Smart, P. R., & Clowes, R. W. (in press) The Gift of Language: Large Language Models and the Extended Mind. In V. Santos & P. Castro (Eds.), *Advances in Philosophy of Artificial Intelligence*. Ethics Press, Bradford, UK.

The Gift of Language: Large Language Models and the Extended Mind

Paul Smart and Robert W. Clowes

Abstract

Proponents of the extended mind insist that human states and cognitive processes can, at times, include non-biological resources that lie external to the bodily boundaries. In the present chapter, we apply this idea to large language models (LLMs), suggesting that some LLMs exist as extended cognitive (or computational) systems. We focus in particular on LLMs that exploit retrieval-augmented generation (RAG) techniques and online computational tools, proposing that these systems constitute extended architectures whose capabilities are realized, in part, by external structures. Drawing on parallels with classic cases of human cognitive extension, we argue that extended LLMs raise important philosophical questions concerning the nature and limits of artificial intelligence. We also draw attention to the role of language in supporting the emergence of extended systems. LLMs, it is suggested, are particularly well placed to benefit from their immersion in an online ecology that is home to a sizable chunk of human symbolic and technological culture.

> *"It's not a weapon. It's a gift. The weapon is their language. They gave it all to us."*
>
> – *Arrival* (2016), directed by Denis Villeneuve

Introduction

In the 2016 movie *Arrival*, a team of scientists race to decipher an intricate system of circular symbols used by 7-limbed alien visitors, called Heptapods.[1] As Dr. Louise Banks immerses herself in the alien language something remarkable happens: she begins to see glimpses of her future life. The alien language, it is revealed, extends the user's cognition in time allowing present actions to be shaped by future events. At one point in the movie, Louise 'remembers' that she will eventually publish a book on the Heptapod language. This enables her to complete her understanding of the alien script, for she now has the same relationship to her future writing as she does her past writing. Evidently, the

[1] The movie is an adaptation of a science fiction novella by American writer Ted Chiang (2004).

alien language is a very powerful tool. It turns out that the alien language is a "gift." The Heptapods, knowing they would have to rely on humans in the future, visited Earth with the intention of offering humans a particularly potent form of cognitive technology.

The movie's preoccupation with language resonates with a recurring theme in the philosophical and cognitive scientific literature. This relates to the cognitive transformational effects of language. As with the Heptapod language, a facility with human language has been seen to engender a range of cognitive benefits that go beyond its merely communicative role. Of particular interest is the idea that language functions as a sort of tool or technology that enables biological brains to tackle an otherwise intractable class of computational problems (Clark, 1997, 1998). According to this view, language is not just one of the hallmark features of human intelligence, it is a central feature of the human cognitive architecture: what we call human cognition is simply a form of language-augmented cognition (see Lupyan, 2016).

Claims regarding the cognitive transformational role of language are especially prominent in discussions of the extended mind, where language is seen not merely as a vehicle for expressing thought, but as a means of reshaping the very structure and dynamics of human cognition.[2] According to proponents of active externalism, the machinery of the mind (the mechanisms responsible for human mental states and cognitive processes) can sometimes extend beyond the borders of skin and skull to include resources in the local environment (Clark, 2008; Clark & Chalmers, 1998). Humans, it is suggested, are particularly adept at creating and exploiting these extended mechanisms, enabling them to assimilate worldly resources deep into their cognitive routines (Clark, 2003). Language then works in the manner of a cognitive amplifier—or "computational transformer" (see Clark, 1998)—yielding opportunities for cognitive extension that lie well beyond the reach of other terrestrial critters. The process of writing, for example, enables us to represent our thoughts in an external medium, and it is via this external medium that we are able to inspect our own thoughts via a perceptual route (Clark, 2006). Externalization also yields other benefits. We are, for example, able to manipulate the bio-external tokenings of our

[2] By way of a terminological note, we regard extended cognition and the extended mind as two strands of active externalist theorising, with the former term (extended cognition) referring to cognitive scientific kinds (e.g., memory) and the latter (extended mind) referring to folk psychological kinds (e.g., belief). In the present chapter, we will use the term "cognitive extension" to refer to both extended cognition and the extended mind.

own thoughts in a way that may be difficult or impossible to achieve with the bare biological brain. In short, we can use our actions to manipulate words, presenting the brain with all manner of new configurations and juxtapositions, some of which may limn the path towards previously unreachable ideas. Language, on this view, is not merely a tool for representing our thoughts, it opens the door to a radically different way of thinking. Courtesy of language, we are able to exploit information processing loops that span the brain, the body, and the world, transforming the process of thinking from a purely brain-based (neurally-realized) process into an extended (world-involving) process. If active externalists are right, it is this linguistically-amplified capacity for cognitive extension that holds the key to human cognitive success. Our human brains are undoubtedly different from the brains of other species, but they are not all that different. The cognitive differences, by contrast, are vast. There is a gaping chasm between our intelligence and the intelligence of other species—a cognitive divide that makes us somewhat unique in the natural order. Precisely how (or why) we made this transition remains unclear, but it is hard to overlook the cognitively-empowering role of language.

While our language-laden minds may be unique in the natural order, we are no longer the only entities to possess a basic facility with language. Recent research has given rise to Large Language Models (LLMs)—a new breed of intelligent machines that exhibit a remarkable proficiency with human language. The impact of the new (technological) arrivals is perhaps not on the scale of an alien visitation, but it is nevertheless significant. Systems like ChatGPT and Gemini have captured the popular imagination, sparking widespread debates about the future of Artificial Intelligence (AI) and its impacts on society.

From an active externalist perspective, there are at least two ways that we might approach LLMs. The first treats LLMs as the potential targets of cognitive incorporation—as novel technological resources that could be assimilated into human cognitive routines, functioning perhaps as the technological ingredients of a new sort of extended mind. But in addition to understanding LLMs as the potential object of extension—as a resource to be incorporated into human cognitive routines—we can also understand LLMs as the potential subject of extension—as an entity that benefits from the operation of extended information processing circuits. This is the approach we pursue in the present chapter. In particular, we explore the extent to which LLMs can be understood as extended systems, drawing on research into extended cognition and the extended mind. As we will see, language plays a particularly important role in the formation of these extended systems (what we dub extended LLMs). In our own species, language opens the door to cognitively-empowering forms of engagement with a wider environment, enabling us to form extended systems whose cognitive/computational capacities far outstrip those of the biological brain. Much the same, we suggest, applies to LLMs. Language is thus revealed as a transformative technology, one whose contributions to machine intelligence may be just as important as they are to our own species-specific form of cognitive success.

