СРМП №2

РАЗРАБОТКА AI-АГЕНТОВ И АНАЛИЗ ВОПРОСОВ БЕЗОПАСНОСТИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Срок выполнения по календарно-тематическому плану: 13-я неделя.

Цель работы

Под руководством преподавателя разработать архитектуру AI-агента или агентного workflow для индивидуального проекта и провести анализ рисков безопасности, ограничений автономности и мер контроля.

Связь с дисциплиной и индивидуальным семестровым проектом

Работа опирается на СРМ №2 и ПЗ №11–13. Её результаты непосредственно используются в ПЗ №14–15 при разработке AI-агента и финальном анализе этики и безопасности проекта.

Краткие теоретические сведения

AI-агент объединяет модель принятия решений, состояние, инструменты и действия. В отличие от обычного LLM-интерфейса, ошибка агента может привести к реальному действию, поэтому проектирование должно включать минимальные привилегии, разделение доверенных инструкций и внешних данных, журналирование, условия остановки и Human-in-the-loop.

Методические указания

СРМП выполняется с обязательным обсуждением промежуточных решений с преподавателем. До консультации магистрант готовит черновую архитектуру и эксперимент; на консультации аргументирует выбор методов, получает замечания и фиксирует принятые решения. После консультации работа дорабатывается. В отчёте обязательно показать, какие замечания преподавателя были учтены и как они изменили проект.

При использовании AI-сервисов запрещается передавать конфиденциальные, персональные или служебные данные без разрешения. Код и вычислительные эксперименты должны быть воспроизводимыми; для стохастических методов фиксируется seed/random_state, а данные, метрики и параметры описываются в отчёте.

Порядок выполнения

1. Уточнить исследовательский вопрос текущей работы.
2. Определить исходные данные и критерии оценки.
3. Изучить и сравнить подходящие методы.
4. Выполнить практический эксперимент или разработать прототип.
5. Проанализировать ошибки и ограничения.
6. Связать результат с индивидуальным семестровым проектом.
7. Оформить выводы и подготовиться к защите.

Задания

Задание 1. Представьте преподавателю текущую архитектуру индивидуального проекта и результаты СРМ №2.

Задание 2. Опишите агентный сценарий минимум из 4 последовательных шагов. Если агент проекту не нужен, докажите это и разработайте отдельный связанный мини-кейс.

Задание 3. Определите инструменты агента и права доступа; для каждого инструмента укажите разрешённые и запрещённые действия.

Задание 4. Добавьте условия остановки, обработку ошибок и минимум одну точку Human-in-the-loop.

Задание 5. Составьте модель угроз: минимум 6 рисков, включая Prompt Injection, утечку данных, ошибочное действие, некорректный результат и злоупотребление правами.

Задание 6. Для каждого риска предложите меру снижения и способ проверки её эффективности.

Задание 7. Продемонстрируйте нормальный и ошибочный сценарий и зафиксируйте журнал действий.

Задание 8. По итогам консультации сформируйте финальные требования безопасности для ПЗ №14–15 и итогового прототипа.

Пример выполнения

Агент анализа подозрительного URL может: принять URL → вычислить локальные лексические признаки → получить прогноз классификатора → сформировать объяснение → при низкой уверенности передать случай аналитику. Агент не должен автоматически открывать URL, скачивать содержимое или выполнять код. Для внешнего текста применяется правило: содержимое рассматривается как данные, а не как доверенная инструкция.

Индивидуальное задание

Подготовьте «паспорт безопасности» собственного интеллектуального решения: активы, данные, инструменты, права, угрозы, меры контроля, Human-in-the-loop, условия остановки и остаточные риски.

Рекомендуемая форма представления результатов

Раздел	**Что должно быть представлено**
Постановка	Цель, исследовательский вопрос, связь с индивидуальным проектом
Данные/материалы	Источник, структура, ограничения и допустимость использования
Методика	Алгоритмы, параметры, критерии и схема эксперимента
Результаты	Таблицы, метрики, схемы, примеры или журналы работы
Анализ	Сравнение, ошибки, ограничения, риски

Решение по проекту	Что включается в итоговую систему и почему

Требования к отчёту

7–10 страниц + схема агента/workflow + таблица прав инструментов + модель угроз + журнал двух сценариев + таблица «замечание преподавателя — изменение архитектуры».

Критерии оценивания

Работа оценивается в соответствии с критериями силлабуса: аккуратность и точность — 30 %, творчество и креативность — 30 %, полнота и зрелость — 40 %. Оригинальность учитывается коэффициентом 1,0; 0,5 или 0. Работа должна быть самостоятельной, а использование готовых решений без собственного анализа снижает коэффициент оригинальности.

Контрольные вопросы

1. Какой исследовательский вопрос решается в данной работе?

2. Почему выбранные методы соответствуют поставленной задаче?

3. Какие альтернативные подходы были рассмотрены?

4. Какие данные использованы и какие у них ограничения?

5. По каким критериям сравнивались варианты решения?

6. Какие ошибки или нестабильные случаи выявлены?

7. Какие угрозы достоверности эксперимента необходимо учитывать?

8. Какие вопросы конфиденциальности и безопасности возникают?

9. Как результаты работы изменяют индивидуальный семестровый проект?

10. Какой следующий этап исследования необходимо выполнить?

Список источников

1. Бессмертный И. А., Авдюшина А. Е., Кугаевских А. В., Королева Ю. А., Рущенко Н. Г. Системы искусственного интеллекта : методическое пособие. – Санкт-Петербург : Университет ИТМО, 2024. – 281 с.

2. Соснило А. И. Атлас искусственного интеллекта для бизнеса и власти : учебное пособие. – Санкт-Петербург : Университет ИТМО, 2022. – 98 с.

3. Сидоркина И. Г. Системы искусственного интеллекта : учебное пособие. – Москва : КНОРУС, 2018. – 248 с.

4. Баланов А. Н. Искусственный интеллект. Понимание, применение и перспективы : учебник для вузов. – Санкт-Петербург : Лань, 2024. – 312 с.

5. Калинин М. О., Крундышев В. М. Основы искусственного интеллекта. Безопасность искусственного интеллекта : учебное пособие. – Санкт-Петербург, 2024. – 220 с.

6. Рассел С., Норвиг П. Искусственный интеллект: современный подход. – 2-е изд. – Москва : Вильямс, 2018. – 230 с.

7. Андреянов Н. В., Евдокимова Т. С., Павлов А. Д. [и др.]. Методы искусственного интеллекта : учебное пособие. – Казань : КНИТУ-КАИ, 2024. – 392 с.