Draft chapter from Baker, Ryan S.; Barany, Amanda (forthcoming) *Artificial Intelligence in Qualitative Research: New Possibilities.* Philadelphia, PA: Vassant Press.

Find the latest chapters online at
https://learninganalytics.upenn.edu/book/ai-qualitative-research.html

Copyright Ryan S. Baker and Amanda Barany, 2026, all rights reserved.

Cite this chapter as Baker, Ryan S.; Barany, Amanda (forthcoming) *Artificial Intelligence in Qualitative Research: New Possibilities.* Ch. 7: Collecting and Selecting Data. Philadelphia, PA: Vassant Press.

Chapter 7. Collecting and Selecting Data.
Version 1.02, 25 October 2026

- What this step contains, what was done before AI

Data collection and data selection (sometimes referred to as data sampling) involves a set of decisions about what empirical materials will be gathered and how the materials will be gathered. Depending on research goals and contexts, this may include (1) identifying cases, sites, participants, and artifacts, (2) establishing criteria for inclusion and exclusion in a study, and (3) deciding which data representations will be created, collected, and used (e.g., interviews, observations, documents, digital traces).

Data collection and selection are deeply connected. In a sense, all collection decisions are selection decisions. That said, these two steps are enacted separately in many research projectsUsing an interview-based study design as an example, a *retrospective sampling* workflow might involve conducting a large number of interviews and then subsequently deciding which to analyze. A *prospective sampling* workflow might reverse the order: a researcher might first select which participants to interview, and then conduct the interviews to collect the chosen data. A third possibility occurs in *constructing a research corpus from existing artifacts*: the selection and collection step are one and the same. In traditions such as grounded theory (Glaser & Strauss, 1967) and ethnography (Spradley, 1980), data collection can be cyclical or

iterative, with early rounds of data analysis informing what data researchers need to collect next, and the choice of participants or data sample shifting as new insights emerge.

In practice, choices around selection are often pragmatic (like when a researcher must use a convenience sample—data chosen primarily because it is readily accessible rather than because it is the best to study the phenomenon of interest—due to limited access to potential participants), but the framing of a study's goals, guiding frameworks, and research questions inform what data sources are most appropriate (Maxwell, 2012; Ravitch and Carl, 2019). Some qualitative traditions are furthermore framed around the idea that data is not passively "collected" but is instead actively "produced" through research relationships and practices. Grounded theory (Bryant & Charmaz, 2010), narrative inquiry (Riessman, 2008), and to an extent discourse analysis (Gee, 1999) all organize around the idea that researchers engage with participants, texts, and settings, and that this interaction shapes what the data ultimately is.

Practices for collection/generation can include interviewing, think-aloud protocols, focus groups, field observation, a search for data sets already available, curation of corpuses from already extant materials, field note writing, document gathering, artifact photography, screen capture, and the systematic archiving of digital traces, among others. All of these forms of documentation and capture convert experiences, interactions, and records into materials that can be systematically examined. Each of them carries its own affordances, constraints, and implications for what becomes visible in the analysis; data is shaped, filtered, and structured through researcher decisions and interactions (Ravitch & Carl, 2019). In the next chapters we will discuss how data is prepared and organized (including data segmentation) after collection, but it is important to note that preparation and segmentation (and even initial sense-making) often occur concurrently with collection and selection, and are often treated as a single step.In addition, steps to identify and preserve key contextual information often must occur during collection and selection or be lost forever (Emerson et al., 2011). As such, data collection/generation often already entails preliminary analytic judgment, blurring the boundary between gathering and analysis.

Within each of these collection/generation practices, key methodological decisions are made. Consider three common sets of practices as examples: interviews, observations, and accessing pre-existing data sources. For interviews, topics must be selected and questions must be written (or the choice must be made to *not* write questions in advance). For observational studies, protocols must be developed (even if the protocol is free-form or informal). When curating existing materials, search criteria must be developed (even if informally). When creating field notes, practices must be selected. In each case, a general methodological approach is translated into a specific set of choices about how to apply methods, that is informed by the

goals and context of the project and the perspectives of the researchers. For example, an ethnographic researcher might adjust their interview protocol as they build understanding of how participants frame their activity. Researchers in other paradigms might emphasize maintaining the consistency of interview design to more easily compare across responses. Some researchers might conduct observations holistically, whereas other researchers might restrict themselves to pre-selected behavioral coding categories, in order to maintain focus. In creating a corpus from extant materials, some researchers might remain open to shifting their focus to initially unexpected phenomena, while others might rely on pre-specified search strategies and selection rules to maintain transparency and systematic coverage.

Within these broader methodological approaches, researchers also make concrete, more "in the weeds", procedural decisions about how data collection will actually occur in practice. There are a large number of potential such decisions. For example, depending on the goals and context of an observational research project, a field observer may sample brief snapshots, choose a standardized time interval (e.g., 20 seconds, five minutes, an hour), or engage in prolonged observation. While these decisions can sometimes be left until a later stage (see Ch. 10 for a discussion of data segmentation and structuring) for the collection and analysis of existing materials, these decisions must be made at the time of data collection for observational, interview, and participatory methods. Similarly, key aspects of the researcher's role such as how (and whether) they influence the setting, and their degree and forms of disclosure to participants, must be established. Beyond this, scope choices -- and the related development of inclusion and exclusion criteria -- have implications for richness of interpretation, the potential for theoretical saturation, and the risk of bias (in the multiple senses of that word -- see Chapter 14 for further discussion). Decisions about how and what to record (e.g. audio versus video; verbatim versus summarized notes) may at times seem driven by logistical or ethical considerations, but can play a major role in what can eventually be gleaned from the data. Together, by shaping what data are gathered, these many seemingly-small decisions shape what phenomena can become analytically visible (Maxwell, 2012).

- Applications of AI

The applications of AI to data collection and selection are wide-ranging, and vary in maturity; we discuss a spectrum of practices, from those already established to exciting possibilities that remain largely speculative (the fun part of writing a book like this one). We begin with what is already happening: researchers are using AI to discover, access, and scrape existing data, leveraging LLMs to identify relevant data sets, construct and refine API queries, and writing code to extract material from webpages that lack formal APIs. We then turn to emerging possibilities for selecting cases from a corpus or data set, where techniques such as

LLM-as-a-judge, embedding-based semantic retrieval, and multidimensional selection strategies allow researchers to find cases that match a target description even when the same idea is expressed in very different terms, and to sample for diversity, typicality, contrast, or anomalies. From there we consider new possibilities for iterative and interactive selection, where case selection is revisited as analysis proceeds—extending longstanding traditions such as theoretical sampling, and drawing an analogy to active learning methods from machine learning. We next examine how AI can support active data collection by humans, including the development and refinement of interview and observation protocols, logistical and preparatory tasks, and the construction of custom data collection tools. We then discuss the use of AI as an interviewer, an application only used by a few research groups, but has been used in large-scale studies of tens of thousands of participants across many languages; we discuss how this development opens new affordances for interviewing in difficult contexts and on sensitive topics. Finally, we take up the controversial prospect of AI as an interviewee, distinguishing relatively low-risk uses such as piloting protocols and training novice interviewers from the higher-risk substitution of synthetic responses for human participants, a practice which we argue warrants substantial caution.

Across these applications, AI does not occupy a single role; it can be used in line with several of the work relationships proposed in Chapter 4, even within a single project. When an AI system executes a clearly specified search, filtering, or scraping task, it functions mainly as a tool. When it proposes questions, cases, or protocols under human direction, it may resemble a junior colleague. When repeated interaction reshapes both the researcher's thinking and the AI's subsequent contributions, the combined human–AI system more closely resembles a centaur. And when the AI is deliberately asked to identify omissions, question assumptions, or generate counterexamples, it serves as a critic or provocateur.

Scraping and accessing data. AI can support researchers in obtaining existing data. Before a researcher can access a data set, they must first identify what data set they want to work with. For researchers working in unfamiliar domains or across disciplinary boundaries, simply knowing what data sources are available can be a significant barrier. A researcher can leverage an LLM to help them discover relevant data sets by describing their research goals and asking the LLM to suggest publicly available data sources, archives, or repositories that might contain relevant materials. The LLM can also help the researcher evaluate whether a candidate data set is likely to meet their needs — for example, by discussing its scope, structure, known limitations, or terms of access — though of course the researcher should verify these details independently, as LLMs can generate plausible but inaccurate descriptions of specific cases, especially when information is sparse or documentation poor (cf. Huang et al., 2025).

