

ИИ для специалистов по Data Science

Динцис Данил Юрьевич



- ✓ д.т.н. “Системный анализ информационных систем”
- ✓ PfMP, PgMP, PMP,
- ✓ ITIL® Expert, ITIL 4.0 managing Professional, Strategic Leader
- ✓ ScrumStudy Agile Master, Product Owner, Certified trainer
- ✓ DASA DevOps Product Owner, Accredited trainer

dinzis@specialist.ru
consult@Dintsis.org
<https://t.me/pmtips>





Основные понятия и термины

**Искусственный
интеллект (ИИ; англ.
Artificial intelligence, AI)**

- (1) наука и технология создания интеллектуальных машин, особенно интеллектуальных компьютерных программ;
- (2) свойство интеллектуальных систем выполнять творческие функции, которые традиционно считаются прерогативой человека





Уровень приложений



На какой стадии применения ИИ ваша компания?



Уровень 0 Мы не используем ИИ и не планируем начинать

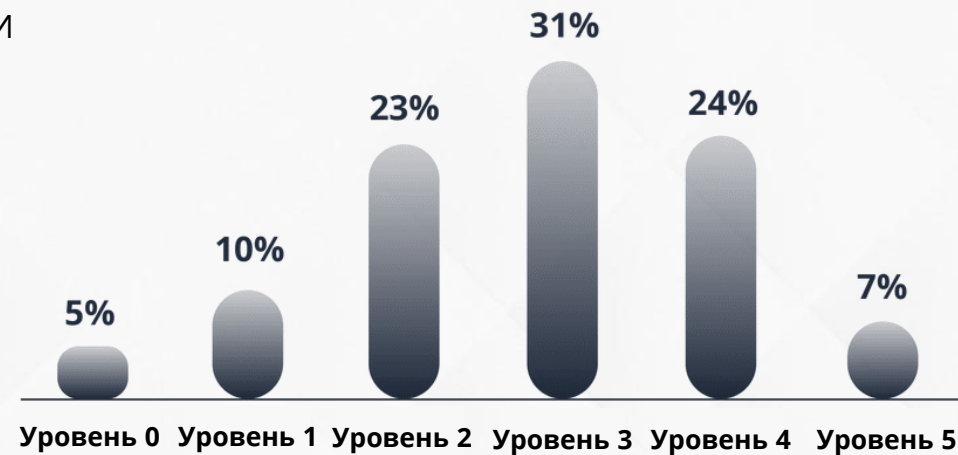
Уровень 1 Мы не используем ИИ, но хотели бы или планируем начать

Уровень 2 Мы используем бесплатные потребительские инструменты ИИ, но не используем возможности ИИ многих имеющихся у нас корпоративных программ

Уровень 3 Мы используем некоторые корпоративные программы с ИИ и не рекомендуем использовать бесплатные потребительские платформы ИИ

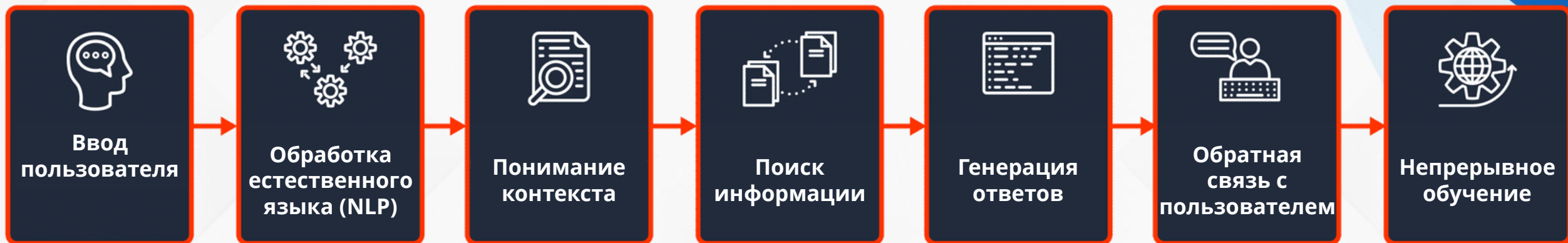
Уровень 4 Мы регулярно используем ПО со встроенными возможностями ИИ и поощряем использование потребительских ИИ при необходимости

