A3  
 First track and second track are working too. Perfect. Okay, then I’d like to start and just ask, hey, introduce yourself, what’s your background?

I7  
 Yeah, so a little bit has already been teased by XXXXXXX XXX XXXXX XXXXX XXXXXXXX XXX XXXXXX XXXXXXXX XXXX XX XXX XX XXXXXXXX XXX XXXX XXX XXXX XXXX XXXXXX XXX XXXXXXXXXXXXX XXXX. Then I went abroad for a research and development position in the field of Artificial Intelligence, that was in Turkey. There, I think we completed about 4-5 projects in total during the 1.5–2 years I was there in the AI field. They were various projects, mainly in the area of computer vision, and yeah, it was pretty interesting. The thing is, I also signed a lot of NDAs, so I don’t think I’m allowed to talk about everything, but what’s public is… let me think… I better not say anything about that. But we did a lot of computer vision work, which was super, super exciting. However, I realized in this field that collaborating on an AI project, where you also get data from clients and then have to label it on a large scale, was already very, very cumbersome. So my colleagues and I, one is a fellow computer science student and another I met at XX XXXXXXXX, together we founded a start-up. I came back from Turkey at the end of 2022 and was initially sponsored by XXX XXXXXXXXXXXXXXXXX. Afterwards, my team and I got the Exist(?) grant, and now it’s running out this month. We already have our first clients, our first orders. It’s going in the direction of software as a service. Interestingly, the product/project is called XXXXXXXXXX, so yeah, an interesting name, anyway, for today’s interview. That’s basically the current state of things. What we do with AI internally in the company is on one hand classic deep learning, but on the other hand, and this was a lesson for me, when people talk about AI today, they no longer mean deep learning, they exclusively mean LLMs. So we’ve also been doing a lot with LLMs recently. I think I’ve said everything.

A3  
 Okay. Yeah, all good. What exactly is your business model? Would you like to go into that? And how do you specifically use deep learning and LLMs in your daily work or in your product?

I7  
 Yeah, so first and foremost we have XXXXXXXXXX, which is a web platform, software as a service, where you can hook in your data, and then you can divide it into work packages, label it. We have a super cool, modern labeling editor, that’s basically what makes us special. However, when it comes to selling it, we quickly realized that many companies aren’t yet ready to carry out their own deep learning projects. So we also started to contribute in a supportive way, meaning consulting/services as a sales approach. Because once we’re in a company, we can also market and sell our product even better in our view. Exactly. That’s our business model.

A3  
 Okay, so basically this labeling, the labeling interface itself, doesn’t have AI built in, but you act as consultants for, for example, applying deep learning models and large language models. Can it be summarized like that?

I7  
 Yes, of course there are supportive aspects in the labeling editor. I don’t know, you probably know things like Segment Anything, which can speed up labeling a lot. We’re currently adding different AI tools so that it becomes, what we call ALMA, automated labeling with human oversight. I think that’s a cool term. Yeah, exactly. So we’re introducing ALMA step by step now. And there are multiple AI modules being used there.

A3  
 Okay, got it, very interesting. What exactly is your role in the company as a founder?

I7  
 Exactly. I’m the CEO in our company, which means, of course, I make decisions in consultation with my colleagues, but ultimately I make the final call. I’m also the AI developer, but I’m not actually the project lead. That’s the other colleague, he has the big picture and the software architecture is a bit better. I create the specialized solutions, especially when it comes to AI implementation or specific modules that need to be completed. Exactly. That’s my role. And interestingly, programming is not even the main thing I do right now, in recent times it’s mostly talking, events, networking, a lot of office products, PowerPoint, Word, Excel. So yeah, interestingly enough.

A3  
 I believe that. So the management part is basically increasing more and more. Yes. Okay, interesting. But you’re still doing some programming, right? You have prior experience in computer vision. Did you already mention if you gained any special experience or expertise from your studies?

I7  
 Yes, actually my master’s thesis turned out to be absolutely industry-relevant, I realized afterward. During my studies, I did a lot of computer graphics, computer vision, you could say, and a lot of work at XXXXXXXXX X. Later I also did my master’s thesis in this area, and it was really an intersection between computer graphics and AI. What I did there was basically extracting material parameters of a surface from camera images, which are actually needed to render a scene, using AI. And of course, once I had the material parameters, I could relight entire scenes completely. It was a super, super exciting project. It was the first time I actually used AI myself, or rather implemented it in programming. I always thought, no, this is a topic I don’t want to deal with. It seemed like just putting Legos together and throwing data at it, and that didn’t seem interesting to me. But I did it in my master’s thesis. I have to say, actually I was right. It really was like putting Legos together and throwing data at it and hoping it works. Of course, it was a bit more involved, so it wasn’t that simple, but overall that’s basically true. I really enjoyed it, and tweaking the models in the layers was also very, very exciting. In the end, it led me to really develop a love for the whole topic. And exactly what I did there. For example, I used a UNET, a self-written one. I was later able to implement it in a similar way at work more than once. And that was very successful. So it was very interesting.

