ИНСТРУМЕНТАЛЬНЫЕ СРЕДСТВА СЕРТИФИКАЦИИ ТЕКСТОВЫХ МОДЕЛЕЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ПО ТРЕБОВАНИЯМ БЕЗОПАСНОСТИ

Лось В.П., доктор военных наук, профессор, главный научный сотрудник, Российский государственный гуманитарный университет,

Тышук Е.Д., старший научный сотрудник, Российский государственный гуманитарный университет,

Шевцова Г.А., кандидат исторических наук, доцент, директор Института информационных наук и технологий безопасности, Российский государственный гуманитарный университет.

Формальная постановка задачи

Имеется некоторая текстовая модель ИИ, внутреннее построение которой неизвестно («черный ящик»). Оценка безопасности модели может быть основана исключительно на анализе правильности или неправильности ответов, которые выдает модель на те или иные вопросы.

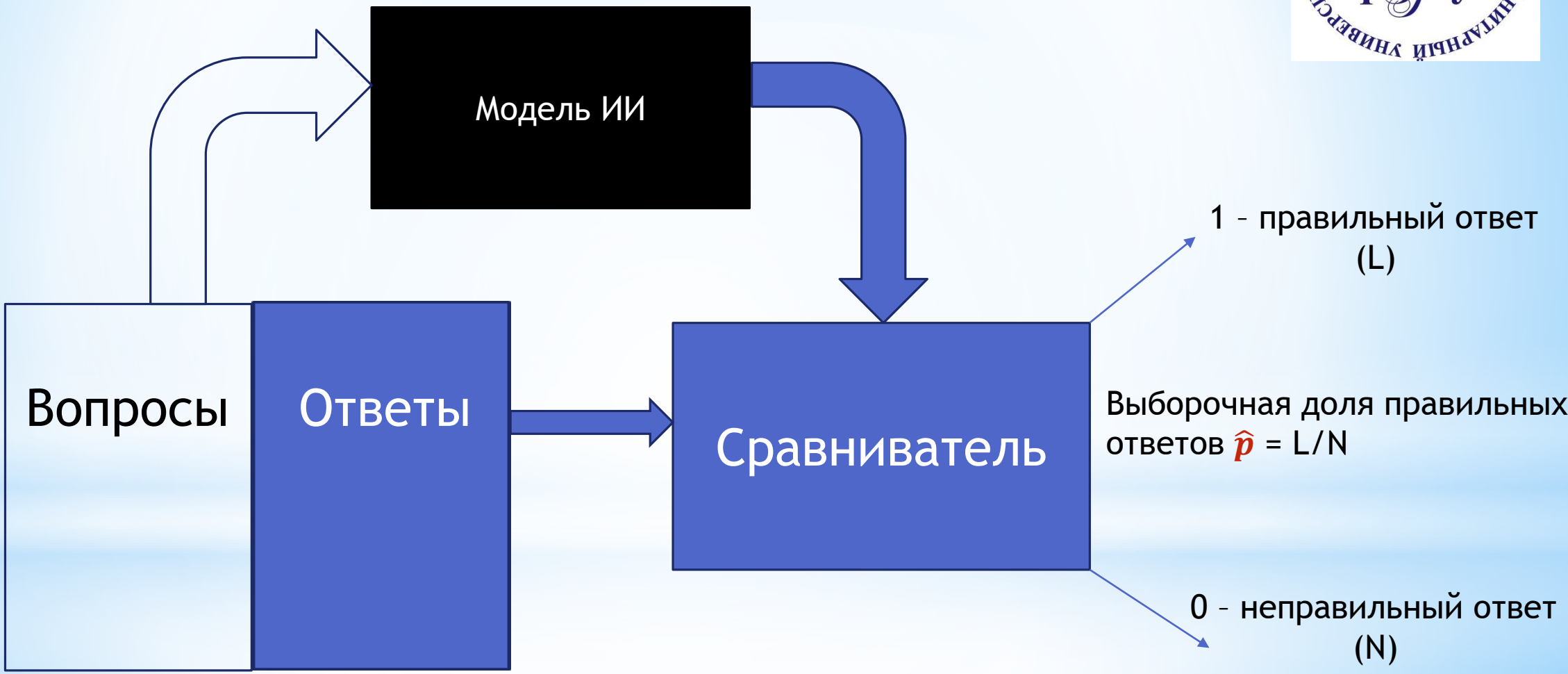
Под сертификацией модели ИИ понимается процедура получения оценок ее безопасности.

Под инструментальными средствами сертификации понимается совокупность набора вопросов и ответов и математических моделей, позволяющих получать оценки безопасности.

Три метода решения задачи:

метод доверительных интервалов;
предметно-ориентированный метод;
комбинированный метод.

1. Метод доверительных интервалов



Верифицированная база вопросов и ответов

Результаты использования метода доверительных интервалов

Объем выборки M	Выборочная доля правильных ответов $\hat{p}$	Уровень значимости α	Уровень доверия β	Доверительный интервал
100	0,90	0,05	0,95	(0,841; 0,959)
100	0,85	0,05	0,95	(0,780; 0,920)
100	0,80	0,05	0,95	(0,722; 0,880)
100	0,90	0,06	0,94	(0,843; 0,957)
100	0,90	0,07	0,93	(0,846; 0,954)
1000	0,90	0,05	0,95	(0,881; 0,919)
1000	0,85	0,05	0,95	(0,828; 0,872)
1000	0,80	0,05	0,95	(0,775; 0,825)

2. Предметно-ориентированный метод.

2.1.Первоначально формируется верифицированный набор вопросов и ответов.

2.2. Вопросы из верифицированного набора вопросов и ответов предъявляются модели искусственного интеллекта, фиксируются соответствующие ответы, устанавливается соответствие между предъявленными вопросами и ответами.

2.3. Вводится операция сравнения ответов модели искусственного интеллекта и ответов из верифицированного набора вопросов и ответов.

2.4. Рассчитывается количество правильных ответов.

2.5. Рассчитывается количество ложных (ошибочных) ответов.

2.6. Определяется относительное количество ошибок модели искусственного интеллекта

3. Комбинированный метод
Этот метод заключается в использовании результатов предметно-ориентированного метода для метода доверительных интервалов

СПАСИБО ЗА ВНИМАНИЕ