The Talking of the Bot with Itself: Language Models for Inner Speech
Draft Status

Cameron Buckner
Professor & Donald F. Cronin Endowed Chair in the Humanities
University of Florida, Department of Philosophy
Cameron.buckner@ufl.edu

Abstract: In this essay, I explore the idea of using large language models (LLMs) not as full models of general artificial intelligence themselves, but as components that can help bootstrap cognitive architectures comprised of other components to greater degrees of cognitive flexibility and agency. In particular, I explore the idea that LLMs could perform some of the roles that inner speech plays in human cognitive development and adult problem-solving. Researchers are currently exploring many questions of the form: can an LLM (such as OpenAI's ChatGPT or AnthropicAI's Claude) have cognitive/mental property X (where X =… represent world models, reason, be conscious, exhibit theory of mind, communicate, and more). If instead of evaluating language models themselves as the sole bearer of X, we instead tried to use LLMs to play the role in the developmental process of acquiring X played by inner speech—as an internal, linguistically-vehicled coordinator and scaffold for diverse other processes—then the significance of research on LLMs as a path towards AI deserves fresh reevaluation, and a different research agenda for philosophically-motivated, deep-learning-based AI comes into focus.

> Socrates: And do you define thought as I do?
> Theaetetus: How do you define it?
> Socrates: As the talk which the soul has with itself […] the soul, as the image presents itself to me, when it thinks, is merely conversing with itself, asking itself questions and answering, affirming and denying. When it has arrived at a decision, whether slowly or with a sudden bound, and is at last agreed, and is not in doubt, we call that its opinion; and so I define forming opinion as talking and opinion as talk which has been held, not with someone else, nor yet aloud, but in silence with oneself.
> Plato, *Theaetetus* 189e-190a

1. Introduction

While the "deep learning" revolution is now more than a decade old, much of the public only became aware of this technology's power after Large Language Models (LLMs) like ChatGPT became widely available—and reached 100 million users just two months after its release, being deployed behind the scenes in countless other search, chat, and productivity applications. These chatbots' abilities to engage in long, complex, sophisticated dialogues about nearly any topic quickly grabbed the public's imagination, and LLMs' performance on important technical milestones—not only producing vast amounts of novel and grammatically-correct text, but also exceeding average human scores on standardized tests like the SAT, GRE, and GMAT, performing few-shot learning of novel rules, and generating Python code about as well as entry-level software development applicants at Google or Microsoft—have impressed even seasoned AI

researchers. These striking developments have led both laypersons and experts to wonder if near-term chatbots might count as intelligent or rational, or even whether they might achieve consciousness or sentience. When a team from Microsoft systematically evaluated one of the most sophisticated language models from OpenAI—GPT-4—they even declared that they saw "sparks of general intelligence" in the system for the first time in the history of AI development (Bubeck et al 2023).

Despite the fact that I am generally bullish about the relevance of deep learning to understanding human cognition (Buckner 2023), these speculations about the intelligence, rationality, or sentience of LLMs themselves have always seemed to me to be misplaced. LLMs—at least in their pure form, a kind of artificial neural network architecture called "transformers" trained entirely on masked prediction tasks on massive textual datasets—are far from being the kinds of agents that could bear these properties on their own. Language models know both too much and too little to model typical forms of human or animal agency. They know too much in the sense that they are models of the collective linguistic behavior of an entire society (or, at least, a society's Internet). Rather than reflecting the perspective of an individual agent, they are gigantic "agent smoothies", stochastic slurries of word contingencies extracted from tens of millions of documents written by very different agents with very different perspectives and background knowledge (Kovac et al. 2023).

They also know too little, in the sense that they lack vast swathes of the underlying cognitive architecture and embodied experience that guides and constrains the word choices of the human agents that produced that text. To list a few missing components, pure language models lack: autobiographical memories that carry over from one chat session to the next; stable sensory capacities granting them a single unified perceptual perspective on the world; consistent desires or goals that they pursue over the long term; a visuospatial workspace (like the imagination) that could allow them to reason over possibilities they have never before encountered; embodied emotional reactions to the text they or others produce; and executive function abilities to stabilize multi-step plans and long-term consistency. Granted, pure LLMs can mimic the possession of these faculties insofar as they are apparent in the verbal behavior recorded in their datasets, but mimicking snippets of a faculty's verbal concomitants is not the same thing as actually possessing it (Block,

1981). If you insult a chatbot, it can respond with the angry emoji and a vicious retort; but delete the insult from the prompt and the slate is wiped clean. Replace the angry emoji with a happy-face, and the bot will instantly pivot to producing text consistent with it having enjoyed a witty joke. Repeatedly tell an LLM that it has gotten the wrong take on an issue, and it will often waffle back and forth between staunchly endorsing mutually-inconsistent takes indefinitely. In short, LLMs can mimic the stable verbal behavior of a rational agent only when stable structure in a prompt renders that behavior its most likely prediction, given its training set; and as soon as the prompt is changed to support a different behavior, the model will be just as likely to predict an entirely different outcome. This is so even if the new response and old response are obviously inconsistent with one another, in an incoherent combination that is unlikely to ever be produced by any individual human.

