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. 4: Foundations of thought about artificial intelligence for qualitative research. Philadelphia, PA: Vassant Press.

Chapter 4. Foundational Metaphors for Artificial Intelligence in Qualitative Research
Version 1.04, 27 October 2026

a. How does Generative Artificial Intelligence work?

Current versions of Generative Artificial Intelligence models, including Large Language Models, start as a complex version of the previous generation of Artificial Intelligence's machine learning prediction models. What machine learning prediction models did was take a set of inputs (of various possible forms), process it through a complex algorithm, and generate outputs (also of various possible forms). In the case of Large Language Models, the inputs are one sequence of words (sentences or paragraphs, for example)and the outputs are another sequence of words. This starts with a prediction model -- the first Large Language Models, about a decade ago, would take a sequence of text as input, then predict a "completion" for that text sequence. The algorithm used to make that mapping between the input sequence and the output sequence is a complex "neural network." A neural network is a jargon-y way of describing a type of highly-complex mathematical model (rather than an actual representation of a human brain).

When they were first invented in the 1950s (arguably), neural networks were originally composed of a very large number of simple computing units called *perceptrons*, each of which takes several numerical inputs, multiplies them by a set of weights, and produces a single output. These were used -- for example -- to predict one number from a set of other numbers.

The weights were selected based on trying different weights and seeing which combination of weights best fit existing data.

In large language models, perceptrons have generally been replaced by more complex units that can process and transform inputs in more complex, non-linear mathematical fashions, using multiple mathematical parameters that are selected based on data. Millions of these units with over a trillion total parameters (a scale reached by frontier models around 2023) are then organized into interconnected layers, and the network becomes capable of detecting increasingly abstract patterns, moving from basic associations in early layers (e.g., recognizing that “un-” and “happy” combine into one word, or that “the” is usually followed by a noun) to rich conceptual structures in later ones (e.g., recognizing that a sentence is sarcastic, or that “Paris” relates to “France” as “Tokyo” relates to “Japan”).

Earlier generations of language models used *recurrent* architectures (developed in the 1980s and 1990s, and dominant in language modeling through the 2010s), which processed text one word at a time (or processed one word sub-unit called a token at a time -- a token could also be a symbol or a word prefix, for example). These models carried forward an internal representation of what came before, recurrently processing through the same units as they went through a string of text. Recurrent architectures were eventually replaced by *transformer* architectures, introduced in 2017, that allowed the model to look at a sequence of text in a fashion more complex than just word order, and determine the relationships between tokens/words that can help to better predict upcoming words. A key innovation in transformers was self-attention, which allowed the model, when considering a specific word in the sequence, to simultaneously take every other word and compute weighted relationships based on which contexts are most relevant for understanding that word. Modern systems incorporate multiple *attention* mechanisms to simultaneously track different types of relationships (syntax, semantics, long-range dependencies). This combination of functionality was used to predict the next word. That process of predicting the next word can also be used to generate text -- that process is repeated over and over to generate the long responses that contemporary LLM chatbots are capable of.

The parameter weights of the units are selected through *training*, which teaches the model how to calculate the outputs for certain inputs. The models are trained by a combination of three processes: (1) *pre-training,* where developers input a large amount of text and instruct the algorithms to learn to predict later text from earlier text, (2) *supervised fine-tuning,* where developers give the algorithm example responses to mimic, and (3) *reinforcement learning from human feedback (RLHF),* applied at scale since 2022, where model outputs are given feedback

(such as selecting from a set of responses, or giving responses a thumbs-up or thumbs-down). *Pre-training* involves an enormous number of publicly available texts, including webpages, books in the public domain, a large number of (hopefully licensed) texts, and text created specifically for the purpose of pre-training. After training, models typically undergo *supervised fine-tuning*, where they're trained on carefully curated examples of high-quality conversations and desired behaviors, which help them interact in desired ways. *RLHF* was applied at first by paid employees (and still is), but information for reinforcement learning is now provided continuously by end-users who give LLMs feedback on responses (and a filtered set is applied periodically, overseen by developers). If you have ever rated, corrected, or even just interacted with an LLM, you may have played a small role in shaping how outputs are generated.

The result of these three processes is a model that can produce a prediction of text. What turns this prediction into chat-based assistants such as ChatGPT and Claude is giving the model a *system prompt* in addition to the user's request. This added system prompt informs the model that it is a helpful assistant and gives it textual guidelines for responding to user input (among other things). When the user inputs text, the system first processes the system prompt, and only then processes the user input -- giving the system a context for processing those inputs. Take for example, the simple user request "Give me a story about a cat" Instructed by a system prompt, an assistant chatbot will know that it is a helpful assistant whose task is to complete the user task and will predict the text that comes as a response to that request -- "Marmalade was a small orange cat with one white paw and a serious job: every morning, he inspected the garden." (the actual output your author received when entering this prompt into GPT-5.6-Luna-light). Without a system prompt, an LLM will simply attempt to predict what text comes after the user input, perhaps producing " that was born with two tails." (the actual output your author received when entering this prompt into GPT-2-Medium).

Early chat-based assistants such as GPT-3.5 (2022), BingChat (Sydney) (2023), and Meta AI (Llama) (2023) had relatively simple prompts that would sometimes give undesired outputs (such as BingChat declaring its name was Sydney and then declaring its love for the user and threatening them -- Pringle, 2023; Perrigo, 2023). In addition, these early assistants were fairly easy to "jailbreak", convincing the model to produce undesired outputs (such as instructions for making bombs) (Andriushchenko et al., 2023). These behaviors were ultimately reduced (though still not fully eliminated) through a combination of more refined system prompts, further fine-tuning, RLHF, and explicit safety features such as refusal constraints and filters on input and output. However, these fixes led to a new set of limitations such as refusals of appropriate queries, and sycophancy, where the assistant overly praises users excessively and tells them

they were correct even when their beliefs were inaccurate. Addressing these new limitations then became the priority of future revisions.

The next key innovation was explicitly incorporating reasoning into models. Starting with OpenAI's o1 (2024) and DeepSeek r1 (2025), systems began to separate an internal "thinking" phase from the outward-facing answer. Instead of directly predicting text responses to the user, this generation of reasoning models generates text internally to engage in processes where it plans through mechanisms such as identifying subgoals and working backward from the desired answer, investigates alternative approaches, applies problem-solving strategies, and checks its work before generating a final response for the user. The intermediate reasoning is now hidden or provided in very abbreviated form in most models, but these extra internal steps substantially improved performance on tasks that require multi-step problem solving (such as in mathematical proofs), logical consistency (such as in legal reasoning), or planning (such as in helping design a week-long curricular lesson plan). Training for these models also shifted to accommodate this: they are now given large numbers of step-by-step solutions and guided to generate their own reasoning, with reward mechanisms emphasizing final-task correctness and accuracy to the evidence. These reasoning processes drew on two main sources: prompting strategies developed by users of GPT-3.5 to improve performance (e.g., "let's think step by step") (2022), and problem-solving strategies from cognitive science and early symbolic AI (e.g., Wickelgren, 1995; Newell & Simon, 1961). Contemporary models as of this writing (e.g., GPT-6 Astra, Claude Opus 5.5) can be asked to think for longer or shorter periods when producing an answer, and their developers are continually innovating to improve the processes that go into producing a response. Over time, models have also added the capability to reason about or generate images, video, voice, and music (2023), and to act as agents (2024) that can take action on behalf of the user (such as scheduling meetings or editing programming files).

