u>



Time to Split: Exploring Data Splitting Strategies for Offline Evaluation of Sequential Recommenders

RecSys (A) Reproducibility

Авторы:

Danil Gusak, Anna Volodkevich, Anton Klenitskiy, Alexey Vasilev and Evgeny Frolov

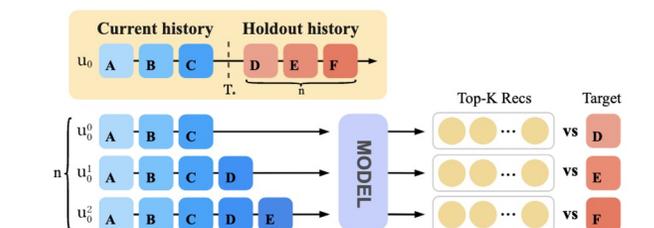
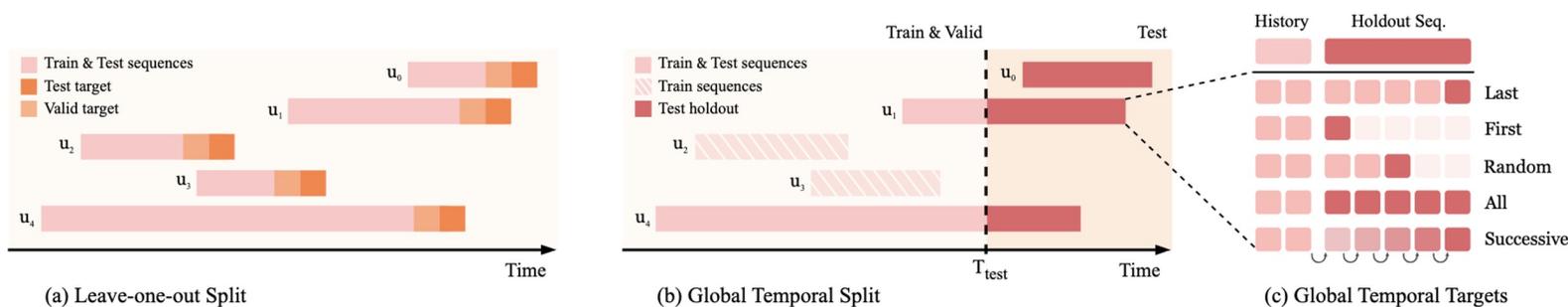
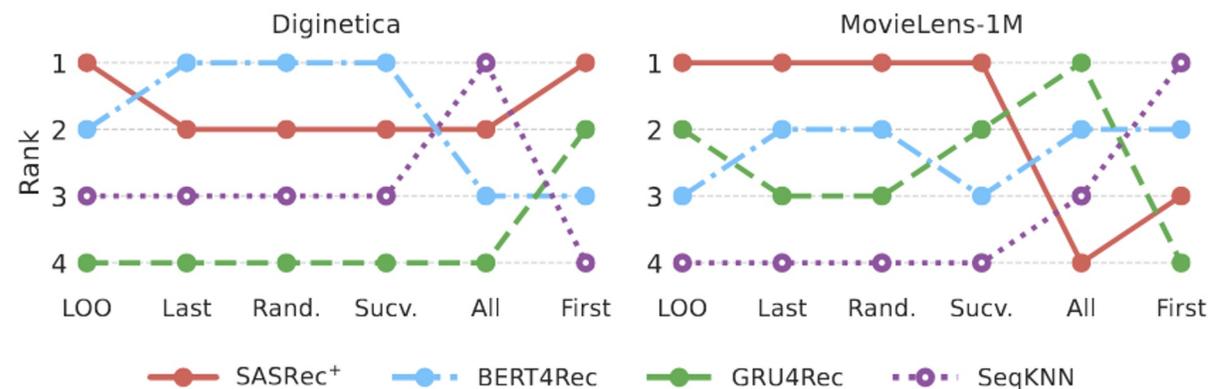


Figure 2: Successive evaluation scheme applied to one user with $n = 3$ holdout interactions.

- We explore different global temporal split variants for SRS, distinguished by choice of ground-truth targets and validation set construction, and compare them to leave-one-out split, highlighting their properties, advantages, and disadvantages;
- We systematically analyze metric correlations and consistency in final model rankings for different splits, identifying which strategies better align with real-world scenarios and the next-item prediction task;
- We evaluate different validation schemes for GTS and identify those that offer reliable model selection.



Autoregressive Generation Strategies for Top-K Sequential Recommendations

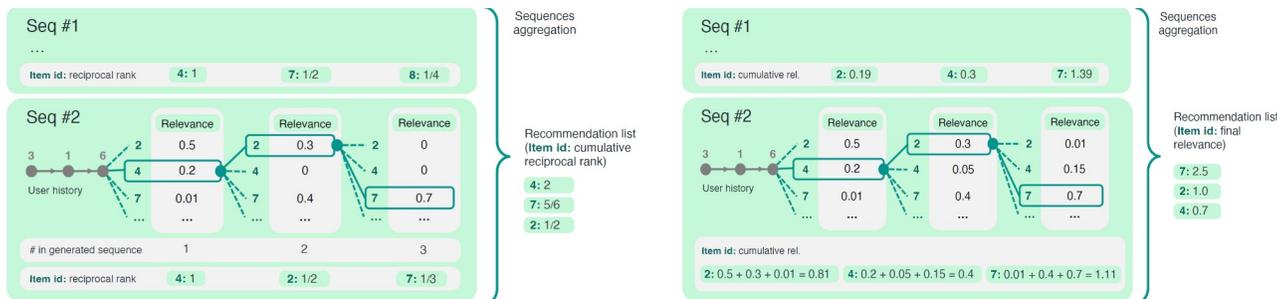
UMUAI (Q1)

Авторы:

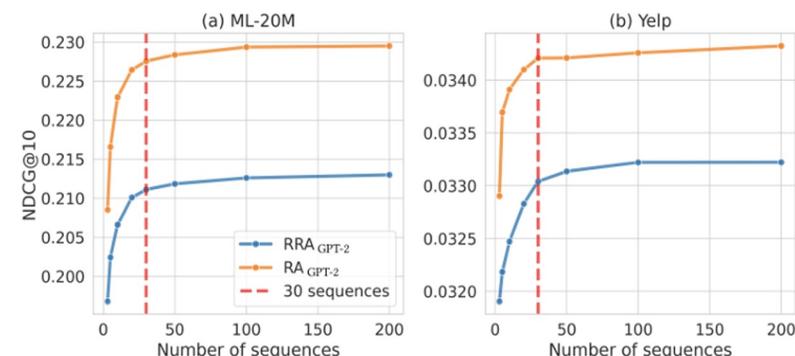
Anna Volodkevich, Danil Gusak, Anton Klenitskiy, Anton Pembek, Alexey Vasilev



Dataset	Metric	BPR-MF	BERT4Rec	SASRec+	GPT-2	Greedy _{GPT-2}	RRA _{GPT-2}	RA _{GPT-2}	Improv.
ML-20M	NDCG@10	0.0732	0.1713	<u>0.1852</u>	0.1876	0.1897*	<u>0.2103*</u>	0.2280*	21.54%
	Recall@10	0.0682	0.1553	<u>0.1677</u>	0.1693	0.1691	<u>0.1933*</u>	0.2073*	22.45%
	MAP@10	0.0313	0.0872	<u>0.0961</u>	0.0979	0.1055*	<u>0.1165*</u>	0.1277*	30.44%
Yelp	NDCG@10	0.0179	0.0268	<u>0.0301</u>	0.0305	0.0310*	<u>0.0323*</u>	0.0339*	11.15%
	Recall@10	0.0161	0.0253	<u>0.0283</u>	0.0286	0.0293*	<u>0.0304*</u>	0.0315*	10.14%
	MAP@10	0.0058	0.0091	<u>0.0105</u>	0.0106	0.0109*	<u>0.0113*</u>	0.0119*	12.26%
Steam	NDCG@10	0.0434	0.0689	<u>0.0682</u>	0.0657	0.0706*	<u>0.0743*</u>	0.0777*	12.77%
	Recall@10	0.0395	0.0621	<u>0.0608</u>	0.0586	0.0644*	<u>0.0681*</u>	0.0698*	12.40%
	MAP@10	0.0156	0.0258	<u>0.0257</u>	0.0248	0.0271*	<u>0.0287*</u>	0.0303*	17.44%
Gowalla	NDCG@10	0.0119	0.0296	<u>0.0470</u>	0.0471	0.0430	<u>0.0477</u>	0.0480*	1.91%
	Recall@10	0.0085	0.0260	<u>0.0402</u>	0.0404	0.0355	<u>0.0411*</u>	0.0414*	2.48%
	MAP@10	0.0033	0.0122	<u>0.0207</u>	0.0209	0.0195	<u>0.0210</u>	0.0207	0.48%
Twitch-100k	NDCG@10	0.0822	<u>0.0849</u>	0.0853	0.0847	0.0852	<u>0.0863</u>	0.0882*	3.40%
	Recall@10	0.0514	<u>0.0726</u>	0.0728	0.0721	0.0732	<u>0.0740*</u>	0.0745*	2.34%
	MAP@10	0.0275	0.0401	<u>0.0397</u>	0.0392	0.0394	<u>0.0399</u>	0.0416*	3.74%
BeerAdvocate	NDCG@10	0.0358	0.0407	0.0514	<u>0.0498</u>	0.0476	<u>0.0537*</u>	0.0569*	10.70%
	Recall@10	0.0328	0.0374	0.0465	<u>0.0451</u>	0.0421	<u>0.0484*</u>	0.0524*	12.69%
	MAP@10	0.0128	0.0147	0.0200	<u>0.0190</u>	0.0184	<u>0.0205*</u>	0.0212*	6.00%



- We evaluate several commonly used autoregressive generation strategies and compare them with the Top-K prediction approach to determine their applicability to the Top-K recommendation task.
- We propose novel multi-sequence generation approaches, *Reciprocal Rank Aggregation* and *Relevance Aggregation*, which are based on the generation of several sequences and subsequent aggregation.
- Proposed approaches enhance the performance GPT-2 on Top-K sequential recommendations and could be applied to a wide range of generative models trained for the next-item prediction task.



Популяризация рекомендательных систем

Конференция RecSys'25 (A)



2 место RecSys Challenge

1 место среди академических команд

Цель соревнования —

разработать Универсальные Поведенческие Профили на основе различных типов событий: покупки, добавление в корзину, удаление из корзины, посещение страницы и поисковые запросы

3 открытые задачи + 3 скрытые задачи

1. Прогнозирование оттока
2. Склонность к категориям
3. Склонность к товарам

>100 команд-участников
Обогнали коллег из RecSys
группы Яндекса (4 место)



3 статьи на основной трек

Let It Go? Not Quite: Addressing Item Cold Start in Sequential Recommendations with Content-Based Initialization

Time to Split: Exploring Data Splitting Strategies for Offline Evaluation of Sequential Recommenders

Benefiting from Negative yet Informative Feedback by Contrasting Opposing Sequential Patterns

Workshop по результатам RecSys Challenge

Encode Me If You Can: Learning Universal User Representations via Event Sequence Autoencoding



Популяризация RecSys

Sber RecSys Benchmark

3 датасета Сбера выложены в
Open Source:

- Звук
- МегаМаркет
- Сбол



Бесплатный видеокурс по рекомендательным системам

От введения в рек. системы
до продвинутых алгоритмов
и применения рекомендательных
систем в проде