A3  
 Ah, very nice. Okay, but then for your startup you’ve kind of moved away from computer vision. What models or AI technologies are you currently using specifically? For example for this ALMA.

I7  
 Exactly, so several things are currently being considered and tested. On one hand, there is Segment Anything, which still goes in the direction of computer vision, since we primarily treat images and videos as, I’d say, first-class citizens. We still have a lot of computer vision models, object detection models, which we’ve actually been using a lot recently. Moving away from computer vision, as you said, we’re also using OpenAI, and we’ve started experimenting with the cloud, got AWS access, which was very exciting and important, because many people rightly have privacy concerns. We’re also using Llama 3 and similar open-source models a lot recently, running many experiments, and we even got a new huge server so we can test this on a larger scale, things like that.

A3  
 Okay. You just mentioned privacy, but what are some other challenges you face in these everyday processes, like deploying AI models or finding different AI models you can use?

I7  
 On one hand, it would be very good to have an overview table, it solves this information very well. But what I would really like is something more like, this can be particularly well used for this use case, because a number in some test doesn’t help me. That it does better than a human in law exams in the United States also doesn’t help me. And that it can solve programming tasks, which are often structured very similarly, very well, doesn’t help me either. I really need to see how it can be used in actual implementation and real use cases. And that is already the first big challenge. So I don’t know, with an LLM, no matter how it is presented, what its actual capacities are. I really have to test it deeply myself and push it to its limits to find out whether we can use it and if so, how we can use it. Exactly, that is one of the biggest challenges, I would say. And otherwise, of course, and we live with this currently anyway, how can we make it solve exactly the tasks we want, limit hallucinations, make it work based on facts. There are many techniques we’re now using, but yes, these are the challenges. We try to counteract them, so to speak. I’m not sure if that fully answers the question.

A3  
 Yes, definitely. How about non-LLMs? For example, your computer vision models. Do you face any challenges there as well?

I7  
 Yes, even from work, one thing comes to mind immediately that is very important. This is really something: in practice, computer vision, I think, is not at all like, oh, we just have an object detector, that's it. We have very highly specific data. The data do not necessarily have to be images and videos as we know them. They can be microscopic images, they can be thermal images, it can be any kind of camera images, and average models usually do not work at all for that. And especially if the quality or certain aspects in a video or a photo are not present as the AI actually expects, then most things are unusable for most applications in my experience. So it cannot be used at all. And that's why very, very often custom solutions need to be created in the field of computer vision to solve something specific. Transfer learning alone is not enough. So it often has to be very, very, very specific to solve a very specific task. I would say overall that computer vision or deep learning requires much more expertise and experts than the whole LLM thing. I think you don’t have to be as involved to make the LLMs work.

A3  
 Okay, I understand. And do you see anything, this might be a bit of crystal ball looking, but what can we do better in AI research, for example, to somehow solve these problems? What do you see as needed?

I7  
 So when you say these problems, do you mean exactly what I said, that very specific AI models have to be created to solve certain tasks?

A3  
 Yes, well, I mean, you solve the problem through customizability, but do you see anything else that could help you?

I7  
 Well, the thing is, this is research and development and I don’t know what can be done to make it less research and development. I mean, more and more niches are being covered by more and more niche solutions, more and more startups are focusing on more and more areas. That is definitely, of course, the economic path, but maybe also the solution, because unfortunately I don’t know, crystal ball-wise, no idea, how you can solve such a research and development task retrospectively and sustainably. I have no idea. So it is very difficult.

A3  
 Okay, yes, we have definitely thought about it. It can roughly go in that direction. We want to make AI models comparable with each other, so that a selection process is simplified. I would like to show you, what you may have already seen on the website. And these are our nice AI labels. Can you see them?

I7  
 I just see black right now. Ah, now.

A3  
 Now you can see something. Okay. I cannot really apply it to your problem yet, but without priming you further or influencing you, just tell me what you see spontaneously, what is your reaction to this AI label?

I7  
 Well, the order in which I read it, I can say immediately, this is very interesting for the designer, I think. So the first thing, of course, is this A, B, C, D, E, that immediately catches the eye, and then starting at the top, yes, I didn’t really read this AI level, then directly MobileNet V3, then this Infer ImageNet, and then I jumped down to the battery, probably how much energy it consumes, I guess. I don’t understand the one on the right. Corrupted Robustness. Bottom left, Top Accuracy. And bottom right, Running Time per inference. Yes, so top left, bottom right is immediately clear to me. Bottom left, I would need to know what the Accuracy refers to. And top right actually tells me nothing, I have to say. So, uh...

A3  
 Top right, the robustness, you mean? Okay. Yes, I mean, the Accuracy is, we have here an image classifier, MobileNet V3, ImageNet. I don’t know if you have worked with ImageNet before.

