A3

Perfect, and we'll start with the first question. Well, it's not really a question, but please introduce yourself briefly.

I3

I'm XXXXXX XXXXXXXX, I'm a student at XX XXXXXXXX, studying computer science in a dual study program. At the same time, I'm also training as an IT specialist at XXXXXX XX, so yeah.

A3

What's your background, your career path? It’s probably not that long yet, but what did you do before?

I3

Yeah, so I’ve been studying computer science for four semesters now. Before that, just regular school-level computer science, a bit of hobby programming, but nothing else really like side jobs or anything.

A3

and probably graduated from high school.

I3

Exactly, A-levels. So that's what it's about, yes, A-levels, quite normal.

A3

Ok, which company do you work for? Can you say anything else about XXXXXX? I've seen that

I3

Exactly. XXXXXX is now a very large software development company that doesn't specialize in any particular area, but works in many different sectors. I am also currently employed in the cross-industry sector. So I don't have a focus on any one industry, I work wherever work is needed.

A3

Okay, Cross Industries, that's like a department at XXXXXX, which is really addressed by everyone. Is that how I can imagine it? Okay, good. Then why don't you briefly describe your role in the company?

I3

Exactly, at the moment I'm a dual student, which of course doesn't mean a big management role or anything. I'm just a software developer, I think I'm now working on my fourth project, mainly internal projects, an administration for us as students, then an administration for the entire car rental business and everything at our company. And right now I'm working on an employee database that contains information about what the employees do and I'm using AI for this, among other things.

A3

Okay, that brings me to the next question. So what are the technologies that you work with? I'll say frameworks, programming languages and most importantly, of course, what kind of touchpoints do you have with AI and machine learning?

I3

Yes, at the moment I'm mainly working with Spring Boot and Angular, so frontend, backend, fullstack and for AI use at the moment LangChain4j, because we then integrate it into Azure OpenAI, so ultimately into ChatGPT, backend and then we work with ChatGPT in there. Otherwise, if possible, I've also learned React in theory and Java quite a lot through university, but mainly Spring Boot with Kotlin.

A3

Okay, can you say again the name of the machine learning model or whatever you mentioned before Azure?

I3

LangChain4J is a library that summarizes the various APIs. In other words, it's an API library so that you don't have to integrate a separate API for each model.

A3

Okay, I actually didn’t know that yet. And does 4J stand for Java?

I3

For Java, exactly.

A3

Exactly. Ok.

I3

It's actually a JavaScript library that was then ported at some point. Although it's not exactly, so it's not a real port, it's actually a further development. I think there are a few other things in it, but I only use it. I didn't develop it.

A3

All right, okay, that makes you a good candidate for this interview. What other previous experience do you have with AI? Maybe this would include whether you've had a course about it at university.

I3

Yes, I already started, I don't know, how long ago was it, about four years or so, before I began studying computer science, to get into PyTorch and all that a bit. I tried to program a few small AIs myself, just the typical things like distinguishing between dogs and cats, nothing big, but a bit of an introduction. At university, I currently do not have a module on that. However, I am writing my seminar paper on the use of AI for creating graphical user interfaces, but I am not developing anything myself, just looking into what possibilities currently exist.

A3

Okay, got it. How exactly do you work with AI? Maybe you could describe a workflow, a typical one that you often use when you interact with an API?

I3

Yes, so the API is very nice and very simple. In the end, it works like this: I can, just like with normal chat clients, enter a system message that gives the context for what the AI should do. Then I can enter a user message, to which the AI reacts accordingly. At the moment, I am not using the AI in a chat context, but rather in a way where I give it a text and the AI is supposed to analyze whether certain data is contained in that text. I really just have to provide two strings, and they are processed directly, so it is very pleasant and very simple.

A3

What is the use case? Of course, you don't have to reveal any business secrets. Or what would be an example from your current projects, for instance in fleet management?

I3

Yes, so for the AI or in general?

A3

No, for the AI. Now more so...

