Citation: Smart, P. R., & Clowes, R. W. (2026) Building AGI One Word at a Time. In I. Serageldin, D. Grigar & F. Hegland (Eds.), *The Future of Text VI: A Myriad of Perspectives*. Future Text Publishing, London, UK.

Building AGI One Word at a Time

Paul Smart and Robert W. Clowes

Artificial General Intelligence (AGI) is often described as the "holy grail" of artificial intelligence. Unlike systems that specialize in the performance of specific tasks—like language translation, image recognition, or strategic gameplay—AGI aims at a more general and flexible form of intelligence. While there is no consensus on what it means for an AI system to qualify as generally intelligent, the term "AGI" is typically understood in relation to human intelligence. In particular, a system exhibiting general intelligence should have the capacity to perform a wide variety of cognitive tasks at a level that matches (or perhaps even surpasses) that of a typical human being.

The appeal to human intelligence is significant, for it directs attention to the mechanisms that underpin our species-specific form of cognitive success. These mechanisms are clearly relevant to the project of explaining human intelligence, but they may also serve as a template for the construction of AGI systems. In this respect, the project of building AGI and the project of explaining human intelligence are not independent endeavours. Progress in one domain can inform the other, particularly when the aim is to emulate—functionally or architecturally—the features that allow human beings to excel at a wide variety of cognitive tasks.

It might be thought that the search for the material bases of human cognitive success has a rather obvious target. If our aim is to identify the mechanisms responsible for human cognitive performances, then surely we need look no further than the borders of the biological brain. This is, of course, true if the relevant mechanisms turn out to be nothing more than neural mechanisms—mechanisms whose borders never exceed the spatial extent of the nervous system. But some theorists insist that human cognitive mechanisms are more than just neural mechanisms. According to the proponents of what has come to be known as active externalism or the extended mind, the mechanisms responsible for human cognitive performances can, on occasion, extend beyond the biological borders of skin and skull, incorporating resources from the wider environment. On this view, notebooks, diagrams, linguistic inscriptions, digital devices, and other artefacts are the potential realizers of human cognitive states and processes. They are, quite literally, part of the machinery of the human mind.

We thus have two visions of human intelligence. The first is what we might call the neurocentric (or non-extended) view. According to this view, human intelligence is tied to the operation of neural mechanisms, i.e., mechanisms that are wholly situated within the brain (or, at any rate, the nervous system). The second view is what might be called the extended view. According to this view, human intelligence is tied to the operation of what are called *extended mechanisms*—mechanisms that extend beyond the ancient metabolic boundaries of skin and skull.

Transposing the neurocentric and extended views to AI yields contrasting images of what an AGI system might look like. From the perspective of the neurocentric view, an AGI system is a self-contained cognitive system that is able to perform a variety of cognitive tasks due to the whirrings and grindings of its inner cognitive architecture (e.g., the operations implemented by a deep neural network). On the extended view, however, an AGI system is a system that is able to perform a variety of cognitive tasks courtesy of the instantiation of extended mechanisms. What makes this latter type of system a fitting candidate for AGI is not so much its capacity to perform cognitive tasks in the 'head'; rather, it is its capacity to extend its own cognitive reach by factoring external resources into its cognitive and computational routines. Such a system is not so much a 'natural-born' master of many cognitive tasks; it is more a specialist at a single cognitive task: the task of

building extended mechanisms. It is this particular cognitive specialism (a capacity to build extended mechanisms) that enables the system to achieve the sort of flexibility and generality required for AGI. AGI, on this view, is a form of *extended intelligence*—a form of intelligence that is rooted in a basic capacity to build (and benefit from) extended mechanisms.

Much of the contemporary research into AGI is focused on the development of increasingly capable large language models (LLMs). This focus is perhaps unsurprising given the centrality of language to human cognition and the impressive performances of systems like ChatGPT, Gemini, and Grok. Attempts to improve the capabilities of LLMs typically involve building models with more internal parameters. The GPT-3 model, for example, was reported to have a total of 175 billion parameters, while GPT-4 is estimated to have approximately 1.75 trillion: a tenfold increase. At first sight, this approach might seem more compatible with a non-extended approach to AGI. After all, the project of building ever-larger LLMs is akin to the project of building ever-bigger 'brains'. But we should not be so quick to dismiss the extended view. While it is true that larger models may be capable of performing a greater number of tasks courtesy of their inner neural nets, scaling may also produce LLMs that are better able to exploit their surrounding environment. In other words, larger models may not simply be more powerful self-contained cognitive engines; they may also be more adept at constructing and leveraging extended mechanisms.

