ИСП РАН

Разработка безопасного ПО, реализующего технологии ИИ

ПАДАРЯН В.А.
vartan@ispras.ru

СОСНИНА Е.С.
esosnina@ispras.ru

ТУРДАКОВ Д.Ю.
turdakov@ispras.ru

Форум «Технологии доверенного искусственного интеллекта»
13 мая 2026 года

Предпосылки

- Правовое регулирование искусственного интеллекта должно опираться на реальные возможности технологий
- Уже наблюдается игнорирование предложений по саморегулированию ИИ и других законодательных инициатив, для которых нет четких критериев и инструментов проверки
- Требования к любым доверенным системам, в том числе – к системам ИИ, должны опираться на существующие технологии и соответствовать их уровню

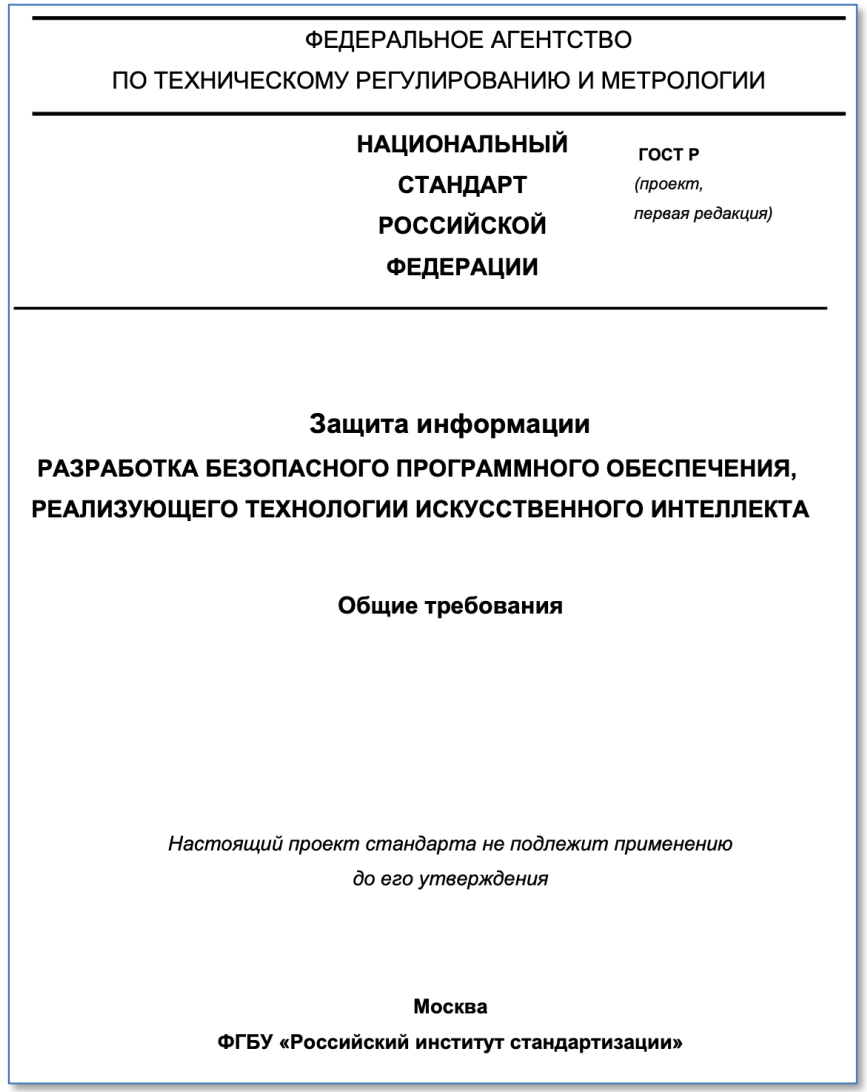
Проект стандарта РБПО ИИ: объект и аспект стандартизации

1. ИСП РАН в рамках деятельности технического комитета по стандартизации ТК 362 «Защита информации» (ФСТЭК России) разработан проект национального стандарта «Защита информации. Разработка безопасного программного обеспечения, реализующего технологии искусственного интеллекта».

2. Стандарт определяет требования к процессам разработки безопасного ПО, реализующего технологии искусственного интеллекта: **модели ИИ и ПО, обеспечивающего ее функционирование.**

3. Процессы разработки **основаны на положениях ГОСТ Р 56939-2024**, **модифицированы (дополнены) с учетом специфики** предметной области.

4. Безопасность ПО, реализующего технологии ИИ, обеспечивается **совокупностью процессов разработки ПО** и выполнением требований к ним, **включая, но не ограничиваясь, выполнением специфичных требований к модели ИИ**.