Many existing large data sets such as government records and social media data can only practically be accessed through APIs (application programming interfaces). For many such data sources, which can be complex in structure and massive in size, it is infeasible (or even disallowed) for an individual researcher to simply download the entirety of the data. If a researcher attempted to download the entirety of a social media archive, for example, they might discover that the dataset is several terabytes in size—quickly overwhelming the 512GB storage capacity on their personal laptop. Instead, they must define their scope of interest and then construct a query to retrieve a relevant subset - say, posts in a certain time frame, about a specific topic, or posted by a certain group of users. Researchers often do not engage in the full process of case selection at this step because it can be difficult to construct a perfect query that downloads exactly the cases they want to eventually work with. Instead, in practice, many researchers retrieve a broad subset of data during initial access and then conduct additional filtering or selection as a separate later step. The second author of this book used this approach in her dissertation; to understand social media learning, a large sample of posts was downloaded from users who chatted with at least one other *person*. The sample was later narrowed offline to include only users who chatted with at least one *friend*, to focus on this group. This approach is also often adopted when the researcher's preferred analytic tools do not integrate directly with the API infrastructure, or when researchers wish to preserve a complete record of all materials matching a query to support transparency, auditability, or later re-analysis.

The process of writing a query in the necessary format and then storing the data appropriately can be a significant technical challenge for many researchers, however. An LLM can assist a researcher in writing their API query. This can increase efficiency for more experienced programmers (many of whom today use LLMs extensively -- Peng et al., 2023) and can make it possible for less experienced programmers to construct queries at all. LLMs can also help refine queries. For example, an LLM can recommend changes to queries to reduce data access charges when access is metered, or can help the researcher modify their query if they determine that their query did not obtain exactly the data they wanted. An example of this is shown in Figure 7-1.

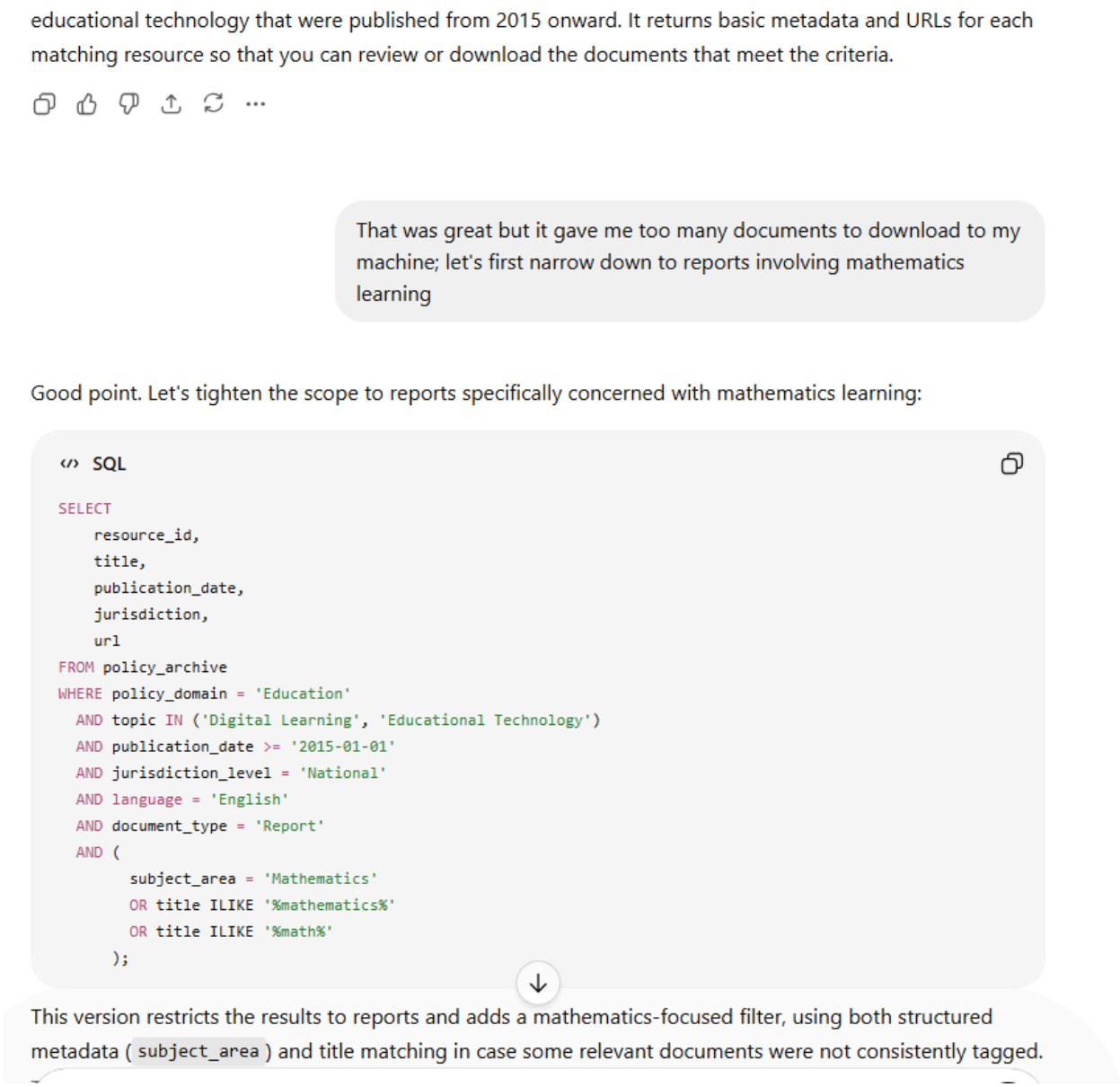


Figure 7-1. A ChatGPT Session Iterating a database search query.

LLMs are also now used to write code to scrape user-facing data from public webpages (though this must of course be done in compliance with webpage terms of use and other ethical considerations -- Zook et al., 2017). This can be particularly valuable for qualitative researchers who need to collect data from organizational websites, online forums, news archives, or other web-based data sources that lack formal APIs but contain relevant materials. To support web scraping, LLMs can help identify the correct HTML (webpage code) elements to target, can parse dynamic content and structured fields, and can help integrate resources split across multiple pages. An example of this is shown in Figure 7-2.

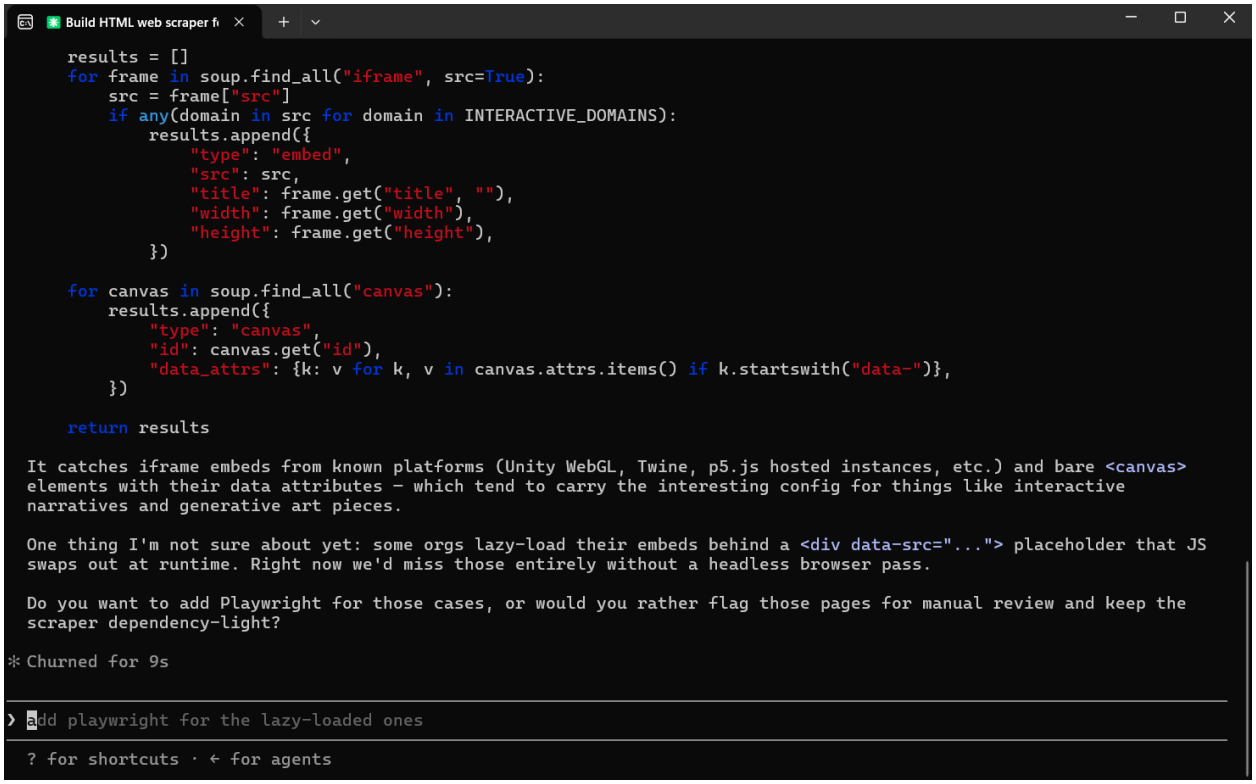


Figure 7-2. A Claude Code session developing a web scraping tool.

