A1  
 Okay, so... Right, perfect. And then I would... Well, before we start, maybe just quickly: in this whole study there are no wrong or right answers. Even if it might sometimes sound a bit like an exam question, we are really just interested in an open conversation. So simply saying whatever comes to mind is the best thing you can do for us. Right, then I would like to ask you to briefly introduce yourself.

I10  
 Yes, I will make it a bit longer, then maybe some of the first questions can already be answered. Yes, XXXXX XXXXX XXXXXX. Professionally, I am working, or rather, I am active in two fields. What does two fields mean? Well, professionally I am doing two things. On the one hand, I am doing software development in consulting. That is the job that pays for my food and my apartment. And then there is the job that I actually want to do, but which does not finance itself. I am developing an interactive, computer-based vocabulary trainer for sign language. Or to put it a bit more simply and less cryptically, so that even non-computer scientists can understand: I am trying to develop a computer program with which one can teach the language of those people who unfortunately cannot hear or speak to people who can. That was originally my XXXXXXXXXXXXXXXXXXXXX. So 13 years ago I developed a concept as part of XXXXXX XXXXXXX, where the point was that when learning signs, you actually get feedback on whether it is correct or not. And as part of this XXXXXXXXXXXXXXXXXXXX the topic was examined. Since we came from a completely different field and originally only wanted to use sign recognition as an example application for something else, we realized that sign recognition or sign translation does not work at all yet. So it was broken down in the context of XXXXXXXXXXXXX to single sign recognition, because that was and still is the technical state of the art. And in order to create an application case, we said, okay, we need a vocabulary trainer. And when we continued working on this, we realized, okay, such a trainer does not even exist. And that is how this project started. I have now been working on it for 13 years and in May I presented a prototype of it at the XXXXXXXXX. I am currently taking part in XXXXXXXXXX XXXX. And through this work I also got into the XXXXXXXXXXX, this XXXXXXXXXXXXXXXXX XXX XXXXXXXX. And that is how I eventually came here.

A1  
 Right, very cool. Definitely a super exciting project. So again a project where I can really get enthusiastic about what computer science can be used for and what kinds of new technologies it can enable. Because I think things like this are super important and of course in capitalism they sometimes get lost, because it only concerns relatively few people, let’s say, and therefore does not generate much money.

I10  
 Actually, I would say, well, what we would actually need is a paradigm shift. So in capitalism you cannot make money with it, that is true, at least for the person developing the program. At least not much. Because basically in capitalism there are different players, for example the state. And if we think about it, in Germany we have 180,000 deaf and hard of hearing people who depend on sign language. If none of us can use sign language, these are all people who cannot do anything. Who just sit around somewhere staring at the ceiling. And that already costs us a lot as a society.

A1  
 That is true, yes, absolutely.

I10  
 In Bangladesh we are even talking about 1.2 million out of 160 million. So they are letting 1.2 million units of human capital rot in a corner, to put it very bluntly. So accordingly, I would not even make it about capitalism, I would rather make it about society, that we as a society or politically simply do not have the will. Although in Bangladesh it even looks better than in Germany. There is no will here either. So...

A1  
 Yes, exactly, but in any case there is human potential, and with technology you can try to build bridges, and that is a very beautiful aspect of computer science. That is why I think it is a very cool project. Right, regarding the XXXXXXX XXXX, we also got in touch with them through the XXX, and then they shared it in their distribution list and so on. For us it is very exciting that we are getting interview partners from all over. We are all somehow based at XX XXXXXXXX, or in my case and in the case of XXXXXX at the XXXXXXXXXXXXXX XXX XXXXXXXXXXXX XXXXXX XXX XXXXXXXXXX XXXXXXXXXXX. That is primarily a federally funded research project, I should add, I left that out earlier in the introduction. So basic research is our daily business, but of course we also do industry transfer, XXXXXXXXXXXX XXX XXX XXXXXXXXXXXXXXXXXXXXXX XXX XXXXXX XXX XXXXXX XXXXXXX XXXXX XXX XXXX XXXXX XXXXXXXX XXX XXXXXXXX XXXXX XXXXXXXX XXX XXX XXXXXXXXXX. But now through the XXXXXXX XXXX we have again gotten in touch with some people we did not have a connection to before. That is really nice. Right, and with that we are already at the topic. So we are a research institute for machine learning and artificial intelligence. To what extent does this play a role in your everyday processes and your work as a software developer, and also in the development of this prototype for sign language and an interactive system for sign language, to what extent do AI and machine learning play a role there?

I10  
 Yes, a very big one, because it is actually simply one of the most important elements in this project. Let me elaborate a bit so that it might be easier to understand. Nowadays, there are basically two scenarios in which people learn sign language. The first one is that you have someone around you who knows sign language. That means you tell them, show me “bus”. And then they stand there and sign “bus”. And then the deaf person sits there and says, yes, that is correct. The second option, which is the one mainly used and which is highly dangerous, is that video CDs are distributed and people stand in front of their mirrors. That means people stand in front of their mirror, watch the video, and copy it. For many things or for some signs that is probably quite practical. However, there are signs that look very similar but have a completely different meaning. Since this here is being recorded like a video, I will not demonstrate it now. But for example, if you take the sign for “football” and rotate your hand 90 degrees, you are asking something very obscene. And once I even went up to a deaf person and asked if we should watch football in the evening and I actually asked him something quite different without meaning to. Well, he knew me. He understood what I actually wanted. He then briefly showed me how to do it correctly. But that was already highly embarrassing and in another context it definitely should not have happened. And what we have in the project is that we add an AI which really checks the correctness of the signs. Is what he is signing really the sign it is supposed to be? So yes, a very big one.

A1  
 Just to set the stage a little, in this conversation we will probably mainly talk about this sign language recognition, translation, AI prototype in the context of software development, meaning your main activity. Is AI and machine learning more in the background or not present at all or...

I10  
 Well, there are projects where there are indeed contact points. But those are A considerably smaller and less complex. That is about predictions or cases where I do not write them myself but mainly take care of the labeling beforehand, meaning this data generation beforehand. So this is definitely the project where I have the most to do with AI.

A1  
 Okay, but still within the context of software development, so we can definitely include that. I just wanted to understand whether it is really sort of an engine only in this prototype or whether it is also present in other cases. I mean, we are now at the point where every second developer already uses GitHub Copilot anyway and then of course you also have AI in software development, but you are not really working with it yourself.

