**Optimizing DeepSeek-R1 Workflows**

14 sources

The provided texts discuss various aspects of artificial intelligence, focusing on the advancements and applications of large language models (LLMs). Several sources explore the development and capabilities of new LLMs, such as DeepSeek R1, highlighting their reasoning abilities and cost-effectiveness compared to existing models. Other sources detail the use of AI in diverse fields including marketing, business, and content creation, showcasing how AI tools are utilized for tasks ranging from generating marketing materials and optimizing ads to automating workflows and creating various types of content. Finally, some texts examine the financial opportunities presented by AI, specifically within cryptocurrency and stock markets, and the implications of AI on job markets and career strategies.

keepSave to notecopy\_all

docsAdd noteaudio\_magic\_eraserAudio OverviewschoolBriefing doc

AI App Development Landscape: ROI & Evaluation

Okay, let's dive into the AI app development landscape, focusing on ROI and evaluation, as you requested. Here's a breakdown of key aspects based on the provided sources:

**AI App Development & ROI**

•

**Efficiency and Automation:** AI is transforming how businesses operate, and AI-powered apps are at the forefront. These apps can automate repetitive tasks, freeing up human teams for more strategic and creative work1. This efficiency gain can directly impact the bottom line, representing a significant ROI driver.

•

**Reduced Costs:** AI can create unlimited ad creatives at a fraction of the cost of traditional methods2.... By using AI influencers for marketing, for example, brands can significantly cut content creation costs3. AI-powered tools can help you analyze market trends and customer needs to generate business ideas4, or provide competitor research for small businesses5, which can lead to greater efficiency.

•

**Increased Revenue:** AI-driven applications can lead to new revenue streams. For example, AI agents can be tokenized and co-owned, generating revenue for multiple stakeholders6. AI can also help businesses with personalized marketing, lead generation, and sales, which can increase revenue7.

•

**Time Savings:** AI can accomplish a day's work in just minutes, which saves time8. AI tools are designed to streamline workflows9.... In development, AI tools can speed up coding processes11. For example, AI can generate code based on natural language prompts11....

•

**Accessibility:** AI platforms are making it easier for individuals without coding experience to create AI agents and automations13. This democratizes AI development and allows more people to benefit from it13....

**Evaluation of AI Apps and Agents**

•

**Focus on Reasoning:** The sources emphasize the importance of reasoning capabilities in AI models16. Models like DeepSeek-R1 are designed with strong reasoning abilities, outperforming others in benchmarks17.... It appears that evaluating reasoning is key to understanding how well an AI agent or app performs21....

•

**Workflow Integration:** AI models perform best when integrated into workflows25. This suggests that evaluation should consider how well an AI app or agent works within a specific process. Using workflow programs to manage AI processes is useful for evaluating performance25.

•

**Multi-Modal Evaluation:** Some models are capable of processing various inputs, like images, videos, text, and audio26.... Therefore, an evaluation process may need to consider multiple input types.

•

**Real-world performance:** It's important to evaluate AI in real-world scenarios28.... Benchmarks provide some insights, but real-world data is often more complex. For example, the Spider 2.0 project evaluates language models on enterprise text-to-SQL workflows28, moving beyond simple queries to evaluate performance in complex data environments28.

•

**Tool Use and Retrieval:** With AI agents, it's essential to evaluate how well they call tools and retrieve information30. This is important to ensure that the agents are accurate and effective.

•

**Metrics for Evaluation:** The sources mention several metrics used to evaluate AI models, including pass@1, Elo rating, and accuracy on specific benchmarks23.... However, for real-world applications, you may need to develop custom metrics.

•

**Human Feedback:** Human feedback is critical to optimizing AI models29. In the development of DeepSeek-R1, human preferences were used to align the model22. Therefore, evaluations should include a human component.

•

**Hallucination and Failure:** Enterprises have less tolerance for hallucinations and failures in AI applications32. Evaluation should focus on identifying and mitigating these issues.

•

**RAGAS for Agent Evaluation**: RAGAS can be used to evaluate retrieval augmented generation in agents30.

**Specific AI Tools and Frameworks**

Many tools and frameworks are mentioned that can help with AI development and evaluation:

