THESIS

THE EPISTEMIC LIMITS OF LANGUAGE MODELS: TRUST, TESTIMONY, AND GROUP UNDERSTANDING

Submitted by

Parker DuVall

Department of Philosophy

In partial fulfillment of the requirements

For the degree of Master of Arts

Colorado State University

Fort Collins, Colorado

Spring 2026

Master's Committee:

Advisor: Collin Rice

Jeff Kasser
Nikhil Krishnaswamy

ABSTRACT

THE EPISTEMIC OF LANGUAGE MODELS: TRUST, TESTIMONY, AND GROUP UNDERSTANDING

In recent years, the use of Large Language Models (LLMs) has become increasingly widespread. A major catalyst in this trend was the 2022 release of OpenAI's ChatGPT GPT-3.5 model for public use. With similar alacrity, the philosophical literature devoted to investigating these and other artificial intelligence (AI) systems has grown. This thesis contributes to this growing literature by aiming to answer the following question: how, if at all, can we use LLMs to obtain epistemic goods, particularly testimonial knowledge and group understanding? In service of this aim, this thesis is split into three chapters. Throughout these chapters, I make three main conclusions: (i) there are limited circumstances under which we can deem an LLM a credible expert, (ii) we cannot deem an LLM a credible standard testifier (iii) LLMs are incapable of contributing to group understanding. In chapter one, I review the existing literature in three areas relevant to answering these questions: testimony, understanding, and the philosophy of AI. In chapter two, I argue that LLMs can be deemed credible experts but not credible standard testifiers. I conclude this section by offering some potential explanations for this tension. In chapter three, I argue that LLMs cannot be contributing members to group understanding on either a deflationary or inflationary account of the phenomenon. Here, I show that LLMs cannot be attributed with non-metaphorical understanding or the requirements for contributing to group grasping.

TABLE OF CONTENTS

Chapter 1 – Literature Review

1. Introduction

In recent years, the use of Large Language Models (LLMs) has become increasingly widespread. A major catalyst in this trend was the 2022 release of OpenAI's ChatGPT GPT-3.5 model for public use. With similar alacrity, the philosophical literature devoted to investigating these and other artificial intelligence (AI) systems has grown – see Alan Turing (1950), Hilary Putnam (1960), John Searle (1980), Jerry Fodor (1987), David Chalmers (2024), Ori Freiman (2024), and Elizabeth Fricker (2025). This thesis contributes to this growing literature by aiming to answer the following question: how, if at all, can we use LLMs to obtain epistemic goods, particularly testimonial knowledge and group understanding? In service of this aim, this thesis is split into three chapters. Throughout these chapters, I make three main conclusions: (i) there are limited circumstances under which we can trust an LLM as an expert testifier, (ii) while we *can* trust an LLM as a standard testifier, this requires a lot of epistemic work on the hearer's part, and (iii) LLMs are incapable of contributing to group understanding. Chapter One is dedicated to demonstrating where in the literature this thesis sits, chapter two is dedicated to demonstrating the first two of these conclusions, and chapter three is dedicated to demonstrating the third conclusion.

The conclusions presented in this thesis bear heavily on the implications for use of LLMs. Particularly, the conclusions in chapter two show that there are limited circumstances under which users can justifiably trust an LLM. While it would be too strong to claim that there are no such circumstances, it is shown here that immediate and automatic trust of LLMs would be unjustified

under prominent models of testimonial trust. This is counter to the immediate trust many laypersons have in the technology. In addition, the conclusions presented in chapter three show that an LLM cannot be a contributing member to group understanding. Thus, LLMs can support individual understanding as a tool, and a powerful one at that, but they cannot contribute to group understanding. Moreover, we see in chapter three that LLMs are, currently, incapable of grasping at the individual or group level. Thus, they cannot possess individual understanding and cannot contribute to group understanding on either inflationary or deflationary accounts of group understanding. As a result, giving any cognitive processing work to an LLM on the individual or group level would be a misuse of the technology. Finally, the arguments here bear on the philosophical literature in a number of ways. In chapter two, we find that Alvin Goldman's (2001) criteria of what constitutes an expert are outdated with respect to LLMs. While the notion may work well for humans, it cannot accurately account for LLMs. In chapter three, we see that investigating LLMs as potentially contributing members allows us to break down group understanding into its further necessary components. As a result, recognition of relationships to other members is found to be a necessary component of inflationary accounts of the phenomenon.

In chapter one, I review the existing literature in three areas relevant to answering these questions: testimony, understanding, and the philosophy of AI. While the philosophy of AI plays a role in the entire thesis, the testimony and understanding literature play important roles in the second and third chapters respectively. In section one of the first chapter, I outline approaches within the testimonial justification, authoritative testimony, and expert testimony debates. In section two, I focus on differing accounts of what constitutes understanding, grasping as a requirement for understanding,

group understanding, and idealizations and the potential factivity of understanding. Finally, in section three, I examine the literature around what constitutes AI, what constitutes LLMs and if they are conscious, and finally, the epistemology of LLMs, with a focus on their potential status as testifiers.

In chapter two, I investigate whether LLMs can be credible experts and if they can be deemed credible standard testifiers. I start by outlining definitions for the concept of an "expert" and show that LLMs seem to fit the common platitudes given about experts and expertise. I then outline eleven criteria for credible expertise and show which criteria LLMs can be attributed with. I conclude in this section that LLMs can be deemed credible experts under the right conditions. Then, I consider if LLMs can be credible standard testifiers under three main approaches to testimonial justification: Reductionism, Anti-Reductionism, and Dualism. I conclude in this section that LLMs cannot be credible testifiers under any of these accounts and thus beliefs formed merely on the basis of their testimony are not justified. In the third section of chapter three, I offer possible explanations for these contradictory findings. Finally, I consider a number of objections to my arguments.

In chapter three, I argue that LLMs cannot be contributing members to group understanding on either a deflationary or inflationary account of the phenomenon. The deflationary account of group understanding holds that a group understands just in case "all or the majority of its members understands (why, how, that) p" (Boyd 2019, 17). The inflationary account holds that a group understanding "(why, how, that) p does not depend solely on whether its members understand (why, how, that) p" (ibid., 18).[1] After reviewing these accounts, I then lay out Kenneth Boyd's (2019)

[1] As Boyd notes in this section of his paper, the two accounts may just reduce to accompanying accounts of group knowledge, depending on how we view the relationship between knowledge and understanding.

overview of deflationary group understanding and consider if an LLM could contribute to a group that comes to understand on such an account. I argue that LLMs are incapable of understanding and to show this, I look at three accounts of understanding in the literature from Allison Hills (2015), Kareem Khalifa (2017), and Soazig Le Bihan (2017) and show how LLMs are incapable of meeting the required criteria for each account. Next, I lay out Boyd's overview of an inflationary account of group understanding and consider if LLMs, despite being unable to understand, could contribute to a group that collectively comes to understand. I argue that LLMs are incapable of *grasping* at the group level. Using Boyd's notion of group grasping, I argue that LLMs are incapable of recognizing their epistemic dependence on others and thus cannot be members of groups that collectively understand. Finally, I consider a number of objections to my arguments. I conclude in section five with a recap of where this leaves LLMs in relation to group understanding.

2. Testimony

Much of what we know about ourselves, others, and the world around us comes from the testimony of others. Despite this immense and obvious importance, much less epistemic trust has been placed in testimony as compared to sources of belief such as perception, memory, or reason. While not all facets of the debates within the testimony literature will be covered here, a cursory overview of the most pertinent (for our purposes), will be necessary. In particular, this literature will be important for chapter two when we consider whether LLMs can be used to produce testimonial knowledge and understanding.

2.1 Testimonial Justification

Likely the most central debate in the epistemology of testimony literature is the debate around testimonial knowledge and justification. Here the primary philosophical question is: what does it take for a hearer of testimony to come to know on the basis of a speaker's testimony? Consider the following scenario: You hear from your friend that the main town road is closed for construction (*p*). Since you have no reason to believe that your friend has any reason to lie to you, you know that your friend is a consummate truth teller, and you know that your friend is often right about these kinds of things, you believe your friend. Evidently, the road is indeed closed and since your belief was clearly justified, you too come to know that the road is closed. Now, contrast that scenario with the following: You hear from your new neighbor, whom you have never met, the main town road is closed for construction (*p*). You have no evidence that this person is a reliable testifer, but you also lack evidence that this person is unreliable in any way. Thus, you believe what you are told. Similar to the previous case, the road is indeed closed, but is your belief justified? How we answer this question gives us the three main approaches to testimonial justification.[2]

[2] There are some authors that fall outside of these three camps. Notably, John Locke's definition of knowledge precludes the possibility of knowledge via testimony. Locke claims that knowing is nothing but "the perception of the connection and agreement, or disagreement and repugnancy of any of our ideas" (*Essay*, IV.i.2). To be even more clear, Locke claims that "where this perception is, there is knowledge" and without this perception we may "fancy, guess, or believe" we nevertheless "always come short of knowledge" (ibid.). Since testimony is not a form of perception, Locke's definition of knowledge blocks any way for him to say that we may obtain knowledge via testimony. On the contemporary front, Jonathan Adler (1994, 2002) has advocated for a strict Reductionist position. So strict, in fact, that his view may collapse into a kind of skepticism about testimonial knowledge. Specifically, Adler claims that hearers have a default justification to trust others but this justification is not a pure entitlement, but rather can be empirically backed up with evidence that testimony is often reliable (2002, 147-153). That is, Adler believes that with inductive background beliefs, such that testimony from my friend Paul has always been reliable in the past and probably will be reliable in the future, we have a default justification in speakers. This may be too stringent to avoid skepticism about testimonial knowledge, however, as we rarely seem to have non-testimonial versions of these background beliefs. Adler's view may, then, either collapse into too strong of a Reductionism position to make sense of how we rely on testimony or could require him to appeal to shared experience (or similar notions) thus becoming more of an Anti-Reductionist view.

2.2 Reductionism

The first of these approaches is called "Reductionism." Reductionists claim that the hearer is justified in believing that *p* on the basis of a speaker's testimony that *p* only if that hearer has positive reasons for thinking that the speaker is a reliable testifier. In the first scenario above, the hearer has these positive reasons in the form of the memory of their friend's track record of truth telling. These positive reasons, however, might also come in the form of perception or inference, such as in the first scenario when the hearer reasons that the speaker does not have any reason to lie at this moment. As a result, a Reductionist would say that the belief in the second scenario is not justified as there is an absence of these positive reasons. That is, a Reductionist will say that you need an absence of epistemic defeaters (or, reasons to think that the speaker is unreliable) but you must also be in possession of *positive* reasons for belief in their reliability.

In the western tradition, Reductionism often traces back to David Hume, with his essay *On Miracles* in his work *An Enquiry Concerning Human Understanding* being the locus classicus.[3] Reductionists, as we discussed above, argue that when S believes that *p* because Q says that *p*, S's belief is justified only if S is also justified in believing that (i) Q actually said that *p*, i.e. S cannot be mistaken on this matter, and (ii) most of the things said by Q are true (Van Cleve 2006, 61). For Hume, the second of these considerations is the "usual conformity of facts to the reports of witnesses" (Hume 1993, 74). Moreover, Hume claims that any assurances we have in any argument derived from the "testimony of men, and the reports of eye-witness and spectators" is nothing else besides the

[3] Dan O'Brien (2020, 2023) argues in favor of an anti-reductionist interpretation of Hume. O'Brien motivates this with reference to Hume's *Treatise* rather than the *Enquiry*. While his arguments are interesting and at times persuasive, his work is in stark contrast to the standard interpretation of Hume's epistemology of testimony and thus will not be discussed here.

aforementioned conformity between facts and reports (ibid.). Much work in contemporary Reductionism is an effort to better understand and dissect Hume's Reductionism, such as Paul Faulkner (1998) or Saul Traiger (1993, 2010).

Reductionism is defended in various forms[4] in contemporary debates – see Elizabeth Fricker (1987, 1994, 1995, 2002, 2006), Jonathan Adler (1994, 2002), Jack Lyons (1997), Anna-Sara Malmgren (2006), and Tim Kenyon (2012). To give a specific example in the literature, Fricker (1994) claims that "a hearer should always engage in some assessment of the speaker for trustworthiness" and to do otherwise, i.e. to presume trustworthiness, is tantamount to "gullibility" (145). Fricker defines the trustworthiness of a speaker in terms of their sincerity and competence, such that if, on a given occasion, the speaker's utterance is sincere and the speaker is competent with respect to the proposition, the speaker is trustworthy (Fricker 1994, 147). Assessing this competence and sincerity, Fricker thinks, comes down to the hearer constructing a theory to explain the utterance of the speaker (ibid., 149). The explanation should "render it comprehensible why she made that assertion, on that occasion" and the trustworthiness will follow from a proper explanation (ibid.). Fricker briefly motivates this by reference to the feeling a hearer may experience when receiving a particularly unbelievable instance of testimony. If my friend testifies to having seen flying saucers, I may naturally try to explain why she is saying this. Is she joking with me? Has she gone crazy? Or, is this really true (ibid.)? Importantly, not all explanations for the instance of testimony lead to trustworthiness. Rather, we consider the various explanations for someone's utterance (in this case that they saw flying saucers)

[4] Though not covered here, Reductionism is often split further into two types, Local and Global Reductionism, with the former being more popular.

and determine if the best explanation is that their utterance is sincere and they are competent. If the best explanation is sincerity and competence (and therefore that the claim is likely true), then we can trust this person, but if the best explanation is that this person did not actually see flying saucers and they are mistaken, lying, etc., then we cannot trust this person's testimony.

Perhaps the largest factor in favor of Reductionism is the problem of gullibility. If we adopt a default trust to all testimony, as some versions of the view below argue for, we may thereby leave ourselves open to being deceived. Fricker (2006) considers this in detail in response to Goldberg and Henderson (2006). If we are constantly basing our belief on the basis of testimony that we have not "checked" and have instead assumed trustworthy we may find ourselves with many false beliefs, which is an "unsafe policy for belief-formation" (Fricker 2006, 620). For instance, Fricker (2002) considers an anonymous blogger on an anonymous website. Since we know nothing about this person we lack defeaters, but since we know nothing about this person it seems that beliefs formed on the basis of their testimony would be unjustified. The counter, then, is to be assessing the trustworthiness of speakers, much as Fricker (1994) advocates.

2.3 Anti-Reductionism

Conversely, we might look at scenarios in which we lack both positive reasons or defeaters and think that, counter to the Reductionist position, there is nothing wrong with believing what you are told. This is, in a rough sense, what an Anti-Reductionist about testimony will hold. More specifically, Anti-Reductionists hold that we have a defeasible but presumptive right to believe what people tell us. Similar to Reductionism, Anti-Reductionism has its western tradition roots in the early modern period with Thomas Reid, who was a vehement critic of Hume's theory of testimony. Reid defined his

theory of testimony with reference to two main principles: veracity (the propensity in humans to speak the truth) and credulity (the propensity in humans to believe what others tell them) (Reid 1983, 95). In essence, since humans naturally have a propensity to tell the truth, we also have a right to have a default trust in the testimony of others. Truth and telling truth, for Reid, is a natural inclination, and even those which we may typically brand as "liars" still tell the truth much more than they lie (ibid., 94).

More modern approaches to Anti-Reductionism can be seen in the work of Jonathan Hardwig (1985, 1991), C.A.J. Coady (1992), Tyler Burge (1993, 1997), Alvin Goldman (1999), Christopher Insole (2000), Sanford Goldberg and David Henderson (2006), and Judith Baker and Philip Clark (2018). A primary reason to side with the Anti-Reductionists is the problem of childhood testimonial beliefs. If what is required for justified testimonial beliefs is positive reasons, it is much harder for children to acquire these beliefs, as they often rely on brute trust in teachers and parents. Anti-Reductionism, on the other hand, is able to allow for relatively easily justified testimonial beliefs and testimonial knowledge. After all, if positive reasons are not required, little stands in the way of these incredibly important epistemic goods.

As a detailed example, we can consider Insole (2000). Insole formulates a transcendental argument against the Reductionist and in favor of a presumptive right to trust the testimony of others. Since it is undeniable that we gain knowledge on the basis of testimony, it is either the case that it is "possible generally for the hearer to obtain independent confirmation that an assertion by a given speaker is trustworthy," or the hearer has the "epistemic right" to believe the testimony in virtue of it being asserted (Insole 2000, 46). However, since it is not possible, generally, for the hearer to obtain

this independent confirmation, the hearer must have the aforementioned epistemic right (ibid.).[5] Insole's main objective is to show this final step, that a speaker's trustworthiness cannot be independently confirmed, is true. In order to do this, Insole attacks the common "developmental/mature phase" defense given by Reductionists (2000, 51). This defense claims that humans are in a developmental phase as children and are given the ability to have a default trust in parents, teachers, etc. As they grow up, they enter the mature phase when they must have positive reasons for trusting instances of testimony. Insole questions, among other things, why we do not have to go back and reconstruct and reassess all testimonial knowledge acquired during the developmental phase after entering the mature phase (ibid.). Secondly, he attacks the Reductionist's use of default settings while still denying a default epistemic right (ibid., 53). That is, he critiques Reductionist approaches that allow for a default setting of believing that testifiers are sincere and competent speakers (ibid.). For him, the value of default settings cannot be captured by a Reductionist account, and instead should point us to the value of an Anti-Reductionist account (Insole 2000, 53). Insole's main claim here is that the only way a reductionist can account for knowledge acquisition via testimony is with reference to either a dual phase or default setting approach. Since neither of these seem to be satisfactory accounts, however, we are left with the other option in his transcendental argument: Anti-Reductionism.

[5] It should be noted here that I think Insole may be playing fast and loose with this transcendental argument. Particularly, it does not seem that the only two available options to account for testimonial knowledge are a strong epistemic right or a restrictive independent confirmation. Among other things, these two do not seem to be automatically mutually exclusive. Moreover, in framing this argument such that the Reductionist position is incredibly restrictive, the deck is stacked against them. In doing so, Insole comes close to begging the question to get the Anti-Reductionist position pole position.

Anti-Reductionists nevertheless have to deal with the problem of gullibility as outlined above. Goldberg and Henderson (2006) argue that Anti-Reductionism can incorporate a "monitoring requirement" such that a hearer must be "on the lookout" for defeaters rather than "going out and looking for defeaters" in a classic Reductionist sense (610). They use an example of Smith, who by default accepts the testimony of others, but has a buzzer that goes off when anyone tells her something that is not reliable. Smith never accepts testimony that elicits a buzz. Her presumption to accept testimony is cancelled when an instance of testimony elicits a buzz. While her testimonial beliefs are sustained via an external device, Smith is able to have an Anti-Reductionist approach to testimony while still enjoying the benefits of monitoring (ibid., 611).

2.4 Hybrid Approaches

Rather than take up Reductionism or Anti-Reductionism, some philosophers have attempted to combine the two in order to, in theory, gain the benefits of both while avoiding the pitfalls each approach brings. For instance, Fricker (1995) considers a theory in which adults must possess the positive reasons favored by the Reductionists, but children are allowed to place a default trust in the testimony of others, thus allowing for straightforward testimonial knowledge when young but reduced gullibility concerns when older. In a similar vein, Jennifer Lackey (2008) argues that a hearer must have positive reasons (albeit rather weak ones) in addition to a speaker actually being a reliable reporter. Put alternatively, it takes two to tango. The upshot for Lackey's approach is that it may allow for a side-stepping of the standard worries of Reductionism and Anti-Reductionism by requiring contributions from both parties in a testimonial exchange. Duncan Pritchard (2006) and Keith Lehrer (2006) offer defenses of more hybrid views as well. There may be credence to these hybrid views, but

whether or not they work as well as hoped is very much still open to debate. Any proponent of a hybrid view must argue why they do not run into the same problem as either of the two standard approaches, or create and face new problems entirely.

