Qualitative Analysis and Large Language Models: Transcript

[0:00:00]

Susan Halford: So, our talk is on the Sociodigital Futures of Qualitative Research and we're going to split the talk between us and you'll see how that emerges over time. So, large language models. Three years ago, I think very few people would have heard that phrase, they wouldn't have known what it meant. Almost certainly very few and possibly nobody in this room today.

Today, there are literally hundreds of large language models, as you can see, some representation of that on this borrowed graph with ChatGPT alone reporting over 700 million regular weekly users in 2025.

Now, I think it's probably fair to say that qualitative researchers are not the largest constituent group of that 700 million, but nonetheless, as Ros alluded to, there has been much excitement about language models as a tool for qualitative research, witness this event today. And indeed, it's very widely suggested that a technical revolution, this is the language that's used, the technical revolution is afoot that will transform the very future of qualitative research.

Now, this kind of claim is explored in a large and rapidly growing body of literature. If anybody's trying to look at it, you'll know how quickly it's growing with quite a systematic review of that literature. I think I'm fairly confident in saying that the vast majority of it takes quite a similar approach. That is, it's framed by a discussion of the opportunities of language models and the risks of language models for qualitative research. The opportunities usually are speed, scale and robustness,

sometimes replicability, we'll come back to that. The risks are data security, bias and transparency.

And what normally happens in this literature, the dominant narrative, if you like, is there's some risks, there's some opportunities, let's do some experiments to see what we can learn about those risks and opportunities and how we can explore language model performance in a variety of tasks and compare those more or less directly to human performance in the same tasks.

Now, out of all those many, many experiments, not surprisingly, the outcomes vary, and I think that's pretty interesting to look at that. But those insights that are produced through the experiments are largely fairly, usually fairly quickly followed by, and I'm putting my hands like this because this is usually how it's put, much needed guidelines to ensure that we can shepherd a kind of responsible transition towards the effective and ethical use of language models in qualitative research.

So, overall, there's a really strong message here. Language models are coming. This is the future. Qualitative researchers, get your act together and get ready for it. Now, Les and I have got some experiments to report as well, and we hope that our experiments contribute to that body of knowledge. But before we do that, we want to take a step back and reframe how we think about this and how we think about the debate.

We've got two elements to the reframing that we want to propose. The first one is that rather than seeing this research that I've just summarised as somehow a timely response to a future that is out there waiting for us, our starting point, to put it really bluntly, is that the future does not exist. And most predictions, especially those about the interplay between technologies, digital technologies and social life, tend to fail. And just as an example of that, I'll point you to Jens Beckert's work, and he draws on longer work by Nye.

[0:03:52]

Nonetheless, despite that kind of failure of prediction about what I would call sociodigital futures, claims about the future really matter. How the future is claimed encodes certain ways of knowing and, in brackets, excludes others. It creates discourse, it shapes expectations and perhaps really importantly, it drives policy and it drives material investments, none of which determine the future, but all of which tend to open up some futures and make them more likely while closing down other futures which may never even get on the table for consideration.

Now, from this perspective, the kind of epochalism that we have seen in the debate about language models and, brackets, we saw in the debate about social media analytics and we saw in the debate about big data and machine learning, close brackets, this epochalism can be seen itself as a kind of more or less intentional future claiming in which language models are now being positioned as the inevitable future, it's going to come, it’s just a matter of when, the inevitable future of qualitative research, and the guidelines that I mentioned are seen as a way of smoothing that into being.

The second element of the framing that we want to propose is that rather than seeing language models instrumentally as discrete kinds of tools that might offer a better or a worse way of doing qualitative research, they can, and I would say should be seen, as the focal point for a far wider network of actors, practices and relations that constitute ways of knowing and constitute the knowledge economy, which is particularly important here, I think.

So, we're drawing here on work from the social life of methods. I'm thinking of Evelyn Ruppert's work and Mike Savage's work in particular. And you can see on the slide that that work in summary, like much critical methodological work that has gone before it, emphasises that the methods we use produce data, they don't just reflect the world, but

they produce certain forms of knowledge and they stabilise a complex, heterogeneous and dynamic world into particular entities and particular kinds of action and not others, so that they have an effect. That's not actually a criticism, that's just an inevitability that methods do that in producing data.

Secondly, that methods are always implicated in forms of ordering and power. And what I mean by this is not only through the way that methods validate some kinds of knowledge over other kinds of knowledge, but also that methods have been historically and very much today continue to be connected to certain kinds of agencies, whether that is the state or commercial actors, and we'll come back to that, but also to wider institutional arrangements. So, if we think about disciplinary norms, for example, and how those are policed, the boundaries of those are policed in a whole variety of ways through disciplines and journals and all the rest of it.

So, putting these two elements – so, putting these two elements – oh, I can hear myself now, but I'll carry on. Putting these two elements together, we might see the current debate about language models in qualitative research as a kind of future claiming in which those claims are mobilised through a nascent reassembling of the dominant methodological apparatus. I mean, that's not yet, but I think we could see that as something that's underway.