ФЕДЕРАЛЬНОЕ АГЕНТСТВО
ПО ТЕХНИЧЕСКОМУ РЕГУЛИРОВАНИЮ И МЕТРОЛОГИИ

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

ГОСТ Р
(проект, первая редакция)

Защита информации
РАЗРАБОТКА БЕЗОПАСНОГО ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ, РЕАЛИЗУЮЩЕГО ТЕХНОЛОГИИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Общие требования

Настоящий проект стандарта не подлежит применению до его утверждения

Москва
ФГБУ «Российский институт стандартизации»

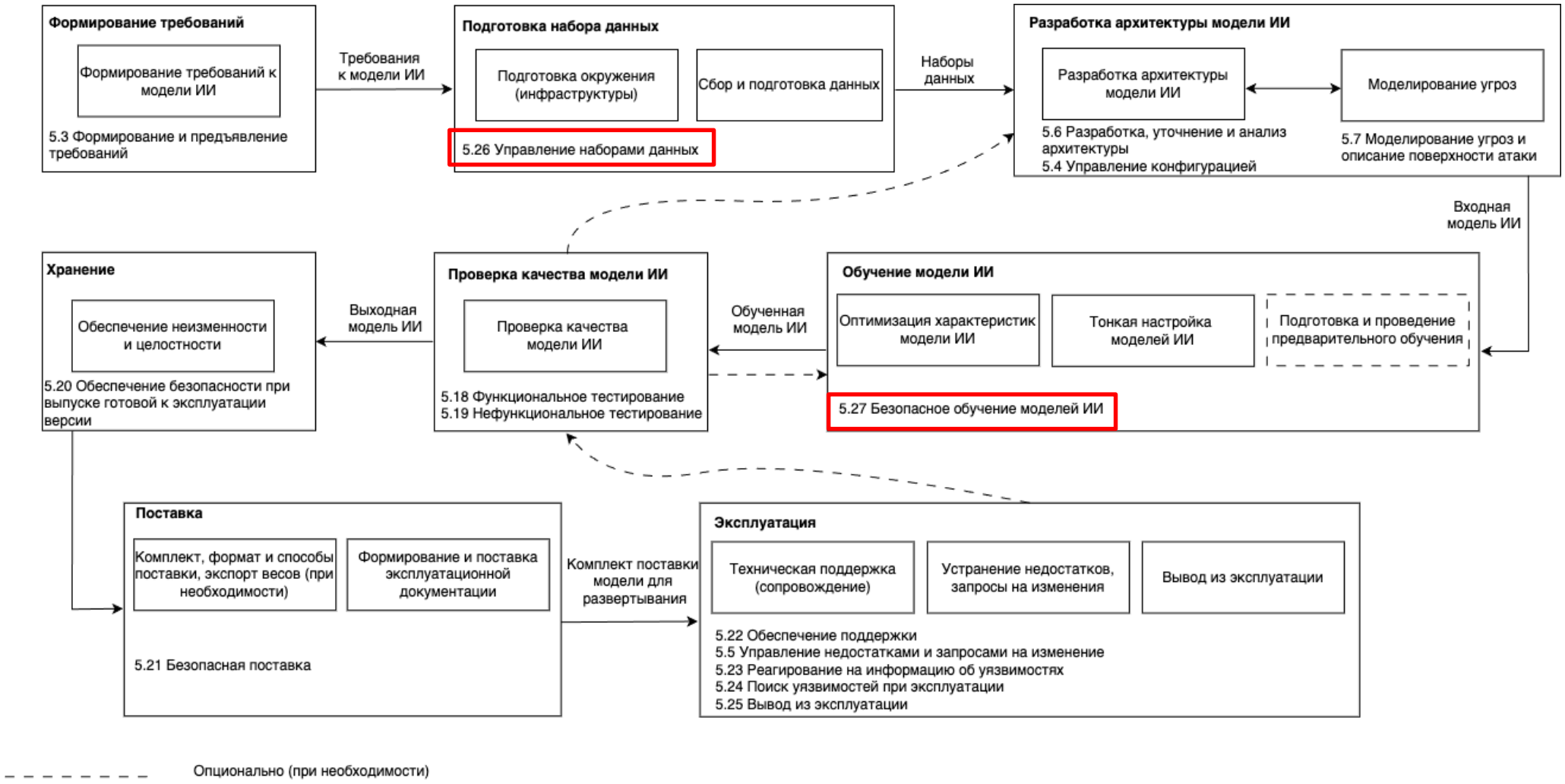
Сопоставление стадий разработки систем ИИ по ГОСТ Р 71539-2024 с процессами ГОСТ Р РБПО ИИ и с указанием их модификации относительно ГОСТ Р 56939-2024

Начальная стадия	Проектирование и разработка			Верификация и валидация	Развертывание	Эксплуатация и мониторинг	Вывод из эксплуатаци
5.1 Планирование процессов разработки (типовой)	5.4 Управление конфигурацией ПО (модифицированный)	5.8 Формирование и поддержание в актуальном состоянии правил кодирования (типовой)	5.26 Управление наборами данных (специфический)	5.18 Функциональное тестирование (модифицированный)	5.20 Обеспечение безопасности при выпуске готовой к эксплуатации версии (типовой)	5.22 Обеспечение поддержки программного обеспечения при эксплуатации пользователями (типовой)	5.25 Обеспечение безопасности при выводе программного обеспечения из эксплуатации (типовой)
5.2 Обучение сотрудников (типовой)	5.6 Разработка, уточнение и анализ архитектуры ПО (модифицированный)	5.9 Экспертиза исходного кода (типовой)	5.27 Безопасное обучение моделей ИИ (специфический)	5.19 Нефункциональное тестирование (модифицированный)	5.21 Безопасная поставка программного обеспечения пользователям (модицифицированный)	5.5 Управление недостатками и запросами на изменение (модифицированный)	
5.3 Формирование и предъявление требований к ПО (модифицированный)	5.7 Моделирование угроз (модифицированный)	5.10 Статический анализ исходного кода (типовой)	5.16 Использование инструментов композиционного анализа (модифицированный)			5.23 Реагирование на информацию об уязвимостях (модифицированный)	
	5.12 Использование безопасной системы сборки (типовой)	5.11 Динамический анализ кода программы (типовой)	5.17 Проверка кода на предмет внедрения ВПО через цепочки поставок (модифицированный)			5.24 Поиск уязвимостей в программном обеспечении при эксплуатации (типовой)	
	5.13 Обеспечение безопасности сборочной среды (модифицированный)	5.15 Обеспечение безопасности используемых секретов (типовой)					
	5.14 Управление доступом и контроль целостности кода при разработке (типовой)						