I3

Yes, it is only for this employee management. It compares different projects to check whether the same technologies and so on are being used, because back then it was a very old project and they did not think that far ahead. That means instead of properly creating things as classes with a proper ID and everything, they just stored plain strings. And because of that, every time you order a project, the technologies are added again, and sometimes someone writes Java, sometimes Java with this version, sometimes Java with that version. So it is just a summary of all that. And yes, mainly it is about checking whether strings match, do not match, or roughly mean the same thing.

A3

Okay, all right, interesting. Definitely a good insight, thank you very much. What would you say are the biggest challenges you have encountered so far in using it?

I3

Yes, the consistency of the results, meaning that if I input the same things, I should always get the same output. Personally, I do not have major issues with it because I do not generate that much, but I notice with some colleagues that they have the problem that the AI often hallucinates something. For example, when you feed it data and ask it to generate a description based on that data, it makes up additional information. And it is important to prevent that.

A3

Okay, all right. How would you, or can you imagine how this challenge could somehow be solved? Especially from the research side in AI models or also in the communication of AI models, for example, what kind of information might still be needed.

I3

Well, if I knew the solution, then we would not have these problems anymore. I can hardly see a solution for the terminology, because AI, the way it works, at least in my understanding, should theoretically be deterministic if you always give the same feedback. But that does not really work, especially with all the LLM models and so on. And for those who don't hallucinate it, you would theoretically need two or three more AIs to check whether it fits together, but I have no idea to what extent that makes sense.

A3

Okay, to follow up, do you sometimes have to choose between different AI models during your work?

I3

Not directly. Since I am not paying it myself, I have to work with what is paid for. Right now, I only have access to the OpenAI models and could choose between GPT-3.5 and GPT-4. In the end, the decision was made based on the cost, so it became 3.5. So theoretically, I did have a choice, but it was not really my own decision.

A3

Okay, I would like to go into that. Let's say money does not matter in this case. What kind of information would you have wanted from the two models to help you decide between them, regardless of the cost?

I3

Yes, there are now quite a lot of benchmarks that test different things, especially when it comes to knowledge in the Dark Dynamics models. It would have been interesting to see how they perform there, and otherwise just some kind of measurement of reliability, general understanding, and to what extent the AI does what it is supposed to do. But I do not know exactly how the benchmarks work for...

A3

Okay, which benchmarks come to mind for the first case?

I3

I absolutely cannot remember the names. I just always see the nice tables where it suddenly says something like 80 percent more or something like that.

A3

Yeah, okay, all right. Good, now I finally have a good transition to the part where I can also show you something. So, I hope you can see my browser. Yes. I would like to make this bigger here. Did I manage to do that last time? I think here. Can you see it in full screen?

I3

Yes. Now I see it in full screen.

A3

Okay, perfect. So, we came up with a solution for this admittedly fictional problem of yours, since you do not often have to choose between models. For this problem, we developed a solution we call AI Labels. And without influencing you any further, I would first like to ask you: what is your spontaneous reaction to this label? What do you see there? What are your impressions?

I3

I see a lot of data. Now I am tempted to scan the QR codes, but in the end there is probably just some nonsense behind them. No, here we are. Yes, a lot of information. I see many different pieces of information. So first of all, there is this ABCDE ranking, which at first gives me the impression that this is a very good AI, according to the label. Below, the different images also indicate something related to power, power consumption, power draw per inference. I cannot immediately picture what corrupted robustness means, but also good, the accuracy. And the running time, meaning how fast it responds.

Then it seems I also have some information about which GPUs are used for it, meaning what it is running on, if I interpret it correctly. Is that A100x8? Right. And then the model behind it, or rather the technology the AI is based on.

A3

Yeah okay, any other impressions on how you find it maybe?

I3

So it looks like I'm looking for the washing machine at the DIY store or Mediamarkt. Very consumer-friendly, if what it says is true.

A3

Yes, that's actually true. I can go into more detail about where this label comes from and how we came to these results. I see the world is ending outside, but we'll ignore that. Do you see a bridge to the problems we may have just mentioned?