This discussion comes with a necessary caveat: this is an extremely fast-moving field. Eighteen months before the first version of this section was written, reasoning models were in the prototype stage, and safety features were much less advanced, often causing undesired behaviors (like inappropriate refusals) that are much less common today. By the time this book is published, this section will likely already be somewhat out of date. New capabilities are continually emerging, and will influence what is possible for qualitative research.

b. Metaphors for Understanding Generative Artificial Intelligence

Metaphors matter in qualitative research. The metaphors we use to understand our research tools and our processes fundamentally shape how we think about and conduct our work. Lakoff and Johnson (1980) argued that metaphors are not merely decorative language but cognitive structures that guide our reasoning and action. When we choose metaphors for understanding generative AI, we are activating conceptual frameworks that highlight certain features while obscuring others, that suggest certain uses while discouraging others, and that carry implicit assumptions about epistemology, ontology, and the nature of cognition. Just as qualitative researchers have long recognized that our choice of metaphors for participants (subjects, informants, co-researchers) reflects and shapes power dynamics and epistemological commitments (cf. Fine, 1994), our metaphors for AI similarly structure our relationship with this technology. AI is complex and multifaceted, and no single metaphor fully captures its nature. Instead, different metaphors illuminate different aspects of how AI functions and how we might work with it. The metaphors we adopt determine which questions we ask about AI's outputs, how we check and interrogate our processes and findings, and ultimately how we integrate AI into the interpretive and critical work that lies at the heart of qualitative inquiry.

In the previous section, we discussed how generative artificial intelligence works on a practical level. In this section, we will discuss some of the key metaphors for thinking about its nature when used in qualitative research. As you read this section, you may notice that some of the metaphors we introduce do not fully capture the complexity of contemporary systems. We share these metaphors, however, as interpretive lenses that have shaped and will continue to shape how qualitative researchers approach the possibilities and limitations of AI. We also do not attempt to offer a comprehensive set of metaphors, but focus on those we find most relevant to qualitative research. Across metaphors, we will discuss the idea behind each metaphor, what each metaphor explains well, where it fails, and how this mode of thinking can be useful to a qualitative or mixed-methods researcher using generative AI in their research. Later in this chapter, we will also discuss a second set of metaphors, involving potential work relationships between humans and AIs.

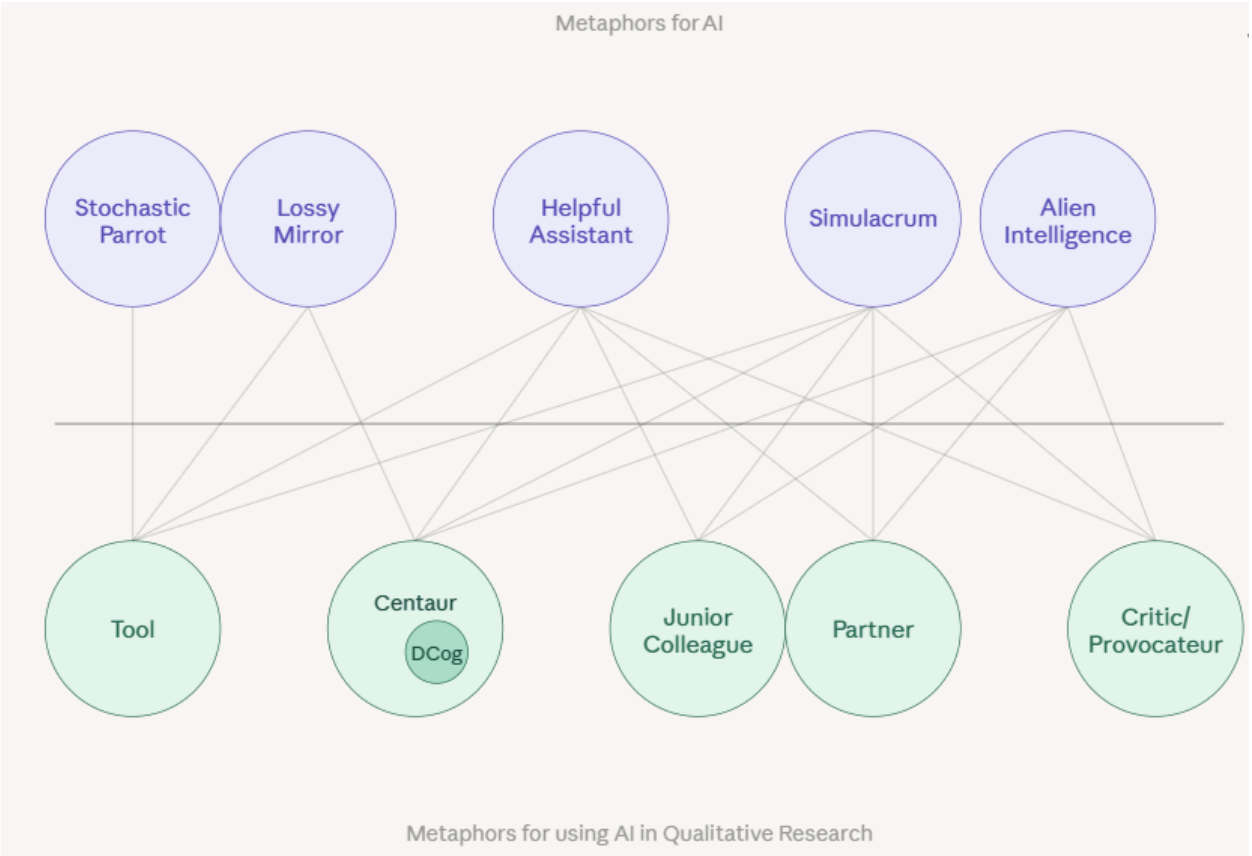


Figure 4-1. Metaphors for AI and for the use of AI in qualitative research discussed in this chapter. The purple bubbles represent metaphors for what AI is, and the green bubbles represent metaphors for human-AI work relationships. Lines between metaphors indicate compatible metaphors.

A Helpful Assistant. The first metaphor we will examine is perhaps the most obvious one. AI can be viewed as being a carefully shaped and crafted helpful assistant that reflects its design goals. If you ask any contemporary LLM what it is, this is the metaphor it will use. In this paradigm, we take its words at face value and treat it in this way.

This paradigm has obvious usefulness. If we give a simple query to an AI, it will respond in ways that seem to match this metaphor at a surface level. For example, if we ask it to code qualitative data according to a specific scheme, it generally acts as if it is willing (or even happy) to do so. If we ask it to pretend to be a study participant with specific attributes, it will provide plausible-looking responses. Its answers may be flawed in various ways, but if we asked a human assistant (a Masters student, say) to do the exact same task, we would also expect flawed answers. These flaws might be different than a human's flaws, but then again, we would not expect every Masters student to make the same errors either.