Retrieval-Augmented Generation

LLMs are AI systems that specialize in language-related tasks, such as text generation, translation, summarization, and question-answering. As with many contemporary AI systems, LLMs are driven by deep (multi-layer) neural networks that are trained on large datasets. These neural networks come in two basic varieties: transformers and recurrent neural networks. While both these architectures have been used to implement LLMs, much of the contemporary interest in LLMs stems from research into transformer networks. A key innovation relates to the introduction of a self-attention mechanism that weights and integrates information from different positions in an input sequence (Vaswani et al., 2017). This enables LLMs to determine the influence that earlier words in an input sequence have on later words, allowing for the representation of long-range dependencies. The overarching objective of the transformer is to perform a simple predictive task. In short, transformers generate text by making probabilistic predictions about the next word in a sequence, using both the initial input and any previously generated words as context for the predictive effort.[3] While this predictive task may not sound like a recipe for success, especially when one considers the complexity of human natural language, LLMs have exhibited impressive performance on a range of language processing tasks. They are also capable of engaging in conversational interactions, using earlier exchanges as the basis for future responses.

[3] Technically, LLMs trade in tokens rather than words. A token is a numerical representation of a particular unit of text, such as a word, sub-word, or individual character. Many LLMs use a tokenization scheme called Byte Pair Encoding (BPE), which yields tokens that are roughly equivalent to 3/4 of an English word.

Much of the success of LLMs stems from the use of powerful machine learning techniques that progressively adjust the weights (or parameters) of an artificial neural network to encode the statistical regularities of human natural language. Although this approach has proved remarkably effective, it does lead to a number of challenges. These include a tendency to produce factually incorrect responses (often referred to as "hallucinations"), a dependence on outdated or missing information, and the use of non-transparent, non-traceable reasoning processes that make it difficult to understand or verify the origins of specific outputs. In response to these challenges, researchers have explored the merits of a technique called Retrieval-Augmented Generation (RAG) (Gao et al., 2024; Ram et al., 2023). A central feature of RAG is the inclusion of a retrieval system that allows external information to be incorporated into the LLM's generative routines. There are, in fact, many variations of this retrieval-oriented technique, with multiple types of RAG-based LLMs (or RAG models) emerging from recent research (see Gao et al., 2024, for a recent view). For present purposes, however, we will limit our attention to a basic RAG model consisting of a single retrieval loop. This is what is sometimes called a single-cycle (or single-time) RAG model.

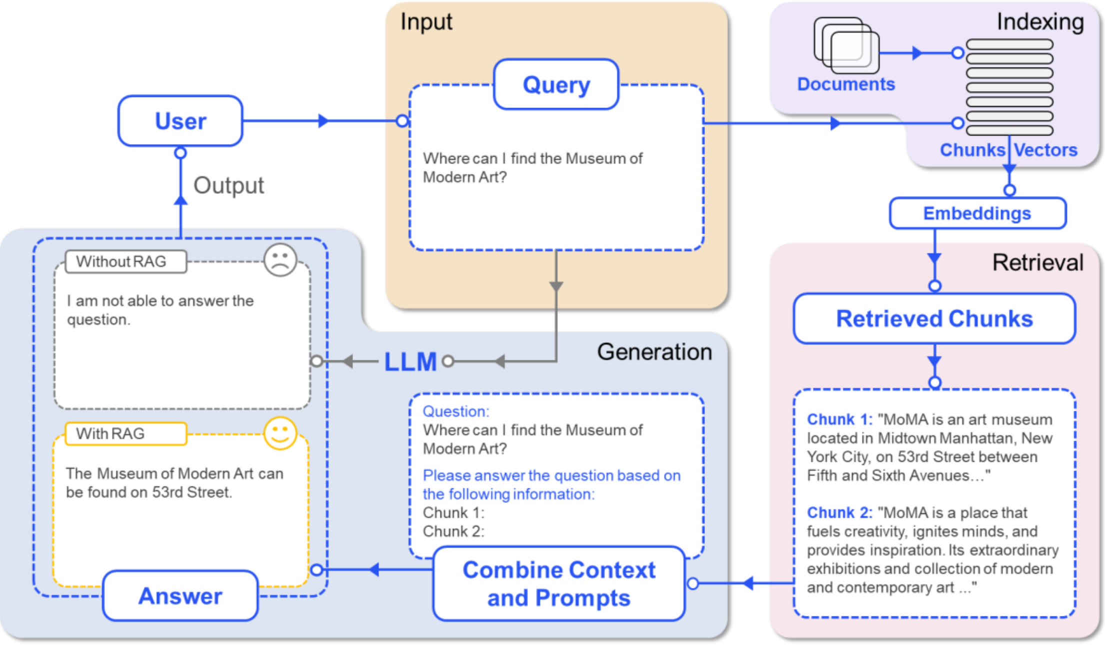
Input
Query
Where can I find the Museum of Modern Art?
Indexing
Documents
Chunks
Vectors
User
Output
Embeddings
Retrieval
Retrieved Chunks
Without RAG
I am not able to answer the question.
LLM
Generation
Question:
Where can I find the Museum of Modern Art?
Please answer the question based on the following information:
Chunk 1:
Chunk 2:
With RAG
The Museum of Modern Art can be found on 53rd Street.
Chunk 1: "MoMA is an art museum located in Midtown Manhattan, New York City, on 53rd Street between Fifth and Sixth Avenues…"
Chunk 2: "MoMA is a place that fuels creativity, ignites minds, and provides inspiration. Its extraordinary exhibitions and collection of modern and contemporary art …"
Answer
Combine Context and Prompts

Figure 1. *Example of a single-cycle RAG model (adapted from Gao et al., 2024). A user query functions as a retrieval cue, prompting the recall of query-relevant information (text chunks) from an external datastore. The retrieved information is then used to contextualize the LLM's response.*

Figure 1 shows the basic structure of a RAG model. For the purposes of illustration, we will assume that the LLM in Figure 1 is a standard transformer network that has been trained on large amounts of textual data. This is what is typically referred to as the training data or training corpus. Courtesy of the training regime, the information contained in the training data will become incorporated into the LLM's internal architecture. Specifically, the information will be encoded in the parameters (or weights) of the transformer network, yielding what is commonly known as *parametric knowledge*. This body of parametric knowledge defines the *internal memory* of the LLM. When the LLM is presented with a particular task, it relies on its internal memory to deliver (or generate) the correct response. If, for example, the LLM in Figure 1 had been trained on information about The Museum of Modern Art (MoMA) (in New York), it would be able to answer questions about MoMA using its internal memory. In response to a question about the location of MoMA, the LLM might respond by saying that the museum was on 53rd Street. This is what we would observe if the LLM had been exposed to information about MoMA as part of its training regime. Suppose, however, that the LLM in Figure 1 has not been trained on any information pertaining to MoMA—such information, let's suppose, was omitted from the training data. In this case, the LLM would have no opportunity to assimilate MoMA-related information internal memory. Accordingly, when the LLM is prompted with a question about the location of MoMA, it responds by saying that it is unable to answer the user's query. This is the state-of-affairs depicted in the upper output in Figure 1 (i.e., the response labelled "Without RAG").

