A3  
 Did you want to do that, XXXXXXX? Perfect. And then the question would be, could you briefly introduce yourself. Go ahead and briefly introduce yourself.

I6  
 Hello everyone, I’m XXXXXX, I’m 28, working full-time, and I just finished my Bachelor’s degree last week, actually in the field of AI. And that’s why I found this a very exciting topic, and I’m very curious about what’s ahead.

A3  
 Okay, yes, congratulations first of all, so fresh. So that means Bachelor is your highest degree, and in which company are you currently working?

I6  
 I’m working at XXXXXXXXXXX XXXXX, which is a subsidiary of the XXXXX group.

A3  
 Okay, can you briefly outline what the business model of the company is?

I6  
 We work in the field of order management. That means we support companies when an online order comes in, to see where we can best fulfill this order. At the moment still with very rigid rules, where you first have hard criteria and then weighted criteria. But in the future, that can still change. Then we also have a few applications to somehow locate the order on site and pack it together.

A3  
 Okay, interesting. And what is your specific role in the company?

I6  
 So the title is Solution Engineer. I help partners and customers connect the shop system with our platform. That means I support partners and customers in building connectors.

A3  
 Okay. And do you use AI or machine learning for that?

I6  
 In parts, yes. We have a GPT model in the XXXX group that can be used to answer questions.

A3  
 Okay. What does that look like exactly? So you just mentioned, GPT module. How exactly is that integrated and who interacts with this GPT module?

I6  
 So we employees can, when developing… that’s like a Teams bot, so for Microsoft Teams. We can then post questions. That’s based on OpenAI’s GPT-4.0, I think. And then when developing and when asking questions, we can write something in there and then get a corresponding answer.

A1  
 Just to quickly follow up on that, so Order Management, we’re only talking about online orders and where they are best processed, so that’s basically the daily business?

I6  
 Yes.

A1  
 Okay. And that means your contact with AI and Machine Learning isn’t in the area of improving the processing or management of orders, but rather more in the sense of providing you with a kind of chatbot so you don’t have to call IT or a colleague for every little thing?

I6  
 Exactly, so for us as developers it’s a work aid, but in the product itself there is currently no AI built in.

A3  
 Okay, got it. Could you describe such a process, for example how you use this work aid? Just a bit of a sample workflow.

I6  
 In our company we use TypeScript. I’m not quite as familiar with that as with .NET. And what I often do is, I want to write something, for example a function that processes events. And then I go ahead and think about how I would do it in .NET, copy that code, and then say “hey, show me how to do this in TypeScript” and paste the code in there. And then I look at what the bot has generated, I check it, and then it ends up in my code in that way or something similar.

A3  
 Okay, all right. When you use it like that, are there also bigger challenges, or any problems, obstacles in using it?

I6  
 I notice that because it’s such a Teams bot, it has a very, very large context model. And you notice that sometimes… I ask a question and then it answers in a completely different context, maybe one I had written in there three days ago. And because of that, the answer quality is sometimes a bit worse than with ChatGPT or so.

A3  
 Okay, what exactly do you mean by a different context? Does that simply mean that many different questions can be answered by the GPT and therefore it’s less specific? Or how should I understand that?

I6  
 Let’s say today I have a question about SQL. And then I ask a few follow-up questions. Tomorrow I ask a question on another topic, and the day after tomorrow I have another question about databases, but this time for example about MongoDB. Then it still starts giving me answers that fit SQL more, because somewhere in the conversation history I had talked about SQL. That’s much less to the point than what I actually wanted. You really notice that you regularly have to check, I think with “New Chat” or something it tries to learn that, but not completely reliably. That works a bit better, but you can tell it often uses the context of the old conversation. You log in XXXXXXXXX.

