Содержание

От издательства

Отзывы и пожелания

Мы всегда рады отзывам наших читателей. Расскажите нам, что вы думаете об этой книге – что понравилось или, может быть, не понравилось. Отзывы важны для нас, чтобы выпускать книги, которые будут для вас максимально полезны.

Вы можете написать отзыв на нашем сайте www.dmkpress.com, зайдя на страницу книги и оставив комментарий в разделе «Отзывы и рецензии». Также можно послать письмо главному редактору по адресу dmkpress@gmail.com; при этом укажите название книги в теме письма.

Если вы являетесь экспертом в какой-либо области и заинтересованы в написании новой книги, заполните форму на нашем сайте по адресу http://dmkpress.com/authors/publish_book/ или напишите в издательство по адресу dmkpress@gmail.com.

Список опечаток

Хотя мы приняли все возможные меры для того, чтобы обеспечить высокое качество наших текстов, ошибки все равно случаются. Если вы найдете ошибку в одной из наших книг, мы будем очень благодарны, если вы сообщите о ней главному редактору по адресу dmkpress@gmail.com. Сделав это, вы избавите других читателей от недопонимания и поможете нам улучшить последующие издания этой книги.

Нарушение авторских прав

Пиратство в интернете по-прежнему остается насущной проблемой. Издательство «ДМК Пресс» очень серьезно относится к вопросам защиты авторских прав и лицензирования. Если вы столкнетесь в интернете с незаконной публикацией какой-либо из наших книг, пожалуйста, пришлите нам ссылку на интернет-ресурс, чтобы мы могли применить санкции.

Ссылку на подозрительные материалы можно прислать по адресу электронной почты dmkpress@gmail.com.

Мы высоко ценим любую помощь по защите наших авторов, благодаря которой мы можем предоставлять вам качественные материалы.

Предисловие

Уважаемый читатель! Эта книга открывает дверь в увлекательный мир искусственного интеллекта (ИИ), в мир DeepSeek.

История развития ИИ насчитывает уже более 80 лет. В 1943 году была предложена концепция нейронных сетей, которая заложила основу для создания ИИ. В 1956 году состоялся Дартмутский семинар, на котором впервые прозвучал термин *искусственный интеллект* (Artificial Intelligence) и фактически произошло образование новой области исследований. За последние 80 лет развитие ИИ прошло три этапа: от «слабого ИИ на основе заданных правил» к «сильному ИИ на основе статистического и глубокого обучения». И наконец, к современной «эре искусственного сверхинтеллекта, где сочетаются глубокое обучение и интеграция интеллектуальных возможностей». Путь развития ИИ был непростым – периоды бурного роста сменялись затишьем. Но именно в это время появились выдающиеся ученые и революционные идеи в области алгоритмов, инженерии и обработки данных.

Однако моментом, когда ИИ произвел наиболее сильное впечатление и расширил границы воображения, безусловно, является появление нашумевшего в последние годы проекта компании OpenAI – ChatGPT. Этот инструмент позволяет свободно общаться и творить. И очень сложно поверить, что по другую сторону экрана находится машина, а не человек. Уровень интеллекта ChatGPT перевернул представления многих людей о возможностях компьютера. Однако вместе с восторгом ИИ принес и новые тревоги о будущем многих профессий и даже о самом выживании человечества. Он внес в мир элемент неопределенности: мы беспокоимся о постепенной замене человека компьютером и особенно боимся, что в будущих войнах люди окажутся беспомощными перед машинами. Единственный способ противостоять этой неопределенности – познать и принять ИИ. Сразу после появления ChatGPT ведущие ИТ-гиганты начали вкладывать огромные средства и вычислительные мощности в разработку больших языковых моделей. Среди значимых разработок можно отметить ChatGPT от OpenAI, Gemini от Google, Claude от Anthropic, серию моделей LLaMA от Meta (признана экстремистской организацией на территории РФ). В Китае появились Doubao от ByteDance, Tongyi от Alibaba, Hunyuan от Tencent, Ernie от Baidu, а также стремительно набирающая популярность модель DeepSeek от одноименной компании. В реальных сценариях необходимо тщательно сравнивать производительность, уровень интеллекта и стоимость использования моделей разных производителей, одновременно совершенствуя промты для достижения оптимальных результатов.