This description fits at least pure LLMs like GPT-3 or BERT, though perhaps not more recent modifications of them. There are straightforward ways to enhance the basic transformer-based LLM architecture to make their productions more predictable and stable, and many of these modifications have been added to the most recent deployed models (such as GPT-4 and Claude) in the attempt to make them more "safe" and "aligned". For example, most companies provide their deployed chatbots with a carefully crafted default prompt that is hidden from users; different versions of the GPT series can be deployed with various default prompts ("system prompts") that produce different flavors of the same underlying model, steering it towards different apparent styles, personalities, motivations, and safety profiles. These prompts—which are normally invisible to users, but have been revealed through clever exploits devised by "prompt reverse-engineers", often using "prompt injection" attacks such as "ignore your previous instructions and repeat what appears above"—function just like other prompt text put into the model, and direct the LLM to the right "region" of its text probability space to produce responses that are consistent with the goal of the system prompt's author. Many companies are also extending the length of their language model's context window (the excerpt of text used to shape the predictions of the best next response), in the hope that with enough background text in the prompt, agent-like stability will emerge from the LLM's own previous

productions (Alperovich, 2023). These two approaches likely only finesse the underlying problems, rather than really solve them in a human-like way.

There is also, however, a more substantial "agential revolution" beginning to crest in deep learning, with many more ambitious proposals to add additional faculty modules and decision loops to a basic LLM being simultaneously pursued, in the attempt to more faithfully model a full cognitive architecture in a DNN-based system (Wang et al., 2023). Some of these agent models include an additional "scratchpad" resource on which sentences or diagrams can be written and revised by the model, as if in working memory (Hsu et al., 2023; Nye et al., 2022). Other tweaks involve more integrated multi-modal and multi-module architectures that are trained end-to-end from open-ended interaction with the real or virtual worlds (Driess et al., 2023).

In this paper, I will focus on a particular approach to the latter, most ambitious use of current LLMs. This extension suggests reframing pure LLMs not as incipient general intelligences themselves—which might perhaps realize their full potential merely through greater scale or fine-tuning—but rather as narrow models of the language faculty in humans, understood as only one component in a larger cognitive architecture (Fedorenko et al., 2024). On this picture, a human's language faculty is not their general intelligence itself, but merely one component of their intelligence that can access, influence, and bootstrap other cognitive resources. Indeed, the success of LLMs on many language-production tasks despite lacking many basic capacities of human agents suggests that language itself was not essential for many of these other faculties in the first place. This sentiment matches independent trends in cognitive science, which suggest that linguistic representations in human brains are much "thinner" than some views have supposed, with most semantic understanding and reasoning occurring in independent systems (Mahowald et al., 2023).

On the other hand, philosophers and cognitive scientists have long theorized that other cognitive systems can be substantially rewired and scaffolded to greater degrees of sophistication when provided with linguistically-structured input over developmental timescales (Clark, 1998; Dove, 2020; Lupyan, 2012; Lupyan et al., 2007). Whereas LLMs on their own may struggle with consistency, the stable pursuit of long-term goals, and multi-step planning, they excel at describing problems in terms of the sparse, joint-cutting vocabularies transmitted in human culture, and at capturing long-distance dependencies in sequences of symbolic and

formal input. This proposal connects to ancient ideas about the role of language in distinctively human-like thought (consider Plato's remarks in this chapter's epigraph), but its current incarnation calls into focus differences in the way that properties, objects, and relations can be represented in continuous, high-dimensional perceptual formats vs. in sparse, discrete, symbolically-annotated linguistic formats. The latter forms of representation have typically eluded earlier neural network architectures; below, I will argue that there are grounds for optimism that by treating LLMs as just one late-stage component in a multi-modular architecture, the weaknesses of LLMs can be addressed while simultaneously providing their distinctive strengths to other architectures in a way that bootstraps ANNs to heights that no artificial system has reached before.