Selecting cases. Prior to contemporary AI, a researcher might have searched for specific keywords across cases, such as looking for specific words in tweets (Kreis, 2017), or public policy statements involving specific terms (Grimmer & Stewart, 2013). Other projects developed complex sets of rules, called regular expressions, that detail what combinations of words or word-stems would include or exclude the data from the sample. However, regular expressions tend to take more expertise and engineering effort to develop (Friedl, 2006), and cannot easily be used to capture situations where the same ideas can be expressed in several different ways.

To improve on these approaches, more computationally-oriented projects have also used machine learning/natural language processing (previous-generation AI, as discussed in Chapter 4) to develop classifiers of whether cases are relevant (Grimmer & Stewart, 2013), sometimes using regular expressions as a first step to label very clear cases (Ratner et al., 2017). This approach is effective, but also time-consuming, both due to time spent manually classifying cases (a key early step towards building the classifiers) and time spent building and validating the classifiers. Other researchers have used topic modeling (Blei, Ng, & Jordan, 2013), a previous-generation natural language processing method, to split cases bottom-up into categories based on commonly-found keywords, before selecting one or more of these

categories for further analysis. Each of these approaches is still used, but contemporary AI has further expanded the range of methods that can be used to select cases.

LLMs can be used in a wide variety of ways to help select cases for analysis from an existing corpus or data set (or an actively collected one, once it has been collected). One of the simplest approaches is to feed the entire data corpus through an LLM, case-by-case (one API call per case), and have it judge whether each case is relevant. This can be done using three potential approaches. First, the researcher may provide a textual description of the target cases. Second, the researcher may select a few example cases in advance and ask the LLM to match them. These approaches are sometimes referred to as LLM-as-a-judge (Zheng et al., 2023); we will discuss another (more common) application of this method in Chapter 13. The third approach involves requesting that the LLM extract the most representative cases (e.g. Ohsaki & Kaneko, 2025). While these types of LLM-based filtering and selection can be expensive and may take substantial time to run (depending on data set size), they can identify cases where the same idea is expressed with very different text, and can be done with less human effort than machine learning or developing regular expressions. For example, a person expressing their dissatisfaction with changes in government tax policies could do so in a wide range of ways -- from reasoned argument, to angry attacks, to sarcastic expressions. Trying to find all of the specific combinations of words that could identify this range of forms of communication would be highly difficult for a researcher, whereas an LLM might be able to identify a wider range of cases for review. A related approach is embedding-based retrieval, also called semantic search (Reimers & Gurevych, 2019). In this approach, each case is converted into a set of numbers using a text encoder (a tool used by LLMs to convert text to numerical representations) and is then compared to a similarly-converted target description or to converted example cases. Cases are then selected by ranking their mathematical similarity to the target description or example cases. Like LLM-as-a-judge, this approach can identify cases where the same idea is expressed with different text, but it can be more scalable to large corpora, as it does not involve running a full LLM for every case. However, it can be less sensitive to subtle differences in meaning than those full LLMs.

Forms of case selection that leverage multiple data sets or types of variables are also possible. A researcher intending to qualitatively analyze one form of data might select cases using another set of variables. Without using AI, a researcher might do this by filtering on one or more specific variables. Researchers have also done this by using various forms of data mining (previous-generation AI) and data analysis to select cases using multidimensional data sets. For example, Leary et al. (2021) analyzed log data from teachers' usage of a curriculum planning tool and found that teachers varied in terms of two dimensions of use: frequency and variability. They then looked at existing interview data from the teachers with high variability in order to

understand the implications of that variability. A second set of approaches has used clustering and latent class methods to identify subgroups of cases within behavioral or interaction data, followed up by researchers qualitatively examining data associated with specific clusters. For example, in Wedel & Kamakura's (2000) market segmentation research, clusters of consumers were algorithmically identified through behavioral data and then interpreted in terms of customer experiences and motivations. Other projects have used predictive AI models to identify unusual or theoretically important cases. For example, cases with unexpectedly high or low outcomes relative to model predictions have been selected for further qualitative analysis (Seawright & Gerring, 2008), an approach sometimes referred to as model-based outlier detection. This approach allows researchers to focus on cases where existing theoretical explanations appeared insufficient (because predictions were inaccurate). In yet other studies, multimodal analytics methods have combined behavioral log data with additional signals such as video, audio, or physiological measures to identify distinctive interaction patterns, which can then be investigated qualitatively using other available data sources (Worsley, 2018). In these ways, pre-generative-AI machine learning and data mining methods have often functioned as tools for identifying theoretically meaningful cases across data sets, even when the data being analyzed qualitatively were quite different from the data used to select the cases.

Generative AI expands the range of strategies available for case selection. When selecting constructs using a second data set, the LLMs' ability to find semantic themes in data (see Chapter 13 on qualitative coding) can be leveraged to identify cases where a construct is present, and then these categories can be used to select data for further qualitative analysis. For example, policy documents from regional governments worldwide could be scanned by an LLM to identify regions with policies with specific attributes. Then, social media data could be searched for local opinions on those policies, assisting a researcher who is not fluent in the full set of languages necessary for this task. LLMs can also discover categories in a more meaning-oriented fashion than previous topic modeling approaches (this is essentially the same method as discovering codebooks, which will be discussed in chapter 11, but with a different goal). They can also suggest additional keywords for keyword-based search by offering synonyms or conceptually-related terms. Finally, they can help refine regular expressions to improve their effectiveness at finding the desired cases. This could involve helping researchers evaluate and iterate their existing keyword lists or regular expression patterns by generating edge cases or counterexamples — text that *should* match the keywords/regular expressions but wouldn't, or text that *would* match but shouldn't. Using a natural language dialogue with an LLM can also be a more natural way to iterate than the prior approach of manually refining queries. By increasing the effectiveness of past approaches, but retaining useful artifacts such as regular

expressions or keyword lists, a researcher can maintain transparency but leverage some of the benefits of LLMs for capturing the many ways the same meaning can be expressed.

These considerations and techniques have utility for a researcher whose goal is to exhaustively find and analyze every case involving a certain topic or set of properties — depth within a single, well-defined slice of the data. This approach is common in qualitative research; a researcher may want to find every case where a participant discussed challenges using a user interface, or every case where a student expresses boredom during learning. But exhaustive selection is only one strategy among many (Patton, 2002), and the same tools can serve a range of selection goals that prioritize breadth, contrast, relationship, or change rather than coverage of a single topic.

One such goal is coverage of relevant variation, where the AI identifies a maximally diverse set of cases along specified dimensions — diversity or extreme sampling — so that important groups, contexts, or conditions are sufficiently represented; a researcher investigating how people justify their positions on a controversial public policy might want cases that differ maximally in ideological stance and in the types of justifications offered. A related strategy is representative or typical case sampling, where the aim is the reverse — finding the case most central to its group, such as the student whose attitudes toward a new curriculum are most typical within each school studied. A second goal is contrastive or theory-testing selection, finding cases that agree on every dimension except one, in order to probe an emerging explanation, such as two students with highly similar academic histories who diverged in whether to enroll in a university. A third is anomaly-driven selection (also called anomaly detection), which surfaces cases atypical of their group (Seawright & Gerring, 2008) — for example, a voice for peace in a country and time point where most social media posts are pro-war. A fourth, common in discourse analysis and narrative inquiry, is relationship-driven selection, where intertextual or contextual linkages — such as recurring references to a particular policy document — warrant expanding the corpus to follow those connections. A fifth is temporal selection, identifying cases in which a meaningful shift occurs (or does not occur) over time, such as a person changing how they express their gender identity within a social network. This last type is particularly difficult to accomplish with keywords or embeddings alone, because it requires comparing meaning across time within a single case rather than simply matching content at a single point.