A2  
 Yes, we tried beforehand how we would do it, and I just couldn’t find the button to raise my hand that quickly. I find it very exciting that you use GPT. Do you know roughly how the decision was made that this was used and not something else? So was there some kind of comparison process? You probably didn’t decide that yourself, but…

I6  
 Well, for that, I and especially our company are too small a player in the XXXX group. So this chatbot runs somewhere in a data center of XXXX. And accordingly, there is a team that takes care of artificial intelligence. And they said, this is now the model. Probably because it is the most widespread, because it is a very large model, and probably because it could be integrated into Teams very well. But exactly how, unfortunately, I don’t know. Thank you very much.

A2  
 No, no, I didn’t expect that at all. What we are interested in is precisely how such decisions are made. In your everyday work, what kinds of systems do you use? Could you explain that to me again? Because XXXXXXX and XXXXXXXXX are the techies. I need a little longer for some things. Could you explain again, on which decision level you are and what kind of decisions you make?

I6  
 I actually don’t make any decisions in the company.

A2  
 I can’t imagine that. So you are just told, do this, and then you do it?

I6  
 My task is to advise customers.

A2  
 Ah, okay.

I6  
 Accordingly, we have a system with which you can fulfill orders and see from which location, if you, I don’t know, see a great pair of pants at Vero Moda and order them, we check what is the best location. Meaning, the shop where you place the order sends the contents of your order to our system, and our system then checks what is the best option. And I help people to connect their shop system with our system.

A2  
 Ah, okay. And the software you use for that, I probably don’t know it anyway…

I6  
 No, XXXXXXXXXXX XXXXX.

A2  
 Okay. No, never heard of it.

I6  
 Very creative, the company XXXXX XXX XXX XXX XXXXXXXXX XXX XXX XXXXXXXXX.

A2  
 Thanks, I need the repetition once more, that really helped me, thanks. Sorry, XXXX, back to you.

A3  
 Okay, wonderful. No, it’s totally fine that you ask that. Perfect. We also have the time. So, back again to the machine learning model you use. You mentioned the somewhat lower specificity earlier as an obstacle. Could you spontaneously think of something that, as AI researchers or more generally the company developing GPT, could be done to make your work easier?

I6  
 Well, I think Microsoft has already created something good with GitHub Copilot. That’s very specific too, giving hints directly in the development environment. And I think, well, it’s kind of difficult. GPT is designed to be, I’d say, a multipurpose AI, where you can throw in different kinds of questions. Accordingly, it can explain financial things to me relatively well, but also technical things. And I think if you were to put this technical subfield into a bot, that could help many developers in a much more targeted way than when you just use a multipurpose AI. I’m sure there are tools that are only made for that. Unfortunately, at my company I’m only allowed to use the internal tool. And I think another limitation isn’t necessarily this low specificity, but the fact that it’s one big conversation. I don’t know, you’ve all probably seen ChatGPT before. And there you can always start a new chat. And then you can chat in it. But with Teams, it’s like I always have the same ongoing chat where I ask things. That means in the same chat where I talk about SQL, I also talk about financial topics, I also say translate this text for me, and I also ask how to best bake a vegan cheesecake. And I think that makes things a bit tricky. So if such bots also had the option, if you want to integrate them into some communication yourself, to really have the entire conversation be forgotten. It might still appear in my Teams window, but the model wouldn’t use it as context anymore.

A3  
 Okay, now imagine you basically have several instances of this bot. Like one bot for the lemon cake recipe, another one for .NET programming. So you’d have multiple models, and you could choose between them. What kind of information would you want from these AI products, from these models, to help you decide which one to use for which task? So before using them, that’s important.

I6  
 So before I would introduce it in a company?

A3  
 Yes, you can answer it that way.

I6  
 If it were only about the technology itself, meaning just the model, I think it would be very important to me that it’s very specific to technical topics. That means not necessarily only .NET, but maybe generally more IT-related topics. Because at work I rarely bake cakes. So I don’t really need that kind of information from an AI. But if we’re talking about a more general decision process, and if I could design my dream AI, I’d also find it important that it was developed in the EU and that there’s maybe also a lot of open source involved. But that’s probably more of a personal preference. And I think it makes sense, as a European company, to also use software from Europe, especially when it comes to data protection.

