A2

Exactly, let’s start the recording. And thank you very much, XXXX. Could you briefly introduce yourself? What’s your background? What do you enjoy in your studies, and what’s your main focus at work? What are you involved with?

I5

Yes, so first of all, I’m currently studying to become a secondary school teacher in computer science and social sciences at XX XXXXXXXX.

A2

Ah, perfect.

A4

That fits well.

I5

Yes, it really does. So I might actually be a good bridge here. Right, and I’m now in my second semester of the master’s program. I wrote my bachelor’s thesis last year. And while I was working on that thesis, I founded a startup. Kind of by accident, you could say. I have a XXXXXXXXXXXXXXXXXX. I don’t know if you’ve heard of it. It’s a relatively new intolerance. It’s trending now. Somehow it seems like everyone has it these days.

A2

I’ve heard of it, yes. I think a friend has it

I5

I’d say it’s kind of like the “new gluten” right now. And I actually developed an app for myself, to help find supermarket products that are tolerable or not. That was the basic idea. I just uploaded it. I’ve made other apps before, but honestly, no one really cared about them. But this one was different. We’ve already gotten tons of emails, with ideas, feedback, and so on. So I went to XXX with it, because I thought there’s definitely more we could do here. And you probably know XXX. I think I also sent you an email. Yeah, so the XXXXXXX XXX XXXXXXXXXXXXXXXX XXX XXX XX XXXXXXXX. It's kind of like transitioning from university into the business world, that whole startup space. Together with them, I applied for the XXXXXXXXXXXXXXXXX XXX. And we got it, well, more like earned it. So now I’m working on the app. I'm still studying on the side, actually, I’m still a student, but the app is also a big focus right now. But it works well to manage both. And then I actually found out about you guys through XXX, because someone posted it in our Slack channel. And that’s how I came across your website.

A2

Cool, that sounds really exciting. But I can’t quite imagine yet what role AI or machine learning plays in the app. Can you explain that?

I5

Yes, in different ways. So first of all, it’s basically machine learning. There are two types of scanners. One is a barcode scanner, like the ones you know. You scan the barcode, like with CodeCheck or similar apps, and then you get a result right away. And then there’s also a kind of text recognition AI, because often you don’t have data from the barcode, so you can directly extract the data from the text, like from the actual ingredient list on the product, and analyze that. That way, it’s not dependent on databases. So that’s the first part it’s more about machine learning, text recognition, and so on. That’s basically the first use case, and it’s already implemented.

But I also want to move later toward something like a food diary. Since everything is being scanned anyway, I’d like to do more with that data maybe something like a coaching feature that suggests which foods to try again to gradually expand your intolerance list. That’s the general idea, but it’s still in the early stages. The other AI parts like the text recognition are already well integrated.

And yes, I’m also working on things like the Nutri-Score and similar systems. That also has a side to it, I think. That’s why I found it interesting, and it probably fits quite well.

A2

Yeah, yeah, that’s the visual element that XXXXXXX used, right. Okay, so yeah, I think that’s text recognition AI and the barcode, right?

I5

Yes, exactly, that’s already part of it.

A2

Yeah, okay. And are these systems that you developed yourself, or do you get them from some databases or something? Sorry, I really have to ask the basics. How does it work?

I5

Right, so first of all, the scanner technology. Scanning the barcode or scanning the text, that comes from Google. It's called the Machine Learning Kit, and it was specifically made by Google for iOS and Android, so it's designed for smartphones. You can use that, they’ve already pre-trained the data and provided a good interface. You could also do the text recognition yourself, of course, but I also wanted to get it done at some point. So I just went with what was already available.

A2

Fair point.

I5

If it’s already there, right? Exactly. And for text recognition, it just takes the output text and analyzes it. And for the barcode scanner, there are databases that I query, basically. And ideally, they already have an ingredient list stored. That’s how it all comes together.

A2

All right. Okay, cool. So, what does your typical workday look like? I guess in a start-up it’s probably even less structured than in a corporate environment, but could you tell us a bit about what your day-to-day looks like?

