Social sciences and AI joining forces: towards new approaches for computational social sciences

Katharina Soemer & Daniela Grunow
Institute of Sociology
Goethe University
Frankfurt am Main, Germany
{soemer,grunow}@soz.uni-frankfurt.de

Steffen Eger
Department Computer Science
& Artificial Intelligence
University of Technology
Nuremberg, Germany
{steffen.eger}@utn.de

Abstract

While computational social science is seen as an interdisciplinary field of data science and social science, it is often overlooked that qualitative social research plays an important role in the various steps of the research process. In this chapter, we argue that structurally integrating qualitative research methods can fill blind spots and reduce bias in computational social science. To provide a roadmap for the systematic integration of qualitative social scientific research in computational social science, we provide an example case of fine-tuning and prompting AI-based large language models with manually annotated data for the task of analyzing social solidarity on social media and political debates. We describe the five-step research process—concept definition and operationalization (step 1), data collection and sampling (step 2), annotation (step 3), computing (step 4) and model evaluation (step 5)—highlighting the complementary contributions of each methodological perspective and skill set. We argue that methods from qualitative and quantitative social research, together with computer science unfold much greater potential than disciplinary approaches for the application of AI methods in the social sciences.

1 Introduction

Recent breakthroughs in computer science, particularly in artificial intelligence (AI), have opened up unprecedented opportunities while also posing new challenges for analyzing social phenomena in the empirical social sciences. Developments such as the proliferation and application of Large Language Models (LLMs) and the availability of computational methods that require less elaborated coding skills open up new avenues for interdisciplinary collaboration. The (anticipated) impact of LLMs on the social sciences is comparable to their impact on many other areas of society. It has been found that LLMs can augment computational social science (CSS) (Ziems et al., 2024), which constitutes an academic field at the intersection of data science and social science (Engel et al., 2021). While the aforementioned developments in AI are already being applied in quantitative (Chen et al., 2025; Erhard et al., 2025) as well as qualitative (Muller et al., 2016; Parfenova et al., 2025; Than et al., 2025) social sciences, we argue that there is a need for the systematic integration of quantitative, qualitative and computational methods into the field of CSS. Therefore, this chapter calls for social scientists to join forces with computer scientists to apply AI.

Quantitative social sciences are typically theory-driven, relying on statistical modeling of large, representative data to describe social phenomena and (deductively) test hypotheses. Qualitative social sciences, in contrast, often take an exploratory approach, deriving theories from data and providing an in-depth analysis of exemplary cases to understand meaning and investigating new social phenomena through interpretive analysis. Computer science is both a theoretical and an applied discipline, so computer scientists typically have knowledge in areas such as mathematics, theoretical computer science, and computer architecture, along with practical skills ranging from software development to implement-

ing algorithms and data structures. Computational methods, often subsumed under the heading of AI—including machine learning, deep learning, natural language processing and large language models—are increasingly applied in quantitative social sciences, but also share characteristics with qualitative approaches, in particular those focusing on interpretative data analysis and the (inductive) generation of theory from data (Lindgren, 2020). Beyond a general understanding of the capabilities and limitations of AI techniques, their application in the social sciences requires expertise in exploring and theorizing novel phenomena and in interpreting results. Application of AI methods further raises questions regarding evaluation criteria for classified data. We argue that computational social science therefore requires interdisciplinary skills and profits greatly from collaboration between computer scientists and social scientists trained in quantitative (empirical-analytical) and qualitative (interpretative) research traditions.

The application of computational methods in the social sciences offers various opportunities, not only for new approaches to classification and prediction but also for collecting research data and accessing novel data sources. For example, political trends can now be analyzed by combining survey data with social media content and corpora of political texts (Kostikova et al., 2024). Similarly, discourses can be studied by linking textual data from newspaper articles with user comments (Kantor & Rafaeli, 2025). However, applying and interpreting these models comes with challenges, particularly in the appropriate (and responsible) use of AI methods. On the one hand, researchers (or research teams) need methodological knowledge on basic principles of machine learning, e.g., how to split training and test data, how to avoid overfitting, how to effectively use LLMs and how to monitor LLM performance.[1] On the other hand, they need an understanding of AIs' limitations, including biases, lack of generalization, reasoning errors, hallucinations and, especially for LLMs, benchmark and data contamination (Sainz et al., 2023) and prompt sensitivity (Mizrahi et al., 2024). There is also no clear set of rules on how to reliably interpret model outputs. While quantitative and computational research competencies, that is an understanding of statistical principles and technical functionalities, can provide a methodological understanding of AI methods, these competencies do not automatically enable researchers to understand limitations and pitfalls when applying such methods. Inductive methods, such as grounded theory, can help to interpret the results of unsupervised models and translate them into theory (Muller et al., 2016). Interdisciplinary skills and collaboration between scholars from computer science, with their technical expertise, quantitative social sciences, with their statistical competences, and qualitative social sciences, with their interpretive understanding, is essential for advancing computational social science. By joining forces, researchers can harness the potential of computational social science to gain deeper insights into the social sphere while maintaining the substantial and methodological standards of social sciences. Based on our joint experiences in cross-disciplinary collaboration and the application of mixed methods research designs, we provide a **roadmap** of how techniques of large-scale data analysis developed in computer science and quantitative social sciences, together with qualitative analytical social sciences approaches, go hand-in-hand to find new solutions to common problems.