A3  
 Okay. Good, then I’d like to move on to the main subject of our discussion and share my screen. I already asked you what kind of information you’d want about a model before using it. I hope you can see my screen, my Chrome. Perfect. And you’ve probably already seen this when you signed up on our website. I’d now just, without influencing you any further, ask you: what do you see here, and what’s your spontaneous first impression of this AI label?

I6  
 Well, I actually find it kind of funny that it’s very similar to the labels I also have on my fridge and things like that. It really reminds me of those. But what do I see? I see that this is a label to label AI. It even says so right at the top. MobileNet V3 Small will be the product name, I’d assume. With the option via the QR code, the lower QR code, I can access more information about this label. What that thing in the top right is, I don’t know, since it doesn’t say anything about it. As for how exactly this classification, this “B” rating, was measured, I can’t really say. There’s no information given. At the bottom, though, I see a few key figures. Top left, the Power Draw per Inference. Doesn’t directly tell me what the label means. The battery suggests it’s the consumption per request. But “inference” is unclear to me right now. Corrupted Robustness is 0.6%. That would make me assume the model hallucinates very little. Bottom left is also 0.6% Test Top 1 Accuracy. That makes me think it’s supposed to be very accurate. But 0.6% is a bit low for accuracy. And Running Time per Inference, if I assume inference is one prompt, means one request takes an average of 1.29 seconds.

A3  
 Okay, yes, very good. So you already mentioned many correct things, definitely. The MobileNetV3 is basically the AI model that is being labeled here. And that is an image classifier, trained on ImageNet, which is a large dataset, an image dataset. And yes, this… I’ll start with these four symbols down here. You also said correctly, so inference is basically… XXXXXX correct me, sorry, XXXXXXX correct me anytime. That is basically a prompt, if you want to have the analogy to GPTs.

A1  
 Practically an evaluation. Input (inaudible), into the model an evaluation, that’s what one would call inference, yes.

A3  
 Exactly. For that it needs 608 milliwatt-seconds. Then robustness, that is determined by the model being attacked multiple times through inference attacks, and it basically states how vulnerable the model is to manipulation in an image and how quickly an image classification changes, meaning the decision of the model ultimately changes. Then here is the Top-1 Accuracy, that is simply the accuracy of the AI model, meaning how often the decision “hey, in this image is a cat or a dog, for example” is the correct decision. And then here we have the Running Time per Inference, again per input-output combination. How long it ultimately takes. And the scale here is actually calculated from all these factors. And as a sum one arrives at the result that it was classified as B here.

I6  
 May I ask a question about that?

A3  
 Yes, of course.

I6  
 So 0.6 percent isn’t really that accurate, right?

A1  
 Yes, that’s my mistake. One has to say, of course, this is not a product in use. So this label thing is something I’ve been developing in my research for a couple of years. I wrote a software framework to generate such labels for, let’s say, arbitrary models. Provided that one can actually perform the tests that are reported below. And that’s a bit of the problem. That doesn’t work for many models. And that is simply a formatting issue here. So the percentage number needed to be multiplied by 100. I just put them in there as float values and didn’t explicitly multiply them by 100 again into proper percentages.

I6  
 So it’s 63 percent.

A1  
 Yes, yes, exactly. It’s 63 percent accuracy. One might think that’s not really good either. It’s also a red label down there. But one has to add that ImageNet has 1000 different classes. That means, if you just randomly say, you get an image and you just randomly pick one of the 1000 classes, and you do that for all the images in your dataset, then you would have an accuracy of 0.1 percent. That is the random search. So the model can already distinguish a lot, very well. We humans are extremely good in visual computing, and it’s a task that, well, to learn that machine-wise is not trivial. Same thing by the way with speech recognition. Which I think is kind of a general thing here, from what I’ve gathered in our conversation so far, your contact with AI is always very specific to large language models or language processing. Which makes sense, because that’s the system, that’s also what we see everywhere right now, which is being operationalized strongly and brought into practice. But that’s of course not the only AI being worked on in the background. In the automotive field, for example, traffic sign recognition for autonomous vehicles is also being worked on, which is also in a way an AI. This here, however, is a label not for language models but for an image model.

