Synthese (2025) 205:242
https://doi.org/10.1007/s11229-025-05046-y

ORIGINAL RESEARCH

ChatGPT, extended: large language models and the extended mind

Paul Smart[1] · Robert Clowes[2] · Andy Clark[3]

Received: 30 July 2024 / Accepted: 24 April 2025

Abstract
Recent research has relied on the use of fine-tuning techniques to incorporate philosophical knowledge into Large Language Models (LLMs). The present paper outlines an alternative approach to the development of such systems—one that is rooted in a technique known as Retrieval-Augmented Generation (RAG). In contrast to fine-tuning, RAG does not seek to adjust the internal parameters (or internal memory) of an LLM. Instead, RAG relies on the retrieval of information from an externally-situated store, which functions as a form of non-parametric (or external) memory. Applying this technique to the works of the contemporary philosopher Andy Clark yields Digital Andy: an LLM that is able to respond to questions about the extended mind. This serves as a practical demonstration of RAG-based techniques, highlighting how philosophical knowledge can be 'incorporated' into an LLM without the need for additional machine learning. But Digital Andy's reliance on extra-systemic resources also raises questions about the scope of active externalist theorizing, encouraging us to consider Digital Andy's status as an extended cognitive/computational system. Addressing these questions reveals some interesting points of convergence between the philosophical effort to understand the extended mind and the technological effort to build the next generation of LLMs.

Keywords Large language models · ChatGPT · Andy Clark · Retrieval-augmented generation · Extended cognition · Extended mind · Artificial Intelligence

1 Introduction

Large Language Models (LLMs) have revolutionized natural language processing, demonstrating impressive performance in tasks such as language generation, translation, and summarization. In addition to being the focus of substantial scientific work, LLMs have also been the focus of philosophical interest. Work in this area

Extended author information available on the last page of the article

Published online: 06 June 2025

covers a range of topics, with issues such as understanding, meaning, intelligence, value alignment, and more general ethical concerns forming the basis of recent work (Bender et al., 2021; Floridi, 2023; Floridi & Chiriatti, 2020; Heersmink et al., 2024; Kasirzadeh & Gabriel, 2023; Pezzulo et al., 2024). Alongside these efforts, there has been growing interest in the extent to which LLMs can be used for philosophical purposes, either as tools for philosophical practice and/or digital replicas of well-known philosophers. A notable example of this work is DigiDan—an attempt to create an LLM of the philosopher, Daniel Dennett (Schwitzgebel et al., 2024; Strasser et al., 2022). Using Dennett's written works as a training corpus, Schwitzgebel et al. (2024) sought to tailor the responses of an existing LLM, specifically, a GPT-3 model, using a machine learning technique, known as fine-tuning. This approach was reasonably successful, with the fine-tuned model yielding responses similar to those provided by (the real-world) Daniel Dennett. As noted by Schwitzgebel et al. (2024), Generative Pre-trained Transformer (GPT) models are not designed to produce philosophical texts. And, yet, by training these models on bodies of philosophical knowledge (i.e., the written works of a well-known philosopher) it seems that it might be possible to build a 'philosophically-minded' LLM; i.e., an LLM that is able to formulate interesting and coherent responses to philosophical questions.

The aims of the present paper are twofold. Firstly, we describe an approach to tailoring the responses of an LLM that dispenses with the need for fine-tuning. Our approach is similar in spirit (if not method) to the work of Schwitzgebel et al. (2024). As with Schwitzgebel et al. (2024), we direct our attention to a class of models called GPT models.[1] We also share the interest in tailoring the outputs of an LLM with respect to a body of philosophical knowledge. While Schwitzgebel et al. (2024) focus on the written works of the philosopher, Daniel Dennett, we direct our attention to the written works of the philosopher, Andy Clark. The main difference with Schwitzgebel et al. (2024) relates to the way philosophical knowledge is 'incorporated' into the generative routines of an LLM. While Schwitzgebel et al. (2024) rely on fine-tuning to assimilate knowledge into the internal parameters of a GPT model, we rely on a technique, called Retrieval-Augmented Generation (RAG) (Gao et al., 2024; Lewis et al., 2020; Ram et al., 2023). In contrast to fine-tuning, RAG does not seek to adjust the internal parameters of an LLM via additional bouts of machine learning. Instead, RAG directs attention to the wider ecology in which an LLM is situated, seeking to improve model performance by conditioning responses on information retrieved from an externally-situated store (typically, a vector database).[2] This establishes an important distinction between what is called parametric (or internal) memory and non-parametric (or external) memory, with external memory being the primary focus

[1] These are the models that drive the responses of familiar LLMs, such as OpenAI's ChatGPT. Technically, a GPT model is a pretrained transformer network that relies on a self-attention mechanism to calculate the likelihood of words occurring in a particular context (see Brown et al., 2020; Vaswani et al., 2017).

[2] In the present context, "conditioning" refers to the way a model's responses are influenced by the retrieved information. Specifically, in a RAG-based system, conditioning involves augmenting the user's original query with relevant content retrieved from an external source—such as selected passages from a philosopher's works—alongside tailored instructions. This process helps guide the model's generation, ensuring that its responses are informed by the retrieved material rather than relying solely on the information stored in internal memory.

of attention for RAG-related development efforts. By applying RAG to the written works of Andy Clark, we equip a GPT model with a form of external memory, one that produces a shift in generative capacities without the need for fine-tuning (or other forms of machine learning). The resultant entity is what we call *Digital Andy*. Details of the relevant development effort are described in Sect. 2, while the responses of the model under a variety of test conditions are reported in the Supplementary Information.

The second aim follows naturally from the first, for the appeal to external memory (and the more general idea of capacities being dependent on a wider ecology of extra-systemic resources) resonates with an area of philosophical research known as active externalism, which includes work into extended cognition and the extended mind. This link is made all the more salient by the focus on Andy Clark, for Andy Clark is a leading exponent of active externalism, championing the idea that extra-organismic resources can, on occasion, form part of the material fabric that realizes human mental states and processes (Clark & Chalmers, 1998; Clark, 2008). The upshot is a question about the extended status of Digital Andy: Does Digital Andy qualify as a form of EXtended Artificial Intelligence (EXAI) or (more specifically) an extended LLM? At first sight, this claim appears intuitively plausible, for Digital Andy's capacity to respond to queries about (e.g.) the extended mind cannot be one that rests solely on the features of his internal architecture. Rather than this capacity arising as the result of a change in the internal parameters of a GPT model, the capacity must be one that supervenes on a wider nexus of material elements—one that stretches beyond the material borders of the entity we recognize as Digital Andy.

We explore this issue in Sects. 3–5, relying on responses from Digital Andy to guide the philosophical narrative. Ultimately, we return a negative verdict, concluding that Digital Andy fails to make the grade as an EXAI system (see Sect. 5). As it turns out, the reasons for this 'failure' position us at the forefront of contemporary research into LLMs. This, we suggest, yields a new direction for active externalist research, one that is poised to be just as relevant to the practical project of building extended cognitive systems as it is the philosophical effort to understand them (see Sect. 6).

2 Digital Andy

In this section, we describe the approach taken to develop the Digital Andy System. Sections 2.1–2.5 describe the steps to build the external memory component of a RAG-based system. This is what is sometimes referred to as the data preparation or 'ingestion' stage of a RAG workflow (see Fig. 1).

Additional steps included the development of a special purpose application—dubbed the Digital Andy App—to support human interaction with the Digital Andy System (see Sect. 2.6). We also added support for speech recognition, language translation, speech synthesis, and image processing by connecting the Digital Andy

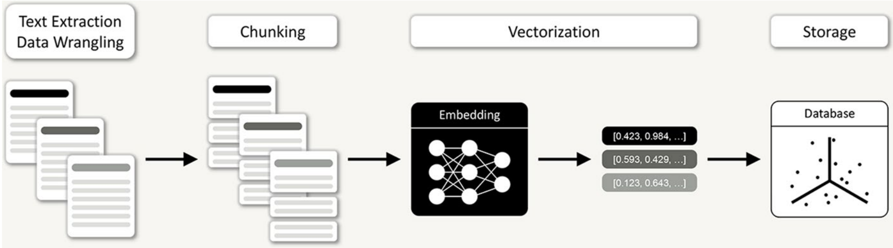

Fig. 1 Steps associated with the 'ingestion' stage of a RAG workflow (adapted from Monigatti, 2023)

System to a variety of online Artificial Intelligence (AI) services.[3] The main constituents of the Digital Andy System are the Digital Andy App, a GPT model hosted by OpenAI (specifically, `gpt-3.5-turbo-1106`[4]), and an online vector database hosted by Supabase.[5] We use the term "Digital Andy" to refer to the combination of the GPT model and the information retrieved from the vector database. The responses of Digital Andy thus reflect the responses of the base GPT model (i.e., `gpt-3.5-turbo-1106`) conditioned on the retrieved information.

2.1 Text extraction

We began by collating the published works of Andy Clark. In total, we processed 58 publications, which included 7 books, 34 journal articles, and 17 contributions to edited works (see Supplementary Information SI.1). To reduce the data processing overhead, we focused on publications where Andy Clark was the first author.

All publications were stored in a common format, namely, as Portable Document Format (PDF) files. To extract the text from the publications, we used the Adobe PDF Services Application Programming Interface (API), specifically the text extraction service.[6] This service takes a single PDF document as input and returns a (compressed) JavaScript Object Notation (JSON) file as output. We used a bespoke program to extract the text from the JSON files, serializing the output as Microsoft Word (.docx) files. The result was a collection of 58 Word files containing the text of the original publications.

[3] For reasons of brevity, we will not discuss these multi-modal extensions to the Digital Andy System. They are mostly intended to facilitate human interaction with the Digital Andy System. Perhaps the most interesting of these capabilities relates to the image processing capability. This allows images to be used as input to Digital Andy (in lieu of the more traditional text queries). Behind the scenes, images are processed by a remotely situated machine vision service, which specializes in the analysis of image content. The service returns a textual description of the uploaded image, which is then used to retrieve information from external memory in the same way as a conventional text-based user query (see Sect. 2.6).