Уровень 5 Мы начали использовать ИИ/ML до того, как ChatGPT сделал его популярным



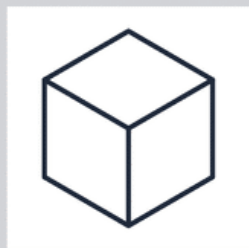


Как работает копилот (ИИ ассистент)

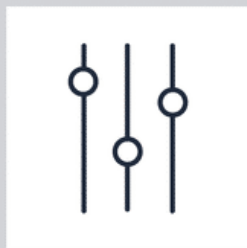
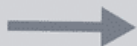




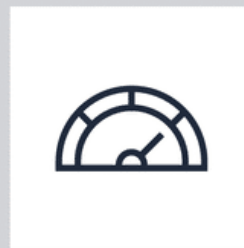
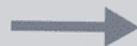
**Реальные
данные**



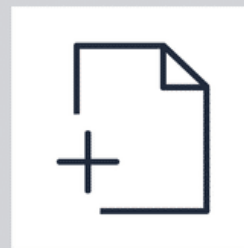
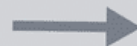
**Обучение
модели**



Настройка



**Отчет и
уточнение**



Генерация



**Синтетические
данные**

	Трудности с данными реального мира	Преимущества синтетических данных
Приватность	Данные реального мира, такие как данные пациентов или клиентов, могут быть строго конфиденциальными. Обмен этими данными с другими командами или их использование для обучения моделей ИИ или разработки приложений может нарушать стандарты конфиденциальности или представлять риск раскрытия персонально идентифицируемой информации (PII) и защищенной медицинской информации (PHI).	Синтетические данные гарантируют, что все конфиденциальные данные будут полностью анонимизированы. В сочетании с другими методами повышения приватности, такими как дифференциальная приватность, синтетические данные могут создавать необратимо приватные данные с математическими гарантиями.
Сбор и маркировка	Собирать реальные данные сложно, долго и дорого, и даже если они собраны, требуется много времени, денег и трудочасов, чтобы обеспечить очистку и маркировку данных для использования в разработке и обучении моделей ИИ.	Используя синтетические данные, организации могут создавать дополнительные маркированные примеры для обучающих наборов за меньшую стоимость, чем при традиционном ручном маркировании. Синтетические данные предоставляются «по требованию», поэтому новые данные могут генерироваться со скоростью разработки, а их маркировка происходит автоматически.
Соответствие нормативным требованиям	Растущее государственное регулирование ограничивает использование некоторых реальных данных. Такие нормативные акты, как ФЗ152, 148, зарубежные - GDPR, CCPA, HIPAA, Закон ЕС об ИИ и другие, затрудняют использование реальных данных в бизнес-приложениях.	Поскольку синтетические данные основаны на реальных, но не являются таковыми, они, как правило, отличаются от исходных данных. Синтетические данные помогают снизить риски несоблюдения этих норм и одновременно делают данные легкодоступными для инноваций.
Предвзятость и разнообразие	Данные реального мира могут содержать нежелательные погрешности. Такие дисбалансы, как раса, класс, пол и т. д., приведут к нежелательной предвзятости в последующих приложениях, для которых используются данные.	Синтетические данные позволяют увеличить число недопредставленных классов, чтобы окончательно изменить распределение в наборе данных. Это используется для устранения нежелательной предвзятости и позволяет организациям создавать справедливые, ответственные и репрезентативные приложения.
Предвзятость и разнообразие	Реальные данные часто оказываются неполными. Из-за ручного ввода данных, игнорирования или проблемных полей, а также из-за переходов системы.	Синтетические модели данных используются для заполнения пробелов в наборах данных контекстно-последовательной информацией.
Объем и поставки	Многие организации сталкиваются с «Уловкой-22»: им нужны ценные данные для роста, но прежде чем получить ценные данные, необходимо вырасти.	В случае проблемы «холодного старта», когда данных не существует, организации создают синтетические данные с помощью простой инструкции на естественном языке или определения схемы.
Новые или смоделированные условия	Реальные данные часто не являются широкодоступными для отдаленных пограничных случаев или ситуаций, таких как мошенничество, когда существует лишь ограниченное количество случаев и недостаточно данных для моделирования.	Синтетические модели данных позволяют организациям обучать модели ИИ на отдаленных, маловероятных или гипотетических пограничных случаях путем генерации данных для новых, неизведанных сред.