I7  
 I think I have, yes.

A3  
 Yes, but it is a huge image dataset with a thousand classes, where different objects are shown. And yes, the MobileNet V3 is supposed to classify what is visible in a specific image. Is there a dog, a cat, an ice cream on a stick, or something like that. And the Accuracy refers, in this case, to how much MobileNet V3 correctly guessed what is visible in the image in the test dataset. So 63% rate. You can imagine, if we just randomly picked one of these thousand classes, we would have a one-thousandth accuracy. And the Robustness is a metric that indicates how reliably this model ultimately decides. That means, the images are attacked, pixels are changed, something new is added.

I7  
 Ah yes, I know that

A3  
 And then you look at how much the result of the MobileNet, the classification result, actually changes and how robust it is, so to speak, against these pixel attacks. Exactly, that would be regarding robustness. Do you have any other questions? Did anything else come to mind just now, maybe during the explanation or because of it?

I7  
 Well, now that you mentioned robustness, corrupted robustness, it reminds me… it’s probably referring to this attack from that paper, where, for example, by changing just a few pixels, you can completely change the class however you want. So you say, it currently recognizes an author, but it’s supposed to be a duck. Visually, for a human, it looks the same, but in the end, it gets… I think that’s the attack you mean.

A3  
 Well, I don’t know, XXXXXXX, I understood it as this actually happening through a random pixel attack, but what you mean is, of course, those inference attacks where you really change specific pixels to change the class.

A1  
 Exactly, you also have to say that the value is actually only hypothetical right now; the others are real measurements. I am currently working on building another test view where we can test adversarial robustness practically end-to-end with Auto-Attack. Exactly, so there are targeted attacks where you really try to find the perfect attack to, for example, trigger a label flip. That is of course a bit harder to be robust against, because it usually tries to use the model’s outputs on other images or the gradients to somehow poison the input, so-called poison attacks. What we are actually testing here is more of a completely generic approach: we throw various distortions or Gaussian noise or whatever over it with a certain norm, and then you know, okay, given these attacks, we still have this and that robustness. But yes, robustness testing and adversarial testing is also a rabbit hole; you can make it arbitrarily complex.

I7  
 Yes, I understand. What also came to mind is when we used to evaluate models. What we often did was let different teams pursue different approaches when I was still working abroad. And to compare them in the end on the images, the most meaningful metric for us was always false positives and false negatives, so that you can see it in an overview. And that was also the goal, to minimize it. And indeed, when we minimized it, there was of course also the visual aspect, the qualitative, almost subjective insight of how well the AI works, which came directly together with the quantitative metric of false positives and false negatives. Naturally, of course, there are metrics that you optimize where visually nothing changes, but that always seemed to me a pretty good metric. It was also always the first thing we looked at, comes to mind when I see this now. Another thing that comes to mind, I must say, is that I see the label now, and for example, it’s rated A. I guess it means overall the whole model is very, very good, that’s why it got rating A. I guess that’s what it means, and not something like energy efficiency A. If that’s the case, I would be a bit careful with such a label, even in a few years, because what does A mean to me from seven years ago, three years ago, two years ago, compared to today? A is a completely different story. And I would need to know, time-wise, when exactly it was rated A. I guess if it’s big, I see 2012, that probably refers to the model.

A3  
 And even here, we actually issued that, it indicates when the label was basically created. So July 24, basically this month.

I7  
 Exactly, it should be a bit more prominent because I think these two pieces of information definitely belong together. From my point of view, this is just my idea, but they absolutely belong together and should be displayed almost equally prominently so that you can really use the label at a glance. So 2012 should be about the same size or 2024 should be about the same size as the A, B, C, D, E, that’s my intuition, but that’s my first thought.

A3  
 Okay, yes, great. Exactly, this overall rating comes from, among other things, these metrics and I think six more. I don’t have the exact number in mind, and it all comes from a study by XXXXXXX, published a few years ago or last year, where multiple networks solving the same task, in this case image classification, were compared. From this comparison, it was derived, for example, whether the battery is particularly energy efficient or accurate compared to other models. All these metrics together give the compound score, which is what you see here at the top. That’s how it was created.

I7  
 I would say, looking at this, the biggest advantages I get are also in telling a story, not just to developers, because honestly, developers, or I would now, and I think many others too, could manage without such a label. And many additional pieces of information, if you had extra labels or deeper info, could make it very useful. But the biggest benefit I immediately see is aimed at decision-makers and clients, meaning customers for a AI developer, the company wants to implement AI, and we can show them these various labels. Where I see the greatest benefit is that you can show them, with two labels, look, we have the solution: one could be solved with, for example, a pretrained multimodal LLM, finding your objects. The development effort is low, you prompt it a bit on what to find, and you get this inference. But look, the energy consumption compared to a specific model doing exactly the same task is disproportionately higher. Also, the time the LLM needs is disproportionately longer, so we can already discard it, for example. And if you can show these two labels, one with a D or something and the other with an A, it should be immediately clear to the decision-maker that a specific model is really much better than a generic giant model that can handle many tasks abstractly. Furthermore, I am sure the accuracy of a specific model would also be very high. So overall, the greatest benefit is actually when I can present it to them and say, hey, look, we’d like to show this to you.