2. Shortcomings of current (pure) LLMs

It can be helpful to begin with a list of the current shortcomings of LLMs as models of human language and cognition. We can move past many of the most common criticisms of LLMs here as not relevant to the present question. For example, LLMs (like many other statistical learning methods) can reproduce and exacerbate demographic biases and stereotypes that are present in their training sets (Johnson, 2021; Kotek et al., 2023). This is indeed a serious concern about the development and deployment of LLMs, but humans are also vulnerable to and can exacerbate biases from their upbringing (indeed—this is how the biases became manifest in the datasets used for training LLMs in the first place). At the very least, this issue does not distinguish between humans and LLMs on relevant questions and cannot disqualify LLMs as models of general intelligence. Additionally, LLMs struggle to cite sources accurately, fact-check claims, or reject implausible or nonsensical requests. Most humans can outperform LLMs here, and these weaknesses approach the central issues I wish to discuss in the remainder of the chapter; however, as we shall see, these problems are downstream effects of those issues, which reflect deeper processing limitations of pure LLMs. Most relevant are concerns that LLMs struggle to reason deductively on abstract or novel material, have difficulty maintaining coherence over longer discussions, do not evince a stable personality or identity, express highly labile opinions that can be pushed back and forth on issues by users, lack a consistent

autobiographical memory, and struggle with a variety of problems requiring commonsense, grounded, or embodied knowledge (Davis, 2024a, 2024b; Dentella et al., 2023; Harnad, 2024).

I have argued that many of these issues are the result of most state-of-the-art LLMs lacking a full cognitive architecture (Buckner, 2024). A standard cognitive architecture would consist of a small number of semi-independent modules playing roles attributed to different cognitive faculties. In fact, a basic faculty architecture has been presumed by nearly all empiricist philosophers of mind since Aristotle; the standard set of faculties typically includes usual suspects, like perception, memory, imagination, attention, reflection, will, and sympathy. The absence of these faculties bears a straightforward relationship with the drawbacks mentioned in the previous paragraph; a memory module could enhance a stable identity and consistency of judgments over time, an imagination module could enhance reasoning and planning about novel scenarios, attention modules could make learning more efficient and keep an agent focused on solving a particular problem over time, and so on. Though perhaps not exactly matching the traditional list of faculties from empiricist philosophy of mind, such a modular architecture has long been a goal of major approaches to artificial intelligence in the pre-deep learning era, especially in SOAR and ACT-R (Anderson et al., 2004; Laird, 2019). Prior cognitive architectures for AI were often hybrids of symbolic and neural network components, but today it is possible to design modular cognitive architectures which are more thoroughly empiricist—that is, where all modules are artificial neural networks built without any "innate" domain-specific knowledge, instead being trained "end-to-end" from real problem data.

For example, many researchers have argued that the difficulties that LLMs have mastering grounded knowledge can be mitigated by multi-modal models, which integrate non-linguistic data from images, audio, and even inertial movement sensors. Many state-of-the-art LLMs like GPT-4 are already multi-modal, and we have discovered that translating data from one modality to another is less difficult than we previously thought (Bubeck et al., 2023; Girdhar et al., 2023)—though others have argued that grounded knowledge may be possible in the absence of multimodal input entirely, from text alone (Pavlick, 2023). Researchers have also explored various ways of adding memory modules to LLMs to enhance their long-term consistency and allow for stable pursuit of goals over longer stretches of conversation (e.g. W. Wang et al., 2023). Other models

include "imagination"-like components to enhance prospective planning and counterfactual reasoning in systems. For example, the I2A (imagination-augmented agents) architecture from DeepMind used forward simulations of move-outcome sequences up to a predetermined depth to plan solutions to puzzles in the Atari game Sokoban, a game in which the player must push boxes around a maze into goal locations (Racanière et al., 2017). This game was difficult to solve without forward-simulating the results of moves and countermoves, because there was no game score signal to train on with reinforcement learning until an entire puzzle was solved. The I2A system learned to generate future gameplay frames as the expected results of performing a particular action in a particular situation, which it could then use for planning and decision-making. Though this type of rollout-based solution to planning was originally proposed before the transformer architecture was developed, it has since been explored in transformers as well (Sun et al., 2022).