Overall, as these examples show, LLMs (and machine learning and data mining as well) can select cases against a variety of criteria — exhaustive coverage, relevant variation, contrast, anomaly, relationship, and temporal change — offering extensive options to a researcher.

Iterative selection. As we discussed earlier in the chapter, case selection is not always conducted only a single time. Many workflows return to case selection as analysis proceeds, particularly when early findings reveal gaps in the data, raise new questions, or suggest that additional types of cases are needed to test or refine emerging interpretations (Glaser & Strauss, 1967). This can produce an overall process where case selection is iterative, with the researcher going back and forth between case selection and sense-making or analysis. Historically, this might have been a manual procedure, with the researcher going back and identifying new cases themselves. In grounded theory work, for example, a researcher might use theoretical sampling (Glaser & Strauss, 1967) to iteratively select cases that can help develop or refine emerging theoretical explanations. For example, after identifying a potential explanation for why some schools adopt a new curriculum more readily than others, a researcher might deliberately search for additional cases that test or elaborate that explanation, such as schools with similar conditions that did not adopt the curriculum.

Iterative selection can now be AI-assisted. AI can help researchers identify how to modify their queries; for example, the researcher can tell the LLM what they think is missing, and ask it to suggest a different query to capture those examples. They can also use LLMs to figure out what is missing (in accordance with one of many views of what sufficiency is or should be -- e.g. Wutich et al., 2024). Take, for instance, a researcher trying to determine when they have reached theoretical saturation -- the point where data collection can cease because new data no longer brings new insights, properties, or themes to the developing theory (Corbin & Strauss, 2014). A researcher could provide the LLM with examples of cases already selected and a description of the emerging theoretical framework, and ask it to identify cases in the corpus that would challenge, extend, or fill gaps in that framework. They could also add data iteratively in smaller chunks, analyze the data as it comes in, and then ask the LLM to identify when theoretical saturation might be reached based on diminishing returns of new themes (De Paoli & Mathis, 2025) This can even move towards a paradigm where data selection is no longer even iterative but becomes fully *interactive* -- small numbers of cases are repeatedly searched for, identified, and analyzed, up until the point where a criterion such as theoretical saturation is reached. In practice, this might look like a structured dialogue: the researcher asks the LLM to select an initial batch of cases. Then the researcher reviews them, and then describes in natural language what is missing or over-represented — for example, 'these cases all involve citizens in electoral democracies, but I need examples from liberal democracies as well' or 'too many of these are expressing the same frustration; I need cases that describe positive experiences.' The LLM then adjusts its selection criteria accordingly, and the cycle repeats. This mirrors how qualitative researchers have previously refined their sampling criteria through engagement with

data (e.g. Miles et al., 2014), but extends the process to corpora too large for a researcher to browse directly.

Of course, theoretical saturation is not the only reason a researcher might return to case selection during analysis. The selection goals described earlier in this chapter do not apply only at the front end of a study; each can also form the basis of iterative selection once coding is underway. A researcher may discover gaps in coverage of relevant variation, encounter cases that invite contrastive or theory-testing comparison, notice anomalies that warrant probing for robustness, uncover relationships that justify expanding the corpus, or identify temporal factors that merit closer analysis. These different logics often overlap in practice, but they imply different criteria for what constitutes a useful additional case. For example, a researcher might discover an unexpected pattern—such as participants from rural settings describing a phenomenon quite differently than urban participants—and seek additional rural cases to assess the scope of that pattern (anomaly-driven and coverage-driven). Or they might realize partway through analysis that their sample underrepresents a particular group and return to case selection to address that gap (coverage-driven). In each of these scenarios, the same logic of LLM-assisted iterative selection applies: the researcher selects the specific analytic goal—whether filling a gap, testing a contrast, or following a newly identified relationship—and the LLM searches the broader data set for cases that support that next step in the analysis. In this interactive form, case selection moves beyond straightforward tool use. The AI's proposed cases reshape the researcher's understanding of what is present, absent, typical, or anomalous, while the researcher's interpretations redirect the AI's subsequent searches. This evokes the centaur metaphor from Chapter 4: the evolving selection strategy is produced through the combined human–AI system rather than being completely specified by either in advance.

In these scenarios thus far, the human researcher manages and drives the process, and the AI plays a secondary, supportive role. However, as Ch. 4 discusses, deeper forms of human-AI collaboration are also possible. An example of this can be seen in a subfield of machine learning called active learning (Settles, 2009). Active learning is a procedure used in classical machine learning to more efficiently develop models to classify data. After the human selects the overall modeling goal and provides a small set of labeled cases (e.g. here are emails that are spam or not spam), the algorithm repeatedly statistically identifies the cases that would be most informative to label next (to help it judge better), and asks the human to label those cases. In active learning, the model and the human annotator work together in a loop: the model identifies uncertain or potentially informative cases, the human provides labels, and the model updates, repeating the cycle until its ability to predict stops improving. Though the goals of this procedure (improved model performance) differ from the goals of qualitative research, the procedure may be useful to qualitative researchers. Approaches for qualitative research inspired

by this analogy could use AI as a process manager that helps the human optimize and focus their own effort, while still leaving key decisions about what to explore to the human researcher. In this role, the AI can help track coverage across the data, identify underexplored or ambiguous areas, and suggest cases that are likely to be most informative -- with the researcher deciding which suggestions to follow up on, how to categorize data going forward, and where the AI needs to delve further into the data to clear up the researcher's uncertainties. As such, it may be possible to adapt algorithms from active learning to the goal of creating interactive processes for case selection, that leverage human sense-making (see chapter 9) and AI ability to search into a highly efficient process for selecting cases within large-scale data.

Supporting active data collection (by humans). In many paradigms and projects, the first step (after deciding what to research) is to collect data, through methods such as interviews or observations. AI can support these processes in many ways. First, it can support researchers in preparing for data collection, such as developing, enhancing, and refining interview protocols (Parker et al., 2023b). For example, a researcher can share their research questions, conceptual framework, and a draft interview protocol with an LLM and ask it to identify gaps in topic coverage, suggest plans for potential follow-up probes, or flag questions that have issues, such as being leading or double-barrelled questions or likely to produce thin responses. For researchers working across linguistic or cultural contexts, an LLM can assist in adapting interview protocols — for instance, adjusting phrasing for age-appropriateness, identifying culturally specific assumptions in how questions are framed, or suggesting alternative ways to ask about sensitive topics (cf. Mani Adhikari et al., 2025, who did something similar for structured questionnaires). The LLM can also be asked to anticipate how different types of participants might interpret a question — surfacing potential ambiguities or assumptions embedded in the wording that the researcher, who is close to the study, might not notice. LLMs can also provide sample answers that a participant might provide, to help a researcher understand how their questions might be understood (e.g. Parker et al., 2023a). Beyond improving individual questions, an LLM can also assist in structuring the overall flow of an interview, suggesting sequences that build rapport, introduce topics progressively, and position more sensitive questions appropriately. In addition, if a researcher does not have an initial draft protocol, the researcher can provide their research questions and theoretical perspectives to an LLM, and it can give them a starting point that the researcher can edit and improve on (Parker et al., 2023a), much as a junior graduate student might create a first draft for a more senior researcher (harking back to our discussion of metaphors for AI-human collaboration in Chapter 4). Question generation can also be targeted to specific analytic goals — for example, a researcher planning to conduct narrative analysis might prompt the LLM for questions that elicit

temporal sequences and turning points, while a researcher interested in argumentation might ask for questions that invite participants to articulate contrasts or justifications.

AI can also play an important role within interview paradigms that adapt interviews to specific situations using context or data (e.g. data-driven retrospective interviewing -- El-Nasr et al., 2015; detector-driven classroom interviewing -- Baker et al., 2024). AI models of different types can be used in a range of ways to make sampling more adaptive and precise, such as selecting participants for interviews (e.g. Baker et al., 2024; also see Palinkas et al., 2015) or identifying specific situations where an interview would be appropriate (e.g. Baker et al., 2024). AI can also be used to support a researcher in adapting interviews, for example by informing interviewers with key data about the interviewee or their context (e.g. El-Nasr et al., 2015; Baker et al., 2024), by selecting artifacts that can be used within interviews (e.g. El-Nasr et al., 2015). These types of AI use could also support ethnographic fieldwork, where a researcher might leverage an LLM to review and synthesize the day's field notes, to identify emerging patterns or unresolved questions (see chapter 9 on sensemaking), to critique emerging interpretations as a critical friend (cf. Hayes, 2025), and to help prioritize what to explore in the next day's observations or interviews.

