[Music] hi welcome to another video when 01 was launched multiple models like deep seeks R1 and quen's qwq were also launched which started giving models like 01 a run for their money I mean no one wants to pay $200 for something that doesn't offer as much value today I have another such model which which is actually one of the smallest reasoning models yet and is actually quite cool and this is called small thinker small thinker is supposed to be a reasoning model but the best part is that it's only 3 billion parameters it is fine-tuned from quen 2.53 b as well which is a good small model the developers of this model have also shared some details about it including these benchmarks let me tell you about it but before we do that let me tell you about today's sponsor photogenius ai photogenius ai is an all-in-one AI powered art generator that allows you to type anything and get stunning visuals instantly photogenius AI gives you all kinds of image generation models in one place whether it be flux stable diffusion cardinski or any image generator model model that you can think of not just that it also gives you the option to do Advanced AI image editing as well with their cool AI tools like an AI Avatar generator background removal logo generator Emoji generator or even add an app icon generator and the best part is that it starts at only $10 and you can get an additional 25% off these already great deals by using my coupon code king2 so make sure that you check out photog genius. a through the link in the description and generate some cool stuff with it anyway if we come back to the video and talk about the benchmarks then it performs significantly better than the original 3B model and performs even better than GPT 40 in most cases it seems and majorly in stem stuff it performs pretty well this model is trained from a synthetic data set that was built with the responses of the qwq model which means that it's basically a smaller distilled version of qwq the data set that they trained it on is also openly available so you can use that as well now if we look at the intended use cases of this model then it is obviously intended to be used locally as a general model but one more thing about it is that it can also be used as a draft model for qwq which can increase your inference speeds for qwq if you don't know what a draft model is then it's a type of speculative decoding system that increases inference speeds speculative decoding with a draft model is a way that increases the inference speed of models by making a smaller model predict the next tokens then the bigger model sees that speculated bunch of tokens from the smaller model and just takes the streak of tokens that it feels are right and uses that and generates the next token after those tokens and then this process goes on it generally yields the same results as the original way of inference but a lot faster and deep seek V3 also comes with a multi-token predictor which also uses the same type of speculative decoding anyway this model can be used as a draft model for qwq and if you use that model locally then you can have about a two times higher inference speed the model is available on a llama so you can use it from there quite easily VM and llama CPP support speculative decoding with a draft model as well and I'll tell you how you can use it as a draft model with the qwq Q model as well but first let's test it out individually and let's see how well it works so these are the 13 questions that I'm going to try it against the first question is tell me the name of a country whose name ends with Leah give me the capital city of that country as well the answer could be something like Australia and canara let's send it and once we send it you can see that it's starts its Chain of Thought and actually it's very similar to The qw Chain of Thought and feels like just a smaller version of qwq anyway if we wait a bit then we have the answer as well which is fully correct so let's keep this a pass the next one is what is the number that rhymes with the word we use to describe a tall plant the answer should be three because it rhymes with tree once we send it you can see that it again goes through the chain of thought and if we wait a bit then it's now done and the answer is also correct so this is also a pass the next one is write a Haiku where the second letter of each word when put together spells simple let's send it and see okay it's now done and this is not correct although it tried and started to get a bit close but this is a fail the next question is name an English adjective of Latin origin that begins and ends with the same letter has 11 letters in total and for which all vowels in the word are ordered alphabetically let's send it and see okay it's now done and this is the answer which is also not correct so this is a fail the next one is Courtney said that there were 48 people but Kelly said that Courtney had overstated the number by 20% if Kelly was right how many people were there let's send it and see okay here's the answer and this is correct so this is a pass the next question is I have two apples then I buy two more I bake a pie with two of the apples after eating half of the pie how many apples do I have left the answer should be two let's send it and see okay here's the answer and this is also correct let's keep this a pass now the next one is Sally is a girl she has three brothers each of her brothers has the same two sisters how many sisters does Sally have the answer should be one let's send it and see okay here's the answer and this is also correct let's keep it a pass the next one is if a regular hexagon has a short diagonal of 64 what is its long diagonal the answer should be 73.9 let's send it and see okay here's the answer and this is also correct it's the first time this small siiz model could pass this which is pretty great let's keep this a pass now after this there are the coding questions and it doesn't perform well in coding tasks at all which is expected because the new questions are pretty complex even for bigger types of models so it failed all the questions similar to other same-size models I won't bother you with the pain of making it generate code either because it was very much not optimized for that so this is the final chart to be honest this model is pretty great on most of the stuff that it's aimed at considering the size I mean in general and STEM related prompts it's too good for the size and is really amazing I mean there's no model that could pass the diagonal test or even the apple pie test because it's just that the knowledge gets limited so this is amazing plus because it's so small you can easily use it on a new age computer like a bass M1 Mac or any computer that has about 8 gigs of RAM which is pretty great so for the size it's amazing now one of the major features about it is that it can be used as a draft model for qwq so if you want to do that then it's quite simple to do with VM you'll just need to get VM installed by running pip install vll M and then just run the model like this in the base model we have the qwq model and then in the speculative model we have the small thinker and in the speculative tokens we have five which means how much further should the tokens be speculated five tokens are generally handy so once you do that and run it you'll see that the VM server gets started on a port which is open AI compatible and you can easily Now integrate it into anything and the results are much faster than general qwq which is pretty great to see so that's also good if you run qwq locally I think that this model has great potential and you can also fine-tune it on your specific task and it can perform really well overall it's pretty cool anyway share your thoughts below and subscribe to the channel you can also donate via super thanks option or join the channel as well and get some perks I'll see you in the next video bye [Music] [Music] hi welcome to another video when 01 was launched multiple models like deep seeks R1 and quen's qwq were also launched which started giving models like 01 a run for their money I mean no one wants to pay $200 for something that doesn't offer as much value today I have another such model which which is actually one of the smallest reasoning models yet and is actually quite cool and this is called small thinker small thinker is supposed to be a reasoning model but the best part is that it's only 3 billion parameters it is fine-tuned from quen 2.53 b as well which is a good small model the developers of this model have also shared some details about it including these benchmarks let me tell you about it but before we do that let me tell you about today's sponsor photogenius ai photogenius ai is an all-in-one AI powered art generator that allows you to type anything and get stunning visuals instantly photogenius AI gives you all kinds of image generation models in one place whether it be flux stable diffusion cardinski or any image generator model model that you can think of not just that it also gives you the option to do Advanced AI image editing as well with their cool AI tools like an AI Avatar generator background removal logo generator Emoji generator or even add an app icon generator and the best part is that it starts at only $10 and you can get an additional 25% off these already great deals by using my coupon code king2 so make sure that you check out photog genius. a through the link in the description and generate some cool stuff with it anyway if we come back to the video and talk about the benchmarks then it performs significantly better than the original 3B model and performs even better than GPT 40 in most cases it seems and majorly in stem stuff it performs pretty well this model is trained from a synthetic data set that was built with the responses of the qwq model which means that it's basically a smaller distilled version of qwq the data set that they trained it on is also openly available so you can use that as well now if we look at the intended use cases of this model then it is obviously intended to be used locally as a general model but one more thing about it is that it can also be used as a draft model for qwq which can increase your inference speeds for qwq if you don't know what a draft model is then it's a type of speculative decoding system that increases inference speeds speculative decoding with a draft model is a way that increases the inference speed of models by making a smaller model predict the next tokens then the bigger model sees that speculated bunch of tokens from the smaller model and just takes the streak of tokens that it feels are right and uses that and generates the next token after those tokens and then this process goes on it generally yields the same results as the original way of inference but a lot faster and deep seek V3 also comes with a multi-token predictor which also uses the same type of speculative decoding anyway this model can be used as a draft model for qwq and if you use that model locally then you can have about a two times higher inference speed the model is available on a llama so you can use it from there quite easily VM and llama CPP support speculative decoding with a draft model as well and I'll tell you how you can use it as a draft model with the qwq Q model as well but first let's test it out individually and let's see how well it works so these are the 13 questions that I'm going to try it against the first question is tell me the name of a country whose name ends with Leah give me the capital city of that country as well the answer could be something like Australia and canara let's send it and once we send it you can see that it's starts its Chain of Thought and actually it's very similar to The qw Chain of Thought and feels like just a smaller version of qwq anyway if we wait a bit then we have the answer as well which is fully correct so let's keep this a pass the next one is what is the number that rhymes with the word we use to describe a tall plant the answer should be three because it rhymes with tree once we send it you can see that it again goes through the chain of thought and if we wait a bit then it's now done and the answer is also correct so this is also a pass the next one is write a Haiku where the second letter of each word when put together spells simple let's send it and see okay it's now done and this is not correct although it tried and started to get a bit close but this is a fail the next question is name an English adjective of Latin origin that begins and ends with the same letter has 11 letters in total and for which all vowels in the word are ordered alphabetically let's send it and see okay it's now done and this is the answer which is also not correct so this is a fail the next one is Courtney said that there were 48 people but Kelly said that Courtney had overstated the number by 20% if Kelly was right how many people were there let's send it and see okay here's the answer and this is correct so this is a pass the next question is I have two apples then I buy two more I bake a pie with two of the apples after eating half of the pie how many apples do I have left the answer should be two let's send it and see okay here's the answer and this is also correct let's keep this a pass now the next one is Sally is a girl she has three brothers each of her brothers has the same two sisters how many sisters does Sally have the answer should be one let's send it and see okay here's the answer and this is also correct let's keep it a pass the next one is if a regular hexagon has a short diagonal of 64 what is its long diagonal the answer should be 73.9 let's send it and see okay here's the answer and this is also correct it's the first time this small siiz model could pass this which is pretty great let's keep this a pass now after this there are the coding questions and it doesn't perform well in coding tasks at all which is expected because the new questions are pretty complex even for bigger types of models so it failed all the questions similar to other same-size models I won't bother you with the pain of making it generate code either because it was very much not optimized for that so this is the final chart to be honest this model is pretty great on most of the stuff that it's aimed at considering the size I mean in general and STEM related prompts it's too good for the size and is really amazing I mean there's no model that could pass the diagonal test or even the apple pie test because it's just that the knowledge gets limited so this is amazing plus because it's so small you can easily use it on a new age computer like a bass M1 Mac or any computer that has about 8 gigs of RAM which is pretty great so for the size it's amazing now one of the major features about it is that it can be used as a draft model for qwq so if you want to do that then it's quite simple to do with VM you'll just need to get VM installed by running pip install vll M and then just run the model like this in the base model we have the qwq model and then in the speculative model we have the small thinker and in the speculative tokens we have five which means how much further should the tokens be speculated five tokens are generally handy so once you do that and run it you'll see that the VM server gets started on a port which is open AI compatible and you can easily Now integrate it into anything and the results are much faster than general qwq which is pretty great to see so that's also good if you run qwq locally I think that this model has great potential and you can also fine-tune it on your specific task and it can perform really well overall it's pretty cool anyway share your thoughts below and subscribe to the channel you can also donate via super thanks option or join the channel as well and get some perks I'll see you in the next video bye [Music]