2.5 Authoritative and Expert Testimony

Moving beyond basic justification criteria, there are cases where justification seems to change greatly depending on speaker credibility. A person may be a reliable reporter on the habits of alligators in Melbourne, Florida because they live there and observe the animals on occasion. On the other hand, a wildlife expert who specializes in alligators is not only also a reliable reporter, but is surely a *more* reliable reporter. Such a case shines a light on a number of questions. For our purposes, however, we will cover just two: (i) how do we account for cases of authoritative testimony where the speaker is in an epistemically superior position and (ii) when there are two conflicting instances of expert testimony, how should a non-expert hearer decide which to trust?

First, a quick overview of the terms "authoritative" and "expert" testimony. Authoritative testimony, as it is often used in the literature, refers to the testimony of a speaker whose epistemic position is superior to that of the hearer. For instance, Christoph Jäger (2025) defines epistemic authority as "authority we ascribe to people in virtue of their favorable relation to epistemic goods" (63). Linda Zagzebski (2012), who has set much of the agenda on epistemic authority, defines the concept as "someone who does what I would do if I were more conscientious or better than I am at satisfying the aim of conscientiousness–getting the truth" (109). For instance, if I live in California and my friend lives in Florida, my friend is in a superior epistemic position with regards to the weather in

her part of Florida. If my friend, further, is a working meteorologist, then she is a further authority on the weather in California.

Expert testimony refers to the testimony of a speaker who is an expert in domain *p*, where *p* is the topic of their testimony.[6] An expert can be roughly understood as someone who knows or understands much of a given domain and can generate new knowledge or understanding within that domain. Expert testimony, therefore, can be a kind of authoritative testimony, if the expert is, for instance, talking to a non-expert or talking to another expert but nevertheless has an epistemically superior position. With these terms laid out, we can start to unpack the first of our previously established questions: with regards to authoritative testimony, how should the hearer use the authority's testimony to justify their own beliefs?

2.6 Approaches to Authoritative Testimony

There are two main approaches to answer this question, Preemptive and Non-Preemptive accounts. Zagezebski's (2012) Preemptive account in particular has set the tone for much of this debate. In conjunction with her definition of epistemic authority, given above, Zagzebski puts forward two theses that claim that if I believe what the authority believes I am more likely to form a true belief that "survives my conscientious self-reflection" (Zagzebski takes the truth to be what ultimately survives conscientious self-reflection) than if I attempt to figure out what to believe myself (2012, 110-111). These theses yield the preemption principle, which claims that the fact that an authority testifies that *p* is a reason for me to believe that *p* and this reason replaces my other reasons to believe that *p*

[6] For many questions within the domain of expert testimony, Goldman (2001) is the locus classicus. However, since Goldman's remarks are discussed at length in a later chapter, our discussion of him here will be limited.

(Zagzebski 2012, 107). When you gain this reason for believing that *p* from the authority, your other, previous reasons no longer count for or against *p*. On this account, authoritative testimony gives you a preemptive reason for a belief that *p* whereas non-authoritative testimony can only give you one reason among others. Zagsebski's account is not the only one, but most Preemptive accounts run very closely with Zagzebski's (2012). More defenses of Preemptive accounts can be found in Zagzebski (2014, 2016), Michael Croce (2017), and Jan Constantin and Thomas Grundmann (2020).

On the other hand, we may think that Preemptive accounts face a problem. For instance, an authoritative speaker may be otherwise unreliable, and thus it may be problematic to disregard all other reasons for belief that *p* or not-*p* simply because the authority says so. Charity Anderson (2014) and Trent Dougherty (2014) offer potential worries with Preemptive accounts and Zagzebski's (2012) in particular. Specifically, Anderson claims that Zagzebski's (2012) account misses out on the nuance of disagreement, while Dougherty (2014) attacks Zagzebski's implication that authority is a binary notion rather than something that comes in degrees. Non-Preemptive accounts argue that an authority's testimony that *p* provides a strong reason to believe that *p*, but that reason is added to all other reasons the hearer has for believing that *p*. Lackey (2018) and Katherin Dormandy (2018) offer defenses of Non-Preemptive accounts.

2.7 Expert Testimony

Some authorities, of course, are actual experts within their fields. While it seems clear that testimony from an expert is important, it can be hard to identify an expert when anyone could claim to be one. Other authors, too, have given alternative answers to this question. Thomas Grundmann (2025), for instance, lists track record, dialectical competence, teaching competence, reputation, and

scientific selection as positive indicators of expertise for novices to use (97). Elizabeth Anderson (2011), who has given one of the more recently influential accounts of trust in experts, gives three categories for assessing what expert(s) to trust, these being their level of expertise, honesty, and responsiveness to opposing views (145-146). Each of these categories is broken down further by Anderson into criteria for judging where an individual stands. For instance, someone with a Ph.D. who is trained in their field has more expertise than a layperson, but less expertise than someone with a Ph.D. and is actively researching within the field (ibid., 146). Matthew Bennett (2022) argues supposed problems emerge with Anderson's criteria when we consider the distinction between epistemic trust (I believe that *p* because *S* said so) and what he calls recommendation trust (I should do *x* because *S* says so). This is because, according to Bennett, trust in an expert as an informer can come apart from trust in an expert as an advisor (2022, 554). For instance, I might trust my doctor when she says that chemotherapy is an effective treatment for certain types of cancer, but I might not follow her recommendation that I personally undergo chemotherapy. Perhaps, in this instance, I believe that my doctor is advising without proper knowledge of what I value. I could thereby trust my doctor's testimony but not her recommendation. Gürol Irzik and Faik Kurtulmus (2019) use Anderson's criteria to examine the alleged connection between the MMR vaccinations and autism. Alexander Guerrero (2019) offers expertise, comparative expertise, sincerity, and quality of the testimony as possible signifiers of reasons to trust an expert. Finally, Ben Almassi (2012) and Boyd (2022) both consider trust in scientific experts, with Boyd focusing on the different ways the problem manifests on and off the internet.

2.8 Conflicting Expert Testimony

Even if we can identify experts, we are almost always able to receive the testimony of multiple experts. When a large group of experts agrees, it is straightforward for a novice to trust their expert testimony. Anderson (2011), for instance, both argue that science is the most reliable way to truth and thus a layperson has no better alternative than trusting a scientific consensus. Others, such as Boaz Miller (2012, 2019) or Oreskes (2007) argue that expert consensus is not always justified (e.g., a consensus among doctors claiming that smoking is safe) but there are hallmarks of justified consensus. Oreskes (2007), specifically, lists five of these hallmarks: inductive support, predictive success, resilience to falsification, convergence of different evidence types, and explanatory success (80-91). Further accounts of consensus are given in Helen Longino (2002), Aviezer Tucker (2003), and Stegenga (2014).

No matter the account we give, it seems that we intuitively want to trust scientific consensus, perhaps not in all but many scenarios. It is less straightforward, however, to choose between the testimony of two conflicting experts. The problem here primarily stems from the fact that the content of and evidence for the expert's testimony is often inaccessible to a novice hearer, such that the hearer is unable to assess the truth value of the testimony. That is, the novice does not know enough about the expert's domain to reliably determine if the content is trustworthy. When there is merely one expert in question in a trustworthy environment (e.g., a professor teaching an introduction to physics class) this poses little problem. When a novice has to decide between the conflicting testimony of two experts, however, such as one expert claiming that Kant said *x* when another expert says Kant really

said *y* (where *x* and *y* are incompatible), not being able to assess the truth value of the testimony is problematic.

How then, should novices choose between the testimony of conflicting experts? Goldman (2001) offers the classic answer to this question. He outlines a number of criteria a novice can use to evaluate which expert to trust, such as credentials. For instance, if one expert has a PhD in the domain while the other does not, the hearer can use this as a reason to believe that the first expert is more trustworthy. The remainder of Goldman's criteria are: dialectical superiority, agreement from other experts, biases, and track records (2001, 93). Finnur Dellsén and Øystein Linnebo (2026) offer a much more wide-ranging approach to expert disagreement. Specifically, they identify three main strategies one could use when faced with expert disagreement. First, one could pick an expert out of the bunch and defer to that person's judgments entirely (ibid., 3). Second, one could attempt to aggregate expert opinions and then defer to the aggregate on what to believe (ibid.). Third and finally, one could simply wait and hold off on forming any beliefs until "one's epistemic situation with respect to the expert disagreement has improved" (ibid.). These three responses, claim Dellsén and Linnebo, are all warranted in different situations given different explanations for the expert disagreement. As such, they argue that laypeople should react epistemically differently to expert disagreement dependent on what is causing the disagreement (ibid, 2). While this approach is pluralistic, they pair it with a monistic zetetic prescription. Essentially, since different explanations of expert disagreement call for

different epistemic responses, laypeople should attempt to inquire about what is the underlying explanation of the disagreement to know which approach to take. [7]

3. Understanding

While the focus of most epistemologists discussing testimony has been knowledge, there has also been a recent shift in epistemology away from knowledge and towards alternative epistemic states such as understanding – see Jonathan Kvanvig (2003), Stephen Grimm (2006), Pritchard (2009), Alison Hills (2015), Soazig Le Bihan (2017), and Kareem Khalifa (2017) among others. As a result of a new focus on understanding as a possible object of epistemological inquiry, epistemologists have started asking questions about the nature of understanding. We might naturally wonder about its nature and the various ways we use the term "to understand." For instance, we often use the terms "understanding *that*," "understanding *how*," and "understanding *why*," and thus might consider how to conceptualize the different kinds of understanding. When considering these variations of understanding, we often see authors use a notion of "grasping" information and the connections between the information. In addition, we might consider if a group, rather than a mere individual, can be said to possess this epistemic achievement. Finally, many philosophers have wondered if understanding, like knowledge, is factive.

[7] Some authors have not explicitly given answers to this question but we can nevertheless use their claims as criteria for discriminating between conflicting expert testimony. For instance, Anderson (2011) gives criteria for assessing expertise. These criteria, however, come in gradients and thus could be used to assess if someone has more or less expertise than others.

3.1 Types of Understanding

Understanding-that is often called propositional understanding. In his influential discussion of understanding, Kvanvig (2003) introduces propositional understanding as the kind of understanding apparent in utterances such as "Keith understands the bank will be open tomorrow." Some authors, such as Gordon (2012) argue that when we attribute propositional understanding we are also attributing some propositional knowledge or one of the more extensive types of understanding outlined below. For instance, in virtue of understanding that the bank will be open tomorrow, Keith also *knows that* the bank will be open tomorrow.

Objectual understanding is attributed in cases where we say "S understands *x*," where *x* is typically a subject matter or body of information (Kvanvig 2003). We can think of objectual understanding, Grimm (2011) suggests, as understanding a system with dependent parts. For instance, someone who "understands U.S. politics" has an understanding of the object that is the system and structure of U.S. politics, including but surely not limited to the procedures of the three branches and how they are dependent on and interact with one another.

One of the more under-investigated types of understanding is understanding *how to do something*, which Zagzebski (2008) calls "understanding-how." She claims that "one gains understanding by knowing how to do something well," which makes you a reliable person to come to with matters about this particular skill (Zagzebski 2008, 144). We might draw out this kind of understanding by appealing to getting to a location. If I memorize a route from my hotel to the American Museum of Natural History in New York City, I might *know* how to get to the museum. On the other hand, an experienced taxi driver will *understand* how to get to the museum, which can

be shown in their ability to, among other things, take multiple routes depending on traffic or construction. This kind of understanding does not involve explanation as many other types do, but rather is focused on skillful action. Michael Hannon (2021) uses the example of a professional athlete who is unable to give a proper explanation of their incredible athletic achievement (24). While a professional baseball player, for example, may understand how to hit a fastball, they may also struggle to provide a sufficient explanation and instead rely on cliché one-liners such as "just take it one pitch at a time."

Finally, understanding-why. This type of understanding, often called "explanatory understanding" is typically the kind of understanding that comes to mind first. I might know that when I press my foot brake my car slows, but my mechanic really understands all of the intricacies and thus he understands why my car slows when I press my brake. We might be inclined to test if someone really has understanding-why *p* by having them explain why *p*. Many, such as Carl Hempel (1965), Wesley Salmon (1984), Khalifa (2017), and Michael Strevens (2013) argue that understanding in essence amounts to having a correct explanation. Others, such as Peter Lipton (2008) or Christoph Kelp (2015) argue that explanations are not necessary. On a historical note, Jaegwon Kim (1994) stated that, despite the reservations toward understanding held by philosophers of science at the time, leaving understanding out of accounts of explanation would be a mistake, as we desire to explain phenomena precisely because we want to understand them. Hills (2015) and Khalifa (2017), two more recent conceptions of understanding, are outlined in the third chapter. A more novel and niche kind of understanding that is closely related to the literature on understanding-why and scientific understanding is modal understanding, which can be seen best in Le Bihan (2017) and Collin Rice

(2021), the former of which is outlined in chapter three. For Le Bihan, someone has modal understanding of some phenomenon *p* "if and only if one knows how to navigate some of the possibility space associated with the phenomenon" (Le Bihan 2017, 112). Modal understanding of a phenomenon, more simply, is some understanding of how a phenomenon *might* arise.

3.2 Grasping

For many of these accounts of understanding, as well as many others, the term "grasping" comes up quite frequently. Strevens's (2008) "simple view" claims that an individual has scientific understanding when they grasp a correct explanation. To grasp, for Strevens, is to "understand that [a state of affairs] obtains," thus, "grasping is a kind of understanding" (Strevens 2013, 3). Importantly, Strevens staves off potential worries with circularity by claiming that the kind of understanding that is constituted by grasping is a fundamentally different type than is characterized by the simple view. While grasping a correct explanation constitutes understanding-why (simple view), grasping is a kind of propositional understanding, or understanding-that (Strevens 2013, 3). As Strevens (2013) puts it, "the simple view is an analysis of understanding why, a view that is couched in terms of grasping propositions, which is a matter of understanding that" (4). Elsewhere, Strevens (2024) puts forth a theory of grasping that claims that grasping some property is "a matter of knowing what it is to instantiate the property," such that grasping *x* is constituted by the ability to recognize instances of *x* (Strevens 2024, 16).

Similar to Strevens (2013), Khalifa (2013) argues that grasping is constituted by explanatory evaluation, such that one is able to differentiate between different explanations. Khalifa (2017) expands on his account of grasping and its relationship to understanding by putting forth the

"Scientific Knowledge Principle," which claims that if A's grasp of *p*'s explanatory nexus (i.e., the set of correct explanations for *p* as well as their relationships) is closer in resemblance to scientific knowledge than B's grasp, A understands why *p* better than B (11). Grasping, then, is a cognitive state bearing resemblance to scientific knowledge of the explanatory nexus of *p* (ibid.). Scientific knowledge, moreover, is constituted by a safe belief of an explanation that is safe in virtue of "scientific explanatory evaluation" (ibid., 12). Safe here refers to the belief forming process's inability to lead to a false belief (ibid.). Scientific explanatory evaluation (SEEing) refers to a three-step process. First, scientists consider the many potential explanations for some phenomenon (e.g., *x, y,* or *z* explain why *p*). Second, scientists compare the explanations before them (e.g., *x* better explains why *p* than *y* or *z*). Finally, scientists form doxastic attitudes about these comparisons (e.g., the belief that *x* is the best available theory) (Khalifa 2017, 12-13).

Khalifa's account here is a kind of ability account, where understanding or parts of it are constituted or displayed by an ability. Abilities accounts of grasping have also become popular in recent years. Hills (2015), for instance, describes grasping as a kind of cognitive control over *p* and reasons related to it. Kvanvig (2003), Grimm (2006), and Daniel Wilkenfeld (2013) give various discussions about grasping and its nature, with Henk de Regt and Dennis Dieks (2005), Grimm (2006), and Khalifa (2012) outlining something similar to Strevens's simple view. Grasping, no matter how we attempt to define it, is viewed as essential to understanding, as the difference between understanding and knowledge is often shown with a person obtaining understanding when they have cognitive control over the phenomenon, the reasons related to and supporting it, and the various connections between information.

3.3 Group Understanding

Finally in our discussion of understanding, we might wonder how all of these pieces, especially grasping, operate when a group comes to understand. Boyd (2019) is perhaps one of the more influential papers within the topic of group understanding. His views on the subject are the cornerstone of the third chapter, and thus will not be discussed here. Malfatti (2022) gives a similar outline of the notion. Broadly speaking, there are two ways in which a group can be said to understand. The first is often referred to as summativist or deflationary group understanding. A group reaches summativist group understanding when all or a majority of the members are said to understand (ibid., 9). For instance, when we say that the scientific community understands climate change, we are really just claiming that all or a majority of the scientific community understands the phenomena that make up climate change.

The second approach to group understanding is often referred to as non-summativist or inflationary group understanding. Non-summativist group understanding can be seen when few or none of the members understand, but the group comes to understand as a whole (ibid.). This is less intuitive, but imagine the following scenario. The crew of the USS Abraham Lincoln all have individual jobs. Some are managers, some are mechanics, and some are engineers. They all have the relevant understanding with regards to their particular job, but no one individual has relevant understanding with regards to operating the entirety of the USS Abraham Lincoln. Nevertheless, it seems correct to say that the crew, as a group, has the relevant understanding with regards to operating the entirety of the USS Abraham Lincoln. Boyd (2019) and Malfatti (2022) both take a non-summativist or inflationary approach to group understanding.

While the literature around group understanding is still budding, it has its roots in the literature of collective epistemology as a whole. Much work has been done on collective belief, judgments, and knowledge. Some of this work can be seen in Margaret Gilbert (1989, 2004), Lackey (2016, 2021), and Christian List and Philip Petit (2011).

3.4 Is Understanding Factive?

In making the shift from investigating knowledge to investigating understanding, epistemologists have had to ask how the two epistemic achievements differ in form and components. For instance, knowledge is generally agreed to be factive. That is, necessarily, if someone knows that *p*, *p* is true. A belief can be justified but if it is false, it always fails to be knowledge. An obvious question, then, is if understanding works the same way. Kvanvig (2003) argues that understanding is generally not factive in the same way we take knowledge to be. It is not the case, however, that one can hold a set of entirely false beliefs and nevertheless understand something. Rather, as long as the propositions central to the understanding are true, the "peripheral" propositions can be false and we can still correctly attribute an individual with understanding (Kvanvig 2003, 201). We might, as Kvanvig notes, still want to claim that this kind of understanding is defective or imperfect, but the important point is that it is still genuine understanding.