I10  
 I did not define that as AI development at all. I think there should be a difference. AI development for me means, I know there are many people who do something else mainly for marketing reasons. But for me AI development really means I develop a model and I train something with it. And everything else is not, for me, that I develop AI. I may use it when I use ChatGPT, but everything else like this, yes, we also do a lot with artificial intelligence, but actually if you only use pre-trained models, for me that is application and not development.

A1  
 Yes, absolutely. One has to be very clear about those differences. That is also something we noticed in this interview study. As machine learners of course we also train our own models and so on. But I also spent a lot of time evaluating pre-trained models because that is also something you can do scientifically. We are now at a time where there are very, very complex models that you simply cannot train locally. For example, language models. Not every company even has the data or the infrastructure to just do that. So right now the market is transforming a bit in the direction that intelligent systems are also being purchased on a large scale to open up internal business opportunities. Which is also fine. And they also have an opinion on AI and they also have an interface or a form of interaction with AI. So of course we also want to talk to them. But yes, one really has to distinguish clearly. So in any case it is cool that you also develop AIs yourself. That is good. Exactly. What are the biggest hurdles in these everyday processes? Where do you see the biggest problems? What costs you the most time in your work with AI or the development of AI?

I10  
 The big problem with the project is that you do not really have a dataset. So what I developed is a supervised AI and for that, of course, you need a dataset. And the biggest problem is simply that this dataset does not exist and sign language is not even international. That means we also have local differences, so there are about 400 languages in the world and about 200 sign languages. Maybe only 75 or 80 of them are really relevant, because if you take Flemish-Belgian, for example, there are 5000 people who use it. So I think you can really ignore that. There is simply no dataset. And this dataset is, of course, a foundation for what we are doing. But it is also a foundation for basic research, which you, XXXXXXX, might be doing in the area of sign language recognition. How are you supposed to do basic research in sign language recognition, or how is anyone else in the world supposed to, if there is no dataset? And that is kind of the discussion, or rather what has been taking up most of the time here for years, that you try to convince people that such a dataset could simply also be the foundation for technology-supported inclusion of these people. And the thing is, you have to deal with people who, A, are already overwhelmed by the word vocabulary trainer, who then keep writing to me about a translation tool. And since they already use that as their basis, they do not understand at all or are not even able to understand why such a dataset would be so important. And the second point is simply that especially in sign recognition there are a lot or actually very many research papers within the scope of bachelor’s and master’s theses. But there are very few real documentations to build upon. That means you often feel like you are already on a kind of basic research level, because there simply are not really many people who deal with this. And with the datasets that do exist... I mean, if we take the ABC, then here we have an A, then we have an E, and then we have an I. For us these are clearly recognizable differences, if I show them like this, whether it is like this or like that. But on the other hand, they are just two fingers being used starting from A. That is, of course, not the easiest thing for an artificial intelligence. And all these example projects that have been done are mainly with something like 25 sample signs, which are then very different. Sometimes the right hand, sometimes both hands, sometimes up, sometimes down. That of course makes everything… yes, that means you do not really have a foundation to build upon.

A1  
 Makes total sense. I think that is always the problem when you try to use AI and machine learning in an area that is also a bit about a marginalized group, because then statistically there is automatically less data available in general. So this goes a little away from our actual conversation, our core discussion, but still I am curious, is there not, like, very often now, I am thinking of the Olympics for example, and all these TV broadcasts and live interpreters who also gesture in sign language, can you not make a connection there and also gather data? Or, well, of course the problem is then probably that there are maybe five people across Germany who do this live interpreting, which means you only have very few different people who are speaking sign language in that moment and for which you then have translations, but can you get access to such data?

I10  
 I would bring in another example, especially in the German context, that would be XXXXXXX XXXXXXXXXXX XXXXXXX XXXXXXXXXX, which is usually interpreted by XXXXXXXX. Incredibly nice people. So if you ever need sign language interpreters, XXXXXXXX XXX XXXX, I can highly recommend them. Also very relaxed in their way of dealing with people. Really nice. I think there are about twelve of them. So you would already have a certain selection of people. The problem is that, on the one hand, an interpreter is never simultaneous. That means even if I just put subtitles underneath and take the data, it is still offset. There is, well, we are also working with the XX XXXXXXXXXXXX, where a student group is dealing with the subject. What they are simply trying in their study program is, nowadays if you do speech recognition, you no longer train the individual phonemes, but you give the system the whole sentence and tell it what it is. And they are currently trying to achieve that based on these data. Whether there will be a result or not, I cannot say at all. I simply do not know yet. But for us the problem is that the signs are not clearly separated. So you would probably need even more data and sort them out in a much more elaborate way. The second problem is signs. If I say a word, then I say a word and in between there is a pause. Usually shorter, but you can also make it longer. And with signs it is different. If I spell ABC for example, it practically flows into each other. That means it makes it even harder to have clearly defined training data. When I go from one word to another, the word is no longer clearly signed. For the deaf person it is, of course, completely understandable and that is also correct to do it that way. But due to this blending of the words it is no longer clearly separated. And the fundamental problem something like XXXXXXXXXX has is that basic words are simply missing, because depending on the topic, I think the words climate protection and climate change could be trained endlessly, because they are simply used very often in the XXXXXXXXXX. Ukraine, Russia, Putin, also no problem. But I think something like bus is already the first problem, because I do not know when that... well, okay, maybe there has been a bus accident, but otherwise the word bus is simply not used. And we are also trying to orient ourselves a little on the European language reference levels, on what we teach. And then it becomes really, really difficult. So sleep, bus, these are all words that would somehow be missing.

A1  
 Yes, of course, I mean, obviously, when you have data from somewhere, you always have the context, that is what we always call it somehow at our institute. So data is one thing, but all data comes out of a certain application context, and of course there is always potentially some bias in it, where for example with language certain words are underrepresented. And then you sometimes notice that again with language models, that they are just like, they are trained on the internet, but not everything is discussed on the internet. Well, kind of, but there are things that show up more often, others that show up less often, and because of that the language models then develop, or also have, the habit of choosing certain phrasings a lot, which then sound like classic internet-speak. And then, for example, research papers generated by language models get filtered out a little bit because of that. There are certain phrasings that are almost cliché for papers, that show up very often, and words like intricate for example, which hardly anyone uses in normal speech, but which are something you can very nicely put into a paper. And yes, there is always bias in it.