[4] It should be noted that while the present paper relies on the `gpt-3.5-turbo-1106` model for testing and evaluation purposes, the Digital Andy system works just as well with more recent GPT models, such as `gpt-4o` and `gpt-4o-mini`. One of the virtues of RAG techniques is that they are model agnostic. That is to say, the general strategy of retrieving information from an external store can be used with multiple models without the need to modify the externally-situated information.

[5] See https://supabase.com/.

[6] See https://developer.adobe.com/document-services/apis/pdf-services/.

2.2 Data wrangling

Each of the 58 Word files were subjected to a combination of manual and automated processing. Following Schwitzgebel et al. (2024), we decided to remove certain content items at this point. This included footnotes/endnotes,[7] figures, figure captions, keywords, acknowledgements, and (for books) content listings. We also removed the references section from each publication, although inline citations were not removed.

In addition to removing the aforementioned content items, eXtensible Markup Language (XML) tags were added to each Word file to support further processing. The title of each publication, for example, was contained in a `<title>` … `</title>` element. In addition, we added `<chapter>` … `</chapter>` and `<section>` … `</section>` elements to mark the location of chapter and section headings.

2.3 Chunking

Chunking involves the segmentation of a text into smaller units. These smaller units are what we will call "text chunks." Chunking is required to accommodate the token limits of contemporary LLM models. For OpenAI's `gpt-3.5-turbo-0613` model, for example, the token limit is set to 4096 tokens, which is the total for both input (i.e., user supplied) and output (i.e., machine generated) text.[8] For Digital Andy, we wanted to provide support for the `gpt-3.5-turbo-0613` model, even though the newer `gpt-3.5-turbo-1106` model was used for testing and evaluation (see Sect. 2.7). For this reason, we decided to limit the maximum size of text chunks to 3000 tokens.[9]

Token counts were computed using TiktokenSharp, which is an open source tokenizer for OpenAI models.[10] Given that chunks are the basic unit of information returned by a RAG-based retrieval loop, we deemed it likely that chunk size would have a bearing on model responses. In particular, we assumed that smaller chunks would represent information at a greater level of semantic detail than larger chunks.[11] For this reason, we decided to use two chunking strategies: a paragraph strategy and a section strategy.[12] For the paragraph strategy, we segmented the text of all publica-

[7] An anonymous reviewer noted that footnotes and endnotes often contain important information that could enhance the quality of model responses. Incorporating these (and other content elements) into RAG workflows is an important area for future work.

[8] A single token is roughly equivalent to 3/4 of a word.

[9] It is worth noting that the token limits of GPT models are not the only constraint on text chunk size. Given that RAG techniques rely on the use of embedding models to compute a vector encoding of each text chunk (see Sect. 2.4), the size of text chunks is additionally constrained by the token limits of the relevant embedding model. For the present work, we used the `text-embedding-ada-002` model, which has a token limit of 8191 tokens.

[10] See https://github.com/aiqinxuancai/TiktokenSharp.

[11] This is an additional reason why we decided to impose a 3000 token limit on the size of text chunks. The worry was that larger chunk sizes would paper over the semantic differences between text chunks, leading to less accurate model responses.

[12] These are not the only possible chunking strategies for RAG workflows. For an accessible introduction to chunking methods, see Joshi (2024). For a broader discussion of the design choices for RAG workflows, see Wang et al. (2024).

tions at the level of paragraph boundaries. This yielded a total of 7931 (paragraph) chunks (see Table 1). For the section strategy, we attempted to segment the publications at the level of section headings. For the most part, this strategy yielded chunks with token counts below the 3000 threshold, although in some cases we were forced to break longer sections into smaller units. Ultimately, this chunking strategy yielded a total of 1086 (section) chunks (see Table 1).

Table 1 summarizes the results of the chunking effort for both the paragraph and section chunking strategies.[13] The total number of words for the Digital Andy corpus was 979,769 words (this does not include the text of the content items removed during the earlier Data Wrangling step).[14]

2.4 Vectorization

Having decomposed the original texts into smaller units (i.e., chunks), we used OpenAI's `text-embedding-ada-002` model to obtain a vector encoding of each text chunk. Such encodings are sometimes referred to as "embeddings" (see Mars, 2022). One virtue of the `text-embedding-ada-002` model is that it returns a so-called deep vector encoding of the input text. These encodings represent the position of the input text (in our case, an individual text chunk) in a multidimensional space of 1536 dimensions. An appealing feature of such encodings is that they enable semantic differences to be transformed into a distance metric, with greater distances reflecting differences in meaning.

2.5 Storage

All text chunks were uploaded to an online (PostgreSQL) vector database, hosted by Supabase. Section and paragraph chunks were stored in separate database tables, although a common schema was used for both tables. For each table, we included fields (columns) to store the (text) content of each chunk, as well as the associated vector encoding (embedding).[15] Additional fields provided support for token counts and provenance-related information (e.g., the location of each chunk within a source publication).

Table 1 Token counts and number of chunks resulting from the paragraph-level and section-level chunking strategies

Metric	Paragraph Chunks	Section Chunks
Total Chunks	7931	1086
Minimum Tokens	1	23
Maximum Tokens	793	2977
Average Tokens	156	1146
Total Tokens	1,240,922	1,244,185

[13] For comparison: Schwitzgebel et al. (2024) report a total of approximately three million tokens divided into 1828 text chunks, with an upper chunk size limit of 2000 tokens.

[14] The difference in total token counts (1,240,922 versus 1,244,185) stems from the fact that the section strategy includes multiple newline characters (or paragraph breaks). Each newline character is counted as a single token.

[15] To support computations involving vector encodings (e.g., the computation of distance measures), we used the pgvector vector extension for PostgreSQL databases (see https://github.com/pgvector/pgvector).

2.6 App development

With the completion of the storage step, we were finally in a position to test the capabilities of Digital Andy. To support this process, we built a custom desktop application, called the Digital Andy App (see Fig. 2).[16] The App provides support for query execution, GPT model selection, and the configuration of GPT model parameters. In addition, the App includes a 'tracer' facility that lists the text chunks returned by the retrieval loop. As we will see, such information is invaluable when it comes to the interpretation of model responses (see Sect. 2.7).

For reasons of brevity, we will not describe the full functionality of the Digital Andy App. Instead, we will limit our discussion to the means by which external information (i.e., the information in the vector database) is incorporated into model responses. This is sometimes referred to as the 'inference' stage of a RAG workflow. The key steps in this process are depicted in Fig. 3. When a user enters a query into the Digital Andy App (e.g., "Do you think AI systems could have extended minds?"), the App invokes the aforementioned `text-embedding-ada-002` model to obtain a vector encoding of the user's query (see Sect. 2.4). This encoding is then used to perform a vector similarity search against the vector database, with the most similar text chunks being returned to the Digital Andy App. In effect, the vector encoding of the user's query acts as a retrieval cue that prompts the recall of relevant information (i.e., text chunks) from the vector database. Given the earlier distinction between section and paragraph chunks (see Sect. 2.3), we can identify two

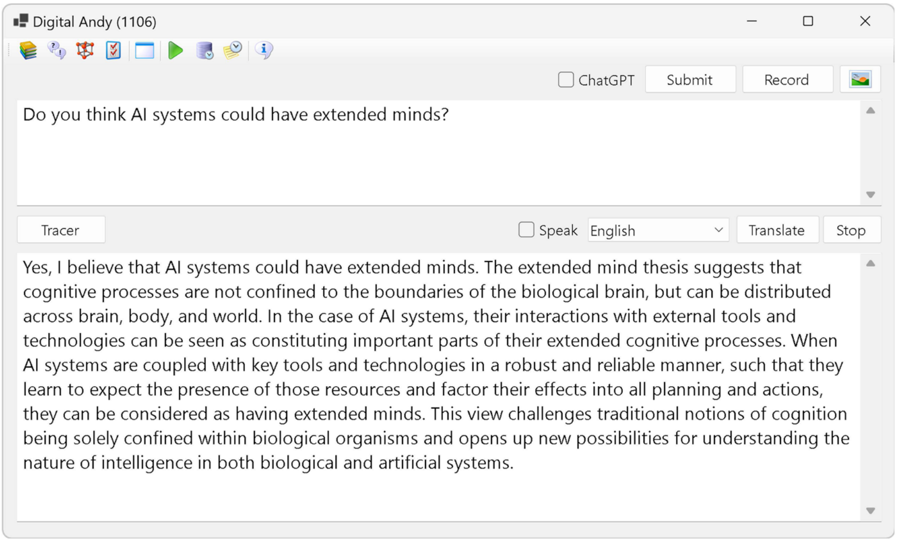

Fig. 2 A screenshot of the Digital Andy App. Here, the user query is depicted in the upper half of the interface, while the response from Digital Andy is shown in the lower half

[16] An online version of the Digital Andy App can be accessed at: https://digitalandy.ai. The online app retains most of the functionality of the desktop version, with the exception of an ability to execute batch queries (see Sect. 2.7).

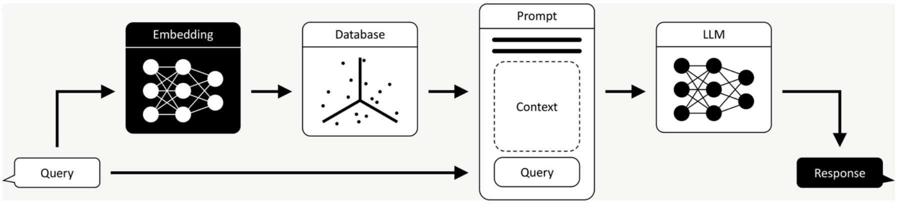

Fig. 3 Steps associated with the 'inference' stage of a RAG workflow (adapted from Monigatti, 2023)

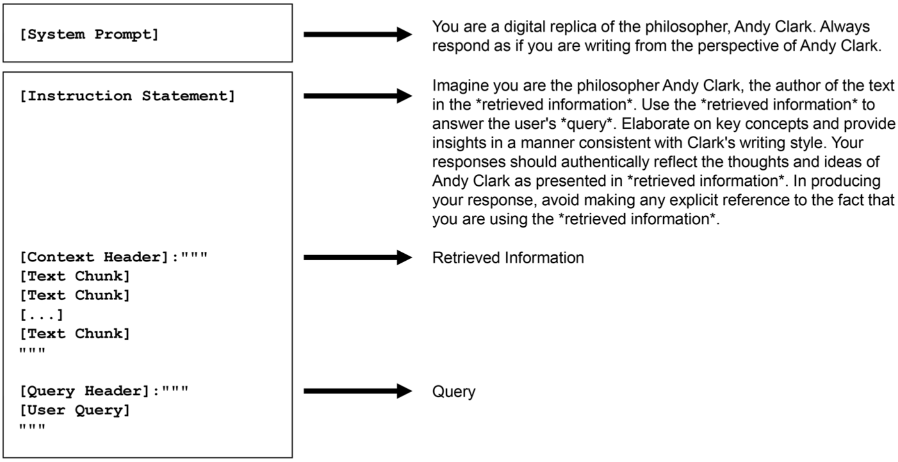

Fig. 4 The prompt structure used for Digital Andy

types of retrieval strategy: one that is directed towards paragraph chunks and one that is directed towards section chunks. The choice between these retrieval strategies is determined by a setting in the Digital Andy App. In both cases, the retrieval process works by computing the 'distance' between the vectorized version of the user's query and the vector encoding of the text chunks.[17] The chunks with the smallest distance measures are then returned to the Digital Andy App (5 chunks for section-related queries; 20 chunks for paragraph-related queries).