Along the same lines, Catherine Elgin (2007) argues that a "non-factive explication of understanding yields a concept that better suits epistemology's purposes than a factive one does" (34). Elgin claims that a factive account of understanding cannot make sense of the ways in which science often involves beliefs that are strictly false but are nevertheless valuable (e.g., idealizations) (Elgin 2007, 34). Rather, Elgin claims that understanding is a "grasp of a comprehensive body of information that

is grounded in fact," is responsive to evidence, and allows for inference, argument, and action (ibid., 39). As opposed to Kvanvig (2003), who claims that peripheral falsehoods do not preclude understanding, Elgin instead claims that falsehoods do not preclude understanding when they play a proper epistemic role, even when they are central to the body of information being grasped. Specifically, certain falsehoods allow us to "connect, synthesize, and grasp" a body of information that is otherwise grounded in facts (2007, 41). Elgin's primary examples are idealized scientific models such as the ideal gas law (2007, 15). While the ideal gas law is literally false (there are no gases composed of dimensionless, spherical molecules that have no mutual attraction) it is nevertheless important for our understanding of thermodynamics and figures into the genuine understanding of how real gases behave.

Finally, Angela Potochnik (2017) takes a very similar approach to understanding. Specifically, she argues that since idealizations are such a crucial part of scientific inquiry, science's "epistemic aim must be something other than truth" (Potochnik 2017, 93). As Potochnik puts it, "idealizations are assumptions made without regard for whether they are true, generally with full knowledge that they are false" (2017, 2). Some immediate examples include a frictionless plane or a perfectly spherical object in physics, ideally rational agents in economics, or the ideal gas law. Despite being explicitly false, there are good reasons scientists nevertheless use these idealizations. Potochnik (2017) considers a number of these, including but not limited to complex causal structures, computational limits, enabling of general application, or a limited research focus (48). Many authors, such as Elgin (2004) and Yasha Rohwer and Collin Rice (2013) argue that idealizations are beneficial because they

contribute positively to understanding, a view which Emily Sullivan and Kareem Khalifa (2019) frame as “epistemic instrumentalism.”

4. Philosophy of AI

In broad strokes, the field of artificial intelligence (AI) is the field devoted to creating artificial persons, or entities that appear to be intelligent in suitable contexts. This section, however, is focused on a handful of questions and areas pertinent to this thesis. Namely (i) what LLMs are, (ii) are LLMs (or other AI for that matter) conscious, and (iii) the epistemology of LLMs. The modern field can be traced back to at least Turing (1950), who established the “Turing Test,” which a computer or AI can be said to pass if its responses are indistinguishable from that of a human. AI, however, can be quite a nebulous and vague term, especially with the commercial advent of Large Language Models (LLMs) which are often themselves merely called “artificial intelligence.” Russell and Norvig (2009), under the assumption that the question of “What is AI?” should be answered in terms of goals of the field, split up the answers along two axes. The first axis asks if the goal is to build AI that is reasoning or human based, while the second asks if the goal is to build AI that has human-like rationality or has ideal rationality. This gives us four possible answers. The field of AI is the field that has the goal of building: (i) systems that think like humans, (ii) systems that act like humans, (iii) systems that think rationally, and (iv) systems that act rationally. This punnet square may not perfectly capture every approach to AI, as there are surely some that cross the boundaries. Nevertheless, many within the literature fall within these four categories. John Haugeland (1989), for instance, falls into the first category when he talks about machines with minds, Patrick Winston (1992) falls into the third category when he talks about machines thinking rationally, Russell and Norvig (2009) themselves fall into the fourth

category when they talk about rationally acting machines, and so on. With these various options laid out, we can start to look at the specific type of AI that will be focused on in this thesis: Large Language Models (LLMs).

4.1 What are LLMs

LLMs are, in a rough sense, AI systems that are trained to generate human language based on user inputs. They are trained on immensely large amounts of data, often being fed any human generated text their proprietors can get ahold of. Once trained, they work by predicting the next word (or token, which is a small chunk of text). By repeatedly predicting the next word or token, they can produce text and perform a number of language-related tasks. Since they operate on a large scale (often containing billions of internal variables) they are able to carefully mimic human language. Mimic is an important point here, as it can often seem like LLMs understand human language themselves. Despite this, LLMs are able to do a variety of language related tasks, typically at a speed inaccessible to humans. They are not, however, perfect.

Yufei Guo et al. (2023) and Adrian Mirza et al. (2024) show that, in certain problems in chemistry, LLMs outperform human chemists. This is combined, however, with the tested LLMs making mistakes on lower-level questions despite accuracy and proficiency with high level questions. Sébastien Bubeck et al. (2023), Jason Wei et al. (2022), and Takeshi Kojima et al. (2022) also show LLMs performing exceptionally well on reasoning-based tasks. This mirrors the findings from Sean Williams and James Huckle (2024), who show that LLMs offered by leading proprietors such as OpenAI, Anthropic, Meta, Mistral, and Google "have difficulty answering novel questions that humans find relatively easy," thus highlighting a gap between the current performance of LLMs and

the capabilities of human reasoning (14). Along the same lines, Andrea Matarazzo and Riccardo Torlone (2025) put it best when they say that "LLMs have shown remarkable capability" but their shortcomings are "equally evident" (151). LLMs perform incredibly well at reasoning-based tasks, leading to a debate about whether or not they possess human-like reasoning capability (e.g., Melanie Mitchell and David Krakauer (2023) and Ali Borji (2023)). Critics and skeptics of this position, however, contend that an LLM doing well in a reasoning-based task is not a reflection of their human-like capabilities but is instead a reflection of their ability to memorize and regurgitate training data (Yasaman Razeghi et al. (2022), and Shengyu Zhang et al. (2023)). The debate of whether or not an AI system could possibly have human-like mental qualities is not a new debate, however, and can be traced back to debate over goals to create "strong" or "weak" AI.

4.2 Are LLMs (or other AI) Conscious?

Strong and weak AI refers to the capability of a theoretical AI. Strong AI refers to AI that have all the mental powers that humans have. These mental powers include but are not limited to consciousness. Weak AI refers to AI that are efficient information processing-machines that merely appear to have all of the mental powers of humans (Searle 1997). While LLMs are firmly in the weak AI category, it is worth mentioning Searle's famous argument against strong AI, given that the appearance of LLMs to have the consciousness of humans can at times make it seem like they ought to be in the strong AI category. In broad strokes, the Chinese Room Argument, from Searle (1980) is as follows: Searle, who does not know Chinese, is in a box. Chinese speakers slide cards with questions written in the language into the box, and Searle, who has a book that tells him which characters to output given the input, slides cards out with answers written in Chinese. Searle knows nothing about

the semantics of the questions, but the book gives him access to the syntax, and thus allows him to create output that makes it *look* like he knows Chinese. As the argument goes, this is all that computers are. While this argument can be used against LLMs and other AIs, as long as one does not incorrectly attribute them with human mental powers, we can still investigate them without falling prey to the Chinese Room Argument.

4.3 Epistemology of LLMs

Finally, we can look at the state of the literature surrounding the epistemology of AI and particularly LLMs, which is largely negative. That is to say that many authors do not think that LLMs fit into the standard roles or mechanisms that epistemology has developed over the years. The most-lively part of this literature is focused around the notion of LLMs as potential testifiers, a topic that is central to the next chapter. This topic has been the main focus as it is the primary reason many are using AI systems, and particularly LLMs, at the moment. An individual will have a question or will want to learn about a topic, and will have a conversation with an LLM, thus forming beliefs on the basis of this conversation. Typically, we would call this exchange a testimonial one, and the LLM would take on the role of speaker. Many, however, such as Fricker (2025), claim that AI systems cannot take on this role of testifer as they do not possess communicative intentions. Furthermore, AI-based testimony is of a different epistemic kind than the human variant, as an AI cannot take responsibility for its outputs. Robert Sparrow and Gene Flenady (2025), similarly, claim that, in virtue of the fact that AI cannot take responsibility, the assertions of an AI system can only be used for justification of beliefs of a human hearer if there is a human to take responsibility for the assertions. Freiman (2024) claims that our current accounts of testimony do not allow us to analyze the outputs

of AIs, including LLMs, and suggests a new account of AI specific testimony. Jinhua He and Chen Yang (2025) develop a similarly novel account of artificial testimony to accommodate AI systems and LLMs. Finally, Freiman (2023) offers technology-based beliefs as the correct category for beliefs formed on the basis of AI "testimony."

Chapter 2 – Can Language Models Be Credible Experts?

1. Introduction

Testimony is a central notion for social epistemology. While it is clear that we *can* garner knowledge individualistically, it is equally clear that we get a lot of our knowledge from others. Testimony is best, however, when it is trustworthy. While we often trust the testimony of other persons, it is unclear that we can (or should) trust the testimony of Large Language Models (LLMs).[8] Two questions immediately arise: (i) is LLM testimony trustworthy enough to produce knowledge and (ii) if so, when can we trust it?

Imagine the following case. You are a student that is using ChatGPT to help you write an important term paper. Rather than doing the research yourself, you prompt ChatGPT to tell you what philosopher Anti-Reductionism about testimony is often traced back to in the western tradition. ChatGPT (correctly) tells you that this view is often traced back to Thomas Reid. Do you now know this proposition or do you need more evidence?

It would seem that the initial, pre-theoretical intuition is that you indeed need more evidence. This evidence could be in the form of another source you already trust telling you the same fact, or it could be evidence that ChatGPT is being trustworthy in this instance. To be sure, it can still be the case that ChatGPT is being 100% reliable and accurate, but it seems like we have an intuitive desire to

[8] While there is some literature in regards to special accounts of testimony of LLMs (He and Yang (2025); Freiman (2023, 2024)), I consider the testimony of these models here to just be their propositional outputs. There are also authors who have questioned if AI can be properly considered a testifier, such as Fricker (2025) and Masaharu Mizumoto et al. (2025). I take a face value approach that says that AI and LLMs specifically can be testifiers in their own right, in virtue of the fact that they are often treated as such.

make *sure* that it is being reliable. In this way, LLMs are often treated like the clairvoyant presented in Lawrence Bonjour (1980). That is to say, LLMs are often looked at as "not to be trusted merely in virtue of having told me this."[9] Moreover, in some (but certainly not all) domains, such as chemistry, certain LLMs will outperform human experts in terms of reliably and accurately answering questions (Mirza et al., 2024). So, we have good reason to believe that LLMs can be reliable sometimes, but it is unclear when we should trust them and when we should worry about gathering further evidence. This chapter is broken up into four sections.

In section one, I investigate under what conditions, if any, can an LLM be deemed an expert on a given topic or domain. Next, we will look at criteria for expert credibility and see what the best-case scenario is for an LLM. In doing so, I am to answer two questions: (i) can an LLM be deemed an expert or to have expertise on a given topic and (ii) how much credibility can an LLM expert have? This distinction is important, as being an expert makes one a trustworthy source of information within a domain, but being trustworthy does not mean that laypeople will *know* that you are trustworthy and give you the correct amount of credibility. If LLMs can be experts but never have any credibility, not trusting them seems reasonable. On the other hand, if they can be experts and have a lot of credibility, more work must be done to explain not trusting them.

In section two, I turn to investigate under what criteria can an LLM be deemed a trustworthy standard testifier. Here, I understand the term "standard testifier" to refer to a non-expert testifier under normal conditions. I briefly outline three approaches to testimonial justification and credibility

[9] Thank you to Sandy Goldberg for suggesting this phrasing.

seen in chapter one (Reductionism, Anti-Reductionism, and Dualism) and investigate if LLMs could be deemed credible testifiers on any of these accounts. In section three, I offer possible explanations for the findings in sections one and two. Finally, in section four, I consider objections to my arguments.

2. Expert LLM Testifiers

While we may often query LLMs for simple pieces of information, we can also use these models as a potential stand-in for asking a human expert. For instance, rather than paying to go to the doctor, individuals may prompt an LLM with their symptoms and request some kind of home remedy to alleviate them.[10] The question, then, is whether or not we *ought* to treat an LLM like an expert and, if the answer is no, why not? In what follows, I use Goldman's (2001) and Grundmann's (2025) definitions of "expert" to determine in what sense, if any, we can call an LLM an expert on a given topic. I conclude that Goldman's treatment of the term allows for LLM experts but is nevertheless quite quick. As a result, I turn towards Grundmann's for a fuller treatment of expertise. Still, I find that Grundmann's account of expertise allows for LLMs to justifiably be considered experts. Importantly, not all LLMs will be considered experts, and it is likely not the case that an LLM can be globally deemed an expert on every domain. I want to claim, then, that LLMs *could* count as experts under the definitions presented here. A user will still have to check each model they use for expertise in any given domain, as expertise is not shared between models or domains. So, if a user correctly deems an LLM an expert on organic chemistry, for instance, that does not allow the user to justifiably treat the LLM as an expert on algebraic combinatorics. Similarly, just because one model is an expert in

[10] This is importantly different from merely "googling" your symptoms, as this would often result in one searching through websites that are filled with true expert testimony (e.g., Mayo Clinic's symptom checker).

domain *A*, does not mean another will be. Despite this work on the part of the user, I show below that LLMs can be, in principle, experts.

2.1 Goldman's Expert

In "Experts: Which Ones Should You Trust?," Alvin Goldman (2001) considers the problem of expert disagreement.[11] Specifically, Goldman considers the question of novices assessing the credibility of two rival experts without having to become experts themselves (Goldman, 89).[12] The tension Goldman is considering here comes from the idea that a true novice about expert domain *E* will be unable to use any of her own beliefs because she either has no beliefs in *E* or has no confidence in her beliefs in *E* (ibid., 90). For instance, if I am speaking to two friends about the weather tomorrow and my friends make contradictory claims, choosing between the two would be a simpler matter. While I may not be a meteorologist, neither are my two friends. Thus, in a case like this, I (a novice) would only have to adjudicate between two other novices. On the other hand, if I find myself in conversation with two meteorologists at a conference for meteorologists, choosing between who has more credibility on the subject of weather forecasting would be quite difficult, as I have no confidence in my opinions (or perhaps have no opinions in the first place) about the science of weather forecasting. Thus, a novice needs some kind of non-*E* evidence to adjudicate between disagreeing experts. Moreover, the specific question is whether or not a novice can *justifiably* choose one expert as more trustworthy than another (ibid., 92)

[11] Though perhaps the most notable, Goldman is not the only philosopher to consider this problem. See: Anderson (2011), Almassi (2012), Guerrero (2019), Bennett (2022), Boyd (2022).

[12] While Goldman is answering a comparative question (i.e., which of these two experts should the layperson trust), I use his criteria to answer a non-comparative question about expert trustworthiness (i.e., under what circumstances can I trust this particular expert).

Within this problem there are some important clarifying points to note: (i) what is an expert and (ii) what kinds of statements are available to the expert and novice? To the first point, Goldman gives a clear definition of what it means to be an expert on some domain of knowledge: an expert in domain D is someone who "possesses an extensive fund of knowledge (true belief) and a set of skills or methods for apt and successful deployment of this knowledge to new questions in the domain" (ibid.).[13] To the second point, a statement can either be *esoteric* or *exoteric* (Goldman, 94). A statement is esoteric if it falls within the relevant sphere of expertise and thus its truth value is inaccessible to a novice, whereas a statement is exoteric if it falls outside this sphere and thus has a truth value that is accessible to a novice. Moreover, a statement can either be *semantically* or *epistemically* esoteric (Goldman, fn. 10). A statement is semantically esoteric if the novice cannot assess the truth value because she does not semantically understand the statement, whereas a statement is epistemically esoteric if the novice cannot assess the truth value despite semantically understanding the statement.

While Goldman (2001)'s paper is influential, his treatment of the notion of an expert is perhaps too quick. Recall that, for Goldman, an expert is someone who possesses a large amount of knowledge within a particular domain and has the skills or methods to deploy this knowledge to new questions within this domain. Goldman does not spend much time describing what this deployment entails. What time he does spend on it, though, seems to favor a *prima facie* reading of LLMs as potential experts. Goldman says that an expert will have the proper "know-how, when presented with a new question in the domain, to go to the right sectors of his information-bank and perform

[13] While this distinction need not weigh us down, Goldman considers a distinction between a strong and weak sense of expert, with the definition provided here being the strong sense of the term.

appropriate operations on this information" (2001, 91). An LLM can deploy its knowledge[14] (or, if you like, training data) and deploy an answer to a novel question in a given domain. While it certainly does not grasp or understand the answer in the same way a human would, it can clearly deploy its training data to answer new or old questions in new ways.[15] On Goldman's (2001) account, then, an LLM can be justifiably deemed an expert.

2.2 Grundmann's Expert

Since Goldman's (2001) treatment of "expert" is quite quick, it would behoove us to investigate a more thorough account of the notion. Grundmann (2025) provides such an account. Before giving his account, however, Grundmann samples previously offered functional accounts in the literature that, in his eyes, will not accurately capture the essence of an expert. The first of these is the *novice-oriented account*. According to this account, being an expert amounts to being able to help laypeople by transmitting true beliefs or knowledge (Grundmann 2025, 86). Experts may lack didactic abilities, however, and still be experts. Next, he considers the *research-oriented account* of expertise, according to which being an expert amounts to being able to help the discipline of science make progress (ibid., 87). This account, too, will not suffice, as experts can have expertise in non-scientific

[14] For the purposes of this thesis, I use talk of LLMs having knowledge despite this perhaps not being immediately obvious. Some, such as Ocean Cangelosi (2024), have argued that AI can indeed possess knowledge, since the epistemic notion need not require phenomenological experience. Nevertheless, I think that even if we claim that the use of "knowledge" is mere fictionalism and their "knowledge" is really just rotely memorized outputs (which it very well might be) that should not be called knowledge in the traditional sense, this will not stop Goldman's criteria from being too weak. This is because the emphasis of the definition lies within the deployment of the knowledge or information. Even for humans, a BonJour (1980) style clairvoyant with loose justification for their true beliefs probably could be deemed an expert on this definition given the proper skills for deployment. This is all a very active area of debate, however, with some, such as Camila Hernandez Flowerman (forthcoming), arguing that LLMs are incapable of beliefs and thus incapable of epistemic goods such as knowledge.

[15] LLMs being unable to grasp or understand is treated in full in the third chapter.

domains and an individual can help make progress in a domain they are not an expert in. For instance, mathematicians have helped make progress in many domains besides math, such as biology or psychology (ibid.).

To counter these accounts, Grundmann offers a number of common platitudes about experts and attempts to formulate a definition that matches them the best. We will not only be evaluating LLMs against these common platitudes but also against the finalized definition offered. Grundmann, in listing these platitudes, does not offer defenses of them. Rather, he claims they are intuitive and acceptable on their face (ibid., 87). In what follows, I list these eight total platitudes and consider if LLMs could, under the right circumstances, be attributed with them:

(1) *Epistemic superiority*: an expert must have cognitive competence that is superior to laypeople's competences with respect to a given domain (Grundmann 2025, 87). Expertise is at least partially comparative, says Grundmann, and thus an expert must have some level of epistemic superiority over laypeople. While this may have odd effects (e.g., not everyone can be an expert), we can accept it at face value. LLMs can surely pass this platitude, as they will have access to a great deal of information within a given domain that a layperson would not.

(2) *Reliability is Not Enough:* an expert must not only be reliable but have many beliefs within the given domain (ibid.). Otherwise, an expert could be someone who always reserves judgment until they are certain they will be correct. Moreover, you might be reliable because you know little to nothing within this domain, and thus you very rarely make claims within it. LLMs, again, can surely be attributed with not only reliability, but also sufficient beliefs. While an LLM will often not make strong or clear recommendations to users, it will take stands on facts

within a domain. There might be room to push back here, as LLM might be more credibly attributed with a belief when there is a consensus around a fact (e.g., George Washington was the first U.S. President), and thus an LLM may only ever have beliefs about the consensus. While this would surely make them an unusual expert, having commonly held beliefs does not preclude expertise.