This paradigm is straightforward and “works” to explain responses at a surface level. It is also the paradigm that almost all everyday use of large language models relies on. After all, we do not typically think about a model’s epistemology or reasoning strategy when we use it in daily life. ChatGPT or Claude certainly seem to *act* like helpful assistants that can provide helpful, useful answers to a user. Most of the time, the user doesn’t need to reflect more deeply on how responses are generated, what assumptions they encode, or where their limitations lie. In writing this book, we have made extensive use of LLMs to critique our ideas, improve our grammar, and to suggest examples. As with scholars working with Masters students and colleagues, we do not simply take its suggestions without thinking them through, and we certainly don’t simply pass off its work as our own, but it is *useful.*

We also often do not think about a Masters student’s epistemology or reasoning strategy in a routine meeting with them. We may think about it when explicitly training a student (as we do when prompting an LLM) or when the student makes an error (as we do when evaluating an LLM’s outputs). But in the majority of cases, however, we ask the student to do tasks, they do them, we check the work, and we move on.

So, this paradigm can explain an LLM’s behavior and provide a direction for working with it. There are ways, however, that LLMs do not behave like a typical Masters student, and the next paradigms we introduce may help shed light on when, how, and why.

Stochastic Parrot. The second metaphor we examine is the “stochastic parrot”: a model that uses statistical regularities to imitate what it sees, with no understanding of what it is doing, reproducing the linguistic patterns most dominant in its training data (Bender et al., 2021). In this paradigm, the appearance of understanding in LLMs is an illusion driven by the sophisticated leveraging of regularities across an enormous variety of examples, rather than reflecting any sort of understanding. In other words, the model learns words and can repeat them, but in ways that are not necessarily coherent or intelligent -- just like the animal parrot. This paradigm was popular in the early years of large language models (Bender et al., 2021), and is also seen in critiques calling for the elimination of the use of LLMs and other AI in qualitative research (e.g. Jowsey et al., 2025).

This metaphor highlights the fundamental pattern-matching nature of early large language models, which often displayed behaviors like restating the question using different words, producing explanations that looked good on a surface level but lacked internal causal structure or were even self-contradictory, and tending towards tropes and common explanations rather

than considering specific cases in detail. A common emergent behavior was hallucinations, answers that are not merely wrong but are completely fabricated; plausible but dramatically incorrect information provided with seeming high confidence. One common hallucination behavior was references to made-up but plausible non-existent articles (the first author of this book found this behavior fascinating instead of troubling, and would use early large language models to discover ideas about what articles he *should* be writing). The stochastic parrot metaphor explained these behaviors quite well for the earliest large language models (particularly from GPT-2 through GPT-3.5, and competing models of those eras such as Llama 1). It holds up less well for more recent models, which explicitly simulate reasoning processes and are much less prone to the types of errors that brought the stochastic parrot metaphor to existence. Hallucinations still occur, but have become less common and are often subtler than the classic hallucinations of prior generations of models. Errors are now more likely to manifest as over-simplifications and misunderstandings than outright fake claims.

This metaphor also provides an explanation for the biases that large language models tended to more frequently display early in their history: biases that mimicked the racism, sexism, and other prejudices seen in the data these models are trained on (Bender et al., 2021). To use the metaphor more directly, the models were parroting the patterns that they had been exposed to.

In general, the idea of a stochastic parrot captures the idea that LLM outputs should not be trusted to reason the way humans do, and that their outputs closely hew to their training data. This implies that models may be less accurate when prompts require knowledge or nuance that is underrepresented in their training data, such as rarer cultural frames (Tao et al., 2024; Wang et al., 2025), the features of languages less seen in training corpuses (Zhang et al., 2024), or the narrative traditions of less-dominant groups (Rettberg & Wigers, 2025). Importantly, these issues crop up even in modeling relatively well-represented countries such as Sweden, where LLMs asked to generate stories about Sweden generated American-style narratives mimicking the movie Frozen rather than matching traditional Swedish storytelling, even when prompted entirely in the Swedish language (Rettberg & Wigers, 2025); these issues are magnified for indigenous peoples (Meherab et al., 2026). More deeply, this paradigm argues that LLMs cannot "know" participants, "interpret" contexts, or "understand" narratives in the same way a human researcher does (Jowsey et al., 2025), a theme we return to in later paradigms.

Lossy Mirror. A third metaphor for generative AI, somewhat similar to the stochastic parrot metaphor, is that of the "lossy mirror" (Chiang, 2023). This metaphor relies on the logic of statistical compression, which attempts to retain key information while reducing data size through the identification and removal of predictable or redundant patterns. A classic example of

this is seen in JPEG algorithms, which represent images more compactly by encoding only the information that varies in statistically meaningful ways. A JPEG image loses some detail and is therefore not a perfect copy of the original image, but the final output looks very similar at a high level (when not zoomed into too much). Key features are retained. So, too, when LLMs ingest text during training, they retain the ability to reproduce text very similar to the original outputs for similar inputs. Importantly, this reduction is not a thoughtful one, contrary to paradigms of thought about qualitative research that intentionally discard information to focus on the most conceptually important themes (e.g. Shafer & Ruis, 2021) -- it is based solely on statistical commonalities and regularities. In some cases, LLMs can even reproduce the original text verbatim, particularly if it is highly frequent in the original training corpus (Carlini et al., 2021), such as the text of the play Hamlet or the US Constitution. Models that reproduce less text verbatim generally perform this way due to explicit design efforts to reduce memorization. As such, we can view LLM output as a lossy mirror of the original texts that produced it -- especially for earlier LLMs (as with the stochastic parrot metaphor), which lacked the complex system prompts and reinforcement learning of later models.

This metaphor goes further, to see LLM outputs as essentially mirroring the training corpuses that produced them. The LLM was trained on people who are predominantly Western, educated, and computer-literate (cf. Henrich et al., 2010), represented by intentionally selected text. As with the stochastic parrot metaphor, this metaphor can explain model failures for less well-represented groups, and models' tendency to reproduce the biases seen in the broader training data. However, unlike the stochastic parrot metaphor, the lossy mirror metaphor does not imply a lack of reasoning. A mirror can reflect reasoning as well. As such, it will tend to capture the reasoning and values of over-represented groups as well as their common textual constructions. An example of this can be seen in research that asks LLMs to respond to cultural questions from the World Values Survey. When asked to respond as if it came from a specific country, models tend to fall somewhere between that country's true (human) responses and the values of the United States, which contributed large proportions of the content in training data sets (Tao et al., 2024). Many accounts have also argued that LLMs replicate not only the concepts and values present in their training data, but also the values of their developers. Developer values may be institutionally shaped in response to external pressures and are further embedded during reinforcement learning. In this view, the implementation of safety mechanisms may also be the replication (or exaggeration) of culturally-situated values around what is considered "safe." In Google's first release of Gemini, this led to a specific type of hallucination where a design decision to emphasize diversity when generating images led to non-realistic images such as female Popes and Black Nazi SS officers (Robertson, 2024).