I3

Yes. Definitely.

A3

in what form?

I3

Yes, I have the top-1 accuracy, the corrupted robustness, I would... what is that actually? Could you explain that?

A3

Sure. Yes, of course. Okay, I can briefly go over the AI label. First of all, it’s important to note that these results, including the color coding, come from various experiments conducted by our colleague XXXXXXX, who unfortunately isn’t here today. He tested different AI architectures, which you can see here. In this case, for example, it would be the MobileNet V3 Small. This is a specific architecture of a neural network. He tested several of them and created various benchmarks, as you correctly mentioned earlier. There are always many benchmarks for different AI models. By comparing various architectures, we arrive at these classifications. In this case, he focused on image classifiers, which you’re probably familiar with, such as ImageNet. Have you had any experience with that?

I3

No, not directly. I have Stable Diffusion running at home, but I don’t use it actively.

A3

Okay all right. Exactly, in this case it is just image classifiers that were compared with each other and we looked specifically at inference and yes, you have already correctly identified the Power Draw Per Inference, so where data shoot through the neural network, just as many milliwatt seconds of power are consumed. Then the accuracy, that's probably what you meant by the benchmarks, simply the accuracy of the best result of the AI model, so how often is the classification ultimately correct. Concerning ImageNet, we are talking about a thousand different classes that can come out and you can imagine that if you just picked a class randomly, then we would have an accuracy of one thousandth. Because then it would just be completely random. Therefore, this is how these 63% are to be read. Relatively high, but it's still one of the worst accuracy scores in comparison. Then what you just rightly asked, the robustness, that's another benchmark. It's about how robust the AI model's decision is against changes in the image. Meaning, you somehow have an image of a dog and make various pixel changes in the image and look at how many pixel changes it takes for the AI model to change its prediction. So how many pixel changes does it take for the AI model to change its prediction. So from...

I3

So, something like random image alteration.

A3

Exactly, so yes. Or rather, yes, I think in this case he did it by accident. Of course, you can also deliberately attack it, which maybe also says something, meaning you can consciously change pixels in a way that, for example, a classification from dog to specifically cat is forced. These are the classic inference attacks. Exactly, and this serves as a benchmark for how robust the model is against attacks. And here we also have the running time per inference, meaning how long it ultimately takes. Yes, you already mentioned it correctly, the hardware and framework that were used. Exactly. Did I forget anything? I hope not. Yes.

A4

The QR codes perhaps?

A3

The QR codes. Yes. You can scan them right?

I3

Yeah, I just checked, some study and a GitHub, I think.

A3

Exactly, this is the study that ultimately compared these different models with each other. It is also important to mention that we did not only use four benchmarks. XXXXXXX actually used quite a few more. And the combination of all the smaller benchmarks leads to the overall score. Okay, good. That was the explanation of this AI model. Now let’s have a look at what we have on the next slide. Yes, good. The use case is briefly shown here. So if someone, like you, uses APIs, it might be interesting to look at something like this before deciding on a specific model. Now, let’s move on to the next label. Can you briefly summarize what you see here?

I3

Yes, I am currently seeing two labels from different image recognition AIs. One is the EfficientNetB4 and the other is the MobileNetV3 Small, which we already talked about. And now I can see quite clearly what differences there are between them.

A3

Okay, can you interpret the differences? Maybe just go with the flow?

I3

Yes, so the model on the left, the new one, initially has the worse rating with a C compared to an A, which stands out at first glance. If you do not go through the values one by one, it has a significantly worse power consumption, much worse power consumption. It has slightly better corrupted robustness, better accuracy, but a clearly worse response time. And now I could weigh what is more important to me: having the better performance score or saving more power and time.

A3

What is your opinion on that? Not in general, but...

I3

The question is what you need. So if it is something specific, like for example assigning employees to projects, I don't know, then I would probably choose the one on the left, because better accuracy means I get better results. If I am at home building something like my Raspberry Pi, a robot, Gladys, or whatever, then the one on the right would probably be enough, because I don't have the power or the time to do a lot of things. In that case, I just want to get quick results, and if they are not perfect, that is not a big deal. So it really depends on the use case.