Основные понятия оценки дискретных моделей



- ✓ Аккуратность (Accuracy) = Правильные($TP+TN$)/Вся выборка $\rightarrow 1$
- ✓ F – мера (не путать с коэффициентом Фишера!)
 - $F=2*(Precision \times Recall)/(Precision+Recall)$, где
- ✓ Точность (precision)
 - $Precision=TP/(TP+FP)$
- ✓ Полнота (recall)
 - $Recall=TP/(TP+FN)$

TP — истинно-положительное решение;

TN — истинно-отрицательное решение;

FP — ложно-положительное решение;

FN — ложно-отрицательное решение.

Контроль отбора проб



- ✓ **LLM** формально не предсказывают одну лексему. Они предсказывают вероятности того, какой может быть следующая лексема, причем каждая лексема в словаре получает свою вероятность. Вероятности этих лексем затем сэмплируются, чтобы определить, какой будет следующая произведенная лексема.
- ✓ **Temperature, top-K и top-P** - наиболее распространенные параметры конфигурации, которые определяют, как обрабатываются предсказанные вероятности токенов для выбора одного выходного токена



Top-K vs Top-P настройки моделей

- ✓ Top-K и top-P (также известные как выборка ядра) - это две настройки выборки, используемые в LLM, чтобы ограничить предсказание следующего лексем лексемами с наибольшей предсказанной вероятностью.



Top-P настройки моделей

- ✓ Выборка Top-P отбирает лучшие лексемы, чья кумулятивная вероятность не превышает заданного значения (P).
- ✓ Значения P варьируются от 0 (жадное декодирование) до 1 (все лексемы в словаре LLM)



Тор-К настройки моделей

- ✓ Выборка Тор-К отбирает K наиболее вероятных лексем из предсказанного моделью распределения.
- ✓ Чем выше $\text{top-}K$, тем более творческим и разнообразным будет вывод модели; чем ниже $\text{top-}K$, тем более спокойным и фактическим будет вывод модели.
- ✓ Значение $\text{top-}K$, равное 1, эквивалентно жадному декодированию.