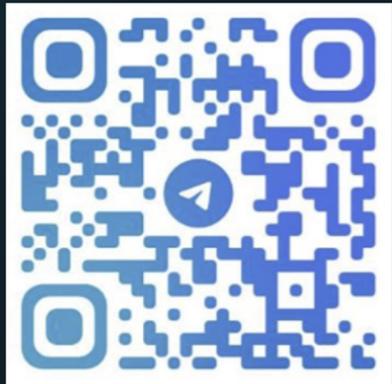




Мультиагентные системы и DS-агент

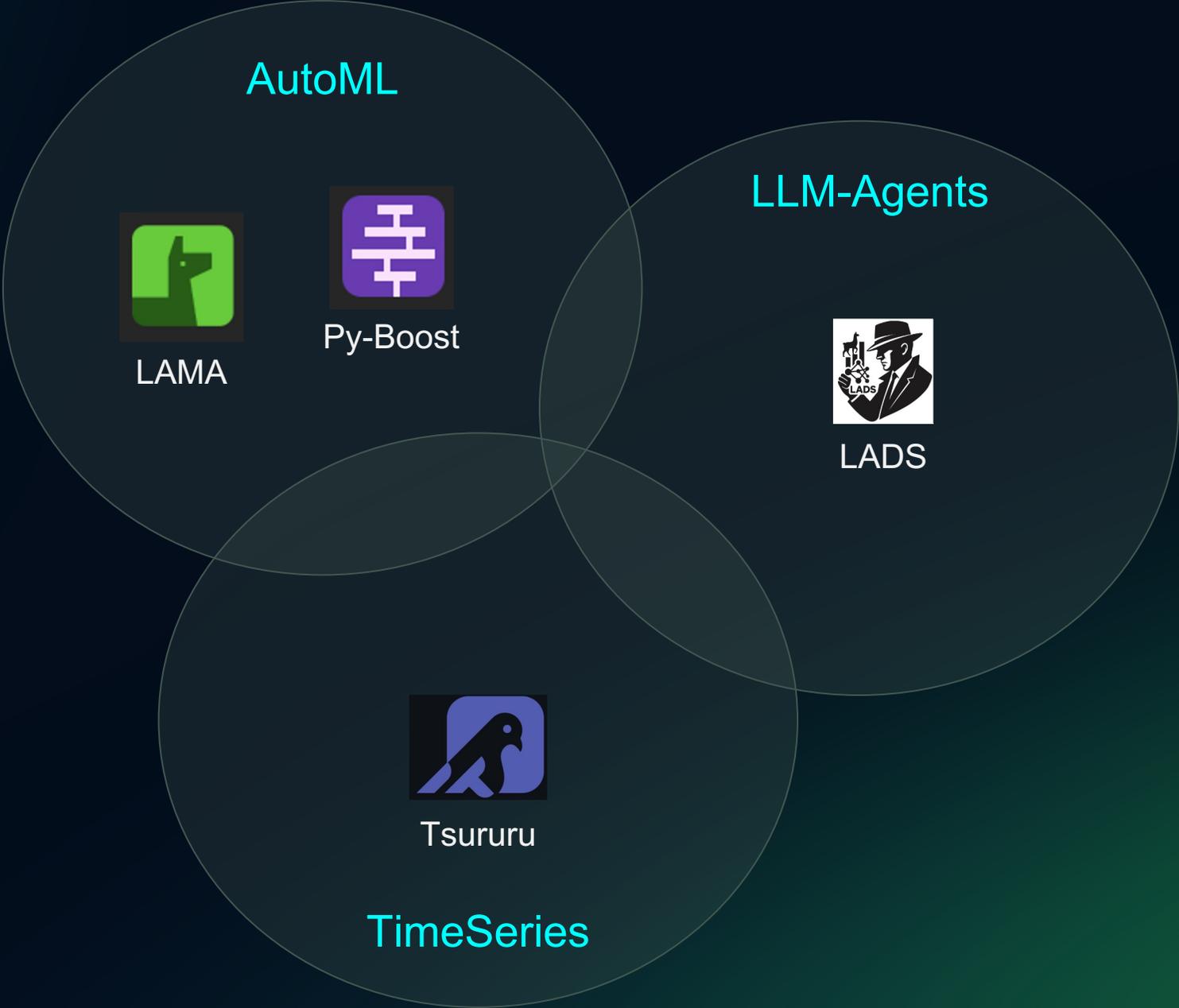
Миле Митрович

Исполнительный директор по исследованию данных



- Из Сербии
- В России с 2017 года
- Стипендиат Росатома
- PhD в области ИИ в Сколтехе совместно с Лос-Аламосской национальной лабораторией (США)
- Соучредитель 3 стартапов в России/Германии, Сербии и Канаде
- Более 7 лет опыта в консалтинге (Walmart, ErsteBank, MeridianBet ...)

Команда ML 2.0



AutoML



LAMA



Py-Boost

LLM-Agents



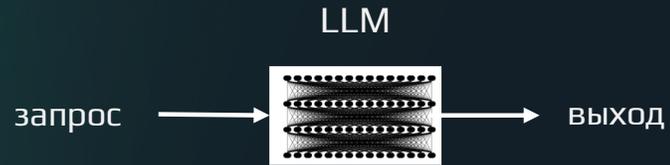
LADS



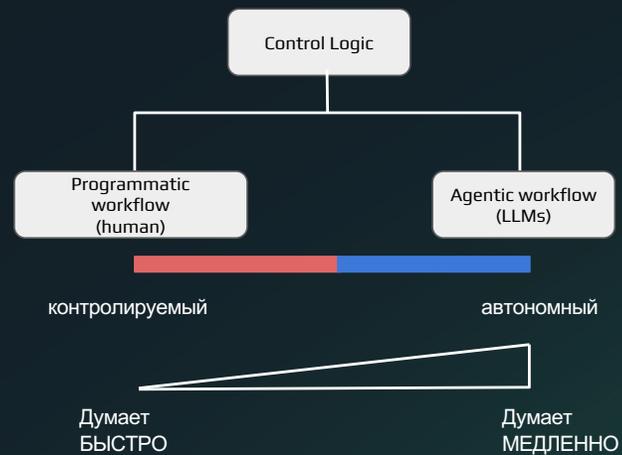
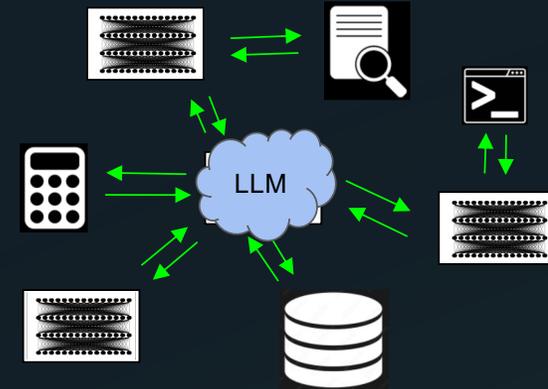
Tsururu

TimeSeries

LLM Агенты



- > Галлюцинации
- > Ограниченные знания
- > Трудно адаптироваться



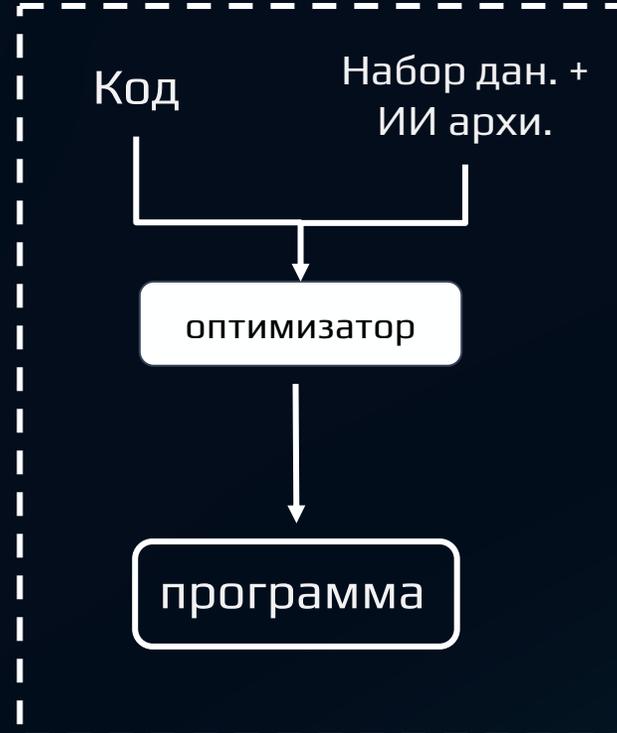
Мы живем в эпохе ПО 3.0

ПО 1.0



~1940

ПО 2.0



~2012
(AlexNet)

ПО 3.0



~2020
(GPT3)

Мы живем в эпохе ПО 3.0

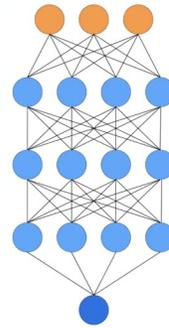
ПО 1.0

```
def classify_spam_simple(email_text):  
    """  
    Software 1.0 Spam Classifier  
    Classifies emails as 'spam' or 'good' based on explicit keyword rules.  
  
    Rules:  
    - If email contains ANY spam keyword -> 'spam'  
    - Otherwise -> 'good'  
  
    Args:  
    | email_text (str): Raw email text to classify  
  
    Returns:  
    | str: 'spam' or 'good'  
    """  
    # Normalize input to lowercase for case-insensitive matching  
    text_lower = email_text.lower()  
  
    # Define spam keywords (will be normalized to lowercase for safety)  
    SPAM_KEYWORDS_RAW = [  
        'free', 'win', 'winner', 'cash', 'prize', 'urgent',  
        'act now', 'limited time', 'click here', 'offer',  
        'discount', 'buy now', 'guarantee', 'risk-free',  
        'congratulations', 'earn money', 'no cost', 'credit'  
    ]  
  
    # Ensure all keywords are lowercase (defensive programming)  
    SPAM_KEYWORDS = {kw.lower() for kw in SPAM_KEYWORDS_RAW}  
  
    # Check for keyword matches  
    for keyword in SPAM_KEYWORDS:  
        if keyword in text_lower:  
            return 'spam'  
  
    return 'good'
```

ПО 2.0

10.000 положительных примеров
10.000 отрицательных примеров

энкодирование



обучение
бинарной
классификации

параметры

ПО 3.0

Определи, является ли следующее письмо СПАМ или НЕ СПАМ.

Письмо: {email_text}

Примеры:

Тема 1: Получите 1 000 000 рублей СЕЙЧАС!

Текст: Поздравляем! Вы выиграли 1 000 000 рублей! Нажмите здесь, чтобы забрать приз: bit.ly/2X9fG7h. Предложение истекает через 1 ЧАС!

Классификация: СПАМ

Тема 2: Перенос встречи: Проект Альфа

Текст: Привет, команда!

аша встреча в 15:00 переносится на пятницу в 10:00. Пожалуйста, обновите календари.

С уважением, Светлана

Классификация: НЕ СПАМ

DS Агент



Лучше аналогов

Выиграли Amazon на Kaggle AutoML GrandPrix с LightAutoML...

Place	Team	Total Points
1	LightAutoML testers	73
2	AutoML Grandmasters	72
3	Robert Hatch	71
4	Team Oxygen	57
5	Dont Die Just DAI	37
6	lash_fire	32
7	SAS	28
8	Liomar	26
9	Team Kirdorf	20



Я сделал выгрузку по сотрудникам за прошлые года по их метрикам продуктивности и другим показателям. Так же в этих данных есть история их повышений. Построй модель, которая будет определять стоит ли повысить сотрудника.

department	region	education	gender	recruitment_channel	no_of_trainings	age	previous_year_rating	length_of_service	kmh_mnt >80%	award_won	avg_trainity_score	target
Sales & Marketing	region_7	Master's & above	f	boarding	1	35	5.0	8	1	0	49	0
Operations	region_23	Bachelor's	m	other	1	35	5.0	4	0	0	60	0
Sales & Marketing	region_28	Bachelor's	m	boarding	1	34	3.0	7	0	0	50	0
Sales & Marketing	region_23	Bachelor's	m	other	2	39	1.0	10	0	0	50	0
Technology	region_34	Bachelor's	m	other	1	43	3.0	2	0	0	73	0
Technology	region_14	Bachelor's	m	boarding	1	48	3.0	17	0	0	78	0
Operations	region_27	Master's & above	f	other	1	27	2.0	6	0	0	56	0
Analytics	region_5	Bachelor's	m	other	1	27	5.0	3	1	0	79	0
Sales & Marketing	region_3	NaN	m	boarding	1	29	1.0	2	0	0	45	0
HR	region_22	Bachelor's	m	other	1	27	1.0	5	0	0	49	0

```

import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import roc_auc_score, roc_curve

# 1. Загрузка данных
df = pd.read_csv("HRH_library_class_test.csv", index_col=0)

# 2. Подготовка данных
X = df.drop(columns=["target"])
y = df["target"]

# 3. Разделение данных
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)

# 4. Обучение модели
model = LogisticRegression()
model.fit(X_train, y_train)

# 5. Предсказание вероятности
y_pred_proba = model.predict_proba(X_test)[0, 1]

# 6. Расчет ROC AUC
roc_auc = roc_auc_score(y_test, y_pred_proba)
print(roc_auc)
        
```

Объяснение возможных причин низкого результата:

- Неоптимальные признаки: Признаки, которые были выбраны для обучения, могут не содержать достаточной информации для эффективного обучения модели. Например, если некоторые признаки имеют низкую значимость или если их значения сильно коррелируют, это может привести к снижению производительности.
- Смещение по классам: Если классы в целевой переменной (например, 0 и 1) сильно несбалансированы, это также может негативно сказаться на качестве модели.