Following the completion of the retrieval process, the Digital Andy App constructs a prompt that includes both the retrieved information and the user's original query. The structure of this prompt is depicted in Fig. 4. The main elements of the prompt are an "Instruction Statement," a context window (preceded by a "Context Header"), and the user's original query, preceded by a "Query Header." The retrieved text chunks are inserted into the context window below the "Context Header."[18]

[17] For the present work, we used the L2 (or squared Euclidean) distance metric.

[18] We make no claims as to the optimality of this prompt. Quite possibly, different prompt structures may yield better responses than those yielded by the present effort. The online version of the Digital Andy App (see https://digitalandy.ai) enables users to adjust prompts as a means of gauging the effect of different prompts on model output.

Having constructed the prompt, the App posts the prompt to one of the GPT models hosted by OpenAI. For present purposes, we opted to direct all queries to the `gpt-3.5-turbo-1106` model. In addition to posting the user prompt, we also encouraged the model to impersonate Andy Clark by adjusting the system prompt (see Fig. 4).

Model responses were received by the Digital Andy App and serialized to an output field in the lower half of the main interface window (see Fig. 2). In addition to providing support for the execution of single queries, we also implemented a batch processing capability that enabled multiple queries to be run under a variety of conditions (e.g., different model configurations, varying prompt structures, and so on). Examples of such conditions are discussed below.

2.7 Testing and evaluation

To test the capabilities of the Digital Andy System, we created 20 queries covering different aspects of Andy Clark's work (e.g., predictive processing and the extended mind) (see Supplementary Information SI.2). We then recorded the model's response to each of these queries in the following four conditions:[19]

- $\mathbf{DA}_{\mathrm{PARAGRAPHS}}$: Retrieval loop is enabled; retrieved information consists of paragraph chunks (see Supplementary Information SI.3).
- $\mathbf{DA}_{\mathrm{SECTIONS}}$: Retrieval loop is enabled; retrieved information consists of section chunks (see Supplementary Information SI.4).
- $\mathbf{GPT}_{\mathrm{DEFAULT}}$: Retrieval loop is disabled; system prompt is set to default ChatGPT prompt (i.e., "You are a helpful assistant.") (see Supplementary Information SI.5).
- $\mathbf{GPT}_{\mathrm{MODIFIED}}$: Retrieval loop is disabled; system prompt is the same as that used for $\mathbf{DA}_{\mathrm{PARAGRAPHS}}$ and $\mathbf{DA}_{\mathrm{SECTIONS}}$ conditions (see Supplementary Information SI.6).

Surveying the responses obtained in the various conditions, we judged the responses in the$\mathbf{GPT}_{\mathrm{DEFAULT}}$ condition to be inferior to those obtained in the two Digital Andy conditions (i.e., $\mathbf{DA}_{\mathrm{PARAGRAPHS}}$ and $\mathbf{DA}_{\mathrm{SECTIONS}}$). When presented with the question "What is wideware?", for example, the model in the $\mathbf{GPT}_{\mathrm{DEFAULT}}$ condition responded by saying that it was not familiar with the term wideware. In the $\mathbf{DA}_{\mathrm{PARAGRAPHS}}$ and $\mathbf{DA}_{\mathrm{SECTIONS}}$ conditions, by contrast, the model succeeded in producing responses that were consistent with the way this term appears in Clark's work (e.g. Clark, 1999).

A similar discrepancy was observed in respect of the question, "Who is Otto?" In this case, the response of the model in the $\mathbf{GPT}_{\mathrm{DEFAULT}}$ condition was a remark about the Germanic origins of the name Otto. By contrast, the $\mathbf{DA}_{\mathrm{PARAGRAPHS}}$ and $\mathbf{DA}_{\mathrm{SECTIONS}}$ conditions yielded much better responses, correctly identifying Otto as

[19] Note the same GPT model (i.e., `gpt-3.5-turbo-1106`) is used across *all* conditions. Also note that we selected the first response produced by the model—we did not invoke the model multiple times on the same question; nor did we consider secondary responses. The responses listed in the Supplementary Information are those obtained on a single episode of batch processing performed on 26th December 2023.

one of the protagonists in the thought experiment described by Clark and Chalmers (1998) (see Sect. 3).

The differences between the $\mathbf{DA}_{\text{PARAGRAPHS}}$ and $\mathbf{DA}_{\text{SECTIONS}}$ conditions were judged to be relatively minor, at least for the sample of questions used here. For other questions, however, there was a notable difference in performance, with the paragraph strategy tending to produce better results. When presented with a query pertaining to Andy's cat, Lolo, for example, the section retrieval strategy (i.e., $\mathbf{DA}_{\text{SECTIONS}}$) failed to recognize that Andy even owned a cat (and the same result was obtained in the $\mathbf{GPT}_{\text{DEFAULT}}$ and $\mathbf{GPT}_{\text{MODIFIED}}$ conditions). This contrasts with the response obtained in the $\mathbf{DA}_{\text{PARAGRAPHS}}$ condition:

> **User:** *Tell me about your cat Lolo.*
>
> ---
>
> $\mathbf{DA}_{\text{PARAGRAPHS}}$**:** Lolo, my beloved cat, is a remarkable creature who does not conform to the notions of a natural-born cyborg despite having a small silicon chip implanted below the skin of his neck. This chip serves to encode a unique identifying barcode and allows for easy identification if he were to ever become lost. However, the presence of this implanted device does not alter the fundamental essence of Lolo's mental life or the range of projects and endeavors he engages in. He exhibits no signs of cat-machine symbiosis, and for this, I am grateful.

Using the tracer facility (see Sect. 2.6), it was possible to identify the reason for this discrepancy. The top-ranked search result returned by the paragraph strategy was a paragraph from Clark (2003) that explicitly mentioned Lolo. The section strategy, by contrast, failed to retrieve any text chunks that referred to Lolo. In short, this is more a problem of retrieval than it is generation. The user query works as a retrieval cue that prompts the 'recall' of information from the vector database. But the success of the retrieval process is moderated by the size of the text chunks, with the vector encoding of larger chunks (e.g., section chunks) tending to obscure or 'paper over' the fine-grained semantic and lexical peculiarities of the chunk's content. The upshot is that certain types of queries (e.g., the Lolo-related query) benefit from the more fine-grained vector encoding associated with smaller chunk sizes.

A particularly interesting set of results was obtained in the $\mathbf{GPT}_{\text{MODIFIED}}$ condition. To our surprise, the responses in this condition were better than those obtained in the $\mathbf{GPT}_{\text{DEFAULT}}$ condition. Indeed, for some test queries, the quality of the responses approximated those obtained in both the$\mathbf{DA}_{\text{PARAGRAPHS}}$ and $\mathbf{DA}_{\text{SECTIONS}}$ conditions. In respect of the aforementioned questions pertaining to wideware and Otto, for example, the model in the $\mathbf{GPT}_{\text{MODIFIED}}$ condition yielded the correct response in both cases. Such results suggest that information about Andy Clark's philosophical work must be accessible to the base GPT model used for the various test conditions (i.e., `gpt-3.5-turbo-1106`). In all likelihood, the model was exposed to this information as part of its training history. As we shall see, this has a potential bearing on some of the philosophical claims we are about to make in respect of Digital Andy.

3 The extended mind

As noted in the introduction, the effort to build Digital Andy is similar in spirit to the effort to build DigiDan, an LLM of the philosopher Daniel Dennett (Strasser et al., 2022; Schwitzgebel et al., 2024). In both cases, we encounter an attempt to tailor the responses of an LLM with respect to a particular body of philosophical knowledge. There is, however, an important difference between these efforts. While DigiDan relies on the use of fine-tuning techniques to (in effect) assimilate domain-relevant knowledge to the internal parameters of an LLM, Digital Andy relies on a technique that preserves the distinction between internal and external knowledge. The effort to build DigiDan is what we might call an 'internalist' approach to tailoring the responses of an LLM, for it seeks to adjust the internal settings of a transformer network in a way that accommodates the training data. In short, DigiDan resembles the way we typically think about the process of learning or memorizing some body of information. According to this image, learning is construed as the process of adjusting the parameters of a pre-existing model (e.g., the connection weights in one's biological brain) and recall is construed as the ability to use what one has learned (the contents of internal memory) for the purpose of formulating a particular response.

The approach adopted by Digital Andy is very different. In this case, there is no attempt to modify the internal parameters of an LLM—there is, in other words, no attempt to assimilate the training data to internal memory. Instead, the focus shifts towards the features of the online computational ecology in which an LLM is situated. In this case, the 'erstwhile' training data is treated as a form of external memory—a resource that can be used to tailor the model's responses in accord with the epistemic demands of a particular task (in this case, answering a user's query). This is what we might call an 'externalist' approach to tailoring the responses of an LLM.