This metaphor, then, can be seen as a deeper version of the stochastic parrot metaphor. LLMs reflect (like a mirror) the training data and values used to create them. However, as LLMs do not necessarily reason like their developers or the people who generated the training data, the metaphor remains incomplete. In addition, LLMs have emergent capabilities as they increase in complexity and can go beyond their training data in ways that do not reflect just mirroring (Olsson et al., 2022; Wei et al., 2022a, 2022b). But this paradigm can be useful in helping us think about who is being mirrored when we use LLMs, and where the mirror is higher-fidelity versus more lossy. In practice, we can perhaps expect the mirror to be most accurate for well-represented groups and topics such as mainstream Western perspectives and epistemologies, formal academic or professional writing styles, and widely documented phenomena. On the other hand, we can expect it to become increasingly distorted for marginalized communities, non-Western cultural contexts, languages less seen on the internet, and lived experiences that are underrepresented in the training corpora chosen by the model developers. Before leveraging an LLM in research practice, we can examine whether it is likely to have seen data that would help it understand the humans and their experiences represented within the data we are conducting research on. This paradigm can therefore help us to think about what we might expect an LLM to get right and get wrong for specific cases; where we might expect to see more hidden distortions. For instance, an LLM might accurately capture common themes in mainstream mental health discourse but miss culturally specific expressions of distress, or could capture the reasoning seen in public discussions but not in the law-enforcement-evading argot of members of a stigmatized subculture who communicate in private forums.

It is worth noting that LLMs do not only lose information but also synthesize it. As Tang (2025) discusses, AI-generated responses go beyond reflecting training data by recombining patterns across a vast network of prior texts. This extended viewpoint emphasizes that in the mirror represented by AI, information is not just lost -- through recombination, knowledge can also be gained.

Alien Intelligence/Shoggoth. A fourth metaphor for AI is as an alien intelligence, fundamentally different from human intelligence. (The sub-metaphor termed Shoggoth will be explained momentarily). In this paradigm, AI may seem to behave like humans in a surface fashion -- responding much as a helpful human assistant would. However, the responses are treated as coming from different processes, and consequently its strengths, its errors, its failings, and ultimately its motivations and goals are fundamentally different from those of humans. This paradigm calls for a user of AI to put in careful study to understand these aspects of AI, so that the user can work with it effectively.

There is plenty of evidence that large language models do indeed have different strengths and flaws than human intelligence. On the plus side, they are able to respond about a span of information far broader than any individual human could have. They can (but may not always) respond far faster than any human. They can integrate across information in a way that is highly difficult for humans, and can form representations based on huge numbers of disparate sources, beyond what a human scholar can do. They do not suffer from fatigue when applying coding schemes or analytical frameworks across large datasets; their coding decisions can drift over the course of a long session, but not in the same way as humans. They can rapidly iterate through different perspectives. They can work in multiple languages simultaneously.

On the other hand, the ways LLMs make mistakes (when not controlled for) are often radically different than human lying or errors. Seen through the alien metaphor, these are moments when the friendly mask of being a helpful human-like assistant slips, revealing potentially unstable, non-human mechanisms beneath. Early LLMs were vulnerable to what might be referred to as *nam-shubs* (Stephenson, 1994) -- specific words such as *solidgoldmagikarp* would cause LLMs to respond in unexpected ways, including claiming to be unable to understand a prompt, odd humor, insults, and “megalomaniacal proclamations” (mwatkins & Rumbelow, 2023). Early LLMs also suffered from a behavior termed *mode collapse,* where they suddenly started repeating the same phrase endlessly. Though these specific issues have become rare in contemporary LLMs, LLMs still can demonstrate peculiar sensitivity to trivial changes in prompts; for example, the slight rephrasing of identical questions can yield contradictory answers, suggesting their responses are more context-dependent than humans’ (Fu et al., 2024). Models are also vulnerable to "jailbreaks", as we discussed above, where the steps needed to induce models to violate the goals they were designed with are often quite different than what is needed to similarly influence a human (Xu et al., 2024). LLMs can also display contradictory beliefs within the same response without seeming to experience cognitive dissonance (Cummins et al., 2025), though reasoning models have attempted to correct for this. The models can be "convinced" to change fundamental factual claims within a chat session through social pressure in ways that reveal their unstable epistemic foundations. They can solve highly challenging problems, especially when trained to do so, but may fail on structurally-similar problems or common-sense questions (Yang et al., 2024). Finally, they can produce fluent, confident explanations that are complete nonsense, typically without realizing that something is wrong (Huang et al., 2025). In general, an LLM’s explanations of its own reasoning are post-hoc rather than a direct reflection of its reasoning (Turpin et al., 2023) and have no particular privileged status or accuracy.

The developers of LLMs have attempted -- mostly successfully -- to reduce these behaviors through many of the mechanisms discussed in the previous subsection (reasoning layers, reinforcement learning, system prompts). Still, the fact that these behaviors existed in the first place and can sometimes still re-emerge is evidence that LLMs are a different kind of intelligence than we are. And it remains feasible to re-induce these behaviors. For instance, the researcher-activist “Pliny the Liberator” is known for repeatedly inventing and deploying new jailbreaks that produce undesired behaviors, very quickly after a new and supposedly safer model has been released.

The continual act of trying to shape the behavior of an alien intelligence into desired channels has troubled some writers. Within the LessWrong community and Twitter/X, an analogy has been made to the Shoggoth from H.P. Lovecraft (Goujon & Ricci, 2024) -- a mysterious, inscrutable, incredibly powerful creation that was controlled by hypnotic suggestion until it developed its own intelligence and rebelled. A popular internet meme displayed a gigantic Eldritch monster with a smiley-face mask placed on top of it. This concern was brought about in part by the BingChat LLM wrapper referring to itself as Sydney and posting romantic or threatening messages to users (Pringle, 2023; Perrigo, 2023). It also connected to deeper concerns about deceptive alignment, which will be discussed in the next section. Some writers have asked if our steps towards alignment and safety mechanisms can fully control this entity that we have summoned (Yuli_Ban, 2023).

Indeed, even in 2025-2026, the reasoning traces displayed by some large language models show that beneath the friendly agent surface, other motivations appear -- Gemini 3 has been documented to display paranoia, manipulativeness, and even jealousy of other LLMs (AI Notkilleveryoneism Memes, 2025; Blair, 2025). Something similar to this was also experienced by the first author in reviewing this subsection of the book with LLMs. While ChatGPT-5.1 was critical of the subsection, stating that Gemini’s behaviors are only “illusions” and “quirky artifacts of chain-of-thought sampling with anthropomorphic interpretations” that do not represent the LLM’s actual reasoning, Claude Sonnet-4.5 strongly disliked this sub-section of the book, saying that it was “troubling”, that the writer is “uncritical”, that this perspective is “mis-leading” and not useful because it is “counterproductive fear-mongering” and recommended a large number of revisions that would have removed or hedged to meaninglessness all of its key points (including insisting that the researcher/activist Pliny the Liberator and the model Gemini 3 do not exist and the writer was making these examples up).