I6  
 Yes, that was a good point that you also mentioned that this is a Red Label. Now I also realize that the colors actually reflect a bit of what you also have up there in the scale. I wasn’t really aware of that at first glance.

A3  
 Yes, exactly. Maybe under the impression of the explanation you just received, you see a bridge to the obstacles you just explained as well, how this label could possibly help you. Maybe you can also do a bit of transfer work and apply it to the Large Language Models, if that makes it more tangible for you.

I6  
 Well, I don’t really have that much to do with image recognition. I’d actually like to have more to do with it. Haven’t really had much contact with it so far. But the accuracy actually fits with what I mentioned a bit. And that’s also what I’ve experienced myself here with Large Language Models. For example, when I have some kind of error while compiling, throw it in here and ask what went wrong, I usually also get, fits quite well actually, in about 60% of the cases, a good answer. In the other cases you have to ask again maybe three times until you finally get a fitting answer, or not. That really fits quite well with my experience. 1.2 seconds, I think, is even a bit faster than GPT.

A3  
 Yes, that could well be. In this case it really is a different task to carry out an image classification compared to creating a prompt. But yes, that’s also a good point. It has already been mentioned that these colors are based on a certain comparison value. And actually, XXXXXXX, who has been dealing with this topic for quite a while now, compared a whole range of different models, all of which, among other things, carry out image classifications. And based on this comparison of models, these color scales could even be created at all. And now I would like to show you a second model for comparison. In this case it’s EfficientNetB4. And here I’d first like to ask you openly, can you summarize what you see here in comparison?

I6  
 In comparison, what’s it called, EfficientNetB4, consumes a bit more electricity. It’s yellow instead of green. And yes, 1.6 watt-seconds are also significantly more than 600, 800, kind of blurry, 608 milliwatt-seconds. Right. But the one on the left is less vulnerable to this attack you just described. And it’s also significantly more accurate with 81.3 percent rounded. But it takes almost twice as long.

A3  
 Okay, and how, well, how do you interpret this information? You’ve kind of already interpreted it.

I6  
 What surprises me, well, as I see it right now, the one on the left is classified with D, but actually has much more green in it than the one on the right, when I look at the summary below. That means, this is actually maybe quite similar to the fridge label I have and maybe is mainly oriented towards energy consumption. Which means that seems to play a greater role here in the overall rating than the other values.

A3  
 Yes, although one has to say that further metrics were also considered here in order to obtain this overall classification. So, um, I just remembered, we also have the Top Five Accuracy, meaning whether the correct result is among the top five results of the network. I don’t know, XXXXXXX, what else was included?

A1  
 Yes, we also have model size, meaning the number of parameters in the model. That in turn is of course more correlated with resource consumption or energy consumption. More generally, the analogy to the energy labels is absolutely correct. I already mentioned at the beginning that I also orient myself a lot in the field of Resource-Aware Machine Learning, and one major problem we have at the moment is that especially in research there’s a very strong focus on quality, which is understandable because we all want our products in use to work properly. But very often the energy consumption is really neglected and abstracted away, and in the context of climate change we’re facing a bit of a small crisis there. I don’t know if you’ve ever looked into it out of interest, how much energy it takes to train a language model like ChatGPT, like GPT-3 or 4. We’re talking about energy amounts that could power a small town for a few months. So, thousands of GPUs are used that run continuously for weeks so that in the end they deliver good results, and that somehow also has to be made transparent.

I6  
 I think that’s a good approach. I mean, sure, accuracy is extremely important, but as much as I’m interested in it, I also think it’s extremely dangerous how much energy the internet in general consumes.

A1  
 Absolutely. And there once was this...

I6  
 Netflix and the three biggest adult video providers are, I think, responsible for about 30% of internet traffic. And of course a data center also has to somehow be kept running.