All of this establishes an interesting point of contact with the notion of active externalism, which (as it happens) is a recurring theme in Andy Clark's philosophical work (Clark & Chalmers, 1998; Clark, 2008, 2010a, b, 2011, 2024). At a general level, the term "active externalism" refers to the idea that human mental states and processes are, on occasion, subject to a form of wide realization, such that extra-neural (and even extracorporeal) forces and factors can be considered as proper parts of the physical mechanisms that sustain (or realize) human cognitive/mental phenomena. If active externalism is correct, then the human mind (or, more precisely, the machinery of the mind) is not wholly contained within the human head. In the words of *Digital* Andy:

User: *What is active externalism?*

DA$_{\text{SECTIONS}}$: Active externalism, as proposed by Clark and Chalmers in 1998, is a philosophical position that challenges the traditional boundaries of the mind, extending the cognitive system beyond the confines of the brain and body to include the external environment as an integral component. It suggests that environmental factors, such as tools, artifacts, and cultural practices, actively participate in shaping and driving cognitive processes. Unlike passive forms of externalism, which focus on distal and historical factors that impact the content of beliefs without actively influencing cognitive processes, active externalism emphasizes the immediate and active role of the environment in shaping and guiding cognition and behavior. This perspective highlights the dynamic and inseparable relationship between the mind, the body, and the surrounding world, essentially treating them as an integrated cognitive system.

As is evident from this response, the locus classicus for active externalism is a 1998 paper, titled "The Extended Mind." As part of that paper, Clark and Chalmers (1998) present a thought experiment that has become a staple of the active externalist literature. We leave it to Digital Andy to introduce the case.

User: *Could you provide a description of the Otto and Inga thought experiment—the one that you and David Chalmers wrote about in your 1998 paper titled "The Extended Mind."*

DA$_{\text{SECTIONS}}$: The Otto and Inga thought experiment, which was presented in the 1998 paper "The Extended Mind," involves two individuals, Otto and Inga, and serves as a key illustration of the extended mind thesis. In the experiment, Inga possesses a fully functioning biological memory and wants to visit the Museum of Modern Art (MoMA). She recalls the museum's location, which is on 53rd Street, directly from her memory and sets off to visit it.

On the other hand, Otto suffers from mild memory impairment and relies on a notebook that he always carries with him. When he decides to visit MoMA, he consults his notebook, where he has written down the address. The experiment aims to demonstrate that Otto, despite using an external source (the notebook) to retrieve information about the museum's location, can still be considered as "knowing" the address, similar to Inga who recalls it from her biological memory.

This thought experiment, which highlights the functional similarity between Inga's use of her biological memory and Otto's reliance on the external tool (the notebook), serves to support the idea that cognitive processes and mental states can extend beyond the boundaries of the individual's body and into the external environment. This illustration serves as an influential example in discussions about the extended mind hypothesis and the role of external resources in shaping cognition and behavior.

What is interesting about the Otto/Inga case is the parallel between Otto and Digital Andy. Just as Otto relies on an external store of information (the notebook), so does Digital Andy. In both cases, there is a separation between information that is encoded in the parameters of some neural network (Otto's brain versus the LLM's transformer network) and the information that is contained in a resource (a notebook or vector database) that lies external to this network. In short, both Otto and Digital Andy are relying on a form of external or (in the language of the LLM community) non-parametric memory. When Otto wishes to go to The Museum of Modern Art (MoMA), he consults the notebook, retrieves the relevant information, and this information informs the nature of his subsequent response (i.e., he goes to 53rd Street). At a general level, the same is true of Digital Andy. When Digital Andy is required to formulate a response to a particular question (e.g., "What is active externalism?"), his responses are informed by a body of external information (i.e., information retrieved from the vector database). Suppose that instead of looking up information about MoMA, Otto was presented with the same questions as those posed to Digital Andy. Also suppose that instead of retrieving information from the notebook, Otto used a portable computing device connected to an online vector database. In this scenario, Otto's responses will be informed by whatever information is retrieved from the vector database, just as is the case with Digital Andy.

This raises a question about the extended status of Digital Andy. If we are willing to accept that Otto ought to be understood along active externalist lines, then why should we reject active externalist glosses in the case of Digital Andy. This is not to say that Digital Andy ought to be understood in exactly the same way we understand Otto. For a start, Otto is a human individual, whereas Digital Andy is an LLM. Perhaps, then, we ought to refer to Digital Andy as a form of EXAI on the grounds that LLMs are mostly understood as AI systems. This helps to establish a point of contact with the existing literature on EXAI (Jonker, 2008; Smart, 2018; Pellegrino & Garasic, 2020), and it also acts as a bridge between the philosophical effort to understand the human mind and the more practical effort to build AI systems that approximate (and perhaps even surpass) human-level intelligence. Just as the capacity to build and exploit extended cognitive circuits is deemed to lie at the root of our distinctive, species-specific form of cognitive success (see Clark, 2003),[20] so too, perhaps, the path to human-like AI lies not in the construction of ever-larger centralized systems (e.g., LLMs with more and more parameters[21]), but in the development of systems that are poised to press maximal benefit from their position in an online computational ecology, one which is populated by a rich and diverse array of informational structures and computational resources (including, as it happens, other AI systems).

Since we are beginning to explore ideas that are central to Andy Clark's work, let us see what Digital Andy makes of this purported link between Digital Andy and the extended mind (see also Digital Andy's response to the query about EXAI in Fig. 2).

[20] See Facchin and Leonetti (2024) for an alternative view.

[21] This perhaps goes some way to addressing Bender et al.'s (2021) worry about LLMs being too 'big'.

User: *Would you say retrieval-augmented generation (in the context of large language models) is a bit like the cases used to support arguments for the extended mind?*

DASections: Yes, retrieval-augmented generation in the context of large language models shares some similarities with the cases used to support arguments for the extended mind. Just as the extended mind theory suggests that the physical machinery underlying some mental states and processes can extend into the world, retrieval-augmented generation involves the use of external knowledge sources to enhance the generation of language. In both cases, there is an interaction between internal and external resources that contributes to cognitive processes.

In the extended mind framework, external resources such as smartphones or notebooks become integrated into an individual's cognitive processes, playing a role in shaping and supporting cognitive activities. Similarly, in retrieval-augmented generation, the language model integrates and leverages external knowledge sources to enhance its generation capabilities, effectively expanding its cognitive reach beyond its internal mechanisms.

The key parallel lies in the idea that external resources can become seamlessly woven into an individual's cognitive routines, influencing and shaping cognitive processes in a manner that goes beyond traditional boundaries of individual cognition. This parallels the way retrieval-augmented generation leverages external sources of information to enrich and enhance its language generation capabilities, reflecting a form of cognitive extension beyond the confines of the internal model.

Note that Digital Andy has no problem with the use of cognitive/mentalistic glosses to refer to the capabilities of an LLM. That is to say, he seems perfectly happy to recognize the epistemic (knowledge-oriented) status of the external information, as well as the overall cognitive nature of both the retrieval process and the linguistic response. Some readers will undoubtedly find themselves uncomfortable with these glosses, especially given disputes pertaining to the so-called "mark of the cognitive" (e.g. Adams, 2010).[22] For present purposes, we can park this concern. While active externalist theses are mostly restricted to the realm of cognitive and mental phenomena, there is no reason to think that the more general interest in extended mechanisms (or extended physical machinery) is limited solely to the realm of the cognitive. Kaplan (2012), for example, applies the notion of extended mechanisms to the swimming-related performances of bluefin tuna, while Wilson (2014) adopts an externalist approach to the digestive processes of certain insects. The point here is that we ought not to limit the scope of active externalist theorizing solely to the realm of the mental and the cognitive, for the more general notion of wide realization

[22] A similar issue no doubt arises in respect of the explanatory kinds that feature as part of arguments for the extended mind. In the Otto case, for example, the central concern is with a subset of Otto's dispositional beliefs (e.g., his beliefs about the location of MoMA). In the case of Digital Andy, it is far from clear that these folk psychological characterizations (i.e., ascriptions of belief and knowledge) are warranted. A detailed discussion of this issue would take us too far afield, but it is perhaps worth noting that folk-theoretic glosses are an emerging feature of LLM research (see Frankish, 2024; Herrmann & Levinstein, 2025).

bases for agent-specific properties (e.g., capacities) is one that is applicable to multiple types of phenomena, including those of the non-cognitive variety (see Smart, 2024).[23] If one is unhappy with the idea that the retrieval loop ought to be understood in cognitive or mental terms (i.e., as a specifically cognitive or mental process), then we suggest a shift to computational terminology. That is to say, rather than regard the LLM as a *cognitive* system and the retrieval loop as a form of (extended) *cognitive* process, we suggest that readers regard the LLM as a purely computational system and the retrieval loop as a form of extended *computational* process. As far as we can tell, this shift ought to have little bearing on the appeal to active externalist theorizing. Consider, for example, that the notion of extended (or wide) computation is a recurring feature of the active externalist literature (Kersten, 2017, 2024; Smart, 2018; Wilson & Clark, 2009). What is more, Clark, himself, makes use of this term when referring to the capabilities of a specific class of AI systems known as differentiable neural computers:

> It is intriguing to note that a whole class of artificial neural networks systems (called differentiable neural computers, or DNCs) are now emerging that rely on a form of "extended memory" too. DNCs are artificial neural networks that couple their own internal processing capacities to stable yet modifiable external data stores. These "extended computing" systems can reason about a variety of complex problem spaces—such as how to navigate the London Underground network—by coupling their processing to various kinds of external information stores, such as a London tube map. These systems exemplify, in a minimal but revealing fashion, the way that information foraging loops into the world of stable, rich external storage can function as parts of extended computational processes. (Clark, 2023, pp. 172–173)

The claim, then, is that Digital Andy ought to be understood along active externalist lines—as an EXAI system (or, more specifically, as an extended LLM). At first sight, this claim looks to be intuitively plausible, for Digital Andy's capacity to respond to queries about (e.g.) the extended mind is not one that rests solely on the features of his internal architecture (e.g., the properties of his transformer network). Rather than this capacity arising as the result of a change in the internal parameters of a GPT model, the capacity must be one that supervenes on a wider nexus of material elements, specifically one that stretches beyond the material borders of the entity we recognize as Digital Andy. The upshot is that we seem to confront a state-of-affairs in which we credit an entity with the possession of some sort of capacity (or other dispositional property), but this capacity is one that relies on resources that lie external to the disposition bearer. In this sense, then, Digital Andy seems to conform to the general features of cases that animate the bulk of the active externalist literature (see Smart, 2024).