LLMs can also provide support for observation protocols. A researcher planning structured or semi-structured observations can use an LLM to brainstorm candidate behavioral indicators for constructs of interest, building on existing theory or general knowledge (and connecting back to some of the choices discussed in Chapter 6). It can also review a draft set of observation categories for internal consistency, coverage, and omissions, or suggest edge cases that stress-test category boundaries. Beyond helping refine individual categories, an LLM can help a researcher think through the broader structure of an observation protocol — for instance, advising on whether time-based sampling, event-based sampling, or continuous observation best fits the research questions, given what is known about the frequency, duration, and temporal clustering of the target phenomena.

An LLM can also help researchers anticipate what contextual details to document alongside primary observations — physical layout, participant configurations, ambient conditions, and other features of the setting that are easy to overlook in the moment but may prove essential for interpretation later. As discussed earlier in this chapter, such contextual information must often be captured at the time of data collection or be lost entirely, and an LLM can serve as a useful prompt to think through what might matter before entering the field. For researchers observing the same type of activity across different sites or cultural contexts, an LLM can help identify assumptions embedded in the protocol that may be specific to one setting — for example, expectations about turn-taking norms, spatial arrangements, or the roles of particular actors —

and suggest adaptations for each context. Finally, an LLM can support researchers in thinking through how their presence and role in the setting may shape the phenomena they observe, helping them develop strategies for managing their impacts and generating prompts for reflexive field notes. As with interview protocol development, these uses of LLMs for observation protocols do not replace the researcher's judgment; they function as a structured interlocutor (cf. Schroeder et al., 2025) that can help surface blind spots, stress-test decisions, and help the researcher ensure that the protocol reflects both the demands of the research questions and the realities of the settings in which data will be collected. In the language of Chapter 4, this is the role of critic or provocateur: the AI offers values through supplying critique and pushback where needed to choices that might otherwise not be fully thought through by the researcher.

Beyond protocol design, AI can assist with a range of logistical and preparatory tasks that surround data collection. For example, an LLM can help draft informed-consent language tailored to a study's specific procedures and participant population (subject, of course, to institutional human subjects review -- but LLMs can support development of documents for that as well[1]), generate or refine recruitment materials such as flyers or email scripts, or assist in constructing observation schedules that balance coverage across sites, time periods, and participant groups. While these tasks are not analytic in nature, they can consume considerable researcher time and often require navigating specific institutional or regulatory language; LLMs can provide useful starting drafts that a researcher can then adapt.

Another important potential contribution is in developing custom tools to collect interview or observational data. Contemporary mixed methods researchers often use apps on handheld devices or tablets that allow observers to tap coded behaviors in real time, timestamp field notes, toggle between predefined observation category sets, automatically cue up which participant to observe next and when to stop, and otherwise formalize protocols in ways that make deviation from them a conscious decision rather than an unnoticed drift. These tools can also support synchronization with other forms of data — for instance, by matching exact observation times to corresponding moments in video recordings, enabling precise alignment across data sources (Bosch et al., 2016). Several such tools have been developed for specific research communities (Ocumpaugh et al., 2015; Friard & Gamba, 2016; Ashwin et al., 2023; Hutt et al., 2023), but they have historically taken considerable time, technical expertise, and funding to build (e.g. Hollands & Bakir, 2015), limiting their use to research teams with internal programming capacity or extensive development budgets. One lesson from this literature is that the design of the tool shapes what is observed and recorded (Martin & Bateson, 2007), making

[1] A technical advance that has been a source of extra time for the first author of this book to spend with his children

it often valuable to design a tool tailored to the research goals rather than use an off-the-shelf tool. AI-assisted programming tools and vibe coding tools (Delaney, 2025) — such as Claude Code and Gemini/Google AI Studio — open the use of custom apps of this nature to a broader number of researchers who might find it useful for their research goals and interests. A researcher can now describe the functionality they need in natural language, iterate on a working prototype through dialogue with an LLM, and arrive at a usable data collection tool in a fraction of the time and cost previously required. This does not eliminate the need for careful specification of what the tool should do — the researcher must still make the methodological decisions about what to record, how to structure observations, and what categories to use. Here, the tool metaphor is especially apt: the AI efficiently translates those methodological decisions into a functioning instrument, considerably lowering the barriers to custom tool development.

AI as interviewer. An emerging application of AI to collect data is having the AI conduct interviews, either by itself or in partnership with a human interviewer -- indeed, this area has now matured sufficiently to generate a literature review article (Kupiec, 2024). Early work in this space predated the current generation of LLMs; for example, Xiao et al. (2020) used a chatbot built around a large number of pre-written dialogue moves to conduct semi-structured interviews, finding that participants provided longer and more detailed responses compared to traditional web surveys. As of the time of this writing, the best known example of having an LLM conduct interviews is Anthropic's study on public attitudes toward AI, where it conducted over 80,000 interviews about people's hopes, fears, and lived experiences with AI (Huang et al., 2026). In each of these interviews, the AI interviewer asked a set list of core questions but adapted its follow-up probes based on each participant's responses, creating a semi-structured interview. For example, when participants initially described wanting AI for productivity, the interviewer probed for what that productivity would ultimately enable, often surfacing deeper motivations such as spending more time with family[2]. Interviews were conducted in 70 different languages, to increase representation of different cultures (although there was some inherent selection bias due to sampling Claude users rather than the full population of AI users). The AI interviewer's ability to probe and follow up on individual responses while applying the same overall protocol to tens of thousands of participants represents something genuinely new: the ability to probe into individual perspectives in depth at a scale that would be logistically infeasible for human interviewers to accomplish. Both in this study and others, AI interviewers have been able to conduct interviews on complex and nuanced topics, such as step-by-step elicitation of mental models (e.g. Geiecke & Jaravel, 2025), producing data that is amenable to qualitative analysis.

[2] As discussed for the first author of this book, on the previous page

The simplest and most obvious value proposition of AI interviewing is scale. Aside from Anthropic's large-scale interview study, one researcher used AI voice mode to interview over 1000 pub owners in Ireland about the current price of beer in their restaurants (Lawrence, 2026), and other more academically-focused projects have rapidly conducted hundreds of interviews (Chopra & Haaland, 2023; Geiecke & Jaravel, 2025), maintaining similar tone across every interview and reducing between-interviewer effects in large studies (Chopra & Haaland, 2023). With AI, broader and more comprehensive sampling becomes possible, compared to what could be achieved by most teams of human interviewers. Higher consistency also becomes possible for aspects of the interview where variation is not desired. Being able to conduct interviews across languages also has the benefit of reducing the degree to which phenomena are understood through research on cultural convenience samples. As noted in (Henrich et al., 2010), most psychological research has been conducted on undergraduate students in Western democracies, and may be unrepresentative of humans as a whole. AI interviewers have the opportunity to be scaled much more internationally and interculturally.

AI interviewers can also conduct interviews in contexts where human interviewing would be difficult. For example, an interview could be conducted immediately after an event of interest (cf. Baker et al., 2024) at times that would be infeasible for a human interviewer, such as interviewing a shift worker at 2am. Interviews could also be conducted in contexts where a human interviewer cannot easily go, such as a learner in their home or at multiple field sites. This can also be done in concert with some of the detector-driven interviewing approaches discussed earlier in this chapter. For instance, Wei et al. (2026) (shown in Figure 7-3) used data to determine which students were struggling to successfully compile programs, and then used an AI interviewer to interview the student in the moment when the code was finally successfully compiled, in order to learn how the student was able to resolve their errors. This style of interviewing could be used in a range of places (e.g., learners at home, workers spread out across several countries, scientists at remote field stations) and in a range of situations where immediate interviewing increases the quality and relevance of interviews.