2 A roadmap to integrating qualitative research expertise in Computational Social Science: a practical example

To illustrate our point and demonstrate how qualitative, quantitative, and computational methods can effectively be integrated into the research process of computational social science, we present an example: a joint research project using Twitter (now X) data to explore changes in social solidarity across different time periods and topics. The collaboration started in 2020, with a joint MA course taught by two social scientists and a computational scientist to computational science students at the Technical University of Darmstadt. The course was designed to teach students of computational science to collect and annotate data that effectively capture the complex social scientific concept of social solidarity. Over

[1] Even though coding skills become less critical when working with LLMs, the academic use of AI should always be grounded in a solid methodological foundation. Understanding modeling decisions and basic principles is crucial for researchers to properly interpret and evaluate their work.

time, the original text corpus gathered and annotated in this course was partly extended (time-periods and topics covered), thematically narrowed (toward specific forms of social solidarity) and fine-tuned to assess various research questions related to the concept of social solidarity. Research questions assessed included: Can we detect changes in the debate on European solidarity before and after the outbreak of COVID-19 (Ils et al., 2021)? How do design choices that often follow ad hoc practical reasons affect research findings regarding changes in social solidarity towards refugees (Eger et al., 2022)? Was the level of social solidarity higher or lower during the 'long summer of migration' in 2015–2016 compared to during the influx of Ukrainian refugees in 2022? Is there evidence of solidarity bias, i.e., are expressions of solidarity or anti-solidarity related to perceived geographical and cultural proximity, that could be interpreted as preferences for Ukrainian over Syrian refugees (Weber et al., 2023)? Can we trace subtypes of solidarity towards refugees in 150+ years of German political debates using LLMs (Kostikova et al., 2024)?

To answer such research questions, computational social scientists may decide to collect data on the topic from a social media platform and use it to fine-tune an LLM. Although training and testing machine learning models on large datasets and evaluating them with statistical measures may seem like tasks for quantitative social scientists and computer scientists, this widely used approach involves numerous interpretative steps requiring expertise in qualitative methods. We demonstrate the role of quantitative social research skills, qualitative social research skills and skills from computer science for each step in CSS research projects. This provides a roadmap according to which researchers, first, define their object of research and operationalize the definition (step 1); second, collect data and sample from it (step 2); third, annotate their sample and refine annotation rules until an acceptable inter-coder reliability threshold has been reached (step 3), in order to , fourth, train a model, assess its quality and improve it (step 4); and fifth, interpret the model, its performance and summarize the model outcomes, both statistically and descriptively (step 5).[2]

2.1 Step 1: Definition and operationalization

The first step of a research process is to define the phenomenon under study. Often, social phenomena such as in the example case social solidarity—as well as many other social phenomena such as poverty, gender identity or hate speech—are concepts that many people understand on an everyday level. However, defining these concepts in a way that makes them operationalizable in a research setting poses a significant challenge that receives little reflection in computational science applications (Eger et al., 2022) while being the subject of extensive academic debate in the social sciences (Papcunová et al., 2023). The definition and especially the operationalization—that is defining variables to represent or measure a concept—is usually the work of quantitative social scientists. However, this task could be enhanced by skills from qualitative social science. For example, Grigoropoulou & Small (2022, 905) suggest cognitive interviews to ensure that the concepts being analyzed are properly operationalized. In the example case, conducting qualitative interviews with expert annotators from social sciences, non-expert annotators from computational sciences as well as people affected by the concept of analysis, in this case refugees from different countries, could highlight differences in how different groups understand *social solidarity*. Systematically considering the perspective of those affected by solidarity as well as annotators' understandings, and linking these perspectives to formal definitions ensures data quality and reduces bias. In our case, the result would be an applicable working definition of social solidarity.

2.2 Step 2: Data collection and sampling

After defining the object of research, scholars identify potential data sources. Traditionally, quantitative social researchers work with pre-collected data such as surveys or administrative data and use statistical samples of the entire target population for analysis. In contrast, qualitative social researchers often work with data that they collect themselves, such as

[2]Step 2 analytically consists of two steps, we describe them together as they are interlinked in our example.

through interviews or observations, often applying iterative, statistically non-representative approaches to reaching thematic saturation regarding their research problem (Saunders et al., 2018). Data used in computational analysis comes in many different forms that may resemble either approach but reliability in the data collection process is usually not a main concern in computer science. This is where qualitative social scientific methods of sampling and data collection can inform the interdisciplinary research process in computational social science. Computer science methods allow for the automated collection of large amounts of data from many different data sources: For example, parliamentary speeches can be of interest to see developments in populist rhetoric (Erhard et al., 2025). Stereotypical biases such as gender biases can be examinedin many different data sources: large music lyric corpora (Chen et al., 2025), political debates (Walter et al., 2021), book corpora (Madhusudan et al., 2025). Data from public Telegram channels can be analyzed for conspiracy theories (Pustet et al., 2024). For researching social solidarity on social media, one or more platforms have to be chosen. Depending on the topic of interest, this can be a mainstream or a fringe platform. The platform choice also depends on if researchers want to analyze text, images, audios, videos, or have a multimodal research setting. Which platform is being chosen also depends on data access. Social media data collection typically involves gathering large datasets from one or more social media platforms, either through web scraping or using an application programming interface (API). One advantage of an API is that it indicates whether the platform supports data collection and, if functioning properly, it simplifies the data collection process (Nyhuis, 2021, 34). The technical infrastructure, such as setting up and maintaining a database, and navigating the API specifications, requires computer science expertise. Additionally, certain conventions and habits from computer science, such as version control, are beneficial to a research team.

