УДК 004.8

Выбор уровня автоматизации контроля ИИ-агентов с учётом оценки стоимости проверки результата

Афанасьев И. С.[1], Афанасьева А. Д.[2], Афанасьев Б. С.[2]

[1] Балтийский федеральный университет имени Иммануила Канта

Калининград

[2] Санкт-Петербургский политехнический университет Петра Великого

Санкт-Петербург

Научный руководитель — Ильяшенко В. М., к.э.н., доцент

Экономический эффект ИИ-агента зависит от затрат на получение приемлемого результата. Цель работы – предложить рекомендации по выбору режима контроля с помощью оценки стоимости верификации на основе анализа качества результата и последствий ошибок.

Стоимость верификации зависит от качества результата, квалификации проверяющего и условий проверки [1]. При этом быстро принятое ошибочное решение и медленно принятое правильное решение могут иметь различную целесообрвзность в зависимости от соотношения потерь от ошибки и затрат на принятие решения.

В разработке компилятор и тесты дают быструю воспроизводимую обратную связь, которая позволяет ИИ-агенту проверять свою работу автоматически, в том числе в цикле. Однако подготовка тестов, среды и инструкций по проверке требует ресурсов. Кроме того, успешное прохождение тестов не подтверждает выполнение всех требований и отсутствие ошибок [2].

Аналогично, автоматически проверяемые задачи существуют и в других сферах: так, в документообороте можно выявлять пропуски полей, нарушения

форматов и заданные противоречия [3]. Однако эти проверки не заменяют содержательную экспертизу договора.

Человеческий контроль все еще остается одним из популярных методов проверки работы ИИ-агента. Так, в исследовании внедренных агентов его использовали 23 из 31 респондента [4]. При этом стоимость ручных, автоматизированных и полностью автоматических проверок зависит от ряда факторов, поэтому является актуальным поиск объективного критерия для её оценки.

Для ручных, автоматизированных и автоматических проверок предлагается сравнивать ожидаемую стоимость принятого результата:

$$C = C_e + C_v + C_f + E(L) + F/N$$

где C_e – затраты на выполнение, C_v – затраты на проверку, C_f – затраты на исправление; $E(L)$ – ожидаемые потери от оставшихся ошибок, F – затраты на внедрение за выбранный период, N – число принятых результатов за этот период. При сравнении задач одинакового типа и уровня сложности для снижения затрат на проверку следует выбирать степень автоматизации, которая обеспечивает минимальную стоимость при выполнении требований к качеству. Необходимо ограничивать недоступные действия и недопустимые последствия, чтобы они были обязательно учтены, независимо от их влияния на затраты [5].

Например, при обработке накладной можно выполнять автоматически проверку наличия реквизитов, форматов и арифметических связей. При этом для содержательной проверки требуется решение специалиста. Объяснение агента помогает эксперту в принятии решения, но ответственность за решение остается на человеке [6].

Для каждого вида ошибки необходимо определить, как их можно эффективно обнаруживать, и в какой момент нужно проводить проверку, чтобы иметь возможность предотвратить или снизить негативные последствия. При этом следует оценивать внедренную систему верификации в течение достаточного времени, чтобы потенциальные ошибки успели

проявиться, так как отсутствие ошибок в краткосрочной перспективе не гарантирует их отсутствие в долгосрочной.

Таким образом, ожидаемая стоимость принятого результата складывается из затрат на выполнение, проверку и исправление, ожидаемых потерь от необнаруженных ошибок и удельных затрат на внедрение. В первую очередь необходимо составить требования к приемлемому результату и ограничить недоступные действия. и недопустимые последствия. Процесс рекомендуется разделить на этапы, результат каждого из которых можно проверить, и оценить, какие шаги целесообразно проверять автоматически.

Список использованной литературы

1. AI Evaluation Should Measure Verification Cost, Not Correctness Alone / Crescitelli V., Immediato G., Persia F., Costantini S. — 2026. — Preprint. arXiv : 2608.08709.

2. Human-In-The-Loop Software Development Agents: Challenges and Future Directions / Pasuksmit J., Takerngsaksiri W., Thongtanunam P. et al. — 2025. — Preprint. arXiv : 2506.11009v1.

3. Костин М. А., Давыдова Д. М., Петров В. Е. Реализация RAG-архитектуры для автоматизированной проверки документов в корпоративных системах хранения с использованием больших языковых моделей // Интеллектуальные технологии на транспорте. — 2026. — № 2 (46). — С. 5–16. — DOI: 10.20295/2413-2527-2026-246-5-16.

4. Measuring Agents in Production / Pan M. Z., Arabzadeh N., Cogo R. et al. // Proceedings of the International Conference on Machine Learning. — PMLR. — 2026. — Vol. 306. — Access mode: https://proceedings.mlr.press/v306/pan26a.html (online; accessed: 2026-09-30).

5. Towards a Science of AI Agent Reliability / Rabanser S., Kapoor S., Kirgis P. et al. // Proceedings of the International Conference on Machine Learning. — PMLR. — 2026. — Vol. 306. — arXiv : 2602.16666v3.

6. Designing meaningful human oversight in AI / Zhu L., Lu Q., Ding M. et al. // AI and Ethics. — 2026. — Vol. 6. — Article 286. DOI: 10.1007/s43681-026-01147-