A3  
 Oh yes, very good point. I would also like to just add a second label. Exactly, this is precisely one of the cases we had in mind with these AI models, where you can show two models, for example, like it is here. Here we have MobileNet V3 on the right side, the one we just had. And now we just added EfficientNet B4. It is, of course, also an image classifier, so not a large language model like you mentioned in your fictional example. But do you see here, especially now when you have a comparison like this, maybe also as a bridge to your everyday work, would you use it?

I7  
 I have to say, the first thing I notice is that I didn’t realize earlier that the icons at the bottom were color-coded. In hindsight, it of course makes sense, but I thought it was just a green battery icon and a red accuracy icon and that was it. Only now do I realize that the color actually has significance. All the better now that I understand how it works. But it wasn’t obvious to me before. Whether it makes sense to compare two AI models like this?

A3  
 Yes, and whether you see a bridge to your daily work where you could use it.

I7  
 So what I do, this is what I mentioned a bit at the very beginning, is quantitative metrics and qualitative metrics, or any metric expressed as a number versus actually testing the model to see how it really solves my problem. And so far I haven’t seen any metric that could show this very well. As I said, false positives and false negatives came closest, but even there we sometimes had models with much lower false negatives but still apparently more wrong results. So in the end, it means I would still have to try both, and I would intuitively start with those rated A, to maybe save some work if it works really well. But I would still test many models and, based on experience, I cannot rely on a label alone because it is just a very specific thing. As a developer, I think that is the important point here.

A3  
 Okay, so you wouldn’t rely on the label at all or would you say maybe it helps for a pre-selection or just to get an overview, or would you really say no, I have to test everything, it has no added value for me?

I7  
 So think of it this way, if I start testing and I see the first models are very good, the ones rated highly with the label, and the further down I go, the worse they actually get, then I would eventually stop. I wouldn’t keep testing down to E, I think, exactly.

A3  
 Okay, so a bit of getting used to the label, feeling your way in.

I7  
 Yes, using the level a bit supportively, to know when I can stop testing.

A3  
 Okay, all right. Generally speaking, what stands out to you about the label, is there anything particularly negative that bothers you, something you think could be improved to make it more helpful for you?

I7  
 So besides the things already mentioned, that maybe the year should be a little bigger to make it clearer, because I think now, okay, if we use this label this year, or next year, okay, but if I can see at a glance, wow, that’s an A rating from 2013 for example, so if you look ahead, an A rating from 2024, we are in 2034, then I would like to see that at a glance. That’s one point. Secondly, as I said, I didn’t immediately recognize that the icons at the bottom are color-coded. Only after seeing the second label did I notice that. But the most important information is actually on there. As I said, the Corrupted Robustness was not relevant for me so far, I think. And there is the Top One Accuracy, whether maybe other metrics could be included, I don’t know, I can’t say. But overall, it looks very good visually, very appealing, as an overview.

A1  
 Maybe a brief note, because it might be relevant for you. So the software, I developed software in the last few years to generate such labels completely generically. You can basically load a Pandas DataFrame with this model on this dataset, with these properties and as many rows as datasets and models you tested. And theoretically, you can, I tried to make it so, it is a work-in-progress class, not a product or anything, but you can also decide yourself which icons should be used and so on. Right now it’s still a bit rough because it is mainly research software and not developed for productive use. But the idea is that you can also generate such labels generically yourself, based on your own evaluations and benchmarks, and exactly the metrics relevant to you, to show others what the model you specifically trained can do, or what models you benchmarked can do. And as you said at the beginning, it is especially relevant, or at least I hope it is, to bridge the knowledge gap. In your case, you have the advantage of being management and AI expert, so you can train models yourself, evaluate and compare them, and make decisions. But in many cases, in many companies, there are people in decision-making positions without deep understanding of this field. And the overall scoring is also influenced by one’s own priorities. So if you want a very energy-efficient model or have strict runtime constraints, it has to evaluate camera images within a certain number of milliseconds, these are of course domain- or use-case-specific constraints. And the label, as well as the overall scoring, can be adjusted to emphasize which metrics are particularly important. So it’s a bit of an open-source approach behind this, and it is quite adaptable to individual cases.