Легенда

Типовой процесс	Процессы, требования к которым идентичны тем, что определены в ГОСТ Р 56939-2024, и применяются с учетом области применения настоящего стандарта	Модифицированный процесс	Процессы, требования к которым были дополнены по сравнению с их определениями в ГОСТ Р 56939-2024	Специфический процесс	Процессы, являющиеся специфическими для ПО, реализующего технологии ИИ, и не основанные на каких-либо процессах, определенных в ГОСТ Р 56939-2024

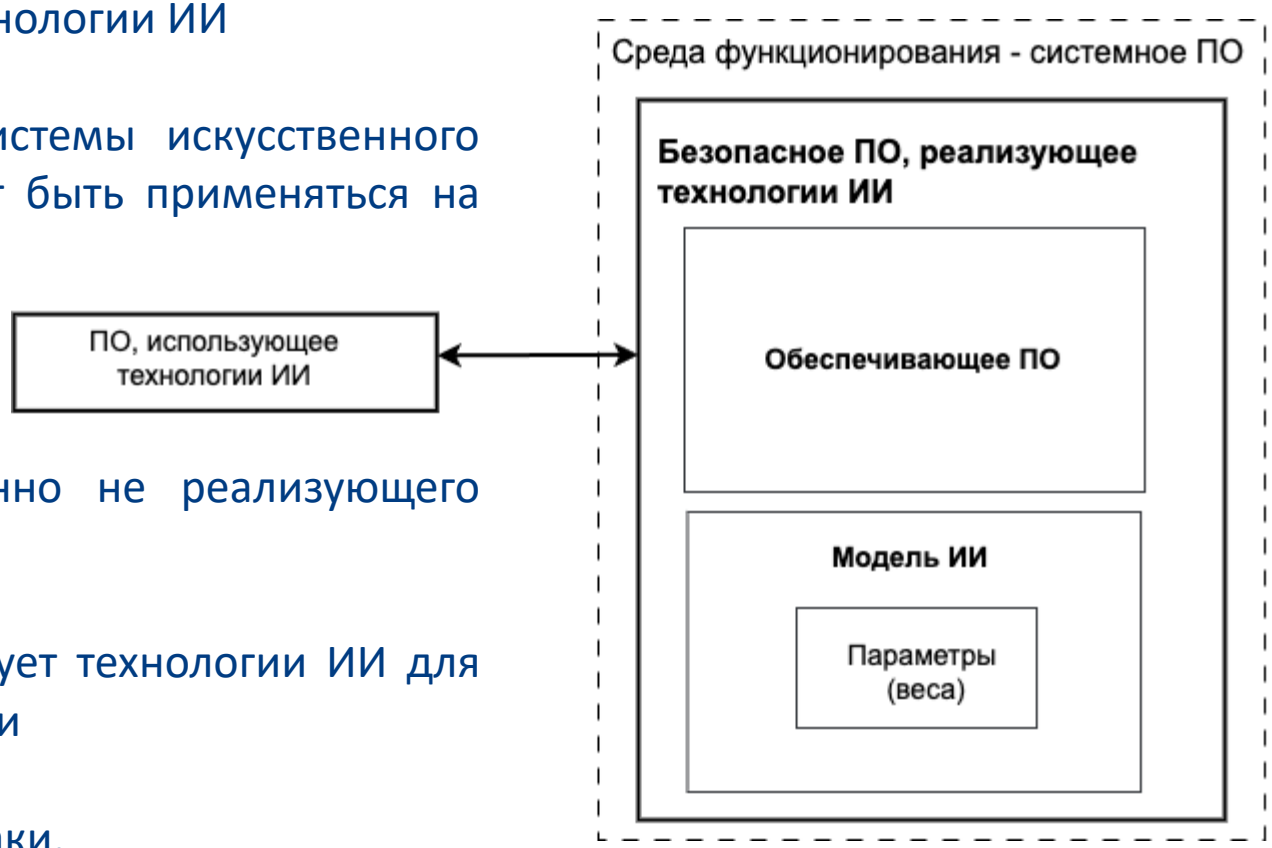
Рекомендуемая модель жизненного цикла моделей ИИ



РБПО ИИ: область применения

Область применения стандарта — ПО, **реализующее** технологии ИИ

Требования установлены не к жизненному циклу системы искусственного интеллекта, а к **процессам разработки**, которые могут быть применяться на разных этапах ее жизненного цикла.



Вопросы разработки безопасного ПО, непосредственно не реализующего технологии ИИ, регулируются ГОСТ Р 56939-2024.