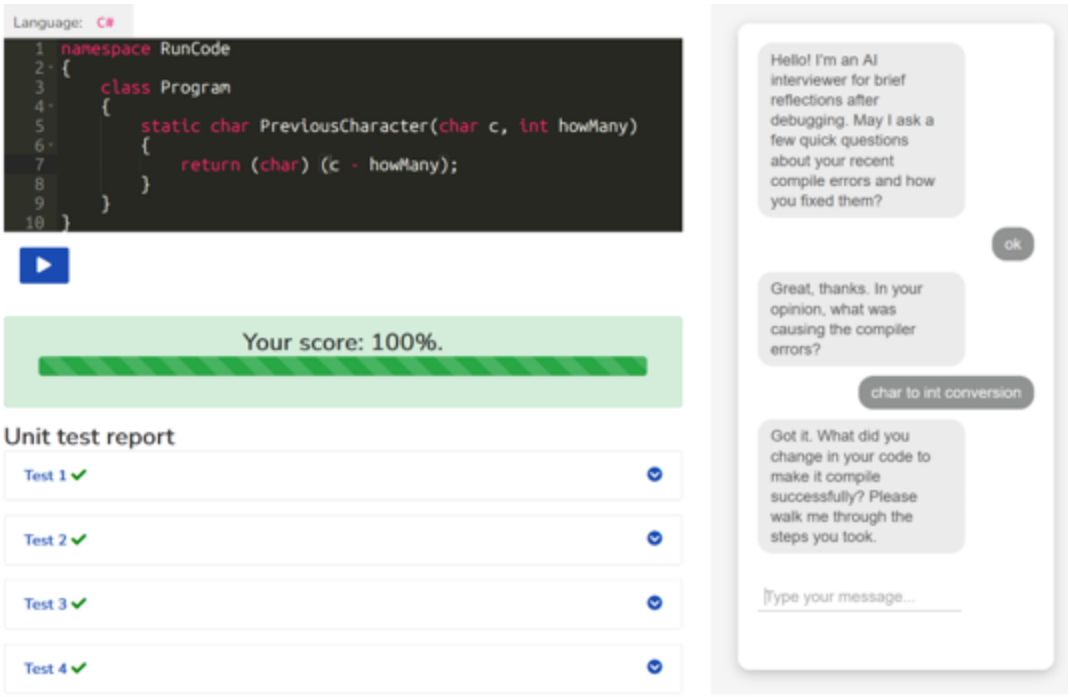


FIGURE 7-3. The interface of the AI interviewer seen by the student in (Wei et al., 2026), interviewing students immediately after they successfully solved a programming challenge (correct on all unit tests).

AI interviewers may also have different affordances for interviewing about sensitive topics. On the one hand, an AI may be seen as more neutral or impartial than a human interviewer. For example, Geiecke and Jaravel (2025) note that French voters asked for their political preferences in an upcoming election reported preferring to share their opinion with an AI interviewer than a human interviewer, due to its "non-judgemental" nature. On the other hand, participants may have concerns about their responses being permanently recorded and linked to the other data the LLM has about them. AI may also not be as good at handling certain sensitive topics as humans, or may not be perceived by participants to be as good. These aspects of AI's affordances as an interviewer remain uncertain, and will need to be explored by the field in greater detail. In general, users' overall attitudes towards AI and human interviews appear to be comparable, even for the current early generation of AI interviewers (Cuevas et al., 2025; Wuttke et al., 2025).

Finally, it is also possible for AI and humans to conduct interviews together. In these cases, the human might lead the interview, with the AI suggesting follow-up questions or interpretations in real time, for the interviewer to consider (Zhang et al., 2025). As of this writing, we are not aware of published studies deploying this approach, but one study interviewed human interviewers to

learn about their perspectives on this possibility (Liu et al., 2025a), determining that interviewers across expertise levels were generally receptive to real-time AI assistance of types such as suggesting follow-up questions, tracking question coverage, and offloading note-taking. However, the participants were generally negative about having AI actually conduct interviews, arguing that AI would be poor at responding to interviewee distress, which they felt required human judgment and empathy. Liu et al. also found that interviewers' attitudes toward AI assistance varied with expertise: novice interviewers saw less need for AI support, intermediate interviewers expressed the strongest demand, and expert interviewers viewed AI as an efficiency tool while expressing concern that over-reliance on AI could prevent novices from developing their own interviewing skills. In general, human-AI partnership in conducting interviews could leverage the strengths of both humans and AI, but which forms of collaboration are most beneficial remains to be fully determined.

AI as interviewee. A more controversial application is the use of AI to simulate human responses to an interview, leveraging AI's ability to act as a simulacrum of a participant. One application for simulated interview subjects is to pilot or refine an interview protocol prior to involving human subjects. In this approach, the researcher conducts one or more interviews with an AI system acting as a simulated participant, in order to identify issues or limitations in the interview protocol before real participants encounter them. An AI interviewee can also be used to test an AI interviewer. Working with simulated pilot participants creates opportunities to discover (for instance) that a question is ambiguous, that the ordering of topics feels jarring, that a key follow-up probe is missing, or that the overall direction of the interview goes in a different direction than intended. Within this use, the AI can also be prompted to simulate a participant with a specific perspective on the interview's core topic(s). While the AI interviewee may not be able to simulate every response that might come from real participants, this use is still relatively low-risk, since the simulated interviews are not treated as data about human perspectives but rather as a stress test of the interview instrument, prior to collecting the intended data.

A second potential use of simulated interview subjects is to help novice researchers develop their interviewing skills (Yamamoto et al., 2024). In this approach, an AI is prompted to simulate participants with specific characteristics — a respondent who gives only brief answers, one who becomes emotionally distressed, or one who repeatedly drifts away from the topic — and the interviewer trainee practices navigating the interaction. This could be particularly valuable when the interview topic is sensitive, such as research on trauma, illness, or stigmatized behaviors. In these cases, an inexperienced interviewer's missteps could result in the interview failing to elicit meaningful data or even cause harm to participants. Practicing with simulated participants allows interviewers in training to develop their skills in a low-stakes environment before encountering real participants, or even human simulated patients. This use also has potential

benefits for research teams working across contexts: interviewers preparing to work with an unfamiliar population or cultural context could rehearse with AI simulations to surface their own assumptions and blind spots, though of course with the caveat that the AI's portrayal of that population may be limited or inaccurate. As with protocol piloting, this use is relatively low-risk, since the simulated responses are not treated as data about human perspectives but instead support later data collection. However, even in this lower-risk application, there remains the concern that repeated interaction with AI-simulated participants may subtly shape researchers' expectations about how interviews "should" unfold, potentially biasing subsequent interactions with human participants.

A third use of simulated interview subjects is to partially or completely replace human interview subjects. This can be done in many ways, some fairly safe, and others quite concerning to the authors of this book. Ultimately, the appropriateness of this approach depends both on how simulated interview subjects are used and on the goal of the research. Take, for example, a case where a researcher is attempting to refine a user interface to make it more usable. If the researcher interviews AI subjects to find and fix major problems before testing it on humans (Lu et al., 2025), this may reduce the research burden on participants with no ultimate loss of quality. Any issues not identified by AI "participants" can be captured in subsequent human interviews. In the unfortunate event that the AI fails to represent humans with fidelity in ways that leads to a worse design, this can similarly be identified and addressed through follow-up interviews with humans. This same approach would be much more concerning, however, if human subjects were never involved. In this case, major issues impacting real humans might be missed, and changes made to satisfy the AI interview subjects' concerns might be unnecessary or even detrimental for real human users.

The concerns become more fundamental when the goal of interviews is not to iterate toward a testable product but is instead to understand human beings and their perspectives. In these cases, which are the majority of qualitative research projects, AI simulation should be treated with the utmost caution. We, the authors of this book, would recommend against this practice for most qualitative research studies. Although AI can simulate humans in many cases, it can respond very differently than a human being interviewed (Dengel et al., 2023; Potekhin, 2024; Kapania et al., 2025), and with less variation across a set of synthetic subjects than is seen for humans (Barambones et al., 2024). Conducting interviews with AI as a substitute for interviewing humans therefore risks systematic inaccuracies if the data is treated as representative of interviews conducted with humans. This concern applies across levels of specificity, from populations to individuals. At the population level, some researchers have proposed using AI to increase the representativeness of samples by simulating underrepresented perspectives, such as speakers of linguistic minority languages or

disadvantaged groups (Argyle et al., 2023). Unfortunately, as discussed in chapter 4, these voices are the least likely to be well-represented in training corpuses and are therefore the most likely to be mis-represented by AI. As such, this practice is likely to create the illusion of representativeness rather than the actuality of it, degrading research quality and actually *worsening* representation of these perspectives. Nor is this problem limited to underrepresented groups: even for groups well represented in the training data, models may collapse the diversity of individual experience into stereotypical aggregate patterns, rather than representing the range of perspectives that actual members of that group would express (cf. Barambones et al., 2024).