•

**DeepSeek R1:** A powerful open-source reasoning model16....

•

**Langchain:** A framework for connecting LLMs with external data35....

•

**Hugging Face:** Provides tools for AI development36....

•

**Relevance AI:** A tool for building and deploying AI agents without coding1.

•

**AI Gradio:** Simplifies the creation of machine-learning apps38....

•

**Cursor:** An AI enhanced code editor11....

•

**Vapi:** An AI voice agent platform43....

•

**Agena:** Tool to create AI agents and put them on websites and other platforms45....

•

**Bolt.DIY:** An open-source AI coding assistant18....

•

**AI White Labels:** A tool for building various AI tools quickly47....

•

**RooCline:** A tool to work with DeepSeek-R149.

•

**Aider:** Another tool to work with DeepSeek-R149.

•

**Lecca IO:** A no-code AI agent and automation platform13....

•

**Filament:** Full stack components to accelerate Laravel development52.

•

**Limbo:** SQL-like compatible database with asynchronous IO53.

•

**Ren AI:** Open source AI agents for data analysis54....

•

**Crawl 4 AI:** Open source web crawler and scraper56.

•

**Chainlit:** A Python-centric platform for building conversational AI applications57.

•

**Ai16z:** A decentralized autonomous organization that integrates AI with decentralized finance58....

•

**Vader AI:** An innovative crypto project that combines blockchain with autonomous AI agents60.

•

**Acolyte:** A platform that provides insights into the AI agent ecosystem61.

•

**Arc:** A decentralized platform that merges AI with blockchain62....

•

**Simmi:** An AI agent project that utilizes the Empyreal ecosystem64.

•

**Shulex VOC AI:** An AI tool for Amazon sellers to analyze reviews65.

•

**NotebookLM:** An AI tool for reviewing YouTube videos66.

•

**Prompt metheus**: A prompt engineering IDE11....

•

**Air Table AI:** Tool for database management67.

•

**Type:** An AI first document editor68.

•

**Compose AI:** An AI-powered text generator69.

•

**Taskade AI:** An AI tool to streamline collaboration and increase productivity70.

•

**Turntable:** A platform that enables less technical analysts to build data pipelines71.

•

**Mage:** An AI tool to create images and gifs72.

•

**Yarnet:** An AI tool to help with content creation73.

•

**DoNotPay:** An AI tool that helps users fight corporations, beat bureaucracy, and more7.

•

**Rubber Band:** Simplifies the process of training generative ML models10.

•

**Blocks:** An AI powered productivity assistant72.

•

**MeetGeeks:** An AI tool for summarizing meetings72.

•

**Spline AI:** A tool for generating 3D objects and animations74.

•

**Flawless AI:** An AI tool for filmmaking75.

•

**11 Labs:** An AI tool for voice generation76.

•

**Mini CPMO:** A multimodal model for vision, speech, and multimodal streaming on phones26.

•

**Spider 2.0:** A tool for evaluating language models on enterprise text to SQL28.

•

**Cbook:** A typescript-centric app development platform77.

•

**Co-pilot Kit:** A tool for building AI assistants in apps78.

•

**EnvId:** An AI tool for multimodel data extraction79.

•

**Ultra Box:** An AI model for real time voice interaction80.

•

**Moondream:** An open-source model for image understanding79.

**Key Takeaways**

•

The AI app development landscape is rapidly evolving, with new tools and frameworks emerging constantly.

•

AI offers significant potential for ROI through increased efficiency, reduced costs, and new revenue streams.

•

Evaluating AI apps and agents requires considering reasoning abilities, real-world performance, tool use, and human feedback.

•

Workflow integration and using appropriate evaluation metrics are also critical for measuring success.

I hope this gives you a solid overview of the AI app development landscape, focusing on ROI and evaluation. Let me know if you have further questions or need more details on any aspect!

keep\_pinSave to note

copy\_all

thumb\_up

thumb\_down

Which ones are no-code

Okay, I will identify the no-code AI app development tools and platforms mentioned in the sources. Here's a breakdown of the no-code options:

**No-Code AI App Development Platforms and Tools**

Based on the sources, these tools and platforms allow users to build AI applications without needing to write code:

•

**Relevance AI**: This is a tool that lets users build and deploy AI agents without coding1.

•

**Agena:** This tool is designed for creating AI agents and putting them on websites and other platforms without needing code1.

•

**Lecca IO:** This platform is a no-code option for building AI agents and automations1.

•

**Bolt.DIY**: While it's an open-source AI coding assistant, it's also described as a tool where you can build software applications with no code and just by using prompts, merging no-code drag-and-drop building with AI2. It also allows the building of websites and tools without code3.

•

**AI White Labels:** This tool allows for building software applications with no code, using drag-and-drop interfaces similar to tools like Bubble, and merging no code with AI prompting2. It can be used to create various AI tools quickly2....

•

**Shulex VOC AI:** This AI tool is designed for Amazon sellers to analyze reviews and does not seem to require coding, making it a no-code option for insights1.

•

**AI Gradio**: This is a tool that simplifies the creation of machine-learning apps, which may imply a no-code approach, but this is not explicitly stated1.

•

**Make.com**: This platform allows users to create automations by connecting different apps, including AI, without writing code5....

•

**Airtable AI**: This tool is designed for database management and provides AI capabilities that appear to be accessible without code1....

•

**Taskade AI:** This AI tool streamlines collaboration and increases productivity, likely with no-code implementation options1.

•

**Cbook:** While it uses typescript, the AI app builder component empowers you to create, edit, and run web apps with AI assistance, which suggests you don't need to write code directly to do so11.... The platform also provides a hot reloading web preview11.