I10  
 I find the sentence length with these things annoyingly short. I am someone who absolutely loves making run-on sentences, like half a page, font size 10 printed. One sentence is totally my thing. I also have no problem reading that. And these language models always make it more like how a ten-year-old would write.

A1  
 Yes, because again, that is also how people often write on the internet, short sentences, and right now we also kind of have a paradigm shift. I recently read that the XXXXX XXXXXXXXXX XXXXXXXXXXX or something like that published that news articles in the future should ideally not have more than 500 characters. Simply because people’s attention span is dropping so badly due to social media, and for example the average watch length on YouTube videos used to be one minute forty, and in the meantime it went down to only 40 seconds until most people drop off again. So that is changing society as a whole right now, right?

I10  
 Well, that we are not exactly the brightest species, I have had that suspicion more than once.

A1  
 Yes, well, okay, but labeling is definitely, well, getting labeled data at that point is a big problem. That makes complete sense. Apart from the fact that of course we would somehow have to put together a dataset, is there still anything you can think of where you would say, that is where AI researchers, machine learning researchers just need to get better, those would be steps that would greatly support the application or the development of artificial intelligence in practice.

I10  
 Well, actually I can just continue with the sentence I just made, I do not feel like our species is the brightest. That is probably less of a computer science job and more one for psychologists and so on. I think we need to catch up a lot in terms of education, especially among the people who make decisions. I think that is a big issue. I know that here in Germany, well I am also used to other countries since I work in XXX XXXXXXXXXXXX and therefore in a border region. I have good contacts to XXXXXXXX XXXXXXXXXXXX XXX XXXXXXXXXXX. I know that here in this country it is particularly tricky to say, because here there was also the quote, why should you need to read and write for the Bundestag. But fundamentally I do think there are just huge gaps in education concerning this AI topic. And I think it has kind of peaked for now in this new EU AI Act. I assume you also read it. I think it was Paragraph 64, 1 and 2, where the glorious thing was written that we must always give all training data to the country… So that in the country where we develop it, we must give the authorities access to the training data. That does not sound so bad at first, when you think, okay, hey, we have something like medical products. And now we are just supposed to give an authority full access to training data. From my point of view that is extremely tricky, because yes, they probably will not have real names in them. But depending on what we are talking about, it is still possible to recognize who it is. And I also wonder how someone even comes up with the idea of writing that into a law, because for me that basically has nothing at all to do with data protection anymore. And for me it also has very little to do with protecting artificial intelligence at companies or with company value and company property. I also do not really feel that the German state is particularly well prepared when it comes to security and similar issues. So I see that very, very critically. But I mean, we are talking about people who are celebrating a law and I believe do not have the slightest idea what they just wrote in there or just signed off. And that is the point for me where the most has to be done, that above all we first teach and then maybe on a societal level explain, okay, what is an artificial intelligence? What can an artificial intelligence do? What can it not do? What do you need to pay attention to? Like, when I write a model and then generate the first file, that I have a 50 megabyte file that can do nothing. I can fill it with 100 cats, with tulips and carnations. It is the same, but I have to fill it in the first place. I think those are steps that have not sunk in. And I know, I think it was seven or eight years ago, when some researchers from America built an AI and filled it with passport photos of homosexual and heterosexual people. And the shitstorm that followed, like oh God, they built a racist AI. No, whether I throw in dogs and cats or them, it does not matter at all. I mean, people should have just looked at the result and maybe thought about, okay, what could we actually do with this insight? Could we maybe apply it to other things? I come from Catholic youth, we heard more about pedophilia than anyone ever wants to hear in their life. One could have used this result for research like that. I mean, it was just an interesting result at first, but no, instead we just go with, they made a racist AI. And then the whole approach, that maybe could have helped other people who have problems that we do not want in society, to put it as neutrally as possible, we just shredded because we are too stupid. That is what I think...

A1  
 Yes, I have to briefly put in a small word in defense of politics, because of course they do invest in basic research. Otherwise I would not be here right now. That means they do make sure to put some money into training experts in the field. But I totally agree with you that the communication gaps are sometimes extreme and we regularly have meetings with individual politicians who visit us as funders, and you really notice how wild it is that this person is sitting on an expert decision-making committee that decides what happens with funds for AI research. And you realize, okay, this person’s background is so far away from the technical level. In politics, the people who sit there, well politicians are not necessarily trained for the things they decide on. And this putting together of expert committees to advise them and form advisory bodies is definitely something that needs improvement. I fully agree with that.

I10  
 I also think they should earn much more, because right now it is just the idiots being washed into politics. The ones who have no chance anywhere else, who never have the chance to make big money, they go into politics because it is their last chance, because it is the maximum they can still get out of life. I mean, in the field of artificial intelligence I do not know people who have huge money. But I know it from the medical field. There is a woman who has a company in the greater XXXXXXXXX area, she can create such wonderful concepts that help reduce deaths from hospital germs and similar things. And she also says, why should I earn in one month what I can earn in two days, and on top of that make myself the laughingstock of people.

A1  
 Yes, that is also part of it. It is definitely not an overly attractive job. It is also not that easy to get into politics anyway. And I also think there are quite a lot of people who are overpaid. But I think that drifts away from the topic. But yes, education is definitely a big point, I fully agree with that. Pedagogy in that area, the knowledge. We just had another interview yesterday where we talked about this. But that is not something that we as AI researchers can really do well at the core. I am not a pedagogue, I am primarily a computer scientist. And of course I think about science communication and so on, but I am not trained for that. And that is where I see pedagogy and the people who shape the school system as being responsible. Something is happening there. Can you think of anything else where you would say that we as computer scientists could build a better bridge?

I10  
 The only thing that comes to my mind spontaneously, but I must admit I have not had much time to think about it because I am more the kind of person who accepts the circumstances and tries to work with them instead of constantly thinking “oh God the system is so bad.” The only thing I have thought about several times myself, though I do not know if it already exists, because the pressure was not that high yet. When I develop these models and especially the tests, and I mean I do hundreds of them every day while I work, what would be cool for me would be if there was something that already evaluates the data a little for me. Like this loss and basically machine-based, it does not even have to be intelligent, maybe some intelligence should be in there. But like this loss and accuracy data, where I just throw it in and it gives me a factor, like this is how good it is. So that you can do things like… right now for example I am at my grandmother’s, when I am away for a week, I could give it the learning rates, the gamma values, let it run for the week and then tell me at the end what the best result was. But I do not know if that exists or if there are actually methods to calculate that. I mean, sure, when you have those beautiful curves, it is relatively easy to program. But what I mostly have is this zigzag stuff because I just have far too little data. So I often just look at it for a minute as a human and see, well, this was better, that was better. That would be pretty cool for my work, it would just make things easier.