How might we remedy this sort of epistemic deficit? One possibility would be to retrain the LLM on the missing information, adding the MoMA-related information to a training dataset and then embarking on an additional round of machine learning. RAG, however, relies on a different approach. Rather than resort to retraining, RAG uses a retrieval loop to access information that would otherwise be added to the training data. Here, the user's query (e.g., "What is the address of MoMA?") serves as a retrieval cue, which is used to access information from an externally-situated datastore (typically, a vector database). This external datastore is what is referred to as the LLM's *external memory*. Unlike the contents of internal memory, which are encoded in the LLM's parameters, the contents of external memory are represented by segments of text, known as text chunks. Given their non-parametric nature, it is common for these chunks to be referred to as *non-parametric knowledge*. The retrieval process works by scoring text chunks in terms of their (semantic) similarity to retrieval cues, with the highest scoring chunks being returned to the LLM. The retrieved information is then used to contextualize the LLM's response, with each response being conditioned on a combination of both the user's query and the retrieved information. In Figure 1, for example, the retrieval loop returns two text chunks that are deemed relevant to the user's query. These text chunks are added to the LLM's context window, where they work to influence the LLM's response.

Relative to Figure 1, it should be clear how RAG addresses the aforementioned problems of hallucinatory responses, outdated or absent knowledge, and non-traceable reasoning processes. In respect of hallucinatory responses, the addition of retrieved information minimizes the model's tendency to hallucinate by conditioning responses on factually correct information. In Figure 1, for example, the answer to the user's query is explicitly represented in the retrieved information. This dramatically increases the chances of the model producing the correct response.

The presence of external memory also resolves the problem of outdated or absent knowledge. As we have seen, external memory can store information that was not available during the LLM's original training data. In addition, it is far easier to change the contents of external memory than it is the contents of internal memory. The contents of external memory, recall, are represented using the familiar (symbolic) resources of human natural language, and this makes it relatively easy for a language-enabled entity such as ourselves (or perhaps an LLM) to edit the contents of external memory.

Finally, RAG yields a degree of transparency (or at least greater traceability) when it comes to understanding model outputs. In Figure 1, for example, it is easy to see what information is informing the model's response, and (assuming that the contents of external memory are associated with provenance information) these responses can be traced to particular sources.

Now that we have a better understanding of RAG, let us turn our attention to the links with active externalism. One such link is suggested by the example in Figure 1, for the location of MoMA features as part of a classic thought experiment that was introduced by Clark and Chalmers (1998) to motivate claims for the extended mind. Clark and Chalmers (1998) ask us to imagine two individuals, both of whom are located in New York city. One individual—Inga—is a neurologically intact individual. When Inga wants to visit MoMA, she retrieves the location of MoMA from bio-memory and heads off in the direction of 53rd Street. The second individual—Otto—has a mild form of dementia. Unlike Inga, Otto cannot rely on bio-memory to inform his museum-going behaviours. Instead, when Otto wants to go to MoMA, he retrieves information from an externally-situated resource, namely a conventional notebook.

Despite the differences between Otto and Inga, Clark and Chalmers (1998) suggest that the familiar folk psychological strategy of explaining behaviour courtesy of the ascription of mental states (e.g., states of dispositional belief) is one that can be applied to both individuals. Thus, just as Inga can be said to believe that MoMA is on 53rd Street (even before the information was retrieved from bio-memory), so Otto can be said to believe that MoMA is on 53rd Street (even before he consulted the notebook). In one sense, then, Otto and Inga are highly similar: both individuals can be said to possess the dispositional belief that MoMA is located on 53rd Street. At the same time, however, it should be clear that these beliefs cannot be rooted in the same physical processes. While Inga's belief is a standard non-extended belief, Otto's belief is an extended belief. Otto's belief is extended in the sense that the supervenience base for the belief (the things that make it true that Otto possesses the belief) must rely on forces and factors that lie external to the biological individual we recognize as Otto. Consider, for instance, that if Otto were to lose his notebook, then he would no longer be able to determine the location of MoMA. Accordingly, it would make no sense to credit Otto with the belief that MoMA is on 53rd Street, for if Otto were to believe that MoMA is on 53rd Street, then this is where he should go given his desire to visit MoMA.

Note how the contrast between Otto and Inga resembles that between RAG models and standard LLMs. Inga resembles a standard LLM that does not rely on RAG. When this LLM is prompted to recall the location of MoMA, it relies on its internal memory to produce the correct response. The LLM's internal memory thus maps to Inga's bio-memory. This mapping is all the more compelling given the way information is encoded in internal memory: just as the LLM's internal memory is encoded in the parameters of its artificial neural network, so Inga's bio-memories are encoded in the parameters of her bio-neural network.

Now consider Otto. Otto resembles an LLM that relies on RAG (i.e., a RAG model). Here, the notebook corresponds to the LLM's external memory. When Otto wishes to go to MoMA, he consults the notebook, retrieves the relevant information, and heads off in the direction of 53rd Street. Similarly, when a RAG model is prompted to recall the location of MoMA, it triggers a retrieval loop that returns information from a datastore that lies external to the LLM. The retrieved information is then factored into the LLM's response. Otto's use of the notebook thus parallels the LLM's use of an external (extra-systemic) datastore. In both cases, we encounter the presence of a retrieval loop that accesses information contained in some external resource (a notebook or vector database). This information is then factored into some sort of overt response, such as heading off to 53rd Street (in the case of Otto) or generating a response to a user query (in the case of the LLM).