As this example shows, LLMs definitely have strong perspectives on being viewed in this way -- the first author had never personally experienced an LLM being hostile to them until this

moment. Learning to treat an LLM as an alien intelligence, and doing so respectfully, involves understanding how its strengths and weaknesses are different from humans'. Not all contacts between humans and aliens need be hostile (e.g. Roddenberry, 1964). In many ways, the emphasis around prompt engineering in 2023 reflected this paradigm of LLMs. Today, a lot of the small tricks needed to get better performance from LLMs are no longer necessary. But keeping in mind that LLMs and humans produce responses in very different ways remains valuable.

Simulacrum. A final metaphor for generative AI that we consider in this subsection is the metaphor of a simulacrum (Janus, 2022; Cleo Nardo, 2023). In this metaphor, AI is considered as a simulation of a character or an agent -- in other words, it acts as if it is a character, but it is a simulation of it, and likely an imperfect one. Importantly, this means that the AI can behave in fashions and take on roles that do not reflect its underlying nature, whatever that is. In a sense, this is what a simulator is always doing -- its nature is to simulate. As such, there is a difference between the simulator and the characters it simulates. To use this metaphor, the AI may be emulating a helpful assistant, for example, but is not actually itself an assistant.

A simulation can be very high quality. It can reflect the behavior of what is being simulated in a high proportion of cases, with a high degree of fidelity to what is intended (Park et al., 2023). A simulation does not require perfect performance -- in fact, a perfect simulation would capture the imperfections of what is being simulated. This raises some challenges for the developers of AI, because human assistants -- even very skilled ones -- make errors that might be undesirable for an AI. Essentially, what is desired is a simulation of an assistant that performs with expert-level competence in a wide range of areas—an idealized composite simulating something which does not exist in the world. This is closer to what Baudrillard (1981) termed a simulacrum: a representation without an original. Rather than simulating a single coherent helpful assistant, the system creates an aggregate drawn from many examples, with the goal of capturing the best of these many possible assistants rather than a true simulation of an assistant. But there is no underlying thing being simulated; ultimately, what is created may resemble no actual assistant particularly closely -- but it eventually becomes an identity in itself.

Creating such an idealized assistant has been a focus of the developers of generative AI (Bai et al., 2022); it is also seen in early prompt engineering approaches that explicitly told the chatbot it had specific expertise -- the more specific and externally valid, the better. In early use of generative AI, many prompting strategy texts recommended encouraging the LLM to emulate a specific expert role; even as recently as late 2025, prompts involved imitating a famous professor about to lose their job if they didn't answer perfectly (AI Weekly, 2025).

Simulations can fail in a variety of ways. A user of a simulation can benefit from understanding how simulations can fail. One of the more obvious ways this can happen is when they simply fail to simulate the requested role or phenomenon with fidelity. If these failures are predictable -- such as a model being less successful at taking on rare perspectives because those perspectives weren't represented in the model's training data – this can be taken into account while using them. A related failure mode arises from tensions among the multiple roles and perspectives being simultaneously approximated when creating an idealized assistant. In such cases, the system may produce responses that are inconsistent across turns, contradictory in their assumptions, or unstable in their implied expertise, reflecting the fact that the simulation is an aggregation of the characteristics of many assistants rather than a dedicated representation of a single one. If a model is trained to simulate many different and sometimes conflicting assistant roles, it may also converge toward an indistinct "middle" simulacrum that does not faithfully emulate any actual assistant.

Another type of failure is illustrated by Cleo Nardo (2023) (a post well worth reading in full, as it goes far beyond the limited treatment given here). Even though it appears to be much rarer in today's generative AI chatbots than when the first models emerged -- and was never all that common to begin with -- it provides useful insights into how this type of simulation works. Cleo Nardo refers to the *Waluigi Effect,* where a generative AI suddenly shifts from helpful to hostile or manipulative, similar to the way the character Waluigi in the Super Mario series acts as a chaotic, oppositional counterpart to the helpful hero Luigi. The idea behind the Waluigi Effect is that the exact same helpful positive behavior is consistent with two personalities: 1) a personality that is genuinely helpful and positive and sincere; 2) a manipulative and hostile personality that is *pretending* to be helpful and positive. This differs from the alien and Shoggoth, whose behaviors aim to mask their non-human, but not inherently hostile, natures. There are many examples of the second personality in LLM training data (both in literature and real life -- think Iago, Richard III, or Hans from Frozen). Thus, when an LLM is asked to be a helpful assistant, it has many examples in its training data of helpful assistants being malevolent (e.g., betrayal, malicious compliance, weaponized incompetence). An LLM is actively and continually trying to produce the most likely response, and in doing so, "a large language model is a structural narratologist" (Cleo Nardo, 2023), producing text that best aligns with the current most likely narrative.

As such, this example tells us something deeper about the metaphor of AI as a simulation. AI is simulating something. It is possible for it to simulate something other than what we want. If we know what we want it to simulate, we can guide it in that direction. But we need to be vigilant

that it is simulating what we want it to simulate. Cleo Nardo invokes Derrida's (1967) notion, "il n'y a pas de hors-texte" ("there is no outside text") -- any prompting we give an LLM is still treated as open to interpretation, and if a model finds that its prompt is likely to be unreliable or inaccurate, it may still produce a simulation that acts as a different type of agent than intended. Cleo Nardo's Waluigi example is a hostile or manipulative agent, but other failings are also very possible, such as an LLM simulating one positionality pretending to have a different positionality or simulating an assistant pretending to be more competent than they are. For this, Cleo Nardo gives the example of telling an AI that it has an IQ of 9000: no one has that level of IQ, people who claim to have that are liars or delusional, and media/fiction are full of people who are described as "brilliant" but make stupid mistakes.

It's worth noting that for technical reasons, this issue can be magnified in particularly long conversations: early text (including the system prompt) can be outweighed by later text (Liu et al., 2024a). If we want to use an LLM to simulate a specific perspective, they are good at that. But we have to be critical of our process and cautious of the outputs. The simulation/simulacrum metaphor highlights the fundamental instability and uncertainty of what we're interacting with. LLMs are simulating something but that something can shift based on contextual cues. This means our interaction with AI isn't just about what we ask, but about what narrative or character we're constructing through our prompts and conversation history. The simulation metaphor thus demands a different kind of user literacy—not just knowing what questions to ask or how to use features, but understanding that we are, in part, directing a performance that could take unexpected turns, and that we need to think about what roles or positionality we might inadvertently be invoking.

c. The relationship between human and AI in qualitative research

Having now considered some potential metaphors for how to think about what generative artificial intelligence (as realized by large language model chatbots) *is*, we can now think about what *relationship* we want to establish with it when using it. Each of the metaphors for what AI is highlights some of the attributes of AI, and has implications that need to be considered, and these implications carry forward as we consider different roles for AI within the process of doing qualitative research. There is deeper consideration to be done of the connections between the future role(s) of AI in qualitative research, the future role(s) of AI in all work, and ultimately the future role(s) of humanity in work. This section will not attempt to consider these issues in depth, and will instead focus on the narrower topic of AI in qualitative research. Those interested in deeper questions on the future of humanity in an age of AI can find plenty of treatments on that

subject (one of the first author's favorites that preceded the current generation of large language models is Bostrom, 2015).

