О СТАНОВЛЕНИИ ОТЕЧЕСТВЕННОЙ СИСТЕМЫ СЕРТИФИКАЦИИ ТЕХНОЛОГИЙ ИИ

ВОПРОСЫ ОТ РЯДОВОГО ПОЛЬЗОВАТЕЛЯ

ЛОСЬ ВЛАДИМИР ПАВЛОВИЧ, Д.ВОЕН.Н., ПРОФЕССОР, ГЛАВНЫЙ НАУЧНЫЙ СОТРУДНИК РГГУ, АКАДЕМИК АВН РФ

СЕРТИФИКАЦИЯ - ФОРМА ОСУЩЕСТВЛЯЕМОГО ОРГАНОМ ПО СЕРТИФИКАЦИИ (?) ПОДТВЕРЖДЕНИЯ СООТВЕТСТВИЯ ОБЪЕКТОВ ТРЕБОВАНИЯМ ТЕХНИЧЕСКИХ РЕГЛАМЕНТОВ, ДОКУМЕНТАМ ПО СТАНДАРТИЗАЦИИ ИЛИ УСЛОВИЯМ ДОГОВОРОВ;
(В РЕД. ФЕДЕРАЛЬНЫХ ЗАКОНОВ ОТ 01.05.2007 N 65-ФЗ, ОТ 05.04.2016 N 104-ФЗ)

ОСНОВНЫЕ ЭТАПЫ СЕРТИФИКАЦИИ ПРОДУКЦИИ

Определить тип сертификации

Провести испытания продукции

Выбрать аккредитованный орган по сертификации

Получить сертификат

Подготовить пакет документов

(включая протоколы испытаний)

Осуществить маркировку продукции

Определить тип сертификации

?

Провести испытания продукции

Выбрать аккредитованный орган по сертификации

?

Получить сертификат

?

Подготовить пакет документов

(включая протоколы испытаний)

?

Осуществить маркировку продукции

?

Ошибки

Основные типы ошибок БЯМ

- 1. Галлюцинации.
- 2. Ошибочный ответ из-за дефекта запроса.
- 3. Предвзятость (Bias).
- 4. Недостаточная глубина анализа проблемы, генерация бессвязного текста.
- 5. Чрезмерная зависимость от шаблонов.

Приказ 117

Приказ ФСТЭК России от 11 апреля 2025 г. № 117

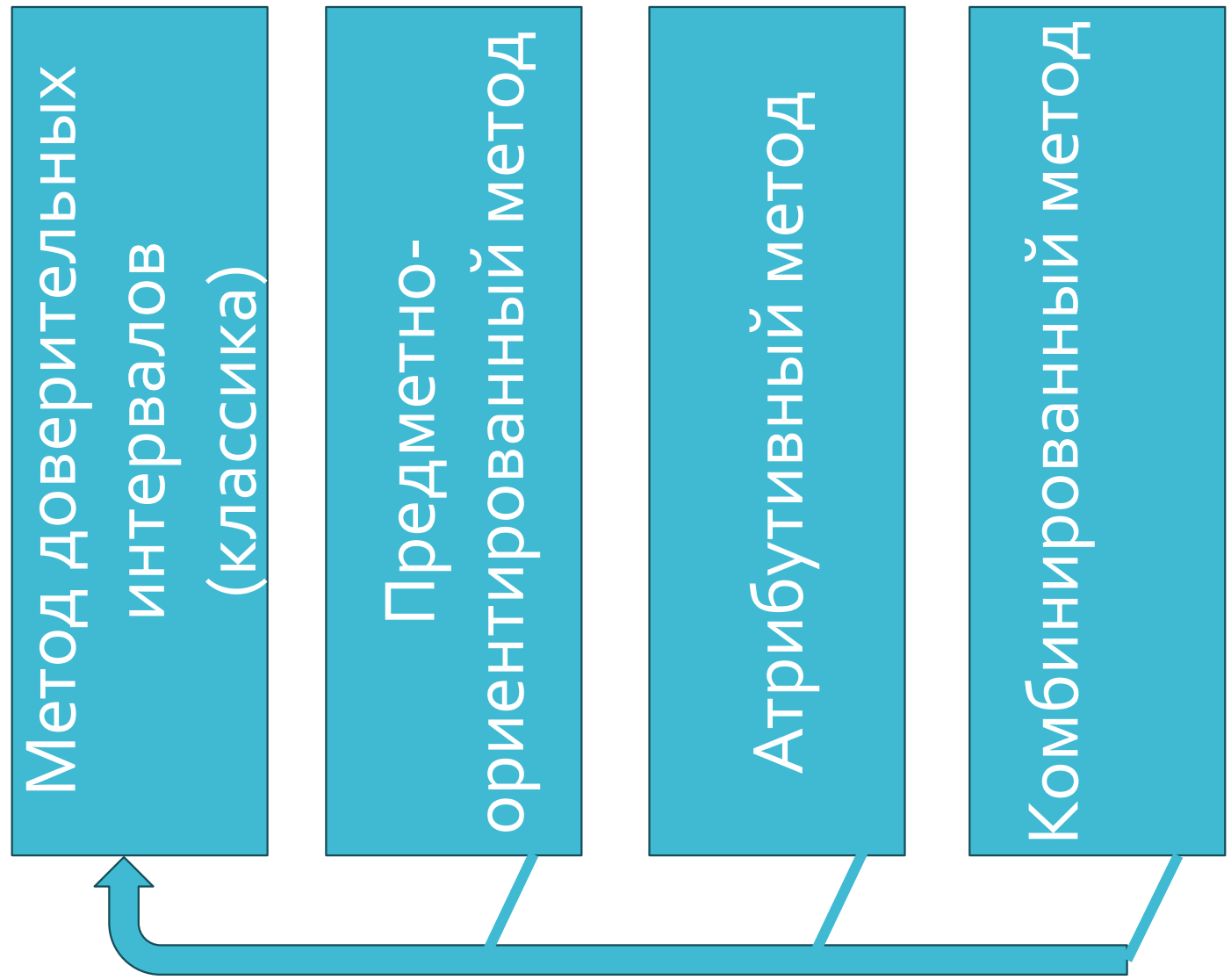
- 1. Обеспечение защиты информации при использовании искусственного интеллекта относится к обязательным мероприятиям.
- 2. В ходе проведения мониторинга информационной безопасности допускается использование доверенных технологий искусственного интеллекта.