I5

Well, I'm still working on my own, so that makes things a bit different. It's actually always different, there’s still university in between. But otherwise, there isn’t really a typical day. Mostly I work on further development, fixing bugs, or for example, right now I have an issue with the ingredient scanner, it often picks up terms like “gluten-free” or “naturally lactose-free,” and things like that. The challenge is filtering those correctly so they’re recognized as "free from" and not as containing lactose, for instance. These kinds of things constantly need fixing. The food industry moves fast with these things. There are thousands of E-numbers and such. So yeah, it’s a lot of bug fixing and replying to emails, but I wouldn’t say I have a fixed routine or anything.

A2

We’re just trying to get a sense of what you do and what it looks like. You just mentioned that there are some problems. What would you say are the challenges you face in your work? You already mentioned the lactose-free issue, but maybe you could speak about it on a more abstract level. What do you find difficult when you’re trying to develop things further? What are the challenges?

I5

Now about AI or in general?

A2

Start with general, and then feel free to talk about AI too.

I5

So, in general, there’s definitely a bit of a data shortage when it comes to food products. I would love to have more barcode data available. There is some manufacturer data too. But usually, it's users themselves who upload the data and make it available to the community because there’s no other way. The manufacturers are very secretive. Especially with discount store products, it’s always carefully concealed who the actual producer is. That’s a bit of a problem.

A2

Yes, we’ve heard that before. The data problem comes up in most of the interviews.

I5

And the funny thing is, discount products are often even better tolerated especially things like cheese or cream cheese, which have very few ingredients. It’s really just the basics you need. The quality might not be the best, but the ingredients are actually ideal. Hardly any additives. So from a gut health perspective, it’s great. But of course, the data is missing. That’s always a bit of a shame. There is a company that issues barcodes, GS1, and they also have a database. It’s called Global Data GDSN or something like that. And technically, all manufacturers are supposed to upload their data there. But the issue is, access is restricted. So most branded products are generally visible, but many choose to opt out of the system and say, “No, we only share data with selected partners XXXXX XXXXX XXXXX,” not just anyone.  
 That means you’d have to negotiate with every manufacturer individually to get access, and that takes effort. And it’s expensive too, access alone costs 5,000€ per year. I’m currently considering whether to go for it or not. There’s a lot of missing data, but the available data is good. That’s kind of the main focus. And in terms of AI, definitely the topic of health... well, nutrition is already part of that space. I don’t want to brand it strictly as a “health app,” because I think food intolerance has also partly become a lifestyle topic. But still, data privacy is extremely important to me, especially because allergy-related data is really sensitive. Just feeding it into an AI like OpenAI, for example, is a tricky matter. And yeah, that’s the big question, how can we make this privacy-friendly? And with the new AI Act, who knows what kind of regulations are still coming our way.

A2

Right, you’re probably in that category too. I’m not even sure if you count as high-risk. It’s probably not a high-risk system yet, is it?

I5

I don’t think so, and I also don’t want to brand it as a medical app or anything like that just as a tool.

A2

Yeah, it serves more as a support function. Okay. So, let me paraphrase and check if I understood you correctly. The challenge is that there’s not always a solution that’s privacy-friendly or rather, there are many solutions, but you already know that data privacy might be at risk. Or is it not clear at all? It’s a bit hard for me to picture that.

I5

Yeah, that’s the main issue. Most of the time it’s just not clear.

A2

Yes, it's not clear.

I5

And often data is used to train the models, and then there’s the question of how GDPR-compliant that actually is, and things like that. No idea. Training in itself is a really sensitive topic anyway.

A2

Yes. So what would make your work easier right now? And if we… of course, if we had access to all the data, that would be an ideal scenario but maybe let’s get a bit more concrete, because no one can realistically make that happen. But what could help reduce these difficulties or obstacles, even just a bit?

I5

That’s a good question. Do you mean in general or specifically about AI? I guess it's still in general…

A2

Feel free to talk generally, and also about AI specifically.