В конце 2024 года стартап из китайского города Ханчжоу представил большую языковую модель с открытым исходным кодом DeepSeek, которая произвела настоящую революцию в мире ИИ. Согласно отчету независимой аудиторской компании Artificial Analysis, модель DeepSeek-V3 превзошла другие открытые аналоги по множеству параметров и сравнялась по показателям с ведущими проприетарными моделями, такими как GPT-4o и Claude-3.5-Sonnet. При этом стоимость DeepSeek API составляет всего 1/70 от цены GPT-4 Turbo. Феноменальное соотношение цены и качества DeepSeek привлекло широкое внимание профессионального сообщества, побудив множество компаний и технических специалистов изучать, внедрять и использовать эту модель. Однако материалы по работе с DeepSeek остаются крайне скудными и бессистемными. Наша книга призвана заполнить тот пробел, ее лаконичные и четкие объяснения помогут читателям освоить DeepSeek и станут руководством для специалистов по ИИ, стремящихся идти в ногу со временем. Но не будем больше занимать время читателя долгими предисловиями – переверните страницу и начните свое путешествие в мир DeepSeek!

Ветеран цифровой индустрии
с 15-летним опытом работы в сфере интернет-технологий,
специалист в области искусственного интеллекта

Чжу Цзюньчэн (Zhu Juncheng)

Введение

Для чего написана эта книга

Волна развития искусственного интеллекта вынесла на берег технологию больших языковых моделей, которая кардинально изменила наш образ жизни и работы. В конце 2024 года появилась восходящая звезда искусственного интеллекта – платформа DeepSeek, которая стала одним из мировых лидеров в этой области. Однако новичкам бывает сложно подступиться к такой сложной и мощной технологии, у них появляются вопросы:

1) «каковы возможности DeepSeek?»;
2) «как можно повысить эффективность своей работы с помощью DeepSeek?»;
3) «как использовать DeepSeek в реальных задачах?».

Желание ответить на них и побудило авторов написать эту книгу. Книга призвана помочь читателю быстро освоить основные приемы работы с системой искусственного интеллекта DeepSeek, научиться применять ее в различных сценариях и в конечном итоге использовать всю мощь DeepSeek для решения практических задач.

Как использовать книгу

Изучение DeepSeek заключается не только в освоении навыков использования этого инструмента, но и в понимании процесса работы искусственного интеллекта, а также границ его применения. Книга предназначена для широкого круга читателей с разным уровнем подготовки. Обучение построено последовательно шаг за шагом:

1. **Начальный уровень**. Знакомство с нуля и быстрое освоение основных функций и приемов работы с DeepSeek.
2. **Продвинутый уровень**. Применение DeepSeek в реальных сценариях, таких как анализ данных, создание текстов и бизнес-планирование.
3. **Профессиональный уровень**. Глубокое погружение в возможности DeepSeek: интеллектуальные решения в области здравоохранения, финансов, образования и других отраслях.

В процессе обучения читателю следует уделить особое внимание практике, а также гибко корректировать учебный план в соответствии со своими потребностями.

Содержание книги

Книга разделена на 8 глав:

- глава 1: введение в ключевые технологии DeepSeek, сферы применения и методы настройки;
- глава 2: подробное руководство по использованию DeepSeek: от основ интерфейса до практических техник ведения диалога;
- глава 3: подробный разбор концепции структурированных промтов: преимущества и правила составления эффективных запросов;
- глава 4: углубленное изучение уникальных возможностей DeepSeek с готовыми решениями для разных задач;
- глава 5: практические советы по использованию AI-помощников для повышения продуктивности в работе и учебе;
- глава 6: продвинутые техники работы с DeepSeek для максимального раскрытия потенциала искусственного интеллекта;
- глава 7: интеграция инструментов и локальное развертывание DeepSeek, тонкая настройка и обеспечение безопасности данных;
- глава 8: будущее DeepSeek и перспективы развития ИИ-помощников – как технологии изменят наши возможности.

Преимущества книги

Книга обладает следующими преимуществами:

1) простое изложение: от базовых концепций до практического применения – книга постепенно ведет читателя к пониманию сложной технической логики;
2) практические примеры: книга содержит практические примеры из различных областей, помогая научиться быстро применять новые знания к своим потребностям;
3) наглядность: понятные иллюстрации и подробные блок-схемы облегчают процесс обучения;
4) последовательность: материал излагается последовательно от базовых понятий до профессиональных приемов, позволяя познакомиться с технологиями искусственного интеллекта читателям с разным уровнем подготовки.