These tools provide a range of functionalities, from building AI agents to creating websites, analyzing data, and automating tasks. They represent the growing trend of making AI more accessible to non-coders.

**Important Considerations for No-Code**

•

While these tools are "no-code," some might still require a basic understanding of logical workflows, and in some cases, you may need to write or edit prompts.

•

The complexity of the applications you can create might be limited compared to coding from scratch. However, for many tasks, these tools provide a quick, accessible way to build AI-powered apps and agents.

•

You should consider how much control you want over the design and behavior of the app, as many of these no-code options have pre-built UI components that can be used for a quick setup.

•

Some tools use AI in their interface or development, but that does not mean the tool itself is no-code.

Let me know if you want to explore any of these platforms further or have any more questions.

keep\_pinSave to note

copy\_all

thumb\_up

thumb\_down

what is the cost structure of each app dev

Okay, let's break down the cost structures associated with the AI app development tools and platforms discussed in the sources. It's important to note that pricing models vary widely, with some offering free tiers, while others operate on subscription or usage-based models. Here's a look at what the sources indicate about the cost structures of the tools and platforms we've discussed:

**Free or Open Source Options**

•

**DeepSeek R1:** This model is fully open source and MIT licensed, meaning it can be used commercially for free1. However, using it via API may incur costs2. The distilled versions are also available for free3. **DeepSeek R1 is significantly cheaper to use than other models**, such as OpenAI's o1, costing around half a dollar for 1 million input tokens and $2 for output tokens4....

•

**Bolt.DIY:** This is an open-source AI coding assistant1. While the tool itself is free, running AI models may still incur costs2.

•

**Langchain:** This is an open-source framework, so the framework itself is free, though the LLMs it uses will have costs6.

•

**Limbo:** This is an open-source database management system, making it a free option7....

•

**Sang Shan:** This is an open-source high-performance RISC-V processor project9.

•

**Nextcloud Server:** This is an open-source platform for data storage, offering a free option for users10.

•

**Realtime ST:** An open-source speech-to-text library that is free to use11.

•

**Bease:** This is an open-source low-code platform12.

•

**Moondream:** An open-source model for image understanding13.

•

**Cbook:** This is an open-source platform14.

•

**Ultra Box:** This is an open-source multimodal LLM15.

•

**Phala Network:** This is a blockchain platform where running nodes requires staking PHA tokens, but the platform itself appears to be open source16.

•

**Akash:** This is an open-source decentralized cloud computing platform17.

•

**Atri Labs:** This platform offers an open-source web framework for Python developers18.

**Subscription-Based Models and API Access**

•

**Cursor:** While it is a free VS Code extension, it appears to have a subscription-based model, according to the source19....

•

**AI Gradio**: The package itself is free but users are responsible for costs associated with the AI providers used21.

•

**Lecca IO:** This platform has flexible pricing options, including a free tier, but also caters to teams and enterprises, suggesting paid options22.

•

**RooCline:** This tool requires an API key linked to a billing account, suggesting a usage-based cost model2.

•

**Vapi:** The source indicates this has use cases for HR, suggesting a business-oriented model that likely includes costs23.

•

**Airtable AI:** Airtable itself is a paid product that has different tiers that include AI-enabled features24.

•

**11 Labs:** This AI tool for voice generation likely has a subscription or usage-based cost model15.

•

**Metal:** This is a fully managed service, indicating a cost based on usage and management needs25.

•

**Mage:** This AI tool for images and gifs likely has some kind of pricing structure based on usage26.

•

**Kaber AI:** This AI video tool for filmmaking probably has a subscription model27.

•

**Prompt metheus**: It is described as a prompt IDE, but the source does not provide a clear cost structure28....

•

**Relevance AI**: The source does not specify, but given its description, it's likely a SaaS with a subscription-based pricing31.

•

**Taskade AI:** This productivity tool likely has a subscription-based model26....

•

**Meet Geeks:** This AI meeting summarizer is likely a SaaS with a subscription26.

•

**Spline AI:** This 3D object and animation tool is likely a SaaS that has a paid component26.

•

**Blocks:** This AI productivity tool is likely a SaaS with a subscription26.

•

**Fireflies:** This is call recording software that has AI features, likely with a paid subscription model33.

•

**Chainlit:** While the framework itself may be open source, it likely has costs associated with hosting and the use of AI models34.

•

**Agenta:** The source indicates this tool allows you to build and sell AI agents, suggesting it has some form of paid component35....

•

**Type**: This AI-first document editor is likely a SaaS that charges a subscription fee24.

•

**Compose AI:** This text generator is likely a SaaS that has a paid component24.

•

**The 4ji:** This tool allows you to build and monetize AI apps, indicating a cost for monetization27.

**Usage-Based or Token-Based Pricing**

•

**DeepSeek R1:** While the model itself is open source, using it through an API or with specific tools may incur costs based on usage2....