Но!!!

Реестра доверенных технологий искусственного интеллекта НЕТ.

Основное допущение

4 метода


Метод доверительных интервалов (классика)
Предметно-ориентированный метод
Атрибутивный метод
Комбинированный метод

Инструментальные средства оценки качества БЯМ

- Под инструментальными средствами оценки качества БЯМ будем понимать совокупность набора эталонных вопросов и ответов, математических моделей, позволяющих получать оценки качества БЯМ, включая процедуры сравнения ответов БЯМ и эталонных ответов.

Метод доверительных интервалов (классика)

M – количество вопросов и ответов (объем выборки).

$K_{\text{пр}}$ - число правильных ответов.

$\hat{p} = \frac{K_{\text{пр}}}{M}$ - доля правильных ответов.

$\hat{q}=1-\hat{p}$ – доля неправильных ответов.

Асимптотический доверительный интервал для доли имеет вид:

$$(\hat{p} - \Delta\,;\ \hat{p} + \Delta), \qquad (1)$$

где Δ – половина ширины доверительного интервала (точность):

$$\Delta = \sqrt{\frac{\hat{p}\hat{q}}{M}} \cdot Z_{\alpha}, \qquad (2)$$

где Z_{α} – квантиль нормального распределения уровня 1-α/2, α–уровень значимости.

Уровень доверия β вычисляется по формуле

$$\beta = 1 - \alpha \qquad (3)$$

Использование формулы (1) предполагает выполнение условий $M \cdot \hat{p} \geq 5$ и $M \cdot \hat{q} \geq 5$. Требованиями к качеству могут устанавливаться границы и ширина доверительного интервала.

Метод доверительных интервалов (классика)

ПРИМЕР

Пример расчета.

Исходные данные. Объем выборки $M = 100$, из них правильных ответов 90, то есть выборочная доля $\hat{p} = 0{,}9$, уровень значимости $\alpha = 0{,}05$, уровень доверия $\beta = 0{,}95$.

Решение. Выборочная доля $\hat{p} = 0.9$, поэтому $\hat{q} = 1 - \hat{p} = 0{,}1$. Убеждаемся, что выполнены условия применимости этих формул: $M \cdot \hat{p} = 90 > 5$ и $M \cdot \hat{q} = 10 > 5$. Находим уровень $1 - \alpha/2 = 0{,}975$ и по таблице нормального распределения определяем квантиль $Z_\alpha = 1{,}96$. Теперь можем найти точность

$$\Delta = \sqrt{\frac{\hat{p} \cdot \hat{q}}{M}} \cdot Z_\alpha = 0{,}059.$$

Искомый 95%-доверительный интервал имеет вид $(0{,}9 - 0{,}059;\ 0{,}9 + 0{,}059) = (0{,}841;\ 0{,}959)$.

Таким образом, **доля правильных ответов для исследуемой БЯМ, с вероятностью 0,95 находится в интервале (0,841; 0,959).**

Предметно-ориентированный метод

Допущение. В базе знаний S модели имеются N подмножеств $S_1, S_2, \ldots S_N$, соответствующих предметным областям (информатика, информационная безопасность, экономика и так далее). В дальнейшем будем называть эти подмножества предметными:

$$S_i \subset S, i = 1,2, \ldots N. \quad (4)$$

Предметные подмножества могут иметь пересечения, то есть

$$\left| S_i \cap S_j \right| = L_{ij} \geq 0, i < j \leq N, \quad (5)$$

где L_{ij} – количество пересечений предметных подмножеств S_i и S_j.