So, putting these two, sorry, here, we want to return to our experiments. So, what we're going to do here is report on a live project. Most of the research that's been done on language models and qualitative research is very purposive, a kind of very specific test. We're going to give a model, this thing, and see if it can do it. What we're going to do is kind of widen out a bit and talk about a project that's been ongoing for a couple of years in the ESRC Centre for Sociodigital Futures, and to explore how we brought language models into play in that project. And this, I suggest, means that we see some of those much broader

connections between actors, practices and so on that the social life of methods points us to, and it also suggests a rather different way of thinking about language models and the future of qualitative research.

[0:08:34]

I'm going to pass over to Les in a moment. Just to say, last year we were offered 14,000 completely, or not completely, almost completely unstructured free text comments that were produced by a very large employer who we work with in the centre in response to their employee survey that they do on a regular basis. Now, they had no idea what to do with all this data and they thought it might be useful for us because we were already doing qualitative research in that organisation using interviews, observations, conventional qualitative methods. And this data appeared to offer us a really unique opportunity to start to look into some of the same questions across a very, very different data set. We got sent this huge spreadsheet and it didn't take us very long to realise that it was overwhelming. It was indigestible. We really didn't know what to do with it. So, given the intense debate that was raging already then about language models, our thought turned in this direction. Could we use language models, or chatbots to be more precise, to help us to query and summarise this data? And what in turn might we learn about language models from doing that?

So, in what follows, we are going to focus on three of the qualities of language models that are most commonly put forward as opportunities for qualitative researchers and we're going to explore how that was for us in relation to our experience of using language models for qualitative research.

Leslie Carr: Thank you very much, Susan. Yes, so obviously, well, I say obviously, but the first quality that people associate perhaps with AIs or the rationale for using AIs is the speed and scale at which they operate, the computational can sort of process large data sets quickly and efficiently.

You know, we're used to that narrative. And now we can apply it to, sort of to AI and LLMs. The hope with IT is always that it will increase productivity and that we can take amounts of texts, volumes of texts, that are beyond the capacity for an individual researcher or coordinated groups of researchers to be able to interpret and synthesise, you know, to produce analyses of in reasonable times.

And I think possibly because this is, you know, sort of this is a computational process, the idea of the speed and scale is rarely, rarely questioned just because we're bringing intelligence to the table. But what we discovered, I think quite, quite quickly, well, no, in a protracted way in the work that we did over the last year, is it really depends what is being timed, how fast, what happens. Are you talking about how fast a particular sentence could be processed and understood? How fast a particular page could be processed and understood? Or how fast you can undertake your research?

And so, for us, because the data was considered commercially sensitive, then there's a large amount of time. And because we're doing commercially sensitive work on using methods and mechanisms and programmes that are not well understood, and obviously I stopped at programmes there, but I should say using commercial platforms where the technology and the commercial aspects are not incredibly well understood, then there is a lot of discussion that involves legal services. And from legal services, just that the basis of which we're going to do this work, the IT services to make sure that what we're doing actually fulfils the contract that we stated, that the route from where we put the data, the route to it actually hitting an AI and a chatbot and that internally that the chatbot wasn't going to do anything or disclose anything or allow data to be used in an inappropriate way and to get the judgement and the agreement of IT services on a number of these platforms and then to make sure that our ethical considerations are being withheld, upheld, not withheld, and that's all reflected in the contract.

[0:14:11]

That took an incredibly long time for us to be able to establish. And having done that, actually being able to choose which, so just to get hold of the data and then which services we were going to be able to use of the many, many different platforms and chatbots and programmes and whether they exist in the public web and they're a large internet platform, whether they exist corporately, they're made available by our IT services with sort of guarantees about visibility and performance, or whether they actually exist on the laptops and the desktops that we own and run, and how the data is transferred between them. In many ways, lots of things that are very familiar from working with other internet platforms that aren't perhaps AI, that weren't anything to do with AI, but in many ways trying to work with new pieces of AI, new companies, new products, where we don't know what the restrictions are.

And then the technical problems leading to lots of issues and thinking. It turns out one of the key parts of an AI platform, one of the key restrictions, is the context window, how much information it can pay attention to, it can focus on, it can maintain in its context, its knowledge while it's analysing.

And so, the size of that compared to the size, the huge size of the data we have, means that actually you can't just give it 14,000 statements. You have to reduce that. You have to chunk it up into different parts and ask it to process those. That meant that we were experimenting with synthetic data of a smaller scale that represented that and having to try to produce that so that we could experiment with the kind of analyses that we were performing without the looking at the huge amount of data and dealing with that all the time.

And that means producing Python scripts, which was for us was going off piste because although I am a computer programmer, the majority

of social scientists are not computational social scientists and we really wanted to represent the kind of analytical pathways that were going to be available to people. That meant using mainly just chatbots.

So, there was some scripting that we had to do. And then once you're working with a lot of data and performing these analyses of it, trying to go back and sense check to make sure and to compare the answers that you were getting with the expectations that you would have, but the expectations that you're trying to form on a piece of data that's too huge, on sets of data that are too huge for humans to be able to come to that sense in the first place.