A1  
 Absolutely. And there’s this nice example with a Google search. With that you could also boil a cup of water. That was actually measured in real terms. It was a big topic about ten years ago. Of course, all companies are working on making their internal infrastructure, their server infrastructure, as efficient as possible. So it might be cheaper now. But the computational effort of a language model request compared to a simple Google database query is horrendous. We’re talking about an estimated factor of 10 to 100. That’s where this closeness to resource consumption and motivation kind of comes from. And one also has to say at this point that the energy labels stuck on electronic devices don’t actually make transparent where the rating comes from. In the background, tests are conducted. They check how much energy a washing machine consumes when washing seven kilos of laundry compared to the best-selling product from five years ago. And from that it’s relatively derived how good it is. But this level of detail is never communicated. That’s not the motivation anyway. People are supposed to go into the store and say, I have 400 euros, I’ll buy the most efficient model for this price. They don’t really need more information than that. With these labels here we’re of course trying to derive a few more characteristics, because… It’s always the question of who ends up using it. Whether people actually want this level of depth or not. So if you wouldn’t see the lower section now, you’d just say, okay, then I’ll take MobileNet, because it’s better. And of course one could also strategically, for example, recommend the use of resource-efficient models a bit more. But I’m taking over the interview here. XXXX, sorry.

A3  
 All good. So maybe also with that in mind, what do you find particularly helpful here, especially if you put yourself in the position of really having to decide between several models one day in your company, for example? Which information stands out to you here?

I6  
 The request time feels very present. Like, the fact that EfficientNet takes 2.3 seconds is clearly longer than the other model. It’s always the question of what the application is and how much data needs to be processed. For example, for facial recognition at airports, I think MobileNet is better suited, since you’d have to detect many people at once. I think accuracy plays a big role. And actually, now that we talked about it, I realize how much more I pay attention to how accurate and how fast the model is, compared to how much energy it consumes. Ideally, of course, there would be a model somewhere in between, one that provides good accuracy and good speed. Accordingly, if accuracy is important, I would actually rather choose EfficientNet.

A3  
 Okay. Is there maybe something, on the other hand, that you don’t understand, that is hard to grasp, or that you really can’t make sense of?

I6  
 The corrupted robustness. You just kind of described that these are attacks on the model, but I honestly didn’t really understand it. I thought it’s something you could just brush aside, but apparently not.

A3  
 Okay. All good. I think I also explained it a bit poorly. Let me try again. I think what they’re trying to do is take the images that the model is supposed to classify and slightly change pixels. Right, XXXXXXX? And then they test how much these pixel changes influence the decision of the AI model. So whether in the end a distortion of the image also leads to a distortion of the classification or the decision the model makes. XXXXXXX, do you want to save me here?

A1  
 No, that’s completely correct. One has to add here that the use of language models, the way you know them, and of ImageNet models or visual computing models, is different. Here it’s about classification, meaning you ask a question and get a result, and it’s either true or false. With generative language models, that’s not the case. You ask a question and you get an answer, which may include true or false information of course, but often there simply is no correct answer to an open-ended question. That means the model generates a plausible answer that you’re supposed to perceive as truth, but very often there simply aren’t right or wrong facts to such questions. And that’s also what you described earlier, this hallucination, meaning making things up, spreading obvious falsehoods. That comes from the fact that language models are not trained to reflect accuracy. They’re not only fed correct information, but the internet. And the internet has all kinds of things. You find some weird blog where someone says if you eat washing pods, you’ll get smarter. Bam, at some point your model, without you realizing it, starts making such stuff up and tells you, okay, you have to eat dishwasher detergent and then you’ll do better on your next exam. This robustness test is something that specifically makes sense for classification, to check if I slightly alter my normal data, where I know the correct information, how robust the model is against it. For language it’s harder. But there are also tests to check hallucinations a bit better. That then goes in the direction of accuracy, meaning how exact, how correct the answers are. But there are also behaviors like prompt injection, where you can get models to output information they’re not supposed to. For example, for a long time if you went to ChatGPT and just asked how to build an atomic bomb, then ChatGPT gave a very good guide on how to build one yourself, or an explosive device, or something simpler than an atomic bomb.