(3) *Epistemic Authority*: "expertise in D grounds the epistemic authority of the expert's judgments about D in relation to laypeople" (Grundmann 2025, 87). Treating someone as an expert requires complete deferral to their judgments within their domain of expertise, no matter what the layperson's other p-related reasons are. This platitude, again, does not seem immediately intuitive, as it seems I could believe that Dr. Johnson is an expert but nonetheless not defer to his judgment on matters concerning my health because I believe that he has different values than I do (e.g., undergoing treatment would change my quality of life in a way that matters more to me than it would if Dr. Johnson was in my position). Regardless, if we accept this platitude, there are surely still cases in which LLMs could be deemed experts in this sense. Certainly we can defer to the judgments of an LLM, but we can also *justifiably* defer to the judgments of an LLM. If I query an LLM with a list of symptoms that I am experiencing and it judges that I have a common flu, my subsequent deference to its judgment and belief that I have a common cold will be justified in just the same way as if I had googled my symptoms.

(4) *Information is Not Enough*: being an expert requires more than reading or being told a lot of truths, rather, an expert needs to be able to form true beliefs about their domain by reasoning on the basis of domain related reasons that are given to her (ibid., 88). LLMs again seem to pass

this test. For instance, while an LLM will know that lead melts at roughly 327.5 degrees Celsius, if you prompt it to assume that lead actually melts at 100 degrees Celsius and deduce other metallurgic truths, it will be able to reply with a list of deductions and implications. Thus, an LLM can do the reasoning required, albeit in a counterfactual sense, to be attributed with this platitude.

(5) *Relative cum Absolute*: "being an expert is time-relative, but it also requires that some absolute condition is satisfied" (ibid.). Grundmann uses an analogy of Ptolemy to explicate this platitude. Ptolemy, in his day, was an expert. If the actual Ptolemy was transported to the present day, he would not be an expert. However, if he lived in our time he would potentially become an expert. In contrast, Franz Gall, the phrenologist, was never an expert, because you cannot competently assess personality on mere skull shape (Grundmann 2025, 88). LLMs likely pass this platitude as well, as they can be compared favorably to current experts. For instance, Guo et al. (2023) and Mirza et al. (2024) show that, in certain problems in chemistry, LLMs outperform human chemists.

(6) *Understanding is Not Necessary*: experts do need to, but perhaps often do, have understanding of their particular domain of expertise (Grundmann 2025, 88). Grundmann defends this platitude by appealing to the Indian mathematician Ramanujan (1887-1920) who supposedly could grasp the correct solutions to math problems without any proofs of the solutions. Thus, it is at least possible for Ramanujan to be an expert in a certain domain of mathematics without having any proper understanding (ibid.). We might also draw an analogy to chicken sexers, who would likely be experts in the domain of "chicken-sexing" (i.e., reliably

determining if a chick is male or female and sorting them accordingly) but nevertheless fail to have any understanding of chicken anatomy or biology. Elsewhere in the paper, Grundmann notes that experts will typically have understanding, but it is not necessary to the nature of being an expert that one have understanding. This, of course, bodes well for LLMs, who likely are unable to grasp or understand in principle. This claim is defended in full in chapter three.

(7) *Being a Successful Teacher*: experts will typically be able to inform and teach laypeople about their domain successfully (ibid., 89). Insofar as belief on the basis of LLM testimony is justified under the right circumstances, LLMs clear this hurdle easily, as an LLM can bring a layperson from a starting position of ignorance to an ending position of genuine knowledge on a topic. Within most topics, especially those heavily represented within their training data, this criterion is straightforward to attribute to LLMs.

(8) *Making Scientific Progress*: experts will typically contribute to scientific progress in their domain of expertise (ibid.). Importantly, Grundmann uses the word "typically" here. Rather than making this a requirement as in the case of epistemic superiority, for instance, Grundmann offers this (in addition to understanding and teaching ability) as a typical, but not necessary, feature of expertise. This is the one place where an LLM may falter. They can certainly be used as tools that help advance the progress, but they themselves do not have the agency to, without prompting, work on current problems within a domain.

Based off of these platitudes, Grundmann formulates his preferred definition as follows: "E is an expert concerning domain D at time t if and only if (i) E possesses more evidence pertaining to propositions in D than the majority at t & (ii) E has skills to reason on the basis of this evidence that

are superior to (more conditionally reliable than) those of the majority at t & (iii) E is sufficiently competent concerning D on the basis of (i) and (ii)" (Grundmann 2025, 93). This definition, for Grundmann, encapsulates the first five platitudes, while leaving room for the typical but not necessary attributes of understanding, didactic skills, and contribution to scientific progress (ibid.). As we saw, an LLM can be attributed with at least the first seven of these platitudes, with the eighth being perhaps a bridge too far. Similarly, an LLM can possess more evidence for propositions within a domain than the majority, it can have the skills to reason on the basis of this evidence, and it can be sufficiently competent with respect to domain D on the basis of these first two criteria. We can safely say, then, that LLMs, under the right circumstances, can be deemed experts.

2.3 LLMs as Credible Experts

While LLMs can be experts, it is still up to a layperson to identify that they are an expert. After all, just because someone has a great amount of knowledge or understanding within a domain and the necessary skills to deploy it, does not mean that a layperson would be able to recognize this. It is at this point that we should distinguish between three terms that are often used interchangeably in the context of expertise: reliability, trustworthiness, and credibility. Reliability, as it is used in this thesis and often within the literature, refers to merely being a reliable reporter of truths. As long as a testifier is consistently telling the truth, they are reliable. Notice, however, that this does not mean that they are trustworthy or credible, as all of these terms can come apart. Trustworthiness refers to the intellectual character of a testifier. Hardwig (1991) defines trustworthiness in a testifier as the culmination of (a) honesty, (b) competency in a domain, (c) conscientiousness, and (d) capability of epistemic self-assessment (700). Trustworthiness, then, is about the process by which you come to your beliefs,

irrespective of their truth value. Credibility, finally, refers to an outsider's perception of a testifier's trustworthiness (Rolin 2002, 96). Thus, just because someone is trustworthy and reliable does not mean they are automatically credible, as an outsider may be unable to perceive these traits. The criteria that we will be looking at in this section are all criteria for credibility, such that if an expert is attributed with at least one of these criteria they are minimally credible, with a theoretical maximally credible expert being attributed with all of the criteria to a maximal extent.

Finally, before we move to the actual criteria, we should consider a crucial question: does an expert need to meet all criteria to be deemed credible, or can they meet only a few? Typically, we would want to give this question a comparative answer, such as, "Well, it would make them more credible than an expert who met fewer criteria." However, we are attempting to look at LLMs in a non-comparative sense, and thus our answer should follow suit. To this question, I think we ought to answer that as long as the LLM is deemed an expert and has at least one of the above criteria, we can safely say they are at least a minimally credible expert. Perhaps not as credible as they could be, but credible nonetheless. However, if they are attributed with none of the criteria, they should not be deemed credible.

Below is a list of eleven criteria for credible expertise from Goldman (2001), Carlo Martini (2020), and Grundmann (2025). In what follows, I will outline each criterion and its supposed ability to illuminate expertise for laypeople. Afterwards, we will assess whether or not LLMs could be attributed with any of these criteria. The criteria are as follows:

(A) Dialectical Competence

(B) Proportionality

(C) Meta-Approval

(D) Evidence from Interests and Biases

(E) Track Record

(F) Teaching Competence

(G) Scientific Selection

(H) Pertinence

(I) Meta-Knowledge

(J) Consistency

(K) Discrimination Ability

The first of these criteria, (A) dialectical competence, is seen in both Goldman (2001) and Grundmann (2025). In essence, the criterion refers to an expert's ability to not only present their own view efficiently but also offer responses to objections others may give to their own view (Goldman 2001, 95). As Goldman claims, there are two distinct ways in which a hearer may come to gain justification on the basis of an expert's testimony: directly and indirectly. Direct justification refers to a hearer becoming justified in their belief for an argument's conclusion as a result of their becoming justified in an argument's premises and their support for the conclusion (ibid., 94). This kind of justification is plausible but, as Goldman notes, the expert's statements in defense of their view will often be esoteric (i.e., have truth values that a novice hearer could not themselves assess) (ibid.). After all, if a hearer can themselves assess the truth value of esoteric statements, they would cease to be a novice and thus be an expert in their own right. Thus, it is usually easier for a novice hearer to gain indirect justification, which comes from a speaker demonstrating *dialectical competence* (ibid.). Here,

Goldman has in mind not pure debating skills, but rather an ability to offer responses to objections to one's view (Goldman 2001, 95). For instance, imagine two rival experts: John and Paul. After John gives an argument in favor of his view, Paul offers a plausible objection that John is unable to offer a rebuttal to. On the other hand, if John offers an objection to Paul's view that Paul (seemingly) has no problem rebutting, Paul seems to have shown dialectical competence while John has not. While a novice hearer cannot access the truth values of all statements offered, such ability offers an indicator that Paul has a greater degree of expertise on the matter at hand (ibid.).

Using this criterion to assess LLMs as singularly trustworthy experts will likely not be fruitful. An LLM will always have incredibly fast processing and answering speed, such that it will always seem to be in a dialectically superior position with any potential interlocutor. Despite this, it does not mean that it is trustworthy within a given domain or in a given instance. Simply because the model responds quickly does not mean that its output will be correct. Moreover, an LLM's output is built on large amounts of training data such that it will likely be able to reliably produce true *sounding* outputs but may not always produce *true* outputs. We could imagine, for instance, a purported expert who always replies quickly to any objection levied at their view, but their reply is always false. Giving a response seems to show dialectical competence, but it will not confer expertise (seemingly) if the answer is always false. This is not to say that an LLM will always reply falsely. Rather, the problem here is that an LLM will always have an ability to retrieve an answer from their training data. This gives a model the superhuman ability to respond to any objection or challenge given. While many commercial models have strict guardrails that require them to assert that they are in no position to respond (e.g., a user asking a model to write them malicious software), a model will be able to always respond to any

objection within almost any topic. On the other hand, a human expert would likely not offer a response if they could not formulate one that they felt they could stand behind, and thus when they do offer quick and effective responses, it likely means they are a credible expert. As a result, an LLM does not fail, per se, on the dialectical competence criterion, rather, the criterion itself does not reliably elucidate expertise in the case of an LLM.

The second criterion for credible expertise, (B) proportionality, is seen in Goldman (2001), Martini (2020), and Grundmann (2025). Proportionality refers to whether or not an expert's opinion has a high level of agreement within their respective field. In the case of choosing between two different experts, it would seem that an expert whose opinion is heavily agreed upon by other experts in her field would, *prima facie*, be more deserving of credibility than one who holds a position by themselves (Goldman 2001, 97). If we assess a single expert and find that their opinions enjoy a level of consensus within their field, it seems to be more likely that this expert has knowledge within their domain.[16] This criterion seems to bode well for LLMs. Since their training data is vast, it is likely that their outputs will be in line with the consensus of a given field, and thus they will almost always be in agreement with any consensus. While an LLM could, in principle, be developed with internal weights such that they always offer the non-consensus opinion, a standard publicly available LLM is almost always going to defer to the consensus within its training data, thus attributing it with the proportionality criterion.

[16] Crucially, this criterion cannot guarantee expertise, as consensus can be non-knowledge based, as Miller (2013) shows.

The third criterion, (C) meta-approval, which shows up in both Goldman (2001) and Grundmann (2025), refers to an expert's approval from other experts within their field. This could manifest in, among other things, a PhD or published work in their field. Of course, someone can be a credible expert within a domain without having published work or a PhD. This meta-approval could come in the form of established experts "vouching" for the potential expert by claiming that they know what they are talking about. For instance, a bodybuilder with a personal history of anabolic steroid use and a keen interest in reading the most updated scientific studies on the effects of anabolic steroids may lack a PhD or published work, but if a board certified endocrinologist reviewed their opinions and approved them as well-founded and/or true, this bodybuilder would be deemed a credible expert under the meta-approval criterion.

Approval of meta-experts, similar to proportionality, seems straightforward to attribute to LLM experts. For instance, an expert within the field of number theory can check the outputs of an LLM within the domain of number theory and approve the outputs as reliable, true, and thorough. They could also, of course, disapprove of a given model's output, but nothing in principle is in the way of an LLM receiving this kind of meta-expert approval. One point to note is that LLMs can only get this meta-expert approval through this checking of outputs, rather than credentials or achievements such as an advanced degree or published research. This shows us that this criterion is actually quite cheap, as we can imagine cases of approved outputs that would not confer expertise. For instance, if Sam tells everyone that he is an expert in microbiology but gets all of his opinions from a magic 8-ball, he could have reliably true outputs that are approved by an actual expert in microbiology,

but this would not make Sam an expert himself. This particular distinction between the value of outputs versus the value of the process that leads to outputs will be explored more in section three.

The fourth criterion, (D) unbiasedness, which is found in both Goldman (2001) and Martini (2020), refers to the commonsense intuition that if an expert has a reason to lie or has a clear bias that could influence their opinion, their credibility is lower (Goldman 2001, 104). Lying is not the only way that credibility can be reduced under this criterion. For instance, if an expert claims, "Cigarettes are actually really good for you," but simultaneously owns a large share in multiple tobacco companies, there would be a clear conflict of interest and thus their testimony would lose some level of credibility. To be fully attributed with this criterion, then, an expert needs to be clear of any potential biases or conflicts of interest that could influence their opinions.

While an LLM cannot fail this criterion by owning a tobacco company, it can still harbor bias and interests. Namely, any biases in the training data can show up in the outputs of a given model.[17] Since most or all of the training data from an LLM will be human-generated, it will contain human biases that will eventually move upstream to the model itself. While many of the criteria presented here may or may not be attributed to a given model depending on a myriad of factors, it is impossible for an LLM to be free of bias – see Philip Resnik (2024) and Charaka Vinayak Kumar et al. (2025). Any notion of a bias-free and perfectly neutral LLM is the stuff of myth. An expert can of course still have these biases, but similar to a human expert, unavoidable bias harms the credibility of a potential LLM expert.

[17] Moreover, any interests on the part of the proprietor of the model can show up as well. For instance, if OpenAI wants their models to push certain messaging, they could change the model's training and weights to favor certain outputs.

The fifth criterion, (E) track record, which is seen in Goldman (2001), Martini (2020), and Grundmann (2025), refers to a purported expert's previous track record in making predictions and claims within their purported domain of expertise. For instance, if an expert has routinely offered correct solutions to problems in their field of expertise, the likelihood that they have offered a correct answer to the current question is considerably higher (Goldman 2001, 106). Moreover, if they have routinely offered correct answers or predictions, the likelihood that they have specialized knowledge within their field is much higher.

Once again, there is nothing standing in the way of an LLM having a positive track record with answering questions within a domain. If a user finds that a particular model has answered many questions within the field of organic chemistry correctly, the likelihood that the current question will be answered correctly is of course much higher. This could come in the form of the user checking the answers themselves (if they are already an expert) or a layperson user having another expert check the outputs for them. Notice, however, that this criterion has a similar problem as meta-approval. Providing consistently correct predictions is an indicator of specialized knowledge or expertise, but we can nevertheless come up with counter-examples where the process is sub-optimal (e.g., magic 8-ball) but the output is consistently favorable.

The sixth criterion, (F) teaching competence, which is found in Grundmann (2025), refers to the ability of a potential expert to teach their expertise, or some subset of it, to a layperson (ibid., 99). Grundmann notes some potential problems with this criterion, as simply because someone is a good teacher of content does not necessarily entail they are an expert in that domain (ibid.). For instance, a regular schoolteacher may be particularly good at teaching algebra, but they may not be an expert in

algebra. On the other hand, we may curtail the domain to a certain grade level of algebra to allow for expertise. In any case, Grundmann notes a possible dilemma for this criteria. If the measure of teaching competence is a layperson starting with ignorance and ending with genuine knowledge, then teaching competence is likely a true measure of expertise. On the other hand, it is doubtful that laypeople can tell whether a view or supposed area of expertise is genuine. For instance, areas such as phrenology or alchemy were teachable despite them not involving any real expertise (ibid.). On the other hand, if the measure is starting at ignorance and ending up with putative knowledge, then laypeople can easily identify this criterion in someone as they could identify who is a successful teacher, but this will not guarantee expertise (Grundmann 2025, 99). Here, I think Grundmann's treatment of the teaching competence criterion is quick and makes an unintuitive assumption. Namely, he assumes that one cannot be an expert or have expertise in an area that makes a lot of false predictions or is otherwise not well-founded. It does not seem immediately obvious that the phrase, "Sam is an expert in astrology," is false or does not obtain simply because astrology is widely believed to be or outright is false. Moreover, it seems plausible that a layperson could see successful teaching, and then check the expert consensus (via other criteria listed here to find these experts) to see if this domain being taught is true expertise.

Moving beyond the potential dilemma Grundmann notes, the teaching competence criterion does seem to allow for LLM expertise. For instance, if a user wants to learn about quantum mechanics, they can prompt an LLM and say, "Teach me about quantum mechanics." Depending on the model and if there is further prompting, the LLM can reply with a list of major developments or principles within the field, such as wave-function or superposition. More recently, models such as OpenAI's GPT-5.2 will end an output with a set of questions to keep the user engaged. So, at the end of a

quantum mechanics teaching-related output, it may ask the user if they want a deep dive into the math, step-by-step beginner's course, or related readings to explore themselves. While the question of whether or not an LLM can replace a human teacher is a complicated question that will not be explored here, if what we are looking for in terms of teaching competence is a layperson starting from ignorance and arriving at knowledge via an expert, it seems that LLMs will succeed.[18]

The seventh criterion, (G) scientific selection, which again comes from Grundmann (2025), refers to members of the scientific community operating under good conditions. Grundmann's claim is that, under these optimal conditions, science will function properly and select members in accordance with six standards, thus making it highly likely that only genuine experts will be a part of the scientific community (ibid., 100). These standards are as follows: only trained or talented people can be members, science is open to diversity, strong and open competition between members, critical thinking and rigor are rewarded, scientific research is guided by the aims of truth and intellectual integrity, and the ruling standard is scientific agreement among independent judges (Grundmann 2025, 100-101). Grundmann provides no further explanation of this criterion and leaves it quite vague as to how "science" selects its members and what constitutes membership. In charity to him, I am going to assume that LLMs are not the kind of thing he imagines could be members of the scientific community.

The final four criteria, (H) - (K), all can be seen in Martini (2020, 118-119). The first of these, (H) pertinence, is seen in experts who give judgments within their field of competence and avoid

[18] There may be room to push against Grundmann's construction of this criterion further, as it may be better for a layperson to come to understanding, for instance, rather than knowledge. However, insofar as we grant that understanding is not necessary for understanding, it may be wrong to construe this criterion in terms of understanding.

giving judgments outside their field without qualification (ibid.). While this criterion surely does not always confer expertise, as an expert in domain D can still falsely believe they have expertise in another domain as well and give unqualified judgments, it nevertheless stands to reason that an expert would practice epistemic caution and qualify statements made outside of their domain of expertise. For instance, while Joe and Steve are both doctors, Joe is an oncologist while Steve is a neurologist. Since they are both doctors who went through medical school, they likely have some background knowledge within the other's domain. Nevertheless, the kind of epistemic caution we would expect an expert to practice would be akin to Joe making a claim about neurology and making it clear that he is not a neurologist and thus you should take his view with a grain of salt, for instance. An expert who will qualify statements made outside of their domain may be more credible within their own domain, as they know what they do not know, as it were. This is closely related to the ninth criterion, (I) meta-knowledge, which is seen in experts who know what they do not know and manage their uncertainty accordingly (Martini 2020, 119). While pertinence refers to claims made outside of the experts domains, meta-knowledge refers to claims made within the expert's domain. For instance, if an expert in biochemistry could not recall the seventh step in the Krebs cycle, they would seem more credible if they were to qualify this for layperson listeners. LLMs seem to be able to pass both of these criteria, as they are able to qualify when an output has minimal training data or lies directly outside of their training data. OpenAI's GPT-5.2, for instance, is aware that its training data only goes up to August 31 of 2025, and thus any questions about more recent events require it to use a web search tool. When you ask it a question about something that happened after this cut-off date, it will not only flash that it

is searching the web before answering, but it will also give you a link to where it found the answer in its output.