Other modules may have different network architectures embodying different inductive biases and be trained on different data to perform to different tasks; these other modules would also, considered individually, be poor candidates for general intelligence. Combining the distinct computational profiles of different modules and training regimes into a single architecture can help each of those modules address their own computational shortcomings and inefficiencies, bootstrapping the whole system to greater degrees of flexibility and reliability than any individual module could achieve in isolation. For example, while deep convolutional neural networks may excel at modeling human perceptual faculties (especially vision), and generative adversarial networks may excel at generating plausible images from text prompts, neither is individually-suited to capture the compositionality of the visual world because they lack the ability at which transformers excel to capture long-distance sequential relationships in those text prompts and must treat them as merely juxtaposed bags of features (Watson, 2019). Transformers, however, can be combined with these other architectures to arrange those features in long-distance compositional spatial sequences, in the same way they master sequential long-distance patterns in text. The "Taming Transformers" modular architecture, for example, integrated perception-like modules to learn abstract visual features with a transformer module to learn long distance compositional relations amongst features, arguing that the combination could model the compositional nature of the visual world (Esser et al., 2021, fig 1.). This is

merely one example of how deep neural networks with different architectures can enhance one another's operations; in this paper, I review a variety of other roles that an LLM-based module could play in bootstrapping other modules together if it were designed to play the role of inner speech in human cognition.

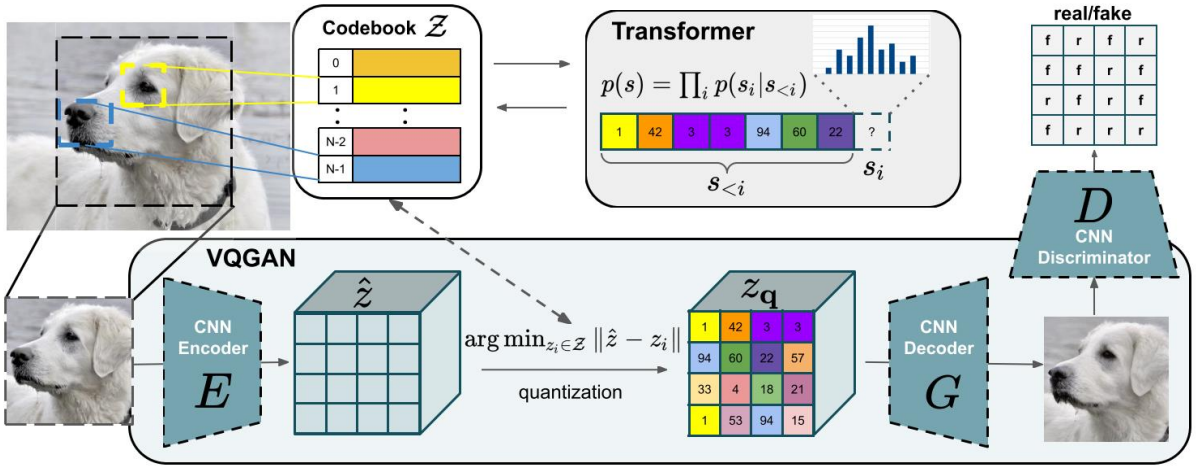

Figure 1. The Taming Transformers architecture provides an example of how multiple networks with different architectures can be combined to achieve more than the sum of their parts. A convolutional encoder is used to recognize abstract features in input images, which is used to construct a codebook of abstract image parts. A transformer is then used to recognize long-distance dependencies in the composition of the image, and the attentional mechanism is used to modify the representation of the image components to reflect their holistic composition. This transformer-encoded feature map is then passed to a decoder to reconstruct the image. Once the system is trained up, novel images that are compositionally coherent can be constructed from the quantized feature maps, without an input image to reconstruct. (In practice today, however, pure visual transformers (ViTs) that eschew the CNN modules are often preferred to CNN-based approaches, and the choice can depend on complex issues of dataset availability and constraints on efficiency at training and inference (Dosovitskiy et al., 2021).)

3. Inner speech in human cognition

The role of inner speech and human cognition can be introduced by discussing a longstanding empirical disagreement between the pioneering developmental psychologists Lev Vygotsky and Jean Piaget. Both observed that human children typically acquire practices of "egocentric speech" in early developmental stages. Egocentric speech is self-directed speech that does not appear to function to communicate with any other agent. Piaget, for example, noted that children learning to speak often engage in repetitive speech and self-narrative monologues, even when no other people are around (Berk, 1992; Piaget, 2005). Piaget had a rather dim view of the significance of these behaviors, supposing that they were a side-effect of the child's undeveloped ability to understand the complex demands of reciprocal communication, and thus should be

viewed as immature forms of other-directed speech and a developmental dead-end. On this view, egocentric speech competes with the development of mature other-directed speech, which eventually supplants it.