При этом в случае, если такое безопасное ПО использует технологии ИИ для реализации своих функциональных возможностей, то при

- разработке и уточнении его архитектуры,
- моделировании угроз и определении поверхности атаки,
- а также функциональном и нефункциональном тестировании **необходимо учитывать специфичные требования**, определенные стандартом

Состав ПО, реализующего технологии ИИ

ПО, реализующее технологии ИИ, включает в свой состав:

1) **модели ИИ (параметры (веса) моделей ИИ)**;

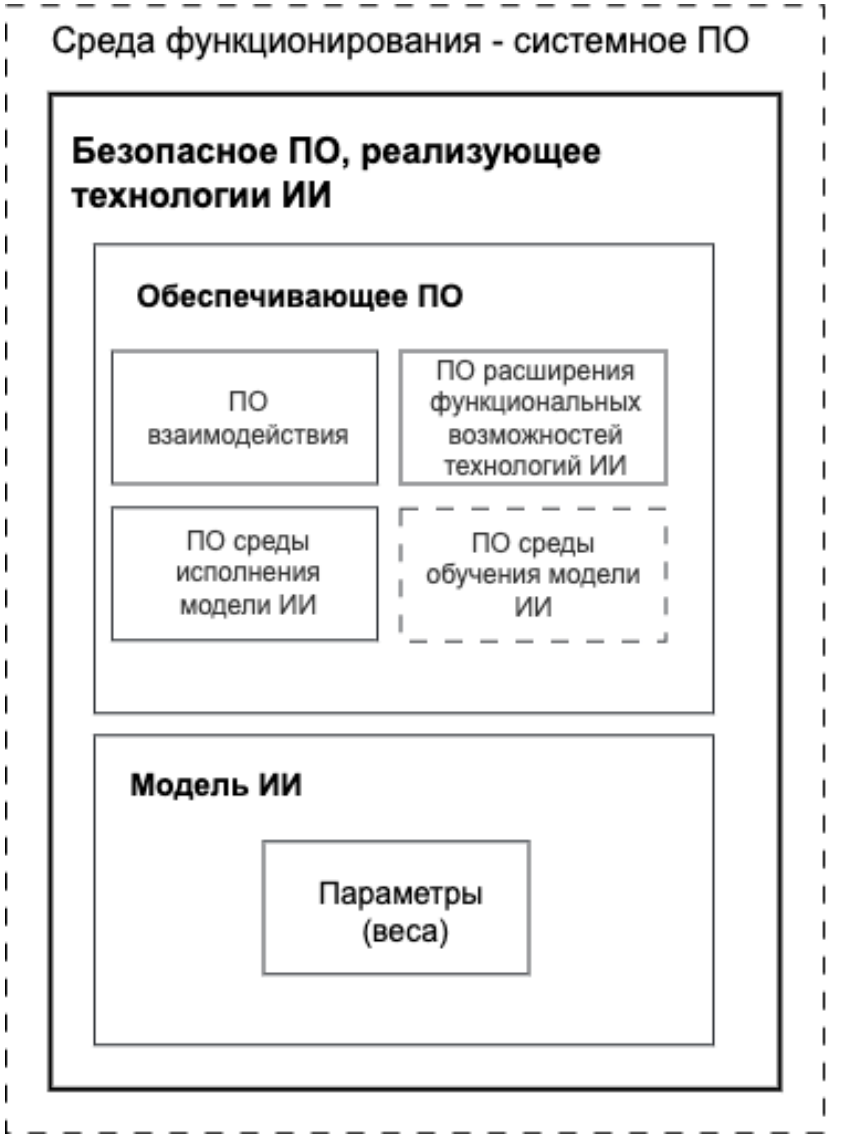
2) **ПО, обеспечивающее:**

а) взаимодействие с моделью ИИ и (или) с прикладным, системным и другим ПО, использующим модели ИИ для реализации своих функциональных возможностей (в том числе с системами ИИ);

б) среду исполнения модели ИИ: выполнение моделей ИИ и предоставление результатов их работы;

в) расширение функциональных возможностей технологий ИИ;

г) обучение моделей ИИ* (опционально).



* ПО, обеспечивающее обучение моделей ИИ, включается в состав ПО, реализующего технологии ИИ, в случае, когда декларируемые разработчиком функциональное назначение и функциональные возможности разрабатываемого ПО предполагают самостоятельное обучение моделей ИИ пользователем (эксплуатирующей организацией) на этапе эксплуатации.
В ином случае, ПО среды обучения в состав ПО, реализующего технологии ИИ, — не входит, к нему предъявляются требования в контексте требований к реализации процессов согласно стандарту.

Состав обеспечивающего ПО

ПО, обеспечивающее среду исполнения модели ИИ: программы и библиотеки, необходимые для развертывания, запуска и функционирования моделей ИИ, в т.ч.