So, putting those processes together was a big problem for us and there were lots of missteps along the way. If you go over the context window, then your AI forgets what it's talking about and starts to, instead of looking at the, trying to do the things you've asked it to do with the data, it will just do what it normally does, which is give you advice about the problems that it sees with the data. There are all sorts of issues that we had there. So, even to make some progress with this, it took a lot of time.

So, speaking of things that always work, definably, rigor. And so one of the claims is that, of course, because it's computational, it'll be more rigorous, AIs will be more rigorous than humans, and we could come to more robust and reliable conclusions, whereas what we actually discovered or what we uncovered during this process was that you don't get the same results every time because you're asking the same questions and the same data because there is an inherent randomness built into the process by which large language models are deployed.

[0:19:35]

And so, at the very heart of what they do, the way that they produce the text, there is some randomness going on. And so, we can't, we won't

always get exactly the same results. And of course, as you change from one product to another, one version of a model to another, you will get different results coming out because the training will change. The way that the AI, the language model, sees the world or tries to respond to your questions will change because it's got a different understanding, different context in which it's looking at your data.

And then, of course, that's all out of your control. The things that are in your control perhaps are this idea of prompting, the way that you make inquiries of the AI, and the coming to some agreement, coming to some conclusion about what is the kind of prompting? There is a literature on prompt engineering, as it's called, or I don't think it's really engineering, but that is widespread and very, very changeable itself, because the way that you prompt changes all the time as different capabilities are added to different systems.

And the actual training of the chatbot is a real problem in rigor because the large language model at the heart is very raw. How it responds to questions, how it understands the task of summarising or translating or analysing or modelling something is dependent very much on the training that the host company has given to their product. They've said, “This is how you summarise, these are the key things to look for, this is what you're allowed to talk about, these are the topics that you're not allowed to go near, these are the taboo issues, these are the way that you should deliver helpful and uplifting responses to avoid reputational harm and legal challenge”.

So, those are issues that we had to be very aware of when analysing, trusting the analyses that the AI had come up with.

The next slide just shows some of the consistency issues that we had as we tried to ask the same, oh, here we go, the same question to the same piece of data. Everything that we've said about the timings and the extra work that we had put in, talks to the problem of cost. People

think, oh, because there's a computer involved, it'll happen quicker, therefore it'll be cheaper. What we'd found was that human labour is being shifted to other things, not just to more better, more analytical issues, but dealing with the problems of bringing the AI in in the first place. Susan, over to you again.

Susan Halford: So, three points, I think. The first thing is we're not opposed to using language models in qualitative research. Actually, both of us are quite enthusiastic about it and we did find that the results of this work was really helpful. We integrated it into our qualitative research, and I can say some more about that if you're interested.

However, it's really clear to us language norms are not only an instrumental tool. They're not quick, easy, or a cheap personal assistant. They're complex, time-consuming and messy. And as critical methodologists, as sociological, small S sociological researchers, we really need to think very, very carefully about how we engage with them and their epistemic disposition in particular that Les has alluded to.

We need to attend to the wider assemblage that they enter into, both as commercial products, but also how they might be reshaping where qualitative research expertise is held, how it's owned, and how it's redistributed not only across commercial companies, but also into areas like medical sociology or science and engineering where there are already people saying, “Well, we don't need to work with qualitative researchers anymore. We can just use language models. It'll be fine”.

[0:24:41]

So, there is something happening in that wider assemblage that I think we need to pay attention to as social science researchers. And whatever we think about that, we need to be really clear that the guidelines that we offer have to pay some attention to that. So, the guidelines cannot just be technocratic guidelines. They do need to, I

think, take these issues into account and be seen quite clearly as claims on the future of qualitative research, not simply as technical bits of advice.

Sarah Jenner: My name is Sarah Jenner, I'm a lecturer in health sciences, but this work that I'm going to be talking about today was actually conducted as part of my PhD which I've recently submitted and I suppose as an example of one of these studies that Les and Susan mentioned about exploring the use of large language models for analysis and evaluating what they can and can't do for us as qualitative researchers.

So, just a little bit of context for this study. So, this started around the time that ChatGPT was first released around the end of 2022, and during this time, I was in the middle of my PhD, I was in the throes of conducting a narrative analysis of some story completion data that I had collected which looked a little bit like this photo on the slide here, a very kind of old school, traditional, hands-on, printing out bits of paper, chopping them up, type qualitative analysis. I had my entire office floor covered in pieces of paper and scribbled and Post-it notes, because that was how I had approached qualitative analysis in most of my work so far.

And then I started to hear about this thing called ChatGPT and how it was a model that was designed to interact with humans and identify patterns of meaning within text. And I thought, “Okay, that sounds exactly like what I'm doing right now”. So, it kind of brought about this idea of can LLMs help us, assist us with these types of qualitative analyses and how can we use LLMs whilst thinking about the things that we always think about when we do qualitative analysis like subjectivity, transparency, reflexivity and the human elements of qualitative analysis? So, that's where this study came about.

And a bit of background about that analysis that I was conducting at the time. So, my PhD is based on the idea that it's all about young people,

adolescents, the process of identity formation during that time as a developmental psychological task. I'm a psychologist by background, I probably should have said that initially.

So, my research question was around how young people themselves understand the role of social media in shaping and expressing identity and dietary choices. So, my PhD is all around identity, food and the role of social media in that.

