

Eurostat: Stats in a Wrap

Innovative approaches in statistics – Part 2: Chatbots and other AI usage

Release date: 18 September 2024

SPEAKERS

Albrecht Wirthmann (Eurostat), Nikos Roubanis (Eurostat), Frankie Kay (Central Statistics Office of Ireland), Eva Charlotte Berner (Statistics Norway), Jonathan Elliott (host)

Jonathan Elliott

Stats in a Wrap, the podcast series from Eurostat.

Jonathan Elliott

Welcome to another edition of Stats in a Wrap, the podcast all about statistics from Eurostat, the statistical office of the European Union. This is the second of two special shows in which we dive deep into the work of innovation at the European Statistical System.

This is the family of organisations all working on official stats in the EU and the European Free Trade Association, or EFTA. We saw in the last podcast how real-time shipping information from a safety agency could be converted into top quality official statistics with far greater speed than could have been imagined possible just a few years ago.

Now we'll be looking at how chatbots at Norway's national statistical Institute - Statistics Norway - could help its statisticians and the Norwegian public get really useful information easily and quickly. So, let's welcome the guests who will be guiding us. First of all, Albrecht Wirthmann. He is the head of methodology and innovation at Eurostat. Albrecht, welcome.

Albrecht Wirthmann

Thank you, welcome to everybody.

Jonathan Elliott

And joining him is Frankie Kay at the Central Statistics Office in Ireland. She heads up the one-stop shop in innovation in artificial intelligence and machine learning, which is a four-year project which has just started. It is an initiative by the European Statistical System, built up as a consortium of NSIs, promoted by Eurostat and funded by the EU. Frankie, thank you so much for joining us again.

Frankie Kay

It's lovely to be here again.

Jonathan Elliott

And we're also welcoming back Nikos Roubanis from the transport unit at Eurostat. Nikos, thank you very much for joining us again.

Nikos Roubanis

Thank you again for inviting me.

Jonathan Elliott

And I'd like to introduce Eva Charlotte Berner, who is at Statistics Norway and heads up their artificial intelligence unit - Team AI.

Eva Charlotte Berner

Thank you for having me.

Jonathan Elliott

Well, it's great to have you here, and I'm looking forward to talking about the amazing work at Statistics Norway shortly with you, Eva. But Albrecht, if I may first come to you.

We heard in the first of these two special podcasts about how the ESS (that's the European Statistical System) Innovation Agenda, as it's called, is pulling together initiatives from all the participating members and hoping to coordinate them into a common framework that has AI (artificial intelligence) and non-AI elements, but I have to ask: isn't AI now the most important technology being offered in innovation and statistics? Do you find it pretty much inserts itself into almost every field?

Albrecht Wirthmann

AI is certainly a very, very important ingredient to innovation, but it's not the only one. However, AI is quite a transversal technology that might influence many aspects of statistics and statistical work: so, classification, producing new outputs, making the processing of statistics much faster to get an overview of vast data sources or data objects, but also communicating statistics. So, it's very transversal, and in this way, it's quite important.

Jonathan Elliott

One of the interesting things in the policy thinking that goes on at the Innovation Agenda is that it is user need, so called, that is driving a lot of this innovation. Government departments, academics, journalists, the general public are just more and more information hungry. And it's hardly surprising in a society that has more and more information available to it. So, it's fair to say that the Innovation Agenda is not just doing innovation for innovation's sake. It doesn't exist in some sort of bubble or laboratory. It's doing it because there's a crying need for it.

Albrecht Wirthmann

We see that society, technology is very much changing, and it creates this expectation of immediate information at hand at the time when you need it. So, when you go on the internet, for example, search for something, you expect that you would immediately have the information that you are searching for. This is also true for statistics. Policymakers ask about certain developments that in the society, in the economy, are ongoing, and there is a pressure on statistical offices to provide faster information.