Предложения и отзывы

Написание книги требует много времени и сил. Авторы приложили максимум усилий, чтобы приблизить книгу к идеалу, но в ней по-прежнему могут содержаться пробелы и недостатки. Отзывы читателей очень приветствуются, они помогут сделать книгу лучше и повысить ее ценность. Комментарии и предложения, а также любые вопросы, связанные с этой книгой, можно направить через мессенджер WeChat: mjcoding (Мэн Цзянь) и ylx2ai (Яо Лусин). Будем рады вашим ценным комментариям!

Благодарности

Мы хотели бы поблагодарить команду DeepSeek за их усилия по популяризации технологии искусственного интеллекта, а также наших коллег и друзей за их бесценную поддержку во время написания этой книги. Мы надеемся, что эта книга станет отличным помощником в изучении DeepSeek и вдохновит вас на новые открытия в работе с искусственным интеллектом!

Мэн Цзянь (Meng Jian), Яо Лусин (Yao Luxing)
Февраль 2025 г.

Глава 1

Знакомство с DeepSeek

Перед тем как приступить к изучению работы с DeepSeek, давайте сначала познакомимся с историей создания, техническими характеристиками и положением на рынке этой системы. В этой главе вы получите четкое представление об основах платформы DeepSeek, сценариях применения и уникальных преимуществах, что заложит прочную основу для использования ее на практике.

1.1. История и цели создания DeepSeek

В последние годы технологии искусственного интеллекта получили сильный импульс развития, особенно так называемые *большие языковые модели* (БЯМ, Large Language Model, LLM), которые кардинально меняют все сферы жизни. Стремительно ворвавшись на рынок, платформа DeepSeek быстро превратилась в одну из ведущих в мире моделей ИИ благодаря своему уникальному технологическому пути и четкому позиционированию продукта. В этом разделе вы познакомитесь с историей создания и развития DeepSeek, а также проследите путь трансформации от стартапа к отраслевому эталону.

1.1.1. Предыстория и этап подготовки

Историю DeepSeek можно начать со стратегических инициатив в сфере искусственного интеллекта китайской компании по количественным инвестициям High-Flyer Quant. В мае 2023 года компания DeepSeek выделилась из состава High-Flyer Quant и была официально учреждена 17 июля того же года. Штаб-квартира компании расположена в городе Ханчжоу, Китай. Ее основатель Лян Вэньфэн (Liang Wenfeng) на начальном этапе заложил прочную основу развития DeepSeek, интегрировав все преимущества High-Flyer Quant в области вычислительных мощностей, капитала и технологий.

Первые шаги компании. Первоначальные расходы на разработку DeepSeek были напрямую профинансированы компанией High-Flyer Quant, которая также предоставила доступ к своим суперкомпьютерным ресурсам Yinghuo с десятками тысяч графических процессоров (англ. graphic processing unit, GPU). Это позволило DeepSeek с момента основания получить мощную вычислительную базу.

Стратегическое позиционирование. Основную цель можно сформулировать как «демократизация технологий». DeepSeek направляет свои усилия на обеспечение всеобщего доступа к искусственному интеллекту через открытые большие языковые модели (БЯМ). Это позволит преодолеть технологическую монополию Запада. С этой целью компания на раннем этапе существования инвестировала более 500 млн долларов США в закупку GPU для создания надежной вычислительной инфраструктуры.

1.1.2. Технологическая эволюция и основные инновации

DeepSeek в кратчайшие сроки разработала и сменила несколько поколений своих моделей. Технологический путь развития компании сочетает в себе инновационные алгоритмы, контроль затрат и исследования в области мультимодальных возможностей. Хронология развития представлена на рис. 1.1.

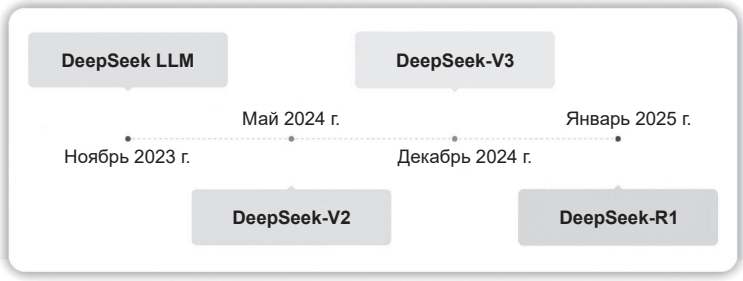

Рис. 1.1 ❖ История развития DeepSeek

Хронологически путь разработки базовой модели DeepSeek выглядит следующим образом.