I5

It's a good question. More staff is, of course, never a bad thing. Other than that, I’m currently doing quite well at solving problems. I usually frame them in a way that I can solve them, maybe that’s the trick, I don’t know. But when it comes to AI specifically… I feel like, and I’ve noticed this during teacher training as well, with AI, everyone kind of knows how it works: you train data, and it gets things right or wrong. People roughly understand that. And then the system continues to work with that learned training data. But the step after that is really difficult. I’ve seen this in my computer science studies too, it’s hard to really represent that part well. In schools, computer science is now becoming a mandatory subject in lower grades, and it’s expected to include AI as a topic. But that bridge between “here’s how it roughly works” and becoming a professional just doesn’t exist with AI. And it's really difficult to develop a proper AI system with just normal, basic computer science knowledge. So I think… that middle ground is completely missing. I don't know if it’s because you need so much data or because you have to specialize so deeply, but somehow there’s no such thing as someone with “medium” knowledge. You either know nothing, or you’re a total expert. That’s how it feels to me.

A2

Yeah, I feel the same way. I’m more on the “I have no idea” side. Not that I need to know, but I find what you just said super interesting. We have a question, I’ll rephrase it a bit: What do you think is important to know about an AI system to get a rough idea of what it can do? What would you say are things one should know about it? I’m thinking, for example, if you’re in a classroom explaining something to students but also in a work context. It could be any situation. Like at work, maybe you want to integrate some kind of image recognition system. What do you need to know about the model? Or imagine you’re in a school lesson, trying to teach students something about AI, what’s the bare minimum they need to understand?

I5

That’s a good question. I think people haven’t really come to a clear agreement on that yet.

A2

Yeah, like I said, there’s no real right or wrong but what’s your take on it? What would you say are important benchmarks? What should people be focusing on?

I5

So, just in general, I’d say you have to understand how data is trained, the process behind it and that there's often a human hand involved too. Like in India, for example, people who farm that data for little money. So, that’s one side, the training and then maybe also the mathematical part, like, if we take ChatGPT, how probabilities are calculated or trained, like which next letter or which next word is picked. That’s what I think is interesting, maybe also in a school context. Just building that bridge to understanding how it’s calculated. I think with text, it’s relatively easy to grasp because you can represent it pretty well. You can say which words are most likely, you look at lots of texts and see the word probabilities. With image recognition, it’s obviously much harder, I haven’t really dug into that yet. You’d have to go much deeper, with the image pixels, I’m not even sure what it’s called, but it doesn’t matter. In any case, you’d have to dive in much more and probably also separate pattern recognition from pattern generation. Yeah, something like that.

And maybe more concretely: if you have some AI model and you want to use it like in a company you absolutely need to know what kind of data it was trained on. Does it even fit my use case? Which data are trained? Because an AI without data is basically useless, may as well forget it. So I think that’s probably the most important thing. The way it technically works isn’t as crucial. But the data behind it that’s what matters. And what really annoys me is when everyone says, “Hey, look, we have 10 trillion data points” but if that’s all garbage, it doesn’t help at all. That’s the big problem. So I’d say quality is much more important than quantity.

A2

Are there any benchmarks for describing data quality? I honestly have no idea, sorry.

I5

Well, I’m not super deep into this either, but I don’t think there are at the moment. Right now, especially with, let’s take large language models as an example, they just take as much as possible because people think that’ll make them better.

A2

So, quantity over everything.

I5

Yeah, exactly. But if you’re mostly pulling from social media, you’re not going to end up with something like a scientific paper. That’s probably the big question, right?

A2

So now, let’s say you’re in a situation where you want to add a new feature to your app, and you don’t want to develop a model yourself, since there are already so many available. What kind of features or benchmarks would you look at when choosing something new? Or do you just try things out?

I5

Well, of course, you have to see what fits. I’d probably also spontaneously say: the costs.

A2

Yeah, totally.

I5

Because if you have a lot of users, it can get expensive pretty quickly.

A2

You're not the only one who’s said that.

