

Why Specialized AI Models Are Challenging the Frontier Labs — With DeepL CEO Jarek Kutylowski

35 MIN · YOUTUBE · [HTTPS://WWW.YOUTUBE.COM/WATCH?V=16UIISJV4NU](https://www.youtube.com/watch?v=16UIISJV4NU)
<https://www.youtube.com/watch?v=16UIISJV4NU>

SUMMARY

Yaric Coutilowski, CEO of Deepell, discusses the growing importance of specialized AI models in contrast to larger, generalized models. While big models have dominated the conversation due to their versatility, specialized models are proving to be more efficient and effective for specific tasks, particularly in language translation and other niche applications.

- Specialized models are rising as they can outperform larger models in quality, speed, latency, and cost for specific tasks.*
- Generalized models may sacrifice accuracy and consistency when handling diverse tasks, while specialized models maintain higher performance for targeted applications.*
- The trend of model routing is emerging, where businesses use a combination of models to optimize performance and reduce costs.*
- Deepell focuses on language translation, helping companies operate in foreign markets by providing accurate and context-aware translations.*
- The advancement of AI has significantly improved the quality of translations, enabling businesses to communicate effectively across language barriers.*
- Voice AI is seen as the next frontier, with potential for real-time translations in conversations, though adoption has been slow.*
- The future of AI is expected to bring transformative changes to society, enhancing human capabilities and fostering global communication.*

Big AI models have gotten a lot of attention, but specialized models are on the rise. Let's talk about how these purpose-built models are challenging the status quo with Deepel's CEO, Yaric Coutilowski, who joins us today in a conversation brought to you by Deepell. Yaric, welcome. >> It's a pleasure being here with you, Alex. Thank you. >> Thanks for being here. Okay, so Deepell is obviously building specialized models. We'll get into it in a moment, but I first want to go through the bigger question about specialized models versus bigger models. You know, the conversation recently has all been about these big foundational models. Bigger is better. all these things, you know, the more compute and the more data you add, the better you're going to get, results you're going to get on a generalized path. but there is a notion now that smaller specialized models are starting to challenge the bigger models, outperform them in some way, be the models that are routed to when a model router decides what to go with, whether it's, you know, bigger or more specialized. So, talk a little bit about why we're seeing this rise in specialized models and where you see this going.

>> Yeah. Yeah, sure. I mean like the fact that generalized models those those big models that they are getting a lot of attention is obviously because they can do a plethora of different things at the same time and for very very many of those tasks that we that we do every day day there there is not enough incentives there it's not really a business model to create something that is going to be specialized we just we just do so many different things every day and and I said like those those generalized models cover those really very very well. but then there's areas in which specialized models can actually outperform those big models and not only outperform on kind of the quality or on the on the performance on this on this particular task but also in this kind of triangle of performance, quality latency, speed and and price. And if you think a little bit about the real applications of models, then then all of those three parameters matter quite a lot. Especially if you're if you're pushing a lot of data through those those models. If you're if you're burning a lot of tokens on a particular use case, if your company is translating millions and millions of documents or translating hours and hours of real time audio because this is kind of how your customer service maybe operates, then it really makes sense to employ those models that can actually be better at this task and also do that that more more more cost effectively. and and then therefore to a large extent in in businesses in in real life applications you're going to see kind of the teams that are setting up the infrastructure really rely also on those on those specialized models.

>> Okay. Okay, but let me ask you a question because this entire generative AI era kicked off in part because of the transformer model and that was built specifically to translate language and what Deepel does is a lot of what you do is language translation, right? You'll help businesses, let's say a an American business that wants to operate in Brazil be able to basically translate their website and their customer service into Brazilian Portuguese and be able to function there. But and you've built these specialized models to do it. But let me put the question to you. If the entire thing, the entire generative AI moment is built on a model innovation that was meant to translate language, then then why would we need something specialized to do that as opposed to the bigger models with that foundational, you know, innovation baked in?