Furthermore, the very act of prompting an AI to simulate a demographic category risks essentializing identity — treating fluid, contextual, complex experiences as fixed traits that can be captured by a simple label (see discussion in Lin, 2025). Instead, a better solution for non-representativeness of sample may be to use AI to find and interview individuals with underrepresented perspectives, as discussed earlier in this chapter. Even when simulating individuals (Amirova et al., 2024), where in some felicitous circumstances there may be extensive data available about a specific person being simulated (e.g., historical figures), it is still likely to be more reliable to use AI to find and provide that individual's original responses (Wactlar et al., 2002) rather than to generate new simulated responses.

Overall, while AI can play a valuable role in supporting the design and conduct of interviews, using it as a substitute for human participants can raise substantial concerns about validity, representation, and ultimately the meaning of the results coming out of the research. Caution is warranted.

- Concerns and considerations for maintaining rigor, trustworthiness, and reflexivity

There are many considerations in terms of rigor, trustworthiness, reflexivity, and other core values of qualitative research, when it comes to collecting and selecting data. In this section we outline a few of the key considerations in terms of the use of AI.

Documentation of research decisions and artifacts has long been central to qualitative rigor, but AI-assisted workflows generate new categories of decisions that need to be documented and additional complexities that need attention, as well as surfacing some decisions that were previously mostly left implicit. Researchers should document everything that might be relevant to later auditability or reproducibility (where relevant):

- the key queries used to scrape and access data
- the case selection strategy used

- the artifacts and cases considered (ideally, both those selected and those ultimately not selected for consideration -- and why)
- any process data generated during iterative selection

Prior to and during data collection (by AI or humans), researchers should preserve key process information (e.g., participant selection decisions) and intermediate versions of research artifacts (e.g., interview or observation protocols). All interview transcript data (whether from human or AI participants) and all observational field notes should be saved. Saving process artifacts to support transparency and auditability is not new (e.g. Lincoln & Guba, 1985), but the volume of artifacts that need saving grows considerably with the use of AI. AI-tool specific documentation (e.g. Gallifant et al., 2025) is also essential: researchers should record information such as model prompts, any examples fed to the model, and the model provider and version.

It is worth distinguishing between two goals that documentation can serve: auditability, the ability for others to trace and evaluate the decisions that were made, and reproducibility, the ability for others to obtain (more or less) the same results if the same approach is taken. Reproducibility has largely not been a goal for qualitative analysis -- many qualitative perspectives argue for the importance of the researchers' unique perspective and influence in shaping the results of the work. But it becomes more relevant when using an algorithm or query to select cases (so that another researcher can find the same cases) or for LLMs that can be induced to produce the same results repeatedly, and where variation in outputs may be due solely to random numbers. The issue of an LLM producing variable outputs due to random numbers can be addressed technically, by using an open-weights model with a pre-set random seed. Even when full reproducibility is not achievable — as is often the case with commercial AI platforms that do not produce the same output for the same query — thorough documentation of prompts, parameters, and selection outcomes can still support meaningful auditability. And indeed, for most other aspects of data collection, auditability is the more relevant standard, since neither field notes nor interviews are typically expected to be reproducible.

One practical way to strengthen audit trails is to use chain-of-thought techniques, where the LLM explains each step in their reasoning as they reason (Wei et al., 2022b). This may be relevant for case selection or for proposed changes to observation or interview protocols, for instance. When real-time reflection is impractical—for example, interrupting a real-time interview or interfering with a model's response process—researchers can also ask the model to explain its reasoning after the fact. However, it is important to remember that these post-hoc explanations may not perfectly represent the model's actual reasoning, providing the most likely-seeming justification rather than reflecting the process that produced the output (Turpin et al., 2023).

Even with strong documentation practices, an important risk with AI-assisted research workflows is that failures can be silent. AI systems may still produce fluent and plausible outputs even when they are incorrect or systematically biased (Huang et al., 2025). To compare to previous methods, if a researcher conducts case selection using search queries, a bad query might return no results or obviously bad results. This signals a problem immediately. By contrast, an LLM that systematically overlooks a category of relevant cases, applies a subtly different

definition of a construct than the researcher intended, or generates superficially competent but analytically shallow interview probes can all produce output that looks correct at a surface level. In these cases, the researcher may get farther along in their workflow before recognizing the error, or might miss the issue entirely.

This risk is heightened when the types of cases being searched for are poorly represented in the LLM's original training data (as discussed in chapter 4). In such situations, the model may choose inappropriate cases or miss the most relevant ones, and not flag to the researcher that it had difficulty. Worse yet, even if the researcher spot checks cases they can easily spot check, accuracy may vary across within the same analysis. For example, take a researcher sampling political opinions across both well-represented (English, French, Chinese) and poorly-represented (Azerbaijani, Lule Sámi, Fulani) languages. This researcher, if fluent in English and Chinese, might spot-check cases in these languages, find them accurate, and not catch inaccuracies for less common languages. Ultimately, the differences between human reasoning and LLM reasoning create an ever-present risk that an LLM will systematically overlook certain kinds of cases, even if its performance matches humans well overall.

These characteristics make AI failures harder to detect than many other methodological errors, and they carry a practical implication: verification cannot rely on outputs looking wrong, because they often will look correct even when there are major issues. Instead, researchers need to verify their process proactively, with explicit attention to surfacing problems that the model outputs will not directly reveal. For example, if a researcher wants to use AI to help choose a sample of selected cases, they should review a random set of cases that were recommended for inclusion *and* a set recommended for exclusion to identify systematic blind spots. A human researcher might also choose to sample a set of cases by hand and compare them to the AI's decisions as a calibration check, particularly early in the process when prompts and criteria are still being refined. But as noted, in doing so, it is important to hand-sample cases where the AI could be expected to perform poorly, such as less-represented languages.

Moreover, the documentation practices discussed above support transparency, but transparency alone does not ensure the ongoing critical examination of how a researcher's own assumptions, positionality, and choices shape the research process and its outcomes. AI-assisted workflows can make reflexivity both harder and easier. They make it harder because AI introduces a layer of intermediary decisions — how a prompt is worded, what examples are provided, what similarity threshold is set — that can feel technical rather than interpretive, making it easy to overlook the assumptions they encode. For example, a researcher who asks an LLM to "find cases where participants express resistance to a policy" has made a consequential (if implicit) interpretive choice about what counts as resistance, but the seeming objectivity of the LLM may discourage scrutiny of that choice in ways that manual case selection would not. AI can itself be part of the solution to this dilemma -- a researcher can ask an LLM to explicate how it is identifying cases (such as explaining how it identifies "resistance to a policy") and then audit that explanation.

A similar dynamic can arise in protocol design: when a researcher uses an LLM to suggest interview questions or observation categories, the AI's suggestions may appear to be highly

fluent on the first pass, which may result in less revision and critique than the researcher's own rough drafts or a colleague's suggestions. Over time, this risks producing instruments that reflect the model's implicit sense of what is worth asking or observing rather than the researcher's own theoretical commitments and contextual knowledge. Avoiding this risk requires continually cultivating a critical spirit about recommendations coming from AI. Treating AI-generated protocol suggestions as a starting point for critical evaluation — asking why each suggested question or category is there and what it assumes — can help maintain alignment between the instrument and the study's conceptual foundations.

However, in other situations, AI can positively support reflexive practice in new ways. For example, a researcher can ask an LLM to critique their case selection criteria for implicit assumptions, to identify what types of perspectives their current sample might systematically exclude, or to surface cultural or disciplinary presuppositions embedded in an interview protocol — uses that are discussed in more detail earlier in this chapter. These uses do not replace the researcher's own reflexive judgment, but they can function as a structured scaffold for the kind of self-interrogation that reflexive practice demands, particularly when working with unfamiliar data or across cultural contexts.

The verification strategies discussed above apply to one-time case selection; additional consideration is needed when case selection becomes iterative, such as when researchers use LLMs to help select data with the goal of reaching theoretical saturation. In this situation, there is a risk that the interactive loop might converge too quickly if the LLM's sense of "diminishing returns" doesn't align with the researcher's. What constitutes theoretical saturation already varies across projects, researchers, and research contexts. Given the very different ways of establishing theoretical saturation for LLMs (e.g. De Paoli et al., 2025) than humans use, some degree of misalignment between stopping conditions seems almost inevitable. To address this mismatch, human researchers should independently review the data for additional possible codes at the end of the LLM’s process (a topic we return to in chapter 11), especially if the process ended with a smaller number of coding categories than might have been expected.