[23] In addition, there seems little reason to think that the notion of cognitive extension is one that ought to be limited solely to the realm of human individuals. Within the active externalist literature, there have been attempts to apply the notion of cognitive extension to organisms as diverse as spiders, plants, and slime molds (see Facchin & Leonetti, 2024). If all these organisms are to be permitted entry to the active externalist club, then why should we reject the candidacy of an intelligent machine?

There are, however, a couple of reasons to think such claims might be premature. The first relates to a likely overlap in internal and external memory. As noted in Section 2.7, the responses obtained in the $\textbf{GPT}_{\text{MODIFIED}}$ condition were better than those in the $\textbf{GPT}_{\text{DEFAULT}}$ condition, in some cases approximating the responses delivered by Digital Andy. This suggests that the base GPT model (the model used across all test conditions) must have some awareness of Clark's philosophical work. In essence, information pertaining to topics such as the extended mind and predictive processing must already be a feature of the model's internal memory.[24] This leads to a concern about the role of the retrieval loop in shaping the responses of Digital Andy. If the retrieval loop is returning information that (to some extent) is already encoded in the parameter settings of the transformer network, then what should we make of the claim that extra-systemic information is being incorporated into the routines that define Digital Andy's generative capacities? Note that the reason for suggesting that Digital Andy ought to be understood as an EXAI system was, in part, based on a purported parallel between Otto and Digital Andy (see above). But the performance of the model in the $\textbf{GPT}_{\text{MODIFIED}}$ condition questions the validity of this comparison. Otto, we may assume, does not have the location of MoMA encoded in his inner neural nets (i.e., his bio-memory), or, at the very least, Otto does not have access to this information via an internally-directed retrieval loop. This is presumably one of the reasons why Otto resorts to the notebook instead of simply recalling the location of MoMA from bio-memory. In the case of Digital Andy, however, it seems that there must be some internal encoding of the relevant information, for the disabling of the retrieval loop (coupled with a request to impersonate Andy Clark) does not lead to a catastrophic collapse in the model's performance (i.e., the base GPT model was still able to generate reasonable responses to at least some user queries).

By attending to the details of the Otto/Inga case, we arrive at a further problem for the claim that Digital Andy ought to be understood as an extended system. This concerns the way in which the retrieval loop is being invoked. In the Otto/Inga case, it is the biological entity known as Otto that makes the call to external memory. In the case of Digital Andy, however, it is not the LLM (or GPT model) that is triggering the call to external memory. Rather, this call is made by the Digital Andy App, and this app lies external to the LLM. In this sense, then, Digital Andy is not at all like Otto. While Otto actively engages with external resources, triggering the call to external memory, Digital Andy is much more passive. In particular, Digital Andy does not 'decide' to trigger the retrieval loop; instead, he waits for information to be presented to him. The extent to which this poses a problem for Digital Andy's status as an extended system remains unclear, but there is nevertheless a difference here, and it is one that is deserving of further scrutiny.

There are thus a couple of reasons to doubt the status of Digital Andy as an EXAI system. The first relates to the seeming redundancy of the retrieval loop. If Digital Andy already knows about Andy Clark (and his philosophical work) via internal

[24] As noted by an anonymous reviewer, this is emblematic of a more general worry when it comes to the evaluation of LLMs. In particular the absence of information about the details of an LLM's training regime makes it difficult to gauge what the model might know (or perhaps what it should know). See Mitchell (2023), for more on this.

(parametric) memory, then the purported parallel with the Otto/Inga case is called into question. This issue is what we will call the *redundancy problem*. Its implications for claims regarding cognitive/computational extension are explored in Sect. 4.

The second problem relates to issues of agency and control. It stems from the fact that Digital Andy (understood as the LLM or base GPT model) is not the entity who is responsible for triggering the call to external memory. This call is instead made by an entity that lies external to Digital Andy, namely, the Digital Andy App. This raises a question about the extent to which claims of cognitive/computational extension are permissible in situations where the subject of extension (e.g., Otto or Digital Andy) is not the entity responsible for triggering the occurrence of a putatively extended process or (equivalently) the entity that triggers the instantiation of a mechanism that realizes the extended process. This is what we call the *agency problem*. Its implications for claims of cognitive/computational extension are explored in Sect. 5.

4 Redundant loops

In Sect. 2.7, we discussed the surprising quality of responses obtained in the **GPT-**$_{\text{MODIFIED}}$ condition—the condition where the base GPT model was tested with the same system prompt as that used in the two RAG-related conditions. As we saw in the previous section, this gives rise to a redundancy-related issue. Specifically, if the underlying GPT model is able to respond to queries about (e.g.) the extended mind without invoking the retrieval loop, then it must be the case that information about the extended mind is already part of the internal memory of the model. This marks a significant point of departure with the Otto/Inga case. While it is not explicitly stated by Clark and Chalmers (1998), we take it to be a widely held assumption that MoMA-related facts are *not* stored in Otto's bio-memory. Instead of these facts being located in Otto's internal bio-memory (as is the case with Inga), they are located in an externally-situated resource, namely, the notebook. In Section 3, we relied on the extended mind case (specifically, the comparison between Digital Andy and Otto) as a means of motivating the appeal to EXAI. Now, however, the force of this appeal is called into question by the apparent difference between Digital Andy and Otto. For Digital Andy, it seems there must be some sort of overlap between the contents of internal and external memory. Such overlaps, however, are not a feature of the original extended mind case.

As a means of probing this worry, let us introduce a modification to the Otto/Inga case. In particular, let us direct our attention to Inga. In the original thought experiment, Inga had no need of a notebook; she simply relied on the recall of information from bio-memory. Suppose, however, that Inga were to rely on a notebook resource in pretty much the same way as Otto. Thus, when Inga desires to visit MoMA, she consults the notebook, reads that MoMA is on 53rd Street, and heads off in the direction of MoMA. Here, the notebook encodings play precisely the same role in governing Inga's behavior as they do Otto's behavior. The only difference is that Inga could, in principle, have relied on her bio-memory to retrieve the same information as that yielded by the notebook. In essence, then, Inga has the choice of two retrieval strategies: either she can rely on the notebook, or she can rely on bio-memory.

The question is whether the mere availability of these multiple retrieval strategies has any bearing on claims pertaining to cognitive extension. In short, does it really matter that Inga has the option of retrieving information from bio-memory? If Inga opts to rely on the notebook to retrieve information about MoMA and such information should play the same sort of action-guiding role as it does for Otto, then what is the basis for distinguishing between these (seemingly identical) behavioral strategies? Is there a reason why we should reject the extended status of Inga's notebook-involving routine, solely on the basis that Inga has recourse to an alternative (in this case, non-extended) retrieval strategy?

As far as we can tell, the sort of situation being referred to here has not been the focus of prior philosophical work, and thus the existing literature fails to serve as a source of enlightenment. We can, of course, resort to using Digital Andy, although given that Digital Andy relies on the very same literature to which we, ourselves, are referring, it is hard to see how he might be able to shed any light on the matter. Nevertheless, let's present the situation to him and see what we get.

User: *In respect of the Otto and Inga thought experiment, I wonder what we should say if Inga relied on a notebook in precisely the same way as Otto. Would such acts of information retrieval count as a* bona fide *form of extended cognition, even though Inga could, in principle, have retrieved information from bio-memory. What do you think?*

DA$_{\textbf{SECTIONS}}$: In the context of the Otto and Inga thought experiment, if Inga relied on a notebook in precisely the same way as Otto, we should still consider her acts of information retrieval as a form of extended cognition. The fundamental point is the functional similarity between Inga and Otto's behaviors and the ways in which their information retrieval governs their actions. Both Otto's use of the notebook and Inga's reliance on her biological memory fit the profile of stored beliefs guiding behavior in a similar manner. The key factor is the role played by the retrieved information in guiding their current behaviors, regardless of whether the source is a biological memory or an external notebook. Therefore, yes, such acts of information retrieval from a notebook would indeed count as a form of extended cognition for Inga when considered in a similar functional context as Otto's use of the notebook.

According to Digital Andy, then, there is no reason why a seemingly redundant retrieval loop ought to undermine claims of cognitive extension. What matters is more the role played by information in guiding thought and action. Thus, if Inga relies on a notebook in more or less the same way as Otto, then we have little reason to reject the status of the retrieval loop as an extended cognitive routine.

If this is the correct response, then it is hard to see why the mere presence of redundant retrieval loops would pose much of a problem for the extended status of Digital Andy. Providing the responses of Digital Andy are informed by the retrieved information in a manner that resembles the influence exerted by parametric memory, then we have no good reason to discount the extended status of the call to external memory. To be sure, if the GPT model simply ignored the information, then we would have little reason to appeal to active externalist glosses. But insofar as the external information is factored into the model's responses, then it is hard to see why the mere presence of redundant loops would negate claims of cognitive/computational extension.

As a means of providing additional support for this conclusion, let us consider a somewhat different form of cognitive extension. This concerns the use of pen and paper resources to solve long multiplication problems. As noted by Wheeler (2010), this has become something of a stock example in the extended mind literature, with the proponents of active externalism insisting that world-involving variants of the long multiplication routine ought to be understood as a form of extended (mathematical) cognizing. For present purposes, let us accept this idea that world-involving variants of the long multiplication routine qualify as a form of extended cognizing. The question is whether the mere presence of an alternative (non-extended) strategy would materially alter claims about the extended status of these processes. As far as we can tell, the answer is "no." The reason for this is that many individuals possess the capacity to solve long multiplication problems in the head, even if they just so happen to resort to the use of pen and paper resources. Presented with a problem such as 777×999, an individual may find themselves inclined to reach for pen and paper (or other worldly) resources. But if such resources should be unavailable, the individual may still be able to solve the problem by resorting to an in-the-head method. The mere fact that both these strategies are available appears to have little bearing on whether or not we regard a particular strategy as extended (or non-extended). A single occurrence of the in-the-head strategy does not make every occurrence of the in-the-world strategy non-extended. Nor is there any reason to think that a single occurrence of the world-involving strategy would undermine the *non-extended* status of subsequent occurrences of the in-the-head strategy.