>> Yeah. I mean like you're you're right the transformer model and language translation I think they're kind of very very very tightly coupled but I think also very early at the at the beginning it was clear that the transformer models can can do more but when they do more when they're made for for different purposes they also lose a little bit of the capability that they had maybe initially when they've been made for translation only this this set of parameters that is available there. This which which kind of determines quite often the capa capacity of the model.

it needs to be divided into very many different things and and therefore if you're keeping the model very much strictly to do one particular task however it's defined in this case language translation it can perform better and it can perform on that also more consistently. I think something that you see with generalized models is that depending on on which kind of input you give to them, they're going to tend to be better or worse. And specialized models have a kind of better layer of consistency. they do they are they're quite often much better as in as in ours that the quality assurance is is really built so that we can make sure that whether it's an email that you're translating or whether it's marketing material or a technical patent application or in all of those cases the language quality is as good as as good as possible. and then also large language models, those that are meant for generalized usage, they're usually trained mainly on English and maybe a couple of languages.

And if you really want to go broad, if you really want to go multilingual, you need models, you need the data and the capacity in those models really to be able to cover all of all of those. >> Yeah, it's so it's interesting. So this large the larger large language models they have that transformer capability baked in right some of the innovations built off of the attention is all you need paper but when you query them for language you're also querying a model that's going to be able to advise you on fitness and science and medicine and and then it sort of when it comes to a specific use case can be so sprawling that as it tries to boil the ocean it might end up getting you know worse performance as you go towards specific areas and that's why there is this movement towards the more specialized models. Am I getting that right?

>> Yeah. Yeah. Specifically as like the the the super important second step of training models. Right now it is not only about giving them access to a large amount of training data. This was kind of where the transformable model has started on just seeing a lot of training data and then being able to replicate the task. the the reinforcement learning aspect that comes in very very heavily nowadays to show those models what we actually want them to do in a in a in more detail. that one covers or like enables them really to solve those real life tasks that you just mentioned like the the like prepare a diet for me or like figure out a d a recipe. And if you run this reinforcement learning step on too many different tasks, the model will be able to do all of that.

But once again, it's going to be very very very broad. And if you focus on making sure that the model understands and knows that it needs to provide the best possible translation for a given text, for a given document, it's just going to be able to do that that in a in a in a better way. >> Yeah. So to go with the big one, you basically you sacrifice you I I'm actually guessing here, but from your your answers, it sounds like you sacrifice some accuracy. You might sacrifice latency as you as opposed to going to maybe a smaller specialized model where you can be sure that you'll end up getting being more accurate and being less latent.

>> Yeah. Yeah. Accuracy. Exactly. accuracy is the is the one part that kind of comes from the specialization of the model. the latency, the speed is just like a very very natural function of the size of the model and the ability of it to just be processed faster on the same kind of kind of hardware and especially in in in some applications that are more real time and speech live speech translation is a big part of that latency is key to the user experience.

It's sometimes it's not even that much about the accuracy. It is quite often also really this combination of accuracy and speed in which we can deliver the translation to make up for for a for for a good experience for for everyone. >> Yeah. If you're waiting for a live translation and you're like sitting there with somebody and the model is you know let me think for a couple minutes. It's going to really ruin the conversation. So >> no you can't have Yeah. you can't have any kind of engaging discussion and and like if you think about use cases where where people actually are using real-time translation right now like in negotiations and like business conversations like there the the the the back and forth of these discussion is extremely important right and so okay so you're building your specialized models in language but there are also specialized models in areas like law that are starting to come out Harvey for instance recently announced that they're going to build their own models.

So do you see this as something that like now that we've seen the capabilities of big large language models that the next step is going to be in almost every discipline we're going to start to see these type of you know specialized models being built out. >> Yeah. I don't know if this going to be really like every discipline out there. I think there needs to be a balance like is the use case big enough? Is the business model for a company big enough to invest into this? Training those models is still expensive, it's complicated, you need to have the capability of doing that. so I think it's going to be limited to to those most important ones, but those most important ones definitely are going to to to run into this direction to not only just make sure that the accuracy and the the the speed are what they are, but also to cover like all of the edge cases that are that are out there that need to be covered for not only a demo and proof of concept, but an application that just kind of runs each and every day in a business.