A3

Okay, the use case. Maybe to come back to XXXXXX again, can you imagine that in a few years you are the boss and you have to make decisions more often about which AI model should ultimately be used, for example in your case which large language model the developers should use. In that context, could you maybe outline a scenario of how you would make a decision using the information from this label?

I3

That will of course mean that I have to build a cost-benefit analysis for myself. How much would I lose through wrong decisions made by the AI? How much would I gain? Right now, I really cannot say which one would be better suited. I am currently also using 3.5. I have colleagues who use 4. I think it is a matter of personal preference for now.

A3

Alright, okay. Good, now I have here, how do you understand the information presented by the label? Do you perhaps have any detailed questions about it, especially in the context of the comparison?

I3

I have to say, this is actually presented very well. So now, after you explained the two benchmarks, there's absolutely no problem understanding anything.

A3

Okay, all clear. Yes, good, then the next question is: Can you compare the labels and...

I3

Yes, I was just in the process of doing that, so it can be compared very well.

A3

Okay, maybe asked differently, what do you find particularly helpful? What stands out to you positively about this?

I3

Especially positive, because… I think it’s pretty cool that the icons not only change the value underneath, but also the color of the icons themselves. This means I don’t just have to check if the number is larger than the other one. 1.6 is actually a bit smaller than 608 but because the battery is yellow, I notice something might be wrong, and I see, oh, milliwatts and watts, well customers do. Otherwise, yeah, this big ABCDE, very prominent, would actually suggest to me that I should take the one on the right, that’s definitely the better option. How helpful that ultimately is, I don’t know, because it depends on the details.

A3

Okay, what additional details would you like?

I3

No. They are given there. So the top-one accuracy and the corrupted robustness are also there, but what matters to me is mainly the results and not the power consumption and the seconds. It's always about the weighting you apply there.

A3

Okay, what do you find particularly difficult to understand, what might seem cryptic?

I3

It would probably be better to interpret it as this large letter, which is weighted differently, compared to how I would weight it, probably.

A3

Okay, sure. So do you feel that you are somewhat dependent on the creator of the label, meaning on what priorities they have?

I3

Exactly, it simply has the bias of finding certain things more important than other values and then assigns the label accordingly, although you can also look at what the label is actually trying to tell you.

A3

Yes, okay, do you maybe have something specific in mind right now on how the label could be designed differently to perhaps solve this problem?

I3

I'll just give an example. You could, of course, simply put several of the rankings up there, so that you have three bars: power efficiency, response rate, and all the other aspects combined, like how accurate the model is. This way, you would have two or three different factors visible at first glance and could then decide which one you want to focus on, instead of first seeing just an overall good or bad rating and then having to dig into the smaller factors.

A3

Ah, okay. I would like to follow up on that. As I said, there is no right or wrong, it is really just about your impression. But how would that then relate to the small labels that are already on it? You mentioned that we have...

I3

Yes, you mentioned that you have many more benchmarks behind it, so that these four labels could be treated as individually weighted labels instead of a single bar, meaning you would have four bars and then, below that, a lot of very small labels. The question is always to what extent this is actually beneficial for the end user, because there would be a huge amount of data to work with.

A3

Okay, yes, I understand. Cool, thank you very much. Definitely a good suggestion. Then I would make a small jump now and ask you: what other forms of communication from AI models come to your mind? Basically, any kind of displays of AI models that you can look through, for example before you use the AI model.

I3

I have never really dealt with that. In my current project, I simply had the list of different AIs with their prices displayed in the Azure portal. And of course, you can also google LLM rankings and find a lot of different studies or, I don't know, advertising websites that say 'please use my program.' So right now, there is quite a lot out there, and you never really know which ones you can trust. Everyone claims that their AI performs better than the others in the benchmark.