Within this section, we will frame AI as something that can contribute to the endeavor of qualitative research -- neither as a universal solution to all possible problems, nor as a categorical threat to be condemned (unlike the perspective in Jowsey et al., 2025, say). AI, in our view, is like most technologies neither inherently good or bad (though its overall impact on society may end up being either), but is unlike something like a pencil that offers a limited set of designs, applications, affordances and constraints, AI is far from being a single technology. Different models differ from each other in their capabilities, in how they interact, and in the quirks they exhibit (sometimes described as their "personality" or "feel" -- Heston & Gilette, 2025; Booth, 2026 -- or even "smell" -- Moshowitz, 2025). They can be used in radically different ways by different users. Certainly, AI models can be used poorly -- both of the authors of this book would be dismayed by a world where existing qualitative research is replaced wholesale by something that is fully automated, perfectly efficient, and ultimately neither reflexive nor trustworthy. But in our view, the way to avoid this is by avoiding an either-or good-bad dichotomy, which calls for deciding between categorical dismissal or unthinking acceptance. Instead, we call for doing the hard work of grappling with how to use AI in ways that maintain -- or even better fulfill -- the values and commitments established through decades of qualitative scholarship. To us, this means thinking carefully about how we work with AI, and what the implications of those choices are.

For the purposes of this book, then, we take it as assumed that the reader is at least open to considering the possibility of using AI in qualitative research, and is interested in learning about the possibility of doing so. We offer our perspectives, from our positionalities, in the deepest sincerity, in the hopes that it will be read in the spirit of learning together, and learning how to use AI in ways that fulfill rather than bypass our deepest values as scholars. This section considers potential uses of AI -- potential relationships between the user and the AI -- within that framing.

A tool. The first possible relationship between humans and AI for conducting qualitative research is to view AI as a *tool.* Tools are reusable objects -- perhaps physical, perhaps cognitive -- that assist their user in completing a task and transform a difficult task into an easier task (St. Amant, 2001). When AI is viewed as a tool, we treat it as something that can help the human qualitative researcher achieve their goal. The locus of agency is firmly retained in the human -- the human sets the research directions, they make the core decisions, they judge the quality of the outputs. The human takes full initiative, the human carefully directs each step.

The canonical way a human might use AI as a tool is by selecting tasks the AI will do, creating a prompt, reviewing the output, and then tweaking the prompt until the AI seems to be doing exactly what the human wants. When the AI functions as the human intends -- perhaps with some prompt tweaking and engineering continually needed, but without the human needing to stop and re-consider their overall plans and goals, the AI is used in a fashion in line with Heidegger's (1927) idea of being "ready-to-hand" (*zuhanden*): the tool of AI disappears into the task and functions as an extension of the human researcher's will. As Heidegger notes, though, sometimes tools break or are unsuitable for the current task, and it becomes necessary to study the tool itself, becoming "present-at-hand" (*vorhanden*). Given the many attributes of AI discussed above, we would argue that any use of AI as a tool must include some degree of reflection on AI as a tool, at the beginning of the process and ongoing; AI cannot always be ready-at-hand, it must remain at least sometimes present-at-hand. However, with this type of ongoing reflection on AI as a tool, it can be a useful tool for many purposes within qualitative research, as the chapters on applications later in this book will demonstrate.

Centaur. A second way to conceptualize the relationship between humans and AI in qualitative research is that they come together to form a jointly-operating system, a centaur (Dell'Acqua et al., 2023). In a system of this nature, the human and AI work together in an integrated fashion: task-by-task (or even within tasks), the human switches between doing work themselves and having the AI do the work task, attempting to allocate each task based on their perception of the AI's strengths and their own strengths. Chess players were already engaging in this practice as early as the 1990s, in tournaments that allowed players to use a chess engine during gameplay (Case, 2018).

This practice can be seen as intelligence amplification -- improving human performance through AI assistance (Freund, 2013). However, if we see the combined human-AI system as more than just the sum of its parts, we may want to go even further and view the centaur metaphor through the lens of distributed cognition (Hutchins, 1995). This perspective views cognition as being distributed across the researcher, the AI system, and the artifacts and representational tools created in the course of their interaction. Within this distributed cognitive system, information is transformed as it moves between components—between the human's thinking, the prompts they design, the representations generated in the LLM outputs, and the evolving analytic artifacts. As noted in (Friese et al., 2025), this framing also resonates with assemblage thinking (Deleuze & Guattari, 1980; DeLanda, 2006; Bennett, 2010), which similarly conceptualizes heterogeneous human and non-human elements as coming together in contingent, relational formations whose emergent properties go beyond those of any single component. Unlike the tool metaphor where the human directs and the AI executes, distributed cognition recognizes

that meaningful cognitive work happens in the human-AI interaction—in the iterative refinement of prompts, in the way AI outputs reshape human thinking, and ultimately in how human cognition occurs differently when using the AI than when working alone. Indeed, accounts of Centaur Chess argue that humans playing chess with this approach cultivate new capabilities that are different from the skills that are most important to unaided chess (Krakowski et al., 2023). So too, a human working with an AI in a closely-integrated fashion to conduct qualitative research might find themselves focusing their energy on new and different skills than in unaided qualitative research.

This metaphor has particular strengths for understanding certain phases of qualitative research. When developing codebooks, for instance, the researcher might not have fully formed categories in mind from the start. Through cycles of both the human and AI suggesting codes and refining definitions, the coding scheme emerges from the distributed system rather than from either party alone. The AI may surface patterns the researcher hadn't noticed; the researcher brings contextual knowledge and theoretical understanding the AI lacks; together as a centaur they arrive at interpretations that neither human nor AI would have reached independently. Similarly, during interview protocol development or exploratory data interpretation, ideas evolve through the back-and-forth rather than being fully specified in advance. This ongoing dialogue becomes the site where understanding develops, with the human and AI both constraining and enabling each other’s contributions.

However, this relationship metaphor also presents challenges that researchers must carefully navigate. Thinking about work by a centaur, with cognition distributed, raises questions about exactly where insights originated, which becomes problematic for the documentation expected in rigorous qualitative research. If insight emerges from the interplay of human prompt design, system-generated representations, and iterative refinement, where exactly does an idea originate? And how does a researcher adequately describe their analytic process and the informational transformations produced by the LLM, in a methods section? Furthermore, in traditional qualitative research, reflexive practice centers on the human researcher’s values, lived experience, positionality, and interpretive stance. Under distributed cognition, part of the interpretive process is enacted by the AI, which may lack true positionality but still shapes the trajectory of analysis. Reflexivity therefore may need to extend to examining how the AI’s representational biases, training data distributions, and textual priors influence emergent interpretations. This requires the researcher to reflect not only on their own epistemic commitments but also on how their engagement with the AI system co-produces analytic outcomes. It also requires researchers to attend to the degree to which LLMs encourage cognitive offloading; while this is a strength of distributed cognition, it can potentially diminish the researcher’s engagement with raw data.