Yeah. And so is there any trade-off then? And all right. So let's say I opt to use a more specialized model for a purpose. Is there any trade-off that I should be aware of, you know, as opposed to going with the bigger model or is it all upside? >> I think it's actually all upside. I mean, other than potentially having another vendor, that's kind of the the the kind of the the question around how do we do model routing within a company? How do we pick vendors? How do we pick suppliers for the different areas of of a business? And the convenience of just working with one big model vendor is obviously there. But I think at the point in time when you start thinking about how to do a particular task really well and and it and it is part of what your business is maybe doing at its core and it's important to your business then then specialized models really make a lot of sense I think you know >> Yar you brought up something that I'm hearing a lot about and our audience is hearing a lot about and it would be great if you could shed some light on which is model routing. So, I think this is more of a recent thing that's come up where companies and people using these tools have seen that maybe asking the biggest model for everything is expensive and not necessarily the best way to get things done. And so now this idea of model routing is in vogue where when you ask a question there is a routing layer that will say oh you know this question would be better for a flash model. this question would be better for a a high-powered model that's going to think a lot or this question would be better for a specialized model like the one we're talking about. So, can you talk a little bit about why we've seen model routing start to become such a popular term in the AI world and where do you think this is going?

Yeah, I think I think the the very very first idea or like the very very first reason why this has been coming up recently so much is the sticker shock on for for many businesses kind of looking at the expenses that they that are being generated by by the usage of of big high-powered models and and I think everybody wants to rationalize how how we're using those models actually in language translation this idea has been coming up already couple of years earlier and we've seen those model routers already in place in in businesses where it was a question like which models excel at given language pairs which offer the best capabilities for a particular type of content maybe and those have been already employed like as said I said earlier I think right now more and more of like in in in in in this in this recent trend, we're going to probably see also also more of this coming up when it's not going to be coming down to cost only, but also to the performance of those models and and making sure that a certain task is really handled in the best way.

>> Yeah. So then where do you see the AI world going? Is it that there's going to be like effectively a router? these big foundational these big high-powered foundational models will handle some tasks but the majority will

go to either smaller or these like flash models or specialized models or is there is it still mostly going to be the big models handle stuff and every now and again things will be routed elsewhere. Yeah, I think I think I think right now we are at a place where a lot of the tasks that we're seeing as common within the business are really well tackled by by the large models and that is primarily coding and software engineering. I think a lot of the kind of workloads that are being run by those models right now is actually going into into coding and therefore as those are the most performant ones for for that particular use case I think this is going to stay there right now but we're seeing more and more of those specialized now going into other directions and if there's going to be new models coming up that are very specialized for co for coding or when the tradeoff between price and and and capability is going to start to shift, we might see more of those use cases actually going going or like being routed away. Yeah. And so when we talk about whether AI is this AI moment can be sustainable, you know, I think a lot of times it's like, well, how many people are going to pay for the 20 or \$200 chat or claude subscriptions? And I always say that, you know, ultimately this technology is going to live or die by its ability to make real change and create business opportunities in the real world. And so just to go back to the deep use case again, that's what you're doing, right? It's helping a com a company that might function well in like let's say I'm just going to use English, an English speaking country and say, "Oh, you want to you want to do your operations elsewhere?" we can use LLMs to actually enable you to set up shop in a country that speaks an entirely different language and make it seamless. So, can you talk a little bit about how you do that?