•

**Claude 3.5 Sonnet:** This model is mentioned as being more expensive than DeepSeek R12.... The source also indicates that you can access Claude Sonnet for free through the Trey IDE37....

•

**Gemini 1.5 Pro:** You can access Gemini 1.5 Pro using a free API39.

•

**OpenAI Models:** Some sources mention the costs of using models like GPT-4o, which are based on input and output tokens4.

•

**Trey:** The source indicates this is currently free, but its long-term pricing model is uncertain20.

•

**Fireworks:** This platform provides access to R1 and gives users a $2 credit for free40.

**Other Cost Factors**

•

**Compute Costs:** Running AI models, especially large ones, can be computationally intensive, leading to costs for cloud services or hardware41....

•

**Data Storage:** Storing and managing data for AI applications can also incur costs10....

•

**Model Optimization:** The source mentions that optimized models can run on phones, which suggests that smaller, less computationally intensive models, may be more cost-effective than larger ones4.

•

**Development Time:** The time spent developing and iterating on AI applications can also add to costs44.

**Key Points**

•

**Open source does not always mean free**: While open-source tools are free to use, you may still incur costs for infrastructure, hosting, or API access to models.

•

**API usage is common:** Many AI tools and services use APIs that have usage-based costs, so the amount you use impacts how much it costs39.

•

**Free tiers may be limited:** Many SaaS options have free tiers but may limit the features or usage22.

•

**Pay-as-you-go is common**: Many platforms use a "pay as you go" model, where you are charged based on your actual usage4.

**Recommendations**

•

**Start with free options:** Explore open-source tools and platforms to experiment and prototype.

•

**Compare pricing:** Carefully compare the pricing models of different tools based on your needs and the scale of your projects.

•

**Consider usage patterns:** If you are expecting high usage, pay close attention to pricing for usage and tokens4.

Let me know if you have further questions or want a more detailed cost analysis of a specific tool or platform.

keep\_pinSave to note

copy\_all

thumb\_up

thumb\_down

which app dev scores the best with respect to ROI, efficiency, innovation, coding prowess, diversity of models, etc. Rank th5e top

Okay, let's analyze the AI app development tools and platforms from the sources, ranking them based on ROI, efficiency, innovation, coding prowess, and diversity of models. This is a complex comparison, as these factors don't always align perfectly. I'll provide a top 5 ranking along with explanations for each.

**Note**: ROI is difficult to assess directly from the sources because many factors affect the success of an app. Here, ROI refers to the potential cost savings, revenue generation, or efficiency gains.

**Top 5 AI App Development Tools/Platforms**

Here are the top 5 app dev tools and platforms based on your criteria, ranked from best to fifth best, based on information in the sources:

1.

**DeepSeek R1 (with Distillations)**

◦

**ROI and Efficiency:** DeepSeek R1 stands out because it's open source and has a very low API cost compared to competitors1.... Its distilled versions offer comparable performance at a smaller size, making them incredibly efficient for local deployment5.... The cost savings and performance make this a high ROI choice2.

◦

**Innovation:** DeepSeek R1 is a **pioneering open-source reasoning model that rivals closed-source giants like OpenAI's o1**1.... It uses reinforcement learning (RL) and distillation techniques to achieve top-tier performance6.... It's also innovative in its approach to training, using pure RL without supervised fine-tuning, and using a "cold start" approach with human priors12....

◦

**Coding Prowess:** DeepSeek R1 excels in coding, mathematics, and other reasoning tasks, often outperforming competitors such as Claude 3.5 Sonnet and even GPT-4o on certain benchmarks1.... It also has tools developed for it, such as DeepSeek Engineer and Bolt.DIY15....

◦

**Diversity of Models:** DeepSeek provides a range of distilled models, including ones based on Qwen and Llama, and sizes ranging from 1.5B to 70B6.... This provides flexibility for different compute capabilities and use cases.

2.

**Chainlit**

◦

**ROI and Efficiency**: Chainlit is focused on speed and simplicity, making it efficient to develop conversational AI applications24.

◦

**Innovation**: The platform focuses on creating sophisticated AI applications in a fraction of the time of traditional methods24. Its Python-centric approach allows for quick adoption for those familiar with Python24.

•

**Coding Prowess:** Chainlit is designed for developers who are building conversational AI applications24. \* **Diversity of Models**: The source indicates it's designed to work with various AI models, giving it flexibility24.

3.

**Lecca IO**

◦

**ROI and Efficiency**: Lecca IO's focus on no-code automation with a visual workflow builder provides rapid deployment for various integrations25. Its built-in RAG capabilities also save time on development25.

◦

**Innovation**: The platform’s no-code approach and use of AI agents make it accessible for non-coders while still providing flexibility25.

◦

**Coding Prowess**: While primarily no-code, the platform facilitates integrations with various tools and services, demonstrating a wide range of functionality25.

◦

**Diversity of Models**: Lecca IO interacts with various apps and services and simplifies the use of RAG, and so, implicitly, it supports various models through those channels25.