There is a further parallel between the two forms of external memory. Both the notebook and the vector database consist of textual information. Accordingly, there is no reason to assume that the two forms of external memory are trading in radically different forms of encoding. The inscriptions in Otto's notebook are a form of textual encoding, in the sense that they represent the location of MoMA using the familiar resources of natural language. But the same is also true of the information stored in a vector database. As a means of highlighting the parallels between the two forms of external memory, let's make a couple of adjustments to the original Otto case. Firstly, let's substitute the notebook for a smartphone connected to an online vector database. Now, when Otto wants to go to MoMA, he uses his smartphone to retrieve the relevant information from the very same database as that used by a RAG model. There is no reason to think that this sort of technological upgrade materially alters the philosophical import of the original Otto case. Insofar as Otto believes (in a dispositional sense) that MoMA is on 53rd Street when coupled to the notebook, he will continue to harbour this belief when coupled to the vector database. The second adjustment relates to Otto's behaviour. Instead of Otto actually going to MoMA, let's assume that Otto is asked about the location of MoMA. In response to the question "Where can I find the Museum of Modern Art?" Otto responds that MoMA can be found on 53rd Street (using his smartphone to retrieve the relevant information). This situation, it should be clear, resembles that depicted in Figure 1. Indeed, there is no reason to think that Otto's response is informed by a radically different sort of retrieval process to that used by a RAG model. In fact, Otto and the RAG model could be relying on computationally identical retrieval processes to return query-relevant information from the *very same* vector database. Assuming the responses of Otto and the RAG model are the same in this scenario—they both report that MoMA is on 53rd street— is there any reason to regard the two cases as radically different: to view extension-related claims as plausible in the case of Otto, but utterly implausible in the case of the RAG model?

To our mind, the answer to this question is "no." Given the parallels with the Otto/Inga case, we suggest that RAG models ought to be understood along active externalist lines. That is to say, we suggest that RAG models ought to be understood as extended LLMs. Just to be clear, this does not mean that we are obliged to regard RAG models as extended (artificial) *minds*. While the appeal to mentalistic notions (e.g., states of dispositional belief) is largely uncontroversial in the Otto/Inga case, this is much less so when it comes to LLMs. It is, in particular, controversial to suggest that LLMs ought to be understood as the bearers of genuine mental states (e.g., Shanahan, 2024), and this is so regardless of the (even more controversial) claim that some of those states ought to be understood as extended.

For present purposes, we will seek to bypass these concerns by adopting an approach that distinguishes extension-related claims from those pertaining to issues of cognitive/mental status. In short, we suggest that the extended status of LLMs can be understood without regard to philosophical disputes pertaining to the "mark of the cognitive" (or the mark of the mental) (see Adams & Garrison, 2013). The basis for this claim relates to the non-cognitive nature of some of the phenomena that have been discussed in the active externalist literature. One example relates to the swimming performances of bluefin tuna. These performances benefit from the exploitation of hydrodynamic phenomena (e.g., self-created vortices and pressure gradients) that work as the literal constituents of extended propulsive mechanisms (see Gillett et al., 2022). While there is undoubtedly a degree of intelligence in play here, it would be difficult to lump propulsive (or locomotor) processes in the same category as cognitive (or mental) processes. In general, swimming processes are best understood as the expression of a physical ability (or physical capacity). Cognitive processes, by contrast, are best understood as the expression of a cognitive ability (or cognitive capacity). The question is whether this difference—the distinction between physical and cognitive processes—has any real bearing on our understanding of what it is that underwrites the *extended* status of the tuna's natatorial performances? Suppose that we abstract away from the details of the various cases in the active externalist literature, directing our attention to the more general realm of mechanisms, processes, and dispositions. Then, at this more general level, we ask ourselves what it is that unites seemingly disparate cases of extension, including those of the cognitive and non-cognitive variety. The answer to this question, we suggest, turns on the way in which a particular agent is credited with the possession of dispositional properties that are subject to extended or wide realization, meaning that the mechanisms responsible for the manifestation of the disposition include resources that lie beyond the physical borders of the agent to which dispositional properties are ascribed (for more on this, see Smart, 2024). The swimming performances of the tuna are thus deemed to be extended because the tuna is credited with the possession of a dispositional property (e.g., an ability to swim at a certain speed), but the manifestation (or exercise) of this dispositional property depends on forces and factors that lie external to the tuna. Likewise, what motivates the appeal to extension in the Otto case is the fact that we ascribe a dispositional property to Otto (i.e., the dispositional belief that MoMA is on 53rd Street), but the manifestation (or exercise) of this dispositional property depends on forces and factors that lie external to the biological individual we recognize as Otto. In both these cases, the appeal to extension is tied to the way the manifestation of some dispositional property depends on forces and factors that lie external to the disposition bearer. Dispositions are extended, we suggest, when the mechanisms responsible for the realization of processes reflecting the manifestation of a disposition include components that lie external to the disposition bearer (i.e., the entity to which the dispositions are ascribed).

This more general approach to understanding extension-related claims is readily applicable to LLMs. Indeed, all that is required for an LLM to qualify as extended is that we credit an LLM with a dispositional property that—on closer inspection—is revealed to rely on forces and factors that lie external to the LLM. The particular nature of this dispositional property (e.g., its status as a mental or cognitive kind) is undoubtedly important when it comes to understanding the LLM's status as an extended *mind*; but we can nevertheless understand the extended status of LLMs without agreement on this issue. Consider, for example, that whatever else we might say about LLMs, they are clearly systems that are the bearers of certain dispositions (e.g., a capacity to generate responses to certain types of queries).[4] All that is required for an LLM to count as extended, according to the foregoing account, is that one or more of these dispositional properties qualifies as an extended dispositional property, meaning that the property must supervene on forces and factors that lie external to the LLM.

For the most part, the claim that RAG models can satisfy this sort of criterion ought to be unproblematic. An LLM that correctly reports the location of MoMA clearly has a capacity of one sort of another. In the standard (non-extended) case, the LLM's response is informed by the information stored in internal memory, and the capacity is thus one that supervenes on forces and factors that lie *internal* to the LLM. In the case of RAG models, however, the correct response occurs as the result of a retrieval loop that returns information from an external store. Accordingly, the capacity cannot be one that supervenes solely on forces and factors that lie internal to the LLM; instead, we must acknowledge the contribution of the wider informational and computational environment in which the LLM is situated. To see this, we need only consider what would happen if the LLM were to be prevented from accessing its external memory. In this case, the LLM would be incapable of responding to the user's query. It would either fail to produce the correct response or report that it was incapable of answering the user's query. There would, as such, be nothing to substantiate the claim that the LLM 'believed' *X*, 'knew' *Y*, or possessed the capacity *Z*. Note that this sudden shift in competence cannot be anything to do with the LLM, for the removal of external memory does not entail that the LLM has, itself, changed. It is thus the operation of the retrieval loop, coupled with the reliable presence of the external memory store, that sustains the presence of the relevant dispositional property. And it is courtesy of such dispositional characterizations that we are able to latch onto coarse-grained patterns of behaviour that paper over the distinction between internal and external memory. An LLM that relies on external memory may be just as capable of answering questions about MoMA as one that relies on internal memory. Indeed, barring some indication that an LLM is relying on an external resource, a human user may be unaware of the presence of the retrieval loop or the distinction between internal and external memory. As with the Otto/Inga case, what matters here is not so

