АРХИТЕКТУРА ЭКОСИСТЕМЫ ГЕНЕРАТИВНОГО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

ФИНТЕХ АССОЦИАЦИЯ

ген ии

Версия v.1.0

Генеративные ИИ-приложения и сервисы

Горизонтальные и вертикальные прикладные решения, доступные одновременно по модели on-premise и в облачном развертывании

Инструменты для генеративного ИИ

Инструменты для эффективного и безопасного промышленного использования базовых моделей генеративного ИИ

Модели генеративного ИИ

Модели и инструменты генеративного ИИ, реализующие базовые фукнции технологии

Инфраструктура

Техническая инфраструктура для обучения и применения (инференса) инструментов на основе генеративного ИИ

Горизонтальные приложения

- Финансы: BloombergGPT, FinanceGPT, J.P.Morgan
- Обслуживание клиентов: Gerwin, radoar
- Продажи: Microsoft, AI SEARCH, Apollo.io, DataRobot
- Продуктивность: copy.ai, RoboGPT, Notion
- Информационные технологии: GitHub Copilot, ∫NNs, AI, CRUDERRA
- Управление знаниями: character.ai, MathGPT, Notion, VISPER, DeepL

Вертикальные/ отраслевые приложения

- Финтех: НейроТекстер, Copilot
- Естественные науки: Recursion, NVIDIA BioNeMo, Isomorphic Labs
- Здравосохранение: zakipoint Health, IBM, ARTERA
- PR/маркетинг: CopyMonkey, YandexGPT 2, Jasper, AI, Tome, Николай Иронов, Adobe Firefly
- Юриспруденция: Humata, LEGALESE DECODER, ailawyer
- Ритейл: Meta
- Транспорт: GHOST
- Робототехника: Google

Кейс-специфичные

- Платформы для чат-ботов: Replika, Cleverbots, Inflection Pi, alta
- Чат-боты (сочетание классических подходов с LLM): botpress
- Агенты: Jay Copilot, Just AI, imbue
- Транскрибация, суммаризация и перевод текста: AI, rev ai, happyscribe, DeepScribe, YandexGPT 2, cohere
- Генеративные презентеры: synthesia, D-ID
- Генерация видео: GEN-2, Sora, LUMIERE

Моделе-центричные

- Разработка модели: GitHub, OpenAI, Google colab, jupyter
- Риск-ориентированное использование моделей: LexCheck, Timely, Cerebrix
- Retrieval Augmented Generation (RAG): Microsoft, Azure AI Search, Just AI, Confident AI, Jay Copilot
- Файн-тьюнинг: Jay Copilot, Just AI, AI

Дата-центричные

- Промпт инжениринг: Spellbook, vellum, Promptify
- Векторные базы данных и эмбедингги: Pinecone, qdrant, Weaviate, chroma ai, milvus

Подготовка и обмен данными

- Федеративное обучение: Google, IBM
- Обезличивание: DATPROF, ataccama, DEEP DOCS
- Гомоморфное шифрование: Google, Microsoft
- Инструментарий для формирования датасетов: sweetviz, Superset
- Синтетические данные: hazy, kaview, Datomize, scale, surge, Synt Data
- Разметка данных: TAGME, Toloka, Label Studio

AI TRISM-инструменты

- Мониторинг отравления и искажения промптов: WHYLABS, Prompt:
- Контроль галлюцинаций: LAKERA, Lasso, WHYLABS
- Контроль деанонимизации данных: LAKERA, CALYPSOAI, vigil ai
- Модерация результатов (анти-байес): LLM GUARD, ROBUST INTELLIGENCE, vigil ai
- Обеспечение безопасности ИИ-приложений: ROBUST INTELLIGENCE, Lasso, Prompt:, rebuff, LLMFuzzer, LAKERA, SLASHNEXT, vigil ai, BurpGPT, LLM GUARD
- Объяснимость результатов ИИ: vigil ai, LLM GUARD, BurpGPT
- Предотвращение кибератак с ИИ: Jericho Security, CALYPSOAI, BRIGHTSIDE, ROBUST INTELLIGENCE, ADVERSA, LLM GUARD

Модели генеративного ИИ

- GPT-модели: saiga_mistral, YandexGPT 2, AI, Gemini, ANTHROPIC, LLAMA 2, GIGA CHAT, DALL·E 3, Bard, Falcon
- Дифузионные модели: Stable Diffusion, Kandinsky 3.0, Шедеврум, stability.ai
- Доменные модели: AI PICASSO, ChefGPT, Николай Иронов, BloombergGPT
- Маркетплейсы генеративных моделей: cloud.ru, Yandex Cloud, Hugging Face

Технологические инструменты

- Инструменты MLOps: DataSphere, mlflow, Kubeflow, SinaraML, SELDON
- Фронтенд для нейросетей: stable-deffusion-webui, text-generation-webui
- NoCode|LowCode платформы: Directual, appflowy, NOCODB, NocoBase
- Оптимизация нейромоделей: run:ai, deci, Skoltech
- Оркестрация моделей по API: Hubble, stack, Patterns

Инфраструктура

- Вычисления для обучения: NVIDIA, HUAWEI, ЭЛЬБИС, AMD, cerebras, TESLA, IVA COGNITIVE, intel, Google
- Вычисления для интерфейса: intel, AMD, HUAWEI
- Программируемые/специализированные интегральные схемы: AMD XILINX, LATTICE, Fudan Microelectronics Group, GOWIN
- Сети связи: CISCO, JUNIPER NETWORKS, NVIDIA, H3C, EDGE-CORE NETWORKS
- Интерконнект: NVIDIA, ИНТЕРКОННЕКТ, Ростелеком
- Облачные решения: #CloudMTS, cloud.ru, Selectel, Yandex Cloud, Azure, aws, Google Cloud
- Анклавные вычисления (изолированные): ВТБ
- Приватные вычисления (Privacy Enhanced Computation): Azure AI Search, Microsoft, СБЕР
- Средства обеспечения базовой безопасности: positive technologies, BI.ZONE, KOA безопасности, kaspersky, TCC

Авторы

Максим Григорьев, Алексей Сидорюк, Евгений Морозов, Валерия Митина, Варвара Ольшницкая

Контрибьютеры

Александр Долбнев (Яндекс), Полина Гришина (Сбер Девайсы), Феликс Скворцов (МТС), Василий Кирсанов (CloudAmplifier), Николай Бушков (Альфа Банк), Сергей Колесников (Тинькофф), Владислав Тушканов (Лаборатория Касперского), Ильяс Киреев (Positive Technologies)

@FINTECHASSOCIATION