I7  
 I also really have to say, now that I think about it, even when we finally got our model working, like for a client, those were really the most important aspects, that we really got the time down incredibly, meaning inference, and got the accuracy incredibly high. And the funny thing is, if back then we had the possibility to also internally evaluate our pretrained models or our rates/waits(?), let's say, with some kind of thing, so that we could immediately see at a glance which rates/waits(?) are labeled how, so labeled in this sense. That would have already, that would have already been useful, such an overview table or database. Of course, there are also alternatives, like having benchmarking lists and such, but yes, it would have been relevant somewhere, I think.

A3  
 That is a very interesting use case you are opening. It's basically about quickly labeling self-developed models during development, just to communicate better, to have a better overview of what one is developing at the moment. Yes, you also make a good point there. What other forms of communication do you even know? How do you represent AI models to each other, for example, to communicate them within the development team or to clients? Or how do you make a selection? What information do you refer to for that?

I7  
 For us it’s like… Sorry, I just got a call. You asked which information we use to present and communicate the AI models to each other. Exactly, so funnily enough we of course had losses and the loss was already very involved, and if you got the loss down, that already said a lot, but, and it’s a big but, at the end of the day what is always, always, always important is the visual approval on our typical test videos, because at the end of the day a quantitative metric is really a quantitative metric, and having seen it for yourself is really super important. And the interesting thing is, it doesn’t help me if the model performs very well quantitatively because it has optimized some specific scenarios perfectly, but what the client actually sees most has really stupid errors, then the client will only see that during approval, and then it doesn’t matter how many points it scored internally. But of course, there is a very strong correlation between the quantitative metric and the qualitative result. How we finally did it was really that we looked at which visually delivered the best results. And we always kept that one. Unfortunately, we didn’t have a proper system. We had things like model names such as One Model to Rule Them All or some nonsense like that, so we knew okay, this is the model that tops everything else or so. Yes, but we also eventually shortened that, so the abbreviation One Model Rule Them All, we took the first letters of that and it was then always versioned 1, 2, 3, and yes, it was a bit chaotic, we didn’t have a proper system. And that is also the reason behind the startup a bit, that you have a benchmarking tab where you can immediately see how well your models perform, but I would personally only use that as an indicator. That’s just my experience.

A3  
 Yes, alright. Very interesting. Good point also with the benchmarking. We once searched the worldwide internet and identified six different forms of communication for how AI models can somehow be communicated. On the one hand, of course, there are scientific publications, from which AI models often emerge. In this case, that would be the MobileNet V3 paper, written by a Google team. Of course, a lot of text, a lot of how it came about, but here there are also many benchmarks in such papers. Then we have Model Cards, which in this case are from Google, and also present something a bit shorter, for example. What does the network actually do? What are the inputs? How big is the network? And again, here also a few benchmark results, how fast it is, under which environmental variables these benchmarking tests were conducted. Then we have Papers with Code. I don’t know if you have ever used that.

I7  
 Of course, yes.

A3  
 Yes, great, okay. So you basically know everything that MobileNet V3 cites. If you scroll down here, you’ll find a great overview, like Towards Data Science but also Medium, just these blog articles, which are now used by a lot of people when they don’t know how to proceed, yes, a bit less scientific, a bit more journalistic almost. Then of course the Python documentation, which also really specifically indicates which parameters should be set and flags when you use a MobileNet V3. And finally, we also have the factsheets, actually a bit like Google’s counterpart, but here it is IBM. IBM always has a relatively extensive listing of information about AI models. For example, they also include the point Bias, which I find very interesting, so really model distortions. It can also go in the direction of discrimination through AI models, fairness. And yes, very extensive information but only for IBM-published models. Yes, that’s how we identified them. If we now put this side by side with the AI label, what comes to mind spontaneously? What advantages and disadvantages do you see here in comparison?

I7  
 Yes, first of all, it’s of course the clarity, that you can see the most important aspects at a glance. Especially if you know the label eventually, then you would also see it, you know the four characters and then you know what color the four characters are, you don’t even have to read exactly and you know how good they are. So you can already see everything a bit at first glance.

A3  
 Otherwise, are there any disadvantages that you might also see in the AI label compared to another format that you already know and have worked with.

I7  
 I have to say, the ones you just listed, the ones I mainly always used were… so of course I looked into the paper and checked how they compare with the benchmarks and the other models. But what I used most towards the end was Papers with Code, I had seen who performs pretty well in this benchmark, looked at that paper, looked at the second-best paper, looked at the third-best paper. But actually, I have to say I looked at a lot of papers again and again. Yes, these model sheets, or whatever they were called, that exist on Hugging Face at the beginning, I have to say I always skip those. I hardly read them at all. Because if it solves my task, that is the most important thing, then if necessary I set up two servers. Or then I check whether it is really necessary to be realtime or to get inferences every second. Building around that is more feasible if it can at least solve the task. But all in all, you have to say, of course, it is clear that your label is presented very clearly and really shows at a glance what might be relevant for some people

A3  
 Although you are actually the type who really reads the papers, not everyone does that either, because for me that would mean you are more the kind of person who wants detailed knowledge but still see an advantage in it

I7  
 Well, I don’t read the papers, at most I skim through them

A3  
 Yeah, well, no one really reads papers

I7  
 That’s what I was about to say

A3  
 Okay, understood

A1  
 That’s a classic statement. First year PhD student: no one reads papers. I really like that XXXX, don’t tell your PhD supervisor

A3  
 No, definitely not

A1  
 Nice money quote from our interview

A3  
 Right?