And people are aware that more and more information is collected through technology that is there. Everything is digitised. Data is collected as you are doing things. The demand is there. It's the art of bridging it. You have to put data into a context, to connect it with other data, and of course, also have the ambition of producing high-quality data.

Person on Street 1

To trust the data, it would be the best to have a human oversight, a human to just look at the results... just as a secondary check. But again, I do believe, because the data is sourced from such a wide, wide variety of places, that it's probably pretty good.

Jonathan Elliott

I was going to ask Nikos a bit about that, because this is very much his world. It's one thing getting data, but ensuring its high quality, whether for training AI algorithms or using in statistics, it's quite another. You learned from your project with maritime safety signals that turning data into statistics that you can really use and that can be really relied upon is quite a challenge. Did you expect the difficulties that you confronted with that to be there?

Nikos Roubanis

The more you work with innovative data, the more you realize that it's difficult to come up with so much specific information on what you are looking for. A lot of innovative sources produce data which are not intended for transport statistics.

They are intended for communication. They are intended for safety. It is very important first of all, to define what you are looking for. And what you are looking for is something that will inform the public. It will inform how policies change and how they affect our life. And this is why we need this data, because besides improving timeliness, they can also give more detail, but they can also produce information where surveys are very expensive to carry out, for example, statistics on cycling.

We need to evaluate how people are using their bicycles and how much and what could be the alternative, and it would be very difficult to carry out passenger mobility surveys with such large samples as to be representative for evaluating distances.

That is where artificial intelligence could give us a hand to pass from disaggregated, basically economics statistics that involve transactions or surveys, into mobility statistics, into making meaningful indicators out of data which are intended for other purposes. That is where I see that artificial intelligence could help the production of European statistics.

Person on street 2

There's a built-in chatbot with Snapchat, and you can ask it simple questions, but I didn't really use it for, like, work. I just, like, played around with it.

Person on street 3

I use ChatGPT for university. So usually for research questions, or just when I have to understand.

Jonathan Elliott

Frankie, this challenge of making data that was not originally collected for official statistics and to make it fit for purpose with AI, is critical, isn't it, as a technology that has real world applications, as Nikos referred to there. Innovative uses of AI are fantastic in the lab, but you've got to then scale them out. You've got to get them from lab to fab, as they say. And so many innovations, particularly in AI, never get beyond the lab, do they?

Frankie Kay

You hit in on a really important an issue, and one of the...you mentioned about the one-stop shop, about me leading that across Europe as part of the consortium. And one of our chief outcomes is to try and get away from the experiment or accelerate the experimentation phase into production. There's a balance here in the statistical world between us wanting to get the most high-quality, accurate estimate that you can get. But one of the...I suppose the traits of quality, to me, is timeliness.

So, you're trying to balance different elements and different aspects of quality to say, well, what is the most accurate in terms of the precision, the data compared to...Well, there's no good getting a statistic out if it's a year too late. We definitely in the one-stop shop, we're looking at trying to provide common resources for different statistical offices so that we can take those models and use them more quickly into production.

And there also needs to be, I suppose, an understanding that these models will continue to evolve. You start off, you develop an algorithm by looking at the data, but you're constantly getting new data in. You need to keep retraining the models in order to make sure they're taking account of the new data that you've got available and updating those models. So, we have to find a way of being more agile in keeping those models up to date.

I think as a community, sometimes, I think we should be bolder. We should be more confident about saying, actually: let's get this out there. Let's get...let's be transparent about what we're doing, which is absolutely key, so that we're being really clear about the strengths of what we're doing, but also the areas where perhaps we're not so sure about and we need to do more research. But we're just being honest and open with people about where we are.

And I think therefore getting these data out and getting collaboration, getting expertise ideas from others as well: the academic world, private sector, think tanks. You know, we can all work together in helping produce better data, because that's what we need to make the best decisions that we can.

Jonathan Elliott

Yes, that's something I learned from the experimental statistics podcast, which is that if you're honest about what you're doing and say: well, these statistics aren't perfect, but they're pretty good, then at least users can make up their own minds, and you can protect the innovation from anxieties about quality and standards.