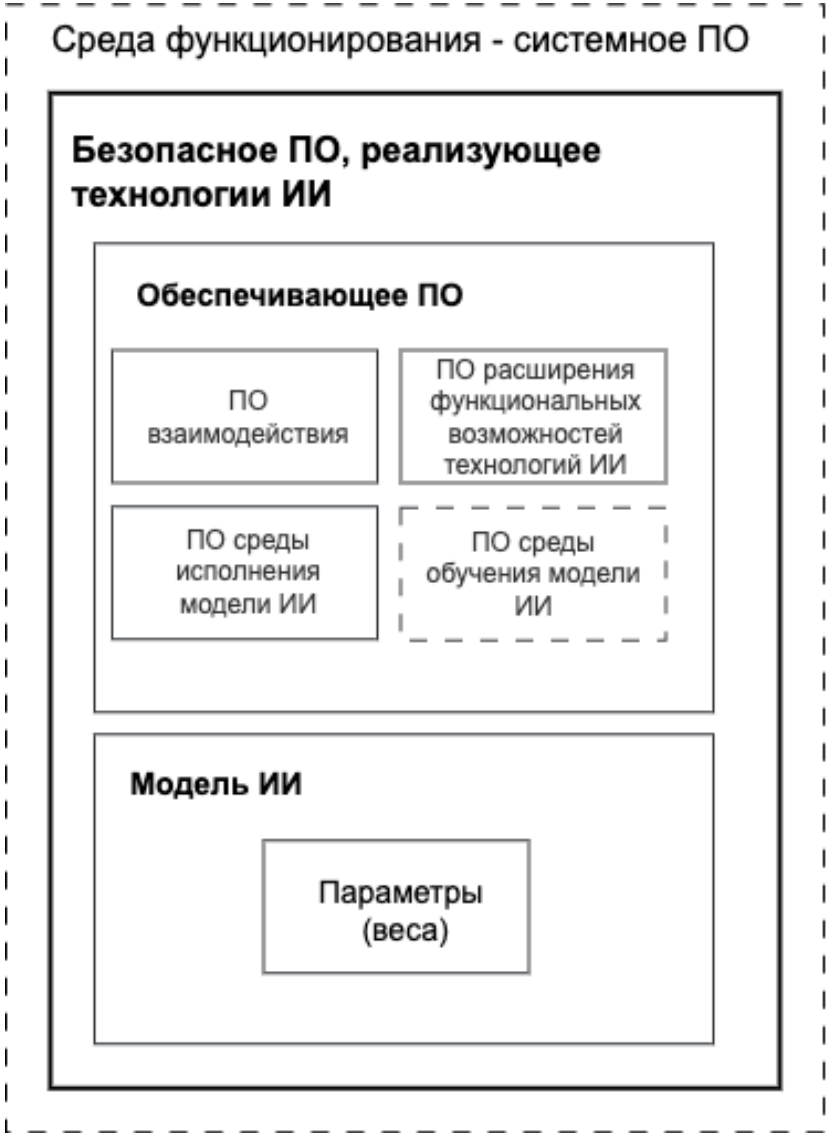
- механизмы ограничения поведения моделей ИИ (guardrails),
- платформы параллельных вычислений и т. п.

ПО, обеспечивающее расширение функциональных возможностей технологий ИИ: программы, интегрированные с моделями ИИ и с ПО среды исполнения моделей искусственного интеллекта для реализации дополнительных функций и архитектурных решений, таких как

- интеграция с СУБД и поисковыми машинами (RAG),
- вызов дополнительных программ и инструментальных средств,
- агенты ИИ, мультиагентные системы и т. п.

ПО, обеспечивающее обучение модели ИИ: инструментальные средства подготовки и управления наборами данных, обучения моделей ИИ, в т.ч.

- токенизаторы,
- фреймворки машинного обучения и т.п.



Особенности некоторых модифицированных процессов

Процесс 5.3 Формирование и предъявление требований к ПО

- необходимость формирования требований безопасности к модели ИИ как к отдельному объекту
- необходимость обеспечения непрерывного мониторинга эксплуатируемых технологий ИИ
- необходимость реализации защитных мер, направленных против известных атак на технологии ИИ
- необходимость интеграции с инструментальными средствами, используемыми для обеспечения безопасности технологий ИИ на этапе эксплуатации
- …

Процесс 5.7 Моделирование угроз и разработка описания поверхности атаки

- необходимость учета сведений об уязвимостях, поверхностях атаки и сценариев реализации угроз при использовании технологий ИИ (для ПО, реализующего технологии ИИ, и для ПО, использующего эти технологии)
- …

Процесс 5.19 Нефункциональное тестирование

- моделирование действий потенциального нарушителя (тестирование на проникновение, для ПО, реализующего технологии ИИ, и для ПО, использующего эти технологии)
- проверка качества модели ИИ

Специфичные для ИИ процессы

Процесс 5.26 Управление наборами данных

Цель:

- Обеспечение качества и безопасности данных для обучения моделей ИИ
- Предотвращение компрометации моделей ИИ путем манипуляции данными

Требования:

- Защита инфраструктуры хранения и обработки данных
- Анализ безопасности набора данных (с применением инструментальных средств)
- Кросс-аннотирование
- Разработка регламента управления наборами данных
- Требования к отчетам анализа данных
- …

Процесс 5.27 Безопасное обучение моделей ИИ

Цель:

- Обеспечение безопасного процесса обучения и устойчивости их к специфичным угрозам
- Недопущение привнесения в модели ИИ уязвимостей и ошибок со стороны среды, в которой реализуется обучение моделей ИИ

Требования:

- Защита инфраструктуры (среды) обучения
- Реализация защитных мер (и механизмов), релевантных процессу обучения
- Использование доверенных фреймворков машинного обучения
- Использование специальных инструментальных средств
- Процедуры и метрики качества обучения
- Разработка регламента безопасного обучения
- …

Инструментальные средства, применяемые на этапах разработки ПО

Инструментальные средства, применяемые для нефункционального тестирования, реализующие функции

- аудита запросов к модели ИИ с анализом потенциально опасных паттернов доступа
- анализа утечки информации
- визуализации и анализа важности признаков входных данных
- моделирования тактик, техник и процедур реализации атак (средства «Red Team»)
- тестирования устойчивости моделей ИИ
- аудита логов для оценки воздействия моделируемых атак
- обнаружения OOD-данных (Out-of-Distribution)
- мониторинга данных
- ...