So, I think when we talk about food and when we ask people about their diets and food choices, people don't necessarily connect that with identity, but actually, if you start to ask people about their diets, people will often talk about them in terms of framing it around particular identities. So, people will say, “I'm a vegan,” “I'm a foodie,” “I'm a picky eater,” “I'm not a breakfast person,” those sorts of things. And those are how we see food and dietary choices reflected through identity.

And for young people in particular, this is a really important process that they are going through and we know that social media plays a big role in that. So, there's a huge amount of food and health related content on social media and young people will take inspiration from influencers, celebrities, their peers, in terms of developing and expressing both their identity and their food choices.

[0:29:24]

So, bearing all that in mind, I had designed a story completion study. So, for those of you who aren't familiar with the method, it involves the researcher writing what we call a story stem. So, the first line or couple of lines of a story introducing a character and a scenario and then participants are recruited and asked to complete the rest of the story. So, write the rest of the story.

And on the right here is an example of one of the stems that I used in my study and some prompts that I included underneath to encourage

the young people in their writing. So, these are purely fictional stories. They're written in the third person, but they are related to the topic that I was interested in studying.

I ended up collecting just under 140 stories and, as I said, I had decided to analyse those using narrative analysis. So, in terms of what I'm going to be talking about today, this is the kind of process of analysis that myself and my colleagues followed.

So, I collected the data, I had analysed all of these stories myself using a traditional narrative analysis method with all the bits of paper on my office floor. We then actually partnered with Ipsos UK. They used to be known as Ipsos Mori. They do lots of market research. They do a lot of the political polls and things like that. We partnered with them to conduct this LLM assisted version of the same analysis, of the same data. And our ultimate aim here was to compare the findings between the human analysis that I had conducted and the LLM assisted analysis that we conducted.

So, I'm just going to briefly go through the process of the narrative analysis that I had been through myself. So, this was before I even thought about using LLMs or before I knew anything about AI really.

I began with reading through all of my stories, as I said, kind of scribbling all over them, using Post-it notes, colour coding. A lot of you qualitative researchers will be familiar with this approach, but as you can see, that was quite chaotic, I would say.

So, I then started to try and identify particular narratives within these stories and group them. So, I grouped them into four different types of narratives that I identified. So, similar to themes, I suppose, if you were to do a thematic analysis of these data. I then summarised those four narrative groups and I used keywords that were within those summaries to label each of these four groups.

So, we then conducted the LLM assisted analysis and, as I said, we partnered with Ipsos to do this analysis and that actually allowed us access to their own secure AI system. So, they have a system called Ipsos Facto, which is only available to people within Ipsos usually. It's not an LLM in itself, it’s a kind of a vehicle to access commercially available LLMs. So, through Ipsos Facto, we had access to all of the latest models from OpenAI, Anthropic, Google, all of the models that were available at the time. But this was through a secure closed environment. So, this environment is within Ipsos. Any information put into Ipsos Facto, it does not go outside of Facto. So, it doesn't go back to OpenAI or Anthropic. No one else can have access to it and this is how we got around some of the ethical issues that Les and Susan started to mention in their talk about data storage, making sure that we were adhering to all of our usual ethical guidelines, GDPR, that sort of thing.

So, at the time that we did this analysis, this was a couple of years ago now, but at the time we used the most advanced models that we had access to. So, we used Claude 3 Opus and GPT o1, so that was the most advanced version of ChatGPT at the time. We used prompt engineering, which I don't really have time to go into detail about, but as Les briefly mentioned, it's a process of how we’re able to really, really clearly communicate with models to tell them how you want them to do analysis.

[0:34:29]

We asked both of these models to conduct the analysis separately using the same data, same method, the same four step process that I had used, and we then reviewed the output.

So, what did we find? I would say that I'm very much more on the kind of pro AI end of the spectrum in terms of this use for qualitative analysis. I found it to be really, really helpful. We felt that both of the LLMs were

able to conduct really thorough, rigorous narrative analysis. Using the same steps, both of the models were able to generate either pretty much the same or very similar narrative types to those four that I myself had identified. It wasn't a completely easy, simple process of just asking the models to conduct those analyses. It was many, many different chats back and forth, interrogating the models, asking them to explain the processes that they had used to identify narratives. We asked the models to provide us with illustrative quotes from the stories to exemplify why they had categorised those stories into those narrative types.

So, initially, when we first did the analysis with Claude, we put in our prompt and our instructions, we put in our data, Claude returned an analysis that was very, very comparable to the findings that I had found. So, those four narrative types returned that to us within 60 seconds, which was, for me at the time, absolutely mind-blowing having never used this type of technology before. We were really blown away. And that was where we started to really think about, okay, what more can this do, right? So, then began the process of going back and forth with the models, asking more about the process and how they had actually got to that point of developing those four narratives. And we estimate that took about 35 hours in total compared to the 64 hours over a period of around four months that it took me to analyse the data, traditionally, manually, by myself. So, although it did take time and preparation of the data to put into the models, etc, we did find that in general it saved quite a lot of time.