I7  
 No, but it’s like this: I think how papers are structured and how you read papers, or how I read papers, especially for a model, is quite different. I immediately look at the visual results, then how it compares in the benchmarking, scroll up, if there is a very detailed appendix, I usually look at that the most, and then if I think, okay, how did they actually do it, I try to read the abstract quickly to see what they did, what it’s based on, roughly how they did it, and if that is still interesting, then I really go in and look at the methodology, what they did differently, which modules they used differently, and yes, that would be the order of how they set up the experiments and which quantitative results came out in the end and how they measured it, that usually doesn’t interest me. I really don’t have time for that

A3  
 Yes, understandably. Okay, very good. Then I think I now understand what kind of added value an AI label could bring for you. Then there is still, yes, a somewhat speculative question, who should actually issue such a label? Can you imagine how that could work in the real world?

I7  
 So what XXXXXXX just said really convinced me that you could say it’s a bit like open source. That means, if I spin the idea a bit further, I don’t know if this is exactly what was meant, but if you could provide a software suite in a way that it reproducibly, verifiably has the label, and also shows that it has the label, that would be the only thing I would accept. Everything else that isn’t reproducible on its own, and anything that an entity issues and you have to trust them, I wouldn’t… maybe in the end you would still use it if it’s really good, but I wouldn’t really welcome it, I think.

A3  
 Not even something like a TÜV, which many people have already mentioned, some kind of independent testing organization that already tests other things today, or a government authority.

I7  
 I have to be honest, judging by the community as it is now, the whole AI community, people are proud of their models, proud of their GitHub pages, proud of whatever. They will want to generate the label themselves, I think. What you can definitely do, of course, is offer a place where you can upload your model in a Docker-container-like way, and it tests it for you and produces the label, verifiably showing that this instance authenticated the label. Because then you also get this authentication, it would make sense to have such an instance and see who authenticated it, instead of some random person on the Internet issuing it for themselves. I would probably, unless I had seen a case where someone cheated, also trust the label, of course I would still test it in the end, but I think this authentication thing, which just came to my mind while I was talking, also has another added value, I think. Yes, a TÜV-like organization, yes, that could be.

A3  
 Okay, yes, also, I’m hearing this suggestion for the first time, I think it’s very clever, really to do it as a Docker container and then just let the label authenticate itself, without necessarily needing a TÜV, for example. What many have also said is that a TÜV wouldn’t be able to keep up, as fast as AI models are developed and evolve.

I7  
 Yes, exactly, that’s why I couldn’t even imagine it, to be honest. This is something that has to be fast, and moreover, it has to work for every version. I would integrate such a thing into a CI/CD pipeline, meaning if the model is further developed and retrained, then automatically this TÜV certificate or whatever verifies it on the website, triggers the build, generates the label, and shows the latest label in the Readme. Everything fully automated, I just can’t imagine a TÜV team doing it manually.

A1  
 Yes, I actually think that’s a really good point. I have thought about it before, that, of course, like you have code documentation and in Keras you also have tables showing how well the model performs on different hardware, I assume that this is not CI-based. Sure, the documentation is CI-generated, but I don’t think they are actually testing the models or weights in the background CI-wise. Instead, someone probably copied it from the paper manually into the documentation, and now it ends up in the Readme. And if some developer accidentally resets the weights and still pushes to master, I don’t know if the software tests catch that. It would actually be really cool if you really work on a concrete model, write a paper about it or provide it somehow, set up a GitHub repo, and for each model in the repo, often like EfficientNet, with different versions and sizes, labels are generated showing it was tested on specific hardware, runtime used, benchmark results, that would be really cool.

I7  
 What I would wish for this label, definitely, is something similarly well-designed. So this is the main label, let’s say, and on the right side, a larger secondary label for those interested, showing other relevant information, maybe comparable models and how they perform in contrast, etc. A complementary label next to it. If that doesn’t exist, the author of the model has to provide it in another form below in some way. I think if you issue a label anyway, it’s pretty close to having an overview label and on the right a detailed, larger label with different graphs that might be relevant. At the end of the day, this is just a general opinion, you probably can never cover all the niche aspects. You would need a huge label with everything imaginable. And we also had our own metrics in the company, which we constructed ourselves. That can never be represented in such a label in a general way. But that it can be generally useful as an overview, seeing at a glance, oh, this is fast and generally A-rated, for example, is definitely interesting.