Person on street 4

I have used an AI chatbot because I like to study architectural styles, and the very unique architectural styles are sometimes very hard to find. But by speaking with a chatbot and with highlighting the different aspects of the styles, it was amazing. In a very rapid time, it gave me the correct answers with lots of good background information.

Jonathan Elliott

Let's go to Oslo now and talk to Eva about what Team AI are doing at Statistics Norway. I don't associate chatbots with statistical inquiry, but I think I'm about to find out more. Eva, just tell us a little bit more about why it's important and why you got going with this chatbot initiative, and the importance, particularly of using open-source software and its development. That's a really important part of this

work that you've been doing. So just sort of set out for us a little bit about what you're trying to do and the open-source element.

Eva Charlotte Berner

So, the reason we wanted to create chatbots is because people really often know what tasks they want to accomplish, but they may not know how to accomplish it. Or they may know what kind of information they want, but not how to get it. And a chatbot can bring...bridge that gap. And the reason we wanted to create it open source is because proprietary large language models - LLM's - can be prohibitively expensive.

They can make you incur costs very, very quickly. And by using open-source technologies, we can wrap that cost right down and democratise access to this technology. The second reason we often want to use open-source technologies is because we control the software pipeline. We can put...take this large language model and put it in jail so that it can only speak to us. And that means that no information is going to leak out of Stats Norway, ensuring that we can use sensitive data in our chatbots.

Jonathan Elliott

So, why do you need that safety?

Eva Charlotte Berner

So, as statisticians, we are quite privileged that we get access to sensitive data, and often there are laws and regulations protecting who has access to this information. So, it is vastly important that we make sure that it does not leave our control, because that would be breach of this privilege.

Jonathan Elliott

Now, Statistics Norway itself is going through a major organisational change. Tell us a little bit about that, the transition to the cloud and what that has meant for the people working there.

Eva Charlotte Berner

So, you're quite right. We're going through a major digitalisation project. It's actually a complete overhaul of our data platform, which means that people that might never have coded before are suddenly expected to interact with R or Python scripts and be able to not just understand them, but use them, which means a completely new way of working. And we believe at Team AI that chatbots can help people in this transition.

Jonathan Elliott

Excellent. So, Python is a programming language, as most people know, but R is also a very specific statistical programming language, isn't it?

Eva Charlotte Berner

Yeah, it's common among data scientists.

Jonathan Elliott

Okay, now going to the chatbots themselves, tell us about some of the prototypes you're working on.

Eva Charlotte Berner

So, one of them, one we've come the furthest with, is a prototype using an algorithm called RAG, which stands for retrieval (R) augmented (A) generation (G). And what we do is we take, for example, a data platform, we take the documentation off the data platform, we web scrape it, and then we reformat it so it's now readable by the large language model.

And then when you ask a question: 'how do I log on to the data platform?', that question is sent through the knowledge base. We collect the relevant information, and then your large language model will provide you an answer based on the relevant information in your knowledge base.

Jonathan Elliott

So, if I've got this right, the RAG goes and scrapes, this strange term scrapes. It actually, basically means gathers together all the data and turns it into something called a large language model, the LLM, which is what AI runs on. Have I...have I got that right?

Eva Charlotte Berner

No, not quite. Not quite.

Jonathan Elliott

Oh, dear. Please explain.

Eva Charlotte Berner

So, you have your large language model, which is also open source, and then you take your knowledge base, and you reformat...it. You embed it. And that is readable by the large language model. So, when you ask a question, it goes into the knowledge base and: oh, this is the relevant information I need. And then it provides you an answer based on the relevant knowledge base.