I5

Yeah, it is expensive. I think the profit margins in this area aren’t as high as you’d expect from the tech world. Normally, the margins are huge. But because of all the computing power and electricity needed, margins drop significantly. And OpenAI and so on, they’re not profitable at all right now. There have already been the first quarterly reports, and they’re quite weak. So I think as long as it’s financially too expensive, or servers are too expensive, it’s hard to implement. For big companies, that’s probably easier than for smaller ones, because budget really matters there.

A2

Yeah. And apart from that, let’s say we’re in a perfect world where everything is Open Access. What would be the next thing you’d look at?

I5

Definitely what kind of data it was trained on, basically what it has learned. That there was some kind of selection process during training or fine-tuning. And of course, whether the output actually works or whether the AI is just bad. I think those three points are the most important.

A2

Yeah, cool. I’d like to move on. I think what you’re doing is super interesting. And from our perspective, we’re looking at how to communicate the capabilities of different models, now that there’s such a variety. And that’s where the idea of AI labels came from. You might’ve seen that in the email or the materials, I think you already had a quick look. I’d like to show you that again now. You should be able to see my screen.

I5

Yes, I see it

A2

I can also put it in full screen.

I5

Yes, sure.

A2

Exactly. I’d just like to ask: what are your spontaneous impressions? What do you see? Just describe it without judgment. If something is unclear or if you like something, feel free to say it all here.

I5

Yeah, so I think this is sort of a certificate for AIs. Like, instead of being on a washing machine, it’s on the AI model. Kind of looks like that too. Should I just describe it?

A2

Yes, just freely describe what you see. Or if something’s unclear, or if you think, “Ah okay, that’s how I understand this information,” just share that.

I5

So first, I see that it’s an AI label, when it was issued. I think the manufacturer and the model. ImageNet maybe, something with images, no idea. The name of the model is probably here. And then this bit looks a bit like the NUTRI-Score label, from A to E.

A2

Yes, that’s definitely what it’s based on.

I5

Exactly. So my first thought is of course: what does A mean, and what does E mean, so to speak? I think with food, I’d say A is good for me or for health, and E is bad. Which is also questionable with the NUTRI-Score, to be honest.

A2

You’re the expert on that.

I5

Yeah, exactly. Since the NUTRI-Score compares items within the same category, like cereals are compared to other cereals, if you're the best cereal, you get an A. That’s why it’s a bit... I think it’s a bit poorly done. And then… You probably know that more from washing machines and such, how much power consumption. So with washing machines, it’s power consumption, how long it lasts, and here it’s maybe how much power is needed, time for the interface, and then accuracy. I think based on tests. Overall, I find the lower part of these energy labels actually quite interesting, because they show real statistics.

A2

Yes, perfect. So exactly, I was just asking about your first impression. I think you already categorized everything pretty well. So here, the model would be MobileNet V3. And the dataset is ImageNet. I don’t know if you're familiar with ImageNet. I think it was in 2012 that the big challenge happened. They had this huge dataset of images. And the challenge was basically to categorize the images as accurately as possible. That’s kind of the idea, and I think it’s been running since 2012. And now it happens every year. I'm not even sure if the challenge is still ongoing. XXXXXXX would know that, for example. But yes, that's the dataset. That's the model. Here's the QR code links to more information about the dataset. You can actually look at what the data looks like. You mentioned that too. And this QR code is about the AI label. That’s the paper XXXXXXX worked on. It explains how it works, it's an automated label.

I5

So the reasoning behind it and the thinking process.

A2

Exactly, it’s an automated process, the label isn’t created manually. And how that all works can be found here. And you actually already interpreted this quite well, the ABCDE, that’s really analogous.

 We also spent a long time thinking about how to represent it. It should be intuitive, something people can understand, but also not just a Nutri-Score. At the beginning, it looked more like the energy label. We said, it can't just be about energy, first everyone called it an energy label, now everyone says Nutri-Score, but okay. This here is kind of the A100, that's the infrastructure, the graphics card it ran on, the same program, but even that can make a difference. If you’ve done the latest update, sometimes the values actually come out differently. And these are basically, so overall, this label is a comparison label, just like the Nutri-Score. So A, green, would mean that, compared to other models, it's the best model. And red would be a worse model in comparison. But we’d say that for accuracy, this is actually a pretty high accuracy. That dataset has thousands of classifications, and if you did it randomly, you'd get an accuracy of something like 0.5 percent. So 63% is actually quite good. But in comparison, it's still not great. That’s what it’s supposed to show.