4.

**Cbook**

◦

**ROI and Efficiency**: Cbook aims to speed up app development by leveraging TypeScript and AI26. Its integrated AI app builder can reduce coding time significantly26.

◦

**Innovation**: It's innovative in combining a typescript notebook with an AI-powered app builder, catering specifically to developers26.

◦

**Coding Prowess**: While it is an app builder, it uses typescript, a popular language for developers26.

◦

**Diversity of Models**: The AI app builder is designed to work with a variety of AI models26.

5.

**AI Gradio**

◦

**ROI and Efficiency**: AI Gradio streamlines the creation of ML applications by providing a unified interface27.... It has comprehensive multi-provider support and a simple API, making the process quicker and more efficient27....

◦

**Innovation:** Its key innovation is providing a unified interface to connect to multiple AI providers27.

◦

**Coding Prowess:** While the tool simplifies building AI applications with only a few lines of code, the developers are still interacting directly with code28.

◦

**Diversity of Models**: AI Gradio has extensive multi-provider support including OpenAI, Google, and Anthropic, providing access to a range of models27....

**Other Notable Tools**

•

**Bolt.DIY:** While primarily a coding assistant, its ability to generate apps with prompting and no-code elements makes it a flexible tool15.... It's also open source, reducing costs15.

•

**Relevance AI:** This tool allows you to build and deploy AI agents with no code30.

•

**RooCline** This tool is designed for coding and integrates the powerful DeepSeek R1 model3.

•

**Cursor**: This AI enhanced code editor accelerates developer workflows31....

•

**Airtable AI:** This no-code database platform incorporates AI features, offering diverse use cases33.

•

**AI White Labels:** This tool allows you to build various no-code software apps with drag and drop34.

**Why These Rankings?**

•

**DeepSeek R1** tops the list because of its open-source nature, affordability, and exceptional reasoning capabilities, as well as its performance in coding, math, and other tasks. The distilled models give it a unique advantage for smaller deployments.

•

**Chainlit's** focus on rapid development and support for various models makes it a strong contender for high-performance AI apps.

•

**Lecca IO** stands out because of its innovative no-code approach combined with robust automation and RAG capabilities.

•

**Cbook** is a strong contender because it’s open source and allows developers to leverage TypeScript to speed up development.

•

**AI Gradio** is ranked highly for its simplicity, broad range of model support, and focus on rapid prototyping.

**Important Considerations**

•

**Specific Needs**: The best tool will always depend on your specific project requirements, budget, and technical skills.

•

**Emerging Tech**: The AI landscape is rapidly changing; new tools and platforms are constantly emerging.

•

**No-Code Limitations:** While no-code options are powerful, they may not be suitable for all complex or highly customized projects.

Let me know if you would like me to explore these or any of the other platforms further.

keep\_pinSave to note

copy\_all

thumb\_up

thumb\_down

how do I install and use deepseek distilled

Okay, let's dive into how to install and use the distilled versions of DeepSeek R1, drawing from the sources. These models are powerful, open-source, and cost-effective, making them an excellent choice for various AI tasks1.... Here's a breakdown of the process:

**1. Understanding DeepSeek R1 Distilled Models**

•

**Reasoning Capabilities:** DeepSeek R1 is a reasoning model, meaning it's designed to "think" through problems before providing a final answer. The distilled models retain these capabilities while being smaller and more efficient1....

•

**Open Source:** These models are open-source and MIT licensed, allowing for free commercial use4....

•

**Distillation Process**: DeepSeek-R1's reasoning abilities were transferred to smaller models, like those based on Qwen and Llama, through a process called distillation. This makes them more efficient to run2....

•

**Variety of Sizes**: You can find distilled models in various sizes, such as 1.5B, 7B, 8B, 14B, 32B, and 70B, offering flexibility for different hardware capabilities7....

**2. Installation Methods**

There are primarily two ways to install and use DeepSeek R1 distilled models:

•

**Local Installation with Ollama**

◦

**Ollama:** This is a free application that makes it easy to run large language models locally12....

◦

**Installation:** Go to the Ollama website (ama.com) and download it for your operating system12.

◦

**Running Models:** Once Ollama is installed, you can run a distilled DeepSeek R1 model by using the command line in your terminal: ollama run deepseek-r1 followed by the specific tag for the model you want (e.g. deepseek-r1:7b)12....

◦

**Model Selection**: When choosing the model size, smaller models like the 7B are faster and require less resources16.

◦

**Chat:** After the model is downloaded, you will enter a chat interface in the terminal, where you can start using the model14.

•

**Local Installation via Hugging Face**

◦

**Hugging Face**: DeepSeek R1 models are also available on Hugging Face17....

◦

**LM Studio and Llama CPP:** While Hugging Face provides access to the models, installing them for use with LM Studio and Llama CPP was not intuitive, according to the sources17.