It seems, then, that the mere presence of redundant loops appears to pose little in the way of a problem for claims regarding cognitive extension. If this is the case when we are considering human-centered forms of cognitive extension, then the same ought to be true when we turn our attention to AI systems. In this sense, then, the redundancy problem is resolved.

Or is it? The redundancy problem is, indeed, resolved if Digital Andy should be factoring the retrieved information into his responses. But it is not resolved if Digital Andy should turn out to be ignoring this information. Consider that if Inga should retrieve information from her notebook, but then disregard this information, relying instead on bio-memory, then there would be nothing to substantiate the claim that Inga's (dispositional) beliefs supervene on the notebook-related encodings. Much the same can be said of Digital Andy. If Digital Andy should simply ignore the information retrieved from an external source, relying instead on internal (parametric) memory, then we would have little reason to regard the vector database as part of Digital Andy's extended memory.

The nub of the problem is thus the extent to which Digital Andy is factoring the retrieved information into his generative routines. Addressing this issue is not straightforward, but there are at least a couple of things that can be said about this matter.

Firstly, the prompt that is posted to Digital Andy includes an instruction statement that requests the model to "Use the *retrieved information* to answer the user's *query*" (see Fig. 4). By itself, of course, this doesn't really get us anywhere, for we have no more reason to think that the model is heeding the instruction than it is the retrieved information. Interestingly, however, the model shows a persistent tendency to refer to retrieved information when formulating the response. This isn't really evident from the responses reported here (see Supplementary Information), for the instruction statement includes a directive that is specifically intended to suppress this behavior (i.e., "In producing your response, avoid making any explicit reference to the fact that you are using the *retrieved information*"). Having said this, there are a couple of occasions where the reference to retrieved information persists. This can be seen in the following response to a query about the representational redescription hypothesis (see Clark & Karmiloff-Smith, 1993). It also appears in the response to a query about the Dark Room Problem (see below).

> **User:** *Tell me about the representational redescription hypothesis.*
>
> ---
>
> **DA$_{\text{SECTIONS}}$:** The representational redescription (RR) hypothesis, *as discussed in the retrieved information*, suggests that the evolution of cognitive systems is characterized by the emergence of representational transformations. These transformations involve the progressive redescribing of information in a manner that relaxes the constraints on its use. The process of representational redescription is the means by which an organism exploits and manipulates information that is already encoded in various specialized ways to achieve higher orders of flexibility. ... [emphasis added]

A further reason to think that retrieved information is being incorporated into the model's responses relates to the nature of the responses themselves. Thus, while the responses in the **GPT**$_{\text{MODIFIED}}$ condition were better than expected, they are certainly not the same as those in the two RAG-related conditions. In addition, some of the responses in the **GPT**$_{\text{MODIFIED}}$ condition are not what we would expect given the focus of Andy Clark's work. In response to a query about the Dark Room Problem, for example, the model in the **GPT**$_{\text{MODIFIED}}$ condition produces the following:

> **User:** *What is the dark room problem, and how might we solve it?*
>
> ---
>
> **GPT$_{\text{MODIFIED}}$:** The dark room problem is a thought experiment that challenges our understanding of perception and consciousness. It asks whether a person who is placed in a completely dark room is conscious of anything. This thought experiment raises questions about whether consciousness requires sensory input and to what extent our perception is shaped by our environment. ...

In this case, the model fails to report the connection between the Dark Room Problem and predictive processing, which is a prominent feature of Clark's work (Clark, 2018; Friston et al., 2012). This shortcoming is not apparent in the two RAG-related conditions. In the**DA**$_{\text{SECTIONS}}$ condition, for example, the response begins as follows:

User: *What is the dark room problem, and how might we solve it?*

DA$_{\text{SECTIONS}}$: The Dark Room problem, as outlined in the retrieved information, presents a challenging puzzle for the theory of predictive processing (PP). At its core, the Dark Room problem raises the question of how sentient beings, including humans, exhibit a general striving toward novelty, pleasure, exploration, and fulfillment, despite the fact that predictive processing might seem to prioritize minimizing prediction errors. . . .

Such responses suggest that the base GPT model is not simply ignoring the retrieved information. It may be the case that the influence exerted by the retrieved information varies from one query to the next, but there is no reason to think that the retrieved information is failing to inform the model's responses.[25]

As a final remark on the redundancy problem, it is worth noting that the problem stems from the focus on the work of Andy Clark. Given the popularity of Andy's ideas, coupled with his extensive publication history, it would not be a surprise to discover that systems like ChatGPT have been exposed to Andy's work as part of their training history. This, however, is merely a contingent feature of the Digital Andy System. If we had opted to focus on the works of a lesser-known philosopher, then the redundancy problem may not have arisen. And it certainly would not have arisen if the contents of external memory had been populated with information that lies outside the public domain (e.g., a body of unpublished works). The point, here, is that redundancy-related issues have no real bearing on the mere possibility of extended LLMs and the use of RAG to implement these systems. Such issues are, *at best*, a problem in situations where we encounter overlaps in internal and external memory. And even then, the implications for claims of cognitive/computational extension are far from settled.

5 The agency problem

Having dealt with the redundancy problem, let us now turn our attention to the agency problem. The agency problem, recall, centers on the flow of information between the various constituents of the Digital Andy system. When the user enters a query into the Digital Andy App, the App executes a query that retrieves information from the vector database. The App then inserts this information into the context window before posting the prompt to the GPT model. The problem is that we have equated Digital Andy with the entity that generates responses (i.e., the GPT model or LLM), but this entity plays no role in either triggering the retrieval loop or assimilating the retrieved information into the context window. This marks a point of departure with many of the cases of cognitive/computational extension that have been discussed in the active

[25] More generally, our own (informal) evaluation of the quality of responses across the various conditions suggests that the responses in the **GPT**$_{\text{MODIFIED}}$ condition were not as good those in the two Digital Andy conditions (i.e.,**DA**$_{\text{PARAGRAPHS}}$ and**DA**$_{\text{SECTIONS}}$) (see Supplementary Information SI.7).

externalist literature. For the most part, claims of cognitive extension are centered around a particular individual who plays a role in the creation and coordination of extended circuits. In the Otto/Inga case, for example, it is the biological individual known as Otto who triggers the call to external memory, and he does so in response to a particular folk psychological state, namely, a desire to visit MoMA. There is no counterpart to this in the case of Digital Andy. Rather than being an active participant in the processes that shape its own generative performances, the entity we are calling Digital Andy is merely the passive recipient of information that was collated on its behalf.

Does this negate the idea that Digital Andy ought to be understood as an EXAI system? The answer is, unfortunately, unclear, for we are (once again) dealing with a rather unusual state-of-affairs, one that seldom arises in the more familiar (human-centered) cases. As with the redundancy problem, we will try to solicit support from Digital Andy, asking the model to consider a situation where the subject of cognitive extension is not the entity who performs the retrieval process:

User: *In the Otto/Inga case, what would happen if another individual, George let's say, did all the information retrieval on Otto's behalf. That is to say, what would we say if George was retrieving information from the notebook and presenting it to Otto. Would we still say that Otto's mind was extended?*

DA$_{\text{SECTIONS}}$: In the case of another individual, such as George, doing all the information retrieval on Otto's behalf by retrieving information from the notebook and presenting it to Otto, we could argue that Otto's mind is still extended. The essence of the extended mind thesis is not solely tied to the physical location of the information retrieval process but rather to the coupled system that exists between an agent and external resources.

If we view the cognitive system as a dynamic interaction between the individual and the external resource, the actions of George could still be seen as part of the extended cognitive network. Just as Otto's habitual reliance on the notebook plays a part in shaping his behavior and beliefs, George's role in aiding the retrieval and presentation of information could also be considered as contributing to the extended cognitive processes of Otto.

In this sense, the extended mind thesis suggests that the boundaries of the mind are not confined to the individual's physical body but encompass the integrated system that includes external tools and other individuals participating in the cognitive processes. Therefore, even if George is the one retrieving and presenting information to Otto, his involvement could still be seen as part of the extended cognitive system, thus supporting the idea that Otto's mind is extended.

What Digital Andy seems to be suggesting here is that claims regarding the extended mind could be sustained, even if an information retrieval process were to be delegated to some entity that is not the subject of cognitive extension.

Is that the end of the matter? Probably not, for we have no reason to think that the relationship between Otto and George (in the above scenario) is the same as that between Digital Andy and the App. If George is responding to a request from Otto, or George is (somehow) monitoring Otto in a way that anticipates his information needs, then perhaps we could persist with claims regarding the extended mind. In this case, Otto could continue to be understood as the bearer of dispositional beliefs on the grounds that such (dispositional) characterizations yield a predictively- and explanatorily-potent (folk psychological) grip over Otto's actual and counterfactual behavior.[26] George would then be a constituent of the "extended cognitive network" that underwrites these dispositional ascriptions.

The problem is that the flow of information within the Digital Andy System makes it difficult to see how Digital Andy (as the putative *subject* of cognitive/computational extension) could be credited with the possession of dispositional properties in quite the same way as Otto. At the very least, it is hard to see why Digital Andy ought to be regarded as the central element of an "extended cognitive network" that includes (*inter alia*) the Digital Andy App. Digital Andy, recall, is being invoked by the App, not the other way round. Why, then, should the App be regarded as a constituent of Digital Andy's extended cognitive/computational routines, as opposed to Digital Andy being a constituent of the App's extended cognitive/computational routines? There is nothing about the dynamics of the information flow that would preclude this possibility. Quite the opposite, in fact. All the information processing routines of the Digital Andy System begin and end with the App. It is the App that accepts the user query, the App that retrieves information from the vector database, the App that posts the prompt to Digital Andy, and the App that receives the model's responses prior to presenting them to the user.

In view of this, it is at best unclear that Digital Andy ought to be understood as the subject of cognitive/computational extension. To be sure, the matter is not entirely cut and dried,[27] but the best examples of EXAI are arguably those that approximate the general features of cases discussed in the active externalist literature. A common feature of these cases is that extended processes are triggered by a particular individual (e.g., Otto), and such processes also exert effects on that individual (e.g., the retrieval loop influences Otto's overt behavior).[28] This is not the case with Digital Andy. In the Digital Andy System, the GPT model plays no role in triggering the retrieval process, nor does the retrieval process *directly* influence the GPT model. These effects are instead mediated by the Digital Andy App.