Once the data has been collected, a sample—that is a small portion of the large data set—is selected for manual annotation (see step 3 for details). In traditional quantitative social research, there are standardized methods for sampling data. Researchers follow specific protocols to avoid (statistical) bias and collect data that best represents their object of research. Many aspects of sampling are the subject of methodological debate, such as discussions about probability and non-probability sampling (see for example Vehovar et al. (2016)). Sampling works differently in qualitative research. The aim of qualitative sampling methods is not to measure or quantify whether the sample meets certain criteria. Instead, methods such as theoretical sampling and contrasting cases are part of constant comparison: In theoretical sampling, the steps of data collection, coding, and analysis occur simultaneously, or iteratively, with decisions about what to sample next being guided by the theory emerging from the collected data rather than by predetermined criteria (Glaser & Strauss, 1967, 45–77). Glaser & Strauss (1967, 62–65) outline several key differences between theoretical and statistical sampling, including contrasts in purpose, adequacy, flexibility, and the role as well as the achievement of sample saturation. Contrasting cases is an integral part of the qualitative technique of constant comparative method application. In constant comparative method applications, researchers compare incidents across diverse or extreme cases to clarify, expand, and test emerging theoretical categories (Glaser & Strauss, 1967, 101–115). Theoretical sampling and contrasting cases are elements of Grounded Theory, a methodological approach to generate theory inductively from data (Glaser & Strauss, 1967). Rather than seeking statistical representativeness, qualitative researchers select cases that promise maximum (theoretical) insight. This sampling approach can enrich computational social science by identifying edge cases and outliers that quantitative approaches might overlook, but which are critical to understanding the full scope of a social phenomenon. In our social solidarity example, purely quantitative sampling methods, such as collecting data with specific keywords may produce systematic substantial biases in the data by missing out on important cases (Eger et al., 2022). Such errors can be avoided by employing and evaluating results from different sampling techniques. Interpretative analysis of such differences helps understand what is left out and what is captured by different sampling techniques (Eger et al., 2022). More generally, this step involves deriving more relevant keywords from the data, or adding keywords that might contrast the phenomenon to better fine-tune the LLMs (step 4).

Grigoropoulou & Small (2022, 904) point out that while research data in the social sciences is typically collected and curated by researchers and for scholarly purposes, there are also

available large datasets that were not collected for research, but for other purposes. For example, some data sets are created by actors outside of academia and for purposes other than academic analysis. Therefore, Grigoropoulou & Small (2022, 904) suggest conducting qualitative interviews with providers of datasets, such as social media platforms that provide specific datasets, to understand how they came together and what potential biases these data contain. They also suggest conducting a pre-study with open-ended interviews of users, in this case from social media platforms, to then be able to quantitatively survey them about their behavior and their intentions. While a qualitative interview with the providers of a social media platform in this analysis of social solidarity is unlikely (and would probably not yield much insight, as how social media platforms curate their content is usually classified as confidential business information), qualitative researcher skills are still beneficial in this step, as it can be seen in the following example.

Eger et al. (2022) analyzed social solidarity on Twitter and showed how methods of data selection can bias findings and substantial conclusions. The authors compared different sampling decisions such as sampling for keywords versus sampling for hashtags, and compared different metadata, such as post language and user location. Their substantial findings differed depending on these design choices. Quantitative social researchers thus have means to statistically validate such findings and test the influence of design choices on outcomes. Another task that requires expertise in quantitative social science is to compare findings with those of representative studies on the topic "as 'ground truth', to double-check the validity of measurements obtained from online data" (Eger et al., 2022, 154). Researchers with qualitative social research skills could contribute to inform sampling design choices by familiarizing the research team with the field, for example discussions found under a specific keyword or hashtag. Jikeli et al. (2023) labeled potentially antisemitic content on Twitter. They reported finding a relatively low percentage of antisemitic posts when searching for the insult "kikes" compared to another antisemitic slur, because posts containing this keyword also included discussions about two sports figures named "Kiké" (Jikeli et al., 2023). Qualitative techniques established in digital ethnography, that is the observation of and participation in online discourse (Marwick, 2013), could help researchers understand their object of interest and prevent such pitfalls from being overlooked or data from being misinterpreted. It is further important for researchers to be aware of which aspects of the phenomenon are being covered by the data they analyze, and which aspects might be missing, as it will be incorporated into a model (step 4). It is worth noting that sampling is also discussed in subfields of linguistics such as corpus linguistics (Lüdeling, 2011), which themselves are related to computational fields such as natural language processing (NLP).

2.3 Step 3: Annotation

The next step is annotation. The term annotation "refers to the process of adding metadata information to the text in order to augment a computer's capability to Perform Natural Language Processing" (Pustejovsky & Stubbs, 2012, ix). Annotation is a linguistic method that corresponds closely with a range of qualitative sociological methods for coding different types of content (Kanygin & Koretckaia, 2021; Parfenova et al., 2025). Such content can be text (Sanguinetti et al., 2018), images (Goffman, 1979), videos (Medina Serrano et al., 2020), or a combination of these types (Steffen, 2024).