A3

I see, okay. There are a few that we have found. Let me quickly click here. So, we have identified these six. I would like to briefly introduce them to you. These are simply different forms of communication for AI models. First of all, we have the classic publication. Many AI models were developed as part of some scientific work, and of course, you can always read a paper if you feel like it. In this case, it is about MobileNetV3, for which we already had the label, created by Google researchers. So that would be one form of communication. It also includes benchmarks, for example. Then we have the Model Card. In this case, I believe it is the Model Card from Google, which also presents various pieces of information about the models. For example, ImageNet, which we also had as part of the information, or what types of inputs are used by the model. You can also see benchmarks here at the bottom, on different hardware environments, or no, in this case, on different runtimes. That is the difference here. Okay. Then we have Papers with Code. I don't know if you have heard of it before.

I3

I haven't yet.

A3

Okay. Maybe for your next seminar paper or something like that, this could be quite helpful. Up here, for example, you can see an AI model, in this case again MobileNetV3. And down here, you would have to scroll down, there is a list showing all the papers that cite this model. So it is a detailed listing of everyone who either uses MobileNetV3 in their work or takes it up as inspiration in some way. Then we have blog posts. You might have come across medium.com or Towards Data Science before. They are also popular. They are usually written by scientists or, in any case, by professionals. They often contain continuous text, but also benchmarks. Then, of course, we have the library documentations, for example from the Python library or TensorFlow. And finally, we have fact sheets. These are from IBM. These are special pages that cover certain categories, which you can see here on the left. For example: bias, how fair is this model? What kinds of inputs and outputs does the model produce? So it is very detailed, you could say. However, these fact sheets were developed only for IBM's own AI models. So, this is what we found out. Now I would like to ask you: What advantages and disadvantages do you see here, spontaneously? What stood out to you, especially in comparison to the AI label we showed you?

I3

With all of them, you have to read a lot. So everything is somehow much more text, much more data, and the label presents this much more practically, much more conveniently. Sure, with some, you always have a table included where things are listed, but none of them are as easy to understand as the label.

A3

That's of course very positive. You can also feel free to be very critical. Don’t you see any disadvantages, maybe something that the AI label does not provide compared to the other forms of representation?

I3

Yes, especially the paper and the Papers with Code, and everything else, obviously provide much more information than the label. So, when I read a paper about an AI, I can probably understand it much better than if I just look at the label. And also blog posts, where there’s always the risk that the author might add their personal opinion, will convey much more information than the label, as long as it’s well written. Yes.

A3

What kind of information would you specifically hope to get from such a blog post? What do you value in this personal opinion?

I3

Then you can take another look at what this AI was actually built for. Of course, image classification is nice, but what ideas did you have in mind, whether you built it for your self-driving car or something else, you will get more information from that, and then you can check if it aligns with your use case or if you want to do something else with it.

A3

Okay, again, so you're also a developer. If, let's say, a situation comes up where you have to decide between two AI models for something like image classification, what do you see as the pros and cons in terms of programming as a developer? Maybe you can think of something additional here?

I3

Yes, I definitely need to take a look at the documentation to see how I can integrate it and whether I can even use it with what I am doing. I will probably also need to take another look at the paper, for sure. Especially the one from IBM would probably be very useful, even though it is only for their models. But from what I just saw, there were some very interesting things included that I definitely need. So yes, everything definitely has its justification for being there. Now the only question is: what do you want to focus on?

A3

Alright, okay. Then, if you have shared all your impressions, I would move on to the next question. It’s a small jump: In your opinion, who should issue such labels?

I3

That's a very good question. Definitely not the publisher themselves. Ideally, it would be done the way most labels in Germany are handled, by having some kind of independent institute. Although there is always the question of how independent these really are. Stiftung Warentest, for example, is already a well-established institution here. Whether something like that should be nationalized is another question, especially considering the current attitude of the government, or rather, the political landscape, towards IT. In theory, you could also establish some kind of standard commission to decide on this. But for that, all the companies would have to agree first. So it definitely has to be something centralized, and not everyone should be allowed to create their own label."

A3

Could you go into more detail again about what exactly the advantage of this centrality would be, maybe in relation to transparency?