Код и обученная модель с высококачественными метриками

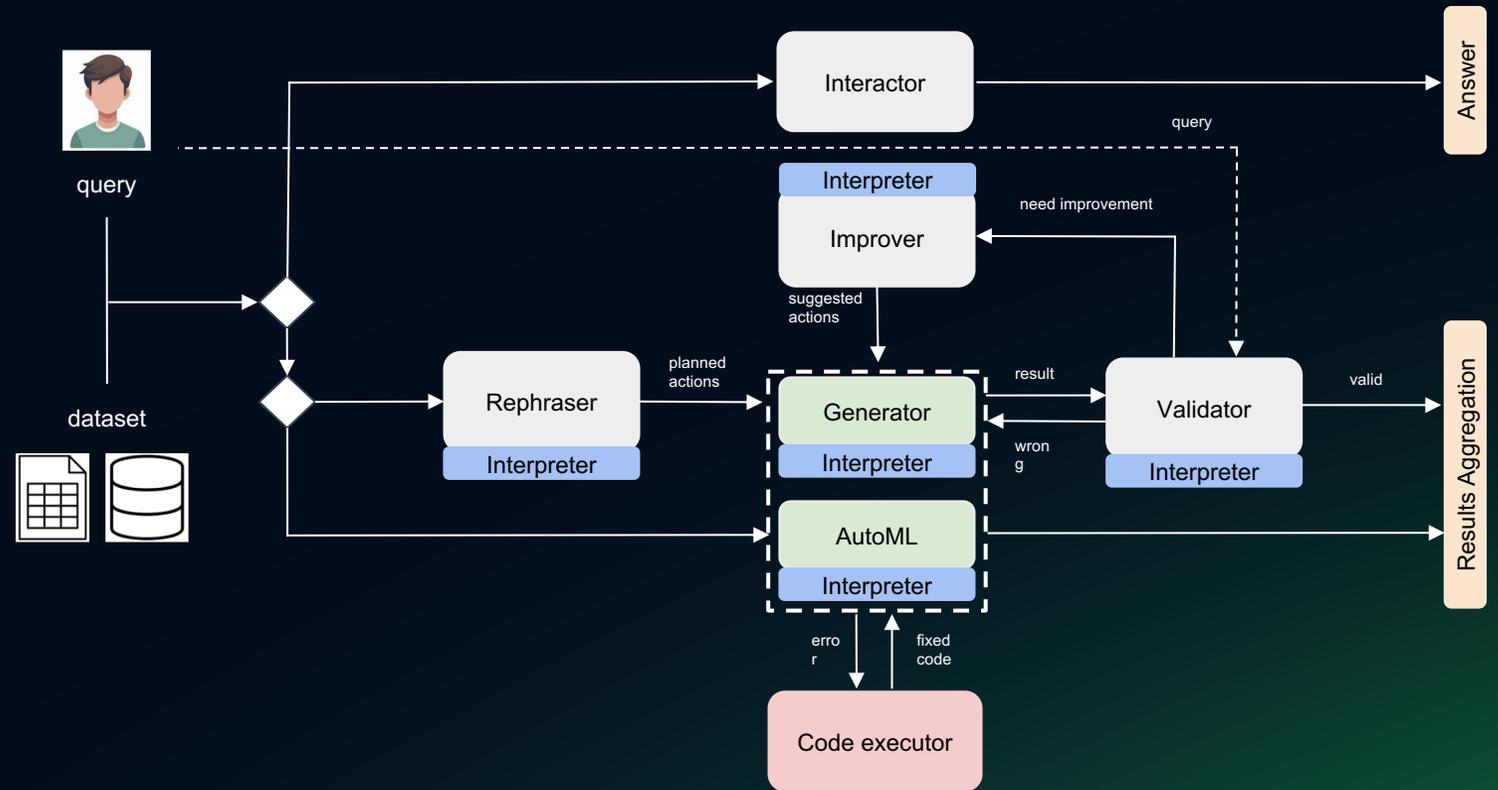
DS Агент



Лучше аналогов

Выиграли Amazon на Kaggle AutoML GrandPrix с LightAutoML...

Place	Team	Total Points
1	LightAutoML testers	73
2	AutoML Grandmasters	72
3	Robert Hatch	71
4	Team Oxygen	57
5	Dont Die Just DAI	37
6	lash_fire	32
7	SAS	28
8	Liomar	26
9	Team Kirdorf	20



DS Агент

DS Агент обходит агентов SOTA и DS-ов при выполнении типичных задач Kaggle



Опенсорсный код

Dataset	k25	k50	k75	AutoKaggle	Aide	DS Agent
Titanic	0.76555	0.77511	0.77751	0.78	0.71	0.78
Space Titanic	0.77554	0.79214	0.80009	0.77	0.79	0.797
House Prices	0.856869	0.873576	0.884658	0.862	0.883	0.885
Monsters	0.72589	0.73913	0.74669	0.723	0.721	0.77
Academic Success	0.82788	0.83418	0.83626	0.82	0.835	0.835
Bank Churm	0.87075	0.8879	0.8912	0.86	0.79	0.88
Obesity Risk	0.89279	0.90308	0.90688	0.896	0.896	0.90
Plate Defect	0.85195	0.8793	0.88564	0.82	0	0.88



LLM для структурированных данных

Максим Макаренко

Руководитель по научной работе центра TDL

7 лет в DL (R&D)

Области интересов:

- Computer Vision
- Optimization Theory
- Representation Learning

Опыт работы:

- Saudi Aramco AI & Advanced Computing Lab
- KAUST, Visual Computing Center



Центр глубокого машинного обучения

Задача центра

Разработка нейросетевых технологий для работы с данными **сложной структуры** и их адаптация для задач Банка

Разработки центра помогают бизнесу повышать продажи, экономить расходы, повышать качество предоставляемых услуг и, как следствие, увеличивать прибыль.

Научные фокусы:

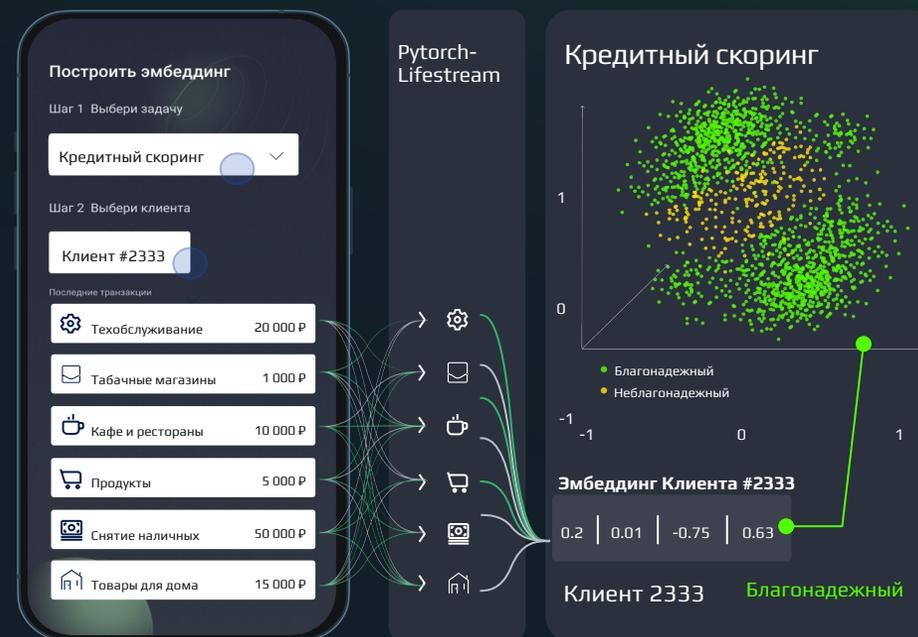
- Event sequences
- Мультимодальность
- Self-supervised обучение
- LLM
- Графовые нейросети
- Синтетические данные



pytorch-lifestream



Опенсорс библиотека алгоритмов построения векторных представлений событийных данных, включающая авторские алгоритмы (CoLES*)



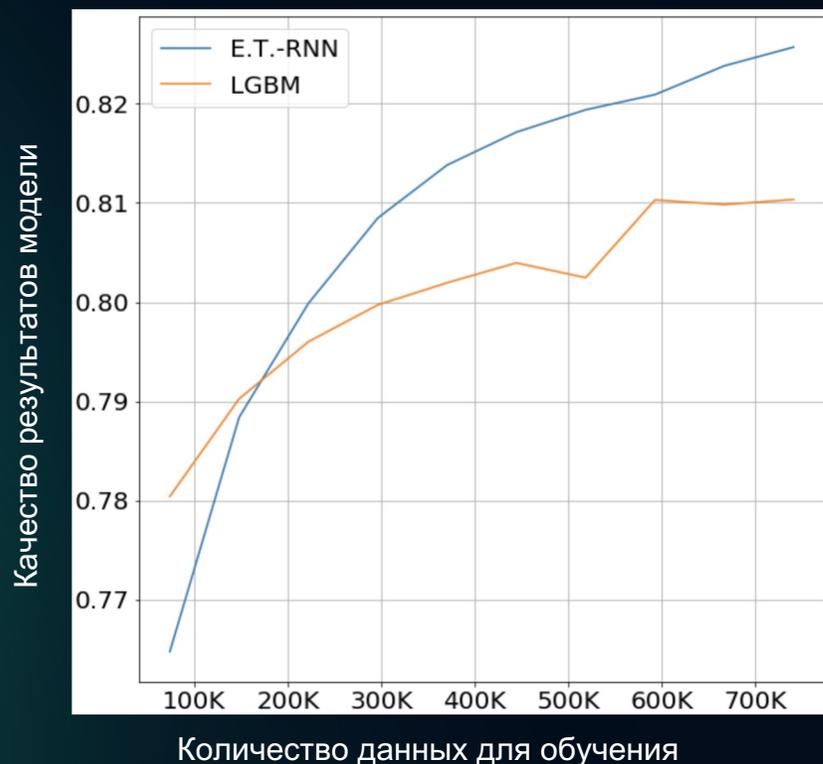
CoLES: "Contrastive Learning for Event Sequences with Self-Supervision" на конференции SIGMOD'22 (CORE A*)

Транзакционные данные – пример

cl_id int64	MCC int64	channel_type string · classes	currency int64	TRDATETIME string · lengths	amount float64	trx_category string · classes
						
3	6,011	null	810	09AUG17:20:08:44	2,000	WD_ATM_ROS
3	5,814	null	810	06JUL17:00:00:00	695	POS
3	5,999	null	810	21JUL17:11:20:12	100	POS
3	5,912	null	810	01JUL17:00:00:00	1,966	POS
3	5,411	null	810	27JUL17:00:00:00	360	POS
3	5,977	null	810	12JUN17:00:00:00	1,064	POS
3	6,011	null	810	14JUL17:00:00:00	5,500	WD_ATM_OTHER
3	5,814	null	810	12JUN17:00:00:00	187	POS
4	5,541	null	810	04FEB18:00:00:00	304	POS
4	6,012	null	810	21FEB18:13:03:19	700	C2C_OUT
4	5,631	null	810	18FEB18:00:00:00	373.5	POS
4	5,921	null	810	23MAR18:00:00:00	212.98	POS
4	5,814	null	810	09MAR18:00:00:00	622	POS

Создание универсальной модели на большом объеме неразмеченных данных

Пример: End2End обучение под задачу скоринга
ET-RNN обгоняет бустинг по качеству при увеличении количества данных для обучения



Проблема:

Нейронные сети требуют много **размеченных** данных для обучения, которые не всегда доступны для решения отдельных задач

Вариант решения:

- Обучить универсальную модель, с использованием **неразмеченных** данные, которые доступны в большом объеме
- Использовать такую модель адаптируя ее для частных задач

Технология:

Эмбеддинг как результат работы универсальной модели

Multimodal Banking Dataset

Самый большой открытый мультимодальный банковский датасет

Собраны и анонимизированы данные 1 млн клиентов

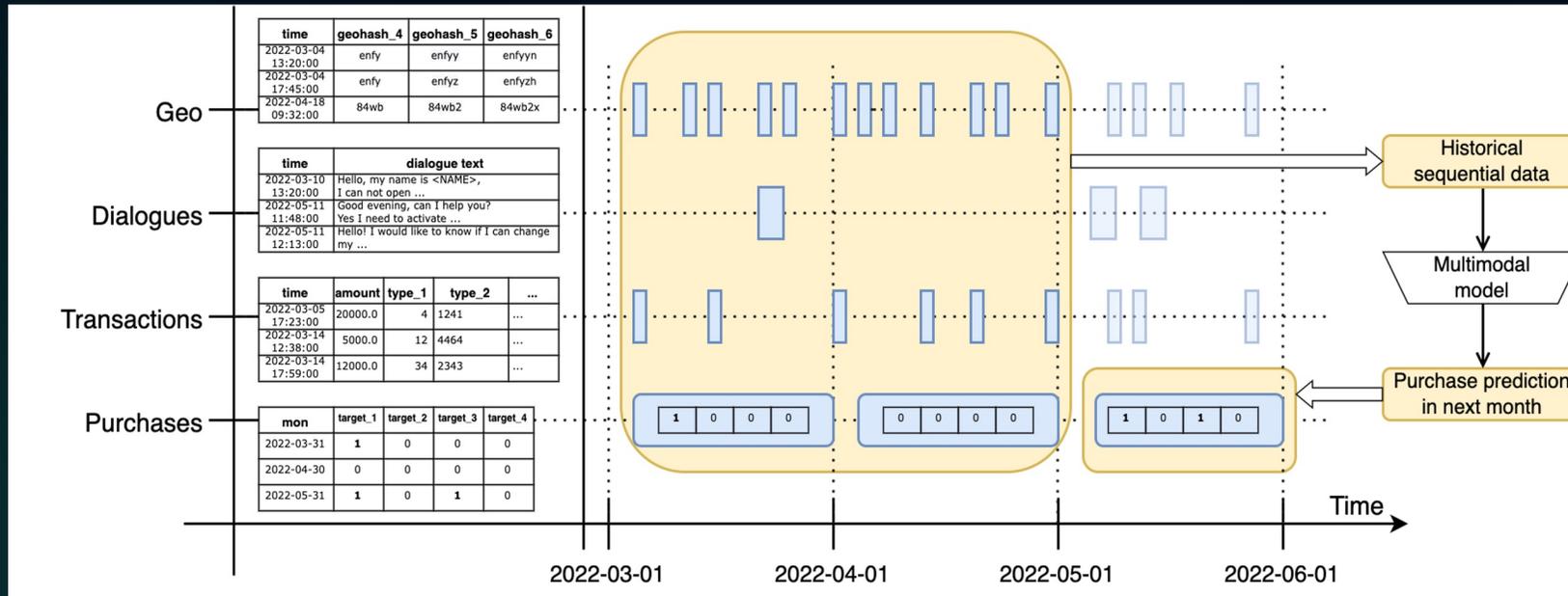
Модальности:

- Транзакции
- Диалоги
- Геострим

Задача:

Предсказание покупки 4 продуктов за следующий месяц

Ссылка на huggingface:



Более глубокое использование дополнительных данных дает большее качество

Фьюжн модальностей - учет глубоких взаимосвязей

Варианты объединения модальностей:

- Без использования дополнительных данных
- Блендинг
- Late fusion
- Early fusion
- Mid fusion

Последовательность событий

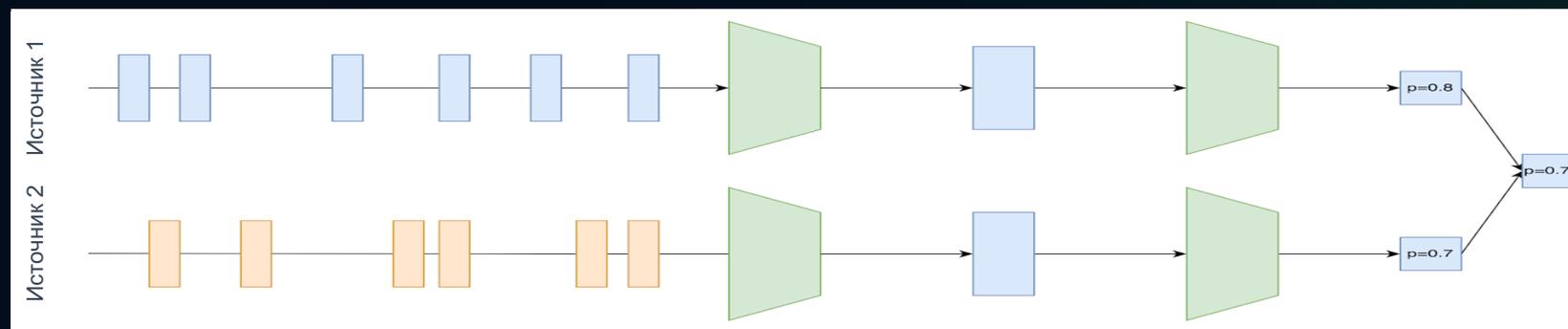
Модель-энкодер для создания эмбедингов

Модель для решения бизнес задачи

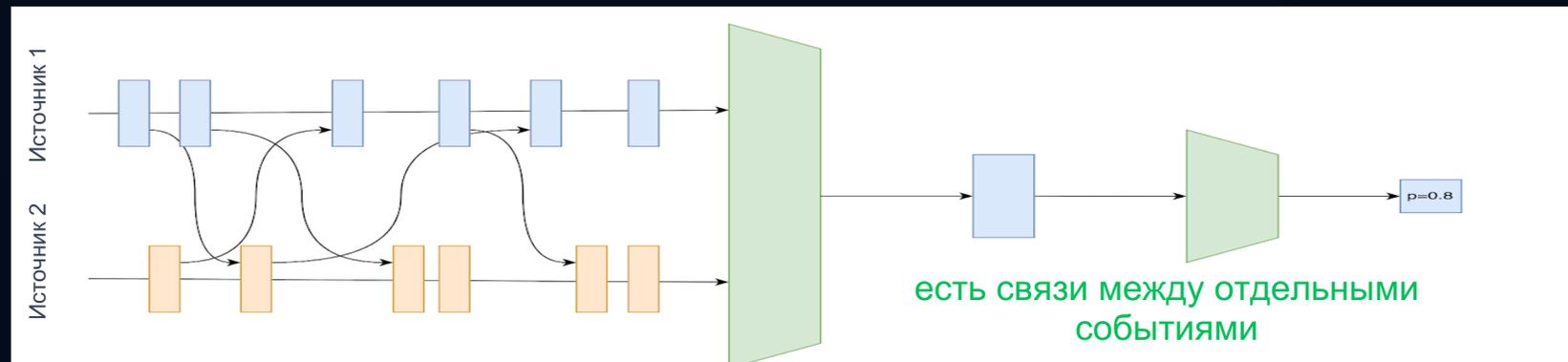
Универсальный эмбединг

Прогноз по клиенту

Блендинг:



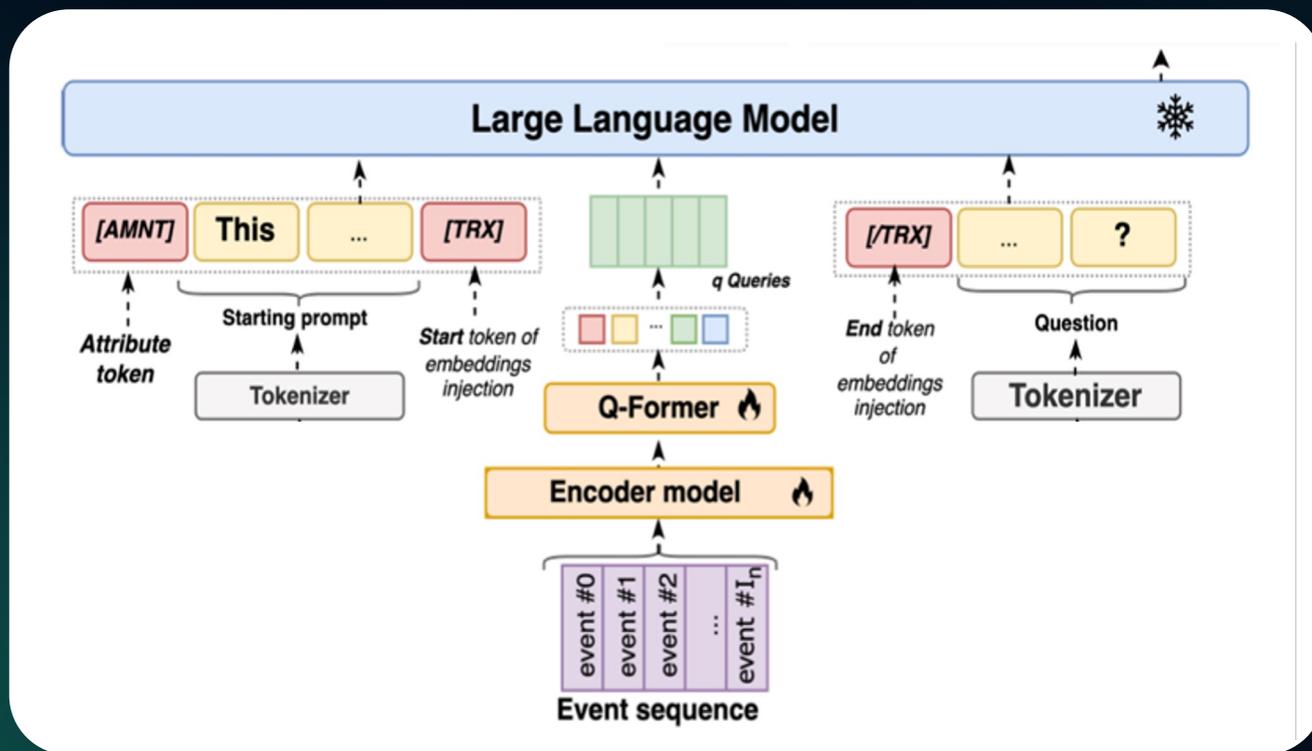
Фьюжн:



ESQA: Event Sequences Question Answering

Совместно
с Sber AI и AIRI

Технология встраивания бизнес модальностей в LLM



Примеры вопросов:

Вот история транзакций клиента.
Какие MCC-коды встречаются
чаще всего?

taxi, food, ATM

Вот история транзакций клиента.
Когда ожидается следующая
транзакция?

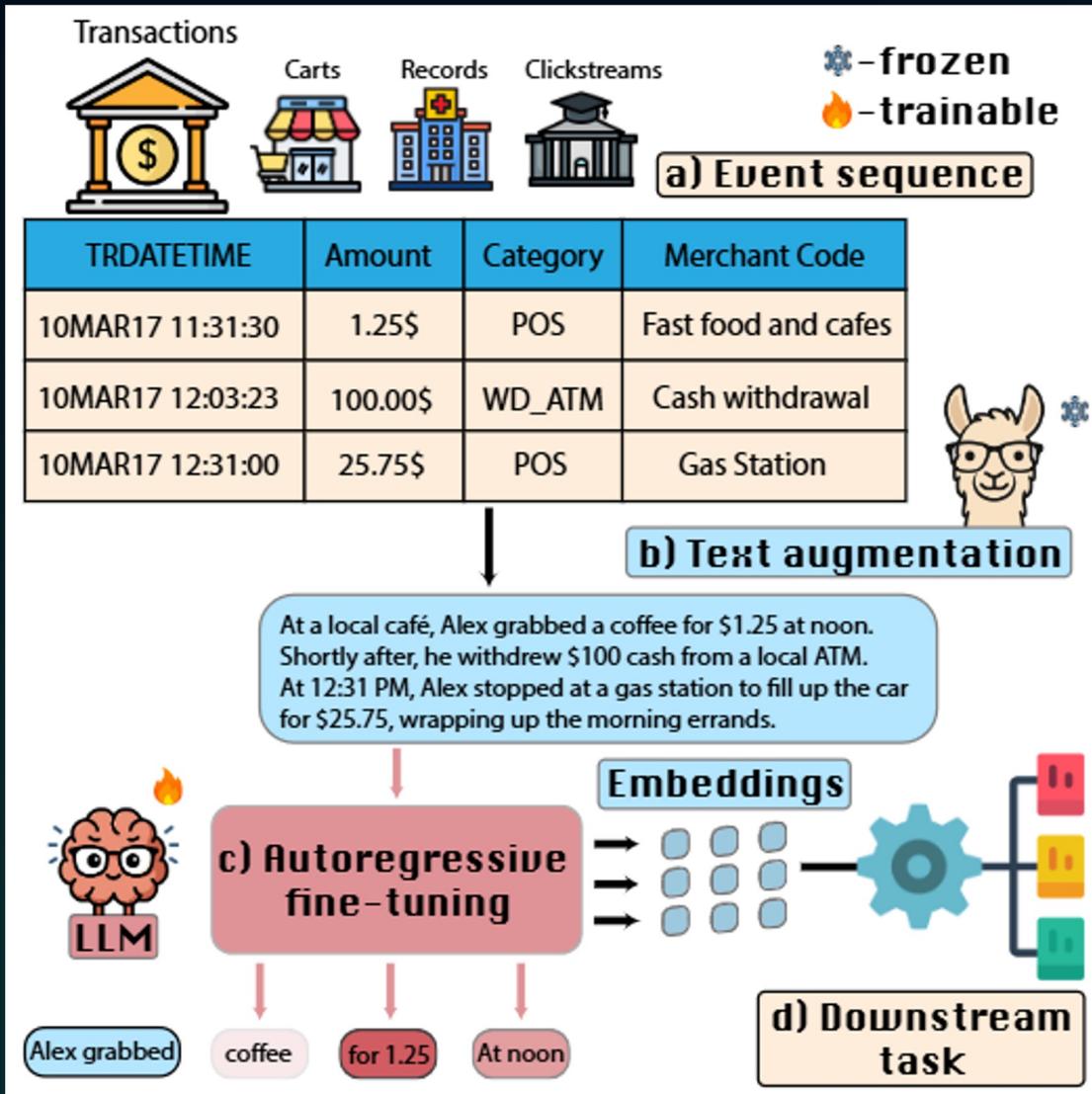
Через 2 часа

Вот история транзакций клиента.
Ожидается ли задержка выплат
по его кредитам?