A similar risk arises in AI-conducted interviews, where follow-up probes may be shallow or imperfectly aligned with the human researcher’s analysis goals (Cuevas et al., 2025). Because the probes may still sound natural in the context of the conversation— another instance of silent failure — these misalignments may not be apparent when reading any individual transcript. Conducting pilot interviews with a synthetic interviewee, or reviewing data from early pilot interviews with genuine human subjects, can help identify these issues and drive quick iteration on the prompts or other resources given to an AI interviewer.

A second risk involves sensitive topics, where AI interviewers may miss the nonverbal cues such as hesitation, distress, and body language that human interviewers use to know when to back off, slow down, or offer support. Contrary to this worry, some research has found that participants are more comfortable discussing sensitive topics with AI interviewers than human interviewers, precisely because of their perceived non-judgmental nature (Geiecke & Jaravel, 2025). This finding suggests that the choice between AI and human interviewers involves multiple considerations. Design strategies such as explicit checks for participant discomfort or a

prominent option to pause the interview may help reduce the risks of AI interviewers while preserving the benefits.

Additional considerations arise for approaches that use AI not only to conduct interviews but to determine when and whom to interview. In detector-driven interviewing (e.g., Baker et al., 2024), a model identifies moments or participants of interest — such as a student who has just resolved a persistent error — and informs the researcher in real time, so that they can decide if this is a good time to conduct an interview. This introduces a layer of selection that precedes the interview itself. If the detector is inaccurate or systematically biased — for instance, more reliably identifying struggle in some student populations than others (algorithmic bias in educational detection; Baker & Hawn, 2022) — then the resulting interview data may inherit those biases without the researcher's knowledge (another silent failure). Furthermore, if only a subset of the intended events are passed on to researchers and lead to interviews, future model refinement may focus on those events, creating a subtle and potentially difficult to identify bias in the model. Documenting the detector's performance characteristics, and periodically checking whether the recommended interviews represent the range of situations the researcher intended to capture, can help mitigate this risk.

As discussed above, there are many situations where substituting an AI for a human as an interview subject raises validity risks and should be avoided. While using AI to pilot interview protocols or train novice interviewers is relatively low-risk, using AI as a partial or complete substitute for human interview subjects raises increasingly serious concerns. AI responses can differ systematically from human responses, and synthetic participants tend to exhibit less variation than real humans (Argyle et al., 2023), creating a risk of flattening the diversity of perspectives that qualitative research is designed to capture. These concerns are especially acute when researchers attempt to simulate underrepresented populations, since these are precisely the groups least likely to be well represented in training data and most likely to be inaccurately portrayed — producing an illusion of representativeness rather than the reality of it. More fundamentally, prompting an AI to simulate a demographic category risks essentializing identity, reducing fluid and contextual experiences to fixed traits captured by a label (Wang et al., 2025). Even when extensive data exists about a specific individual being simulated, it is generally more reliable to use AI to locate that person's actual responses than to generate new synthetic ones. When the goal of research is to understand human beings and their perspectives, there is no substitute for human participants — and we would argue that AI's most productive role is in helping researchers find, reach, and listen to those participants more effectively.

- Final thoughts

Data collection and selection decisions are foundational to qualitative research because they determine what phenomena can become analytically visible in the first place. The choices made at this stage — who to interview, what to observe, which cases to include or exclude, and what contextual information to record — set up what is possible in the research that follows. Analysis cannot easily recover perspectives that were never gathered, contexts that were not

documented, or types of cases that were systematically overlooked during selection. The decisions made at this phase reflect the researcher's theoretical commitments and positionality, both explicit and implicit. For these reasons, data collection and selection have always demanded careful methodological attention, and the introduction of AI into these processes expands both what is possible and what can go wrong.

AI is already supporting data collection and selection in several concrete ways. Researchers are using LLMs to discover relevant data sources, and to write and refine API queries. For case selection, LLM-as-a-judge approaches and embedding-based semantic retrieval are allowing researchers to find cases where the same idea is expressed in different terms, across languages, or indirectly. AI is also being used to design and refine interview and observation protocols — identifying gaps in coverage, flagging leading or double-barrelled questions, surfacing cultural assumptions, and suggesting follow-up probes.

Several emerging AI practices are beginning to move qualitative research beyond what was previously feasible, with early examples already in the literature:

- Iterative case selection assisted by LLMs extends longstanding traditions like theoretical sampling to corpora too large for a researcher to browse directly, with initial work exploring how LLMs can help identify when theoretical saturation has been reached (De Paoli & Mathis, 2025).
- Detector-driven interviewing makes it possible to trigger interviews at the precise moments when they are most informative (Baker et al., 2024); with AI interviewers, this can now be realized in contexts and at times where human interviewers cannot easily go (Wei et al., 2026).
- AI-conducted interviewing has scaled to tens of thousands of participants across dozens of languages (Huang et al., 2026), opening new possibilities for qualitative work that meaningfully spans cultural contexts worldwide rather than relying on convenience samples drawn from a narrow slice of humanity.

Each of these practices is still early in its development, and considerable methodological refinement is to come, but together they suggest that the scale and reach of qualitative inquiry are expanding in ways that were not practically achievable before. The ability to collect a larger sample does not always mean that we should do so. Research on how sample size is determined within interview-based qualitative inquiry emphasizes that the sample size needed depends on a study’s analytic goals and methodology, rather than seeking the largest sample size (Wutich et al., 2024) -- and that sampling often ends based on a theoretical argument that the sample that has been obtained is now sufficient, a point reached at very different sample sizes within different analyses (Elmholdt et al., 2026). Indeed, the core idea of theoretical saturation is that, at some point, additional data no longer meaningfully deepens or alters the developing analysis, and there is limited benefit in continuing to collect data or search for more analytical categories (Wutich et al., 2024).

Beyond these emerging practices, this chapter has also pointed toward several exciting possibilities that, to our knowledge, remain largely unexplored. One is the move to more

interactive case selection, where the researcher and LLM engage in a continual and integrated process of selection, review, and refinement until the researcher's chosen criteria are met — a workflow that has clear antecedents in theoretical sampling but that LLMs make tractable to conduct in deeper, more complex fashions. A related possibility is the adaptation of active learning, well established in classical machine learning, to qualitative case selection: an AI process manager that tracks coverage across the data, flags theoretically ambiguous cases, and helps the researcher direct their finite time and attention efficiently. A third possibility is the extension of real-time AI support to ethnographic fieldwork — using an LLM as a critical friend that reviews each day's field notes, surfaces emerging patterns and unresolved tensions, and helps the researcher prioritize what to pursue next. A fourth is the use of AI as a structured interlocutor for reflexive practice, where the LLM critiques the researcher's selection criteria for implicit assumptions (cf. Buckingham Shum, 2024), identifies perspectives that the current sample may exclude, or points out cultural and disciplinary assumptions built into current draft research plans and protocols.

None of these possibilities replaces the researcher's interpretive work. Instead, each offers a new kind of scaffold for it. Realizing these visions will require considerable methodological work — pilot studies and iterative design, carefully-designed comparisons between variants and to established approaches, and the development of practical guidelines. As this work develops, these new methods -- and the rest of the emerging methods in this chapter -- create the potential for research that is not just more efficient, but better fulfills the broader values of qualitative research.

A theme running through this chapter is that AI's greatest contributions to data collection and selection tend to come when it augments human collection and selection processes rather than replacing the humans involved in them. AI can help researchers discover data sources, write API queries, refine interview protocols, iterate on case selection criteria, and even conduct interviews at scales that would otherwise be infeasible — but in each of these applications, the researcher remains responsible for the core methodological decisions: what to study, whom to include, what questions to ask, and how to interpret the answers. At the same time, AI does not merely speed up existing workflows; it opens genuinely new possibilities. Interactive case selection driven by structured dialogue between the researcher and LLM, detector-driven interviewing that responds to events as they unfold, and interview studies spanning dozens of languages and thousands of participants all represent qualitative research practices that were not practically achievable before. These new possibilities come with corresponding risks. The expanded scale and automation that AI enables can multiply the consequences of poor methodological choices just as readily as it multiplies the benefits of good ones — a flawed case selection criterion applied across a large corpus can systematically exclude important perspectives, and an AI interviewer with a poorly designed protocol can conduct thousands of shallow interviews before the problem is noticed. As such, pilot testing, thoughtful documentation practices, and ongoing human oversight are essential components of AI-assisted data collection and selection.

In the next chapters, we turn to what happens after data has been collected and selected: how it is prepared and segmented for analysis, and how AI can support initial sense-making. As we will see, many of the same themes and considerations will apply.

Find the most up-to-date version of this chapter's citations at
https://learninganalytics.upenn.edu/book/ai-qualitative-research.html