I3

Yes, everything is evaluated according to the same types, and not in a way where one person says, 'I focus on giving answers as quickly as possible,' and another says, 'I focus on consuming as little electricity as possible.' Otherwise, everyone could adjust things so that they always end up with a big green number, simply by arranging everything so that the important aspects are emphasized, while the poorly performing ones are scaled down so much that they hardly show.

A3

Couldn't you also critically ask whether, with a central authority for example, something similar could happen, like with the Nutri-Score, where it still ends up…

I3

Of course, that is possible. That is why you would have to make sure that these labels are set up in such a way that, if the score is good, everything is truly good, and that you cannot somehow bypass the system by adjusting certain things. With the Nutri-Score, for example, there are seven groups overall. So with AI, you would also have to be careful not to compare apples and oranges.

A3

Yes okay, absolutely fine. What do you think in general about the idea, maybe also the obligation, to certify AI models?

I3

That's rather difficult because I can't, as just any developer uploading their fine-tuned Stable Diffusion model to Hugging Face, deal with all sorts of certificates. That’s a bit complicated. But of course, as soon as you start offering things, it would be nice if you could do that. Although I rather think that the label has to establish itself and be recognized, rather than people being forced to use the label. You could use it nicely for marketing, like saying, 'We have Label A, and the others only have Label C.'

A3

Ah, okay.

I3

Test winner at Stiftung Warentest and so on, are exactly the same advertising promises.

A3

So, voluntary self-regulation, that's what it means.

I3

Exactly, and if they don't have it, then they probably have something to hide.

A3

Okay, yes, interesting. So you would find it helpful, maybe also in relation to your job. I know it is a bit difficult because you are not yet making these decisions yourself. But in general, would you find it helpful?

I3

Yes, absolutely.

A3

Okay, good. Would the label still need to be adjusted in some way, for example to fit your specific background knowledge? I hadn't asked that from this perspective yet. Should there perhaps be different forms of labels?

I3

So the label definitely needs to be adjusted. Half of the QR code is missing. But no, theoretically there should also be a table at the bottom or on the back, where all the detailed results are listed, so that you can directly view all the key values through the central testing authority, without having to go into the paper first. But apart from that, based on what I would base my decision on, everything necessary is there.

A3

Okay, so a kind of opt-in where you simply get more information. Can you imagine that the other direction, that someone would initially want less information, could also be something people want? Maybe among your colleagues, for example, or among people who have no background in computer science?

I3

I mean, a simple ranking system, really just A to E, is of course much nicer and easier. And they probably also won't really know what to do with information like what accuracy, robustness, or power draw actually mean. Is 600 milliwatt-seconds a lot or a little? You probably don't need that there. And which graphics cards it runs on is probably pretty irrelevant to them. But especially when you move towards developers, it becomes quite interesting.

A3

Wonderful, I am actually done with the interview guide. XXXXX, did I forget anything?

A4

I actually don't think so.

A3

I think we also weaseled through it relatively smoothly.

A4

The only thing that comes to mind, and unfortunately I forgot to ask at the right moment, is that you placed a strong focus on Top-One Accuracy. We also tested Top-Five, for example. Would that change your decision, if you consider it in relation to the power draw?

I3

I don't have enough direct knowledge about image classification to say how useful that would be. I just need some kind of value that tells me how often the AI makes the correct decision that we want. Whether that is this 1% or 5%, I would have to take another look beforehand to understand what it actually means.

A3

So, the Top-5 accuracy basically just means that among the top five results, or the five results the AI considers most likely, the correct result, meaning the correct class, is included. This means that the Top-5 accuracy is always higher than the Top-1 accuracy, because the Top-1 accuracy only measures whether the best result is also the correct result.

I3

Okay, that makes sense. There’s also the use case and everything, but then the top 1 accuracy is definitely better suited, especially since that is really the golden value you want to have.

A3

Yes. Ok. Yes, thank you very much. Wonderful.

A4

Then I would stop the recording.

A3

Yes, I think we can do that according to the guide.