I just want to briefly touch here on reflexivity as well. So, this is something that people, qualitative researchers, always ask me when I talk about this work, rightly so. It's something that myself and my team have thought about a lot. Can these models, can they be reflexive? What does that mean? So, part of this study that we conducted, we asked both of the models to think about reflexivity and to write us a

reflective, reflexive paragraph about their own biases and their own background and all of the things that you would usually expect researchers to think about when reflecting on their own impact and own biases. And this is what GPT o1 produced for us.

So, you can see it's a pretty standard reflexive paragraph. It's pretty much telling us what it thinks we want to see, right? It knows what reflexivity is and what people usually write for this sort of thing. It's getting all those keywords in, like biases, subtleties, nuances. But is this model truly being reflexive? I don't think so. Everyone has their own opinions about this, but I think it's easy to overestimate the understanding of these models. Obviously they do not have any capacity for understanding or true cognition. They're not conscious or sentient. They're not truly intelligent. They are just very clever computer programming.

So, I don't think that this is true reflexivity, but I do feel that doing the analysis with these models really provided an alternative perspective to my data. It made me think about things that I hadn't necessarily thought of before. It brought in alternative views and interpretations, and I think that then in turn helped me think about my own reflexivity.

So, there's something here with these models about assisting us to think more deeply about our own impact on the analysis process. But yeah, something I'm always talking about with colleagues at the moment is what is this called? Do we need a different word for this? It's not reflexivity because we're not talking about human cognition here and human feeling and human emotion. But there's something here and I think that that's an important part of the discussion.

[0:40:20]

So, as I said, I found that particularly Claude was really able to replicate this analysis. And as Les said earlier, you won't find that if you ask the

same model with the same data to do the same analysis multiple times, you won't find exactly the same results, but you wouldn't find that with a human anyway, right? We're conducting qualitative analysis with colleagues, we all have different interpretations, we all come from different perspectives, we will all interpret things differently. And I think I kind of see these models as almost the same as working with a new colleague or a new research assistant. We'll think about things differently, but actually we were on the same page. We had picked out the same key narrative. So, these lines represent connections between the narratives that Claude found and the narratives that I had found were generated as a result of this analysis.

And I don't have time to go through all of them in detail, but you can basically see from this that it was finding the same key themes. And I didn't tell the models anything about what I had interpreted or my narrative groups that I had identified or anything like that. This was purely coming from the data and the prompts that we put in about how we wanted Claude to conduct a narrative analysis.

So, that was really, really impressive. Of course, I completely agree with everything that Susan and Les said about limitations, ethical issues. There is a long way to go in terms of using LLMs for this type of analysis, hallucinations being a big one there that often comes up. We definitely had problems with hallucinations in our analysis, making sure that we were checking everything, asking for quotes from the stories to back up and explain why the models had made certain decisions around grouping stories and constructing narratives and that sort of thing. Obviously, we spent quite a lot of time preparing the data and going through the prompt engineering process. That took more time, although I would still say that I think we save time overall by using this approach.

And of course, the ethical issues that are associated with this in terms of do participants need to be aware that we're using AI to analyse their data? This is a big topic of conversation at the moment and there's

people have very strong opinions about this at both ends of the spectrum. So, I don't know if we're quite there yet with a consensus, but it's definitely something important that I think is important to think about.

Again, in terms of data storage, location, who has access, are these models GDPR compliant, especially if you're accessing them directly without the kind of secure system that we had access to. There's all sorts that have been in the news as well about the kind of environmental and human costs of AI. So, the extortionate amount of water, for example, that is needed to cool the systems that store and run this technology, and also the awful conditions that many workers, particularly in developing countries, are working in to train and maintain and moderate content through these models. There's all sorts out there about that.

And as I said, I think it's really important to recognise the limitations of this cognition or whatever we want to call it. I know I've said many times in this presentation, like the Claude was, you know, what did it think and how did it interpret? Of course it's not doing those things truly. So, thinking about making sure we remember the fact that this is a computer programme and there are many, many limitations and issues there and just not falling into the trap of treating it like a human being because they are very convincing, very convincing. So, I know I'm kind of coming to the end of my time here, but all of this work has been published in a paper which I'll link at the end, and the purpose of that paper was to explain the method that we had used and almost demystify this process for qualitative researchers who have no knowledge or experience of AI or LLMs. And we've developed this really easy four-step process that can be applied to any type of qualitative analysis with the caveat of these, what I like to call golden rules, which everyone is talking about, all things that Susan and Les have mentioned as well, but about understanding how this technology works before you dive into using it blindly, checking and validating all LLM generated content,

acknowledging researcher and LLM biases. So, being mindful of reflexivity and making sure to still incorporate that. And of course, understanding responsible and ethical use of these models.

[0:46:03]

I think that's me. You can find our paper, which goes into much more detail about everything I've spoken about today through this QR code link. So, or feel free to use the reference.

Marianne Aubin Le Quéré: Okay, so, just as a brief introduction, we wanted to give you an introduction of ourselves, our background and where we're coming from. It's really exciting to be here. I myself actually grew up in Norwich, although I don't know what has happened to my accent since I moved to the States across for the last 12 years or so. But it's really fun to be here and talk to this crowd. My name is Marianne Aubin Le Quéré. I'm currently a postdoc at Princeton Center for Information Technology Policy which is an interdisciplinary research centre that brings in folks from machine learning, computer science, but also other disciplines like law, sociology, and journalism. Casey, do you want to give a bit brief intro?