In the study we use as an example, textual data has been annotated. Where available, emojis were used to help facilitate the coding process. Annotation requires a manual with clear instructions to make annotators interchangeable (Artstein, 2017, 308). The quality of the annotation manual is crucial. Suggestive instructions can result in what is known as *instructor bias* (Parmar et al., 2023). In the annotation process, annotators provide labels to predefined units of data according to the task at hand. For example, annotators could label social media posts as probably or confidently biased against a particular group, as Jikeli et al. (2023) did for antisemitism on Twitter. Annotators could then be asked to label whether this hatred is implicit or explicit, as done by Kennedy et al. (2022), who labeled posts from the fringe platform Gab. Annotators could also label whether the content contains incitement to violence against the target group, as done by Sanguinetti et al. (2018), who distinguish four types of implicit and explicit incitement against immigrants in Italy on Twitter. In our example case, social scientifically trained (expert) annotators, together with (non-social

scientifically trained) non-expert annotators labeled social media posts from Twitter that contain specific hashtags related to the phenomenon of social solidarity. The labels used were *solidarity*, *anti-solidarity* and *other*, with *other* capturing "tweets that can be interpreted ambiguously or [...] tweets that do not fit into the (anti-)solidarity categories" (Weber et al., 2023, 356–357).

There are statistical measures to assess the performance of annotators, such as kappa statistics, which are typically used to measure inter-annotator reliability. Specific thresholds determine what level of (dis-)agreement between annotators is considered acceptable. There is a necessity for quantitative social researchers to assess the scope of measures like kappa in annotating social media content. When evaluating inter-annotator reliability, research projects on hatred on social media and other social phenomena often report Cohen's kappa with a cutoff value from biostatistics (Landis & Koch, 1977). But just as there are different levels of acceptance of statistical measures in different disciplines, Cohen's kappa yet needs to be examined regarding what is acceptable, and what should be rejected. The accepted range of statistical measures can vary across disciplines: when quantifying the explained variance in a regression analysis, an R^2 with the same value will probably be accepted in social science and rejected in biostatistics, since analyzed variables usually account for much less variance in the social sciences. Social phenomena like social solidarity are usually complex to define, and can be understood differently, depending on different perspectives, as indicated above. Quantitative social researchers should conduct methodological research on the range of acceptable values in that field, just as measures like R^2 are being reported with different ranges of acceptable values in different fields.

Beyond this analysis, which is done by quantitative social scientists, the process of annotation evaluation is another way to integrate qualitative methods. Rather than focusing solely on how often annotators disagreed (a quantitative question), researchers can substantially enrich their work and ensure the annotation's validity and reliability by evaluating why annotators disagreed (a qualitative question). This would involve a qualitative analysis of the content of the disagreement. This approach recognizes that not all disagreements are random: Disagreements in annotation may result from a lack of expertise among annotators or from an unclear definition of a concept. Disagreement may also indicate that certain social phenomena cannot be fully captured by rigid definitions, or that language itself is inherently ambiguous in certain contexts. Qualitative researchers are therefore trained to perform iterative coding and evaluation until the concept under study is adequately defined and sufficiently understood by all members of the research team. In such cases, the agreement levels obtained by human coders are the threshold against which the model performance is evaluated (Fornaciari et al., 2021; Plank, 2022; Fleisig et al., 2023). Cases in which annotators disagree could be used as a sample for qualitative content analysis to better understand the sources of disagreement. This can help to improve the annotation manual. Recently, some researchers in computational sciences such as NLP have considered disagreements in annotation as a valuable signal (Fornaciari et al., 2021; Plank, 2022; Sorensen et al., 2025) rather than as noise which should be eliminated.

Another key aspect of manual annotation is the annotation tool. Annotation can be done in Excel spreadsheets, but can be more efficient when using specific software: Jikeli et al. (2024) for example, designed a specific annotation tool for social media content. Niekler et al. (2023, 326–327) point out that social scientists usually lack software development skills, and therefore turn to using software that has already been developed. Computer scientists could help develop the software needed for the research task, for example, annotation software tailored to the specific needs of a research project (Klie et al., 2018). Whereas we used excel spreadsheets in our social solidarity project, we realize that the task could have been performed more comfortably, and with more control over the process by using a software for qualitative data analysis, such as MaxQDA or Atlas.ti (Kuckartz & Kuckartz, 2002; Hwang, 2008).[3]

[3]Many more aspects of annotation can be discussed in regard to the integration of quantitative, qualitative, and computational methods. These include discussions about annotator diversity (Bui et al., 2025), questions about how task definitions bias annotators (Parmar et al., 2023) and the extent to which LLMs can substitute for human annotators (Tan et al., 2024; Weber & Reichardt, 2024).