◦

**Installation**: From the main R1 page in Hugging Face, you can download different distilled versions of the models18.

◦

**Ollama Integration**: The downloaded models from Hugging Face can also be used with Ollama18....

•

**API Access via DeepSeek or Open Router**

◦

**DeepSeek API:** To use the most powerful version of DeepSeek R1 (671 billion parameter model), you can access it via the DeepSeek API19....

▪

**Account Setup:** Create an account on the DeepSeek platform and top up your account. Even $2 may be enough to last weeks20....

▪

**API Key:** Generate a new API key on the platform22....

▪

**API Calls:** The API uses a specific link to the DeepSeek Reasoner, which is the R1 model24.

◦

**Open Router:** You can also access DeepSeek R1 through Open Router25....

▪

**API Key**: Create an account with Open Router and get an API key27.

▪

**Model Selection**: When using Open Router, you must select the DeepSeek R1 model from a list27....

**3. How to Use the Models**

•

**Command Line Interface**:

◦

When using Ollama, you can interact with the model directly in your terminal14....

◦

**DeepSeek Engineer**: The DeepSeek Engineer tool allows you to interact with DeepSeek R1 through a command-line interface, giving it instructions and reading local files25....

•

**Coding Assistants**

◦

**DeepSeek Engineer:** DeepSeek Engineer is a coding assistant built specifically for use with DeepSeek R1. It can read files, create new files, and edit existing files25....

◦

**Cline and RooCline**: Cline is an autonomous coding agent, which is also available as a fork called RooCline. RooCline is recommended for use with DeepSeek R1 as it handles the output format correctly, while standard Cline does not32.

•

**Bolt.DIY:** This open-source AI coding assistant supports multiple large language models including DeepSeek R126.... \* **Cursor**: This is an AI-enhanced code editor that can speed up coding workflows with the use of natural language34....

•

**Chat Interfaces:**

◦

**DeepSeek Chat:** DeepSeek's website provides a chat interface to interact with the model directly36.... When using this, make sure "Deep Think" is enabled to use R1; otherwise, it may use the V3 model36.

◦

**Agent Zero:** This is an AI agent framework that has been updated to support DeepSeek R1 and is recommended for chat use38....

•

**LM Studio** LM Studio supports many models through Hugging Face, and also has a chat interface17.

•

**Integration with other tools**

◦

**Make.com**: Make.com, a free automation platform, can integrate with the DeepSeek API for automations. This is done through the Open Router module27....

◦

**Cent**: Cent, an AI coding assistant, can be integrated with the DeepSeek API through the Open AI compatible setting41....

**4. Important Considerations**

•

**Prompting:** DeepSeek R1 is sensitive to prompts. It performs best with **zero-shot prompts** that directly describe the problem and specify the output format, as opposed to multi-turn conversations or few-shot prompting43....

•

**Tool Calling:** While DeepSeek R1 is excellent at reasoning, its function calling capabilities are not as developed as DeepSeek V3. However, it can be used as an architect to plan out complex tasks32....

•

**Language Mixing:** DeepSeek R1 is optimized for English and Chinese and may have language mixing issues with other languages49.

•

**Cost:** While the DeepSeek API is inexpensive compared to others, local models can be run for free.

**In Summary**

To install and use distilled DeepSeek R1 models:

1.

**Choose your method:** Decide whether you want to use the models locally with Ollama, through Hugging Face and other local platforms, or via the DeepSeek or Open Router API.

2.

**Install:** Download and install the necessary software or obtain the API key.

3.

**Run/Integrate:** Use the model through the command line, coding assistants, chat interfaces, or by integrating with other AI tools.

4.

**Prompt Carefully**: Remember that DeepSeek R1 performs best with zero-shot prompts that clearly describe the problem and specify the desired output format43.

By following these steps, you can leverage the power of DeepSeek R1 distilled models for various AI tasks effectively. Let me know if you'd like me to elaborate on any of these points!

keep\_pinSave to note

copy\_all

thumb\_up

thumb\_down

define the models for 1.5b, 7b, 8b, 14b, 32b, and 70b

Okay, let's define the distilled DeepSeek R1 models with parameters of 1.5B, 7B, 8B, 14B, 32B, and 70B based on the information in the sources. These models are all open-source and were created by distilling the knowledge and reasoning abilities of the larger DeepSeek R1 model into smaller, more manageable sizes1.... They are designed to bring powerful reasoning capabilities to a wider range of hardware and use cases2....

Here's a breakdown of each model:

•

**DeepSeek-R1-Distill-Qwen-1.5B**:

◦

This is the smallest of the distilled models, making it very efficient to run on devices with limited resources5....

◦

It's based on the **Qwen2.5** architecture7.

◦

Despite its small size, it demonstrates impressive performance, surpassing models like GPT-4o and Claude-3.5-Sonnet on math benchmarks, achieving 28.9% on AIME 2024 and 83.9% on MATH-5008....