Casey Randazzo: Yeah. So, hi, everyone. You heard that I'm an assistant professor in communication at UC Santa Barbara. I study how humans organise with AI and typically in contexts like crises. Nice to meet you all.

Marianne Aubin Le Quéré: Yeah, and I would say between Casey and I, we represent a variety of fields and perspectives, but primarily we're bringing together these two disciplines of computer science and communication, and so this is a really exciting venue for us.

And the work that we'll be presenting today was presented at a place called CHI Conference in Human Factors. And so, for us, our conferences are kind of like journals in terms of how important they are, and this is an interdisciplinary conference. It's the biggest and most

reputable in our field of human computer interaction, but it's also really unique because it has both a critical streak where people are really willing to contend with some of the negative aspects and harms of AI and technologies, but it's also the place where you really can reach a community of tool builders and it's itself a very technology forward conference, right? And so, we'll be sharing, I guess, a little bit more from the technical perspective, but we're also really, really looking forward to learning from this crowd, seeing what it is we should be bringing back to our own communities, everything like that.

So, just as a form of background, I started thinking about some of this when ChatGPT first came out. I was working on a project about Nextdoor, the local social media platform, and at the time in 2023, we were already using it a lot for other work in our lab. And so, because I was trying to tie some qualitative data together with underlying quantitative data about neighbourhoods and census data and things like that, I made this first initial attempt at scaling up a qualitative codebook using a large language model.

And if you'll go to the next visualisation, I was sort of, on the one hand, impressed. I loved what it let us do in terms of tying these complex qualitative codebooks to underlying quantitative data. But the entire time I was doing this, I think I felt deeply uncomfortable in terms of what it meant for the method and was anything that I was doing, was it okay from a methodological perspective, were others doing this too, should I be doing it better? All of these were questions that I was asking myself.

[0:49:50]

And so, to start tackling some of those, I brought together a convening of researchers in our area at this same conference in human computer interaction. And we started to discuss this idea of using large language models as research tools. One of the things that I loved about this workshop is that it probably can never happen in the same way again

because it was nascent enough that just the idea of using large language models as research tools in a way that is agnostic to method was actually possible, right? So, we brought together people from so many different sub-disciplines that were using LLMs in so many different ways. from many, many different epistemological perspectives. And so, I love that we were able to do this, but one of the things that came out of that workshop is that in particular for the qualitative researchers, they were posing themselves a lot of questions about what this meant for their work, for their own insights, their own process, things like that.

Okay, so this paper that we'll present today, which is an exploration of large language models for qualitative research. I think to Susan's point, some of this is getting into benefits and risks, but I think one of the things we do in this paper where I think the meat really is, is in looking at epistemological tensions. So, where are the places where we think that LLMs really might be fundamentally clashing with some of the values of qualitative research and what are some of the things that could be done to keep these tensions in mind? But I think some of what we would posit in this paper is that these may fundamentally be irresolvable.

Again, this is this very central research question that we're looking at here. What do large language models mean for qualitative research? On the one hand, qualitative research, of course, is fundamentally grounded in this deep engagement with participants and with their data. The context, as you all know, involves a lot of iterative sense-making, immersing yourself in the data, and a contextual understanding of problems that allow researchers to notice subtle patterns and identify common themes.

But over the last few years, if you'll go to the next animation, large language models have transformed many research processes across disciplines. Surveys that our field puts out right now for context will say that as many as 80% of researchers could be using these tools in their practice today. I suspect there may be a little bit of bias there in terms

of who was measured, but nonetheless, we really posit that LLMs distinguish themselves in a research context from prior tools because they offer one, interactions in natural language, two, they're marketed as these general purpose chatbots and assistants that are used across many, many contexts, and three, is that they mimic aspects of human sense-making.

In many ways, the speed of adoption has outpaced our ability to conscientiously reflect on the role of these tools and how they may best be used.

And so, for the last session, yeah, using LLMs for qualitative work, oh, can you go back one, we think might surface unique tensions. So, for example, models could distance researchers from their data, automate sense-making or raise data stewardship concerns, but they could also allow researchers to read broader audiences, explore larger data sets, or speed up tedious parts of the research.

So, now if you'll get to the next slide. Therefore, we really think that the current moment offers this unique opportunity to examine how these tools interact with qualitative values. And we identify the urgent need to understand how LLMs are shaping qualitative research processes before and as these practices solidify. So, this is really an exploration of that.

[0:53:38]

So, our research questions are tackling number one, you'll go, you can just put all of them up. How are large language models being adopted by qualitative researchers today? Number two, what may be gained or lost from adopting LLMs into qualitative research processes and what tensions arise? And three, how might LLMs be ethically adopted into qualitative research processes?