A1  
 Okay, cool. That actually goes in a direction we are also working on. So what you are basically looking for is an AutoML solution. There is a lot of research and technology in AutoML, and it works well when you have a lot of data or very well-curated datasets. But when you are in highly specialized cases, with very different data that also bring very domain-specific problems with them, like weird loss curves or something, AutoML suffers. Because AutoML always assumes, well in machine learning in general you need a lot of data for it to work well. The more data you have, the less human interaction you need. So the less domain expertise you need when developing the model. But when you are in a very specialized case, you need a developer who really knows the data well. From my work I always see this when we collaborate with companies and develop very specialized solutions. Data and business understanding is 80 or 90 percent of the work. Only the rest is machine learning. First you have to understand and curate the datasets, understand how they are embedded in the company context, that is where the effort goes. And at the end you just throw 20 SK Learn models or neural networks or whatever on it, test them all, look at the results and choose the best one. But to make that a little easier we have been working on it, and that is basically the core reason why we are doing this study, because we are asking ourselves: does it work and does it have practical relevance. For that I would share my screen. And I would like to ask you to give a spontaneous reaction directly.

I10  
 a spontaneous reaction welcome to Germany. basically I am hungry now because this ABCDE reminds me of the Nutri-Score

A1  
 Is that positive or negative

I10  
 Since I have not had breakfast yet, right now negative. May I ask questions?

A1  
 Yes, absolutely, please. That is exactly why we are sitting here.

I10  
 So we are talking about image recognition I guess because it says ImageNet above. And powerdraw per inference is then the milliwatt value when I recognize an image?

A1  
 Exactly, well not a single image, but a batch of images, because in image recognition you usually run the data in batches, since that is much more energy efficient on GPUs. I think in this case it was a batch size of 32, so 32 images are loaded onto the graphics card, classified with the model and that then consumes 600 milliwatt seconds, which is about 0.6 watt seconds.

I10  
 Corrupted Robustness is, I suppose, the recognition rate.

A1  
 Yes, something like that. That is something we basically had in almost every interview, but of course it is a very good learning for us. So the Top One Accuracy at the bottom left is practically the classification quality over the test data. With ImageNet we have a thousand classes and I think in the test datasets there are like ten or a hundred images for each class. That makes 50,000 images, so 50 per class, and 63% of them are classified correctly. Simply as right or wrong classification, not like Top Five meaning is it among the best five, but just is it classified correctly, 60%. With Corrupted Robustness these images are additionally overlaid with filters, for example with a kind of snow filter or noise or a perspective distortion like a camera zoom, to check whether the prediction remains constant or robust when the input is perturbed, just to test the generalization of the model a bit. And you can see that the accuracy of course gets worse.

I10  
 And Running time per Inference is probably the run through of one like that

A1  
 Exactly, so just like above with powerdraw, it takes 1.3 seconds to classify and consumes 0.6 watt seconds.

I10  
 That is not really fast though, is it?

A1  
 1.3 seconds for 32 images is

I10  
 Well I would have expected more from the graphics card. That is more like a CPU value I would have expected.

A1  
 Yes, I have to say I would also not bet my hand on the values below being 1000% correct.

I10  
 That is probably made up anyway.

A1  
 No, they are not made up, this was actually measured somewhere at some point, but it has been one or two years since I did that project. In that project we measured a lot of different models for ImageNet, also on different hardware. In this case it is on an A100 node, so on a cluster machine with A100 GPUs in it. But I think it is normalized per inference, so I think it actually is

I10  
 A single image.

A1  
 No, a batch of images, 32 images. So it could be, I mean, 32 images in 1.3 seconds I do not find extremely slow.

I10  
 Yes, but I still would have expected a lower value if you have such an A100 construction. Well, I also do not know how large the images are, right?

A1  
 Yes, with MobileNet those should be 280 by 280 pixels. But then you also have factors like whether the preprocessing is included or not, usually parts of it are also executed on the GPU. So, as I said, in any case this was...

I10  
 We do not want the model to speak, but the label.

A1  
 Exactly, so we do not need to get lost too much in technical depth here. First of all, it is basically about the label.

I10  
 There are two QR codes. If I scan the upper one, what do I get?

A1  
 That one leads to the software framework and the paper in which I introduce this concept, and theoretically you can also generate your own labels there if you upload your own benchmarks, so you can somehow generate labels accordingly. The lower one leads to the paper in which MobileNet is introduced. So practically the paper for this specific model.

I10  
 Which factors are included in the A to E?

A1  
 The lower factors and additionally a few more. So we measured a few additional metrics, such as the number of parameters of the model, size when you write the model to disk, also a Top-5-Accuracy, which is a slightly different evaluation metric than Top-1. That one is always a bit higher, so the question is whether the correct class is among the top 5 predictions. That means we evaluated a few more metrics and we summarized and weighted all metrics to derive an overall scoring. And in doing so we distinguished between metrics that relate to the quality of the model and those that relate to the resource consumption of the model. And in this case we balanced it out so that both are equally weighted for the overall evaluation. And that is why you can also see at the bottom that the Top-1-Accuracy is red, which means it is relatively poor compared to other models.

I10  
 Oh those are colors? I am color blind, so for me it all looked the same, only the Corrupted Robustness, that was... By the way, I am not the only one who is color blind. That is 9.8 percent of men in Germany. So maybe one should reconsider whether it is really a good idea to make this in red and green. Especially at that size it is not really... well actually... I am technically color weak and not color blind. That means I can usually recognize it on the computer. And at the top with A to E I also saw that there were different colors, right and left. But for the Top One you really have to tell me. If it is so close, it gets really hard. That is 9.8 percent of men and 1.2 percent of women in Germany alone.

A1  
 At this point one also has to say that this is of course nothing that is market ready yet. A usability study like that would definitely be useful. It is super important feedback for sure. I mean, with energy labels everything is also color coded somehow, but the information is never only that. There it also says A, B, C, D, E everywhere, just like at the scale above. One could easily add that at the bottom as well, just put a little letter next to it in the right color.