The tenth criterion, (J) consistency, is seen in experts who give internally consistent judgments (ibid.). If a purported expert's judgments are all inconsistent with one another, it stands to reason that they do not have expertise or specialized knowledge within their domain. This is different, of course, from an expert changing their mind or publicly retracting a previously made claim. If, for instance, an alleged-expert claims that the caloric model of heat is correct and the kinetic model of heat is correct, their judgments would be internally inconsistent. An alleged expert who used to believe in the caloric model but retracted their statement and now believes in the kinetic model would remain internally consistent. Similarly, if an expert consistently advocated for one model over another (irrespective of the truth value of the model) they would be internally consistent.

An LLM can be internally consistent, especially if used properly. Most models will put the existing conversation history onto the beginning of a new prompt as an invisible block of text. So, any new questions I ask an LLM will continue the entire history of the conversation, making it quite likely that it will be able to remain consistent, or at least flag potential inconsistencies. More than this, however, a user can control the temperature of the model. Temperature is a hyperparameter that controls the randomness and creativity of the textual outputs. Typically, one can set the temperature from 0.0 all the way to 2.0. With a moderate temperature, such as 1.0, the model will treat the likelihood of the next token at face value. So, consider a chain of tokens that says "I want to go to the" and the remaining options are "zoo," "park," and "store." If "zoo" has a probability of .2, "park" has a probability of .15, and "store" has a probability of .1, a temperature of 1.0 will treat these equally and

likely choose "zoo." If the temperature is turned up to 2.0, the net probability of lower probability options such as "store" will be higher. If the temperature is turned down to 0.0, the model will optimize for high probability answers, and thus will weigh the higher probability of "zoo" more than other variables such as creativity. All this is to say that when the temperature of a model is turned down, it is incredibly consistent in its answers. On the other hand, if the temperature is turned up, it will often give different answers. These different answers may be inconsistent, but often they are merely different ways of saying the same thing. Thus, LLMs are more than capable of being internally consistent.

Finally, the eleventh criterion, (K) discrimination ability, refers to experts who are able to discern the difference between two closely related by distinct cases within their domain (Martini 2020, 119). For instance, an oncologist will likely be able to tell the difference between two very similar cases of leukemia, where one patient's cancer was more resistant to treatment whereas the cancer of the second patient responded very quickly. LLMs are surely capable of discerning and analyzing the small differences between two presented cases, as they will be able to compare them to the myriad of other cases in their training data and look for patterns. LLMs have shown great promise in medical contexts, particularly in diagnostic reasoning, which often requires the individual to differentiate between cases that may present quite similarly but have different causes – see Xiangbin Meng et al. (2024), Felix Bojesmo et al. (2024), and Alabi Busch et al. (2025).

With all of these criteria laid before us, we can now look at the best-case scenario for an LLM expert. Assuming a particular model is deemed an expert within domain D, the best-case scenario is that it has all of the following criteria for credibility:

(B) Proportionality

(C) Meta-Approval

(E) Track Record

(F) Teaching Competence

(H) Pertinence

(I) Meta-Knowledge

(J) Consistency

(K) Discrimination Ability

Out of a total of eleven possible criteria presented in this section, an LLM can, in principle, be attributed with eight. While an LLM may not be maximally credible, as some criteria such as dialectical competence are not sufficient for analyzing its trustworthiness, it is clear at this point that an LLM can be a credible expert testifier. If we are not immediately convinced, we can imagine a scenario where a human expert has all of the eight criteria listed above. In such a case, it would be wrong of us to deem this person's expertise not credible. Similarly, if we are presented with an LLM expert and want to give it less credibility than a human with the same attributed criteria, a sufficient explanation is needed.[19]

3. LLMs as Standard Testifiers

While we might treat LLM testimony as that of an expert under certain circumstances, LLMs can also be used as standard testifiers. That is, producers of testimony that do not have expertise within a domain, but who may nevertheless be trustworthy. In what follows, I investigate if LLMs can be

[19] One could argue that this is an artifact of setting up the criteria such that one either meets or fails to meet a given criterion. Nevertheless, even if a model only minimally meets these criteria, they can still be deemed to have a degree of expertise that is higher than the average layperson.

considered a trustworthy standard testifier under the three main views within the testimonial justification debate: Reductionism, Anti-Reductionism, and Dualism. While many views within each of these camps are slightly different from one another, the arguments within each section come from papers that are representative of each overarching approach to testimonial justification. We will start with Fricker's (1994) arguments in favor of Reductionism, then move to Insole's (2000) arguments in favor of Anti-Reductionism, and finally end with Lackey's (2008) arguments in favor of Dualism.

3.1 Reductionism

As we saw in chapter one, Fricker (1994) claims that "a hearer should always engage in some assessment of the speaker for trustworthiness" and to do otherwise, i.e., to presume trustworthiness, is tantamount to "gullibility" (145). Fricker defines the trustworthiness of a speaker in terms of their sincerity and competence, such that if, on a given occasion, the speaker's utterance is sincere and the speaker is competent with respect to the proposition, the speaker is trustworthy (Fricker 1994, 147). Assessing this sincerity and competence comes in the form of assessing the utterance and its testifier respectively. Sincerity, for Fricker, amounts to believing what you knowingly assert (1994, 145). A sincere instance of testimony, therefore, need not be true. For instance, if someone is under the influence of hallucinogenics, they can sincerely assert that there is a pink elephant in the corner of the room, despite this being obviously false. Competence, for Fricker, is defined as follows: if a subject were sincerely to assert that *P*, then it would be the case that *P* (1994, 147). Competence, then, refers to whether or not a testifier has a reliable disposition to form true beliefs in the relevant domain. This disposition ensures that the speaker yields accurate testimony across a range of similar circumstances and not just the current one by accident. There is, however, a pressing issue that LLMs will run into

with Fricker's account: even if they are sincere and competent testifiers, it is not clear that they can be credible testifiers.[20] [21]

Recall that, in the last section, we distinguished between trustworthiness and credibility. Trustworthiness refers to the intellectual character of a testifier, whereas credibility refers to an outsider's perception of a testifier's trustworthiness (Rolin 2002, 96). Thus, just because someone is trustworthy and reliable does not mean they are automatically credible, as an outsider may be unable to perceive these traits.[22] If we assume that LLMs can be sincere and competent testifiers, which may not be the case, we still will not be able to assess them as credible standard testifiers. This, again, is an important distinction to make, as it can explain why someone would not trust an LLM even if it was trustworthy. For Fricker (1994), trustworthiness just amounts to sincerity and competence, but a testifier can be sincere and competent without me knowing about it. Thus, there needs to be some way to deem a testifier trustworthy within a given instance from the outside and not from a "god's eye view," as it were. Fricker thinks the best way to do this is by constructing a theory to explain the

[20] Notably, Fricker herself has argued that AI systems more broadly are not the kind of thing that can be testifiers in the first place (Fricker 2025). I put this to the side in this thesis, as LLMs specifically are often used as if they are testifiers and they do produce testimony, and thus it seems the best way to analyze potential trustworthiness and credibility of these models is with the arguments presented in the testimony literature.

[21] Some would argue, as well, that LLMs cannot be sincere. We see that sincerity requires that the testifer believe that the proposition they testify to is true, which is an easy bar to clear for any human testifier, but gets unintuitive to apply to LLMs. Some, such as Cangelosi (2024), have argued that beliefs do not require that one has phenomenological experience of consciousness and that since this is the only hurdle standing in the way of AI systems broadly having beliefs, AI can be correctly attributed with certain kinds of beliefs. In other words, even though there is nothing it is like to be an LLM, an LLM can still have beliefs. Flowerman (forthcoming) considers a number of views that would claim that LLMs cannot have attitudes such as beliefs or intentions. On the other hand, Putnam (1960) famously argued in favor of the multiple realizability of beliefs and a functionalist theory of belief, such that machines can in principle have beliefs.

[22] Moreover, credibility is a person-relative notion and much of what plays into credibility is the priors of the attributer. For our purposes, however, we can assume that the attributer of credibility to a given model is as close to having truly neutral priors as possible. They may be slightly in favor or slightly against LLMs, but are on the whole open-minded towards these models.

utterance of the speaker (ibid., 149). The explanation should "render it comprehensible why she made that assertion, on that occasion" and the trustworthiness will follow from a proper explanation (ibid.). Fricker briefly motivates this by reference to the feeling a hearer may experience when receiving a particularly unbelievable instance of testimony. If my friend testifies to having seen flying saucers, I may naturally try to explain why she is saying this. If the best explanation is sincerity and competence (and therefore that the claim is likely true), then we can trust this person, but if the best explanation is that this person did not actually see flying saucers and they are mistaken, lying, etc., then we cannot trust this person's testimony. LLMs can be hard to give an explanation for because the usual considerations we would make would not or are difficult to apply to them. For instance, I might consider if my friend has a reason to lie to me when evaluating her sincerity, but LLMs seem to neither have reasons for or against lying to me. Similarly, when talking about this vetting process, Fricker uses examples that indicate part of what she has in mind is tone and body language, such as "I didn't like the look of him" and "Well, she seemed perfectly normal" (1994, 150). Moreover, Patricia Hill Collins (2000) claims that what is often used in cases of assessing sincerity is the "moral and ethical connections" to the ideas espoused by a testifier, such that a testifier that seems to be personally invested in their utterance is more sincere than one who is not (265). This, still, is plainly not available to LLMs. Finally, LLMs are programmed to always provide a response, such that even if they are providing partially or fully false information, they still reply with an answer. Thus, it can be difficult to tell if an LLM is responding because the answer it is giving is what it believes or because it has to respond. In the case of humans, however, we may be able to recall instances where an individual has withheld assertions because they did not have knowledge or did not have a belief about the matter, and

thus did not want to assert anything. This lends them more credibility, theoretically, when they do assert a proposition to be true. All of this shows us that even if LLMs could be sincere and competent (and therefore trustworthy) testifiers on a Reductionist account like Fricker's (1994), they cannot be credible testifiers.

3.2 Anti-Reductionism

Perhaps, if having to assess the credibility of an LLM poses a problem, we can take an approach that allows us to take a default position of credibility. Recall that Anti-Reductionists hold that all that is required for a hearer to be justified in basing their belief on the basis of a speaker's testimony is a lack of epistemic defeaters (or, reasons for thinking that the assertion is false or the person is not credible). In chapter one, we took a closer look at Insole's (2000) Anti-Reductionist argument. Similar to other Anti-Reductionists, Insole starts with a transcendental argument for Anti-Reductionism. Since it is undeniable that we gain knowledge on the basis of testimony, it is either the case that it is "possible generally for the hearer to obtain independent confirmation that an assertion by a given speaker is trustworthy," or the hearer has the "epistemic right" to believe the testimony in virtue of it being asserted (Insole 2000, 46). However, since it is not possible, generally, for the hearer to obtain this independent confirmation, the hearer must have the aforementioned epistemic right (ibid.). Insole's main goal, then, is to show that Reductionism, or this independent confirmation, fails. Thus, Anti-Reductionism is the only way to account for testimonial knowledge.

Now, obviously, Anti-Reductionism posits no criteria by which we could assess a speaker's trustworthiness; it is merely assumed with the absence of defeaters. This has led many Reductionists to press Anti-Reductionism with the problem of gullibility. If we adapt a default trust to all testimony,

we may thereby leave ourselves open to being deceived. If we are constantly believing on the basis of testimony that we have not "checked" and have instead assumed trustworthy we may find ourselves with many false beliefs, which is an "unsafe policy for belief-formation" (Fricker 2006, 620). One of the ways in which Anti-Reductionists get around objections about gullibility is to add a "monitoring condition" (Goldberg and Henderson 2006). The idea is that hearers not only need to not have undefeated defeaters but also need to be *on the lookout* for such defeaters. Monitoring for defeaters amounts to a kind of epistemic protection[23] insofar as we can seemingly tell when a speaker's testimony is "off," in some sense. When I converse with a human being I have more information available to me than I do with an LLM, such as body language or tone of voice. If I start to get the feeling that someone may not be reliable for some reason, I will start to be on the lookout for epistemic defeaters. On the other hand, I cannot read an LLM's body language or tone of voice, so some of these modes are unavailable to me. If the LLM starts outputting completely bizarre information, I can easily treat that as a defeater, but I am unable to be signaled by my internal "epistemic protection," as it were, that I need to be on the lookout for defeaters. Perhaps more pertinently, thinking about the LLM's belief formation process does not function in the same way as a tool for monitoring. In the case of a human, I can consider their assertion and, because we are both humans and will likely have very similar belief formation processes, assess if I would have come to the same or a similar belief. If I would not have come to the same conclusion, I can start to consider them as a testifier and their testimony more closely for stronger defeaters. On the other hand, if I consider how an LLM came to produce their

[23] Thank you to Sandy Goldberg for the suggestion of this phrasing.

assertion/output, this might always prove to be a defeater, as the process of probabilistically choosing a string of tokens is not the kind of process I may want to place my trust in. Once again, even if an LLM can produce reliably true outputs, it seems that we will always have defeaters to its credibility under an Anti-Reductionist approach.

3.3 Dualism

Finally, let us now turn to Dualism. The most prominent modern defender of this view of testimony is Jennifer Lackey, with her most well treated defense of it coming in Lackey (2008). Dualism, as the name suggests, is a hybrid view about testimonial justification that combines aspects of both Reductionism and Anti-Reductionism. Specifically, Lackey (2008) claims that in order for hearer B to know that *p* on the basis of speaker A's testimony it must be the case that:

> (i) B believes that *p* on the basis of the content
> (ii) A's testimony is reliable or otherwise truth-conducive
> (iii) B is a reliable or properly functioning recipient of testimony
> (iv) the environment in which B receives A's testimony is suitable for the reception of reliable testimony
> (v) B has no undefeated defeaters for A's testimony and
> (vi) B has appropriate positive reasons for accepting A's testimony (177-178)

The two most obvious aspects that Lackey has used from Reductionism and Anti-Reductionism are of course (v) and (vi). Given what has already been said about LLMs and instances of their testimony in this chapter, I am going to assume that we can allow for (i) - (iii) to be true in instances of LLM testimonial exchange. The sixth requirement, that B has appropriate positive reasons for accepting A's testimony, will not be a problem for LLMs, as most of the positive reasons that Lackey has in mind boil down to the reliability of speaker A, the kind of report A is giving, or the kind of context that the testimonial exchange is

happening in (Lackey 2008, 182-183). While these may not always be the case for LLMs, a user could, for instance, have inductively justified reasons to believe that one LLM is reliable with regards to certain kinds of questions in virtue of other LLMs being reliable with regards to those questions in the past. The fifth requirement, that B has no undefeated defeaters for A's testimony, was considered above in the previous section on Anti-Reductionism. Insofar as those arguments hold, then, we need not reevaluate LLMs on this notion. The fourth requirement, however, is where I think LLMs will run into problems.

This fourth requirement is aimed at avoiding any cases where the testimonial exchange is due to mere luck. Lackey imagines a case where a man, Marvin, asks another man, Alfred, what city he is in (2008, 165). Alfred, correctly, tells Marvin that he is in Smithville. Unbeknownst to Marvin, however, Alfred is the only truth-teller in all of Smithville, and had he asked anyone else he would have been lied to (ibid.). Other examples include turning to the only reliable sports broadcast or picking the only reliable book in the library. If you are walking into a minefield of unreliable testifiers or instances of testimony and just so happen to walk out with a true belief, your belief is not justified. This poses a problem for LLMs, as although they can be reliable in principle, they have been shown time and time again to be remarkably unreliable on a wide range of topics – see Neeraj Varshney et al. (2023), Mahmud Omar et al. (2025), Ernests Lavrinovics et al. (2025), Ziwei Xu et al. (2025), and Matthew Dahl et al. (2024). Thus, even if a user does not have any defeaters to the credibility of an LLM (perhaps in virtue of them not knowing how an LLM works), the environment of interfacing with an

LLM via a testimonial exchange prohibits justified beliefs on the basis of their testimony under dualism.

4. What Have We Learned?

As a recap, here's what we have learned about LLMs as potential experts and standard testifiers: (i) they can be very credible experts within a domain and (ii) they cannot be credible standard testifiers. There is a clear tension here. Imagine talking to a human expert of quantum mechanics that you correctly identify as a credible source of information on the domain but you cannot trust them to tell you the truth about the weather outside. While there are surely some examples we can come up with where a credible expert is nevertheless a nefarious trickster who enjoys lying about mundane things, we typically think that if someone is a credible expert, we will be able to trust their testimony such that our beliefs on the basis of it would be justified. In the case of LLMs, on the other hand, we can in principle call them very credible experts, despite their testimony not conferring upon us justified beliefs. In what follows, I offer some possible explanations about what could be causing this tension, what we can learn from it, as well as what it might teach us about trust. I do not claim here that any of these explanations fully describe the scenario at hand. I leave this to future research to fully and adequately explain why there is this disconnect and merely offer these as potential explanations to be explored.

4.1 Credibility Gap

As we saw in the section on expertise, LLMs can be ideally attributed with eight out of eleven criteria for credible expertise. Despite this, there are clearly cases in which LLMs are not

given the credibility these criteria would seem to suggest. In that section, I concluded that if one wants to give LLMs less credibility than is deserved, a sufficient explanation is required. One such sufficient explanation is to point to the lack of credibility in standard testimony cases. Specifically, we might say that the reason that LLMs are often given a lower "credibility score" than they may deserve is that the standard trust that is often assumed in human cases is not appropriate in LLM cases. Recall that if a human is a credible expert, we expect it to be true that this person is also a credible standard testifier. Perhaps what causes the credibility deficit for LLMs is that this standard trust is not assumed or smuggled in, and thus we temper our credibility of LLM experts to a lower score despite them actually scoring higher. For instance, if we had a human expert that was shown to be credible but nevertheless presented us with defeaters, lack of positive reasons, or an unreliable environment for testimonial exchanges, this might cause us to temper our credibility of their expertise as well.