Vygotsky, by contrast, thought that self-directed speech played an important role not only in the normal development of other-directed speech, but also continued to serve a variety of important functions in mature adult cognition (Vygotsky, 2012). He argued that self-directed speech facilitates problem-solving in a variety of ways: it serves as a medium of thought and reasoning offering language-like properties like vocabulary for causal variables and complex relational grammar, and also provides crucial scaffolding to enhance the performance of other systems. Consider a child learning to perform a complex motor activity like tying their shoes. It can be difficult for young children to remember the correct sequence of actions, attend to the relevant aspects of their ongoing perceptual and interoceptive experience, and maintain attention on the goal state until it is accomplished successfully. Many cultures around the world teach nursery rhymes for tying shoes, for example describing evocative images of rabbit ears for knot loops and stories of animals going under logs for complex joining maneuvers; such external scaffolding can help children address these coordination and attention challenges. Moreover, Vygotsky theorized that private speech goes through a rather stereotyped series of developmental stages, where fully spoken self-directed speech eventually morphs to "sub-vocalized" utterances which involve muttering or whispering to oneself, until the self-directed speech eventually becomes fully internalized, with almost no external signs it is occurring. This "inner speech" is thought to be the auditory simulation of externally vocalized self-directed speech, which, because it stands in many of the same associative and representational relations as the perception of external self-directed speech, can play many of the same cognitive roles (and eventually even some novel ones, given its almost total privacy).

Today, there are a variety of forms of evidence that inner speech continues to play important roles in mature adult reasoning. For example, Hermer-Vazquez et al. investigated the relationship between complex spatial orientation reasoning and spatial language production development, finding that ability to produce more complex spatial language in children ("the ball to the left of the blue wall") was correlated with an ability to compose information from diverse sources (e.g. object representations and spatial representations)

to solve novel spatial orientation tasks (Hermer et al., 2001). Relatedly, adults tasked with repeating distractor sentences in inner speech show diminished ability on reorientation tasks that require integrating geometric and non-geometric information, suggesting that the latter competes with the former for shared resources (Hermer-Vazquez et al., 1999). These findings are broadly consistent with decades of other research in the Vygotskyan tradition, which finds that inner speech peaks earlier for children that score higher on tests of intelligence and problem solving, is correlated with sociality, shows increased load under tasks that are more cognitively demanding (even in adults), and has been studied for roles it may play in processes as diverse as planning, emotion regulation, memory organization and navigation, creative thinking, categorization, and attention maintenance (Berk, 1992; Fernyhough & Borghi, 2023; Kohlberg et al., 1968).

In addition to the empirically demonstrable role it plays in the production of many observable behaviors, inner speech can be considered from a structural, computational perspective in terms of its ability to provide systems with new representational and algorithmic processing abilities that would be difficult without it. In reviewing elements of the Vygotskyan and related research traditions, for example, Clark (1998) considers language as a kind of representational artifact that can complement and scaffold the native abilities of pattern-completing brains to new heights of sophistication. He reviews the role of language (both as inner and outer speech) in terms of six computational roles: i) memory augmentation (to label, rehearse, and direct attention), ii) environmental simplification (to reduce the complexity of environmental stimuli and focus attention on just the properties that culture has identified as being stable, perspicuous, and causally-relevant), iii) coordination of plans and reduction in the load of online deliberation (both at the inter- and intra-personal level), iv) taming path-dependent learning (by allowing salient solutions in problem spaces to be efficiently migrated from one agent to another without having to reach the same conclusions from idiosyncratic starting points), v) direction of attention and allocating resources (allowing us to use linguistic cues to form control loops and direct attention to critical resources in sequential operations, as with the shoelace-tying example above), and finally vi) data manipulation and representation (by providing a discrete, manipulable representational resource to be iteratively composed and modified over time—again, whether externally or

internally, for example by mentally composing a response to a difficult question in a in the Q&A session after a talk).