I10  
 So for me no, I also did not even know that these things had different colors.

A1  
 Yes, but there you additionally also have, like at the scale above, you can see that what is highlighted is also larger. So at the top scale you can tell that the right one is A and the left one is C, because the size is different and because the letter is shown next to it. And that way you also get an order. With energy labels it is like that too, there it is a vertical scale from A to, I am not sure, E I think as well. And then on the right you have like a cursor that shows which row you are on in the scale. But yes, that is super important feedback for sure.

I10  
 Now that there are two next to each other I do at least notice with power draw per inference that they are different.

A1  
 Exactly, so in the end the color scale at the bottom is the same as the one at the top, from green to red. Inspired of course by energy labels and Nutri-Score somehow, since those are already familiar systems. But you could easily add a small letter next to the badges at the bottom to show, also for people who are color blind, what the rating is here.

I10  
 So I am just supposed to say whatever comes to my mind.

A1  
 Exactly. And also with regard to whether this could in some way be helpful in your everyday processes and your work, or something where you would say, ah, that would actually be useful if such a system were established.

I10  
 I think, the way it stands, it would be extremely tricky to establish this, because we already talked about the topic of education, especially for people in decision-making positions. What my personal feeling about this graphic is, and this is also the reason why I do not pay any attention at all to the Nutri-Score at Rewe, is this summarizing. Like this overall A, this overall C, I am really not a fan of that. I do not think it is bad that values are listed at the bottom, like this Corrupted Robustness and Top-1 Accuracy and also this Running Time per Inference. Capturing these three values in terms of labeling would not even be bad, I think. But whether they need to be given letters, I do not know. I mean, this Power Draw per Inference is, for me, more something for page 2. Because I really think, for me the three most important values for a model are simply this Top-1, Corrupted Robustness and the Running Time. I think these are actually the three values. I know this is now a bit against your studies. What does against your studies mean? Not against, but rather outside of your studies, just from my own perspective. Those are the three values. It would already be practical, if you have such models, when you are on Kaggle, you would always have these three values. That way you could really compare them. Like, okay, that would help to compare these models. For page 2, or page 2 in a figurative sense, something like Power Draw per Inference would definitely be important. Once you have selected three models, you also start comparing them. But for me the really important values, and the most important value of a neural net for me, is always this Top-1 Accuracy. Followed by Corrupted Robustness. And then, for applicability, also this Running Time per Inference. I also think it is a bit unfortunate that here for ImageNet it is calculated in batch. I mean, of course you can do that. But I think it should be calculated on the value individually. Simply because of what points more towards application cases. Like if I now think of XXXX XXXXXXXXXXXXXX, a friend of mine from XXXXXXXXXX, who wants to detect cancer. I mean, she cannot wait until 32 doctors have uploaded an image and then we calculate it. Instead, the other value is, I think, more interesting. But personally, what I do not like, that is my personal opinion, is that there is this overall A and this overall C. Because I do not know, A is probably considered better than C, I guess. And I really do not think so when I look at this thing. For me the left one is much, much better than the right one. And if I then imagine, there is some idiot in the Bundestag who has to decide something, he will definitely take the right one, because it has an A on it. And then we end up with some crap with 63% recognition rate, even though it could have been 81. Just because someone who can count to three and has already reached the end of the road in their head with machine learning, sees an A.

A1  
 Well, one has to say, this is also something, if you read the paper I wrote about it and which is also supported in my software framework, these are super application-specific questions. There are application cases where you have a fixed limit for inference time, because you know, we have to push through so and so many gigabytes of data per day. That means we need a response time of under one second, for example. Not in the case of this concrete image classification ImageNet case, but more generally, when you talk to people, what are your priorities, what matters to you in the end, we often have the case that people say, you know, 10% accuracy more or less is not really that important. What matters is that the response time is high, because in the end, every third decision is checked by a human anyway. On the other hand, we now have something like 20 years of research in machine learning and artificial intelligence, where people are very motivated to report that they have the new best model. And then you see, ah, they are really 3% better on this or that dataset than all other models. And then you test it locally and find out, yeah, but they also consume twice as much energy. You have to question critically then, is this really an improvement of the state of the art or are we just throwing more resources at the problem? And especially with regard to climate change, that is highly critical.

I10  
 On the other hand, if we bring in again the little Chinese girl from XXXXXXXXXX with her cancer detection, wow, those 3% could matter, and since this is also about cervical cancer, I think XXXXX would definitely find it great to have those 3% when she uses it. That could already decide her life, right?

A1  
 Absolutely. And that is why we also do not say we must not use resources, but rather we must be resource-conscious. And if in an application case accuracy and reliability of the system have the highest priority, then that is the way it is. Then that is also fair. And then you could also simply say, all metrics except accuracy and maybe robustness or so are completely irrelevant. Or you just say accuracy, and then there are also, especially for such cases, technologies to estimate such a prediction confidence. Like, with such and such probability, this really is cervical cancer or not. And then you again have other metrics on top of that. So the question of which metrics are relevant and how relevant they are, that is absolutely a use-case-specific question. And in this whole framework you can of course also adjust the weighting of the metrics below. And say, in this case, based on my own conviction, and in ImageNet we do not have this case. Nobody does cancer detection based on ImageNet. Because ImageNet as a dataset does not provide that. There are no scans or anything like that in it. It is just random pictures of the Internet. So in this case I said, okay, we weight all metrics equally and look for a fair comparison between quality and resource consumption. But of course you always have to adapt this specifically in application cases. I am a thousand percent with you on that. But at the moment a lot of people only report accuracy and let everything else fall under the table. And I find that questionable. That cannot be the idea of research, that we only look at quality.

I10  
 That is also true for research, I mean, research should also be open to results. And in research I agree with you. But I think that in the free market, the models that are currently being developed, in 90 percent of the cases accuracy is the value.

A1  
 Yes, but it then gets pretty interesting when you, I mean, we also had this in interviews, and you are basically totally right that most people say, the main thing is that the thing works. But then there are also many people in management who backtrack when they find out, oh wow, okay, so this is only a little bit better. As a human you hardly notice the difference in the language model, it is supposed to be a little better, but as a human you do not notice it. But it costs us three times as much to use the model, because OpenAI charges correspondingly high fees for it, which in turn is mapped in the background to electricity consumption. So the bigger the model, maybe also the more competent the model, the more costs it also causes through electricity consumption and cooling of the hardware and so on. And that of course directly translates into license costs. And when people choose language models, many are completely overwhelmed by the options. You go on Hugging Face, there are like 50 different models, which one should we use? And our hope is that making these decision processes a bit more transparent, maybe such labeling could help with that.