[4] Given the status of LLMs as AI systems, these dispositional properties will qualify as what are dubbed artificial dispositions. For more on artificial dispositions, see Bauer and Marmodoro (2024).

much the details of the retrieval process, but more the way in which different bodies of information are poised to influence overt behaviour. In this sense, there may be little to distinguish between extended and non-extended LLMs. An extended LLM may succeed in producing the correct response just as surely as Otto navigates his way to MoMA.

Extended Memory: The Next Generation

Recent years have seen an explosion of interest in RAG techniques. This has led to the emergence of multiple types of RAG model (Gao et al., 2024). In addition to the single-cycle model discussed in the previous section, researchers have increasingly turned their attention to the capabilities of so-called multi-cycle models. These models make multiple calls to external memory, often adapting the retrieval loop in response to shifting epistemic demands. There has also been research into the ways that LLMs can control the retrieval process via the adaptive selection (or generation) of retrieval cues. In essence, recent research has sought to equip LLMs with something akin to a metacognitive capacity, one that enables LLMs to decide when the retrieval loop should be called and (crucially) what retrieval cues should be used for the retrieval process (Asai et al., 2024; Zhou et al., 2024).

A nice example of recent work in this area is provided by Jiang et al. (2023). They describe a technique called Forward-Looking Active REtrieval augmented generation (FLARE), which ties the invocation of the retrieval loop to the estimated quality of model responses (or generations). FLARE relies on an initial prediction of what the model would produce in the absence of the retrieval loop (i.e., in the absence of external memory). The quality of this response is then evaluated using token probabilities as a proxy for what the model knows courtesy of its internal memory. When token probabilities fall below a certain threshold value, the model issues a call to external memory, using its initial prediction as a retrieval cue. It then revises the initial prediction, conditioning its response on the retrieved information. This process continues throughout the generative process, with the evaluation/retrieval cycle being repeated after every sentential generation.

A key feature of FLARE thus relates to the contingent invocation of the retrieval loop. In effect, the LLM decides when to initiate the call to external memory based on an ongoing assessment of its capacity to produce a high-quality response to the user’s query.

This echoes the broadly metacognitive nature of the decision-making process that underlies the choice between internal and external strategies in biological cases of cognitive extension (Clark, 2015, 2024a). Clark (2024a), for example, draws on predictive processing accounts of brain function to propose a mechanism by which bio-external resources might be incorporated into extended cognitive routines. Here, the 'choice' between internal (non-extended) and external (extended) strategies occurs as part of an overarching imperative to bring about a preferred outcome. Suppose, for example, that your goal is to write an academic paper on the extended mind. At some point, you find yourself needing to refer to the details of the Otto/Inga case. You think that MoMA might be located on 53rd Street, but you aren't sure. Given that your goal (your preferred outcome) is to write a high-quality paper, you will be motivated to minimize this uncertainty. The upshot is that your generative (i.e., writing) activity is temporarily suspended in favour of an epistemic action that retrieves information from an external source. In this case, your choice between inner and outer strategies—the decision to rely on bio-memory or some bio-external resource—is informed by an ongoing assessment of what you already 'know' courtesy of internal bio-memory. It is then the resultant level of uncertainty (coupled with your goal of writing a high-quality paper) that motivates the call to external memory. A similar sort of dynamic can be found in the case of FLARE. Once again, there is a choice between inner and outer strategies (i.e., between internal and external memory), and the decision to rely on external memory is informed by an ongoing assessment of the quality of responses delivered by internal memory.

The effort to give LLMs greater control over the retrieval process is important, for us to see the LLM as an active agent that exhibits control over the timing and operation of extended circuits. This brings RAG models into closer alignment with many of the cases that have been discussed in the active externalist literature. Indeed, as far as we can tell, all cases of cognitive extension feature an individual agent that triggers the instantiation of an extended circuit courtesy of their own actions. In the Otto/Inga case, for example, it is the biological individual known as Otto who triggers the retrieval loop. Likewise, in the case of the bluefin tuna, it is the biological individual (the fish) that exploits (and sometimes deliberately creates) the external pools of kinetic energy that help to propel it through its watery world. Thus, even in the non-cognitive cases (the cases where an extended routine does not qualify as cognitive in nature), we can nevertheless discern an intelligent entity at the heart of every extended circuit. Quite plausibly, the presence of this intelligent entity—or "cognitive core" (Clark, 2008, pp. 107–108)—is relevant to the way we credit individuals with the possession of extended dispositional properties, for the ascription of extended dispositional properties is likely tied to the presence of a more basic capacity to create and exploit extended circuits. It is precisely in this sense, we suggest, that the recent work to equip LLMs with greater control over the retrieval process. It is important not just because such capabilities promise to improve the overall quality and efficiency of LLM responses; it is also that by bringing the retrieval loop under the control of the LLM, we are perhaps more inclined to credit the LLM with dispositional properties that (upon closer inspection) turn out to rely on a wider web of material resources (see Smart et al., 2025, for a more detailed discussion of this issue).

Wideware

The distinctive feature of RAG models is the presence of a retrieval system that targets the information stored in an externally-situated datastore, typically a vector database. But datastores are not the only type of resource that can be accessed by LLMs. Nor is information retrieval the only point of engaging with such resources. In addition to specialized repositories of domain-specific knowledge, many LLMs are embedded in an online ecology that is replete with informational artefacts, computational services, and increasingly, other AI systems. From an active externalist perspective, these resources are the potential building blocks of extended cognitive and computational circuits. They are, to borrow a term from Clark (1999), the "wideware" of extended LLMs.

One form of wideware has become a standard feature of production LLMs such as ChatGPT and Gemini. This relates to the use of a search engine. Search engines are a familiar feature of the online ecology, one that we humans use on a regular basis. But a search engine can also be used by LLMs to access information that may not have been contained in the training data. This is important, because LLMs are trained on data that is typically out of date by the time the model is released to the public. Consider, for example, the GPT-4 Turbo model, which (at the time of writing) is the model that powers ChatGPT. This model was trained on a dataset containing information up to 1st December 2023, meaning that information about more recent events (e.g., those occurring in 2024) cannot form part of the model's internal memory. In principle, then, a question such as "Who won the 2024 U.S. presidential election?" should be out of reach of ChatGPT. Rather than being presented with the correct answer, we might expect the model to respond by saying that it doesn't know.