To appreciate this point, it is worth noting that many of the features that make LLMs a compelling target for AGI research are plausibly tied to the operation of extended mechanisms. That is to say, the main reason why LLMs appear to provide us with our best chance of achieving AGI is precisely because they already trade in various forms of extended intelligence. Consider, for example, the way in which an LLM uses its own linguaform outputs to guide and constrain subsequent processing. When prompted to “think step-by-step,” LLMs generate intermediate token sequences—chain-of-thought (CoT) traces—that decompose complex tasks into manageable sub-problems. In effect, the model uses its own outputs as a form of linguistic scaffolding, one that enriches the context for subsequent generations and progressively steers the model in the direction of a correct solution. Empirically, this approach outperforms the more direct, "one shot" approach to response generation. When the model is prevented from 'talking' itself through a problem, it is forced to make a dramatic leap in inferential space, using only the resources of its inner neural network. No surprise, then, that as the complexity of the problem increases, the greater the chances of the model falling short. The progressive generation of intermediate tokens helps to resolve this issue, with each token serving as a sort of stepping stone through inferential space. Linguistic scaffolding thus pays substantive cognitive dividends, allowing the model to tackle problems that might otherwise prove infeasible. Such benefits mirror those attributed to human writing practices. When we write, sketch, or otherwise externalise our thoughts, we create stable, manipulable structures that feed back into our reasoning processes. Such feedback loops—which run outside the head—can often help us perform tasks that might be difficult (if not impossible) were we to rely solely on the resources of the biological brain.

Another example of extended intelligence stems from recent work into augmented or tool-using LLMs. The core idea, here, is that an LLM is able to interact with external resources by generating the textual (and, more specifically, programmatic) commands required to invoke those resources. Suppose, for the sake of example, that an LLM is presented with a question that lies beyond the scope of its internal knowledge (i.e., the knowledge contained in its training data). If the LLM were to rely solely on its internal architecture, then it would either produce the incorrect answer, or respond by saying that it doesn't know. But the LLM can also rely on the wider environment to address this epistemic shortcoming. In particular, it can issue a call to a search engine, generating search terms to

guide the search, and factoring the search results into its response. The same applies to situations where an LLM is required to perform tasks that it would otherwise be unable to perform. Consider an LLM that lacks the capacity to interpret images. Such a model is, in effect, visually blind; yet it can circumvent this limitation by calling on an external vision system. By generating a sequence of API calls, the model can progressively analyse the image contents, with the results of one call informing the structure of subsequent steps. Perceptual success, in this case, stems not from the LLM's capacity to understand or interpret an image courtesy of its inner neural architecture; rather, the LLM possesses a capacity to create an extended mechanism that leverages the functionality of a remotely-situated (or, at any rate, external) resource. Once again, it is the LLM's facility with language that provides the basis for this particular form of cognitive success. Rather than being an all-powerful, self-contained cognitive engine, the use of language enables the LLM to function more like an intelligent conductor, orchestrating calls to the wider environment in a manner that befits the demands of different tasks.

A third, and in some ways even more striking, example of extended intelligence occurs when LLMs generate executable code. Consider a case where an LLM is tasked with counting the number of occurrences of the letter "r" in the word "strawberry". Such tasks are notoriously difficult for LLMs, in part because they rely on tokenization schemes, such as Byte Pair Encoding (BPE), that operate above the level of single characters. But the model can effectively bypass these constraints by using its facility with language (and other linguaform structures) to generate a sequence of programmatic statements that solves the problem via a different computational route. Presented with the aforementioned problem, for example, an LLM might generate the following code:

```
word = "strawberry"
count_r = word.count('r')
print(count_r)
```

This code can then be executed by a code interpreter and the result returned to the model. The upshot is that the LLM is now able to solve a class of problems that might be difficult or impossible to solve using a neural network architecture. The model may not be particularly good at solving mathematical or logical problems in the 'head', but it can nevertheless tackle these problems by drawing on its core competence—a facility with language—to create an ad-hoc computational tool that delivers the requisite functionality. In such cases, it hardly seems correct to say that the LLM is incapable of performing tasks such as those of the letter-counting variety. The same applies to any other computational task. Given sufficient knowledge of a programming language, an LLM can, in principle, perform any task that can be performed by a conventional symbol-crunching digital computer. And this is despite the fact that an LLM—as a subsymbolic entity—is congenitally ill-equipped to perform these tasks.

Across all these cases, a facility with language plays a central role in enabling LLMs to build and benefit from extended mechanisms. Much the same has been said about the role of language in human intelligence. As with LLMs, a facility with language opens the door to potent forms of cognitive and computational extension that radically reshape the space of human thought and reason, enabling us to tackle problems that would easily defeat the linguistically- and technologically-unaugmented brain. In the end, the story of AGI may turn out to be a continuation of our own cognitive story: a story of how language paved the way for a new, barely natural, form of intelligence, one whose cognitive limits have yet to be fully determined.