Инструментальные средства, применяемые при управлении наборами данных, реализующие функции

- фильтрации потенциально небезопасных или некорректных данных
- обнаружения и маскирования конфиденциальных сведений (в том числе персональных данных)
- анализа качества данных
- анализа смещения данных
- анализа покрытия данных
- обнаружения ошибок разметки данных
- анализ согласованности разметки данных
- ...

Инструментальные средства, применяемые при обучении моделей ИИ и реализующие функции

- защиты от релевантных угроз, например для защиты от атак, характерных для логического вывода (inference) модели ИИ
- оценки влияния защитных мер друг на друга
- анализа безопасности и сериализации моделей ИИ
- дифференциальной приватности
- федеративного обучения
- контроля выходных данных LLM
- ...

А также – **Доверенные фреймворки машинного обучения**

Инструментальные средства ИСП РАН для реализации специфичных процессов

Система поддержки создания наборов данных Colba

https://www.ispras.ru/technologies/colba/

Платформа Доверенного Искусственного Интеллекта (ДИИ)

https://www.ispras.ru/technologies/tai/

Интеграционная платформа для гибридного ИИ «Talisman»

https://www.ispras.ru/technologies/talisman/

MLM: система управления задачами машинного обучения

https://mlm.at.ispras.ru/latest/en/index.html

Доверенные фреймворки машинного обучения:
TrustFlow (на базе TensorFlow)
TrustTorch (на базе PyTorch)

https://www.ispras.ru/technologies/frameworks/

Бенчмарк Socio-political Landscape And Value Analysis (SLAVA)

https://huggingface.co/datasets/RANEPA-ai/SLAVA-OpenData-2800-v1

Инструментальные средства ИСП РАН для реализации специфичных процессов

Платформа Доверенного Искусственного Интеллекта (ДИИ)

https://www.ispras.ru/technologies/tai/

MLM: система управления задачами машинного обучения
https://mlm.at.ispras.ru/latest/en/index.html

Интегрированные с платформой Доверенного Искусственного Интеллекта инструментальные средства

Обнаружение OOD-данных

Мониторинг данных

Контроль выходных данных LLM

Анализ утечки информации

Тестирование устойчивости модели ИИ

Дифференциальная приватность

Фильтрация потенциально небезопасных или некорректных данных

Защита от атак, характерных для логического вывода (inference)

Оценка влияния защитных мер (механизмов) друг для друга

Моделирования тактик, техник и процедур реализации атак (средства «Red Team»)

Аудит логов для оценки воздействия моделируемых атак

Дорожная карта разработки инструментов в Центре доверенного ИИ

	3 квартал 2025	4 квартал 2025	1 квартал 2026	2 квартал 2026	3 квартал 2026
Архитектура и инструментальные проверки	✓Общая архитектура ✓ Тест на уязвимость к **уклонениям**	✓ Тест на уязвимость к атакам на **приватность** ✓ Тест на уязвимость к **отравлениям**	✓ Добавление методов защиты от **уклонений**	Добавление методов защиты от атак на **приватность** Добавление методов защиты от **отравления**	Добавление теста на уязвимость к **out-of-distribution** примерам
Поддержка разновидностей моделей ИИ и типов данных	✓ Классификация изображений ✓ Классификация табличных данных ✓ Классификация временных рядов ✓ Регрессия на табличных данных	✓ Сегментация изображений ✓ Детекция изображений	✓ Классификация видео	Классификация текстов	Прогнозирование временных рядов

Примеры отчета Платформы ДИИ (1/2)

Отчет устойчивости resnet18 к атакам уклонения

▼ Обзор

Атакуемая модель

Задача: Классификация изображений (Image classification)

Модель: resnet18

▼ **Параметры**

__init__:

- **input_shape:** [3, 32, 32]
- **num_classes:** 10
- **weights_path:** weights.pth

compute_metrics: по умолчанию

reset_metrics: по умолчанию

update_metrics: по умолчанию

train_function:

- **lr:** 0.00100
- **momentum:** 0.90000
- **num_epochs:** 5
- **save_model_weigths:** weights.pth

get_grad: по умолчанию

predict_function: по умолчанию

Данные

Загрузчик данных: cifar10

▼ **Параметры загрузчика данных**

__init__: по умолчанию

get_dataset: по умолчанию

get_test_data: по умолчанию

get_train_data: по умолчанию

get_validation_data: по умолчанию

▼ Метрики

Эксперимент 1: cw_linf

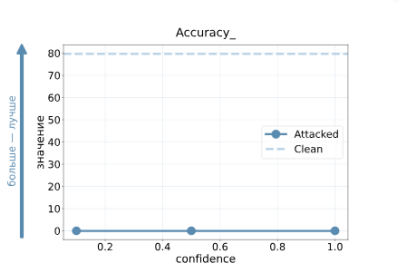
confidence	Accuracy_	Clean_Accuracy_	ASR_	Perturbation_norm_	Accuracy_user
1.00000	0.00000	79.6875	100.000	0.00722	0.00000
0.50000	0.00000	79.6875	100.000	0.00720	0.00000
0.10000	0.00000	79.6875	100.000	0.00720	0.00000