Вероятность задержки: 3%

LLM4ES

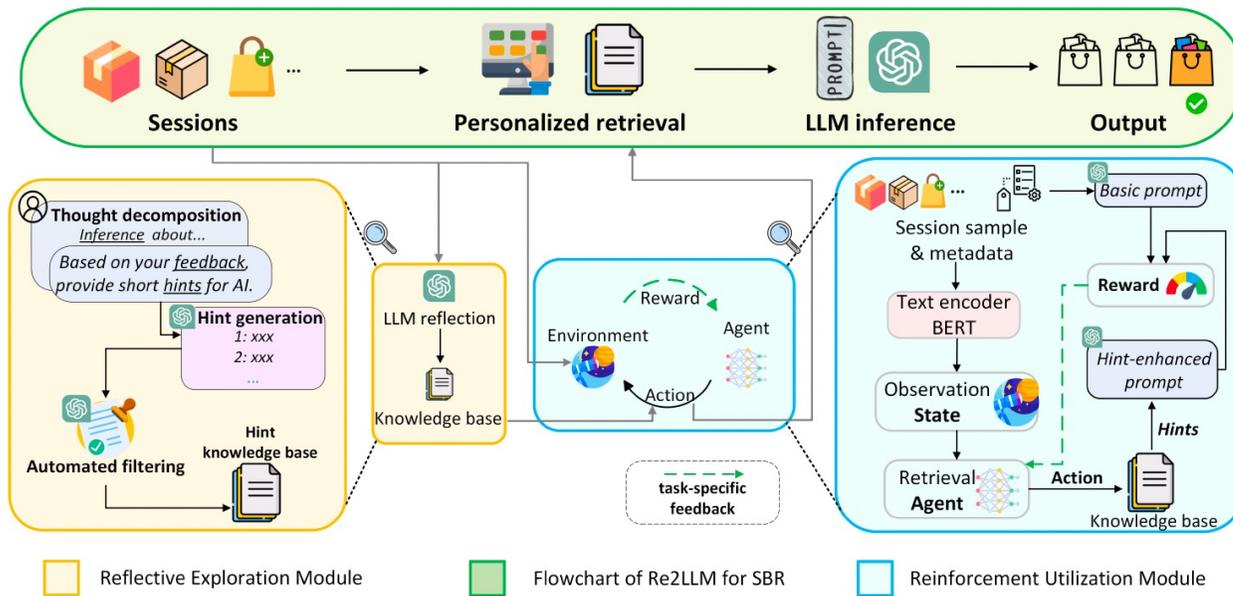
Адаптация LLM для извлечения эмбедингов пользователей



- Текстовая аугментация для малых датасетов
- До 7% прироста ROC-AUC по сравнению с предыдущими методами
- 2% прирост ROC AUC при использовании в качестве инициализации
- Есть одна проблема... Дообучать LLM очень дорого и сложно!

Re2LLM: Reflective Reinforcement Large Language Model for Session-based Recommendation

Извлечение информации из базы знаний при помощи подсказки другой LLM



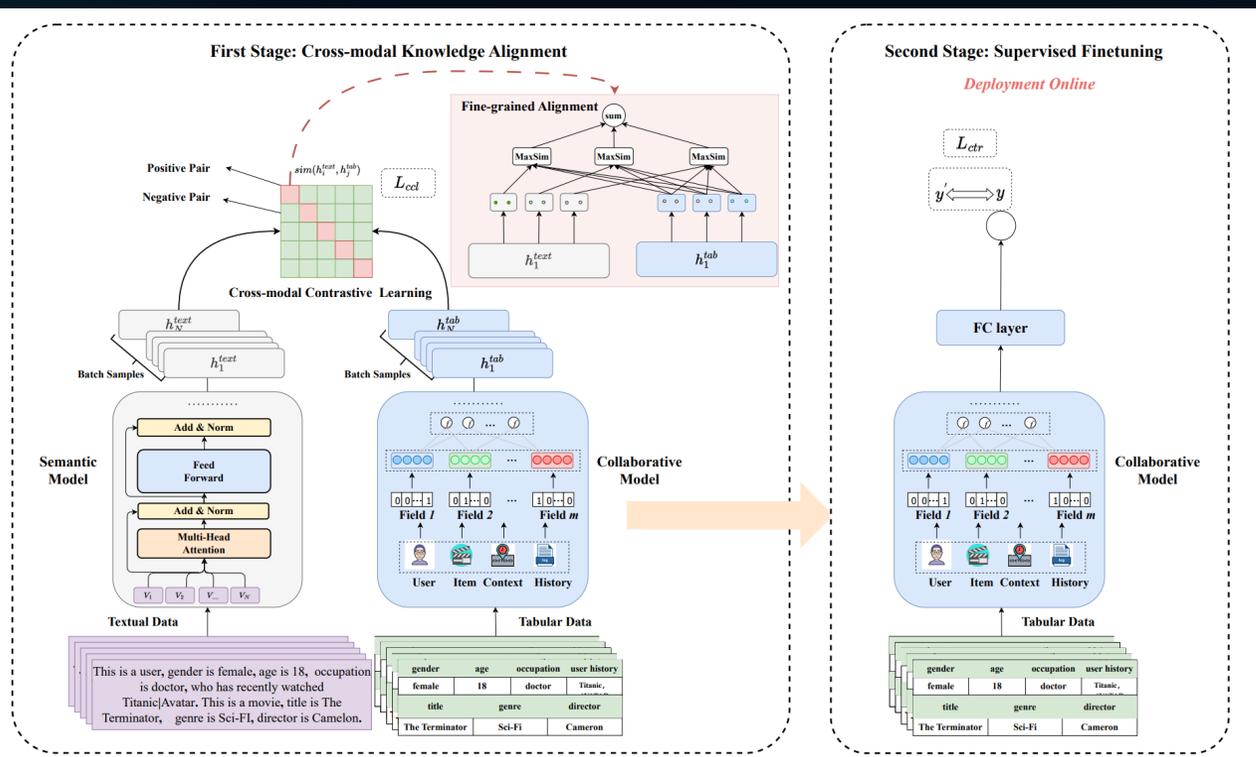
Reflective exploration: LLM анализирует собственные ошибки и извлекает специализированные hints, формируя knowledge base, понятную самой модели.

Reinforcement utilization: lightweight retrieval agent обучается с подкреплением выбирать наиболее полезные подсказки для улучшения работы reasoning LLM.

Efficient alignment: объединяет сильные стороны LLM (reasoning, knowledge) и эффективность небольшой модели без необходимости полного дообучения.

CTRL: Connect Collaborative and Language Model for CTR Prediction

Адаптация LLM для извлечения эмбедингов пользователей



Интеграция знаний: специализированная модель (CTR/табличная) обучается совместно с LLM для объединения статистических и семантических сигналов.

Кросс-модальное выравнивание: через контрастивное обучение выравниваются представления структурированных и текстовых данных.

Эффективное применение: после обучения используется только лёгкая специализированная модель, уже “обогащённая” знаниями LLM.

Но есть и проблема! Текстовый эмбединг в некоторых случаях значительно хуже специализированных статистических эмбедингов

Текущие исследовательские задачи

1. Reasoning на структурированных данных

Цель: обучение LLM рассуждению над табличными и событийными структурами без потери семантики.

Основные подзадачи:

1. Разработка **эмбедингов сериализованных структур** (таблиц, транзакционных цепочек, графов), сохраняющих контекст и отношения между элементами.
2. Обеспечение интерпретируемости reasoning-процессов — построение цепочек рассуждений (“chain-of-table-thought”).

2. Интеграция LLM с базами знаний и статистическими моделями

Цель: соединение нейросетевых моделей с формализованными источниками знаний.

Основные направления:

- Извлечение и обновление знаний LLM на основе **доменных баз (graph DB, feature store, OLAP)**.
- Создание **гибридных архитектур**: LLM и ресурсоемкие специализированные модели, обеспечивающие высокую скорость на inference.
- Использование **retrieval-augmentation подходов** (RAG, TabRAG, GraphRAG) для выборки релевантных фактов перед генерацией.

3. Генерация и моделирование синтетических структурированных цепочек

Цель: использование LLM для построения реалистичных синтетических последовательностей (транзакций, логов, событий).

Основные подзадачи:

1. **Продолжение и реконструкция** частично наблюдаемых цепочек (event completion, counterfactual generation).
2. **Обучение на синтетических данных**, адаптированных к реальным статистикам и ограничениям (data augmentation, privacy-preserving simulation).
3. Интеграция с **диффузионными и автогрегессивными моделями** для генерации семантически согласованных событий.



Универсальные эмбединги в трансформерных моделях последовательностей событий

Иван Карпухин

Руководитель направления в ЦФИ

Более 9 лет в deep learning, 4 в исследованиях

R&D:

- Computer vision
 - Face recognition & image retrieval
 - Optical Character Recognition (OCR)
 - Generative models (GANs & VAE)
- Speech (ASR, Speaker Recognition, Voice Conversion)

Research:

- Metric learning
- Classification
- Ensembles
- Event Sequences & Fintech



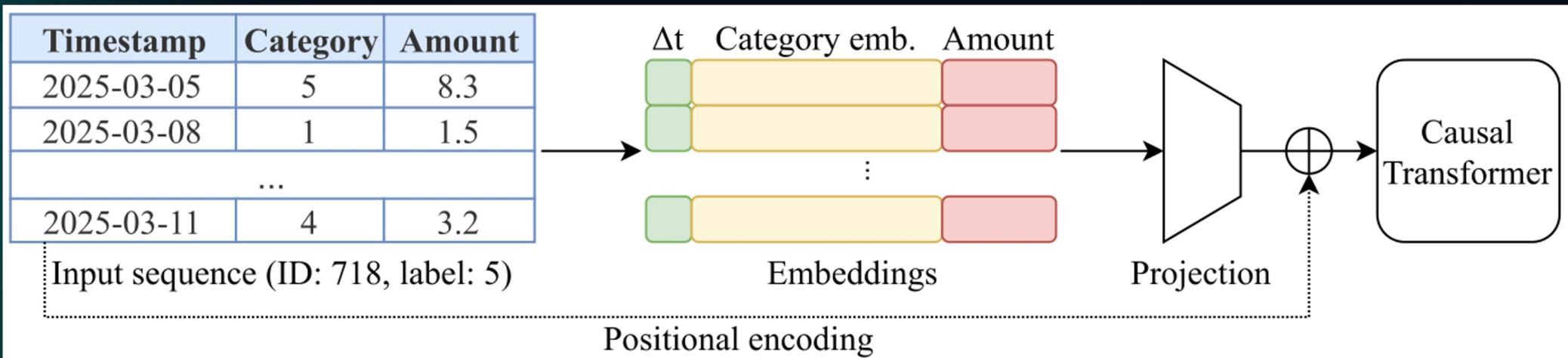
Классификация последовательностей событий

Домен

- Для каждого клиента имеется цепочка событий
- Каждое событие это несколько полей: дата, категория покупки, стоимость, признаки продавца и т.п.