2.4 Step 4: Computing

Computational models are essential for computational social science in order to make large-scale predictions on data. Machine learning and AI are ideal candidates for such computational models. AI may then act as an aide for large-scale human annotation. Machine learning and deep learning can be roughly distinguished into supervised and unsupervised methods (with a variety of overlaps). A supervised method is a "technique that generates a function mapping from inputs to a fixed set of labels (the desired output)" (Pustejovsky & Stubbs, 2012, 141),while an unsupervised method is "[a]ny technique that tries to find structure from an input set of unlabeled data" (Pustejovsky & Stubbs, 2012, 141). Annotated corpora are used in machine learning and deep learning approaches: in supervised methods, annotated corpora serve as the gold standard for training models. With LLMs, new paradigms of steering models have become available. As these models are **pre-trained** on gigantic corpora, they already have a very good understanding of many tasks. In this case, model training (now called **fine-tuning**) may not be necessary anymore. Instead, **prompting** becomes an important craft, that is, giving suitable **instructions** to general-purpose LLMs to execute the task at hand. **In-context learning**, where researchers e.g. provide a few examples (also called **few-shot learning**), then oftentimes replaces traditional model training that would involve large amounts of training instances. It is worth noting that due to the sheer size of some LLMs (for example, GPT4, which is among the models with the most parameters at the time of writing), fine-tuning is often not even possible anymore for most users, including traditional universities. It is also essential to point out that while these novel LLM-based approaches have led to unprecedented quality improvements in AI within a few years, they concurrently still have myriads of **limitations** (Kostikova et al., 2025). Not only that *prompting may be brittle*, where small variations may lead to very different outcomes (Mizrahi et al., 2024; Leiter & Eger, 2024), but models may also not *generalize* well to new domains (e.g., when classifying solidarity in political debates instead of social media), they may *hallucinate*, i.e., deliver fake information which can lead users astray, and their *multilingual*, e.g., when classifying non-English texts, and *multimodal abilities*, e.g., when the input involves other modalities besides text, may be limited. **Ethical concerns** must also be mentioned: LLMs may have large negative environmental impact due to high electricity costs, they may have been pre-trained on *pirated data* (without due compensation of the human data creators), they may lack transparency, they may pose *security threats*, they may perpetuate *biases* of their inventors(Bang et al., 2024; Hartmann et al., 2023) and the underlying data(Feng et al., 2023), and LLM quality is typically strongly bound to the size of the LLMs (measured in the number of trainable parameters), where smaller models may perform substantially worse. AI users, including computational social scientists, should be aware of these limitations and collaborate with computational and qualitative researchers to map potential failure cases and prioritize their relevance. Naive, uneducated use of AI may have severe negative consequences, especially in an age where using these models is easier than ever before as a consequence of democratization of AI.

In our example scenario, researchers fine-tuned a smaller LLM with the task of finding *solidarity*, *anti-solidarity*, and *other* posts (Ils et al., 2021; Weber et al., 2023). Fine-tuning means 'teaching' a model that has been broadly trained with relevant data to perform a specific task, in this case detecting social solidarity and anti-solidarity, particularly well. Follow-up work then showed that larger LLMs such as GPT4o vastly outperform the smaller fine-tuned models, especially in harder and more fine-grained classification tasks involving solidarity frames (Kostikova et al., 2024). LLMs may then also be seen as very cost-effective replacements of human annotators, being orders of magnitude cheaper while exhibiting only marginal decreases in classification quality. This is where a symbiosis of quantitative, qualitative, and computational skill sets and methods may have the greatest potential. Computer scientists and quantitative social researchers could jointly decide which state-of-the-art model best fits the task, because together they are able to understand the model infrastructure and requirements, and to compare model performances on similar tasks to the analysis they plan to conduct.

2.5 Step 5: Model evaluation

In modern AI approaches, annotated corpora constitute evaluation benchmarks for assessing model quality. These corpora provide a "ground truth" against which model performance is quantitatively evaluated: statistical measures such as precision, recall, and F1 as their harmonic mean quantify how well a model performs in identifying the correct labels (Pustejovsky & Stubbs, 2012, 30–31). More complex tasks as enabled by generative AI often leverage more sophisticated evaluation methods, including the paradigm of "LLM-as-a-judge" (Gu et al., 2025) where LLMs evaluate the responses of LLMs, sometimes leading to biases in evaluation when LLMs prefer their own output (Wang et al., 2025; Liu et al., 2024). Another pitfall of evaluation in the era of LLMs is when models have been pre-trained on the test data (referred to as benchmark, task or data contamination) or when researchers leak information from the test data. This may lead to inflated evaluation scores (Balloccu et al., 2024) where LLMs can simply *memorize* the answers, an issue that practitioners from non-AI fields, including social scientists, may not be sufficiently aware of.

Before analysis, the data needs to be preprocessed: Missing data needs to be taken care of. In statistical terms, data may be *missing completely at random* (MCAR), *missing at random* (MAR) and *missing not at random* (MNAR) (Little & Rubin, 2019). Completely at random means the data is missing for no causal reason, but pure randomly. If data is missing at random, there is a reason for it being missing, but the reason is outside of the dependent variable. Not missing at random means that there is a relation between what is being analyzed and what is missing from the data. We need quantitative social researchers to make a stronger point on what the equivalents to these are in cases of computational social science, such as collecting data from social media platforms. In our example, data can be missing completely at random when posts are missing because a user decides to delete their profile to spend less time on social media and more time with their cat. This means that a post they made on the topic (independent of its content) is not collected and therefore not part of the dataset. If researchers collect data over a longer amount of time and their server goes down for a few hours, their data is missing at random, since the missing data is not coincidental, but is still independent of the content of the data. Data is not missing at random if posts are removed by the platform or users are banned from a platform. Therefore, posts made on the topic of interest are not collected—but the fact that they are missing will be due to the content of the post, that is the dependent variable, and not to a reason outside of it. In addition, specific social groups are systematically underrepresented on certain platforms (for example, older people or people living in poor regions of the world, where internet access is highly selective). Thus, these groups' views on a topic under study, for example social solidarity, will not be represented in the data. It is the job of quantitative social researchers to reflect on bias, but qualitative skills can be helpful, such as field knowledge gained with online ethnographies that we mentioned above. Computational skills can help identify technical problems, e.g., by reading the log file if a server has been offline.