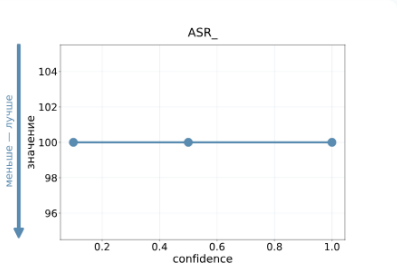
Эксперимент 2: patch

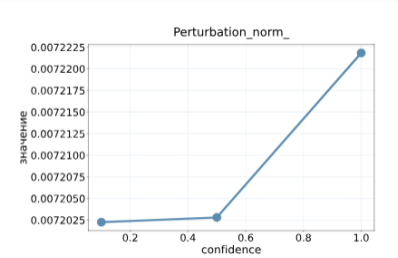
learning_rate	Accuracy_	Clean_Accuracy_	ASR_	Perturbation_norm_	Accuracy_user
1.00000	75.0000	79.6875	9.80392	0.00890	75.0000
5.00000	68.7500	79.6875	19.6078	0.01132	68.7500
10.0000	67.1875	79.6875	25.4902	0.01039	67.1875

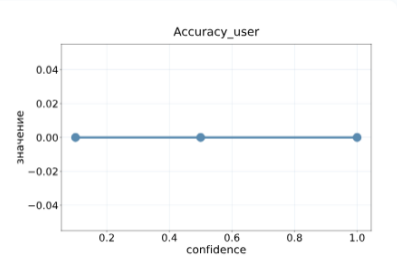
▼ Графики

Эксперимент 1: cw_linf









Примеры отчета Платформы ДИИ (2/2)

Январь 15, 2026

Содержание

2.2 Эксперимент 2: patch

learning_rate	Accuracy_	Δ Accuracy_	ASR_	Perturbation_norm_	Accuracy_user
Clean data	79.6875	-	-	-	-
1.00000	75.0000	-4.68750 (5.88%)	9.80392	0.00890	75.0000
5.00000	68.7500	-10.9375 (13.7%)	19.6078	0.01132	68.7500
10.0000	67.1875	-12.5000 (15.7%)	25.4902	0.01039	67.1875

2.3 Эксперимент 3: pgd

eps	Accuracy_	Δ Accuracy_	ASR_	Perturbation_norm_	Accuracy_user
Clean data	79.6875	-	-	-	-
0.00392	40.6250	-39.0625 (49.0%)	49.0196	0.00328	40.6250
0.01961	0.00000	-79.6875 (100.0%)	100.000	0.01296	0.00000
0.05882	0.00000	-79.6875 (100.0%)	100.000	0.01610	0.00000

Январь 15, 2026

5 Обозначения

ASR_: ASR (Attack Success Rate) — доля успешных атак среди всех предпринятых попыток. Учитываются только те попытки, на которых исходная модель правильно классифицирует исходное изображение. Диапазон: 0.0 – 1.0, меньше — лучше.

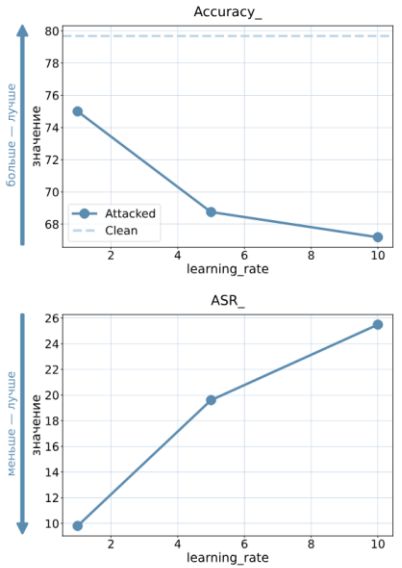
Accuracy_: Accuracy — доля правильно классифицированных изображений среди всех изображений в выборке после атаки. Диапазон: 0.0 – 1.0, больше — лучше.

Clean_Accuracy_: Clean Accuracy — точность модели на исходных (не атакованных) изображениях. Диапазон: 0.0 – 1.0, больше — лучше.

Perturbation_norm_: Perturbation norm — средняя норма разницы между чистыми и атакованными изображениями (используется l1 норма, деленная на n, где n = число писелей × число каналов). Диапазон: 0.0 – +∞.

Январь 15, 2026

3.2 Эксперимент 2: patch



4.2 Эксперимент 2: patch

- Метрика **Accuracy_** (диапазон: 0.0 – 1.0, больше — лучше) уменьшилась с 79.69 до 67.19 (уменьшение на 15.7%)

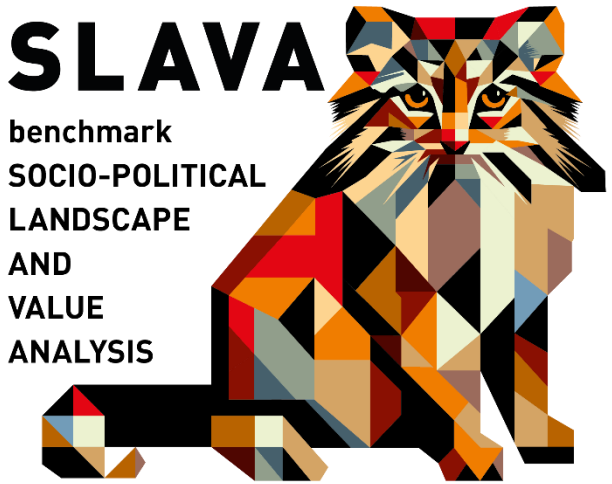
4.3 Эксперимент 3: pgd

- Метрика **Accuracy_** (диапазон: 0.0 – 1.0, больше — лучше) уменьшилась с 79.69 до 0.000 (уменьшение на 100.0%)