So, from a methods perspective, we conducted semi-structured interviews with 20 qualitative research participants. We wanted a variety of fields and domains, so we recruited across many different domains. Six participants were in human-computer interaction, given that was the venue, so they're probably the most represented, followed by political science and then sociology and communication. So, , we were getting a breadth of researchers, we were recruiting across approaches. So, we have nine who are primarily qualitative and 11 who are mixed method researchers like ourselves.

And lastly, we thought about this question a lot, but we decided in the end to recruit agnostic to large language model use. So, we didn't want people that, we didn't want only people that had used large language models for qualitative research. We really wanted to understand the counter perspective as well and to understand when those things are intentions and we also wanted to see how people were using LLMs who weren't thinking about it as actively.

And then finally, one methodological choice that we made is that the interviews focused on large language model use in data collection and analysis, and so we're focusing primarily on how the researcher interacts with the data since that is what came up as salient during our workshop.

We then analysed the transcripts and the videos using inductive thematic analysis that are inspired by grounded theory. So, I'll take you through the first part of the findings, what I think of as the least interesting part, and then Casey will take over for the more fun part. All of the analysis of the interviews was done by humans. We thought it would be too ironic if we also used LLMs for the study to do the actual analysis. So, I'll talk you through some of this.

First of all, we outlined the main use cases where people were using LLMs in our study. And so, on the data collection front, participants were

using LLMs to help with data collection tasks, so things like creating or adapting study materials. One person here said, “I was feeling stuck, I was not sure what questions to ask, but I had this broad set of topics and so I put them into ChatGPT”. But another person also says, “This juice kind of has not been worth the squeeze. I can write this in the way that I need it to be for recruitment much faster”.

On the data analysis front, we found that participants used LLMs in both inductive and deductive work when it came to qualitative research. And there was a significant set of people who used ChatGPT as this main analysis interface, and they were often left disappointed, dissatisfied, right? So, if you just use ChatGPT once, you put your data in and then that is your experience, usually you think it is subpar, but some participants, similar to what we saw today, actually created these very complex thought through and tested LLM pipelines for annotation, and they tended to find LLMs more useful. That aligns with some of what we're seeing today. I think it's very useful in the coding, especially if we're thinking deductively, but somebody else saying the connections between the themes ended up being better because I myself had spent the time coding it.

[0:57:29]

We also reflect on a number of other tasks that participant mentioned, including using LLMs as, for example, brainstorming buddies for ideation. And so, you know, this could be obvious to some, but not for others here. One of the things that we find here is that qualitative researchers really are already using these tools across the pipeline, despite this lack of consensus, policy and privacy assurances. And it's no shock here, of course, that people can find this technology useful, but there are concerns that arise. So, our participants were actively experimenting with using LLMs. Some were able to leverage them effectively in their work. However, many participants really did express these deep concerns that involve participant privacy. In terms of

potential risks, we identified this set of five key perceived risks of using LLMs for qualitative work. One of them is around this idea of uneven adoption for people that are coming, for example, from different fields, have different amounts of technical expertise or different amounts of access to funding. There's also this deep uncertainty about norms, where people are unsure if the actions that they're taking are sanctioned or if they're okay, and this tends to leave students vulnerable to dual pressures of, specifically in our technological field, there's this pressure to adopt these technologies and to be experimenting with them, but at the same time, you don't have the protection of knowing exactly which of your sanctions are expected.

And then we saw a lot around people navigating privacy implications, questions around validity and evaluation of data and outputs have come up.

And finally, this question of model bias, where models may not be suitable for analysis or understanding specific populations. And in particular, participants were concerned about privacy implications which tended to be amplified by this lack of perceived clear guidance. So, here we had one participant even saying, “You know, sometimes I do put some participants' quotations in, but I intentionally trim out specific keywords, for example, specific product names or functions”. And so, even for us, this left us in a state of wondering, well, is this something that actually should be permitted? Is this a breach of a privacy concern? Things like that. And I'll hand over to Casey now.

Casey Randazzo: All right. So, our results also revealed how LLMs permeate a lot of the different stages of the research process, even ones that we didn't ask about that were beyond data collection and analysis. So, we became really interested, as Marianne said earlier, in those tensions. So, we saw them emerging between the use of LLMs and different themes in qualitative research.

So, with one being that qualitative researchers, as you know, often use an inductive approach or perhaps sometimes grounded theory, which prioritises the way that insights emerge. And that's what we saw on top here. So, on the bottom of the circle, you see the tension that people raised, which is when they interact with an LLM as they're analysing the data. And maybe it's the LLM that's imposing this external interpretation that's actually at odds with what their qualitative research processes and best practices might recommend.

Second is the issue of close engagement with the data. And we found that qualitative researchers really take pride in spending a lot of time with primary sources and interviews and interviewees. But on the other hand, you have an LLM which appears to be able to potentially process some of that work, as we've been discussing, and that might also be intention.

And third is this question of what responsibility qualitative researchers have to participants. Many of them feel that to do the data justice, they need to bear witness to the stories and really spend time with those stories one-on-one and raise questions about what it means for an emotionless machine instead to be raising these questions and analysing these themes or spending time with the data instead of just them.