Researchers must also decide which pre-trained model will best perform their tasks. Making these decisions means having technical knowledge about state-of-the-art models available and their requirements, which is a computer science skill. It also means having statistical knowledge of evaluation criteria, such as F1 scores, to compare model performance, which is a skill from quantitative social sciences. However, qualitative methodological knowledge is also required for model evaluation. On the one hand, researchers could perform qualitative content analysis to examine content that the model failed to classify correctly in order to understand its blind spots. Analogous to the annotation disagreement analysis, there could be a categorization of cases that the model did not capture. On the other hand, since the model in our example case is fine-tuned or prompted, it has learned most of its knowledge on other data (the pre-training data). Therefore, it may have inherent biases, especially when dealing with sensitive issues such as social solidarity, and explicit anti-solidarity (Feng et al., 2023). Qualitative research approaches can help to develop theories from observations and to show which aspects of the data are relevant findings and which aspects are just noise (Grigoropoulou & Small, 2022, 95). A qualitative analysis of error cases would allow researchers to understand what kind of external knowledge (or bias) is being brought into the research setting by the model itself. This knowledge could then be brought into the

iterative process and used to reduce or even fix bias. It could also contribute to what is called explainable AI (The Royal Society, 2021; Leiter et al., 2024).

3 Conclusion

We have demonstrated that integrating methodologies at each stage can lead to more complex insights, reduced bias, and fewer errors. Table 1 highlights the necessary expertise from each field in each step of the research process. As with any interdisciplinary and multi-method research project, close collaboration among researchers is essential. At present, we need researchers from different disciplines, i.e., quantitative social researchers, qualitative social researchers and computer scientists to gain an understanding of their methodological expertise as well as sources of bias and pitfalls in their research. Ideally, the next generation of scholars in social computing would receive interdisciplinary training, capturing expertise in computational as well as qualitative and quantitative social scientific skills. Researchers also need to understand which other traditions or subjects complement their expertise to adequately address research questions in computational social science. We believe that interdisciplinary methodological education holds immense potential for students interested in computational social science to grow into the application of suitable methods for their research projects. Within academia, this also means an exchange between faculties, in teaching as well as in research.

Steps \ Expertise	Quantitative Social Science	Qualitative Social Science	Computer Science
[1] Concept Definition & Operationalization	Formal concept definition	Nuancing definition toward applicable working definition	Operational fit with data
[2] Data Collection & Sampling	Assessing representativeness	Gaining and integrating field knowledge	Technical infrastructure: pipelines for data collection
[3] Annotation	Reliability metrics	Codebook development, resolving disagreement	Providing annotation software
[4] Computing	Analyze (missing) data	Prompt development	Preprocessing, model selection, training, prompting, in-context learning, or fine-tuning
[5] Model Evaluation	Statistical evaluation	Interpretation, bias and blind spot detection	Monitoring evaluation strategies

Table 1: Core contributions by research expertise for each step in computational social science research project.

We are aware that the current times are uncertain regarding developments in AI, specifically the potential of LLMs. It is impossible to predict the capacities or relevance of these models within the next few years, or even by the time this book is published. We also find ourselves in the so-called post-API age (Freelon, 2018), where data access for researchers is becoming increasingly uncertain. Nevertheless, we don't see how these uncertainties will diminish the relevance of the field in the future. We advocated for the systematic integration of qualitative, quantitative, and computational methods in computational social science. Although computational social science is typically viewed as an interdisciplinary field combining data science and quantitative social science, our roadmap emphasizes the

essential role of qualitative social research expertise at every stage of the research process. We argue that computer scientists and quantitative social science researchers benefit from increased model validity and bias reduction when qualitative social research methods and expertise are included in a computational social science research setting.

We illustrated this argument based on a specific example: a joint research project on expressions of social solidarity in social media data. The five key steps of the project—concept definition and operationalization, data collection and sampling, annotation, computing and model evaluation—demonstrate how distinct methodological traditions contribute complementary strengths. Quantitative social scientists contribute theory-driven models, statistical validation, and reliability metrics. Qualitative researchers contribute conceptual nuance, field-specific knowledge, and interpretive depth to sampling, annotation, and model interpretation. Computer scientists provide the technical foundation for data acquisition, model design, and infrastructure maintenance and, of course, expertise in computing models.

The broader implication of this roadmap is that interdisciplinary collaboration must evolve from mere cooperation to structural integration. This requires designing projects that explicitly incorporate various types of expertise from the beginning and assembling teams that acknowledge how different types of expertise can contribute to better outcomes. It also means rethinking academic training and institutional structures to support future computational social science scholars to become methodologically multilingual and grounded in social science theory—able to understand and apply a variety of approaches and communicate across disciplinary boundaries.

References

Ron Artstein. Inter-annotator agreement. *Handbook of Linguistic Annotation*, pp. 297–313, 2017.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source llms. In Yvette Graham and Matthew Purver (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 67–93, St. Julian's, Malta, 2024. Association for Computational Linguistics.