>> Yeah, I mean like that there there there's a lot that goes into that and and the way that we think about our how we're helping our customers, it is it is on the one hand internal and on the other hand external. Like if you if you think about a big international organization that has offices all around the world, the the language question always comes up and we might kind of assume that everybody speaks English, but like honestly that's usually not really the case. No, they don't. And if they think they do, maybe they don't really do that well. and and and there is studies that show how much actually gets lost in the context and how much productivity suffers from that and how often like an employee within a meeting just doesn't speak up u because they feel maybe not confident enough and the in in English or the language that is being spoken and just like this great idea gets gets lost. so so we help companies make sure that internally they can forget about all of those all of those all of those boundaries and it also helps them to hire the best people doesn't matter where they're sitting like whether they're in France and speak only French and and can contribute to a company that is English speaking that makes a big difference that that allows them for getting the best talent on board at a moment where maybe this talent is not so easy to find. And then when it comes to the external appearance, in order to be able to serve customers in a in a in a new market and pretty much each and every company tries to be a global company nowadays, it was necessary to find people who are speaking that language both in maybe sales and customer service and preparing the technical documentation and and being able to translate this technical documentation.

Well, and that that means a lot of upfront investment cost maybe even at the moment when we don't know whether the market is going to be performing well for for us. We we don't know some of the specificities there and I think right now it is much easier to just start doing that. U bring in an AI tool to translate all of the documentation that is out there. bring in DPL as an AI tool to make sure that salespeople can communicate with the with the with the new market.

make sure that customer service is just transparently translated in the background whether that's over chat, text,

email or or or voice with with with like like a with a phone call or or something like that. so and that really makes a difference for for for those companies and enables them to open up those markets much easier and then in some cases it is really about situations in which there was already a process which allowed the company to operate in a different market but it's just clunky.

It's just based on a lot of human translation happening in the background which has very long lead times. I've we we have customers who have optimized their whole product delivery cycle but the localization aspect the translation aspect at the at the end of this product development was always the reason why things could not be shipped to the market and now with bringing AI into this they basically have an instant solution to that problem.

>> Interesting. And how much has the growth and of capabilities of LLMs enabled you to do this job? Like and talk talk a little bit about how we've seen better LLMs and what they've enable deep to do in terms of like going from a point A to where you are today. >> Oh. Oh, totally. I think there's there's this like this big stack of of use cases and the complexity and how how hard they really are to to solve with with AI. And I think it maybe starts somewhere at the bottom with like sending out spam emails like you really don't have you you really don't need to have like the best translation for for that. anything just kind of basically works like we know how those emails look and at the at the top range of that is probably regulated documentation that needs to be compliant where like there's legal liabilities behind behind all of that. Maybe think like a think like a leaflet that is that is being distributed with a with with medicine things like that. And as the quality of models has been rising over the last years, we've been able to unlock more and more and more of those of those use cases. And honestly, this is always something that is really and truly complicated for our customers to find out is the model quality good enough for doing this in particular job. And this is also then our responsibility as DBEL to come in and help our customers find out. what is the quality level that they require? What is the error rate that they're seeing on those? What is what is the reasons for this error rate?

Maybe optimizing the whole setup there and at the end bringing them a solution which solves as big of a chunk of all of those use cases that that that they have. So where were you like a couple years ago compared to where you are today in terms of being able to attack those use cases? >> No, it's it's night and day I would say. I mean like with in general I would say with neural networks with AI and and we've we've launched with Debbell in 2017. That was like basically the point in time where where where neural networks have been coming up. That was the major step in AI translation.

before that all of the solutions really didn't work. I think that's just general consensus. I think that 2017 step was was a milestone. but after that things got better and better and better and like right now if you look at the quality of those solutions if you if you take a well-written document and if you look at a language pair u that is like one of the major language pairs like let's say English to French or English to Spanish or u some some some of the some of the big big languages out there the translation is going to be pretty much flawless sometimes it's not going to be perfect Not because the model is not perfect, but because it exposes problems in the source document. Like it it kind of actually shows it's been written in an ambiguous way. It was written in a way that is actually hard to being under understood. So sometimes like we actually catch the the mistakes that that that

have been made during writing a document.

>> I think I need to use it for my writing to make sure I'm being clear. It actually it actually really helps because if you're looking at something that you have been writing like your brain is really skewed towards accepting that because you have just written it and and you know it and then I and then I translate it into into let's say I've written something in German and I translate it into English and and I see the translation is clunky and and then I notice okay that there there's actually there's a problem problem in there. The models can figure out a lot of that and they can help you improve the text to some extent, but there's a level of ambiguity that they might not be able to resolve because they don't know you and they don't know what you actually wanted to write.