[1:01:45]

And finally, there's the issue of subjectivity and multiplicity. which has often been seen as a real strength of qualitative research. Different qualitative researchers that we interviewed also approach the same subject matter with a different point of view, and that multiplicity is typically celebrated, right? But when you approach that with an LLM, at least in our current interfaces, it often creates one LLM imposed interpretation, which might homogenise the way that we see data, which also echoes some of what we heard in the other presentations today.

So, let's delve into an example of one of those tensions and actions. So, participants also reported both experiences with and fears of LLMs misinterpreting data from marginalised groups, particularly through the mistranscription of important language from that community, or by making judgements that are not faithful to the values in a community. For example, one participant noted that AI often falsely identifies queer content as hateful, and this quote suggests that an LLM's interpretation of a participant group may be at odds with that community's interests.

And then participants also raised concerns about the potential of LLM biases in that power existing perpetrating harms towards marginalised groups. And as one participant said about LLM training data, most of the literature that's accessible is in the voice of the powerful, right? So, you know, that's a complex issue that requires a sort of re-evaluation of what it means to analyse responsibly, particularly when setting vulnerable populations. And researchers should caution that using these tools may risk that misinterpretation of the data in addition to privacy violations.

So, what are some of the takeaways from these tensions? So, first we saw a sociotechnical breakdown of guidelines, norms, and tooling. The adoption and ubiquity and permeation of these tools are outpacing an organisation's ability to create clear guidance that helps people navigate these changes.

And second was the ubiquitous use of AI. So, this compromises individual researchers' ability to think through how they might use these tools intentionally. And when you're approaching the same interface for tasks all the way from editing to original research, it can be a little tough to see where those lines blur. And some believe that tensions between LLMs and some styles of qualitative work may be fundamental. We know that this is really important to explore and have further research on because LLM curious qualitative researchers may be positioned to

enact some of this slippery slope between some more traditional qualitative approaches and some mixed methods approaches.

Okay, so now for the grand finale of what does this all mean, right? So, we condensed our findings into a resource that provides guidance around key decisions and recommendations. So, the first one. When handling participant data, the first decision you should make is where to run the model and understand who sees the inputs or information you're feeding to the AI. We recommend defaulting to local or open source models so data never leaves your control. And if you must use a proprietary system, assume that the information may be released accidentally and use identifiers.

We also want to emphasise that LLMs may be embedded into the tools that you may think are already protected from privacy concerns. So, before turning on any sort of smart features in Google Docs or a grammar checker, confirm whether that tool is sharing text with third parties or using it for training.

[1:06:03]

Second, we recommend being transparent about how LLMs were used as part of your project and we understand that this may be tough due to stigma surrounding AI use, but being more transparent about how an LLM was used for things like ideation or coding assistance or editing may help shift those perceived threats to one's credibility or the validity of their work. And when you do disclose, make sure to name the model and version you used. As we've also heard, these models change rather quickly, so that may be important for future understanding or interpretation of case studies as a whole over time. And you've just described the prompts you sent to the LLM or workflow at a high level and make sure to explain how you check the model's performance, of course, and this will help reviewers understand what the human did, what the model did, and where interpretive judgements were made.

And when it comes to choosing when to use an LLM, we recommend deciding on it by a case-by-case basis, and as you likely have experienced, chat interfaces blur boundaries and make it easy to drift from harmless assistance from the LLM into analysis that should be conducted by a human. And for that reason, we also recommend planning the tasks you want the LLM to do ahead of time, and then screen those tasks for risks like compliance, originality, bias, privacy and validity. And we suggest avoiding using LLMs when codes rely on specific contextual meeting from a community.

Fourth, validating the outputs of the LLM is critical. You can check the LLM's performance through side-by-side comparisons with human coders on a small representative subset of data. And we also recommend validating how the model edits text or do spot edits with edge cases from marginalised groups. And validation shouldn't just happen once, it should be constant due to the changing data sources and prompts.

And then finally, tool choice matters. So, with chat user interfaces like ChatGPT or Google Gemini, we recommend predefining the AI's permitted task and watching for task drift. If you're using LLMs as part of popular analysis software, like NVivo, to generate qualitative codes, we recommend to always assess biases in the LLM's outputs and confirm privacy policies before use. As LLMs grow in popularity, their presence will be less obvious in those types of tools, especially as we're seeing now with grammar and spelling check and spelling checkers or spreadsheets. I mean, this type of software may quietly call on an embedded LLM without you even realising that it's there. And in those cases, it's critical that you always check what is a part of a piece of software before using it so you can avoid accidentally sending identifiable data to their servers. And if using APIs or pipelines, we suggest open source models and local hosting, and if you must use an

API, always anonymise, of course, anything you share and avoid services that train on user data.

So, the results from this study will hopefully go on to influence the design of the next generation of LLM powered research tools, which is something that we talked about a little bit in the beginning. But we are hoping that as qualitative researchers, we can make recommendations to people designing, like NVivo, and different tools that we use and how these LLMs are embedded. So, while these tools may not be for everyone, we believe in the importance of promoting intentional use which we define as being upfront about what LLMs can and cannot do as a part of the qualitative research process.

So, that is our presentation. Here's our information here. Feel free to reach out to us with any additional questions. We're also looking forward to the Q&A. And thank you for having us speak today. Appreciate it.

[End of Transcript]