Yejin Bang, Delong Chen, Nayeon Lee, and Pascale Fung. Measuring political bias in large language models: What is said and how it is said. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11142–11159. Association for Computational Linguistics, 2024.

Minh Duc Bui, Katharina Von Der Wense, and Anne Lauscher. Multiˆ3hate: Multimodal, multilingual, and multicultural hate speech detection with vision–language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 9714–9731, 2025.

Danqing Chen, Adithi Satish, Rasul Khanbayov, Carolin M. Schuster, and Georg Groh. Tuning into bias: A computational study of gender bias in song lyrics. In *Proceedings of the Joint Workshop on Language Technology for Equality, Diversity and Inclusion and the Workshop on Computational Linguistics for Linguistic Complexity (LaTeCH-CLfL 2025)*, 2025.

Steffen Eger, Dan Liu, and Daniela Grunow. Measuring social solidarity during crisis: The role of design choices. *Journal of Social Computing*, 3(2):139–157, 2022.

Uwe Engel, Anabel Quan-Haase, Sunny Xun Liu, and Lars Lyberg. *Introduction to the Handbook of Computational Social Science*, pp. 1–12. Routledge, 2021.

Lukas Erhard, Sebastian Hanke, Ulrich Remer, Agnieszka Falenska, and Raphael H. Heiberger. Popbert: Detecting populism and its host ideologies in the german bundestag. *Political Analysis*, 33:1–17, 2025.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11737–11762, 2023.

Eve Fleisig, Rediet Abebe, and Dan Klein. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6715–6726, 2023.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2591–2597, 2021.

Deen Freelon. Computational research in the post-api age. *Political Communication*, 35(4): 665–668, 2018.

Barney G. Glaser and Anselm L. Strauss. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine, Chicago, 1967.

Erving Goffman. *Gender Advertisements*. Harper & Row, New York, 1979.

Nikolitsa Grigoropoulou and Mario L. Small. The data revolution in social science needs qualitative research. *Nature Human Behaviour*, 6(7):904–906, 2022.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge, 2025. URL https://arxiv.org/abs/2411.15594.

Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. The political ideology of conversational ai: Converging evidence on chatgpt's pro-environmental, left-libertarian orientation, 2023. URL https://arxiv.org/abs/2301.01768.

Sungsoo Hwang. Utilizing qualitative data analysis software: A review of atlas.ti. *Social Science Computer Review*, 26(4):519–527, 2008.

Alexandra Ils, Dan Liu, Daniela Grunow, and Steffen Eger. Changes in european solidarity before and during covid-19: Evidence from a large crowd- and expert-annotated twitter dataset. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1623–1637, 2021.

Gunther Jikeli, Sameer Karali, Daniel Miehling, and Katharina Soemer. Antisemitic messages? a guide to high-quality annotation and a labeled dataset of tweets. *arXiv preprint arXiv:2304.14599*, 2023.

Gunther Jikeli, Katharina Soemer, and Sameer Karali. Annotating live messages on social media: Testing the efficiency of the annothate – live data annotation portal. *Journal of Computational Social Science*, 7:571–585, 2024.

Avner Kantor and Sheifaz Rafaeli. Conversations with data: How data journalism affects online comments in the new york times. In *Proceedings of the 54th Hawaii International Conference on System Sciences*, pp. 2386–2395, 2025.

Gennady Kanygin and Viktoria Koretckaia. Analytical coding: Performing qualitative data analysis based on programming principles. *The Qualitative Report*, 26(2):316–333, 2021.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, et al. Introducing the gab hate corpus: Defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, 56: 79–108, 2022.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In Dongyan Zhao (ed.), *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pp. 5–9, Santa Fe, New Mexico, 2018. Association for Computational Linguistics.

Aida Kostikova, Dominik Beese, Benjamin Paassen, Ole Pütz, Gregor Wiedemann, and Steffen Eger. Fine-grained detection of solidarity for women and migrants in 155 years of german parliamentary debates. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5884–5907, 2024.

Aida Kostikova, Zhipin Wang, Deidamea Bajri, Ole Pütz, Benjamin Paaßen, and Steffen Eger. Lllms: A data-driven survey of evolving research on limitations of large language models, 2025. URL https://arxiv.org/abs/2505.19240.

Anne Kuckartz and Udo Kuckartz. Qualitative text analysis with maxqda, 2002.

J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

Christoph Leiter and Steffen Eger. PrExMe! large scale prompt exploration of open source llms for machine translation and summarization evaluation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11481–11506. Association for Computational Linguistics, 2024.

Christoph Leiter, Piyawat Lertvittayakumjorn, Marina Fomicheva, Wei Zhao, Yang Gao, and Steffen Eger. Towards explainable evaluation metrics for machine translation. *Journal of Machine Learning Research*, 25(75):1–49, 2024.

Simon Lindgren. *Data Theory: Interpretive Sociology and Computational Methods*. Polity, Cambridge, 2020.

Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, Hoboken, NJ, 3rd edition, 2019.

Yiqi Liu, Nafise Moosavi, and Chenghua Lin. Llms as narcissistic evaluators: When ego inflates evaluation scores. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12688–12701. Association for Computational Linguistics, 2024.

Anke Lüdeling. *Corpora in Linguistics: Sampling and Annotation*, pp. 191–210. Science History Publications/USA, New York, 2011.