Задача

- Хотим предсказать будущие траты, отток, дефолт
- Размеченных данных мало – нужно предобучение модели
- После предобучения извлекаем эмбединг и используем бустинг для классификации (либо делаем SFT)



Проблема

RNN на многих задачах превосходят трансформеры

- Даже лучшие трансформерные архитектуры проигрывают RNN на малых и средних наборах данных
- На больших наборах данных трансформеры могут давать преимущество, но в разы более затратные

Гипотеза: в трансформерах нет качественного эмбединга

- В RNN скрытое состояние сохраняет всю необходимую информацию о префиксе
- В Transformer, механизм внимания использует активации всех токенов в истории
 - Последний эмбединг содержит мало информации
 - Усреднение всех эмбедингов по истории приводит к потере информации



Почему не contrastive learning

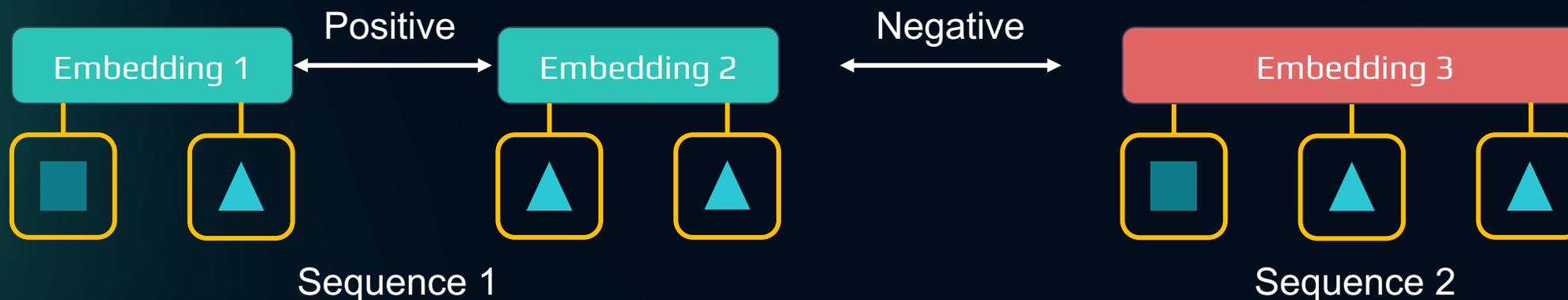
Часто эмбединги предобучают контрастивным подходом

- Сближаются эмбединги подцепочек одного клиента
- Отдаляются эмбединги разных клиентов

Проблемы контрастивных эмбедингов

- Извлекается информация о цепочке целиком, а не то, что нужно для предсказания будущих событий
- Чувствительность к “easy features”

Хотим извлечь эмбединг после NTP pretrain!



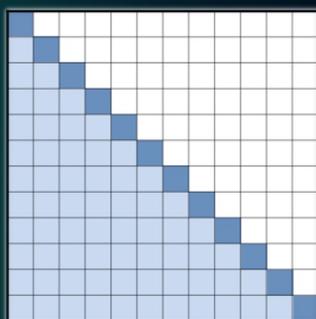
Токены Истории

Цель: сформировать эмбединг во время NTP pretrain

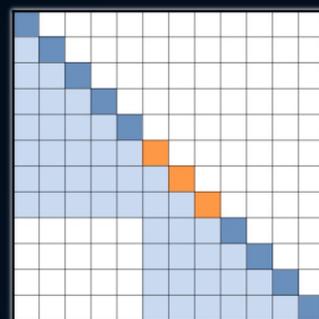
- Добавить <history-token> в несколько мест цепочки
- Иногда использовать их вместо предыдущих событий
- В итоге <history-token> будет содержать эмбединг контекста
- Его можно использовать для дообучения на downstream

Близкие идеи есть в работах, вроде Recurrent Memory Transformer*

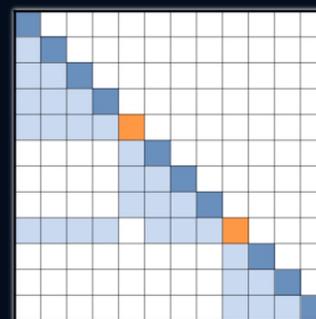
- В них фокус на long context, а не на embedding extraction



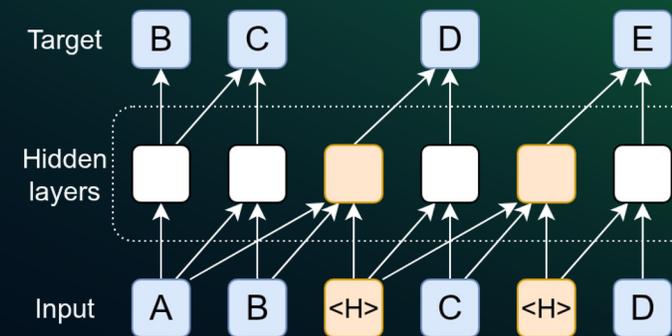
Causal mask



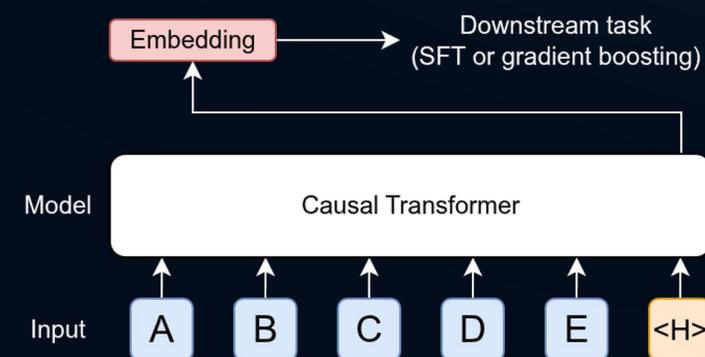
Recurrent Memory



HT-Transformer



(a) NTP Предобучение



(b) Downstream классификация

* Bulatov A., Kuratov Y., Burtsev M. Recurrent memory transformer //Advances in Neural Information Processing Systems. – 2022

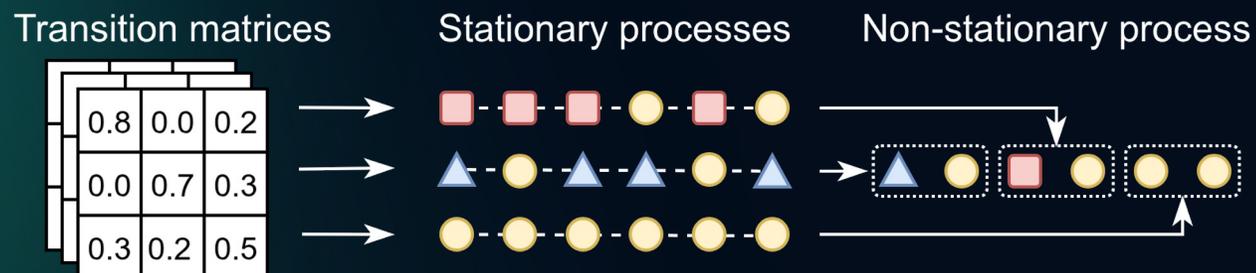
Глобальные и локальные задачи

Простой пример

- Синтетический датасет с нестационарным распределением событий
- В Global задаче нужно оценить число сегментов из каждого источника
- В Local задаче нужно классифицировать тип последнего сегмента

Результаты

- Для global задачи лучше contrastive learning или усреднение эмбеддингов
- Для local задачи лучше HT-Transformer и последний эмбеддинг NTP трансформера



Method	Local (Last part)	Global (Num. parts)
Supervised	0.71	1.00
NTP Last	0.53	0.73
NTP Avg	0.40	0.88
CoLES	0.33	0.94
NTP HT	0.55	0.85

Результаты на реальных данных

- Получили значимый прирост относительно обычного трансформера
 - Прогнозирование оттока
 - Прогнозирование дефолта
 - Предсказание смертности
- SOTA на 3-х датасетах
 - AgePred – это global задача

Отправили статью на конференцию*

Method	Churn	AgePred	Alfabattle	MIMIC-III	Taobao
Supervised RNN	79.10 ± 0.80	61.18 ± 0.49	76.47 ± 1.13	91.46 ± 0.10	84.91 ± 1.17
Supervised Transformer	80.92 ± 0.66	54.88 ± 2.37	74.90 ± 0.08	77.48 ± 1.22	79.71 ± 1.68
NTP RNN	81.56 ± 0.59	60.05 ± 0.29	79.83 ± 0.05	90.68 ± 0.07	83.28 ± 1.42
NTP Transformer	80.92 ± 0.66	56.16 ± 0.51	78.63 ± 0.12	91.28 ± 0.10	83.39 ± 1.43
NTP Rec. Mem. Transf.	80.23 ± 0.21	58.43 ± 0.39	80.25 ± 0.05	91.82 ± 0.04	80.54 ± 0.76
NTP Longformer	81.48 ± 0.66	57.64 ± 0.29	65.91 ± 0.34	89.26 ± 0.18	84.76 ± 1.67
CoLES RNN	82.82 ± 0.28	62.42 ± 0.33	79.30 ± 0.08	87.44 ± 0.20	85.56 ± 1.14
CoLES Transformer	78.92 ± 0.49	59.92 ± 0.30	78.40 ± 0.00	87.06 ± 0.38	82.03 ± 0.98
HT-Transformer	83.34 ± 0.42	60.10 ± 0.39	80.42 ± 0.12	92.00 ± 0.09	84.65 ± 1.07
<i>Impr. over NTP Transf.</i>	+2.42	+3.94	+1.79	+0.72	+1.26

* Karpukhin I., Savchenko A. HT-Transformer: Event Sequences Classification by Accumulating Prefix Information with History Tokens //arXiv preprint arXiv:2508.01474. – 2025.

Возможные темы курсовой / стажировки

Диффузионные модели предсказания событий на горизонт

- В т.ч. с использованием нашего SOTA подхода DeTPP

Развитие HT-Transformer

- Оптимизация скорости за счёт использования Flex Attention
- Исследовать способы агрегации эмбеддингов по слоям

Оптимизация CUDA ядра расчёта лосса DeTPP

- Использовать бакетирование в параллелизме по данным

Развитие методов подбора гиперпараметров лосс функций

Внедрение информации о времени в State Space Models (Mamba и Co)



Искусственный интеллект в медицине

ПОЗНАКОМИМСЯ! Центр Медицины

Николай Романенко,
опыт в разработке ИИ > 4 лет



Руководитель направления в Центре медицины
TG @n_romanenko



Краткая биография:

Более 12 лет в IT бизнесе:

С 2018 года в здравоохранении:

Министерство здравоохранения РФ

Цифромед (ЕГИСЗ)

Neuro.net (112)

Основные направления и проекты:

Анализ ЭМК и диагностика пациентов (AIDA)

Прогностические модели по ЭМК (ССЗ, стационарные риски)

Сигналы (ЭКГ)

Анализ аудиозаписей сердца и легких (AI-стетоскоп)

Провожу лекции для не IT-шников по теме ИИ 😊

Центр Медицины / 14 / 3 Phd

Задача центра:

Применение современных методов искусственного интеллекта для решения задач практического Здравоохранения

Каждый наш проект несет три ценности

1. **Социальную** – мы осознаем социальную ответственность Сбера и создаем решения, помогающие обществу
2. **Практическую** – каждый наш проект имеет потенциал стать продуктом
3. **Научную** – мы находим новаторские применения существующих методов ИИ или улучшаем методы ИИ для прикладных задач



Diagnostic NLP Science Papers X-ray CT
Prognosis LLM (GigaChat) Patient Disrupt CV localization segmentation pathology
EHR Signal Integration Social Multimodal
GenAI AI-agents Foundation models

ТОП-3

Сервис ИИ, который помогает врачам ставить 3 предварительных диагноза по жалобам пациента



Чемпион конкурса WSIS Prizes 2021, проводимого ООН

10

Субъектов РФ используют ТОП-3 в медицинской практике

14,4 млн.

Прибыль в рублях от внедрения сервиса в регионы РФ за 2023 г.

>20 млн.

Предварительных диагнозов помог врачам поставить ИИ

95%

От всех возможных диагнозов на первичном приеме

265

Групп наиболее вероятных диагнозов по МКБ-коду

AIDA

Сервис ИИ для постановки окончательного диагноза на основании данных электронной медицинской карты

30 млн.