>> Yeah, exactly. Like you can look at something that you've written and take a look at it 10 15 times and still not catch that typo. but then you phase shift a little bit and you're like, "Oh, okay. Actually, that idea made no sense. That was spelled poorly. I use the wrong there or something like that. And I know when you when you translate into German, there's no mercy, right? That language. >> Yeah. Yeah. There's there's there's no mercy. It's a complicated one. Yeah.

>> For poor grammar. I'm in the middle of learning German. So, congratulations. >> I've learned this the hard way. Thank you. Well, not no congrats yet. so let me let me ask you that the the So, language is one thing, right? Let's say I'm a US company. I want to operate in Brazil. I can probably get my company to function, my website to function, maybe some of my customer service in Brazilian Portuguese. But then there's also laws, regulations, customs. So, does language get you half of the way or how far, you know, can this take you when you're trying to operate somewhere else?

>> I I think it gets you pretty far. And I think it gets you also pretty far because you can start leveraging local partners in at this point in time already. like you can you can engage that Brazilian law firm that is going to help you in some aspects or that are that are local and you can engage them in a in a in a good way and and maybe they speak English so then kind of this is easy but maybe they're not and in those situations you can you can already start far quicker there's going to be definitely things that you're going to have to set up in your new market it's not only language there's also as other aspects but you can you can get there I think and and hey even like we're doing in our in our local markets for example in Asia we we are talking to journalists we are talking to our customers and all of those interactions are being translated by our by our technology and our AI.

So, like if I would be doing this podcast with a with a with a in in a in a Japanese market, we would be totally doing that with DL running the language layer in the background. >> Oh, okay. So, that that sort of brings me to that thing I've been thinking about as we've been talking because we've talked about customer service and we've talked about communicating across borders and latency and and I wonder, you know, how much of AI do you think is going to move towards voice from where it is today, which is which is text? I was recently speaking with Greg Brockman, the OpenAI president at our summit. and I think the point where he got most excited was talking about like new bidirectional voice models.

where you'll be able to instead of like go turn by turn have a conversation like we're having now where input

and output are somewhat seamless. And to me it seems like the the path forward in AI includes a lot of a lot of voice and and where's ground zero for that? It's going to be in areas like translation where you know let's say you know you were speaking German and I was speaking English we'd be able to have a conversation and the AI would just translate for us right away which would be the coolest thing and I know there's like some attempt to bake this into glasses already by companies like like Google and others. So so where do you think the voice side of AI is going especially when it comes to your use case?

>> Yeah, I think when it comes to translation and languages like this is really the next frontier. I think we really have gone so far in text and documents by now. there's like a lot is possible. >> I think we're still in the early years when it comes to real time speech when it comes to having those conversations as engaging as quick as fast and and like really making sure that that the quality is at as at the right level. so so there is a lot that can be done and also on the other hand there's really a lot of conversations that we probably are not having right now because this this barrier is there. So I'm really looking forward to a world where where those conversations are going to be to be happening. When it comes to voice in like in a in a more broader sense when I thinking about AI and voice and seeing voice as an input interface I'm I'm kind of split twofold I have to say.

like we we like those models are really great those conversational models and if you start talking with them in situation like you're driving a car or you're just cooking something and and and you can't type it seems magic. it seems like something totally different comparing to what we've had until now. but then on the other hand like there's been so many attempts at establishing voice input into computers, phones for the last maybe decade like we've had voice assistants coming up but like it never really caught on. There's something about us humans maybe not wanting to really talk so much to to to to technology.

just preferring writing because it's I don't know like we have a little bit more time to to to just think while we're writing. I don't know exactly where like what this is coming down to. but I think the adoption of voice AI until now hasn't been as fast as the technology would have actually enabled enabled that because those conversational models are out there and they are already pretty good but yet we're not using them so much.