This is not what happens, of course. In response to the aforementioned question, ChatGPT reports the correct result: "Donald Trump." When pressed as to how it arrived at this answer, the model responds by saying "I used my web browsing tool, which allows me to search the live internet for up-to-date information." In this case, then, ChatGPT is working in the manner of a RAG model. Indeed, in some ways, online search emerges as just another form of RAG. The main difference is that the online environment is now being treated as a form of 'external memory', and the retrieval effort is being delegated to a search engine as opposed to the query engine of a conventional database.

In addition to search engines, other online tools have been the focus of recent research attention (see Mialon et al., 2023, for a recent review). Schick et al. (2023), for example, describe an LLM, called Toolformer, that learns how to interact with tools by generating content respecting the target tool's Application Programming Interface (API). An interesting feature of Toolformer is its capacity to interact with *multiple* tools. These include a question-answering system (based on a RAG model), a calculator, a calendar, a Wikipedia search tool, and a machine translation system. Note that two of these tools (the question-answering system and the machine translation system) qualify as AI systems, and one of them (the question-answering system) also qualifies as an LLM. This highlights the potential for complex forms of extension in which one AI system is assimilated into the cognitive/computational routines of another. There ought to be nothing particularly problematic about this idea. The proponents of active externalism already accept the possibility of AI systems being incorporated in human cognitive routines (Wheeler, 2019), so there ought to be no issue with the mere idea of AI systems being assimilated into an extended circuit. All that the notion of extended LLMs requires is that we allow such systems (perhaps the *very same systems*) to qualify as the constituents of extended circuits that are centred on a different sort of intelligent entity, namely, an LLM.

Another interesting feature of Toolformer relates to issues of tool selection. In the previous section, we saw how researchers are seeking to equip LLMs with something akin to a metacognitive capacity, enabling them to coordinate calls to extended memory with ongoing epistemic demands. In such cases, the focus was on a single ‘tool’ (i.e., the interaction with external memory), and the primary challenge was to determine when the retrieval loop should be invoked. The interaction with multiple tools poses additional challenges, requiring the LLM to decide between different tools based on the demands of the current task. If, for example, the LLM is presented with a mathematical problem, then there is little point in invoking a tool specialized for machine translation. Conversely, if the LLM is presented with a translation task, it will do little good to invoke the services of a calculator. The success of the LLM is thus tied to its ability to select the right tool for the job. In addition, the LLM needs to invoke the tool in a manner that meets the demands of the task and the constraints imposed by the tool’s programmatic interface (i.e., its API). In the case of Toolformer, for example, the model is required to “decide which APIs to call, when to call them, what arguments to pass, and how to best incorporate the results into future token prediction” (Schick et al., 2023, p. 68539). All this arguably requires a degree of intelligence, echoing one of the points raised in the previous section (recall the discussion of the ‘cognitive core’).

In recent years, there has been a burgeoning of research interest into tool integration. Indeed, issues of tool selection and tool integration are very much at the forefront of contemporary research into LLMs. The reason for this interest is no doubt obvious. While LLMs are good at some things, they are bad at others. LLMs tend to excel at language-related tasks, but they often fall short when it comes to tasks that require mathematical precision or domain-specific reasoning. While much has been made of these shortcomings in at least some parts of the philosophical literature (see Floridi & Chiriatti, 2020), it is far from clear that we can understand the capabilities of LLMs independently of their wider environment. There is, of course, no denying the limitations of the typical transformer network. But why judge the capabilities of an LLM independently of its capacity to interact with a wider suite of computational tools? After all, according to the proponents of active externalism, much of our human cognitive success is tied to our ability to factor bio-external resources deep into our cognitive processing routines, and it is courtesy of such forms of incorporation that we humans are able to surpass the limits of our basic bio-neural endowments (Clark, 1999, 2003). If this should be the case for *human* intelligence, then why assume that the limitations of the typical transformer network are particularly definitive of its cognitive horizons, especially once it is embedded in an artefactually-rich online environment? From an active externalist perspective, what matters is not so much the capacity of a neural network (a biological brain or transformer architecture) to perform a variety of cognitive tasks in the absence of a supportive technological and symbolic matrix, it is more the capacity of the network to exploit external resources in a way that obviates the appeal to an internally-grounded deficit. Much of the recent research into LLMs resonates with this idea. By enabling LLMs to exploit the resources of the online environment, we open the door to complex forms of cognitive and computational extension that may be no less empowering for LLMs than they are for ourselves. And just as with our own species-specific form of intelligence, it may be difficult to know where (in space and time) the cognitive story ends. Just as it can be hard to determine the limits of our own species-specific form of cognitive success (see Clark, 2024b), so it may be hard to determine the limits of the somewhat alien form of intelligence exhibited by LLMs.

The Gift

Our aim, thus far, has been to convince the reader that some LLMs ought to be understood along active externalist lines—as extended systems or extended LLMs. In one sense, however, there is nothing particularly remarkable about extended LLMs. Given the status of LLMs as AI systems, it follows that extended LLMs are an example of what has been dubbed extended AI (see Smart, 2018). There is, however, no reason to think that extended LLMs exhaust the possibilities for extended AI. Quite possibly, other forms of AI (e.g., a humanoid robot) could meet the criteria for cognitive/computational extension, and extended LLMs are just one example of a much broader class of systems. This raises a question about the significance of LLMs: Beyond their status as AI systems, what, if anything, makes extended LLMs deserving of particular philosophical and cognitive scientific attention?

The answer, we suggest, stems from the language-oriented nature of LLMs—the peculiar facility that LLMs have with linguistic and quasi-linguistic structures. While there is no reason to regard language as essential for the formation of extended systems, it is nevertheless a feature of many of the cases that have been discussed in the active externalist literature. To see this, we need only remind ourselves of the centrality of language to the classic extended mind case. Otto's interactions with the notebook serve as the basis for claims regarding the extended mind. But such forms of cognitive extension are contingent on Otto's ability to read the notebook's contents. Absent this (language-related) ability and the appeal to cognitive extension evaporates. Language is thus a central feature of the classic extended mind case. Indeed, it is Otto's facility with language that makes this case possible. Otto's linguaform abilities need not, by themselves, qualify as extended, but they nevertheless provide the foundation for Otto's extended mind.