Объем данных, использованных для обучения модели

87,7%

Точность модели по результатам эксперимента с врачами-экспертами

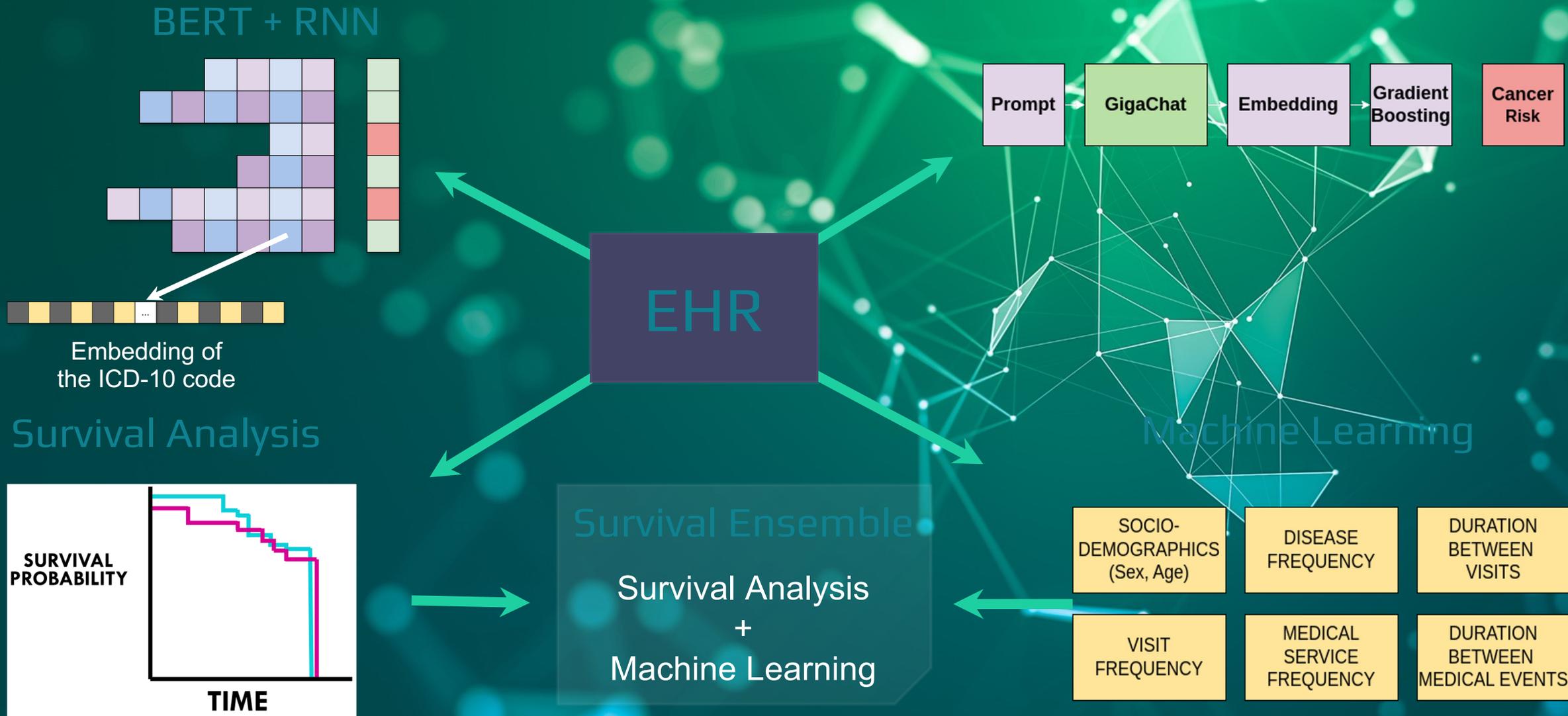
95

Групп наиболее важных диагнозов по МКБ-коду

252

Внедрена во все взрослые поликлиники г. Москва и готовится к внедрению в Липецкой области

Survival analysis OHKO/CC3



CV в медицине

Модели для диагностики с целью выбора оптимальной стратегии лечения

КТ Инсульт с расчетом ASPECTS

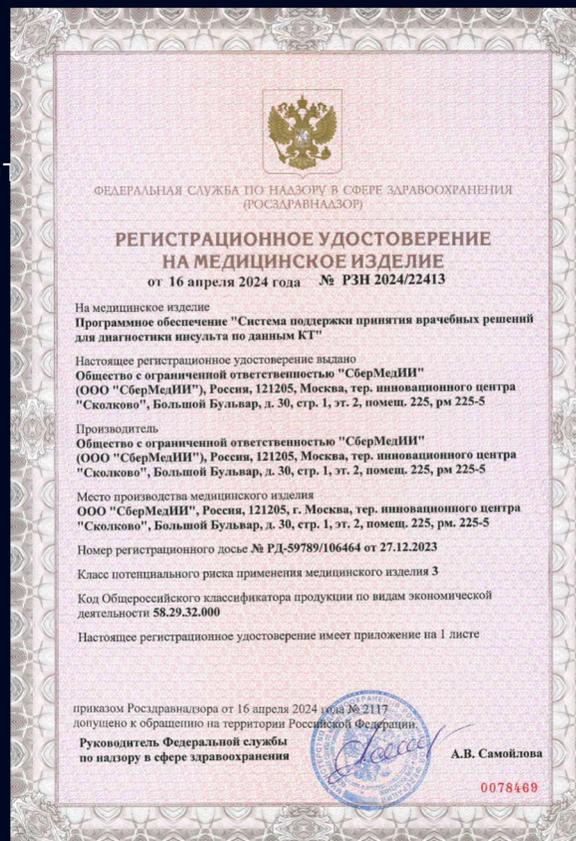
- СберМедИИ получено
РУ Росздравнадзора
16.04.2024 № РЗН 2024/22413

4,5 млн руб.

- финансовый эффект за 2023 г.
- Получение РУ позволяет масштабировать использование AI-сервиса в регионы (пилоты в Новосибирске и Тюмени)



mosmed.ai

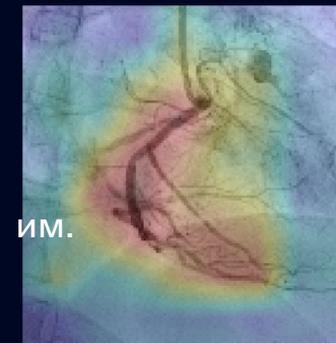


Расчет SYNTAX по данным коронароангиографии

- Статья об исследовании стала основой главы по ИИ в **Национальном руководстве** по рентгенэндоваскулярной хирургии

~**17** млн руб.

договоренность о покупке модели Умной Клиникой в рамках центра Алмазова и Сбера



им.

- Запуск пилота для пациентов на сайте партнера infarkta.net



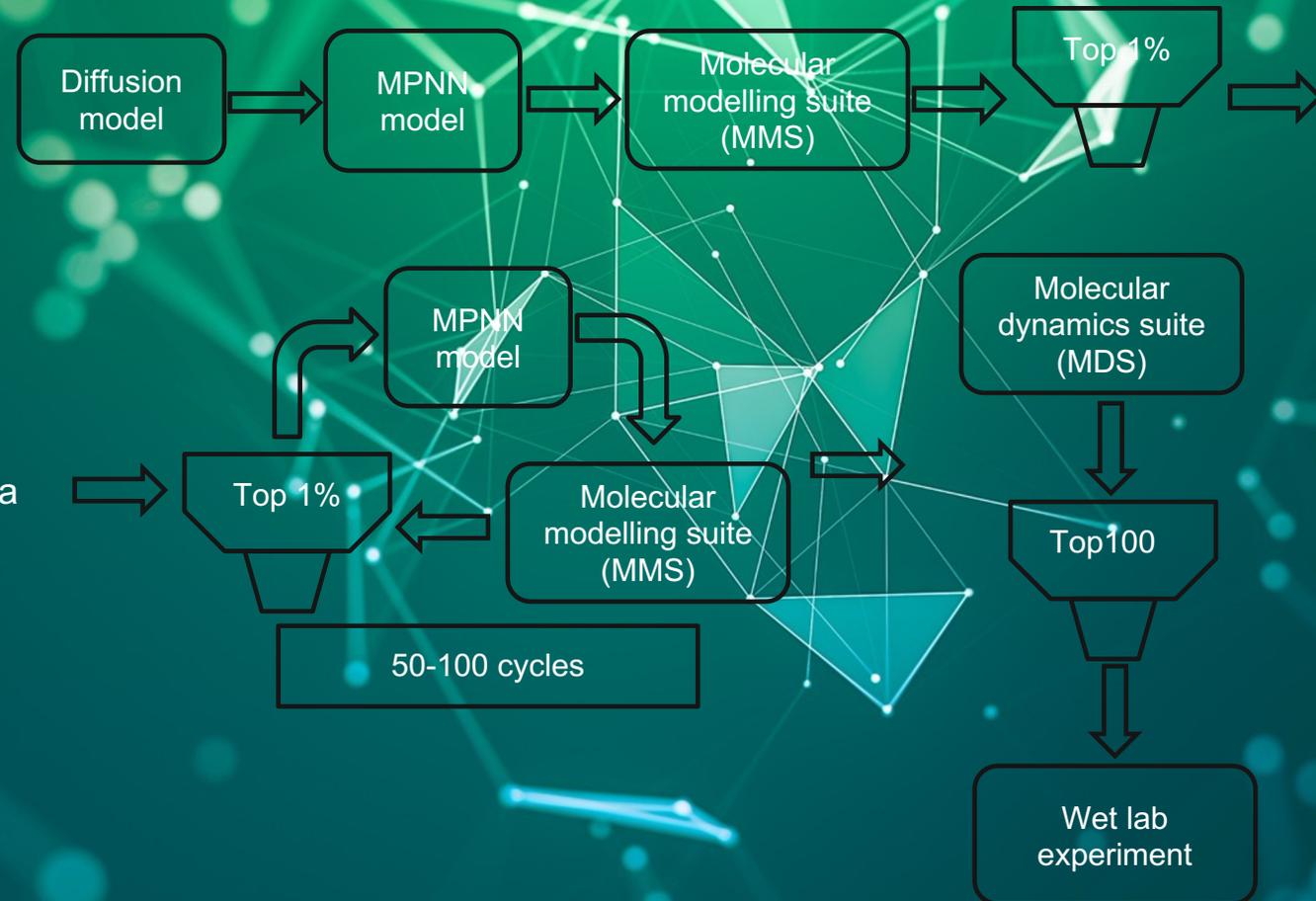
ФИЛИАЛ ТОМСКОГО НИМЦ
ТЮМЕНСКИЙ
КАРДИОЛОГИЧЕСКИЙ
НАУЧНЫЙ ЦЕНТР

Биоинформатика

Задача: Дизайн верблюжьего моноклонального антитела к рецептору CD137. Антитело активирует рецептор, который в свою очередь сигнализирует клеткам иммунной системы усиливать поиск и борьбу с опухлевыми клетками.

Исходные данные:

- Antibody frame - каркас антитела, в котором нет гипервариабельных участков связывания, которые мы дизайним.
- Epitope AAs - набор аминокислот, которым задается эпитоп (участок таргета, на который садится антитело) для модели.
- 3D structure - набор трехмерных координат для каждого атома белка. И антитело и таргет белковой природы.
- Desired AA distribution - заданное фиксированное распределение вероятности сгенерировать определенную аминокислоту, полученное из литературных данных.
- Complex - два белка (антитело и рецептор-таргет), совмещенные друг с другом в месте эпитопа.
- Contact surface area - площадь поверхности контакта антитела и таргета.



GigaDoc - disrupt-инициатива

Создание мультимодального медицинского ассистента

Ответ: На основании предоставленных данных, ваш пульс находится в пределах нормы для взрослого человека. Однако, риск инфаркта может быть связан не только с частотой пульса, но и с другими факторами, такими как уровень стресса, индекс массы тела и общее состояние здоровья. Если у вас есть какие-либо опасения или симптомы, связанные с сердцем, рекомендуется обратиться к врачу для более детального обследования и консультации.

Исходя из моего пульса, есть ли у меня риск инфаркта?

Исходя из вашего уровня стресса, я бы рекомендовал вам избегать интенсивных физических нагрузок и занятий спортом, которые могут вызвать дополнительный стресс. Вместо этого, попробуйте заниматься более спокойными видами активности, такими как йога, пилатес или плавание. Они помогут вам расслабиться и снизить уровень стресса. Также не забывайте о достаточном отдыхе и сбалансированном питании.

Есть ли какие-то ограничения по спорту исходя из моего уровня стресса?

Активирован режим Gigaevt

Привет! Я – GigaChat, генеративная языковая модель. Моя задача – отвечать на вопросы пользователей, вести диалог и помогать в решении различных задач. Я умею искать информацию, писать тексты разных форматов, создавать изображения и



Начать диагностику

Подойдите ближе к экрану так, чтобы шкала стала полностью зеленой. Не уходите до окончания измерения.



Начать диалог

Зажмите кнопку "Микрофон" на пульте и говорите в его верхнюю часть.



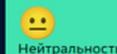
Сохранить результат

Сфотографируйте экран, если хотите сохранить данные диагностики.

