ПАВЕЛ ВАСИЛЬЕВ

Москва, Россия

+7(922) 176-0851 ◇ vasilievpavel4@gmail.com ◇ github.com/PavelPaha ◇ tg ◇ LinkedIn

ОБРАЗОВАНИЕ

Студент, Школа анализа данных Яндекса
2024–2026
Ключевые курсы: Классическое машинное обучение, Компьютерное зрение, Обучение с подкреплением, Обработка естественного языка.

Фундаментальная информатика и информационные технологии, УрФУ (Екатеринбург)
2022–2026
Ключевые курсы: Математический анализ, Линейная алгебра, Алгоритмы и структуры данных, Теория вероятностей, Статистика, Машинное обучение, Сети и протоколы интернета, Программирование на C# и Python, Git, Проектная деятельность.

НАВЫКИ

Технические навыки	C#, C++, Python, PostgreSQL, Git, NumPy, PyTorch, Hydra, DVC, Wandb
Исследовательские навыки	Чтение и воспроизведение экспериментов из статей на arxiv.org
Soft Skills	Умение работать в команде

ИССЛЕДОВАТЕЛЬСКИЙ ОПЫТ

Federated Learning School (MZBUAI × YSDA)
Изучение смеси экспертов, баланс между целевой функцией и функцией баланса, анализ работы разных роутингов для экспертов.
· Наставники: Martin Takac, Aleksandr Beznosikov

Текущие исследования: Внедрение и анализ интерпретируемости sparse-архитектур (MoE) в унифицированные мультимодальные диффузионные модели.

Работаю в команде энтузиастов из ШАД, и мы исследуем унифицированные трансформерные архитектуры и способы повышения их эффективности для генеративных задач. Мы концентрируемся на разреженных архитектурах, где разные эксперты специализируются на отдельных модальностях. Также мы разрабатываем shared эксперты, которые позволяют "переиспользовать" информацию из разных модальностей.

ОПЫТ РАБОТЫ

СКБ Контур Стажировка
Младший инженер-программист (C# Backend) *Екатеринбург, июль 2024 — сентябрь 2024*

· Команда: Hosting and Deployment.
· Поддержка динамических сущностей для деплоймента.
· Работа с распределёнными системами.
· Продукт: система, похожая на Kubernetes, для внутренних сервисов компании.
· Длительность: 2 месяца (full-time).

ООО «Яндекс» Стажировка
Инженер NLP *Москва, апрель 2025 — июль 2025*

· Команда: разработка голосового помощника Алиса / команда NLP / подгруппа Scripted Models.
· Разработка функциональности function calling для Алисы.
· Длительность: 2 месяца (full-time).

ПАО «Сбербанк»
Middle DL Engineer *Москва, август 2025 — настоящее время*

· Команда: Gigacode — модель автодополнения кода.
· Разработка ассистента для автодополнения кода в IDE.

ПРОЕКТЫ

Инженерный проект: InferD — распределённый движок инференса LLM.

Реализовал децентрализованный движок для инференса больших языковых моделей на нескольких узлах. Разработал механизм балансировки нагрузки с использованием распределённых хеш-таблиц. Идея вдохновлена подходами из статей SWARM и Petals (arxiv).