I10  
 But what I find with this overall value, especially when you now also say that one can weight it differently, then I find that thing up there, meaning this overall letter rating, even more problematic. So at the very least it should also be stated on the first page which weights each of these four values has. So that this is at least clear, because I mean, I can then influence idiots in their decisions simply by changing the weighting.

A1  
 Yes, but people do that anyway. I mean, if now, imagine someone develops AIs for the management level, presents that in some kind of steering meeting, says, look, we tried out ten models, these are the properties, we would recommend taking this or that or decide freely, but you can always sugarcoat it in a way that fits how you want it and let your own priorities or convictions flow into it. But I agree with you...

I10  
 That is the right way a boss, I will not call it manager now but boss, has an employee, gives him a task, he gets a result back, and if the employee says this is the solution, then the boss, that is why he is also called managing director, he does not run the place because he decides or knows everything, he is not God, he is only the managing director. But if I now go more in the direction of politics, well I really believe them capable, with the kind of education they have, of deciding during operations that things with E may no longer be used. I do believe our politics could come up with such stupid decisions. And yes, if then TÜV says okay we weight everything equally and then we have a model that consumes an unbelievable amount of energy but unfortunately recognizes cervical cancer with 99% accuracy, then suddenly it is forbidden because someone came up with something. So I think labeling makes sense, but then please label all four and not just an overall one.

A1  
 Yes, that can definitely be assumed. That is also super important feedback. I said right from the start that we want opinions on this, and this is far from being an accepted, TÜV-certified seal or anything or a decision-making basis. At the moment it is just three papers I wrote and a GitHub repo with two stars. That is the level of market maturity we have right now.

I10  
 If you send it to me, you will get a third star. I actually do not find it uninteresting. The idea of classifying this, no really, the idea of classifying it or capturing these values, especially with Kaggle, you have these models and one can see it as an overview, but I am an absolute fan of the four lower values. The computer scientists mainly look at the four lower values anyway, and the upper ones.

A1  
 That is exactly the point. I have been writing papers and researching on this topic for a few years now, and what I realized at some point is that the target audience of such labels are not machine learners, because they will just say, send me the paper and give me the code. They do not want a label, they want to understand in depth what is going on. That means the target group, for whom such a label is meant, is completely different. Also behind the Nutri-Score or behind the EU labels there are complex test procedures or evaluation criteria. You can be critical of those, that is also fine. But the end user who simply buys a new washing machine in the supermarket does not need to understand how this label came about. They also do not want to understand that. They do not want that depth at all. Of course, in our case now or in your case, you have the technical understanding and you are also interested in it. You say, such a general rating does not give me enough information. I do not know where it comes from, I do not know how it is put together. That leaves room for manipulation. Absolutely right. But in some places that is the only thing that could be relevant. Not necessarily in the case of AI, because the analogy limps. No one goes into a store and buys an AI. But everyone can somehow try out OpenAI online, not download it but try it online, or maybe also download a language model locally, although that already requires a bit more expertise. So this level of communication, the way it is communicated, now you also say one should show how the lower ones are weighted. Yes and no. I understand that for technicians this would be totally relevant and they would want to know it, but if you put too much information on it, then it is no longer a label, it becomes too complex. That is a very strong trade-off, and I personally believe that reporting on these topics must be individualized. Much more strongly. And that is something we do not have at the moment. At the moment we have papers and GitHub repos, and you only understand those if you have been dealing with it for five years.

I10  
 I mean, what one could alternatively do, if I now look at it, is maybe split it into three labels. If you look at the lower ones, then you could have an accuracy label, an energy label, and a time label. I think that would be, I do not know which values you take in the background,

A1  
 There are not that many more either. Altogether we examined ten values. As I said, model size in number of parameters is still included, because that is somehow relevant especially for embedded and edge devices. How much memory is available? Does the model realistically even fit on there?

I10  
 Yes, but then maybe energy and memory, accuracy and time, so that you make a three-part thing out of it.

A1  
 Yes, that would also be conceivable.

I10  
 I mean, then you could also have the values without icons, just under the labels. I would already like that much better, that you simply, I mean, what I think is cool is, I look at it and I see the one and 63%. And I mean, if I then have something where I can rely on the fact that these 63% are correct, then I do not need to read through a paper with 63% anymore, instead I just throw it away and keep looking and then I find the 81% and remember that. So that is what I actually find really cool about it. And if that were on a website, I will stay here in this cancer field because I simply think that is incredibly relevant, and then at the bottom of the website the AI label would be shown or whatever, maybe even in the imprint. I mean, then already now, I have a wife who has cancer, I look at four things where you can test something like this. You already look at the label and say, okay, here 84% costs five times as much, but okay, it has the 84% label thing, that is where she wants it sent. So that would be cool. Or if it is just like, okay, these have the costs, the three cost less, those were more expensive, and one of them has an A and B level, I call it A level, also for that, but then it is like, I think there are values mixed together that do not belong mixed together, because they are simply different things, but it is also, I am also particular.

A1  
 Yes, but that is very good, that is why we are doing the interviews. We want individual feedback on it. Just for completeness. You probably mean something like this, where you have a comparison, and that is actually...

I10  
 Oh I like that thing at the bottom, the properties (unintelligible).

A1  
 That is in the end the little tool I developed somehow, a small dashboard where you can display benchmark results from different datasets. This here would be the ImageNet version and then you can hover over a model and here for example you have accuracy versus power draw nicely, so on the x-axis power consumption, on the y-axis accuracy…

I10  
 Those are different models. Yes, that took me a moment.

A1  
 Exactly, those are different models and if you now hover over a model, then the corresponding label is generated, and as I said this is all super prototypical and nothing where I would say it is market ready, but I have it like that so that when people visit and want to see what I do in my research, I can somehow show it.

I10  
 So top left for something like Kaggle I would find super cool.

A1  
 Yes, exactly.

I10  
 If I were the CEO here and you my employee, I would also expect it top left.