I5

The thing is, it actually makes sense. Maybe the Nutri-Score isn’t such a bad comparison, because the idea of comparing within a single category makes a lot more sense than it does with food. Just because one muesli scores well doesn’t mean it’s healthy for me. But with AI, that’s exactly what you need. I don’t need to compare a text AI to an image AI. That’s actually really smart. Sorry, didn’t mean to interrupt, but...

A2

No, hey, it’s good, everything you’re saying is important. Exactly, that’s how it works. And I think, in theory, you probably could compare a text AI to an image AI, but then the comparison just doesn’t make sense anymore, right? Like, the conclusion becomes kind of meaningless if it just says something like “63%” here. Exactly. So, these are four dimensions shown here. I think you already mentioned two of them yourself, like, how much energy does one inference use? What’s the top-1 accuracy? I believe they looked at up to top-10 accuracy in total. XXXXXX analyzed that.

A4

Yeah, I think it’s Top 1, Top 5, and Top 10.

A2

Right, so different levels of accuracy. Then there’s “Corrupted Robustness”, how resistant or secure is the model against attacks? Like, if I change one pixel in an image, how accurate is it still? Or how much does it change if I alter, say, ten pixels? That’s what corrupted robustness means. And then there’s the running time per inference, how long does it take to spit out a result? Exactly. We just said this is a lot about comparison. That’s the background of why we thought about presenting it like this. And this here is an example comparison. Can you just describe, using the same idea, what you see here, what kind of associations you have, how you understand this information?

I5

I think we’ve already seen the one on the right, and the left one is the new one, right? I’d say the one on the left is obviously a bit worse, since it only has a C rating. And it’s yellow, while the one on the right has significantly better energy consumption and running time.  
 But the model on the left has much better accuracy. Yeah.  
 So I guess you’d have to weigh whether you’d rather use more power for better accuracy, or go for a model that maybe performs decently across all parameters. You probably always have to make trade-offs. If you give the model less power, it might be less accurate. But of course, if you allow more computation, it can perform better. So finding a healthy middle ground is key.

A2

Exactly, these values are definitely interrelated. You’re totally right.  
 And what we’re seeing here is that there aren’t just those four values going into the top-level Nutri-Score, I think there are ten in total. Don’t ask me exactly which ones, I don’t have them all in my head right now. But we could check the paper. We’ve got the QR code.

A4

I think model size is one of them too, or something like that.

A2

Right, and from those ten, the average value is taken, which then results in the C or the A. And that's how we arrive at the score. Is there anything you noticed now, something you still don't quite understand, or something that doesn't feel very intuitive?

I5

Now that you're explaining it, I'd say it actually makes sense.

A2

I should have asked the question earlier.

I5

No, but...

A2

Looking back, did you find anything a bit odd or confusing?

I5

I think... well, I think sometimes, like here, I have "Corrupted Robustness" and I wasn’t really sure what exactly that meant, just based on the wording. But once you know what it is, then it’s easy to understand, yeah.

A2

Yeah. You're not the only one saying that. Everyone brings up "Corrupted Robustness." I don’t think it’s that cool. Could you imagine a better way to present it, something more detailed maybe, or just different? You said you already found the bottom part important. Maybe more of that, maybe less? What are your thoughts?

I5

In terms of user experience, I’d say that if more power is required, the battery should also show that it needs a bit more power compared to when it’s green, for example. Like, the bar should be completely filled when it’s red or something like that. It’s just an intuitive thing, if you were to approach it as an app. Or with time, for instance, the pointer could be at the beginning for shorter time and far at the end for longer time.

A2

So an additional visual support besides the color.

I5

Yes, exactly.

A2

I think XXXXXXX thought about that too, but I believe with Corrupted Robustness he didn’t know how to implement it.