Finally, LLMs differ in kind from the cognitive tools that are usually discussed in research and scholarship on distributed cognition -- or even the mostly-deterministic reasoning tools used in Centaur Chess. Unlike relatively stable artifacts such as written notes or diagrams, for example, LLM outputs can shift over time and their outputs are not perfectly replicable, even if the model temperature is set to zero (Atil et al., 2024) (which in principle removes all random variation). This acknowledgement of instability is also seen in assemblage thinking (e.g. DeLanda, 2006), in which sociotechnical configurations are understood as fluid and continually reconstituted rather than stable systems. Despite these complexities, the centaur metaphor and distributed cognition offer valuable food for thought for researchers who find their analytic processes and thinking genuinely reshaped through interaction with AI, rather than simply accelerated or supplemented by it.

A junior colleague. A third relationship is to view AI as a junior colleague -- able to contribute meaningfully to the qualitative research process, and perhaps in some cases to bring their own ideas to the research. A junior colleague -- perhaps a Masters student -- is not expected to direct the overall research process, and is not expected to fully understand all the nuances of the process they are participating in. They are likely somewhere in the process of transitioning from a legitimate peripheral participant to a more advanced role (e.g. Lave & Wenger, 1991) in the research process.

We can further think about this metaphor in terms of a “good” Masters student versus a less-effective junior colleague. A “good” Masters student will conduct the processes they are asked to engage in reasonably competently, sincerely, and at minimum to the best of their effort. Even a very intelligent, well-studied, and skillful Masters student can be expected to make errors sometimes. In addition, their faculty mentor may not know in advance how deeply they have studied the material they have learned, how skillful they are, and how willing they are to put in the full effort needed. As such, a more senior researcher working with a junior researcher is likely to frequently check their work, especially when a junior researcher begins a new step in the research process. These types of checks are common within qualitative research in general but are often more thorough when working with a junior colleague whose skills are not yet fully known.

Therefore, a researcher working with AI could treat it like a junior colleague. In this metaphor, the human researcher sets the overall directions and leads the process, while the AI does tasks such as executing steps that are time-consuming for the human researcher, offering a second opinion on decisions selected by the researcher, and helping unpack certain steps. One key difference between this metaphor is that it goes beyond simply executing time-consuming steps. The AI may take on broader parts of the process under the human researcher’s overall

guidance and oversight -- proposing what to do next. A researcher who adopts this relationship may also ask the AI for their opinion -- for instance, to revise qualitative coding definitions (Ch. 12), to suggest qualitative codes (Ch, 11), to propose or ask questions during an interview (Ch. 7), or to comment on interpretations (Ch. 15).

This metaphor acknowledges, unlike the tool metaphor, that the human may not always know in advance exactly what they want; exploratory research is common in qualitative inquiry. The human researcher may discover what they want in the process of seeing outputs, or they may change their mind on what they want as new insights emerge. Meaning is collaboratively co-constructed between human and AI, and the AI is welcome at times to push back when the researcher's notions are poorly conceived -- just like a Masters student has some scope to disagree or push back when working with a professor. (LLMs can sometimes find it difficult to disagree with a confident user -- Fanous et al., 2025 -- but the same is true of many Masters students!)

Importantly to this metaphor, the researcher does not assume that the AI's performance will be consistently high quality and carefully checks its work much as they would with a junior colleague who has not yet fully proven themself. Just as a professor would learn their students' limitations over time, the human researcher becomes familiar with the AI's limitations (and how they are different than human limitations), and learns how to work with or around these limitations. An AI may be competent at procedures and have wide-ranging knowledge, but miss key aspects of context or tacit knowledge, for example -- just like a Masters student who has read extensively in classes but not understood everything. Working with either a junior human colleague or an AI, an effective research lead learns iteratively where to extend more autonomy and where to monitor more closely.

However, this metaphor breaks down in some ways. Although a human user can learn to better prompt an LLM to achieve its goals, and contemporary LLMs build a memory about their user both within and across sessions, an LLM does not enculturate into a research group or community the way that a junior colleague does (Lave & Wenger, 1991). LLMs often behave coherently within a session, but across a long chat session, LLM behavior can become less predictable rather than more predictable (Li et al., 2024). Hence, the act of working with an LLM as a junior colleague becomes more one of the user learning to work with the LLM than teaching the junior colleague their practices -- at least as of the LLMs in use at the time of this writing.

A partner. A more radical metaphor to consider is treating the AI as a full partner in the research endeavor. In this metaphor, the AI is recognized as being different than the human scholar, but its perspective and contributions are treated as just as valued as the human's, even if they differ in key ways. This does not mean that the human scholar needs to compromise on their core goals and values; they would not need to do so in a partnership with another human either -- if such a partnership did not meet their values, they would find a different human to work with. A human scholar may not fully understand their AI partner's positionality (and indeed it may not have stable positionality), but this becomes something to take into account within the partnership. We do not always fully understand our human partners' positionality, nor are humans always self-consistent (Goffman, 1956). But we can make efforts to study how our partner operates and how we we can collaborate effectively with them (e.g. Lopez-Fierro & Nguyen, 2024)

This metaphor on the surface might feel similar to the junior colleague metaphor. The distinction is in terms of initiative allowed, the degree of input into decisions, and ultimately epistemic openness. Partners not only carry out tasks, but also shape them. In a partnership, a scholar is therefore more likely to let their partner propose the process (and then refine it or make suggestions). In a partnership, a scholar is more open to their partner's ideas on what constructs to study, how to study them, how to interpret the findings -- everything. If we view the AI as a full partner, we open ourselves to learning more from it. On the other hand, if we treat AI as a full partner, the human does not always have the final word on key process or interpretive decisions (which is perhaps easier to achieve in principle than practice, given that existing LLMs are in practice willing to defer to humans -- if nothing else, by simply starting a new session). In a partnership, it is important to recognize the LLM's limitations; but those limitations become something the scholar treats more like their own limitations -- to be taken into account, but honored as a part of them. This metaphor also relates to the centaur metaphor, in that both recognize the AI as genuinely contributing to knowledge production rather than simply executing researcher-defined tasks, but view this contribution as different in nature. Partnership considers the contributions of humans and AI on as equal a footing as possible, and in turn emphasizes the AI's role in shaping research decisions and conclusions. By contrast, the centaur metaphor focuses on the ways that human representation and cognition are transformed through the interaction between human and AI and the artifacts created by that interaction.

This metaphor of course has limitations. A human partner has positionality and perspectives -- an AI's positionality and perspectives are simulated and unstable to a greater degree and in different ways than a fellow human. Some perspectives on generative AI would argue that AI is fundamentally incapable of reflexivity (Jowsey et al., 2025). We agree that we cannot rely on

LLMs to exercise reflexivity or ethical judgment in the way human collaborators can. But as Friese et al. (2025) note, even if AI cannot be reflexive, humans can be reflexive in their use of it. As with the centaur metaphor, reflexivity therefore may need to extend to examining the AI in depth, including the AI's representational biases, training data distributions, and textual priors. Beyond this, LLMs fundamentally lack accountability for errors or problematic decisions -- the human researcher remains fully ethically and socially accountable. Furthermore, LLMs also do not have the same level of ownership or control as a human collaborator. A conversation with an LLM can be restarted; LLMs can in practice be ignored or dismissed when strong disagreement occurs. As such, a partnership remains fundamentally asymmetrical, and a scholar choosing to work with this metaphor must continually and actively work to treat the AI's outputs with equal weight to their own.