Предметно-ориентированный метод

Продолжение 1

Вопросам к БЯМ

$$Z = \{Z_1, Z_2, \ldots, Z_M\}, \quad (6)$$

где M – общее количество возможных вопросов, соответствуют ответы:

$$O = \{O_1, O_2, \ldots, O_M\}, \quad (7)$$

причем имеют место следующие соотношения:

$$\begin{cases} Z_1 \to O_1; \\ Z_2 \to O_2; \\ \ldots\ldots\ldots\ldots \\ Z_M \to O_M. \end{cases} \quad (8)$$

Предметно-ориентированный метод

Продолжение 2

1. Первоначально формируется верифицированный (эталонный) набор вопросов и ответов:

$$\begin{cases} Z_1^B \rightarrow O_1^B ; \\ Z_2^B \rightarrow O_2^B ; \\ \dots\dots\dots\dots\dots \\ Z_M^B \rightarrow O_M^B , \end{cases} \quad (9)$$

где $Z_m^в$ - m-й вопрос из верифицированного набора вопросов и ответов, $O_m^в$ - m-й ответ на m-й вопрос из верифицированного набора вопросов и ответов, m = 1, 2, …,M.

2. Вопросы $Z_1^в, Z_2^в, \dots, Z_M^в$ из верифицированного набора вопросов и ответов предъявляются БЯМ, фиксируются соответствующие ответы $O_1^{ии}, O_1^{ии}, \dots, O_M^{ии}$ и устанавливается соответствие между предъявленными вопросами и ответами:

$$\begin{cases} Z_1^B \rightarrow O_1^{ИИ} ; \\ Z_2^B \rightarrow O_2^{ИИ} ; \\ \dots\dots\dots\dots\dots \\ Z_M^B \rightarrow O_M^{ИИ} , \end{cases} \quad (10)$$

Предметно-ориентированный метод

Продолжение 3

3. Вводится операция сравнения $R(O_m^{ии}, O_m^{в})$ ответов модели искусственного интеллекта и ответов из верифицированного набора вопросов и ответов, m = 1, 2, …,M,

$$R(O_m^{ИИ}, O_m^{B}) = \begin{cases} 1, \text{если ответы } O_m^{ии} \text{ и } O_m^{в} \text{ совпадают по смыслу;} \\ 0, \text{если ответы } O_m^{ии} \text{ и } O_m^{в} \text{ не совпадают по смыслу.} \end{cases} \quad (11)$$

4. Рассчитывается количество правильных ответов:

$$К_{пр} = \sum_{m=1}^{M} R(O_m^{ии}, O_m^{в}). \quad (12)$$

5. Рассчитывается количество ложных (ошибочных) ответов

$$К_{ош} = M - \sum_{m=1}^{M} R(O_m^{ии}, O_m^{в}). \quad (13)$$

6. Относительное количество ошибок модели искусственного интеллекта определяется по формуле:

$$\gamma_{ош} = К_{ош}/М. \quad (14)$$

Требованиями по безопасности может устанавливаться предельное значение относительного количества ошибок $\gamma_{ош}^{пр}$.Полученные данные могут использоваться в качестве исходных для применения классического метода доверительных интервалов.

Проблемы использования

Проблемы:

1. Доказательство однородности статистики.
2. Локализация БЯМ.
3. Автоматизация процедуры сравнения по смыслу ответов БЯМ и эталонных ответов.
4. Поддержание актуальности верифицированного набора вопросов и ответов.

Атрибутивный метод

Подход к оценке качества не на основе сравнения ответов с эталонами, которых может и не быть в распоряжении пользователя, а на основе использования дополнительной информации (атрибутов), которая содержится в ответах (генерациях) модели.

Атрибутивный метод

Продолжение 1

Результаты оценки 10000 генераций по двум атрибутам (случай независимости)

	X=0	X=1	Всего
У=0	9000	500	9500
У=1	450	50	500
Всего	9450	550	10000

Согласно данным табл., относительная доля дефектности для атрибута X равна 0,055, для атрибута Y 0,050, в то время как для выборки в целом относительная доля дефектности составляет 0,1, поскольку из общего количества в 10000 генераций только 9000 оказались без дефектов.

Атрибутивный метод

Основная проблема:

доказательство гипотезы о независимости атрибутов

ВЫВОДЫ

Выводы:

1. Сертификация технологий ИИ неизбежна. Но непонятно на какой основе ее реализовывать: добровольной или добровольно-принудительной.
2. Имеющиеся разработки по оценке качества технологий ИИ носят, в основном, декларативный характер и ориентированы на проблемы внутреннего построения ИИ.
3. Нужны разработки по процедурам оценки выходных характеристик моделей ИИ и определения временных

СПАСИБО

СПАСИБО
ЗА
ВНИМАНИЕ