I5

If I were designing it as an app, that’s the approach I’d take. And apart from that, I think it’s not overloaded with too many elements, you can see everything at a glance.

A2

Yes, absolutely legitimate. I’d also say… XXXXX, is there anything missing? Did I forget something?

A4

I was just wondering… Is there anything generally missing on the label? Would you say there’s a piece of information that’s important to you but not shown on the label?

I5

Maybe what kind of AI it is. Like, in broad terms, what kind of output it gives. Like, is it an image? I don’t know if you can tell that just from the model, or if it could be made clearer whether it’s a text AI or a multimodal model or one trained specifically for something. So, maybe that.

A2

Yes, I think XXXXXXXX’s answer would be something like: based on ImageNet. If you know ImageNet, you know this is about image classification. But yeah, you need to be familiar with ImageNet for that. So yes, that kind of information could probably be visualized somehow as well.

I5

Right, and I mean the output, too. Like, maybe it just recognizes patterns, I upload a photo of a bird, and it gives me the name of the bird as text. Or maybe it actually generates another image. You know what I mean?

A2

So yeah, yeah, I got it.

I5

But the other thing would work too, of course. So both, input and output, so to speak, again.

A2

Yeah, got it. Right. Cool. Haven’t heard that before. Cool. At least I haven’t.

I5

Maybe the idea is nonsense, but that’s just something that came to my mind.

A2

No idea is nonsense. XXXXX, can I continue? Cool. I only have one screen, the interview guide, I should’ve printed that out. Right. So there are other ways of communicating about these systems. And we’ve just made a list of the types of ways that exist. I’ll walk you through what all of these are real quick. You probably know most of them. And then, in the next step, we want to think about the advantages and disadvantages, and where these AI labels fit into the broader context of communicating about models. So, sure, the classic one would be scientific communication. That would be, like, we just saw the model. The label was for MobileNet V3. Here’s the paper that goes with it. And then there are also Model Cards. You can actually find similar information there, like what’s shown on the label, just in table form. And that’s because the Model Card is based on the publication. So I assume you’re familiar with that. Papers with Code, I think that’s like a database. For example, let’s say ImageNet was used. On Papers with Code, you’ll find all publications that trained a model using ImageNet listed and displayed. So you can basically look at everything there is related to that dataset and what was trained with it.  
 And then there are also the classic blog posts. I think Medium is a very well-known outlet where a lot is written and explained about different models, also in text form. Library documentation, I’m not sure where he got that from. There are different libraries where everything is documented and listed. And then there are Factsheets, that’s something that only IBM does. And they only make them for their own models. So you won’t find them for just any model, but whenever they’ve developed something, they also release a factsheet. And in those, a lot is listed, what the model can do, and also, for example, bias, like how…

I5

Yeah, I just noticed that too, yeah.

A2

I think that’s also not very common. Exactly. Those are the communication formats we're familiar with. And now this label is being introduced. Where do you see certain advantages and disadvantages? For example, what do you find valuable and important in these standard formats that the label might be missing? And also the other way around.

A4

Maybe before we get into that, should we first talk about whether you’re already familiar with any of these communication formats in general, or if you’ve worked with any of them before? We actually forgot to ask that earlier.

A2

Good question. Thanks, XXXXX.

I5

That’s true, yeah. Let me think. I’ve seen scientific papers like that before, in that sense.

A2

Would be a shame if you hadn’t.

I5

Yeah, though with AI, it’s often not that scientific. I mean, it looks like it is, but then it’s just...

A2

Arxiv paper, yeah.

I5

Yeah, especially with all that closed AI stuff. And of course, blog posts, Medium and so on, I’ve seen those quite a bit. That one in the middle at the top.

A2

Model Cards

I5

Exactly. I’ve worked on something similar before. What’s also quite standard are all those benchmark tests, comparing different AIs, looking at their performance, what processors they use and so on. It’s kind of hard to tell the others apart just from the screenshots.

A2

Papers with Code. But if you don’t know it, that’s okay.