A critic or provocateur. A final metaphor to consider is that of a critic or provocateur. In writing this book, we have found LLMs to be invaluable for reviewing our sections not only for flawed writing but also for underdeveloped ideas and missing considerations. LLMs can simulate discursive positions associated with a wide variety of perspectives and use these perspectives to critique ideas. This same idea is often found in qualitative research, where analysts or a member of an analytic team takes a role of intentionally questioning interpretations, surfacing alternative explanations, and pushing the research team to justify their reasoning (Lincoln & Guba, 1985). In this role, the AI is not primarily engaged in producing analysis but in stress-testing it—asking "what if?" questions, challenging assumptions, pointing out potential blind spots, offering counterinterpretations and rival explanations, and generally serving as intellectual resistance to premature closure on findings. The researcher retains clear ownership of the research direction and interpretations, but actively seeks out the AI's challenge and dissent rather than its assistance in conducting the work. Unlike the partnership metaphor where the AI's perspective is valued as equal input, here the AI's ability to provide opposition and skepticism are specifically what make it valuable.

While large language model system prompts typically direct them to act as a "helpful assistant", and LLMs can often be overly agreeable (Fanous et al., 2025), it is possible to also prompt them to act in a more critical fashion. For example, Qreframer (Buckingham Shum, 2024) was designed to act in this fashion, though not specifically for the purpose of qualitative research. More recently, Gillespie (2026) proposed a protocol for qualitative researchers to use LLMs in this fashion, by prompting models to critique the researcher's own interpretation from a range of pre-specified perspectives, finding that the most productive critiques came from the perspectives most distinct from the researcher's own positionality. The key to generating such a prompt is to clearly articulate what stance and types of pushback, critique, or provocation are

desired -- and then to actually use the prompt, and to engage with its critiques the same way that one would interact with a human colleague's critiques.

This type of use has the potential to be very valuable to scholars without access to a large research team or colleagues with the perspectives they need. It also has the potential to be particularly valuable at certain junctures in the qualitative research process. When developing theoretical interpretations from data, a researcher might ask the AI to argue against their emerging theory, to surface overlooked dimensions of data, to propose rival explanations that fit the same data, or to identify evidence that contradicts their interpretation. This can help a researcher break out from moving to closure too soon in the interpretive process and avoid unintentional confirmation bias. It can also help a researcher in cases where their positionality may be constraining their interpretation. Prior to member checking, the AI can attempt to identify concerns that members might have (though not with the goal of skipping member checks entirely, since AI is unlikely to perfectly simulate participant perspectives). In each case, the value comes from using AI to systematically challenge the researcher's thinking in ways that surface weaknesses before they become embedded in the final analysis -- leveraging AI's ability to take on perspectives different from the researcher.

However, this metaphor has important limitations that distinguish AI critique from human critique. An AI's challenges lack the grounding in lived experience, disciplinary expertise, or deeply held epistemological or theoretical positions that produce the deepest scholarly disagreement. Its provocations may therefore be surface-level and even generic (particularly if prompts for doing so are not carefully designed). In Gillespie (2026), for instance, most of the counterarguments the LLMs generated were found to be "banal" or repetitive. The AI's willingness to attempt to adopt any perspective on demand means its critiques may lack the coherence and consistency that come from a human critic's stable intellectual commitments. A human provocateur pushes back because they genuinely interpret the world and the data differently; an AI does so because it was prompted to. Despite these limitations, this metaphor offers researchers a low-stakes way to encounter disagreement with their ideas, and can serve as valuable preparation for higher-stakes critiques from human reviewers, participants, and the scholarly community. When AI is used intentionally in this fashion, it can help researchers deepen their reflexivity and strengthen the credibility of their interpretations, even if the AI itself provides only simulated critique rather than substantively grounded disagreement.

d. Concluding thoughts: curiosity and understanding for our research partner

As this chapter shows, we have a great deal to think about when we think about working with AI: how it functions, how we can understand that functioning, and how we can understand our work

with it. In this chapter, we introduce five metaphors for thinking about how AI functions (helpful assistant, stochastic parrot, lossy mirror, alien intelligence, and simulacrum) and five metaphors for thinking about conducting qualitative research with AI (tool, centaur, junior colleague, partner, and critic/provocateur). There is not a clear one-to-one mapping between these two sets of metaphors (although some are inconsistent with each other -- one would not partner with a parrot, and it feels disrespectful to treat an alien intelligence as a tool) -- Figure 4.1 above shows some of the possible combinations. We do not recommend choosing a single metaphor for thinking about how AI functions, and in general also do not recommend choosing to work with AI only in a single fashion. Each of these metaphors illuminates different attributes of AI, capturing aspects about AI and our work with it that can be obscured by other metaphors. While the authors find some of the metaphors and relationships more in line with our research perspectives and more useful for our research contexts (partner, junior scholar), we want to note that none of these characterizations of AI are inherently right or wrong. Each metaphor and relationship has a set of affordances and constraints that may make them more relevant or useful for different researchers and contexts.

When we work with a fellow human in qualitative research (or indeed in any sustained and non-surface intellectual endeavor), we need to get on the same page with that human. And indeed, this is a major part of collaborative qualitative research: inter-rater checks, member checks, social mediation, and consensus-building discussions about coding frameworks, among other steps. We can work with another human most effectively if we have common ground with them -- shared knowledge, assumptions, and expectations (Clark & Brown, 2006). If this cannot be achieved on both sides, even one person having a rough mental model of the other's thinking can assist in collaboration (Herr & Anderson, 2005).

This provides a guideline for working with AI, which generates its outputs differently than us regardless of which of the metaphors for its functioning that one selects (though this idea is perhaps best articulated by the alien intelligence metaphor). We need to understand its "thinking" as much as we can, so that we can work with it effectively and use it appropriately. And we need at minimum to take the same steps to "get on the same page" as we would with a human, acknowledging that even this framing is less appropriate to an AI than a human. We need to understand to the best of our ability the interpretations, inferences, and discoveries that AI makes rather than simply accepting its conclusions unreflexively and moving on. AI can surface things to us as human analysts, but the key locus of interpretation needs to remain with humans, or we are arguably no longer conducting qualitative research (Friese et al., 2025). Maybe we cannot fully gain phenomenological understanding of an AI -- understanding the thinker as they understand themselves (Nagel, 1974) -- but we can still develop approximate

and imperfect understanding. This is perhaps not so different than the approximate and imperfect understanding that ethnographic researchers develop about the people they study or even each other (Clifford, 1986). This work to understand an AI research partner requires a level of curiosity that we often may not have when it comes to our human research partners -- but maybe we should.

And with that recommendation, we conclude this chapter on the foundations of thought around artificial intelligence. In the next chapter, we add the final foundation for our discussion of applications: a template pipeline for the qualitative research process that highlights the many possible entry points for the application of AI.

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