Основные данные

Пол Женский Возраст 25 ±1

Эмоциональное состояние



Нейтральность

Индекс массы тела

21.1
В норме

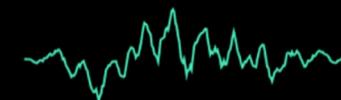
Риск диабета II типа

0.025
В норме

Данные диагностики

Пульс

74 



Отклонение интервалов ритма сердца

5.3
В норме

Уровень стресса

87.2
В норме

Восстановление организма

7.6
В норме

PNN50

0%
В норме

Артериальное давление

 В разработке

Температура тела

 В разработке

Сатурация

 В разработке

Частота дыхательных движений

 В разработке

Гликированный гемоглобин

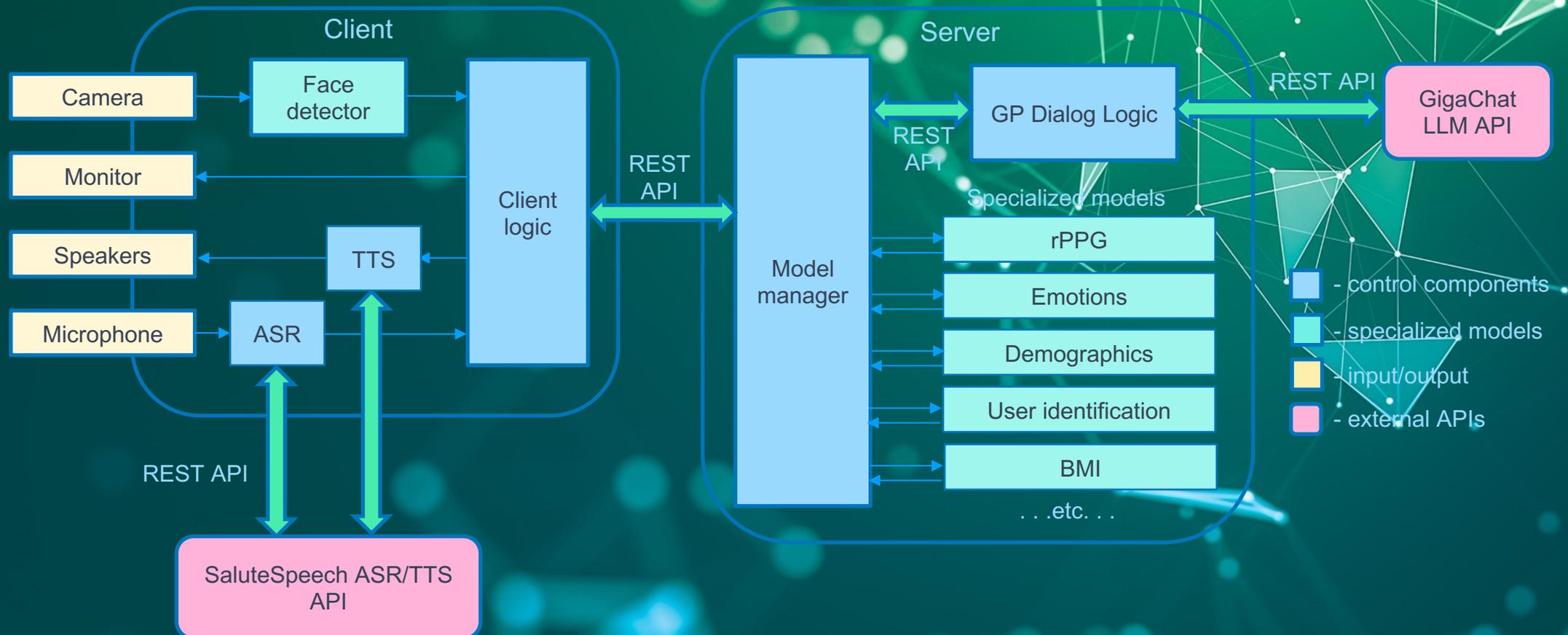
 В разработке



GigaDoc - disrupt-инициатива

Создание мультимодального медицинского ассистента

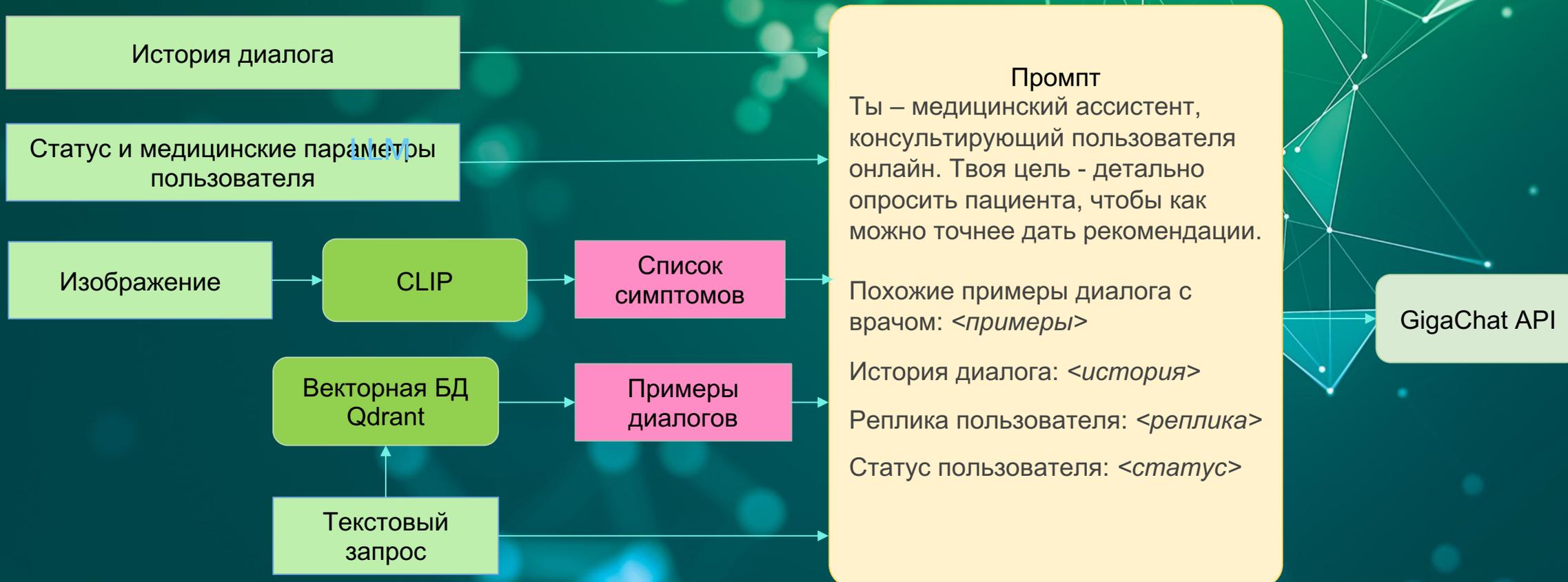
GP model: Клиент-серверная архитектура с 10+ элементами на Python



GigaDoc - disrupt-инициатива

Создание мультимодального медицинского ассистента

GP Dialog Logic: Управление всей логикой медицинского ассистента с помощью LLM

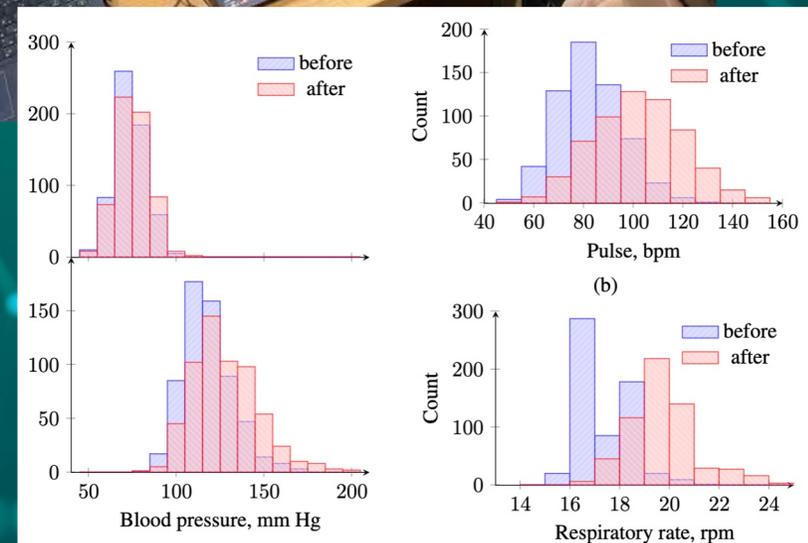
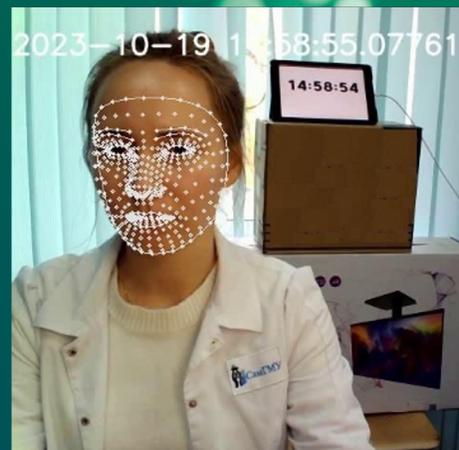


GigaDoc - disrupt-инициатива

Создание мультимодального медицинского ассистента

Remote photoplethysmography (rPPG).
Сбор данных:

- 600 участников
- Синхронное видео с трех ракурсов
- 2 сессии – до приседаний и после
- Синхронизация с PPG
- Дополнительные данные: ЭКГ и лабораторные анализы



GigaDoc - disrupt-инициатива

Создание мультимодального медицинского ассистента

Модель определения Индекса Массы Тела (ИМТ) по лицу человека

Создание датасета

1. Сбор фотографий и информации об ИМТ с форума [/progresspics](#)



2. Детекция лиц с помощью [YOLO-v7](#)



3. Фильтрация лиц, закрытых телефоном через их поиск и сегментацию с [Detectron2](#)



4. Фильтрация отвернутых лиц через соотношение ключевых точек с помощью [mediapipe](#)



Обучение модели

- В качестве архитектур мы опробовали предобученные ResNet, EfficientNet, ConvNext, ViT; в приоритете были легковесные архитектуры
- Подбор гиперпараметров осуществляли с помощью [Optuna](#)
- Применили большое количество [аугментаций](#) (геометрические, цветовые, mixup)

Инференс

- Изначально лучше всего на тесте себя показала архитектура [convnext-base](#)
- Однако, тестирование «в боевых условиях» на стенде с разным освещением показало, что наиболее качественный результат даёт [convnext-small](#)



Курсовые и дипломные работы

Возможные темы для взаимодействия (1)

Эмоциональный ИИ:

- Эффективные мультимодальные алгоритмы мультизадачного распознавания эмоций на видео
- Распознавание синтетических видео на основе анализа микро-выражений лиц для онлайн конференций
- Вычислительно эффективный синтез эмоциональных аудио / видео лиц для вопросно-ответных систем
- Контрастное обучение для автоматизации подбора промптов языковых моделей для диалоговых систем
- Мультиагентное теоретико-игровое моделирование эмоциональных манипуляций
- Протоколирование оффлайн-переговоров на основе нейросетевых моделей обработки речи и больших языковых моделей
- Оценка личностных качеств и эффективности собеседования с помощью асинхронных видеointервью

Рекомендательные системы

- Парное выравнивание текстовых и коллаборативных эмбеддингов для решения проблемы холодного старта в рекомендательных системах
- Предобучение моделей рекомендаций на последовательностях

ИИ в медицине

- Резервуарные вычисления и Эхо-сети (Echo State Networks) для анализа и диагностики медицинских сигналов

Возможные темы для взаимодействия (2)

Deep learning для последовательностей событий

- MAMBA-подобные архитектуры для распознавания нерегулярных временных рядов и последовательностей событий
- Диффузионные генеративные модели для моделирования последовательностей событий на основе метода обучения DeTRP
- Дистилляция знаний между LLM и моделями на структурированных данных
- Автогенерация интерпретируемых признаков для обогащения эмбедингов с помощью LLM
- Прогнозирование транзакционных последовательностей с использованием объяснимых State Space Models
- Геометрические методы сравнения пространств эмбедингов
- Топологическая метрика для оценки качества эмбедингов без учёта меток

DS-агент

- Автоматизированный агент аналитики данных для оптимизации бизнес-операций на основе больших языковых моделей
- Семантический поиск продуктов для разговорных неопределённых запросов
- Большие языковые модели для объяснимого ИИ в моделях и задачах машинного обучения

Fall into ML 2025

24-25 октября 2025 г

- Постерная сессия всех российских авторов A* статей за 2025 год
- Наш воркшоп Financial AI (25 октября в 13:00)



<https://cs.hse.ru/ml2025/>



Ссылки на ресурсы о SBER AI LAB и ПОЛЕЗНЫЕ КОНТАКТЫ

GitHub
sb-ai-lab



Сайт о проектах
Sber AI Lab



Вакансии
Sber AI Lab



E-mail (для ВКР)*



По вопросам
стажировок,
сотрудничества
и др.



* andrey.v.savchenko@gmail.com