>> Yeah. Yeah. So maybe it's just the the form of use issue as opposed to a technology issue. >> Yeah. Yeah. And I don't like my wife is saying like these models are creeping her out a little bit when he's when she's starting talking talking to them. I'm I'm fine with that, but like she doesn't like that that that that an input model somehow. >> Yeah. where do you stand on the AI device then? Do you think that that will be successful like an AI wearable or whatever OpenAI is brewing?

I I think it makes a lot of sense. I think kind of getting devices as small as possible and kind of as as near to us as as possible. Specifically also in the case of of language translation. that that makes a lot of sense. I'm I'm a big advocate of the fact that for like realtime translation we actually have all of the devices that we need like the airports that we have like the phone that we have that actually suffices for for that. but having more data and having devices embedded with us all of the time and in in a situation like that can also gather a lot of data about us and and therefore be more more context aware. I think that that that is pretty cool.

>> But can't the phone play that role? >> Yeah, to an extent yes. I think and I think if you really need the kind of compute capabilities and the connectivity capabilities that you will need for for for serious AI like even even the the best phones that we have right now are are can can only run extremely small models at this point in time. so so we're going to need a big device for that. But when it comes to like gathering data and having the context of what's happening around us, the the phone might be just like a bit too restricted physically because it's sitting in our pocket. so having glasses, having like sensors in our body that might actually help in and determining what is happening in the physical world around us. And quite often when models do not get things right is because they just do not understand what is just happening here.

like what is what is the real life word word that they cannot see >> because they're limited to the prompt that they just got. >> Totally. All right. I I want to end with this. You know, we're we're definitely in a moment where it's very interesting. AI the capabilities are increasing. a lot of people are becoming like AI pill or having I guess they call it AI psychosis where they're just like they're at a point where they just can't live without AI. But then we're also in a society where there is growing uneasiness around the technology.

where people are worried about data centers, you know, being built near them or whether AI is going to, you know, change their jobs, take their jobs, alter reality as they see it today. And you know, I'm curious to hear your perspective on where this is all heading. And if there's a world where AI actually is, you know, flips that narrative among the skeptics, you know, what might do it? >> Yeah. I mean like I think it is clear that we are at the beginning of a huge huge huge transformation in how our world looks like and and we had a couple of those technological revolutions already in the past. I I think that this one might be probably the biggest one that we had until now because intelligence is just so core to everything that we are as humans and having another form of intelligence at this world is just like a completely new thing for us and therefore a lot's going to change and we're talking about a lot about jobs but I think it's not going to really boil down to jobs but it's also going to boil down to capabilities that we have as as humanity like we we are while SpaceX is being able to launch rockets into space.

They're not that we're not yet able as humanity to to build like big spaceships that would be able to carry like a larger number of people somewhere. And the engineering hurdles that come with that are probably just too big for us to be able to solve them as humanity alone. So, we're going to need some assistance with that. And I think AI can be really the the solution to that. but at the same time I I I see how these big changes can be unnerving to to to a lot of people like there's moments when when I feel like this progress is so fast that I that I get also intimidated by that too. And and I think at this situation I think like really thinking also about language translation this is just an amazing use case of AI. If you think about the fact that whole civilizations have been built on our ability to communicate with each other, exchange ideas, move the world forward, resolve issues that might lead to conflicts. having the ability of every person on this world talk to another person on this world. I think that's a that's a really beautiful application of of AI and and I think that makes it really worth not only from a business perspective but also from a human perspective to build this company at least for me.

>> Agreed definitely and I'll tell you there are certainly people in my life who I would love to be able to have seamless trans AI translation you know as we speak and you're right there are communications that get lost and

ideas get that get lost without it. So it is one of those positive visions that I think we could all believe in. Yark, look, great speaking with you. I learned a lot from you today. Thank you for coming on the show, >> Alex. Thank you for having me. It's been pleasure.

>> Definitely. All right. And if folks want to learn more about Deepell, where do they go? >> They go to dell.com. That's that's that's super easy. >> There you go. All right. Thank you, Yark. Thank you, everybody, for watching and we'll see you next time here on the channel.