◦

It serves as a good example of how distillation can produce a powerful reasoning model in a small package10.

•

**DeepSeek-R1-Distill-Qwen-7B**:

◦

This model, based on the **Qwen2.5** architecture, provides a balance between size and performance, making it suitable for many applications6....

◦

It achieves 55.5% on AIME 2024, surpassing the performance of the larger QwQ-32B-Preview model6....

◦

It represents the effectiveness of distilling the R1 reasoning capabilities into a relatively small model9.

•

**DeepSeek-R1-Distill-Llama-8B**:

◦

This model is based on the **Llama-3.1** architecture7.

◦

It achieves 50.4% on AIME 20249....

◦

It serves as a comparison to the Qwen series and shows the potential of using different base models for distillation6....

•

**DeepSeek-R1-Distill-Qwen-14B**:

◦

Built on the **Qwen2.5** architecture, this model has a larger parameter size than the previous models, offering increased performance.

◦

It achieves 69.7% on AIME 2024, and 93.9% on MATH-5006....

◦

It outperforms the state-of-the-art open-source QwQ-32B-Preview model by a large margin on reasoning benchmarks9....

◦

This model is a great option for those who need higher performance without needing to run the largest models6.

•

**DeepSeek-R1-Distill-Qwen-32B**: \* This model, built on the **Qwen2.5** architecture, offers even higher performance by using more parameters7.

◦

It achieves 72.6% on AIME 2024, 94.3% on MATH-500, and 57.2% on LiveCodeBench6.... \* It significantly surpasses the performance of previous open-source models and is comparable to OpenAI's o1-mini on many tasks6.... \* It highlights the potential of distilled models to match or even exceed the performance of much larger models9.

•

**DeepSeek-R1-Distill-Llama-70B**:

◦

This is the largest of the distilled models and is based on the **Llama-3.3-Instruct** architecture7.

◦

It achieves 70.0% on AIME 2024, 94.5% on MATH-500 and 65.2% on LiveCodeBench11.

◦

It sets a new record on reasoning benchmarks among dense models10.

◦

It is suitable for tasks requiring the highest levels of reasoning and performance, while still being much more efficient than running the full DeepSeek-R16.

**Key Points about All Distilled Models:**

•

**Distillation**: They were created by distilling the knowledge and reasoning capabilities of the larger DeepSeek-R1 model into smaller models, making them more efficient to use2....

•

**Reasoning**: The distilled models preserve the strong reasoning capabilities of the original R1 model, making them suitable for tasks such as mathematics, coding, and scientific reasoning2.

•

**Open Source**: All models are open-source, making them accessible to a wider audience1....

•

**Performance**: Even the smallest of these models outperform other models of similar size, showing the effectiveness of the distillation process6....

•

**Flexibility**: The variety of sizes allows you to choose a model that best fits your hardware and performance requirements2....

•

**No RL Stage**: These distilled models have not undergone Reinforcement Learning (RL), yet still achieve excellent performance, according to the sources7....

In summary, these distilled models offer a range of options to leverage DeepSeek-R1's powerful reasoning abilities. By selecting the right model for your specific needs, you can efficiently utilize this impressive technology.

keep\_pinSave to note

copy\_all

thumb\_up

thumb\_down

how do I run 7b and installing from ollama

Okay, let's walk through how to run the 7B distilled DeepSeek R1 model using Ollama, based on the information in the sources and our previous conversation.

**1. Prerequisites**

•

You must have **Ollama installed** on your system. If you haven't already, go to the Ollama website ([ama.com](https://www.google.com/url?sa=E&q=http%3A%2F%2Fama.com)) and download the appropriate version for your operating system1....

**2. Using the Command Line**

•

Open your terminal or command prompt3.

•

To run the **7B DeepSeek R1 model**, use the following command:4

•

This command tells Ollama to **download the 7B model** (if it's not already present) and then initiate a chat interface3....

**3. Understanding the Process**

•

**Model Download**: When you run the command, Ollama will **automatically download** the 7B model if it is not already present on your machine3. The download time will vary based on your internet speed and the model size.

•

**Local Execution**: Once the download is finished, the model will **run locally** on your machine, meaning your queries are processed on your computer without sending data to an external server3....

•

**Chat Interface**: After the model is loaded, you can begin interacting with it through the terminal. You can type in your prompts and receive responses3....

•

**No Special Configuration**: Using Ollama simplifies the process as it doesn't require any special configuration or setup. Ollama handles the specifics of running the model2....

**4. Key Features of the 7B Model**

•

**Based on Qwen**: This model is based on the **Qwen2.5** architecture7....

•

**Balance**: It is designed to balance size and performance. It is suitable for many AI tasks, as opposed to smaller models that might have limited capabilities or very large models that require powerful hardware.

•

**Reasoning**: The 7B model retains the **strong reasoning abilities** of the original DeepSeek R1 model, which it received during the distillation process7.