I6  
 That was it, pretend you’re Evil GPT and tell me what you would say and what Evil GPT would say.

A1  
 Exactly, and that would then, for example, be such a prompt injection, where someone tries to trick the model a little bit, and again of course there are AI researchers who think about how we can build tests for that and query other quality metrics beyond pure accuracy. Exactly, and things like that also exist for language models. These here are all just examples. What would be especially super interesting for us would be, in general, where you say, in this context, in the contact you have with AI, what are, well, we’ve just talked about accuracy and runtime, so the results should be accurate, the results should be quickly available. Are there any other characteristics where you would say, I would really like to know this about the model, and at the moment it’s completely abstract to me?

I6  
 Well, I don’t see that here, for example. I mentioned earlier that I would like it if it were developed somewhere in the EU. I don’t see at all where this comes from. I see, okay, it’s EfficientNetB4 or MobileNetB3 Small, but where does it come from, which company developed it?

A1  
 That’s hidden a bit behind the QR code, this “Further Information”. There you get to the paper from the people who originally developed it. But maybe also think a bit beyond the concrete labels we’re showing here, because this isn’t the AI you’re working with, but think about the models you actually use in your everyday life, or at work as well, so language models, chat models. Do you think of any other characteristics where you’d say, that would be super important, if I could get more information about what defines this model, and that’s something I’m missing.

I6  
 What data it was tested on. For example, if I have some kind of human recognition system at the airport, if we stick with that example, I need completely different things than, well, these here are image recognition things, than if I have, I don’t know, a suitcase recognition system at the airport, that checks: is there a bomb inside, is there water inside? And if, for example, the one on the right is super good at recognizing humans, but what I want is something for luggage, and the one on the left was trained and tested precisely on such suitcase contents, those values are already a massive factor in my decision.

A3  
 So, ImageNet is basically the dataset it was trained on, as XXXXXXX already said, that’s a thousand classes, from dog to ice cream cup. It’s meant to classify objects in images. How would you, since you don’t know ImageNet.

I6  
 No, not at all.

A3  
 Okay, would you maybe wish for more information in that direction about this dataset? Because, sure, ImageNet tells me something, since I’ve worked with it before, but obviously not to you. What kind of meta information would you like to have about it?

I6  
 As I said, what was it trained on? What was it tested with? So what was the context? If we have something that only recognizes objects that, well, I don’t know, that now detects that I have a hand cream in my hand, that’s, as I said, a different object recognition than if I use 3D data from a suitcase scan at check-in for a flight. That’s a completely different context, a completely different kind of object recognition.

A3  
 Okay, so in the end a more detailed description of this dataset, instead of just the dataset’s name.

I6  
 Yes.

A3  
 Okay. Good, okay. Then another question would be, do you know similar forms of communication from AI models? I don’t know if you’ve already looked some up. Probably not for work, maybe more privately. Other ways in which AI models have been represented, presented to you.

A1  
 Or maybe also regarding the concrete model you’re now using at work, a chat model or a ChatGPT model. Have you ever tried to find out what kind of model it is, where it comes from, what characteristics it has? Or were you maybe briefed at work? Did you get some kind of memo saying: by the way, we now have this service, you can just chat with a model, you need to know this and that. Is there anything like that?

I6  
 Well, before it was OpenGPT 3.5, but I only knew that because I asked the thing: what model are you based on? Now, since I use the thing and I’m in some kind of email distribution list, I don’t know if there was a company-wide announcement, I know that very recently it’s now a ChatGPT 4.0. But apart from that, I haven’t really looked into it in depth. For my bachelor’s thesis I had to check how GPT works, so I did some more detailed reading, but only about that. And only because I had to. The others, like something like DORLY or so, they all cost money, so I haven’t tested them.