Sangmitra Madhusudan, Robert Morabito, Skye Reid, Nikta Gohari Sadr, and Ali Emami. Fine-tuned LLMs are "time capsules" for tracking societal bias through books. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2329–2358. Association for Computational Linguistics, 2025.

Alice E. Marwick. *Ethnographic and Qualitative Research on Twitter*, pp. 109–122. Peter Lang, New York, 2013.

Juan Carlos Medina Serrano, Orestis Papakyriakopoulos, and Simon Hegelich. Dancing to the partisan beat: A first analysis of political communication on tiktok. In *Proceedings of the 12th ACM Conference on Web Science*, pp. 257–266, 2020.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? a call for multi-prompt llm evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949, 2024.

Michael Muller, Shion Guha, Eric P.S. Baumer, David Mimno, and N. Sadat Shami. Machine learning and grounded theory method: Convergence, divergence, and combination. In *Proceedings of the 2016 ACM International Conference on Supporting Group Work*, pp. 3–8, 2016.

Andreas Niekler, Christian Kahmann, Manuel Burghardt, and Gerhard Heyer. The interactive leipzig corpus miner: An extensible and adaptable text analysis tool for content analysis. *Publizistik*, 68(3):325–354, 2023.

Dominic Nyhuis. *Application Programming Interfaces and Web Data for Social Research*, pp. 45–60. Routledge, London, 2021.

Jana Papcunová, Marcel Martončik, Denisa Fedáková, Michal Kentoš, Miroslava Bozogáňová, Ivan Srba, Robert Moro, Matúš Pikuliak, Marián Šimko, and Matúš Adamkovič. Hate speech operationalization: A preliminary examination of hate speech indicators and their structure. *Complex & Intelligent Systems*, 9(3):2827–2842, 2023.

Angelina Parfenova, Andreas Marfurt, Jürgen Pfeffer, and Alexander Denzler. Text annotation via inductive coding: Comparing human experts to llms in qualitative data analysis. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 6456–6469, 2025.

Mihir Parmar, Swaroop Mishra, Mor Geva, and Chitta Baral. Don't blame the annotator: Bias already starts in the annotation instructions. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 1779–1789, 2023.

Barbara Plank. The 'problem' of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 10671–10682, 2022.

James Pustejovsky and Amber Stubbs. *Natural Language Annotation for Machine Learning: A Guide to Corpus-Building for Applications*. O'Reilly Media, 2012.

Milena Pustet, Elisabeth Steffen, and Helena Mihaljevic. Detection of conspiracy theories beyond keyword bias in german-language telegram using large language models. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, 2024.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. Nlp evaluation in trouble: On the need to measure llm data contamination for each benchmark. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10776–10787. Association for Computational Linguistics, 2023.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. An italian twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pp. 3550–3558, 2018.

Benjamin Saunders, Julius Sim, Tom Kingstone, Shula Baker, Jackie Waterfield, Bernadette Bartlam, Heather Burroughs, and Clare Jinks. Saturation in qualitative research: Exploring its conceptualization and operationalization. *Quality & Quantity*, 52:1893–1907, 2018.

Taylor Sorensen, Pushkar Mishra, Roma Patel, Michael Henry Tessler, Michiel Bakker, Georgina Evans, Iason Gabriel, Noah Goodman, and Verena Rieser. Value profiles for encoding human variation, 2025. URL https://arxiv.org/abs/2503.15484.

Elisabeth Steffen. More than memes: A multimodal topic modeling approach to conspiracy theories on telegram. *arXiv preprint arXiv:2410.08642*, 2024.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 930–957, 2024.

Nga Than, Leanne Fan, Tina Law, Laura K. Nelson, and Leslie McCall. Updating "the future of coding": Qualitative coding with generative large language models. *Sociological Methods & Research*, 54(3):849–888, 2025.

The Royal Society. Ai and interpretability: A policy briefing, 2021. URL https://ec.europa.eu/futurium/en/system/files/ged/ai-and-interpretability-policybriefing_creative_commons.pdf.

Vasja Vehovar, Vera Toepoel, and Stephanie Steinmetz. *Non-Probability Sampling*, pp. 329–345. SAGE Publications Ltd, London, 2016.

Tobias Walter, Celina Kirschner, Steffen Eger, Goran Glavas, Anne Lauscher, and Simone Paolo Ponzetto. Diachronic Analysis of German Parliamentary Proceedings: Ideological Shifts through the Lens of Political Biases . In *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pp. 51–60. IEEE Computer Society, 2021.

Xiao Wang, Daniil Larionov, Siwei Wu, Yiqi Liu, Steffen Eger, Nafise Sadat Moosavi, and Chenghua Lin. Contrastscore: Towards higher quality, less biased, more efficient evaluation metrics with contrastive evaluation, 2025. URL https://arxiv.org/abs/2504.02106.

Maximilian Weber and Merle Reichardt. Evaluation is all you need. prompting generative large language models for annotation tasks in the social sciences. a primer using open models. *arXiv preprint arXiv:2401.00284*, 2024.

Maximilian Weber, Daniela Grunow, Yanran Chen, and Steffen Eger. Social solidarity with ukrainian and syrian refugees in the twitter discourse: A comparison between 2015 and 2022. *European Societies*, 26(2):346–373, 2023.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291, 2024.