I5

So, I haven’t worked exclusively with something like that. And often, I’ve worked directly with the manufacturer’s documentation, so to speak. That’s usually different each time. Exactly.

A2

Yeah. Then back to my question. Where do you find any advantages or disadvantages?

I5

Well, what just caught my eye was the one from IBM, that actually looks pretty cool. I only know the first sections here, like in the picture, so to speak. But what I noticed, the way all the headings are arranged, sounds quite interesting because I think it’s just relatively short sections. Just, exactly, if it all continues like this, then you can read through it pretty quickly. And I think you also get, well, overall it’s nicely done, I’d say. I don’t know how it looks completely, but in terms of structure and design, you always have a good overview, and aspects like bias are also pretty cool. Because that kind of tells you something about the data, like how it was trained or whether only biased data was thrown in or something. So that definitely caught my eye during the presentation. In itself, of course, scientific papers are always difficult, I think. It makes sense to look at the individual ones. But if you want to compare quickly, it’s a huge effort to look through them, I would say. Most of them are not just one page long. You can read the abstract, but if you want to go a bit deeper, then it makes sense. And they’re obviously not written in a way that’s very engaging for structure, I’d say, they’re still scientific writing. And often that makes it much harder to quickly get an overview.

A2

So accessibility is limited there.

I5

Exactly, exactly. Definitely, because you first have to work your way through it. And blog posts are usually written in a somewhat more casual tone. I mean, the headline *Everything you need to know* is already a bit of that social media wording for AI, so to speak. That’s always nice for getting an overview. Compared to your AI label, what stands out to me is that it’s really just a simple and clear thing, not something where I have to read a lot, but I can immediately see the most important parameters at a glance, and how it compares to others, without having to look at bar charts with tons of tests and whatnot. Instead, the comparison is already done for me, and it definitely stands out in a positive way.

A2

You can also say, where do you see the limits when you look at such a comparison?

I5

I think it’s hard to compare because the six forms are very text-heavy and go quite deep in some parts. But scientific papers, well, they’re naturally meant to go deep, that’s the point, and that’s what people want. Now maybe what IBM and the blog post do is a bit more superficial, and that’s also hard to compare because the formats target different audiences. If I think of IBM, the customer is someone who just wants to buy or use the product, while scientific papers aim to inform the research community or let them use the findings further. The target groups are probably extremely different, so you can’t really compare them directly. For every use case, you need a different format anyway.

A2

For which target group do you think such an AI label would be useful?

I5

That’s a good question, I’ve wondered about that myself. I’d say it’s probably definitely for people who want to use AI. For example, with OpenAI or Google’s text-based AIs, you can directly select models, there are tons of models, like ChatGPT 4, 4.0, mini, and so on, and the same with Google in many variations. Having a quick glance at a label helps me understand how much power it uses, how accurate it is compared to others, so I can quickly decide which model to look at more closely, especially if I want to work cost-efficiently or need a lot of power or want to have it available, whatever it may be. So I think it’s perfect for beginners to quickly see which model to focus on first. That’s how I’d define the target group.

A2

Yes, would it also be relevant for you in your work right now, or would you say you already need deeper information?

I5

Well, I think it wouldn’t be bad, especially if you just need a simple text AI or something that summarizes something, like for tasks in the app that summarize user entries or something like that, it would be interesting to see that at a glance. That’s actually something I struggled with, you really have to dig into all these models to figure out which one to use, because the model names themselves don’t say what they do or what they are capable of. So you have to read up elsewhere to find out which one is better or something. So, for a beginner, it’s just simple. What I noticed just now is, for example, I once thought about which AI can translate best, so I just typed into Google “which AI translates best,” also in English or something. And it’s hard to find benchmarks that really try contexts or tasks, because often you have school tests that get benchmarked or some AI tests, there are tons being developed and tested, but use cases like translating well or creating recipes or whatever are hard to benchmark. So you basically have to try it yourself to see what fits well, but you can’t figure it out from school tests or anything like that.

A2

Yeah, well anyway, the benchmarks, if I understand correctly, are always on the test dataset, like how well does it perform on the test dataset?