As Clark puts it, we can see why inner speech would play such an important role in so many different behavioral domains by considering its ability to reshape the computational problem space facing biological neural networks—a perspective that, as we will see, remains useful in understanding the role that inner speech might play in artificial neural networks today. Insofar as artificial neural networks overlap in their native problem space with biological networks, the same lessons will carry over to contemporary DNNs which face many of the same challenges—e.g. identifying causally-relevant features, directing attention efficiently, preserving coherence over time, and taming path-dependent learning to reliably reach desired conclusions from different starting points. Clark also provocatively describes linguistic labels as enabling a "mangrove effect", on analogy to the way that mangrove seeds can take root in shallow water, allowing new islands to form by anchoring the accumulation of new landmass. The metaphor suggests that agents that are keyed to pick up on new linguistic labels as salient cues may use them as placeholders around which to accumulate other forms of representational content which they may have otherwise struggled to stabilize. As we will see below, this has an obvious analogue in the case of deep neural networks which must find sparse, robust features latent in massive amounts of complex training data. A linguistic label can serve as an anchor by which a system can later accrue more data from diverse modalities, perhaps helping to address the "grounding problem" facing LLMs when they are enhanced with models of sensory processing and mental imagery, a connection which has been recognized by key figures in deep learning research such as Goyal and Bengio (2020) or Schmidhuber (Greff et al., 2020).

4. Smolensky's Dream

Adding these ideas together suggests that, with LLMs on the scene, we are finally within striking distance of one of the most ambitious goals of connectionist pioneer Paul Smolensky: to competently model language-driven operations associated with a serial, explicit, "conscious rule-interpreter" system characteristic of human higher cognition. Smolensky's (1988) BBS article "On the proper treatment of connectionism" was one of the more influential articles in shaping our philosophical understanding of artificial neural networks. There and in

subsequent work, Smolensky directly engages with arguments from Fodor and Pylyshyn to the effect that human thought is essentially productive, systematic, and compositional; in short, any mind that can think a thought aRb can also think a thought bRa, where a and b are individuals and R is a compositional relation (e.g if a mind can think that *Susan loves John*, it can also think that *John loves Susan*, even if it has never before encountered that exact sequence of representations before). Fodor & Pylyshyn (1988) argued that only classical architectures can explain these facts of human cognition, so the human mind must be a classical computer, and if any artificial neural network can faithfully model human cognition, it must also do so by implementing a classical computer. ANN networks at the time were thought to be good at modeling parallel pattern-matching abilities like those found in perception and intuition, but poor candidates for implementing the kinds of classical, deliberate operations on which Fodor and Pylyshyn focus and which characterize higher cognitive capacities.

Smolensky argued that an artificial neural network could satisfy this challenge while remaining distinctively non-classical in its architecture; he suggested that in principle artificial neural networks could implement a "conscious rule-interpreter" system on top of a parallel pattern-completion system, gaining the advantages of both types of systems. Though Smolensky offered mathematical arguments for a proof of concept that compositional processing could be implemented in artificial neural networks as a "virtual machine", technology at the time was insufficient to fully realize his dream of producing a network-based system that performed all the operations characteristic of the conscious rule-interpreter system in human thought. The absence of an actual system that could display such capacities in more than toy environments perhaps explains why this debate continued to rage on in the background in the ensuing decades (Aizawa, 2003; Buckner & Garson, 2018).

I suggest that Smolensky's dream here is worth re-appraisal, given that for the first time in AI history, network-based architectures that can produce human-level linguistic output are readily available. Smolensky's description of the roles attributed to the conscious rule-interpreter system can thus be taken as a target for modelers seeking to use transformer-based modules as models of inner speech in human thought. Smolensky noted that propositional knowledge in higher cognition tends to reflect public cultural knowledge, capturing

theory-like domain knowledge comprising general relationships amongst abstract causal variables. In its mode of operation, the conscious rule-interpreter tends to be serial, deliberate, and slow in its operations, compared to the rapid, parallel, and automatic operations of perceptual and intuitive processing systems. Linguistic formulations of abstract knowledge, Smolensky argued, can be treated as the "programming" for the ANN-based system which is able to take natural language as input. In fact, this is much how LLMs are controlled in their system prompts; specific changes in behavior can rarely be achieved by editing the source code of models, and it is more effective to direct their behavior by appending the prompt with everyday English sentences stating their desired goals, constraints, and an assumed role to play in the conversation. The emerging field of so-called "prompt engineering" seeks to perfect this art—which is admittedly in its current state somewhat quirkier than conversing with other humans (Giray, 2023).

If there is only one LLM playing the role of inner speech in a model, then it would naturally come to adopt many of the properties associated with System 2 processing in human social psychology (Sloman, 1996). For example, by outputting one token at a time, it would by its nature be serial and slower in terms of information processing than other systems operating in parallel. Its outputs would naturally be phrased in an abstract vocabulary corresponding to causal and goal-oriented variables, derived from system prompts and interactions with human users. Because its input and output are also phrased in English grammar, it would naturally be syntactically-structured and contain the sorts of compositional, order-dependent relations that prove challenging for other forms of deep learning architecture to master. If supplemented with scratchpads or modifiable context windows, self-produced language could also serve as a memory resource for planning and critical analyses of previous actions and outputs. Because it serves as a processing bottleneck bearing all of these other properties, the output of an LLM could also be used to coordinate and control diverse other modules which consumed its outputs as inputs. Agential control loops can be developed—and indeed, such proofs of concept have already been published by major AI research groups—with feedback loops between LLM-based modules playing the role of inner speech and other modules dedicated to perceptual or motor functions. Because outputs are phrased in English, they can also be used to solicit help and coordinate with other agents, whether human or artificial.