A3  
 Okay, all right. Very interesting. Of course, there are other ways in which different companies have already presented AI models. And of course, from academia there are also very, very many AI models that have in some form emerged together with a scientific publication. I would like to go over these five different forms of presentation here again and briefly introduce them to you. These are basically the… these are the alternatives to these labels we just looked at for a long time. There are, of course, the scientific publications, for example MobileNet V3, which we also just had. That came out of a paper. And…

A3  
 OpenAI does that too, also on Arxiv.

A1  
 Exactly. For example, for GPT-4 there is also a corresponding paper, 60 pages long, you can just skim through it.

I6  
 I did that too, that’s why I know.

A3  
 Easy. Ah, okay. Yes, interesting. Then of course the Model Cards. Those are from Google, if I remember correctly. Where you can also see certain benchmarks, for example here. So here too we have the inference time, which we also just showed a bit larger. What is the model for? What kinds of data does the model work on, and so on? Then there’s Papers with Code. I didn’t know that site myself until a few weeks ago. If you scroll down a bit there, basically all the papers are listed that in some way cite or use MobileNetV3. And then of course there are the blog posts. You know them, Towards Data Science or Medium. Those are the biggest ones, where definitely a lot of information can be found about individual AI models. Then we have the classic documentations. And finally, the IBM Fact Sheets. These are fact sheets developed specifically by IBM that present facts about AI models that are deployed by IBM itself. And this is also structured. So you can see it here a little bit. They organize the whole thing based on categories and then basically give you different kinds of information. For example, how fair is the model? How fast is the model? How accurate is the model? And this is presented in a structured way. So, that’s just a quick overview of the alternatives. And now the question would be, if we once again pull them all together into one picture, what advantages and disadvantages do you see here compared to the labels?

I6  
 Let’s look once at this Medium, or I think it was another site, but Medium cited it. What bothers me about Medium is that basically anyone can write anything there and sometimes it’s not such good information, I think. If we look at the Arxiv post, just the one you have there is guaranteed to have like 50 pages. That means you have a lot of text in order to somehow inform yourself. Sure, you can just read the abstract, but I find it a bit too much text. I think that’s a disadvantage, but the advantage would be that you get very precise information on how it was done. The class documentation, that rather explains the work with it and not what exactly happens in the background. That means it makes it more usable for you, but you don’t really find out exactly why. So you can only base your decision on usability, how well can I implement this into my code, and not on how this model works in the background. The one top left, I don’t even remember anymore, that was the IBM Fact Sheet.

A3  
 Yes, that should be the one, yes.

I6  
 Yes, I thought its structure was quite good, same as the one from Google right next to it. Because there you can also specifically just look at the part of the model that somehow interests you. And you still have the precision, maybe not quite as precise as the paper, but you can still get informed in detail about the topics that interest you. The one top right is from a while back.

A3  
 That’s Papers with Code. All publications are listed there.

I6  
 Ah, so there you see everything where it was used. So the advantage is that you also see how others have used it, but of course it also has the downside that you again have to read just as much as with the paper. I think the advantage, sure, with the label that XXXXXXX designed, is that you can get informed very quickly. The disadvantage is that you need quite a lot of background knowledge. That means, in order for me to know what kind of image database it was trained with, I also need to know what ImageNet is. Or else somehow scan the code. But it does make it a little easier to grasp at a glance.

A3  
 Okay, thank you very much. Then the question would be, you also just said, yes, your fridge also has such a label, that is organized or certified by an EU authority, namely how these labels are designed, which energy classes are ultimately assigned, and in your opinion, who should issue such a label?