I5

I think in the beginning they mainly took school exams, as far as I understood, and then at some point they also developed other special AI exams.

A2

Yeah, cool. I have to check the next slide. No, we don’t need that yet. If you now imagine a label “in the wild,” what would be important? Who should issue something like that? Would you generally pay attention to that, or is it somehow, whether it’s TÜV, a university, or the company itself maybe not so important at first, but the main thing is that there is some kind of label? So who creates something like that? Who should create it?

I5

Well, I mean, I’m thinking about… Oh, my headphone just fell out… I mean, with the NUTRI-Score, I think each manufacturer creates it for themselves, so there’s no institution calculating it. I think with energy labels, I honestly don’t know much about that, but I think those are also issued by the manufacturer, if you just state maybe how much… It might also be that TÜV checks it a bit, I don’t know if you know more about that. But I think mainly the manufacturers do that, and my idea was, since you said they are automatically created or can be automatically created, maybe there are open source tools that anyone can transparently download and apply to their models, and then you can directly issue this kind of certificate. That’s probably easiest because then you don’t need an extra organization or whatever to issue it. But then the question is how well it would be accepted, which is also harder. And if it’s open source, then everyone can do it and try it out directly.

A2

So, Open Source would mean that someone from the scientific community created it, and then people can use it.

I5

Exactly, and the industry can then use it. Or of course, also private individuals. I don’t know how complicated this creating process is, I don’t know that yet. But theoretically, private people could also use it. I don’t know if that works, to also control it, so to speak.

A2

What do you mean by controlling it?

I5

Well, the manufacturer runs this test with their own products. But users or AI users could also test it for themselves. They would download this AI labeling and then test it with the AI, so to speak.

A2

Ah, gotcha. To double-check that what they showed me is really true. Okay.

I5

Well, that would of course create transparency, and you could quickly check for yourself, ideally, if it works as I imagine.

A2

Yeah, I actually don’t know, but I like the idea. Okay, so if it’s somehow Open Source, or maybe even a company created the AI label, like if IBM offered something like this, would that make a difference compared to if it’s from XXXXXXX at XX XXXXXXXX?

I5

Yeah, good question. For me, for validity, it probably doesn’t matter. But practically, it might matter if someone from XX XXXXXXXX issues it, or a big company or something.

A2

Okay, so as long as it’s Open Source, it’s transparent, and that’s enough?

I5

I'd say yes

A2

Yeah, yeah. No right or wrong. Cool, XXXXX, any other questions? Do you have anything else on your mind?

A4

Yeah, how about a general certification, like making it mandatory? Do you know what I mean?

I5

Yeah, like from the lawmakers, right? Yeah, that’s possible. I don’t know how the EU AI Act looks regarding that, whether something like that is planned, I have no idea.

A2

It is planned. Certification is planned.

I5

Yeah, well, I mean, if it’s really just some kind of AI label, it can’t hurt, unless maybe you want to hide something. But I think such a label is of course a good compromise. So manufacturers don’t have to fear that they’ll have to disclose something with closed models that they might not want to. There is already enough information on the site, as consumer protection. Exactly. But I think the interesting thing is really only if it shows something, it’s kind of a B2B label, so I think consumer protection is different here than in B2C, or towards the customer. Because this label strongly resembles normal consumer protection labels, I’d say. And the question is, of course, whether something like that would actually be law, or if something like that is not legally required at all, but only applies for business customers and transparency. That’s what I’m wondering about right now.

A2

I think that’s basically how it goes in general. So, AI systems should be certified. So, does ChatGPT, which you have on your phone, need a certificate or not? And if yes, who should issue it? Those are all questions that come up. And this is a bit separate from the AI label.

I5

Yeah, right, yeah.

A2

XXXXX?

A4

I think that’s it.

A2

Do you have anything else, XXXX? So, we’ve kind of covered our questions pretty well, but maybe we totally forgot something here that you think you’d like to get off your chest?

I5

I don’t think so, I told you everything important.

A2

Cool, then thanks for participating. I’m going to stop my screen sharing now, and I’ll stop...