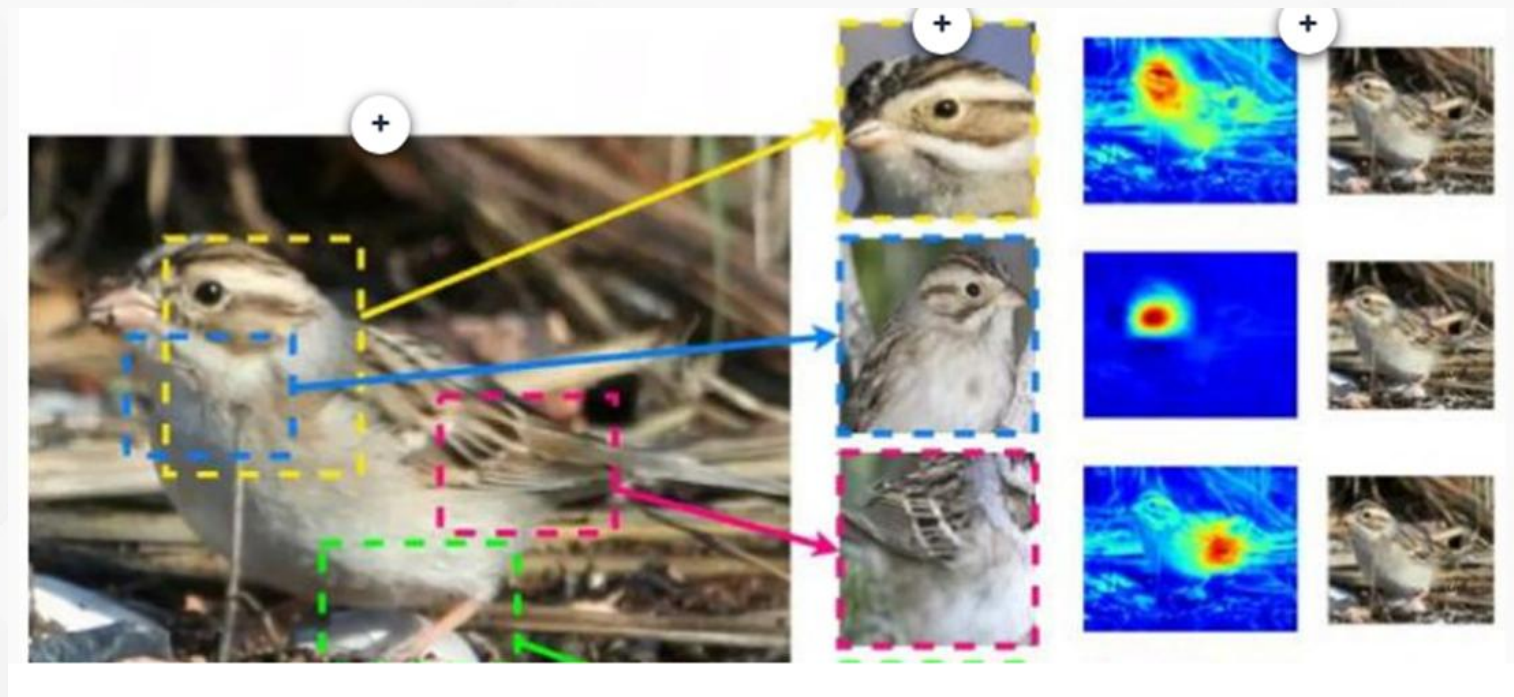
Температурные настройки



- ✓ Температура регулирует степень случайности при выборе.
- ✓ Более низкие температуры подходят для подсказок, которые ожидают более детерминированного ответа, в то время как более высокие температуры могут привести к более разнообразным или неожиданным результатам.
- ✓ Температура 0 - жадное декодирование, детерминированный подход с явно выбранным маркером наибольшей вероятности
- ✓ Температура, близкая к максимальной, как правило, создает больше случайных результатов. И по мере того как температура становится все выше и выше, все токены становятся одинаково вероятными для того, чтобы стать следующим предсказанным токеном

Раскрытие «черного ящика»

- ✓ Модели ИИ, особенно сети глубокого обучения, часто называют «черными ящиками», потому что их внутренняя работа сложна и трудна для интерпретации.
- ✓ ИИ стремится открыть этот черный ящик, визуализируя, как и почему ИИ принимает решения, такие как выявление признаков, которые привели к классификации птицы.



Важность функций и приоритетные области

Методы ИИ могут выделять, на каких частях изображения сфокусировалась модель ИИ для принятия решения. Например, он может выделять крылья, клюв или узоры оперения птицы, давая представление о том, что модель «заметила», определяя, что объект действительно является птицей.



Практические Рекомендации



Для сложных задач разбейте подсказку на более мелкие, управляемые шаги.

Такой подход гарантирует, что каждая часть задачи будет тщательно проработана и может повысить точность конечного ответа.

Например, если вы работаете над комплексным проектом, вы можете начать с того, чтобы попросить GPT составить план или ключевые идеи.

Как только они будут доработаны, следуйте подсказкам, чтобы расширить каждый раздел по отдельности.

Этот пошаговый подход позволяет искусственному интеллекту предоставлять более целенаправленные ответы, а вам — направлять его на протяжении всего пути, уточняя и развивая каждый ответ.