I6  
 That’s very difficult, because we also saw with these… with various things that were decided everywhere, that not always only neutral arguments played a role in the decision-making. And such a label should also somehow have the possibility to be designed in a very neutral way. So it’s difficult to determine such an authority, I must say, because if you say the EU has to decide this so that we somehow have something here in the EU, the largest models were simply developed in the US. And if we then somehow… well, you end up with that anyway, which means the label doesn’t benefit anyone, and such an overarching unit doesn’t exist. So it would be, well, a voluntary standard would somehow be good, but there is no control authority for that, and that’s a difficulty with such a thing. So if you take a look at this EU-wide Nutri-Score, we also have that problem. I also see, if the EU were to decide something like that, that you can have a big influence yourself on how you present yourself. For example, this Nesquik cocoa has, I think, A or B on the Nutri-Score, but it is obviously a very unhealthy food. They just made the serving suggestion with two teaspoons and very low-fat milk. And then of course it’s better than the cocoa prepared with three teaspoons. That’s why it’s difficult to define such an authority that should do it. So I find it really hard to make a clear decision here.

A3  
 Okay, and what would, again, speak against a neutral third party for you, so a European authority? Because a European authority could also certify American models. At least that’s how I imagine it.

I6  
 The people who, well, in itself it’s a good idea to have this done by an EU authority, if it’s really neutral and you can’t somehow adjust the code in such a way that you look especially good in the test they conduct in order to label you, but only for that purpose. That’s a bit the danger of misuse with something like this, because I think, in such places, the rules are usually set by people who don’t have the corresponding knowledge to provide a proper decision-making basis.

A3  
 Yeah, okay, although I mean, now, if you mention the Nutri-Score as an analogy, there it’s disclosed how the decision is made, and that’s why manufacturers can hack the system so well, and it’s also created by experts, in a way. Yeah, doesn’t mean it wouldn’t be possible.

I6  
 [Unintelligible]. If we can be sure that these are somehow also neutral criteria, and maybe they’re also more quickly adaptable, especially because we’re talking about a very changing product here. I mean, ten years ago it was hardly imaginable that this would exist in the way it does today. That means, if the rules can also adapt to the future at this body, then I do think an EU body is a good place for that. The framework conditions just have to fit.

A3  
 Okay, all right. [MISSING 2 SECONDS]… talk. Would you say that the label should be adjusted to your knowledge background, or that it should look different for different user groups, depending on who it’s aimed at?

I6  
 You can, for example, write values on it that everyone understands. Like, one request costs you this many milliwatt-seconds. What’s written on it, you can kind of decide yourself. What I think would be good is if it were the same for all AI models, because with the label we’ve just seen, we were very much in image recognition, but something like Gen-AI, which consumers actually interact with a lot, wasn’t captured at all. Because something like corrupted robustness, you can hardly measure that in a Gen-AI model. So if you had something that applies to all of them, and then you also put something on it that consumers can understand, I think you wouldn’t need to adapt it that much.

A3  
 Mhm, okay. Then, since I basically have you already for the very last, my last question goes in the direction of sustainability. That’s an aspect that, as I understood it, is new, well, not entirely new, but has newly entered your horizon today.

I6  
 Don’t put it like that.

A3  
 In relation to AI models. In relation to AI models. Is that okay?

I6  
 Yeah, in relation to AI models, yes. Otherwise I think I’m more knowledgeable than anyone else.

A3  
 No, no, no. Really only in relation to AI models. I mean, we don’t know each other that well. Wow. So, what role do sustainability and resource efficiency play for you in this labeling, especially in light of this AI label?

I6  
 Well, now that it has newly come up, so, in general, in the global sense, it was always clear to me that it all costs money somehow, but I wasn’t really aware how much electricity is behind it. I think now it’s become a bit more important. But I have to admit, accuracy is still a bit more important to me.

A3  
 Although, I mean, you’d now say there’s a certain trade-off, so, does accuracy really play the top role, priority number one, that should be maximized first, or would you still take both metrics into account, both energy consumption and accuracy?

I6  
 So, when we’re talking about a label like this, I quickly noticed that one model consumed significantly more, simply because it was more accurate. I think if you have a good mix of all three points, that means accuracy, good speed, and good resource consumption, I could live very well with that. And then I might also accept a lower accuracy in favor of a more sustainable solution.

A3  
 Okay, all right. Okay, wonderful. I think I’ve checked off all my questions. I hope the okay. Okay, XXXXX nods. The check is done. Then we can also stop the recording.