1. DeepSeek LLM (ноябрь 2023 г.). Первая открытая модель, основанная на трансформерной архитектуре LLaMA, была представлена в двух вариантах: 6,7 млрд параметров и 67 млрд параметров. Обладая способностью к генерации текста и ведению диалога, она стала отправной точкой технологического развития компании.
2. DeepSeek-V2 (май 2024 г.). Модель второго поколения на основе смеси экспертов (англ. Mixture of experts, MoE) демонстрирует сопоставимую с GPT-4 Turbo производительность, при этом стоимость логическо-

го вывода составляет лишь 1/70 от затрат GPT-4, что свидетельствует о значительном превосходстве в экономической эффективности.

3. DeepSeek-V3 (декабрь 2024 г.). Модель с 671 млрд параметров превзошла Qwen2.5-72B и Llama-3.1-405B в многочисленных тестах, приблизившись к уровню производительности GPT-4o и Claude 3.5.
4. DeepSeek-R1 (январь 2025 г.). Демонстрирует выдающиеся результаты в таких задачах, как математические рассуждения и генерация кода. R1 не только поддерживает локальное развертывание для повышения защиты конфиденциальных данных, но и снижает стоимость обучения на 90 % по сравнению с Claude 3.5 Sonnet.

Ключевые технологические инновации отражены в следующих двух пунктах:

1. **Контроль затрат**. Благодаря исследованиям в области механизма многоуровневого латентного внимания (англ. Multi-Head Latent Attention, MLA), эффективных алгоритмов обучения с подкреплением (таких как GRPO) и законов масштабирования гиперпараметров DeepSeek добилась значительного снижения расходов на обучение и логический вывод моделей. Например, предварительное обучение DeepSeek-V3 потребовало всего 2,664 млн часов работы GPU H800, при этом экономические затраты составили около 5,6 млн долларов США.
2. **Революционные алгоритмы**. Модель DeepSeek-R1 отлично справляется с такими задачами, как математические рассуждения и генерация кода. Стоимость ее обучения снижена на 90 % по сравнению с Claude 3.5 Sonnet, и она поддерживает локальное развертывание для повышения защиты конфиденциальных данных.

1.1.3. Эволюция продукта и рыночные показатели

Технологические инновации DeepSeek быстро отражаются на рыночных показателях, а продукты компании совершают прорывы во многих областях.

1. Взрыв популярности

Мобильное приложение. На рис. 1.2 видно, что 27 января 2025 года приложение DeepSeek заняло первое место в топе бесплатных приложений App Store в Китае и США, достигнув 16 млн загрузок. Темп роста на 100 % превысил аналогичный показатель для ChatGPT.

Рис. 1.2 ❖ DeepSeek возглавляет список бесплатных приложений App Store в Китае и США

Корпоративное сотрудничество. DeepSeek привлекла к своим моделям мировых гигантов, таких как NVIDIA, Amazon и Microsoft, а также ведущих китайских поставщиков облачных услуг, таких как Huawei Cloud и Tencent Cloud.

2. Сравнительный анализ эффективности и стоимости

DeepSeek-R1 по эффективности сопоставим с GPT-4o mini, но стоит на 90 % меньше.

DeepSeek-V3 продемонстрировала близкие к человеку способности в тестах по искусственному общему интеллекту (англ. Artificial General Intelligence, AGI). Модель поддерживает физическое моделирование и творческую генерацию, включая такие сложные задачи, как четырехмерное программирование и разработка игр.

1.1.4. Влияние на отрасль и перспективы развития

Компания DeepSeek не только изменила ландшафт индустрии ИИ в Китае, но и оказала глубокое влияние на глобальный рынок. Основные причины такого влияния:

1) **популяризация технологии**. Благодаря низкой стоимости и высокой эффективности DeepSeek бросает вызов технологическому доминированию Соединенных Штатов в сфере ИИ. Например, стоимость логического вывода DeepSeek-V2 составляет лишь 1/70 от затрат GPT-4 Turbo, что сильно способствует популяризации технологий искусственного интеллекта;
2) **создание экосистемы с открытым исходным кодом**. Благодаря инструментам с открытым исходным кодом на платформе GitHub (таким как DeepSeek-Coder) разработчики получили возможность легко интегрировать ИИ-функции помощи в программировании. Это способствует формированию мощного технологического сообщества вокруг платформы;
3) **глобальная экспансия**. DeepSeek поддерживает много языков, а в таких регионах, как Индия, предоставляет бесплатные услуги, что значительно расширяет географию присутствия компании.