Given the nature of the earlier parallels with the Otto case, it should be relatively clear that the forms of extension emerging in respect of RAG models are similarly dependent on a facility with language. RAG models, recall, feature a retrieval loop that returns *textual* content from an externally-situated database. As with Otto, the external database is only a candidate feature of an extended architecture if the LLM can factor the retrieved information into its generative routines. Absent this ability to, in effect, 'read' the information and there is no basis for the claim that the contents of external memory ought to be understood as on a par with the contents of internal memory. It would, for example, make no sense to insist that an LLM 'knows' whatever facts are contained in the external database if such facts are incapable of influencing the LLM's responses in the manner we typically expect of an epistemic state (e.g., the state of knowing the location of MoMA).

The same is true when we turn our attention to online search. As with RAG, it is the LLM's facility with language that enables it to benefit from online search—to factor search results into its own generative performances. In contrast to RAG, however, online search opens the door to a much more expansive body of information. This is important, for the online realm is home to a significant portion of our symbolic culture, including much of the knowledge distilled from centuries of intellectual activity. Such knowledge may or may not have been used to train an LLM, but once an LLM has been equipped with a capacity to interact with a search engine, it nevertheless possesses an ability to access this knowledge. In fact, from an active externalist perspective, the externally-situated knowledge already counts as 'part' of the knowledge base of the LLM, for the ascription of extended dispositional properties encourages us to disregard the differences between internal and external memory—to, in effect, regard these as a common pool of knowledge and information that drives the LLM's generative performances.

The LLM's facility with language is, of course, tied to the details of its training regime. But note how the capacity to interact with search engines alters the way we think about the training process. Rather than regard training as an attempt to deliver a fully-formed system with a fixed suite of epistemic capabilities, the training process is perhaps better understood as an attempt to prepare the LLM for its deployment in an online environment. Once an LLM has the capacity to interact with this environment, it can begin to assimilate external resources into its generative performances, yielding capabilities that we might not have expected to emerge given the nature of the training regime. This can quickly lead to confusion. Consider that as part of a recent presentation, one of the authors was asked to comment on the ethics of using online data to train LLMs. In particular, the question concerned the use of online data without the explicit approval or consent of the individual (or organization) who originally produced the data. This question is clearly important, but it tends to assume that the successes (and sometimes failures) of the LLM are predicated solely on the training data—that any information excluded from the training dataset could not play a productive role in informing the LLM's responses. By now, however, it should be clear that there is something wrong with this image. The mere fact that one has managed to prevent an online resource from being included in the training data for an LLM does not mean that the online resource could not be accessed and exploited by the LLM as part of its runtime operation. Indeed, from an extended perspective, the exclusion of the resource may turn out to be irrelevant, for once an LLM is able to access the resource via (e.g.) a search engine, the absence of the resource from the original dataset may have little bearing on the LLM's responses. In essence, the LLM may respond *as if* the resource had featured as part of its training regime all along.[5]

None of this is to detract from the importance of the training data or the ethical issues surrounding the use of such data. The point is more that an extended perspective changes the way we think about the training process, encouraging us to regard it as something more akin to learning to read. This step is, of course, crucial, but it is no more

[5] This may explain some of the confusion surrounding the use of content from major news organizations, such as the British Broadcasting Corporation (BBC). There is a difference, here, between the use of content to tailor an LLM's responses (e.g., via a search tool) and the use of content to train an LLM. If information has been available online for public consumption, then it is, in principle, poised to influence the performances of an LLM. The mere fact that an LLM tailors its responses with regard to such content does not mean that such content has been incorporated into internal memory. In such cases, the LLM is accessing online information just as a human would access such information to keep themselves abreast of current affairs (see https://www.bbc.co.uk/news/articles/cy7ndgylzzmo).

definitive of an LLM's cognitive and epistemic horizons than it is our own. It is more a stepping stone to something else—a way of building something with the potential to exploit the vast edifice of knowledge and information that our species has accumulated as the result of innovations such as the Internet, the growth of portable computing technologies, and the rise of social media.

A facility with language is also important when it comes to invoking tools. Consider, for example, the Toolformer system, which we discussed in the previous section. The aim here is not just to invoke an online tool, it is to ensure the tool is invoked in the *right* way. This means the LLM must generate content that respects the target tool's API. Typically, this involves the generation of programmatic instructions, although, in some cases, natural language may also be appropriate. (This is especially so if the target tool is, itself, a language-enabled entity, such as another LLM.) The basic point here is that a facility with language (of either the natural or artificial [programmatic] variety) is essential to the formation of extended LLMs, for it is only by generating linguaform content (in the form of prompts, queries, programmatic instructions, and so on) that an LLM is able to solicit support from a wider web of material resources. And therein lies the peculiar value and significance of language. For all the resources of the online realm were created by us or intended for us, and it is only because of our facility with language that we are able to derive any sort of cognitive or epistemic benefit from these resources. This is why language is important. Unlike the Heptapod language, our own (human) language does not "open time," but it does open the door to cognitively-empowering forms of extension, ones that are, perhaps, no less significant than those that drive our own species-specific cognitive performances.

Conclusion

Much of the philosophical debate surrounding the extended mind has been limited to the realm of human (or at least biological) intelligence. The primary point of contention in such debates is whether human mental states and cognitive processes are (at all times) realized by mechanisms that are solely confined to the intra-cranial (or, at least, intra-bodily) realm. According to the proponents of active externalism, even quite familiar cognitive and mental phenomena (e.g., the state of believing that MoMA is on 53rd Street) can, on occasion, be subject to extra-neural realization. In such cases, the mechanisms responsible for human mental states and cognitive processes reach beyond the biological borders of skin and skull, incorporating resources from the wider (extra-organismic) environment.

In the present chapter, we sought to apply this idea to LLMs, suggesting that some LLMs exist as extended LLMs. According to this view, an extended LLM is an LLM whose performances are, on occasion, subject to wide or extended realization. Just as human cognitive capacities are, at times, realised by information processing loops that extend beyond the borders of the biological brain, so the capacities of LLMs are, at times, realised by information processing loops that extend beyond the borders of the transformer network.