And we have different prototypes. So, one is based on the documentation for a data platform, but we're also working on a prototype where you can upload your own knowledge base, and it'll do all the 'embedding them' maths for you, so that you can use it on, for example, an annual report for a company, and then chat with the LLM about your annual report. Or if you've got an Eurostat document that you need to be read quickly, because you have a publication in two hours, and 'what is the format I need?!', then you can chat with it quickly.

Jonathan Elliott

So, it sounds like a researcher who can go off to a library and get everything you need, except the chatbot could probably do it...

Eva Charlotte Berner

Exactly!

Jonathan Elliott

...in about nine seconds. I think if you ask that from a human researcher, they will probably burst into tears. Doesn't that then need people to ask the right question? I'm just wondering, because that in itself can be quite a precise art, can't it?

Eva Charlotte Berner

I think that we shouldn't undervalue people who are not digitally capable. They might well have the ability to ask relevant questions. They just might not have the skill to get onto the database, formulate an SQL query and get the data.

Person on street 4

There is a risk with artificial intelligence because it can get smarter than humans if we let it. But I think so long if there's a kill switch to, like, shut off the AI, we're safe.

Person on street 5

Am I concerned about AI in the future? Yes, because I think that it may have an effect on creativity and independence of thought.

Jonathan Elliott

Talking a little bit earlier about the ethical concerns of AI, can you just tell us about what Statistics Norway is doing around things like privacy and data security and so on, and how you at Team AI are handling that?

Eva Charlotte Berner

There are several concerns; many actually, when you think about AI. But one of the ones I want to highlight is biases. Because these large language models are trained on a text corpus, and that text corpus is created by humans. At the moment, it is highly difficult to create a large language model that is less biased than the data it is trained on. We can correct it, but it doesn't work as well as we'd hoped.

And so, I mean, you can look at the results from ChatGPT or from Gemini, where both OpenAI and Google have been in lots of hot water for trying to bias-correct the large language models. It's very, very difficult to get correct.

And that's something we're working with, and it's another challenge, because we want our language models to answer in Norwegian. And Norwegian is a Germanic language, which means that we have gendered nouns. And that will sway the way in which English language tests for bias.

Jonathan Elliott

Oh, my goodness, that's fascinating. Are you saying then that the gendering of nouns in particular languages makes the language models work differently?

Eva Charlotte Berner

It might - because there has been so little research on minority languages with different grammatical building blocks that are not English.

Jonathan Elliott

One thing I hadn't realised was that the way that English dominates large language model science generally, and perhaps text to speech generation, particularly, is that it entirely assumes the English language. I didn't realize that. I mean, can you explain a bit more? I thought large language models operated irrespective of national languages, that they were more really about mathematics.

Eva Charlotte Berner

Well, more and more language models are getting better at minority languages, but the results so far, as far as I've tested, are simply not good enough in minority languages. So, how we're solving that in Stats Norway is we're doing the reasoning and the thinking in English, and then after we have an answer, we're translating it. So, we have a twostep because the reasoning is so much more powerful in English.

Jonathan Elliott

Let's just talk quickly about what difference this could make to people working with stats every day within stats Norway - so your colleagues, subject matter experts, people who are there who may well have previously spent a lot of time digging around for information or filtering, sifting, and sorting it. They can now get so much more done that's more high value. Can you tell us a bit about how AI is really releasing human talent to do what it's really good at?

Eva Charlotte Berner

My dream for AI in this regard is to have no unnecessarily boring tasks; that anything that AI can automate that doesn't imply a reduction quality AI should automate. I hope that AI can reduce tedium for subject matter experts and allow statistics products to be produced more efficiently and allow them to spend time being subject matter experts, because their competency and their expertise is so valuable.

Jonathan Elliott

Yes, yes, I mean, do you find that you're having to kind of do a little bit of education for your colleagues and to help them kind of get up to speed with AI?

Eva Charlotte Berner

I think the adaptation of AI and ML is always going to be difficult, because as statisticians we like transparency in our statistical production. And AI and machine learning...well, in its essence, it's maths, and most people get a bit of math's anxiety when they're presented for abstract mathematics. And in many ways, machine learning, and at least AI just predicts the next words. So, that is difficult.