Практические Рекомендации

При необходимости укажите соответствующий контекст и данные в запросе. ИИ работает лучше, когда у него есть необходимая база, поэтому кратко опишите любой необходимый контекст для вашего запроса.

Например, запрос может включать дополнительный контекст или прикреплять связанную информацию, такую как PDF-файл или набор данных, что позволяет GPT анализировать или ссылаться на определенные детали.

Вместо того, чтобы спрашивать: «Обобщите глобальные тенденции», более эффективным вопросом с учетом контекста было бы следующее: «Используя прилагаемый отчет, предоставьте краткое изложение текущих глобальных экономических тенденций с акцентом на развивающиеся страны».



Практические Рекомендации

Попросите ИИ «отразить» или «рассмотреть» различные аспекты, если вы ищете вдумчивый ответ.

Такие формулировки, как «рассмотреть разные точки зрения» или «предоставить анализ», помогают смоделировать более глубокие рассуждения.

Например, пользователь может работать над предложением клиента. Он поручает GPT пошагово рассчитать затраты, учесть скидки и объяснить добавленную стоимость тех или иных элементов пакета.

Этот метод эффективен для задач, связанных с несколькими факторами или решениями.

Поощряя модель «обдумывать» каждый аспект перед ответом, можно получить более полный и логически структурированный ответ, особенно для задач, включающих сценарии, расчеты или стратегические рекомендации.

Используйте ролевую модель и промпты от GPT



Настройте ответы GPT, создав «пользовательский GPT», настроив запрос в качестве определенной роли (например, «действовать как аналитик данных» или «притворяться, что вы менеджер по продукту»).

Такой подход заставляет ИИ адаптировать свой тон и реакцию в соответствии со специализированными потребностями.

Попросите GPT задать вопросы, например, «Что бы ты посоветовала уточнить еще?», «Какие вопросы ты бы задала себе сейчас?»



Доверие и подотчетность

Объясняя решения ИИ, AI помогает укрепить доверие к системам ИИ, показывая, что они не просто выдают случайные результаты, а вместо этого основаны на различных закономерностях и логических структурах. Это имеет решающее значение для отраслей, где понимание принятия решений с помощью ИИ может повлиять на справедливость, предвзятость и подотчетность.

Развивая доверие к ИИ.

ИИ и блокчейн

Децентрализованный ИИ



- ✓ Децентрализованный ИИ относится к системам искусственного интеллекта, которые работают в распределенной сети, а не полагаются на центральный орган или сервер. Такой подход обеспечивает большую автономность, конфиденциальность и безопасность при обработке и обработке данных.
- ✓ Это расширяет возможности пользователей, предоставляя им контроль над своими данными, снижает риск возникновения единых точек отказа, повышает устойчивость системы, поощряет совместную работу и инновации в различных сетях, а также способствует прозрачности и подотчетности в процессах принятия решений с помощью ИИ.
- ✓ Децентрализованный ИИ важен, потому что он решает несколько критических проблем, связанных с традиционными системами ИИ, таких как конфиденциальность данных, предвзятость и монополизация технологий. Распределяя власть и контроль над ИИ, он способствует созданию более справедливого технологического ландшафта.



Децентрализованный ИИ отличается от традиционного ИИ несколькими ключевыми аспектами

- ✓ **Владение данными:** В традиционном ИИ данные часто хранятся и контролируются центральным органом, что приводит к опасениям по поводу конфиденциальности и неправомерного использования. Децентрализованный искусственный интеллект позволяет пользователям сохранять право собственности на свои данные, повышая конфиденциальность.
- ✓ **Вычислительная мощность:** Традиционный ИИ полагается на централизованные серверы для обработки, что может создавать узкие места и уязвимости. Децентрализованный ИИ распределяет задачи обработки между несколькими узлами, повышая эффективность и отказоустойчивость.
- ✓ **Принятие решений:** Традиционные системы ИИ часто работают как черные ящики, что затрудняет понимание того, как принимаются решения. Децентрализованный ИИ способствует прозрачности, поскольку решения могут быть отслежены до отдельных узлов и их входных данных.
- ✓ **Сотрудничество:** децентрализованный ИИ поощряет сотрудничество между различными заинтересованными сторонами, обеспечивая более разнообразный спектр входных данных и точек зрения. Традиционный ИИ часто ограничивает совместную работу избранной группой разработчиков и организаций.