A1  
 And then you can also see that the best model usually does not have a real, actually that is a Pareto front, in research you would speak of multi-objective optimization, so there is not the one best, but rather different selection criteria, and you have to weigh or balance them against each other, and the best trade-off here is then made by MobileNets, but that always depends on where you set your own priorities. I can also gladly send you the repo if you want to take another look at it in peace, we actually still have a short section in the interview that we need to quickly discuss now so that we finish on time.

I10  
 I also find it really exciting. So that is why, the interest is really not just for the interview, but beyond that.

A1  
 Yes, very cool, I’m happy about that. So, as I said, I’m currently writing my dissertation on the topic, and exactly, there will definitely be more on this in the coming months. If you develop AI or work with AI, where do you get your information from? We’ve already mentioned Kaggle a few times, so Kaggle is definitely on the list. Are there any other websites or forms of representation that you look at when searching for information?

I10  
 I would split it once between using AI and developing AI. Using AI for me is more like ChatGPT or similar. With that, I also look at newspaper articles, I think, as one of the first points of reference. Just googling what newspapers write about it. Not that I take every newspaper seriously, but fortunately, or especially Wikipedia is quite cool because it lists sources. If there are critiques of certain models, the source is usually referenced, or it comes from a study, which is usually mentioned in the newspaper articles. Actually, when I use AI, someone has often already complained about it. And if so, why? I then look at why they complained and whether I am actually interested in why they complained, or if it was just like, yes, we can detect cancer well but it uses too much electricity. That’s a point where I might think, yes, I’d like to survive and the electricity doesn’t matter to me. When I develop AI, it’s mainly Kaggle. I really value the Morpheus tutorial by Cedric Messner, who always wrote complete code wonderfully. Normally, or the way I approached this AI, was to take one of Messner’s next tutorials and enrich it with Kaggle to break it down a bit. But with the sign language project, I really have to say that I hit my limits on Kaggle because they had nothing on it. Then it’s really just tedious googling and reading PDF files, hoping someone else did something. On page 7, you might find some result from some Asians. What helped us in the project was... You probably know him, XXXXXXX XXXXXXX, who is in the cooperative. You probably also did an interview with him because he, like me, was once linked in a negative way and told to go ahead. He helped me a lot because I had a communication basis with him where I could say, okay, this is the current status. Then I met him for half an hour on Zoom, and he just bombarded me with ideas. And they all worked for me. So also this personal exchange. And honestly, the way we now recognize these signs was at Easter with the family. It was my father who somehow came up with an idea. He especially got the idea because of energy efficiency, as a small funny anecdote that fits here. He comes from a time, he is 28 years older than me, where resources were limited. He thought about how to do it as efficiently as possible and came up with the idea of post-estimation, to rely entirely on it. So we are even quite energy-efficient compared to most other models. But that is more by chance.

A1  
 Yes, but it’s even nicer if you can also do it efficiently. In our research, of course, we also looked at what alternatives already exist on the market, maybe already established, to provide information about AI models. Then we, of course, first look at a classic publication, right? PDF, 40 pages, in this case all about MobileNet. On the right side, we have a ModelCard. This is something now also used on HuggingFace, this large repository for language models in particular, where you get information about the model at a glance, a few tables. Here you can see, for example, that it was tested on a Samsung Galaxy S23, with some benchmarks, inference time, because MobileNet was especially developed for use on mobile phones, hence Mobile. Then we have Papers with Code. I don’t know if you know it. Maybe something to keep in mind because it’s a bit like Kaggle with a nice bonus factor. You can see for all kinds of datasets and models the papers and GitHub repos in which they were discussed. But it’s completely user-contributed, so anyone can write anything there. It’s not always very tidy, a bit chaotic, but for some datasets you can find pretty good overviews. On the right, a blog post about MobileNet. This goes a bit in the direction of a newspaper, although blog posts are, of course, simpler because they have no journalistic standards, although nowadays online it all kind of blurs. Many journalistic outlets also post user stories and such. I think the difference between a newspaper article and a blog post is kind of blurring in the digital age. On the left, the documentation as in Keras from TensorFlow, if you want to use the model in practice. That’s also what I used to test all the models myself. And on the right, not for MobileNet specifically, but in general, IBM Factsheet. IBM has licensed it, patented it, and only provides it for IBM products. But it’s very similar to the Model Cards originally proposed by Google, giving a simple operating manual for a model. How to use it, what it does, what the performance indicators are. If you compare it, maybe you can comment a bit. What comes to mind, also in relation to our labeling approach? What are the pros and cons?

I10  
 What always annoyed me about these articles, in the middle bottom, is that due to the scientific approach, they have an incredible amount of text and it takes forever. Sometimes I had to read four or five pages to see, okay, they are not doing it the way I need. That’s really annoying. A Model-Card, having something like a Model-Card as a starting point, yes, a Model-Card, especially to clarify the input data. That’s actually the most important thing for the models, to describe the input at the beginning so you can immediately filter, okay, what can I use and what not.

A1  
 So a bit about the specification of the interface. So what goes into the model, what comes out of the model...

I10  
 Yes, so basically I find here this Model Card, I’m just looking at it a bit on the side or just opened it and scrolling down. Is that the one at the bottom left?

A1  
 No, in this case it is this one right here, so

I10  
 Yes, exactly, that’s what I had in mind too.

A1  
 This is now here on HuggingFace for example, there you have a bit like how to use it? Model Stats, Model Checkpoint, Input Resolution, Number of Parameters, Model Size. That is clear, it is about image classification, so a bit of this specification is included and you also have directly a kind of operating manual for how you can use it.

I10  
 That is cool again. I didn’t know HuggingFace at all, I know HuggingChat and played around with it a bit, but I didn’t know that HuggingFace also has all these things.

A1  
 Yes, Hugging Face is, if you look at it, also an overview, similar to Kaggle, there are just so many different datasets and for each dataset many different models are suggested and mainly Hugging Face has grown in the last few years because they provide open source language models and a lot of...

I10  
 [Unintelligible]

A1  
 Yes, exactly. That’s how you know most of them. Image generative models and language generative models. HuggingFace is very big for that. And what is funny is they also have these interactive Spaces where you can just try out models. So, I have no idea what this is. Fake Insects. Start playing. Okay, this is a bit like a game. You click the real Insects. Oh, another image is loading and so on. For example, this is now using an image generative model to display invented insects. But you also have...