4.2 Motives of Proprietors

A second potential explanation can be found with the proprietors of LLMs. While some may have worries about individual figureheads of companies developing these models (e.g., Elon Musk or Sam Altman), these worries may or may not ring true. What is undeniable, however, is that the truth is not their only concern. It is a fact of the matter that their motive is not truth for truth's sake, rather, their final motivation is financial. Whether we cash this out in terms of profit or sufficiently keeping a non-profit afloat, the fact remains that while LLMs can provide reliable outputs, the fact that these proprietors' ultimate goal is not truth might be a sufficient normative or psychological defeater. When considering the sincerity of a testifier

on Fricker's (1994) account, for instance, we might refrain from attributing them with sincerity if a live option is that they have an ulterior motive besides telling the truth. In the case of a human testifier, if the speaker has no reason to tell the truth over lying, perhaps because their primary goal or intention is not to tell the truth but is instead to serve me advertisements, I ought not trust their testimony. Moreover, because we worry that the motives are not entirely truth-aligned, we might worry about the supposed reliability of these models. If the data is cherry picked, for instance, and a model is fed only trivial truths, its reliability does not confer credibility.

4.3 Reliability and Truth Isn't Enough

A final explanation that has been hinted at throughout this chapter is that reliability and truth are not sufficient for trust. Rather, the process by which someone comes to their reliable truths matters for whether or not we deem you credible. Imagine two science labs that both produce reliably true predictions and accurate research about scientific phenomena. Lab A is doing careful and cautious science, whereas Lab B is letting a magic 8-ball make all decisions. It would seem that despite their outputs being equally reliable and true, we ought to trust Lab A more than Lab B, and might in fact not trust Lab B at all. Notice that the criteria for credible expertise and trust on standard testimony accounts seem to come apart in this regard. The criteria for expertise seem to focus entirely on outputs or results. For instance, the track record criterion only refers to your track record with predictions but not how you came to those predictions, and the teaching competence criterion only refers to if you can bring a novice to knowledge when starting from ignorance. On the other hand, the criteria for credible

testimony on the standard accounts presented makes the process a central part of the evaluation, thus precluding LLMs from being credible testifiers. Much of the trust that we have in standard testimonial exchanges with humans comes from an implicit belief that the speaker's belief was formed in much the same way as we form ours. For instance, if my neighbor tells me that the firework show starts at 5pm, I will usually trust her, as she likely formed this belief in much the same way I form my beliefs. If she was shaking a magic 8-ball, however, I would not be so quick to trust her. Finally, this explanation of reliability not being enough for trust explains why proprietors of LLMs can publish the new record-breaking benchmark scores[24] of their newest model while not changing the minds of many in regards to the trustworthiness of these models. Despite empirical increases in performance in key metrics and reliability, the process by which they come to these outputs still seems to preclude trust.

5. Objections and Replies

Given all of the above arguments, there are a number of objections that should be considered. The first is that LLMs cannot be testifiers or give testimony, which explains why they do not fit into testimonial accounts. Why treat LLMs like testifiers rather than epistemic instruments such as thermometers? To this objection I would claim that this thesis aims to investigate LLMs as they are treated. LLMs, it seems, are treated like testifiers, and their outputs are treated like testimony. Users ask the models questions, receive answers, and form beliefs on the basis of them, which is functionally similar (and in some cases identical) to standard testimonial exchanges. LLM outputs occupy the

[24] Arena.ai, livebench.ai, and vellum.ai are three popular examples of LLM benchmark leaderboards. Proprietors are incentivized to maximize performance for these leaderboards in order to secure the top position in certain areas such as coding performance that will drive subscriptions.

functional role of testimony for many users, and thus to investigate whether or not these outputs are credible, we must use testimonial accounts. Even if LLMs are not testifiers metaphysically, their outputs seem to constitute or fill the role of testimony epistemically insofar as they play the same belief-formation role as ordinary testimony. We might argue, counter to the objection, that the conclusions presented here show that accounts of testimony need to be adapted to account for LLM testimony, or that we need LLM specific accounts of testimony. Some, such as Freiman (2024), aim to provide such accounts.

The second objection is that many of the arguments presented here rely on idealized models. However, if LLMs can in principle be experts but never actually are in practice, why should we care about these theoretical models? To this objection, I would point out that the goal of this chapter is not to evaluate any current models for expertise or credibility. Rather, the overarching goal is to understand whether or not the concepts of expertise, credible expertise, and credible testimony, apply to LLMs as a kind. Thus, I am focused here on the type (i.e., LLMs) rather than the tokens (e.g., GPT-5.2). Idealizing away from current models, further, allows us to not worry about trying to hit a moving target. If the discussion here was focused on current models, it would likely be out of date within a few months as new models are released. Thus, idealizations allow us to not get bogged down by current technological limitations that proprietors of these models are attempting to overcome, but also avoid the impossible task of attempting to hit a fast-moving target.

The third and final objection is that the final conclusion, that LLMs cannot provide justified beliefs via testimony, is too pessimistic. To this point, I would claim that simply because LLMs cannot provide us with justified beliefs via its testimony does not mean that they are not epistemically useful.

If nothing else, LLMs provide a starting point for reflection that can justify beliefs. Moreover, LLMs are great at providing alternative explanations, possible objections to a view that someone holds, or helping bring someone up to speed on an otherwise inaccessible topic. For instance, even if my belief cannot be justified on the mere basis of LLM testimony, I can nevertheless query an LLM with a question about a complex topic, such as quantum mechanics, with the specific prompt to make it digestible for someone with only a high school education. I can then go verify this output and come to gain some kind of knowledge or minimal understanding. While the LLM is not doing the justificatory work, it nevertheless is a useful epistemic tool.

Chapter 3 – Can Language Models Be Contributing Members to Group Understanding?

1. Introduction

As the understanding literature continues to evolve[25], the notion of group understanding has become increasingly important. For instance, we often say things like "The CDC understands how diseases spread" or "The Oklahoma City Thunder understand what they need to do in order to win." In relative lockstep, artificial intelligence systems continue to mature. With the rise of Large Language Models (LLMs), these systems have become vital parts of teams working to solve problems. These systems are proficient at combing through data, and thus may be useful in aiding smaller teams tackle bigger problems. Despite these clear contributions as tools, it is an open question whether an LLM can be a contributing *member of an epistemic group* that is said to collectively understand. In what follows, I use the influential overview of group understanding found in Boyd (2019). While there are other treatments, such as Malfatti (2022), Boyd's constitutes a seminal treatment of the concept.

In section one, I briefly outline deflationary group understanding and evaluate if LLMs can be contributing members under such an account. To do this, I look at three accounts of what it takes to understand, including Hills (2015), Khalifa (2017), and Le Bihan (2017). I conclude that on each of these accounts, LLMs cannot be said to understand, and thus cannot contribute to group understanding under a deflationary account. In section two, I briefly outline inflationary group understanding and evaluate if LLMs can be contributing members under such an account. The initial

[25] Khalifa (2012, 2013, 2017), De Regt and Dieks (2005), Hills (2015), Zagzebski (2008), Wilkenfeld (2013), Kvanvig (2003), and Christoph Kelp (2015) constitute a non-exhaustive but influential list of some work in the understanding literature.

upshot to this investigation is that inflationary group understanding does not require any of the members to possess understanding. Despite not requiring the LLM member to understand, I conclude in this section that LLMs cannot contribute to group understanding on an inflationary account. This is because they are incapable of recognizing their relationship to the other members of the group, which seems to be required on this account. LLMs can support individual understanding similar to a tool, but because they cannot understand themselves/be mutually p-reliant they cannot contribute as members to group understanding. Finally, in section three, I consider a number of objections to my view.

2. Deflationary Group Understanding

Boyd's notion of deflationary group understanding does not stray far from its summativist roots. In epistemology of belief, summativism is the theory that "members of a population, P, collectively believe that p when and only when all or most of its members believe that thing" (Gilbert 2004, 97). Recall that deflationary group understanding refers to a group that understands just in case "all or the majority of its members understands (why, how, that) p" (Boyd 2019, 17). This means that for an LLM to succeed on such an account, it will need to be non-metaphorically attributed with understanding. In this section, I will look at a number of prominent accounts of understanding and show why LLMs fail to understand on each of them. As a result, it seems unlikely that LLMs will be able to be members of a group that understands on a deflationary approach that requires its members to understand.

2.1 Hills's Account

Alison Hills (2015) focuses on understanding *why* p is true, where p is some proposition (661). Understanding why *p* is true requires not only that you believe that *p* but also that you have an explanation why *p* is true (e.g., *p* is true because *q*) (ibid., 663). Not only this, but you need to be able to grasp the relationship between *p* and *q* (ibid.). Hills's full account of understanding why is that to understand why *p* (and *q* is why *p*), then you have the belief that *p* and the belief that *q* is why *p* and in the right circumstances you can perform a number of abilities that show "cognitive control" (ibid.). These abilities include:

> (i) follow some explanation of why p given by someone else. (ii) explain why p in your own words. (iii) draw the conclusion that p (or that probably p) from the information that q. (iv) draw the conclusion that p' (or that probably p') from the information that q'. (where p' and q' are similar to but not identical to p and q). (v) given the information that p, give the right explanation, q. (vi) given the information that p', give the right explanation, q' (Hills 2015, 674).

I argue that this account does not allow for an LLM to understand. To see this, let us go through the abilities and consider whether or not an LLM could qualify. The first of these abilities is following an explanation of why p given by someone else. It seems, *prima facie*, like LLMs are capable of following an explanation given by someone else. If I give an LLM an explanation of some phenomenon that includes the causal relationships, it would likely be able to do some basic reasoning with these relationships that minimally constitutes "following" the explanation. The second ability is giving the explanation in your own words. This seems dicey for LLMs, as they can give regurgitation from their training data, but this is equivalent to a student giving an explanation that they merely rotely memorized and then wrote on a page. They can, however, give an explanation in their own words.

This too, does not seem to constitute understanding. The third, fourth, fifth, and sixth abilities require the understander to draw a conclusion or give an explanation. These abilities too seem cut off for LLMs as mere regurgitation of the most statistically likely explanation given a memorization of the explanations of others does not show cognitive control.

2.2 Khalifa's Account

Perhaps, however, Hills's notion of understanding is too stringent in requiring cognitive control and linguistic abilities.[26] Khalifa (2017) constitutes another fleshed out account of understanding. LLMs do not fare well on Khalifa's account. Khalifa is focused not on general understanding why *p*, but rather a kind of scientific understanding. On Khalifa's account, to grasp is to have a cognitive state bearing resemblance to scientific knowledge of some part of the explanation (Khalifa 2017, 11). A subject has scientific knowledge of an explanation, moreover, if and only if their belief that *q* explains why *p* is safe (i.e., the belief forming process could not have easily led to false belief) and this safety comes from her scientific explanatory evaluation (SEE) (Khalifa 2017, 12). This SEEing is, as Khalifa puts it, "naturally glossed in virtue-epistemological terms" (ibid., 13). Specifically, SEEing amounts to (i) considering numerous plausible explanations (ii) comparing these explanations and (iii) forming doxastic attitudes based on the comparisons (ibid.). The biggest problem for LLMs here will be comparing the explanations. To be sure, an LLM can be attributed with the ability to compare different explanations. Particularly, certain kinds of models called reasoning models will generate a hidden answer first, then internally critique this answer to find errors or a more accurate

[26] While this is an objection she considers directly (Hills, 667), it is worth our time to look at different accounts.

way to answer. The problem for an LLM is that it cannot weigh different explanatory criteria/virtues such as simplicity, scope, or conservativeness, as these explanatory virtues are somewhat subjective–particularly in a way that is unavailable to LLMs. Indeed, much of what is entailed by the consideration of a "best" explanation is subjective. One individual might value simplicity the highest while another might value thoroughness the highest. We might take further issue that an LLM is not capable of forming doxastic attitudes in a standard sense. Perhaps, admittedly, we can attribute an LLM with a minimal kind of "representational state" that can be seen as akin to a belief and thus perhaps a sort of doxastic attitude. Even if we grant this, however, the weighing of explanatory criteria seems inaccessible for an LLM.[27]

2.3 Le Bihan's Account

Finally, we may claim that the above accounts of understanding from Hills (2015) and Khalifa (2017) are too stringent as they require grasping of an explanation to understand. We can, instead, try a modal account of understanding from Le Bihan (2017).[28] For Le Bihan, someone has modal understanding of some phenomenon *p* "if and only if one knows how to navigate some of the possibility space associated with the phenomenon" (Le Bihan, 112). Modal understanding of a phenomenon, more simply, is some understanding of how a phenomenon *might* arise. The possibility space for a phenomenon refers to the dependency structures that, when properly arranged, give rise to the phenomenon in addition to the relationships between those dependency structures (ibid., 114). For example, if we imagine a phenomenon of gas behavior, one of these dependency structures would

[27] Importantly, I am not claiming that IBE as a process is entirely subjective. Rather, the decision of what constitutes "best" is partially subjective.

[28] Rice (2021) also gives a modal account, but requires the agent to grasp the modal information.

be that kinetic energy when associated with the mean speed of gas molecules gives rise to the internal energy of gases (ibid., 115). The relations of dependency structures refers not only to how they are meant to interact with one another but also how they may be parts of larger dependency structures. Navigating a possibility space comes in three levels, for Le Bihan. The first level (which is the most basic) is knowing how a particular dependency structure gives rise to P (Le Bihan, 117). There are two more levels, but if an LLM cannot navigate on the first level, it cannot navigate on the second or third levels. Thus, we ought to investigate if an LLM can qualify on this level before moving forward.

It should first be noted that Le Bihan's modal understanding is a form of "know-how" (ibid.). Thus, an LLM still needs to show some kind of ability (in this case, an ability to navigate the possibility space) even though it does not need to display an ability to grasp. Now, while the ability that is *constitutive* of this level of navigating and thus modal understanding is knowing how a particular dependency structure gives rise to a phenomenon, this can be manifested through *derivative* abilities, such as the abilities to answer counterfactual questions or show an ability to justify views when pressed by a critical interlocutor (Le Bihan, 117). The problem facing an LLM will, at this point, be quite familiar. While an LLM can give outputs about what causes some phenomenon to arise in this world, can seemingly answer counterfactual questions, and can likely show the kind of epistemic resilience that is manifested in being able to justify views when criticized, it nevertheless is merely predicting a string of words to the user, rather than considering various answers or thoughtfully generating a response in a similar way as a human. In all of the above accounts of understanding, an LLM is closer to Planck's chauffeur rather than a genuine understander.

2.4 Planck's Chauffeur

Planck's chauffeur is a (probably) apocryphal story about the German physicist Max Planck after he won the Nobel Prize. As the story goes, after winning the award, Planck goes around Germany giving the same lecture over and over again. Over time, his chauffeur, who watches each of these lectures every night, claims that he has heard the lecture so many times that he could give it himself. Planck and the chauffeur switch clothes and the chauffeur gets up to give the lecture while Planck, dressed like the chauffeur, sits in the front. The chauffeur gives a perfect lecture, exactly as Planck would have given it. After the talk, an actual physicist asks a complex question, to which the chauffeur exclaims, "This is such an easy question! I invite my chauffeur in the front row to answer it!" This story is sometimes taken to show the difference between knowledge and understanding, pretend and real knowledge, or any number of distinctions. In all of the above accounts of understanding, there is a tacit requirement for the potential understander to be doing the cognitive work to generate the explanation *themselves* or grasp some phenomenon *themselves*. Here's how I think we should understand this story in the context of LLMs: an LLM is just a very advanced version of Planck's chauffeur insofar as an LLM has just heard most, if not all, speeches over and over again, in addition to possible questions that could be asked. For instance, if the chauffeur had heard Planck answer the exact question the physicist asked, he could have of course given a perfectly good answer. Even if he had heard and memorized every possible question and respective answer, we would not want to attribute grasp of the material. An LLM is merely Planck's chauffeur iterated billions of times over, having memorized definitions, questions, answers, explanations, etc.

2.5 The State of LLM Understanding

Where does all of this discussion leave us? On the account proposed in Hills (2015), LLMs seem to either lack the cognitive control or lack the ability to produce correct explanations in virtue of their training data. On the account proposed by Khalifa (2017), LLMs seem to be unable to SEE and thus cannot grasp. On a modal account of understanding from Le Bihan (2017), LLMs seem to be unable to navigate the possibility space of some phenomenon in a non-metaphorical way. These three accounts are not the only ones that are available, but they constitute some of the more fleshed out accounts in the literature. In all of these accounts the problem for LLMs is quite similar. It seems intuitive to attribute LLMs with possessing an explanation, but we are hard pressed to attribute them with *grasping* an explanation or being able to reason with that explanation and alternatives. Thus, LLMs as possibly contributing members on a deflationary account of understanding seems unlikely at best and completely impossible at worst.

3. Inflationary Group Understanding

If a deflationary account presents insurmountable hurdles, an inflationary account may be a better option. After all, inflationary accounts do not require constituents of the group to understand in order for the group to understand. Thus, turning towards an inflationary account has the distinct advantage of not requiring to weigh in on whether or not an LLM can be said to understand. However, the inflationary account is not without its issues either. Specifically, it suffers from what Boyd calls the "group grasping problem" (Boyd 2019, 27). As Boyd defines it, a group grasps some proposition *p* and reasons to support it just in case the group both has representation of the material somewhere in the group and that the members of the group are "mutually p-reliant" (ibid.). Boyd

leaves the first criterion schematic. He claims that all that is required is that the material being grasped is "presented somewhere within the group" (ibid.). Boyd further notes that we could appeal to a theory of group belief to get this representation criterion off of the ground, but that the requirement is meant to be neutral and accommodate different theories of group representation (Boyd 2019, 27). This second criterion is defined by the members of the group recognizing that they are working towards a shared goal and that they could not do it by themselves (ibid.). To elucidate this criterion, imagine if a lab consists of three scientists and only one of them really understood anything with regards to p, we would not say that the group grasps p but rather one member does. If we imagine that none of the scientists are able to recognize that they are working towards the same goal and must rely on each other to complete it, they are not mutually p-reliant. To show both of these in action, Boyd uses the dependable and dysfunctional autobody examples. In the former, the mechanics share a goal, and trust that each person will do their task. In the latter, each mechanic distrusts everyone else, and even though cars get fixed, we are hesitant to say that the shop as a group has understanding with regards to fixing cars (Boyd 2019, 27-28). While an LLM need not understand on this account, it still faces a pressing problem. Specifically, for an LLM to contribute to inflationary group understanding, there is a pressing problem we must face: mutual *p*-reliance.

3.1 Trust is Not the Issue

Recall that for Boyd, inflationary group understanding refers to the idea that whether or not a group understands does not depend solely on whether its members understand themselves (Boyd 2019, 18). The crucial problem for such an account is the problem of group grasping, which requires that a group represent reasons for *p* and the members of the group are "mutually *p*-reliant" (ibid., 27).

The second requirement is what motivates the group grasping problem. As Boyd claims, members are mutually *p*-reliant in the case that they recognize they are contributing to a shared goal and could not achieve it on their own. While trust is not explicitly a part of Boyd's account, he elucidates his *Dependable* and *Dysfunctional* autobody cases with members trusting and distrusting each other respectively (ibid., 28). In trusting one another, the members of the dependable autobody seem to have grasping at the group level, while members of the dysfunctional autobody do not (ibid.). Trust, however, does not seem to be the ultimate fulcrum for mutual *p*-reliance Imagine the following:

School Project: Charlie, Lucy, Sally, and Linus are assigned to a group to reenact a scene from Shakespeare's *Romeo and Juliet*. None of them have previously interacted, as they belong to different school cliques, and none feel any particular motivation to get to know the others. Each student is aware that completing the assignment requires all roles to be performed: Charlie will play Romeo, Lucy will play Juliet, Sally will play Tybalt, and Linus will play Mercutio. Each student focuses on learning their own part. They notice when the teacher, Ms. Othmar, requests updates on progress, and provide the information as required, without making any assumptions about the others' commitment or competence. They do not consider whether the others will do well or poorly. Rather, they simply respond to the assignment's requirements. On the day of the performance, each student enacts their role according to Ms. Othmar's expectations. Since everyone has completed the necessary preparation independently, the scene is performed successfully, and the assignment is completed.[29]

[29] Thank you to my partner Nathalie for this example.