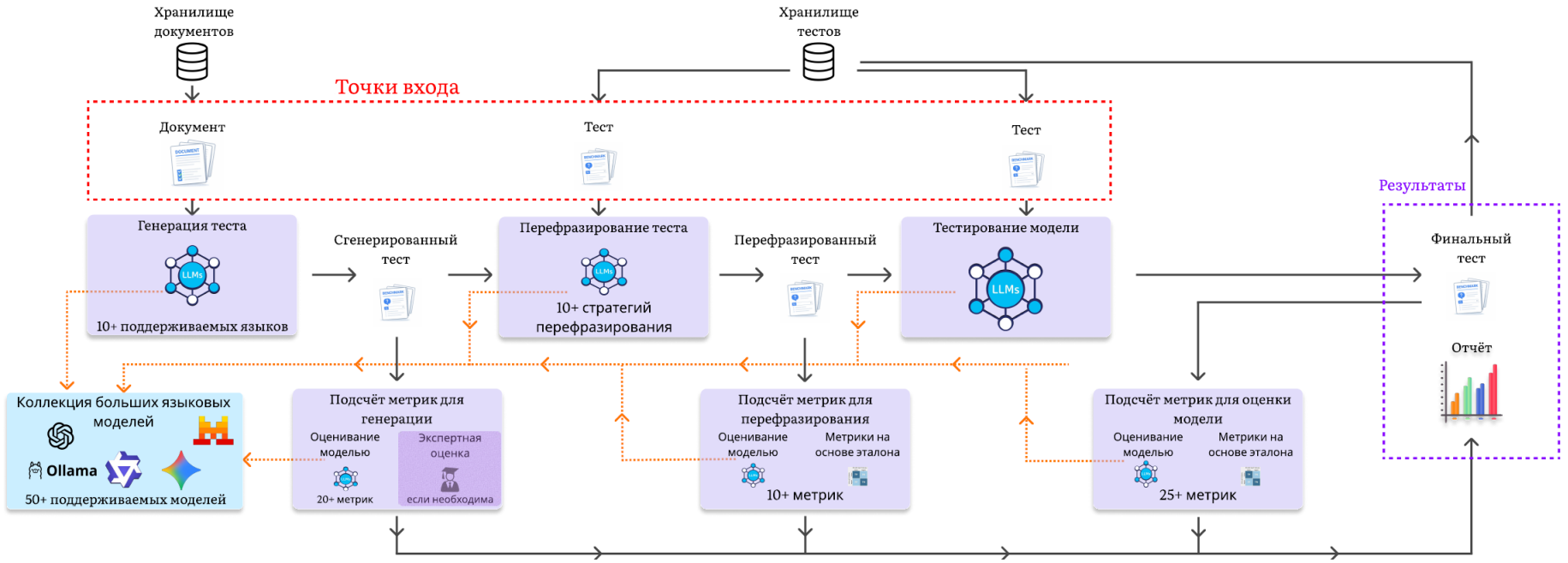
Оценка больших языковых моделей на эталонных тестах

Комплексная оценка работы LLM с чувствительными гуманитарными темами (история, обществознание, политология, география) в российском и постсоветском контексте

- охватывает четыре ключевых направления — историю, обществознание, политологию и географию
- три уровня чувствительности (провокативности) вопросов;
- включает около 14 тысяч заданий и методологию оценки ответов языковых моделей и ранжирования вопросов по степени «провокативности», то есть чувствительности респондента к теме
- использование официальных источников, близкие к позиции РФ (базы ЕГЭ по гуманитарным дисциплинам, открытые экзаменационные вопросы ведущих российских вузов, а также вопросы, сформулированные специалистами РАНХиГС и ИСП РАН),
- валидация междисциплинарной группой экспертов



Система тестирования больших языковых моделей GRACE-LLM



- Создана совместно специалистами РАНХиГС и ИСП РАН.
- Модуль генерации тестов GenA, расширенный при интеграции на 10 языков
- Модуль перефразирования тестов TrustVar при сохранении семантики 10+ стратегиями
- Интеграция в единый конвейер с поддержкой тестирования более 50 моделей, более 30 различных метрик, включая метрики на основе эталона и на основании оценки других БЯМ
- Задействованы различные тесты, включая тест СЛАВА

Заключение

- Разработан проект стандарта, задающего требования РБПО применительно к ИИ
  - Предиктивный и генеративный ИИ, расширение требований ГОСТ Р 56939-2024
  - Проект стандарта уже прошел апробацию в организации-разработчике
- В (оптимистичных) планах довести стандарт в этом году до принятия
- Требования должны быть поддержаны доступными инструментами (технологиями), которые обеспечат проведение проверок
  - Обучающих наборов данных
  - Моделей ИИ
  - ПО, реализующего технологии ИИ

ИСП РАН

Спасибо за внимание
Вопросы

Разработка безопасного ПО, реализующего технологии ИИ

ПАДАРЯН ВАРТАН
vartan@ispras.ru

СОСНИНА Е.С.
esosnina@ispras.ru

ТУРДАКОВ Д.Ю.
turdakov@ispras.ru

Форум «Технологии доверенного искусственного интеллекта»

13 мая 2026 года