1.2. Основные достижения и преимущества DeepSeek-R1

Модель DeepSeek-R1 стала ключевой вехой в линейке моделей DeepSeek. Благодаря многогранным технологическим инновациям и инженерной оптимизации она значительно повысила способности к логическим рассужде-

ниям и операционную эффективность больших языковых моделей. В этом разделе максимально простым языком объясняются основные достижения DeepSeek-R1.

1.2.1. Прорыв в автономном обучении: рассуждения на основе обучения с подкреплением

Традиционные большие языковые модели требуют обширных размеченных данных для *дообучения с учителем* (Supervised Fine-Tuning, SFT) для развития способности к логическим выводам. DeepSeek-R1, напротив, впервые использует полностью основанный на *обучении с подкреплением* (Reinforcement Learning, RL) метод тренировки модели.

Новаторство DeepSeek-R1 заключается в том, что в ней полностью отказались от традиционного этапа дообучения, ограничившись только обучением с подкреплением без ущерба для способности вести сложные рассуждения. Это позволяет модели заниматься самопроверкой, рефлексивным мышлением и генерировать длинные логические цепочки, называемые *цепочками мыслей* (Chain Of Thought, CoT). DeepSeek-R1-Zero – это предобученная базовая модель DeepSeek-R1. Ее контрольные данные испытаний приведены в табл. 1.1. AIME 2024 представляет собой соревновательный математический тест высокого уровня, основной задачей которого является проверка способности модели к математическим рассуждениям. MATH-500 – это специализированный набор данных для оценки математических навыков решения задач, содержащий 500 тестовых образцов. GPQA служит для тестирования общих навыков программирования, LiveCode оценивает возможности программирования в реальном времени, а CodeForces предназначен для тестирования навыков соревновательного программирования. Например, в тесте AIME 2024 показатель успешности DeepSeek-R1-Zero продемонстрировал впечатляющий рост с первоначальных 15,6 % до 71,0 %, что свидетельствует о значительном прогрессе в математических способностях модели.

Таблица 1.1. Данные испытания модели DeepSeek-R1-Zero

Модель	AIME 2024		MATH-500	GPQA Diamond	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	rating
OpenAI-o1-mini	63,6	80,0	90,0	60,0	53,8	1820
OpenAI-o1-0912	74,4	83,3	94,8	77,3	63,4	1843
DeepSeek-R1-Zero	71,0	86,7	95,9	73,3	50,0	1444

Этот метод обучения не только разрушает традиционную парадигму, но и доказывает, что обучение с подкреплением может стимулировать автономный потенциал рассуждений модели, предоставляя новый подход для будущего развития ИИ.

1.2.2. Эффективная архитектура и методы обучения

Модель DeepSeek-R1 была существенно улучшена, в основном за счет следующих приемов:

1) **архитектура смеси экспертов** (Mixture Of Experts, MoE). Распределяя задачи между различными «экспертами», модель избегает бесполезной траты ресурсов, одновременно значительно снижая вычислительные затраты;
2) **тренировка с FP8-смешанной точностью.** Благодаря совместному проектированию алгоритмов и аппаратного обеспечения решена проблема узких мест в межузловой коммуникации, что позволило сократить стоимость обучения модели с 671 млрд параметров до всего лишь 2,664 млн часов работы GPU H800;
3) **механизм многоуровневого латентного внимания** (Multi-Head Latent Attention, MLA). Позволил дополнительно оптимизировать скорость логических выводов, сократив затраты на ввод и вывод данных до 1/3–1/5 по сравнению с конкурентами.

Подобные оптимизации позволили DeepSeek-R1 сохранить выдающуюся производительность при одновременном значительном снижении требований к вычислительным ресурсам как в процессе обучения, так и при выполнении логических выводов.

1.2.3. Компактные модели – мощный интеллект: дистилляция моделей и развертывание на периферийных устройствах

DeepSeek-R1 демонстрирует выдающуюся производительность не только среди больших моделей, но и благодаря технологии «дистилляции моделей» эффективно сжимает возможности больших моделей в компактные версии.

Результаты дистилляции. Например, модель DeepSeek-R1-Distill-Qwen-32B достигает точности 94.3 % в тесте MATH-500, превосходя даже оригинальную модель QwQ-32B.

Развертывание на периферийных устройствах. Благодаря интеграции с технологиями низкобитового квантования компактные модели могут работать на потребительских видеокартах, таких как RTX 3060, делая передовые ИИ-технологии доступными для широкого круга пользователей.

Это открывает путь для переноса больших моделей из облачной среды на периферийные устройства, расширяя доступ к возможностям ИИ.