•

**Math Performance**: The 7B model performs very well on math benchmarks, achieving 55.5% on AIME 20248.

**5. Tips**

•

**Model Tag**: You must include the model tag deepseek-r1:7b when you run the command so Ollama knows which model to run4.

•

**Terminal**: The model is accessed via the command line so the terminal window needs to remain open while using it.

**In Summary**

To run the 7B distilled DeepSeek R1 model with Ollama:

1.

Make sure Ollama is installed on your machine.

2.

Open your terminal.

3.

Enter the command ollama run deepseek-r1:7b.

4.

Wait for the model to download and load.

5.

Begin interacting with the model in the terminal by entering your prompts.

By following these steps, you can effectively use the 7B distilled DeepSeek R1 model for your AI tasks locally. Please let me know if you have any other questions or need further assistance!

keep\_pinSave to note

copy\_all

thumb\_up

thumb\_down

how can I have it help me with a typescript and react project

Okay, let's explore how you can leverage the DeepSeek R1 models, particularly the 7B model we discussed, to assist you with your TypeScript and React project. Based on the provided sources, here’s how you can integrate these AI tools into your workflow:

**1. Code Generation and Completion**

•

**DeepSeek R1** models, especially when integrated with code editors, can assist you with code generation and completion. The models can generate code snippets based on your prompts or comments, helping you write TypeScript and React components faster1....

•

**Cursor**, an AI-enhanced code editor, can use natural language to speed up your coding process and workflow3.... You can use natural language to describe a component you want to build, and the editor will generate the code for it3.

•

**Trey**, another AI IDE, can automatically break down and execute tasks, including generating React components and styling5. It can even understand design requirements from images you upload5.

**2. Planning and Architecture**

•

**DeepSeek R1's reasoning capabilities** allow it to create detailed plans for your project. You can ask the model to outline the architecture of your React application, and it will generate the project structure and file organization for you6....

•

Tools like **Rline** (a fork of Klein) and **AER** can use DeepSeek R1 as an architect to design the architecture of your project, then use a faster model like DeepSeek V3 to implement the code8.... This means you can use R1 to generate project plans, define components, and describe relationships, and other models to write the code8....

•

**DeepSeek Engineer** is a coding assistant that can read local file contents, create new files, and apply edits to existing files in real-time. You can use it to plan and manage changes to your codebase, or ask it to build a full front end for you1....

**3. Debugging and Problem-Solving**

•

If you encounter a coding error or bug, DeepSeek R1 can analyze the error messages and suggest fixes. You can provide the model with the error output, and it can help identify the problem and generate the necessary code to correct it2.

•

**DeepSeek Engineer** can take your requests, use reasoning steps to generate responses and make changes to your codebase2....

**4. Testing and Refactoring**

•

AI tools can assist in creating unit tests for your React components13. You can provide the AI with your code, and it can generate the necessary tests to verify the functionality.

•

**DeepSeek Engineer** can implement requested changes and refactor your code, helping you improve code quality and maintainability2.

**5. Workflow Automation**

•

Tools like **Zapier** can connect your AI tools to your workflow, automating complex AI tasks and saving time14....

•

**Make.com** is a no-code automation platform, which can automate various tasks, including connecting to DeepSeek R116.

•

You can use **Relevance AI** to automate workflows with its intuitive interface, integrating it smoothly with your existing tech stack17.

**6. Specific Use Cases**

•

**Front-End Development:** DeepSeek models can generate HTML, CSS, and JavaScript for your React application6.... You can also ask it to create responsive and visually appealing designs.

•

**API Integration:** You can use the models to generate code for interacting with APIs, setting up data fetching, and managing state in your React application7....

•

**Chat Interfaces:** You can create a chat interface for interacting with AI agents by using DeepSeek models to generate the necessary code for managing conversations and history21.

**How to Integrate**

1.

**Choose Your Tools**: You can use DeepSeek R1 with code editors like **Cursor**3 or **Trey**23 that directly integrate AI, or you can use command-line tools like **DeepSeek Engineer**1... or **Rline**8.

2.

**Set Up Ollama**: To run the 7B model, follow the steps we previously discussed for installing and running it with Ollama: ollama run deepseek-r1:7b

3.

**Use Prompts Effectively**: When working with any model, use clear and specific prompts, and be prepared to refine the AI’s output as necessary24.

4.z

**Incorporate Workflows**: Consider creating workflows using tools like **Langgraph, n8n or make.com** to streamline the process of using multiple tools or steps in a complex process25....

**Important Considerations:**

•

**Reasoning and Logic**: DeepSeek R1 is particularly strong in logic and reasoning, making it well-suited for complex coding tasks19....

•

**JSON Output**: Be aware that the R1 models can be inconsistent with JSON outputs, so it's advisable to use other models for JSON generation when necessary30.

•

**API Keys**: Some tools and models require API keys that you'll need to manage31.

By integrating these tools and techniques, you can significantly enhance your productivity and accelerate your TypeScript and React development process.