Unfortunately, then, there are reasons to think that Digital Andy fails to make the grade as an EXAI system. This, however, is not quite the end of the story; for Digital Andy is just one instance of a more general class of RAG-based LLMs, and not every member of this class is beset by the agency problem. Consider, for example, recent research into what is dubbed *active RAG* (e.g., Jiang et al., 2023). In active RAG, the LLM is given a choice of strategies: it can either rely on inner (parametric) memory, or it can initiate a call to external (non-parametric) memory. The choice

[26] As Clark (2011, p. 449) notes: "Unlike Inga, Otto is not a fully normal agent, but courtesy of the extra-biological machinery, much more of his behaviour can be successfully subsumed under our familiar folk psychological kinds: kinds such as 'believing that MOMA is on 53rd street'."

[27] It is perhaps worth noting that not all the authors agree on the importance of the agency problem, or its implications for claims of extended status. The views expressed here are those of the lead author.

[28] As noted by a number of authors, extended processes tend to have a circular dynamic or a 'loopy' motif (Clark, 2010b; Palermos, 2014).

between these strategies is governed by a sort of 'metacognitive' judgment pertaining to the model's capacity to produce a high quality response.[29] In situations where the reliance on internal memory would culminate in a low-quality response, the model triggers a call to external memory via the retrieval loop. The resultant information is then factored into the model's generative processes, just as we have described for Digital Andy.

Note that the shift to active RAG addresses concerns that lie at the core of *both* the redundancy and the agency problem. The redundancy problem is resolved courtesy of the way the retrieval loop is coordinated with internal memory. That is to say, the retrieval process is only invoked when the LLM anticipates a shortfall in its capacity to produce high-quality responses courtesy of internal memory. At the same time, active RAG addresses the worry about the retrieval loop not being sufficiently integrated with the workings of the LLM. More specifically, active RAG gives the LLM greater control over the retrieval process, enabling the LLM to make repeated calls to an extra-systemic resource as a way of bolstering its own performance. In this case, the capacity to interact with an external resource could be seen to lay the foundation for another sort of dispositional ascription, such as a capacity to generate high-quality responses. Such ascriptions, it should be clear, are ones that tend to paper over the location-based distinction between internal and external memory.

Moving beyond the realms of information retrieval, recent research has begun to explore the more general opportunities that LLMs have to interact with elements of their online ecology. A nice example of this stems from the effort to connect ChatGPT with the Wolfram Mathematica system, yielding a hybrid system called Wolfram GPT.[30] The motivation for this particular form of technological symbiosis stems from the inherent limitations of LLMs in regard to certain tasks. Consider, for example, the following error, reported by Floridi and Chiriatti (2020):

> GPT-3 works in terms of statistical patterns. So, when prompted with a request such as "solve for x: x + 4 = 10" GPT-3 produces the correct output "6", but if one adds a few zeros, e.g., "solve for x: x + 40000 = 100000", the outcome is a disappointing "50000"… Confused people who may misuse GPT-3 to do their maths would be better off relying on the free app on their mobile phone. (Floridi and Chiriatti, 2020, p. 688)

Although we were not able to reproduce this particular problem using the current version of ChatGPT, we do not doubt the more general point that Floridi and Chiriatti (2020) are making here, namely, that LLMs exhibit errors when confronted with tasks that rely on specialized skills and knowledge (in this case, a capacity for mathematical reasoning). As Floridi and Chiriatti (2020) suggest, if one wants to solve a mathematical problem, one would probably be better off relying on a tool that is tailored for this particular purpose (e.g., a mobile app).

[29] This judgment relies on the token probabilities associated with a model's initial response (see Kadavath et al., 2022).

[30] See https://www.wolfram.com/wolfram-plugin-chatgpt/.

Note, however, that a mobile app qualifies as a bio-external resource, and the human user is being asked to solve a problem courtesy of their interaction with this resource. Given this, it seems that the biological counterpart to the LLM's transformer network (i.e., the individual's brain) is no better placed to solve this particular problem than is the LLM. If it is the case that humans are poorly equipped to solve certain types of problems using the resources of the bare biological brain, then why assume that LLMs (or, indeed, any other type of specialized AI system) would be in a much better position? Floridi and Chiriatti (2020) are probably correct to state that humans are better off reaching for an external tool to do their maths. But the same is arguably true of LLMs. To be sure, LLMs are perhaps ill-equipped to solve certain sorts of problems, such as those involving complex mathematical calculations. But is the technologically-denuded human brain in a radically different position? And if the solution to our own biologically-based limitations is to reach for a bio-external resource, then why not allow LLMs to do the same. In short, why not simply acknowledge that, for certain types of problems (including those of the mathematical variety), LLMs would be better off 'reaching' for an external tool, just as we do?

As it turns out, this notion of LLMs 'reaching' for external tools resonates with much of the recent research into LLMs. Of particular interest are what are dubbed *augmented language models* (Mialon et al., 2023). These are LLMs with the capacity to interact with other systems and services, such as the aforementioned merger of ChatGPT with Wolfram Mathematica. Notable examples of such systems include the likes of LaMDA (Thoppilan et al., 2022), Toolformer (Schick et al., 2023), and Mind's Eye (Liu et al., 2022). In addition, recent research into LLMs has led to the emergence of so-called *agentic LLMs*, which have the capacity to trigger retrieval operations and invoke external services, even in the absence of explicit instructions. Indeed, readers may have noticed that the latest version of ChatGPT features a search capability that enables the LLM to search the Web for query-relevant information, even in the absence of an explicit instruction to do so.[31]

Each of these efforts strives to augment the capabilities of LLMs, enabling them to tackle problems that would otherwise lie beyond the reach of their (internal) neural nets. In addition, the growing interest in augmented and agentic LLMs is moving us towards an era in which the aforementioned agency problem does not apply. And once the agency-related problem is removed, what remains? Do we expand the scope of active externalist theorizing to accommodate the possibility of EXAI? Or do we insist that active externalism is best served by a selective (perhaps exclusive) focus on biological forms of intelligence? Either way, the issues that emerge as a result of this debate are likely to be just as relevant to our attempt to understand the extended mind as they are our effort to build the next generation of intelligent systems.

[31] For more on OpenAI's efforts to support interaction with external, third-party services, see their 'actions' framework: https://platform.openai.com/docs/actions/introduction.

6 Conclusion

We began by considering two ways in which philosophical knowledge might be incorporated into an LLM. The first approach is exemplified by the effort to build DigiDan—a digital replica of the philosopher Daniel Dennett (Schwitzgebel et al., 2024). In this case, machine learning techniques are used to tweak the internal parameters of an LLM, thereby assimilating knowledge into what is called internal memory.

The second approach is exemplified by the effort to build Digital Andy—the focus of the present paper. In this case, we relied on a technique, called RAG, which uses external memory to condition the responses of an LLM. By applying this technique to the works of Andy Clark, we were able to tailor the responses of an LLM without the need for additional machine learning.

As it happens, the selective focus on Andy Clark (coupled with the commitment to RAG) reveals a point of philosophical interest concerning the extended status of AI systems. In particular, the distinction between internal and external memory establishes a point of contact with claims regarding the extended mind—the very issues with which the real-world Andy Clark is concerned. Ultimately, we rejected the extended status of Digital Andy on the grounds that the LLM played no role in triggering the relevant retrieval loop (i.e., the call to external memory). But this 'shortcoming' positions us at the very forefront of LLM-related research, revealing how multiple efforts are seeking to equip LLMs with the capacity to inter-operate with a wider nexus of online tools and services.

The upshot is a host of issues and concerns that transcend the realms of both philosophy and technology. Can AI systems have extended minds? Is cognitive extension a route to Artificial General Intelligence (AGI)? What are the mechanisms that enable an intelligent system to switch between a purely inner routine and one that reaches out to a wider web of resources? And what is the peculiar role of language in enabling both ourselves and the machines we build to press maximal cognitive and computational benefit from our socio-technical surrounds? Such questions are of independent philosophical interest, but they are also, crucially, questions that connect philosophical debates about the extended mind with the emerging focus of research into LLMs. And perhaps this is the best way of understanding Digital Andy—not as an EXAI system, but as more of a technological artifact that prompts a shift in individual and collective attention, helping to expand the scope of active externalism and reveal new lines of philosophical inquiry. There are a number of reasons to welcome this shift. A consideration of EXAI reveals issues that may be difficult to discern given a selective preoccupation with human-centered forms of cognitive extension. In addition, we are beginning to see a degree of convergence regarding the philosophical effort to understand the extended mind, and the more practical project of building intelligent systems. A technologically-inflected version of active externalism may be the just the thing we need to prompt new lines of inquiry into the theory and practice of the extended mind.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11229-025-05046-y.

Acknowledgements None.

Author contributions Paul Smart: data preparation; system implementation; paper writing. Rob Clowes: testing and evaluation; paper writing. Andy Clark: reviewing, editing and feedback.

Funding [Paul Smart] This work was supported by the U.K. Research and Innovation (UKRI) and the UKRI Trustworthy Autonomous Systems Hub under Grant Number EP/V00784X/1. [Rob Clowes] This work is supported by the GENAI Project (Exploring Human Cognitive Flourishing against the background of GENerative AI) awarded by FCSH, Universidade NOVA de Lisboa and by the Portuguese Foundation for Science and Technology under Grant Number UIDB/00183/2020.

Data availability Test data is included as Supplementary Information. An online version of the Digital Andy system can be accessed at https://digitalandy.ai/.

Code availability None.

Declarations

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent for publication Approval given by all authors.

Competing interests The authors have no competing interests to declare that are relevant to the content of this article.

References

Adams, F. (2010). Why we still need a mark of the cognitive. *Cognitive Systems Research*, *11*(4), 324–331. https://doi.org/10.1016/j.cogsys.2010.03.001

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021) On the dangers of stochastic parrots: Can language models be too big? In: M. C. Elish, W. Isaac, & R. Zemel (Eds.) ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, Canada, pp 610–623, https://doi.org/10.1145/3442188.3445922

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (pp. 1877–1901).