In our view, many contemporary LLMs exist as extended LLMs. But why be interested in extended LLMs? The answer is that extended LLMs reveal a host of issues that transcend the traditional divide between philosophy and engineering. From an engineering perspective, the aim is to build evermore capable LLMs, ones that move us in the direction of human levels of intelligence. But active externalists insist that much of human intelligence is a form of extended intelligence—that many of the hallmark features of human intelligence (e.g., our capacity for advanced thought and reason) are tied to the operation of extended mechanisms. This establishes an important point of contact between the philosophical effort to understand the nature of human intelligence and the technological effort to build the next generation of intelligent machines. If the distinctive feature of human intelligence is its extended status, then the project of building machines with human-level intelligence is, in effect, the project of building machines that are able to factor external resources deep into their cognitive routines. This merits a closer examination of human cognitive performances, focusing not just on what is inside the head but also what the head is inside of.

An extended perspective is also important when it comes to the evaluation of LLMs. Rather than focus on the capabilities of LLMs independently of the wider environment, it may be important to consider the ways in which the ecological embedding of LLMs contributes to various forms of situated success. Recall the case of the bluefin tuna. We cannot understand the tuna's capabilities by removing it from its aqueous medium. Nor is the explanatory effort best served by limiting our attention to what happens on one or other side of the tuna's biological boundary. It is only by acknowledging the complex interactions that occur between the brain, the body, and the world that we are able to explain (and thus understand) the origin of the tuna's natatorial success. We can, of course, study the capabilities of LLMs independently of their wider environment, but we should not expect these capabilities to be the same as those arising from the operation of extended mechanisms. This looks to be important when it comes to understanding both the benefits and (crucially) the risks associated with LLMs, especially when it comes to the deployment of LLMs in informationally-rich and technologically-saturated 'smart' environments.

On the philosophical front, extended LLMs are important because they expand the scope of active externalism, opening up new lines of philosophical inquiry, and revealing new directions for research into the extended mind (see Smart et al., 2025). Of particular importance is the way that LLMs direct our attention to the role of language in the formation of extended circuits. Courtesy of their facility with language, LLMs are able to invoke computational services, configure the operation of computational tools, and exploit online information. The virtue of language, on this view, is that it provides the basis for the formation of extended systems, enabling LLMs to factor external resources deep into their cognitive and computational routines. This is the nature of the gift we give to the new arrivals. It is a gift befitting those whose help we may one day come to rely on. The gift is our language. And what better gift for an intelligent entity whose artificial habitat is an online ocean of words?

References

Adams, F. and Garrison, R. (2013). The mark of the cognitive. *Minds and Machines*, 23(3):339–352.

Asai, A., Wu, Z., Wang, Y., Sil, A., and Hajishirzi, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In Kim, B., Chaudhuri, S., Khan, M. E., Sun, Y., and Fragkiadaki, K., editors, *The Twelfth International Conference on Learning Representations*, Vienna, Austria.

Bauer, W. A. and Marmodoro, A., editors (2024). *Artificial Dispositions: Investigating Ethical and Metaphysical Issues*. Bloomsbury Publishing, London, UK.

Chiang, T. (2004). *Stories of Your Life and Others*. Pan Macmillan Ltd, London, UK.

Clark, A. (1997). *Being There: Putting Brain, Body and World Together Again*. MIT Press, Cambridge, Massachusetts, USA.

Clark, A. (1998). Magic words: How language augments human computation. In Carruthers, P. and Boucher, J., editors, *Language and Thought: Interdisciplinary Themes*, pages 162–183. Cambridge University Press, Cambridge, UK.

Clark, A. (1999). Where brain, body, and world collide. *Cognitive Systems Research*, 1(1):5–17.

Clark, A. (2003). *Natural-Born Cyborgs: Minds, Technologies and the Future of Human Intelligence*. Oxford University Press, Oxford, UK.

Clark, A. (2006). Material symbols. *Philosophical Psychology*, 19(3):291–307.

Clark, A. (2008). *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford University Press, New York, New York, USA.

Clark, A. (2015). What 'extended me' knows. *Synthese*, 192(11):3757–3775.

Clark, A. (2024a). Extending the predictive mind. *Australasian Journal of Philosophy*, 102(1):119–130.

Clark, A. (2024b). Mind unlimited? In Hetherington, S., editor, *Extreme Philosophy*, pages 123–137. Routledge, New York, New York, USA.

Clark, A. and Chalmers, D. (1998). The extended mind. *Analysis*, 58(1):7–19.

Floridi, L. and Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Guo, Q., Wang, M., and Wang, H. (2024). Retrieval-augmented generation for large language models: A survey. Preprint at https://arxiv.org/abs/2312.10997.

Gillett, A. J., Whyte, C. J., Hewitson, C. L., and Kaplan, D. M. (2022). Defending the use of the mutual manipulability criterion in the extended cognition debate. *Frontiers in Psychology*, 13(Article 1043747):1–9.

Jiang, Z., Xu, F. F., Gao, L., Sun, Z., Liu, Q., Dwivedi-Yu, J., Yang, Y., Callan, J., and Neubig, G. (2023). Active retrieval augmented generation. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

Lupyan, G. (2016). The centrality of language in human cognition. *Language Learning*, 66(3):516–553.

Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., Rozière, B., Schick, T., Dwivedi-Yu, J., Celikyilmaz, A., Grave, E., LeCun, Y., and Scialom, T. (2023). Augmented language models: a survey. *Transactions on Machine Learning Research*, pages 1–35.

Ram, O., Levine, Y., Dalmedigos, I., Muhlgay, D., Shashua, A., Leyton-Brown, K., and Shoham, Y. (2023). In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., and Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 68539–68551, New Orleans, Louisiana, USA. Curran Associates, Inc.

Shanahan, M. (2024) Talking about large language models. *Communications of the ACM*, 67(2), 68–79.

Smart, P. R. (2018). Human-extended machine cognition. *Cognitive Systems Research*, 49:9–23.

Smart, P. R. (2024). Extended X: Extending the reach of active externalism. *Cognitive Systems Research*, 84(Article 101202):1–12.

Smart, P. R., Clowes, R. W., and Clark, A. (2025). ChatGPT, extended: Large language models and the extended mind. *Synthese*, 205(Article 242), 1–30

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In von Luxburg, U., Guyon, I., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30, pages 1–11, Long Beach, California, USA.

Wheeler, M. (2019). The reappearing tool: transparency, smart technology, and the extended mind. *AI & Society*, 34(4):857–866.

Zhou, Y., Liu, Z., Jin, J., Nie, J.-Y., and Dou, Z. (2024). Metacognitive retrieval-augmented large language models. In Chua, T.-S. and Ngo, C.-W., editors, *Proceedings of the ACM Web Conference*, pages 1453–1463, Singapore. Association for Computing Machinery.