Jonathan Elliott

These chatbots aren't ready to go yet, are they? But they're very much in development and planning. So, what do you hope are going to be the next steps? What's the future?

Eva Charlotte Berner

So, now we're testing out different deployment strategies - how to launch them, to make them efficient - but we hope to have a prototype deployed for internal use pretty quickly. And the reason for that is that although our prototypes may not be perfect, we at least then get out of an R&D limbo where a lot of AI projects die.

Jonathan Elliott

Ah yes, which is what we were talking about before. This is the lab to fab headache which bedevils quite a lot of this innovation. Frankie, just tell us a bit about how you're working with Eva and her team. And, you know, the extent to which this is an example of the things that the one-stop shop can be doing.

Frankie Kay

The work that we're doing is what we're calling the sort of AI ecosystem. So, this is about understanding what AI and ML is being used across many different countries, and then bringing that together so that we can make that available - so that people know about all the different work that is going on. We're also carrying out some individual projects around new cases, not specifically on chatbots, but we will then make those resources available.

We'll also look to provide training as well. We're going to be showcasing the work that we do in various different conferences and, you know, looking at using social media as part of what we're trying to do is really create a community around AI and machine learning so that we can connect people up together, understand what they're doing and share those resources.

So, yeah, it's that idea of bringing this community together, understanding what each other is doing, making everything we're doing available, other than the raw data, to be really clear - that has to be kept safe and secure and so forth, of course, but all the methods, the best practice, the frameworks, the training, is made available so that we can learn from people like Eva. Because in Ireland, we're using chatbots potentially to support running our census.

So, if citizens have got queries, instead of having to ring a helpdesk or talk to one of our field officers, they can...they can go online and they can say, you know: 'send me a form', or 'give me a new code'. So, you know, we're looking at how we can use chatbots in that context as well.

Jonathan Elliott

Nikos, do you think chatbots have the potential to give people in-depth access to information which otherwise would be time-consuming or require uncommon technical ability? How do you view the chatbot revolution if I can put it that way?

Nikos Roubanis

Well, knowing from inside the complexity and the depth and the detail of information we produce, I think that every way of making more accessible the data and more easily understandable and more easily discoverable, I think it's worth investing at. And artificial intelligence is an area where it is very promising to work together with the scientific community for dissemination, but also for other purposes.

Jonathan Elliott

Fantastic. Well, we look forward to talking about it in further detail as things develop. It's a hugely interesting and very fast-moving area. We've run out of time, unfortunately. It remains only for me to say thank you very much to everybody and for their amazing contributions. In particular, really fascinating, to Albrecht Wirthmann, Head of Unit Methodology; Innovation in official statistics at Eurostat. Thank you so much for your contributions today.

Albrecht Wirthmann

Thank you, it was a pleasure.

Jonathan Elliott

And Frankie Kay, the Chief Information Officer at the Central Statistics Office Ireland, and also leading the one-stop shop. Thank you so much for joining us.

Frankie Kay

It's been a pleasure. Thanks very much for having me.

Jonathan Elliott

Nikos Roubanis, Head of Unit at Transport statistics, Eurostat. Thank you so much for bringing your perspective today.

Nikos Roubanis

Thank you, too.

Jonathan Elliott

And finally, Eva. Eva Charlotte Berner, thank you so much for your contribution. You're the leader of Team AI at Statistics Norway, and we're really grateful that you came in to tell us all about the chatbots today.

Eva Charlotte Berner

Thank you for having me.

Jonathan Elliott

If you've enjoyed Stats in a Wrap, don't forget to follow us on social media and share our adventures with friends and colleagues, where the show can be found, on Spotify, Apple and all the usual places. And of course, join us for the next edition, when we'll be dishing up more flavoursome insights from Eurostat, this time about how people use Eurostat products and data in their work. Join us then, but for now, goodbye!