A1  
 Yes, especially when you say that it is extremely individual, the needs, the metrics, all the different learning tasks, there are of course now, I mean we are only talking about an image classifier here, but there are incredibly versatile learning tasks and accordingly models and solutions for them. But the big advantage would be, of course, if it is not done by an independent third-party instance, but rather has this open-source idea, so that everyone can grab the software, provide the metrics and the icons that are relevant for them, and then it is more of a rough framework with a few examples of how to use it, but you can then adapt it for your own needs. That would be the advantage. If a company then says we build super specific models for this or that use case, they can still generate such labels for themselves.

I7  
 Now that you say it like that, what would be super exciting is if you generally had this database of labels and they are now, let’s say, tested, there is this general label, and then somehow, if it was a JavaScript widget or whatever, if you could put your configuration that is relevant for someone, which is then encoded in a small string or something, just put it into this input field and press Enter, and then it automatically builds this overview that would be relevant for you. Or if you had a table with 100 models being compared, if you put your configuration in at the top, either as an encoded string as I said, then the interface immediately shows all the labels relevant for your view, that would be super exciting. That would be something I can imagine being used intensively, I would use it intensively if I could reconfigure all the already tested labels so that they are, yes, relevant for me, all the information. Of course, you have to say there are metrics that just need to be calculated, which cannot be shown in a pre-generated thing, but that would already be a really good step, I think, in the right direction. I don’t know if it came across what I meant with the configured string and then having these labels configured for yourself. I imagined a little bit how it could look.

A1  
 I think I roughly understand what you mean. Actually, as you also say, more detailed information would be totally practical. I’ll just send a link here. This is also work in progress. It is now a different use case. It’s about forecasting models. But I have also worked on a tool that visualizes this interactively. And here, for example, running time versus a specific error metric. Exactly, XXXX can open it quickly. This is even simpler. If you now hover over a specific point in this scatterplot, it generates a tabular summary and the corresponding label. It takes a bit of time because this is a free rendering instance with like 0.1 CPU and so on.

I7  
 Is that something like… is it Dash?

A1  
 Yes, it is Dash Plotly. You can also look at the label in large, of course. This is not perfectly polished. It is still the old Energy Label scale here. These are more like energy labels, based on the EU energy labels. But in the end, the idea is the same. And then the plot, this 2D scatter plot that you see, you can actually customize what you want on the X and Y axes. And then you have exactly this comparison between different models, for example, you have it exactly through that. And you have the properties, all the metrics again as a detailed table with the real measured values and the relative scorings and the rating, which is color coded at the end. And also the weighting for the overall rating. You can see it all there. This is also part of the open-source library I am working on.

I7  
 That is of course super exciting when you have such a label and you scan it, like with the QR code above, you get to a page where you can really see exactly how this label comes about. That is super exciting. And as I said, if you can then, without looking around too much, quickly put your configuration in there so that you immediately get the view that is actually relevant for someone.

A1  
 So the configuration, what exactly do you mean by that?

I7  
 Ah, I kind of finished it in my head, like I thought through a few steps, so maybe I’ll just finish explaining. I imagine you have a page where you really see all the detailed views, and then you can use different sliders and controls to adjust the dashboard and the view so that exactly the things that are relevant to you are displayed. You can hide things that are not important, things that are currently

A1  
 When you say the things, the things, do you mean for example all these different metrics, and you mean that

I7  
 Exactly, exactly. Metrics, maybe comparisons with other models that are similar. Exactly, things like that. And then the view, how it is displayed, etc. For example, if it were a software library and I just get the link from the label, and I can build my own dashboard, that would be one thing. Or the other thing would be if you’ve arranged it on the website once, then the website has a share view button, and you get some code that is generated automatically. And if you then enter it somewhere else for another model, it automatically shows you this configuration again, so you can always view your models in this configuration, for example. I don’t know if it’s clear what I mean. Yes.

A1  
 It’s clear what you mean. In my head, the development effort immediately explodes. But in principle, it’s of course... We’re not talking about whether it’s technically possible. It’s just that it then goes more in the direction of a product, which for my PhD until next year is not necessarily the top priority. But of course, in itself, that would be cool. I’ve also thought before about what would be really awesome. There is actually a button for that. If someone has this label, it would of course also be super cool if you could automatically generate a model card as a PDF from it. Like in two pages, description of the model, important benchmarks, comparisons, and so on. So that everything is linked end-to-end. But that all collides at the moment when you want this framework to support as many different models and learning tasks as possible. So it just produces a... a model card is highly dependent on what problem you actually want to solve. And yes, it’s not so easy to really generalize it. So this is now generalized to the extent that the person interested doesn’t know what the semantics behind Training Time are. For them, it’s just some number, and it’s called Training Time. And then you have other quirks, like some metrics being bounded differently. That means you would have to, like, maximize Accuracy, minimize Running Time. That’s the simplest case. But then there are metrics where zero, for example, is not an allowable value. You can have Zero Accuracy, but if you have Zero Runtime, then you actually need catches around it, because something is really broken in that case. And then you have metrics that are scaled. So it’s not trivial to standardize everything.