Преимущественные области применения децентрализованного ИИ



- ✓ **Здравоохранение:** децентрализованный искусственный интеллект может способствовать безопасному обмену данными о пациентах между поставщиками медицинских услуг, что позволяет улучшить диагностику и планирование лечения. Он также может способствовать разработке лекарств за счет анализа обширных наборов данных без ущерба для конфиденциальности пациентов.
- ✓ **Финансы:** В финансовом секторе децентрализованный ИИ может улучшить обнаружение мошенничества и оценку рисков. Анализируя шаблоны транзакций в распределенной сети, он может выявлять аномалии более эффективно, чем традиционные системы.
- ✓ **Управление цепочкой поставок:** децентрализованный искусственный интеллект может оптимизировать операции в цепочке поставок, предоставляя информацию об уровнях запасов, прогнозировании спроса и логистике в режиме реального времени. Это приводит к повышению эффективности и снижению затрат.
- ✓ **Управление энергопотреблением:** ИИ может быть использован для оптимизации потребления энергии в децентрализованных энергетических сетях. Анализируя данные из различных источников, он может прогнозировать спрос на энергию и соответствующим образом корректировать предложение, способствуя устойчивому развитию.
- ✓ **Маркетинг и реклама:** децентрализованный искусственный интеллект может анализировать поведение потребителей на нескольких платформах, что позволяет предприятиям создавать целевые маркетинговые кампании. Это приводит к более высоким коэффициентам конверсии и улучшению взаимодействия с клиентами.
- ✓ **Автономные транспортные средства:** децентрализованные сети искусственного интеллекта могут поддерживать разработку автономных транспортных средств, обеспечивая обмен данными между транспортными средствами в режиме реального времени. Это повышает безопасность и эффективность на дорогах.

Снижение стоимости при использовании ИИ на блокчейне



- ✓ **Совместное использование ресурсов:** В децентрализованной сети участники могут делиться своими вычислительными ресурсами. Такой подход к совместной работе снижает потребность в дорогостоящих централизованных центрах обработки данных, снижая общие затраты на инфраструктуру.
- ✓ **Периферийные вычисления:** децентрализованный искусственный интеллект часто использует периферийные вычисления, где обработка данных происходит ближе к источнику данных. Это сводит к минимуму задержку и снижает пропускную способность, необходимую для передачи данных, что приводит к экономии средств.
- ✓ **Эффективная обработка данных:** используя федеративное обучение, децентрализованный ИИ позволяет обучать модели на локальных устройствах без передачи больших наборов данных на центральный сервер. Это снижает затраты на передачу данных и ускоряет процесс обучения.
- ✓ **Динамическое масштабирование:** децентрализованные сети могут динамически масштабировать ресурсы в зависимости от спроса. Такая гибкость позволяет предприятиям платить только за необходимую им вычислительную мощность, избегая затрат, связанных с избыточным выделением ресурсов.
- ✓ **Снижение лицензионных сборов:** Многие децентрализованные решения ИИ имеют открытый исходный код, что может устранить или значительно снизить лицензионные сборы, связанные с проприетарным программным обеспечением. Это делает расширенные возможности искусственного интеллекта более доступными для предприятий любого размера.
- ✓ **Снижение затрат на обслуживание:** В децентрализованных системах нагрузка по обслуживанию распределяется между участниками. Это может привести к снижению эксплуатационных расходов по сравнению с обслуживанием централизованной инфраструктуры, где полную ответственность несет один субъект.

Спасибо за внимание!

Ваши вопросы...

<https://www.specialist.ru/course/ds0>

dinzis@specialist.ru
consult@Dintsis.org

<https://t.me/joinchat/Ekkyg1Hrtyb1vjBseB4nSA>

<https://t.me/pmtips>