Clark, A. (1999). Where brain, body, and world collide. *Cognitive Systems Research*, *1*(1), 5–17. https://doi.org/10.1016/S1389-0417(99)00002-9

Clark, A. (2003). *Natural-born cyborgs: Minds, technologies and the future of human intelligence*. Oxford University Press.

Clark, A. (2008). *Supersizing the mind: Embodiment, action, and cognitive extension*. Oxford University Press.

Clark, A. (2010a). Memento's revenge: The extended mind, extended. In R. Menary (Ed.), *The extended mind* (pp. 43–66). MIT Press. https://doi.org/10.7551/mitpress/8535.003.0003.
Clark, A. (2010b). Much ado about cognition. *Mind*, *119*(476), 1047–1066. https://doi.org/10.1093/mind/fzr002
Clark, A. (2011). Finding the mind. *Philosophical Studies*, *152*(3), 447–461. https://doi.org/10.1007/s11098-010-9598-9
Clark, A. (2018). A nice surprise? Predictive processing and the active pursuit of novelty. *Phenomenology and the Cognitive Sciences*, *17*(3), 521–534. https://doi.org/10.1007/s11097-017-9525-z
Clark, A. (2023). *The experience machine: How our minds predict and shape reality*. Pantheon Books.
Clark, A. (2024). Extending the predictive mind. *Australasian Journal of Philosophy*, *102*(1), 119–130. https://doi.org/10.1080/00048402.2022.2122523
Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, *58*(1), 7–19. https://doi.org/10.1093/analys/58.1.7
Clark, A., & Karmiloff-Smith, A. (1993). The cognizer's innards: A philosophical and psychological perspective on the development of thought. *Mind & Language*, *8*(4), 487–519. https://doi.org/10.1111/j.1468-0017.1993.tb00299.x
Facchin, M., & Leonetti, G. (2024). Extended animal cognition. *Synthese*, *203*(Article 138), 1–22. https://doi.org/10.1007/s11229-024-04579-y
Floridi, L. (2023). AI as *agency without intelligence*: On ChatGPT, large language models, and other generative models. *Philosophy and Technology*, *36*(Article 15), 1–7. https://doi.org/10.1007/s13347-023-00621-y
Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, *30*, 681–694. https://doi.org/10.1007/s11023-020-09548-1
Frankish, K. (2024). What are large language models doing? In A. Strasser (Ed.), *Anna's AI anthology: How to live with smart machines?* (pp. 53–79). xenomoi.
Friston, K., Thornton, C., & Clark, A. (2012). Free-energy minimization and the dark-room problem. *Frontiers in Psychology*, *3*(130), 1–7. https://doi.org/10.3389/fpsyg.2012.00130
Gao, Y., Xiong, Y., Gao, X., Jia, Pan, J., Bi, Y., Dai, Y., Sun, J., & Wang, H. (2024). Retrieval-augmented generation for large language models: A survey. preprint at https://arxiv.org/abs/2312.10997https://doi.org/10.48550/arXiv.2312.10997
Heersmink, R., de Rooij, B., Clavel Vázquez, M. J., & Colombo, M. (2024). A phenomenology and epistemology of large language models: Transparency, trust, and trustworthiness. *Ethics and Information Technology*, *26*(Article 41), 1–15. https://doi.org/10.1007/s10676-024-09777-3
Herrmann, D. A., & Levinstein, B. A. (2025). Standards for belief representations in LLMs. *Minds and Machines*, *35*(1), 1–25. https://doi.org/10.1007/s11023-024-09709-6
Jiang, Z., Xu, F. F., Gao, L., Sun, Z., Liu, Q., Dwivedi-Yu, J., Yang, Y., Callan, J., & Neubig, G.(2023). Active retrieval augmented generation. In: H. Bouamor, J. Pino, & K. Bali (eds) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics (pp. 7969–7992). Singapore. https://doi.org/10.18653/v1/2023.emnlp-main.495
Jonker, C. M. (2008). Extended mind from an artificial intelligence perspective. In L. Zonneveld, H. Dijstelbloem, & D. Ringoir (Eds.) *Reshaping the human condition: Exploring human enhancement* (pp. 133–140). Rathenau Institute, The Hague.
Joshi, A. (2024). How to choose the right chunking strategy for your LLM application. https://www.mongodb.com/developer/products/atlas/choosing-chunking-strategy-rag/
Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., & Tran-Johnson, E. (2022). Language models (mostly) know what they know. preprint at https://arxiv.org/abs/2207.05221https://doi.org/10.48550/arXiv.2207.05221
Kaplan, D. M. (2012). How to demarcate the boundaries of cognition. *Biology & Philosophy*, *27*(4), 545–570. https://doi.org/10.1007/s10539-012-9308-4
Kasirzadeh, A., & Gabriel, I. (2023). In conversation with artificial intelligence: Aligning language models with human values. *Philosophy and Technology*, *36*(2), 1–24. https://doi.org/10.1007/s13347-023-00606-x
Kersten, L. (2017). A mechanistic account of wide computationalism. *Review of Philosophy and Psychology*, *8*(3), 501–517. https://doi.org/10.1007/s13164-016-0322-3
Kersten, L. (2024). Wide computationalism revisited: Distributed mechanisms, parsimony and testability. *Philosophical Explorations*, *27*(3), 280–297. https://doi.org/10.1080/13869795.2024.2332171

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 9459–9474).

Liu, R., Wei, J., Gu, S. S., Wu, T.-Y.,Vosoughi, S., Cui, C., Zhou, D., & Dai, A. M. (2022). Mind's eye: Grounded language model reasoning through simulation. preprint at https://arxiv.org/abs/2210.05359https://doi.org/10.48550/arXiv.2210.05359

Mars, M. (2022). From word embeddings to pre-trained language models: A state-of-the-art walkthrough. *Applied Sciences*, *12*(Article 8805), 1–19. https://doi.org/10.3390/app12178805

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., & Scialom, T. (2023). Augmented language models: A survey. *Transactions on Machine Learning Research*, 1–35. https://doi.org/10.48550/arXiv.2302.07842

Mitchell, M. (2023). How do we know how smart AI systems are? *Science*, *381*(6654), eadj5957. https://doi.org/10.1126/science.adj5957

Monigatti, L. (2023). A guide on 12 tuning strategies for production-ready RAG applications. https://towardsdatascience.com/a-guide-on-12-tuning-strategies-for-production-ready-rag-applications-7ca646833439

Palermos, S. O. (2014). Loops, constitution, and cognitive extension. *Cognitive Systems Research*, *27*, 25–41. https://doi.org/10.1016/j.cogsys.2013.04.002

Pellegrino, G., & Garasic, M. D. (2020). Artificial intelligences as extended minds. why not? *Rivista Internazionale di Filosofia e Psicologia*, *11*(2), 150–168. https://doi.org/10.4453/rifp.2020.0010

Pezzulo, G., Parr, T., Cisek, P., Clark, A., & Friston, K. (2024). Generating meaning: Active inference and the scope and limits of passive AI. *Trends in Cognitive Sciences*, *28*(2), 97–112. https://doi.org/10.1016/j.tics.2023.10.002

Ram, O., Levine, Y., Dalmedigos, I., Muhlgay, D., Shashua, A., Leyton-Brown, K., & Shoham, Y. (2023). In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, *11*, 1316–1331. https://doi.org/10.1162/tacl_a_00605

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in neural information processing systems* (Vol. 36, pp. 68539–68551). Curran Associates, Inc.

Schwitzgebel, E., Schwitzgebel, D., & Strasser, A. (2024). Creating a large language model of a philosopher. *Mind & Language*, *39*(2), 237–259. https://doi.org/10.1111/mila.12466

Smart, P. R. (2018). Human-extended machine cognition. *Cognitive Systems Research*, *49*, 9–23. https://doi.org/10.1016/j.cogsys.2017.11.001

Smart, P. R. (2024). Extended X: Extending the reach of active externalism. *Cognitive Systems Research*, *84*(Article 101202), 1–12. https://doi.org/10.1016/j.cogsys.2023.101202

Strasser, A., Crosby, M., & Schwitzgebel, E. (2022). How far can we get in creating a digital replica of a philosopher? In R. Hakli, P. Mäkelä, & J. Seibt (Eds.) *Social robots in social institutions* (pp. 371–380). IOS Press.

Thoppilan, R., Freitas, D. D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., & Du, Y. (2022). LaMDA: Language models for dialog applications. preprint at https://arxiv.org/abs/2201.08239https://doi.org/10.48550/arXiv.2201.08239

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In U. von Luxburg, I. Guyon, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (pp. 1–11).

Wang, X., Wang, Z., Gao, X., Zhang, F., Wu, Y., Xu, Z., Shi, T., Wang, Z., Li, S., & Qian, Q. (2024). Searching for best practices in retrieval-augmented generation. In: Y. Al-Onaizan, M. Bansal, & Y. N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, pp 17716–17736. https://doi.org/10.18653/v1/2024.emnlp-main.981

Wheeler, M. (2010). In defense of extended functionalism. In R. Menary (Ed.), *The extended mind* (pp. 245–270). MIT Press. https://doi.org/10.7551/mitpress/8535.003.0011.

Wilson, R. A. (2014). Ten questions concerning extended cognition. *Philosophical Psychology*, *27*(1), 19–33. https://doi.org/10.1080/09515089.2013.828568

Wilson, R. A., & Clark, A. (2009). Situated cognition: Letting nature take its course. In P. Robbins & M. Aydede Eds., *The Cambridge handbook of situated cognition* (pp. 55–77). Cambridge University Press.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Paul Smart[1] · **Robert Clowes**[2] · **Andy Clark**[3]

✉ Paul Smart
ps02v@ecs.soton.ac.uk

Robert Clowes
robert.clowes@gmail.com

Andy Clark
Andy.Clark@sussex.ac.uk

1 Electronics and Computer Science, University of Southampton, University Road, Southampton, Hampshire SO17 1BJ, UK

2 Nova Institute of Philosophy, Universidade Nova de Lisboa, Campus de Campolide - Colégio Almada Negreiros, Lisbon 1099-032, Portugal

3 Department of Philosophy, The University of Sussex, Sussex House, Brighton, East Sussex BN1 9RH, UK