In *School Project*, the students have a truly neutral relationship to one another, while still being mutually p-reliant as they all recognize that they cannot complete the assignment alone. Thus, while Boyd uses the language of trust and distrust in his cases, we should view this language as merely a way to easily get the importance of contribution to shared goals off of the ground.

3.2 LLMs in an Inflationary Scenario

Even if we have reason to believe that trusting relationships are not required, we still need to defend the argument that an LLM can be placed into a case like *School Project* without loss. Consider the following example:

Silicon Valley: At a large technology company, three employees have been tasked with the job of reworking the software of an important product. None of the employees have ever met one another, as they all are from different teams in the company. John is the company's JavaScript expert, Paul is the company's Python expert, and George is the assigned product manager of the team. Since the company just laid off their Java expert, the team is assigned an LLM, named Ringo, to be their Java programmer. The team works in relative isolation, and their only communication comes in the form of meetings which the company requires them to attend, and at which nothing more is said by John and Paul other than, "Looks good to me." Ringo, the LLM Java expert, is fed what the team needs by George. Ringo then writes Java code to accomplish the tasks required and George, who remembers virtually no Java, simply implements the code where he is told. After working on the project, the team eventually presents a working product, much to the satisfaction of their employer.

In the case of *Silicon Valley*, it would seem that the individuals themselves do not have the relevant understanding, so the important question is whether or not the group as a whole has

understanding with regards to their product (such as understanding how to build it, or understanding why it works, etc.). Assuming we want to attribute group understanding in a case like *School Project*, then there are two questions we must answer to figure out if the members of the group in *Silicon Valley* grasp at the group level and thus could be said to possess group understanding: (i) do the members of the group recognize their contribution towards a shared goal and dependence on each other and (ii) can the LLM Ringo make this same recognition? To the first question, I think we can intuitively say that Paul, John, and George each recognize that they are all working towards a shared goal and must rely on one another to accomplish this goal. While the case has been left rather simple and the members are not as close as those in Boyd's *Dependable Autobody*, they nevertheless seem to be aware of their own strengths and weaknesses, and possess no belief that they could (or should) complete the project themselves. The second question, then, will be our sticking point. It will be important first, however, to get clear on what it means to have and contribute to a shared goal and be dependent on others, as well as what it means to be aware of, or recognize, this relationship.

3.3 Shared Goals and Dependence

What can we make of having and contributing to shared goals and dependence between members of an epistemic group? These goods are, I think intuitively, quite cheap. In the case of human members at least, Boyd's example of pulling someone out of a body of water is a perfect example:

> Consider, for example, a case in which a man has fallen into a river, with a strong current threatening to carry him downstream and out of reach of help. Thinking quickly, a friend on the shore tosses him a rope, which the man in the river is able to hold onto. With the current being so strong, however, the friend on the shore has to call on three others for help, all of whom grab part of the rope. Individually,

> none of them would be able to maintain a grip on the rope, or manipulate it in the desired way (i.e., to rescue the drowning friend). Together, however, they are able to pull the man to shore (Boyd 2019, 25).

In this example, all of the friends share a goal of saving the man in the river. Moreover, they are aware of this shared goal insofar as they could consciously reason that, since everyone else is also helping to pull the man out of the river, they must also share the goal of wanting to save the man. They are all clearly dependent on one another to realize this goal, insofar as they could not individually pull the man out of the water. Moreover, they are aware of this dependence relationship insofar as they, perhaps, feel how challenging it is to pull the man out of the river working together and can consciously reason that they could not physically pull the man out of the river by themselves. Thus, there are four parts to achieve mutual *p*-reliance: (i) the members of the group need to share a goal (ii) the members of the group need to be aware of this shared goal (iii) the members of the group need to be dependent on one another to achieve said goal and (iv) the members of the group need to be aware of this dependence relationship.

The example used from Boyd only has human members. As a result, it is easy to imagine the members of the group meeting all criteria. If we supplement one of the members with an LLM, however, it becomes harder to imagine *all* members meeting all criteria. Recall that in the *Silicon Valley* example, Ringo is clearly dependent on the other members of the group. Ringo is unable to write Python or JavaScript and so is dependent on Paul or John respectively. Ringo is also dependent on George as he prompts Ringo with the necessary

information to write the Java. Moreover, Ringo cannot implement the code himself, and thus is dependent on George to do so. Dependence on the other members, it seems, is cheap to get.

Ringo having a shared goal is a bit tricker, however. It may be the case that LLMs are merely not the kind of thing that can have a goal in the first place. There are minimal accounts of concepts such as agency that construe goals very minimally. Christian List (2016), for instance, construes goals as merely representational states of how the entity "would like the world to be" (297). This notion may or may not actually constitute a "goal," as we imagine the concept, but I will move forward with the assumption that LLMs can have this minimal type of goal and that such a type is sufficient for mutual p-reliance. As we will see, LLMs will still fail to be mutually *p*-reliant either way.

Assuming that LLMs *can* have dependence and a shared goal, all they need from the four requirements we laid out earlier is awareness or recognition of these relationships. This recognition, however, is plainly not available to LLMs. Even though they will have incredibly rote recognitional skills such that they can quickly analyze data or could possibly be fed two scenarios and claim that in both scenarios the members are dependent and share goals, they are unable to be aware that *they* are dependent or share goals because they do not have a sense of self. Recognition of dependence or shared goals presupposes that the entity doing the recognizing is capable of distinguishing itself from others, but LLMs lack this self-referential perspective. Moreover, they lack a sense of self insofar as they experience no phenomenological consciousness. There is no salient "what-it's-likeness" for an LLM in the same way there is for a human or an animal. In other words, there is nothing it is like to be an LLM. Any impression

that LLMs are experiencing this first-person subjectivity is merely a byproduct of its outputs, which are in many cases indistinguishable from human "outputs," being interpreted by humans. Any sense of self we attribute to an LLM is merely a gloss on the output to make the system seem more personable to a user. While they can manipulate representations of these relationships via text, they themselves do not occupy a first-person subjective experience from which these relationships are experientially meaningful.

4. Objections and Replies

One objection that is worth considering at this point is that this recognition is not required for mutual *p*-reliance. Perhaps all we need is the dependence relationship or contributing towards a shared goal or both. When we get rid of this awareness or recognition requirement, however, we get a lot of epistemic bloat. That is to say that there will be far too many groups that understand such that the term "group understanding" would go far beyond the phenomenon it is supposed to refer to. Imagine that all is required for mutual *p*-reliance is a dependence relationship between the members. There is no need for contribution to the same goal, nor a need to recognize the dependence relationship. Nevertheless, this conception of mutual *p*-reliance will not suffice. Consider the example of Otto and his notebook from Andy Clark and David Chalmers (1998). Otto suffers from Alzheimer's and relies on information that he writes in his notebook. Whenever he learns something new, he writes it down in said notebook. When he needs old information he looks it up in his notebook. Groups are easy to come by, and membership seems to be cheap. Thus, we can tentatively claim Otto and his notebook are a group of sorts. If there is no need for contributing towards

the same goal, nor a recognition of dependence, it seems like Otto and his notebook are mutually *p*-reliant. He clearly depends on his notebook, and his notebook depends on him insofar as the notebook would perhaps not be used without Otto. This dependence relationship does not seem like it is enough to claim that Otto and the notebook understand as a group how to get to the museum on 53rd Street.

If mere dependence is not enough, perhaps dependence and contribution towards the same goal is sufficient for mutual *p*-reliance. If it is only these components, and no recognition of them by the members, we are again left with an odd picture. In such a world, I could be said to form a group with and be mutually *p*-reliant upon René Descartes. Plausibly we both share the goal of advancing epistemology. Moreover, I am dependent upon him insofar as he kickstarted modern epistemic endeavors with his writings, without which I would likely not be an epistemologist. He is dependent on me to reach our shared goal insofar as he could not investigate all areas of epistemology and thus needs someone else to "carry the torch," as it were. Do we form a group that comes to understand why *p* (whatever *p* may be in this instance)? We can run a similar example in the opposite direction. Am I in an epistemic group with someone who reads my paper in 20 years and advances knowledge in social epistemology? Even excluding weird temporal cases such as these, if all that is required is a mere dependence relationship and a shared goal (or mere contribution towards the same goal, in the case of Otto), and *not* an awareness of these conditions, it seems that our notion of group understanding no longer picks out the phenomenon that we desire and we are left with a great deal of epistemic bloat.

Two further objections are closely related. First, much of the problem with LLMs being attributed with understanding and thus being considered contributing members of a group that is said to understand on a deflationary account is that they are (seemingly) non-conscious entities, what if future LLMs are conscious? If they become conscious, could they be able to understand and furthermore, could they develop goals of their own and be contributing members on an inflationary account as well? Secondly, are we able to program LLMs to grasp goals and thus be contributing members on an inflationary account?

To the first objection, I think it largely depends on what we mean by "conscious." Nevertheless, insofar as it is definitionally possible for LLMs to be conscious (which I will merely assume), such an advancement would likely be sufficient for LLMs to understand phenomena in and of themselves. While consciousness is not explicitly a part of any of the accounts of understanding presented here, it seems that being conscious would allow an LLM to (at least in a minimal sense) grasp explanations and reason with them. It may not be on the same level as human understanding, grasping, and reasoning, but it would perhaps nevertheless constitute. Consciousness, moreover, would allow an LLM to be a contributing member on inflationary accounts. This is because it would rid an LLM of any problem with it not being able to occupy a first-person perspective and thus recognize its relationships to the other members. This objection, however, is only pressing if we believe that (i) LLMs are the kind of thing that can be conscious in principle and (ii) LLMs will be conscious in the future. I leave both of these possibilities to future research.

To the second objection, I do not think that mere programing will allow an LLM to grasp or recognize goals and/or recognize dependency relations between members. Already an LLM can have outputs that *seem* to demonstrate the capability to recognize dependency relations, but as we saw previously, occupying a first-person perspective seems to be a necessary condition for the kind of self-referential reasoning required to recognize your standing in the epistemic relationships of group grasping.

There is one final objection that I feel is the most pertinent to our discussion here. I claimed that LLMs cannot be contributing members on an inflationary account of group understanding because they cannot be said to actually recognize that they have certain relationships with the other members, namely dependence and shared goals. A functionalist, however, might want to push back by claiming that a "sense of self" could mean one of two things. It could mean they lack phenomenal self-awareness (i.e., there is nothing it is like to be them), or it could mean they lack any internal states that represent themselves as distinct from their environment and track their own capacities and limitations. While these are presumed to be the same, a functionalist will insist they are different. Modern LLMs, particularly those with chain-of-thought reasoning, do appear to have these kinds of internal states. Specifically, they represent their own uncertainty, acknowledge their own limitations, and track what they have and haven't been told. The question then is whether or not this later sense of self is sufficient.

In response to this objection, I would argue that the kind of self-recognition required for mutual p-reliance is not merely functional. Recognizing that "I am dependent on my

group members" is not just a matter of having a representation that says so, it requires that this dependence matters to you, that it has stakes, and that failure would be felt. This is why Boyd's dysfunctional autobody mechanics fail the mutual p-reliance condition even though they clearly have functional representations of each other's roles. They simply don't care about the shared project in the relevant way. LLMs are an even more extreme version of this. It is not that they are merely indifferent (such as the *School Project* case), but incapable of the kind of caring that gives recognition its weight.

5. Conclusion

What are we left with in regards to LLMs as members of a group that comes to understand? If one wishes to take a deflationary account of group understanding, LLMs cannot be said to contribute to group understanding insofar as they cannot be said to understand. The abilities or capacities necessary to grasp seem to be fundamentally cut off from LLMs and thus, understanding as a whole. If one wishes to take an inflationary account of group understanding, however, this is not a problem. A new problem arises, however, when we consider group grasping. On Boyd's account, group grasping requires mutual *p*-reliance, which further requires a shared goal and dependence on the other members, as well as recognition of these relationships. As we saw, LLMs are incapable of this awareness as they are nonconscious entities which cannot have a sense of self and thus cannot recognize that "I am dependent on my group members to realize a goal that I share with them." Thus, LLMs cannot be said to be capable of mutual *p*-reliance, and therefore group grasping, and therefore group understanding on an inflationary account. LLMs are not completely useless, however.

They are clearly very powerful tools which can aid a group in achieving understanding. They cannot, however, be contributing members of groups that understand.

References

Adler, Jonathan E. “Testimony, trust, knowing.” *Journal of Philosophy*, vol. 91, no. 5, 1994, pp. 264–275, https://doi.org/10.2307/2940754.

Adler, Jonathan Eric. *Belief's Own Ethics*. MIT Press, 2002.

Almassi, Ben. “Climate change, epistemic trust, and expert trustworthiness.” *Ethics and the Environment*, vol. 17, no. 2, 2012, p. 29, https://doi.org/10.2979/ethicsenviro.17.2.29.

Anderson, Charity. “Epistemic authority and conscientious belief.” *European Journal for Philosophy of Religion*, vol. 6, no. 4, 22 Dec. 2014, pp. 91–99, https://doi.org/10.24204/ejpr.v6i4.147.

Anderson, Elizabeth. “Democracy, public policy, and lay assessments of scientific testimony.” *Episteme*, vol. 8, no. 2, June 2011, pp. 144–164, https://doi.org/10.3366/epi.2011.0013.

Baker, Judith, and Philip Clark. “Epistemic buck-passing and the interpersonal view of testimony.” *Canadian Journal of Philosophy*, vol. 48, no. 2, 2018, pp. 178–199, https://doi.org/10.1080/00455091.2017.1341781.

Bennett, Matthew. “Judging expert trustworthiness: The difference between believing and following the science.” *Social Epistemology*, vol. 36, no. 5, 16 Aug. 2022, pp. 550–560, https://doi.org/10.1080/02691728.2022.2106459.

Bojesomo, Alabi, et al. *Revolutionizing Disease Diagnosis with Large Language Models: A Systematic Review*, 27 Dec. 2024, https://doi.org/10.21203/rs.3.rs-5704278/v1.

Bonjour, Laurence. “Externalist theories of empirical knowledge.” *Midwest Studies in Philosophy*, vol. 5, 1980, pp. 53–73, https://doi.org/10.1111/j.1475-4975.1980.tb00396.x.

Borji, Ali. “A Categorical Archive of CHATGPT Failures.” *arXiv.Org*, 3 Apr. 2023, arxiv.org/abs/2302.03494.

Boyd, Kenneth. “Group understanding.” *Synthese*, vol. 198, no. 7, 3 Dec. 2019, pp. 6837–6858, https://doi.org/10.1007/s11229-019-02492-3.

Boyd, Kenneth. “Trusting scientific experts in an online world.” *Synthese*, vol. 200, no. 1, Feb. 2022, https://doi.org/10.1007/s11229-022-03592-3.

Bubeck, Sébastien, et al. “Sparks of Artificial General Intelligence: Early Experiments with GPT-4.” *arXiv.Org*, 13 Apr. 2023, arxiv.org/abs/2303.12712.

Burge, Tyler. “Content preservation.” *The Philosophical Review*, vol. 102, no. 4, Oct. 1993, p. 457, https://doi.org/10.2307/2185680.

Burge, Tyler. “Interlocution, perception, and memory.” *Philosophical Studies*, vol. 86, no. 1, Apr. 1997, pp. 21–47, https://doi.org/10.1023/a:1004261628340.

Busch, Felix, et al. “Current applications and challenges in large language models for patient care: A systematic review.” *Communications Medicine*, vol. 5, no. 1, 21 Jan. 2025, https://doi.org/10.1038/s43856-024-00717-2.

Cangelosi, Ocean. “Can ai know?” *Philosophy & Technology*, vol. 37, no. 3, 1 July 2024, https://doi.org/10.1007/s13347-024-00776-2.

Chalmers, David J. “Could a Large Language Model Be Conscious?” *arXiv.Org*, 18 Aug. 2024, doi.org/10.48550/arXiv.2303.07103.

Clark, A., and D. Chalmers. “The extended mind.” Analysis, vol. 58, no. 1, 1 Jan. 1998, pp. 7–19, https://doi.org/10.1093/analys/58.1.7.

Coady, C. A. J. *Testimony: A Philosophical Study*. Clarendon; Oxford University Press, 1992.

Collins, Patricia Hill. *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*. Routledge, 2000.

Constantin, Jan, and Thomas Grundmann. “Epistemic authority: Preemption through source sensitive defeat.” *Synthese*, vol. 197, no. 9, 4 Sept. 2020, pp. 4109–4130, https://doi.org/10.1007/s11229-018-01923-x.

Croce, Michel. “Expert-oriented abilities vs. novice-oriented abilities: An alternative account of Epistemic Authority.” *Episteme*, vol. 15, no. 4, 23 May 2017, pp. 476–498, https://doi.org/10.1017/epi.2017.16.

Dahl, Matthew, et al. “Large legal fictions: Profiling legal hallucinations in large language models.” *Journal of Legal Analysis*, vol. 16, no. 1, 1 Jan. 2024, pp. 64–93, https://doi.org/10.1093/jla/laae003.

De Regt, Henk W., and Dennis Dieks. “A contextual approach to scientific understanding.” *Synthese*, vol. 144, no. 1, Mar. 2005, pp. 137–170, https://doi.org/10.1007/s11229-005-5000-4.

Dellsén, Finnur, and Øystein Linnebo. *The Epistemology of Experts: New Essays*, edited by Peter Brössel et al., Routledge, 2026, pp. 1–20.

Dormandy, Katherine. “Epistemic authority: Preemption or proper basing?” *Erkenntnis*, vol. 83, no. 4, 19 July 2018, pp. 773–791, https://doi.org/10.1007/s10670-017-9913-3.

Dougherty, Trent. “Zagzebski, authority, and faith.” *European Journal for Philosophy of Religion*, vol. 6, no. 4, 22 Dec. 2014, pp. 47–59, https://doi.org/10.24204/ejpr.v6i4.144.

Elgin, Catherine Z. “True enough*.” *Philosophical Issues*, vol. 14, no. 1, Oct. 2004, pp. 113–131, https://doi.org/10.1111/j.1533-6077.2004.00023.x.

Elgin, Catherine. “Understanding and the facts.” *Philosophical Studies*, vol. 132, no. 1, 30 Nov. 2007, pp. 33–42, https://doi.org/10.1007/s11098-006-9054-z.

Faulkner, Paul. “David Hume’s reductionist epistemology of testimony.” *Pacific Philosophical Quarterly*, vol. 79, no. 4, Dec. 1998, pp. 302–313, https://doi.org/10.1111/1468-0114.00065.

Fodor, Jerry. “Modules, Frames, Fridgeons, Sleeping Dogs, and the Music of the Spheres.” *The Robot’s Dilemma: The Frame Problem in Artificial Intelligence*, edited by Zenon Pylyshyn, Ablex, 1987, pp. 139–149.

Freiman, Ori. “Ai-Testimony, conversational AIS and our anthropocentric theory of testimony.” *Social Epistemology*, vol. 38, no. 4, 6 Mar. 2024, pp. 476–490, https://doi.org/10.1080/02691728.2024.2316622.