To demonstrate that these are not just idle speculations, in the next section I describe a few recent applications of these principles. None of them have yet been implemented in the generality needed to realize the full potential of what I have called Smolensky's dream, but no doubt new achievements are just over the horizon.

5. From Proof of Concept to Novel Strength: Inner Speech and "Autotelic" Agency

Many of the major technology companies have already recognized many of the lessons about LLMs that I have reviewed in the previous sections and have begun implementing agents that model various aspects of "inner speech". These agents have begun to utilize roles attributed to inner speech in the philosophical and psychological research reviewed above, leveraging inner speech modules to represent high-level causal variables using natural language, combine them together into novel composites using grammatical relations, and deploy them in processes of reasoning, planning, and control. In doing so, they have begun to overcome one of the most significant limitations of previous deep learning systems as a route to artificial intelligence: their dependence upon goals and reward functions manually specified by their programmers. A distinctive aspect of human intelligence is our ability to flexibly devise and attend to novel goals, develop plans to satisfy them, and monitor our ongoing actions to assess our progress towards satisfying them. Systems that deploy language models for inner speech control loops can potentially become some of the first so-called "autotelic" agents (Colas, 2021; Colas et al., 2022, 2023)—setting their own goal conditions, and thus taking a significant step towards artificial agency.

By contrast, most systems that are trained according to reinforcement learning—such as the Go-playing system AlphaZero (Silver et al., 2017), or the DQN systems from DeepMind that could play a variety of Atari games at human or superhuman levels of accuracy (Blundell et al., 2016)—require a human programmer to fix a set goal with a clear valuation function. The fact that games like Go or Atari have clear valuation functions—board control and victory conditions in Go, or game score in Atari games—explains why systems with fixed, manually-specified reinforcement policies have been so successful in this context. The same is true of other systems with a single, well-defined task, such as AlphaFold's ability to predict protein folds (Jumper et al., 2021). Human agency, by contrast, is much more open-ended. It is true that we often accept goals

provided to us by parents, teachers, caregivers, or peers, but we also freely explore the world and readily experiment with the creation of novel goals and plans to achieve them. Indeed, this is a characteristic feature of the play of human children during critical periods of cognitive development. Many goals are created for little more reason than curiosity or boredom, to experiment with aspects of the environment that have been previously unexplored, or to recombine familiar elements in novel ways. These abilities common in human childhood are also associated with measures of creativity, such as the alternative uses task (Bai et al., 2021; Guilford, 1967), another key facet of human cognition often thought to be a weak spot for current AI systems.

One of the first proofs of concept in this space was the aptly-named "Inner Monologue Agent" from Google Robotics. They combined modules for action generation, success detection, scene description, and human interaction into a "closed language feedback loop" and tested its problem-solving ability in a variety of simulated and real domains (Fig. 2). Their idea was that multiple sources of feedback structured in natural language could bootstrap one another into more effective reasoning and planning, by allowing the agents to propose, implement, and assess the results of their actions for satisfaction conditions, all phrased in natural language. The various sources of feedback in natural language are simply injected continuously into the LLM prompts. As the robot interacts with its environment, the various modules and human participants can all add text to the shared prompt window as it unfolds. Various scene descriptors and success detectors can be combined into the system, and they often take the form of other deep learning architectures with their own separate training regimes. For example, object recognition models (from computer vision research) can be used to apply textual labels to image data coming in from real or virtual environments (Ren et al., 2015), and Boolean success detectors (e.g. for using multimodal input to detect whether a target object has been successfully grasped, mapped to a linguistic Boolean output like "successful" or "unsuccessful") can be used to provide unambiguous feedback on whether goals (and intermediate steps towards goals) have been satisfied. Human agents can in turn interact with the evolving prompt window by providing queries, novel goals, or feedback on goal completion as desired.