I10  
 I know HuggingChat, I didn’t find it that great when I used it, but as a little experiment and especially that it is open source, I thought it was quite cool

A1  
 Exactly, like this MidJourney is one of the biggest image generative models, they were also directly here, yes, and you can also use them, all these Spaces here, they run in the background, if you want to use them a lot you also have to pay because it links to some offshore computer resources, so they also have a cloud-based infrastructure thing attached, which is quite cool

I10  
 Yes, then I think Model Card, although we might want to talk about the layout. So this, what is at the top, Model Stats, Input Resolution and so on, that would actually be, for me, what should be at the very top.

A1  
 Yes, that is a bit the problem, as you said before, there is no standard. Everyone does it differently and because of that it is not so easy to always find the information you are looking for. On Kaggle it looks different than on Hugging Face, it looks different than on, I don’t know, the Keras documentation and

I10  
 Yes, the worst are these scientific articles, there you just search yourself silly.

A1  
 Exactly, that is, that is also a different target group. These are people who want to take an hour to read it calmly and really understand it. They are not the users who just want to try it quickly. And, um.

I10  
 I don’t want to spend that much time either, but I want to know if I can avoid spending that much time. I mean, I also watch Netflix, check out the summary, and think, okay, the romantic comedy from India isn’t really my thing.

A1  
 Yes, that happens a lot now, because many people, when they write a paper, also have a short README on GitHub, which goes a bit more in the direction of a model card, or they write a blog post about it. So a lot is happening, and we are kind of fitting in with the labeling. Maybe you can elaborate on that. Do you see any connections to what we have proposed here?

I10  
 Well, especially if I go to this model card, I mean, if you open that Hugging Face page, you can see it perfectly... just a model thing. I mean, we have up there… Why are you putting yourself down? Stay up there. We have Space Shuttle, 86 Rockets, and so on. On the right side, we already have information. Downloads last month. Not really many. But up there, three scores with energy efficiency, speed, and recognition. I can imagine that perfectly. Maybe not in this label, which anyway looks like it’s meant to be stuck on some machine. But up there, these three scores with ABC, and you click and then get the parameters that calculate them. So this list, the A lights up big. And if we are on Accuracy, I click and a popup opens with Top 1, Top 5, and what other values were there for Accuracy? I can’t even speak anymore. With the weighting of these things behind it. I can imagine that perfectly. That would really be helpful. I open it, I see C, okay, gone. Or I want to do it on mobile, I open something, I see Energy is E, then oh god, I don’t even have to look further. It’s just an insight.

A1  
 But that is actually quite similar to just displaying the label here or the lower part of the label. I mean, of course, this is embedded in Hugging Face. The idea is that you could just put this on any PowerPoint slide and see at a glance, okay, it’s about the dataset, that’s the name of the model. Here are more details, and these are the important KPIs, the important metrics. I do see a similarity. Of course, the way it’s displayed can be changed.

I10  
 No, but what I would take over, or what I think is very good if it’s adopted, are the values you collect. That would really be… I mean, I miss that on Kaggle all the time. I have to scroll all the way down and see what they did, and they mostly only show Top 1 and so on. But that would give access to information. Just having it on the paper, I think, is like Oahh… because we are just one, so no, it’s just…

A1  
 Yes, that is…

I10  
 Basically, you’re not really saying anything if you just show ABC, because if you don’t include the weighting, it’s just like, yes, we used a bit of ink. Honestly, I’ve never cared about it either, I have no idea what the Nutri-Score is, it never interested me, and it doesn’t affect my purchase decision whether the Nutri-Score is A or E.

A1  
 What about energy labels?

I10  
 For the fridge. I pay attention to that. I also think that’s a good label, because apart from the name being wrong, I feel it really is about energy. It’s a value that gets compared, it’s not like we take seven values, turn on the mixer, and get a letter out.

A1  
 Yes, but of course, that’s semi-true. In the background, they test different things. There’s always a main factor, like how much water and electricity is used to wash seven liters of laundry. Not for the fridge, but for the washing machine. For the fridge, I think it’s about the temperature of the apartment, how much electricity is used to maintain the fridge at half load at a certain temperature.

I10  
 Yes, but there it really is only about electricity.

A1  
 Sure, yes, yes, of course, yes, of course, yes.

I10  
 And that is a comparable value for me. And as long as I have five refrigerators where I trust the brand that the fridge will actually work, then I would also look at the energy efficiency thing and say, yeah, the best one, then I'll take that one too. Where I wouldn’t do it, for example, is screens. I mean, I have screens that have a short response time, I have screens that have a long response time, so I just do different things with them. Or here, where I am right now, I simply don’t care if it’s faster or slower, whether it takes two or eight milliseconds, it doesn’t matter to me at all when programming. A label doesn’t help me there, because I have no idea which of them has the best energy efficiency, and above all, I care about other values there. So when it comes to one thing, I think labeling is good. But as soon as it’s mixed, it’s just somewhere, then it’s simply annoying that you can’t see directly what it actually is.

A1  
 Yes, I completely understand. I think that’s a good conclusion. So I can definitely live with that. XXXXX, do you have any questions or comments?

I10  
 She’s very quiet anyway

A1  
 Yes, we actually did it so that we always say, one leads the interview and the other kind of makes sure that if questions are left open, which you might miss yourself, then only intervene, and you don’t talk over each other the whole time. But...

I10  
 The more I think about it, split those things up and make three of them. Or four if you want, if you feel that Accuracy, Speed… Why can’t I talk today? Accuracy, Speed, and Energy Efficiency separately. Then it’s helpful. Then you look at it and think, yeah, because otherwise it’s just a mix and a mix is annoying.

A1  
 Yes, no, it’s a good conclusion. We have kind of taken the approach, or I have so far, that we simply want to transparently highlight different factors, but of course, you could also say, really make separate labels. It’s just nicer to have it at a glance, than if you now have different labels, but in the end it’s also a bit a matter of preference. That’s why we are doing the study and usability was not really the core focus, but rather just general feedback. I would definitely like to thank you at this point for your attention and for the conversation. I found it very pleasant and very interesting takeaways. We are now pretty much done with the interview phase. Over the next few weeks we will analyze and code everything in depth and then write a paper about it, which you won’t want to read. But no, of course I will…

I10  
 That can be forwarded.

A1  
 It’s about a few

I10  
 Since I already know what’s in it and know that the topic could be interesting for me, I’ll take the hour

A1  
 Exactly, so we will also make sure to prepare it in the form of a blog post or something like that, I think, so that it’s easier to get through. In the paper, it’s mainly about the…