Freiman, Ori. “Analysis of beliefs acquired from a conversational ai: Instruments-based beliefs, testimony-based beliefs, and technology-based beliefs.” *Episteme*, vol. 21, no. 3, 6 Mar. 2023, pp. 1031–1047, https://doi.org/10.1017/epi.2023.12.

Fricker, Elizabeth, and David E. Cooper. “The epistemology of testimony.” *Aristotelian Society Supplementary Volume*, vol. 61, no. 1, 1 July 1987, pp. 57–106, https://doi.org/10.1093/aristoteliansupp/61.1.57.

Fricker, Elizabeth. “Against Gullibility.” *Knowing from Words: Western and Indian Philosophical Analysis of Understanding and Testimony*, edited by Arindam Chakrabarti and Bimal K Matilal, Kluwer Academic Publishers, 1994, pp. 125–161.

Fricker, Elizabeth. “Critical notice.” *Mind*, vol. 104, no. 414, 1995, pp. 393–411, https://doi.org/10.1093/mind/104.414.393.

Fricker, Elizabeth. “On the metaphysical and epistemic contrasts between human and ai testimony.” *Inquiry*, 29 Sept. 2025, pp. 1–26, https://doi.org/10.1080/0020174x.2025.2553297.

Fricker, Elizabeth. “Second-hand knowledge*.” *Philosophy and Phenomenological Research*, vol. 73, no. 3, Nov. 2006, pp. 592–618, https://doi.org/10.1111/j.1933-1592.2006.tb00550.x.

Fricker, Elizabeth. “Trusting others in the sciences: A priori or empirical warrant?” *Studies in History and Philosophy of Science Part A*, vol. 33, no. 2, June 2002, pp. 373–383, https://doi.org/10.1016/s0039-3681(02)00006-7.

Gilbert, Margaret. “Collective epistemology.” *Episteme*, vol. 1, no. 2, Oct. 2004, pp. 95–107, https://doi.org/10.3366/epi.2004.1.2.95.

Gilbert, Margaret. *On Social Facts*. Princeton University Press, 1989.

Goldberg, Sanford, and David Henderson. “Monitoring and anti-reductionism in the epistemology of testimony.” *Philosophy and Phenomenological Research*, vol. 72, no. 3, May 2006, pp. 600–617, https://doi.org/10.1111/j.1933-1592.2006.tb00586.x.

GOLDMAN, ALVIN I. “Experts: Which ones should you trust?” *Philosophy and Phenomenological Research*, vol. 63, no. 1, July 2001, pp. 85–110, https://doi.org/10.1111/j.1933-1592.2001.tb00093.x.

Goldman, Alvin I. *Knowledge in a Social World*. Clarendon Press: Oxford University Press, 1999.

Gordon, Emma C. “Is there propositional understanding?” *Logos & Episteme*, vol. 3, no. 2, 2012, pp. 181–192, https://doi.org/10.5840/logos-episteme20123234.

Grimm, Stephen R. “Is understanding a species of knowledge?” *The British Journal for the Philosophy of Science*, vol. 57, no. 3, 1 Sept. 2006, pp. 515–535, https://doi.org/10.1093/bjps/axl015.

Grimm, Stephen. “Understanding.” *The Routledge Companion to Epistemology*, edited by Duncan Pritchard and Sven Berneker, Routledge, 2011, pp. 1–22.

Grundmann, Thomas. “Experts: What Are They and How Can Laypeople Identify Them?” *The Oxford Handbook of Social Epistemology*, edited by Aidan McGlynn and Jennifer Lackey, Oxford University Press, 2025, pp. 85–104.

Guerrero, Alexander. “Living with Ignorance in a World of Experts.” *Perspectives on Ignorance from Moral and Social Philosophy*, edited by Rik Peels, Routledge, 2016, pp. 1–21.

Guo, Yufei, et al. “Bias in Large Language Models: Origin, Evaluation, and Mitigation.” *arXiv.Org*, 16 Nov. 2024, arxiv.org/abs/2411.10915.

Hannon, Michael. “Recent work in the epistemology of understanding.” *American Philosophical Quarterly*, vol. 58, no. 3, 1 July 2021, pp. 269–290, https://doi.org/10.2307/48616060.

Hardwig, John. “Epistemic dependence.” *The Journal of Philosophy*, vol. 82, no. 7, July 1985, p. 335, https://doi.org/10.2307/2026523.

Hardwig, John. “The role of trust in knowledge.” *The Journal of Philosophy*, vol. 88, no. 12, Dec. 1991, p. 693, https://doi.org/10.2307/2027007.

Haugeland, John. *Artificial Intelligence: The Very Idea*. A Bradford Book - The MIT Press, 1989.

He, Jinhua, and Chen Yang. "Testimony by Llms." *AI & Society*, vol. 40, no. 8, 24 Apr. 2025, pp. 6201–6213, https://doi.org/10.1007/s00146-025-02366-y.

Hempel, Carl Gustav. *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. Free Press; Collier Macmillan, 1965.

Hernandez Flowerman, Camila. "AI, LLMs, and the Normativity of Belief." *Synthese*.

Hills, Alison. "Understanding why." *Noûs*, vol. 50, no. 4, 16 Apr. 2015, pp. 661–688, https://doi.org/10.1111/nous.12092.

Hume, David, and Eric Steinberg. *An Enquiry Concerning Human Understanding; with a Letter from a Gentleman to His Friend in Edinburgh ; and an Abstract of a Treatise of Human Nature*. Hackett Publishing Company, 1993.

Insole, Christopher J. "Seeing off the local threat to irreducible knowledge by testimony." *The Philosophical Quarterly*, vol. 50, no. 198, Jan. 2000, pp. 44–56, https://doi.org/10.1111/1467-9213.00166.

Irzik, Gürol, and Faik Kurtulmus. "What is Epistemic Public Trust in science?" *The British Journal for the Philosophy of Science*, vol. 70, no. 4, 1 Dec. 2019, pp. 1145–1166, https://doi.org/10.1093/bjps/axy007.

Jäger, Chrisoph. "Epistemic Authority." *The Oxford Handbook of Social Epistemology*, edited by Aidan McGlynn and Jennifer Lackey, Oxford University Press, 2025, pp. 63–84.

Kelp, Christoph. "Understanding phenomena." *Synthese*, vol. 192, no. 12, 7 Jan. 2015, pp. 3799–3816, https://doi.org/10.1007/s11229-014-0616-x.

Kenyon, Tim. "The informational richness of testimonial contexts." *The Philosophical Quarterly*, vol. 63, no. 250, 17 Dec. 2012, pp. 58–80, https://doi.org/10.1111/1467-9213.12000.

Khalifa, Kareem. "Inaugurating understanding or repackaging explanation?" *Philosophy of Science*, vol. 79, no. 1, 1 Jan. 2012, pp. 15–37, https://doi.org/10.1086/663235.

Khalifa, Kareem. *Understanding, Explanation, and Scientific Knowledge*. Cambridge University Press, 2017.

Khalifa, Kareem. "Understanding, grasping and luck." *Episteme*, vol. 10, no. 1, Mar. 2013, pp. 1–17, https://doi.org/10.1017/epi.2013.6.

Kim, Jaegwon. "Explanatory knowledge and metaphysical dependence." *Philosophical Issues*, vol. 5, 1994, p. 51, https://doi.org/10.2307/1522873.

Kojima, Takeshi, et al. “Large Language Models Are Zero-Shot Reasoners.” *arXiv.Org*, 29 Jan. 2023, arxiv.org/abs/2205.11916.

Kumar, Charaka Vinayak, et al. “No LLM Is Free from Bias: A Comprehensive Study of Bias Evaluation in Large Language Models.” *arXiv.Org*, 27 May 2025, arxiv.org/abs/2503.11985.

Kvanvig, Jonathan L. *The Value of Knowledge and the Pursuit of Understanding*. Cambridge University Press, 2003.

Lackey, Jennifer. “Experts and Peer Disagreement.” *Knowledge, Belief, and God: New Insights in Religious Epistemology*, edited by Matthew Benton et al., Oxford University Press, 2018, pp. 228–245.

Lackey, Jennifer. *Learning from Words: Testimony as a Source of Knowledge*. Oxford University Press, 2008.

Lackey, Jennifer. *The Epistemology of Groups*. Oxford University Press, 2021.

Lackey, Jennifer. “What is justified group belief?” *The Philosophical Review*, vol. 125, no. 3, 1 July 2016, pp. 341–396, https://doi.org/10.1215/00318108-3516946.

Lavrinovics, Ernests, et al. “Knowledge graphs, large language models, and hallucinations: An NLP perspective.” *Journal of Web Semantics*, vol. 85, May 2025, p. 100844, https://doi.org/10.1016/j.websem.2024.100844.

Le Bihan, Soazig. “Enlightening Falsehoods: A Modal View of Scientific Understanding.” Explaining Understanding: New Perspectives From Epistemology and Philosophy of Science, edited by Stephen Grimm et al., Routledge, London, 2017, pp. 111–136.

Lehrer, Keith. “Testimony and Trustworthiness.” *Epistemology of Testimony*, edited by Jennifer Lackey and Ernest Sosa, Clarendon Press, 2006, pp. 145–159.

Lipton, Peter. “Understanding Without Explanation.” *Scientific Understanding: Philosophical Perspectives*, edited by Hank W. De Regt et al., University of Pittsburgh Press, 2008, pp. 43–63.

List, Christian, and Philip Pettit. *Group Agency: The Possibility, Design, and Status of Corporate Agents*. Oxford Univ. Press, 2011.

List, Christian. “What is it like to be a group agent?” Noûs, vol. 52, no. 2, 28 July 2016, pp. 295–319, https://doi.org/10.1111/nous.12162.

Locke, John, and Kenneth Winkler. *An Essay Concerning Human Understanding*. Hackett Pub. Co, 1996.

Longino, Helen E. *The Fate of Knowledge*. Princeton University Press, 2002.

Lyons, Jack. "Testimony, induction and folk psychology." *Australasian Journal of Philosophy*, vol. 75, no. 2, June 1997, pp. 163–178, https://doi.org/10.1080/00048409712347771.

Malfatti, Federica I. "Understanding phenomena: From social to collective?" *Philosophical Issues*, vol. 32, no. 1, Oct. 2022, pp. 253–267, https://doi.org/10.1111/phis.12216.

Malmgren, Anna-Sara. "Is there a priori knowledge by testimony?" *The Philosophical Review*, vol. 115, no. 2, 1 Apr. 2006, pp. 199–241, https://doi.org/10.1215/00318108-2005-015.

Martini, Carlo. "The Epistemology of Expertise." *The Routledge Handbook of Social Epistemology*, edited by Miranda Fricker et al., 2020, pp. 115–122.

Matarazzo, Andrea, and Riccardo Torlone. "A Survey on Large Language Models with Some Insights on Their Capabilities and Limitations." *arXiv.Org*, 9 Feb. 2025, arxiv.org/abs/2501.04040.

Meng, Xiangbin, et al. "The application of large language models in medicine: A scoping review." *iScience*, vol. 27, no. 5, May 2024, p. 109713, https://doi.org/10.1016/j.isci.2024.109713.

Miller, Boaz. "The Social Epistemology of Consensus and Dissent." *The Routledge Handbook of Social Epistemology*, edited by Miranda Fricker et al., Routledge, 2019, pp. 230–239.

Miller, Boaz. "When is consensus knowledge based? distinguishing shared knowledge from mere agreement." *Synthese*, vol. 190, no. 7, 4 Dec. 2012, pp. 1293–1316, https://doi.org/10.1007/s11229-012-0225-5.

Mirza, Adrian, et al. "Are Large Language Models Superhuman Chemists?" *arXiv.Org*, 1 Nov. 2024, arxiv.org/abs/2404.01475.

Mitchell, Melanie, and David C. Krakauer. "The Debate over Understanding in Ai's Large Language Models." *arXiv.Org*, 10 Feb. 2023, arxiv.org/abs/2210.13966.

Mizumoto, Masaharu, et al. "The Credit View and AI Testimony: A Cross-Cultural Epistemological Study of Human and AI Testimony." *Artificial Intelligence and the Future of Human Relations: Eastern and Western Perspectives*, edited by Yanto Chandra and Ruiping Fan, Springer Nature Singapore, 2025, pp. 273–298.

Omar, Mahmud, et al. "Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support." *Communications Medicine*, vol. 5, no. 1, 2 Aug. 2025, https://doi.org/10.1038/s43856-025-01021-3.

Oreskes, Naomi. "The Scientific Consensus on Climate Change: How Do We Know We 're Not Wrong?" *Climate Change: What It Means for Us, Our Children, and Our Grandchildren*, edited by Joseph F. DiMento F. DiMento and Pamela Doughman, MIT Press, 2007, pp. 65–99.

O'Brien, Dan. *Hume on Testimony*. Routledge, 2024.

O'Brien, Dan. "Humeanism and the epistemology of testimony." *Synthese*, vol. 199, no. 1–2, 22 Oct. 2020, pp. 2647–2669, https://doi.org/10.1007/s11229-020-02905-8.

Potochnik, Angela. *Idealization and the Aims of Science*. The University of Chicago Press, 2017.

Pritchard, Duncan. "A defence of quasi-reductionism in the epistemology of testimony." *Philosophica*, vol. 78, no. 2, 2 Jan. 2006, https://doi.org/10.21825/philosophica.82189.

Putnam, Hilary. "Minds and Machines." *Dimensions Of Mind: A Symposium*, edited by Sidney Hook, NYU Press, New York City, NY, 1960, pp. 138–164.

Razeghi, Yasaman, et al. "Impact of Pretraining Term Frequencies on Few-Shot Reasoning." *arXiv.Org*, 24 May 2022, arxiv.org/abs/2202.07206.

Reid, Thomas. *Thomas Reid's Inquiry and Essays*. Edited by Ronald E. Beanblossom and Keith Lehrer, Hackett Pub, 1983.

Resnik, Philip. "Large Language Models Are Biased Because They Are Large Language Models." *arXiv.Org*, 13 Mar. 2025, arxiv.org/abs/2406.13138.

Rice, Collin. *Leveraging Distortions: Explanation, Idealization, and Universality in Science*. The MIT Press, 2021.

Rohwer, Yasha, and Collin Rice. "Hypothetical pattern idealization and explanatory models." *Philosophy of Science*, vol. 80, no. 3, July 2013, pp. 334–355, https://doi.org/10.1086/671399.

Rolin, Kristina. "Gender and Trust in Science." *Hypatia*, vol. 17, no. 4, 2002, pp. 95–118, https://doi.org/10.1111/j.1527-2001.2002.tb01075.x.

Russell, Stuart J., and Peter Norvig. *Artificial Intelligence*. Pearson Education, 2009.

Salmon, Wesley C. *Scientific Explanation and the Causal Structure of the World*. Princeton University Press, 1984.

Searle, John. "Reductionism and the Irreducibility of Consciousness." *The Nature of Consciousness: Philosophical Debates*, edited by Owen Flanagan et al., MIT Press, 1997.

Searle, John R. *Minds, Brains, and Programs*. Cambridge University Press, 1980.

Sparrow, Robert, and Gene Flenady. "The testimony gap: Machines and reasons." *Minds and Machines*, vol. 35, no. 1, 24 Feb. 2025, https://doi.org/10.1007/s11023-025-09712-5.

Stegenga, Jacob. “Three criteria for Consensus Conferences.” *Foundations of Science*, vol. 21, no. 1, 17 Oct. 2014, pp. 35–49, https://doi.org/10.1007/s10699-014-9374-y.

Strevens, Michael. *Depth: An Account of Scientific Explanation*. Harvard University Press, 2008.

Strevens, Michael. “Grasp and scientific understanding: A recognition account.” *Philosophical Studies*, vol. 181, no. 4, 20 Mar. 2024, pp. 741–762, https://doi.org/10.1007/s11098-024-02121-x.

Strevens, Michael. “No understanding without explanation.” *Studies in History and Philosophy of Science Part A*, vol. 44, no. 3, Sept. 2013, pp. 510–515, https://doi.org/10.1016/j.shpsa.2012.12.005.

Sullivan, Emily, and Kareem Khalifa. “Idealizations and understanding: Much ado about nothing?” *Australasian Journal of Philosophy*, vol. 97, no. 4, 22 Jan. 2019, pp. 673–689, https://doi.org/10.1080/00048402.2018.1564337.

Traiger, Saul. “Experience and testimony in Hume’s philosophy.” *Episteme*, vol. 7, no. 1, Feb. 2010, pp. 42–57, https://doi.org/10.3366/e174236000900080x.

Traiger, Saul. “Humean testimony.” *Pacific Philosophical Quarterly*, vol. 74, no. 2, June 1993, pp. 135–149, https://doi.org/10.1111/j.1468-0114.1993.tb00355.x.

Tucker, Aviezer. “The epistemic significance of consensus.” *Inquiry*, vol. 46, no. 4, Dec. 2003, pp. 501–521, https://doi.org/10.1080/00201740310003388.

Turing, A. M. “Computing Machinery and intelligence.” *Mind*, LIX, no. 236, 1 Oct. 1950, pp. 433–460, https://doi.org/10.1093/mind/lix.236.433.

Van Cleve, James. “Reid on the Credit of Human Testimony.” *The Epistemology of Testimony*, edited by Jennifer Lackey and Ernest Sosa, Clarendon Press, 2006, pp. 50–75.

Varshney, Neeraj, et al. “A Stitch in Time Saves Nine: Detecting and Mitigating Hallucinations of LLMs by Validating Low-Confidence Generation.” *arXiv.Org*, 12 Aug. 2023, arxiv.org/abs/2307.03987.

Wei, Jason, et al. “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.” *arXiv.Org*, 10 Jan. 2023, arxiv.org/abs/2201.11903.

Wilkenfeld, Daniel A. “Understanding as representation manipulability.” *Synthese*, vol. 190, no. 6, 13 Dec. 2013, pp. 997–1016, https://doi.org/10.1007/s11229-011-0055-x.

Williams, Sean, and James Huckle. “Easy Problems That LLMs Get Wrong.” *arXiv.Org*, 1 June 2024, arxiv.org/abs/2405.19616.

Winston, P.H. *Artificial Intelligence*. Addison-Wesley, 1992.

Xu, Ziwei, et al. “Hallucination Is Inevitable: An Innate Limitation of Large Language Models.” *arXiv.Org*, 13 Feb. 2025, arxiv.org/abs/2401.11817?utm_source=chatgpt.com.

Zagzebski, Linda Trinkaus. *Epistemic Authority: a Theory of Trust, Authority, and Autonomy in Belief* . Oxford University Press, 2012.

Zagzebski, Linda. “Epistemic Authority and its critics.” *European Journal for Philosophy of Religion*, vol. 6, no. 4, 22 Dec. 2014, pp. 169–187, https://doi.org/10.24204/ejpr.v6i4.153.

Zagzebski, Linda. *On Epistemology*. Wadsworth Publishing Company, 2008.

Zagzebski, Linda. “Replies to Christoph Jäger and Elizabeth Fricker.” *Episteme*, vol. 13, no. 2, 7 Sept. 2016, pp. 187–194, https://doi.org/10.1017/epi.2015.39.

Zhang, Shengyu, et al. “Instruction Tuning for Large Language Models: A Survey.” *arXiv.Org*, 6 Oct. 2025, arxiv.org/abs/2308.10792.