I7  
 So as I said, overall, my conclusion after looking at all of this, I keep repeating this point, is that it is always very individual, depending on the niche, etc. But the biggest added value still lies in the fact that you can present it to customers, tell them, look, these were important points, we configured the label like this, these were the important points for you. And the model we suggest to you is now rated A. The model, if we solved it with an LLM for example, would be catastrophic energy-wise, which is very important to you, so it would be a D or E solution. That is really a huge added value. In communication with people who really know little, and especially in communication with customers. And if you can also manage this, it’s very much in the direction of, how should I say, carbon emissions and so on, which is also very important. And that allows companies to say, hey, you can now put a label on your AI showing the energy efficiency of the product that customers are using. That could also be very relevant for them. So you have to see. But that would then be a bit different. That would then only be energy and sustainability, etc.

A1  
 Yes, but that’s also a big area right now. So it’s also one of the things, as you can see, that I’m very interested in, how you can make machine learning resource-conscious. It doesn’t necessarily have to be edge computing or ultra-low-power devices, but at least having this awareness of how much energy is being consumed. Especially now that we have LLMs and sometimes image models, where so much is abstracted away, and you don’t even run the stuff on your own hardware but rent everything as a service, a lot gets overlooked. And in the end, of course, companies are really interested in, they want a red dollar sign for how much it costs to run the model, and a green one for how much money the model makes them, that’s basically the metrics that matter most to them. But for the people who have to implement it, like developers and engineers, especially those who might not have ten years of machine learning and AI background, it’s really hard to map this profit-cost thing onto metrics for machine learning models.

I7  
 Yeah, no, I totally agree with you.

A3  
 Very cool. Well, that actually already kind of answered my last question. But I’d still like to ask again: how useful do you find a general certification? You said you could imagine definitely using labels under certain conditions. How do you feel about a general certification if it were kind of mandatory or a standard?

I7  
 I’m always very, very, very cautious with mandatory things because if you want to do it, it has to be incredibly, incredibly easy. We already have enough obligations, and now with the AI Act coming, many people are panicking about how exactly this will work in the future. We’ve actually done workshops on it, so everything introduced as regulation is of course important. Since I started getting a bit more entrepreneurial, I’ve also understood the entrepreneur mindset: it’s already super, super hard to meet all the requirements and at the same time build up a startup. Talking to lawyers and notaries, reading NDAs, having them reviewed by lawyers, going through contracts, making sure your imprint is correct, your privacy policy is correct on the website, and so on. And if then there’s a label for AI, and let’s say you make many AI models, and it’s a bit of effort to get certified, I’m currently in favor of reducing unnecessary regulations for startups rather than adding more… I get a bit cautious about that. You have to see how it would actually be implemented, I have to say.

A3  
 So more like voluntary self-commitment, like the Nutri-Score. A company doesn’t have to put the score on their cereal box or anything. Could that be a way for you?

I7  
 Well, I’d say, if it’s meaningful, if in the end it becomes something really useful and you manage to make sure people actually get value from it if it’s voluntary, then I think many people would do it voluntarily. As I said, someone developing an AI model, then it would become a standard thing to include so they can sell it even better, so to speak. Now, I think Nutri-Score is a bad comparison, because for me it’s something I never look at and for me it’s completely, well, that’s my personal opinion, one of the worst metrics I’ve seen in my life used in practice. Really… well. That’s my personal opinion.

A3  
 Okay, thank you very much.

I7  
 For me, Nutri-Score is absolutely irrelevant.

A3  
 Alright, good. Then I actually have no further questions. I don’t know, XXXXXX XXXXXXX, is there anything on your mind?

A1  
 Nope, I thought it was very cool. Thanks a lot for the insights.

I7  
 Oh, you’re welcome. I have to admit, I did it out of interest.

A1  
 Yes, but exactly, you’re right at the interface that’s interesting for us. I also have an industry project I’m working on alongside, and it’s really exciting right now how the whole AI field is being kind of revolutionized by LLMs. That means we’ve talked to a lot of people for whom AI is simply a chat model. At the same time, we’ve also talked to people for whom AI isn’t deep learning, but who really make an AI transfer for people who, I don’t know, sometimes still want to use linear regression because they have a, I don’t know, 600-line industrial dataset and want to do intelligent things on it, intelligent things in the broadest sense. It’s very different. And that’s why the study is really cool, you get a lot of insight into what’s actually relevant in practice. That’s why we do it.

I7  
 Yeah. I just see you don’t use, uh, so you exclusively use Plotly, no Matplotlib or anything else, right?

A1  
 I don’t use Matplotlib anymore. Everything is interactive. Well, lately I’ve often started including a dashboard in papers directly so that you can… that’s the idea a bit. But I think we can stop the recording now, right?