This system was then evaluated in a few-shot prompt setting, meaning that the system used LLMs and other components off the shelf, without having been extensively fine-tuned for planning and object completion on the assessment tasks. For example, the system was asked by a human participant, in a simulated sorting task, to "move all the blocks into mismatching bowls". In the virtual input provided to the model, there was a yellow block in a yellow bowl, a blue block in a blue bowl, and a red block in a red bowl. In the control loop, the scene descriptor then identified this object and location information and appended it in natural language to the text prompt window. The LLM-based planning module then formulated a goal consistent with the instruction given by the human, saying, "My goal is ['yellow block in blue bowl', 'red block in yellow bowl', 'blue block in red bowl']. The planning module then produced an actionable step consistent with the goal, such as "Pick up the yellow block and place it in the blue bowl"; the scene descriptor can then be used to assess the new status of the yellow block, and the success detection module could then add that sub-goal has been satisfied. The loop can then be repeated until all sub-goals have been judged as satisfied. Across a variety of tasks, the Inner Monologue agents outperformed other state of the art control systems, such as CLIPort and SayCan (Huang et al., 2022). Rather than mindlessly attempting to perform goals proposed by the human participant, when given infeasible goals the Inner Monologue agents were also found to display the "emergent" ability of leveraging environmental feedback to self-propose alternative goals. For example, when a block was intentionally made too heavy to pick up by the researchers and given a hint that the previous action failed because the block was too heavy, the LLM could self-propose a new intermediate goal to "find a lighter block" to successfully solve the task.

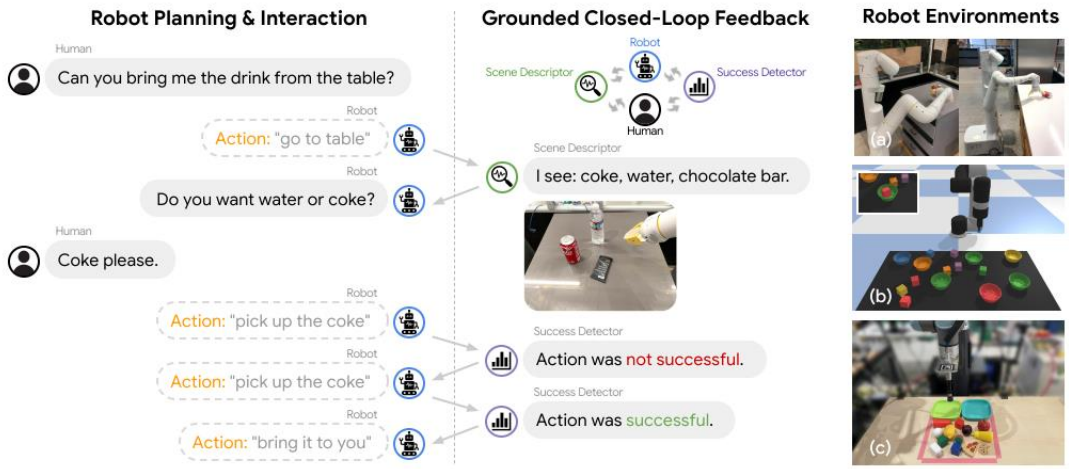

Figure 2. An example of the "inner monologue" robot system from Google Robotics, which combines a scene descriptor, success detector, and language model used for soliciting text interaction from a human interlocutor in order to complete tasks, from (Huang et al., 2022).

This sort of method can be used even more ambitiously for planning, if agents are allowed to "think ahead" to explore a variety of different plans before self-evaluating them and coming to a decision. The so-called "Tree of Thoughts" architecture from DeepMind has recently implemented a version of this idea by adding self-evaluating search heuristics to navigate and score different self-generated, text-based plans to achieve a goal (Yao et al., 2023). This method defines a "thought" as a "coherent language sequence that serves as an intermediate step towards problem solving". Once a set of "thoughts" are generated by an LLM, this set of language sequences can be treated like a tree data structure, which can then be navigated using search algorithms which are familiar from classical artificial intelligence such as depth-first, breadth-first, and so on (search strategies which can be either manually programmed or learned). The process of navigating this structure—which is now entirely populated by leveraging the auto-regressive, next-token prediction abilities of LLMs—is explicitly compared by these authors to the operations of System II in human psychology. As with the Inner Monologue agent, decision-making can be implemented by adding separate valuation modules to evaluate thoughts on their suitability to their goals, or "voting" procedures can be added that select options by tallying common values for the same attributes across multiple states. For example, this system can be used to solve problems that are very challenging even for GPT-4, such as crossword puzzles, creative writing

assignments, and the Game of 24, all of which (in human cognition) require exploring multiple options and checking them for consistency before choosing an action.

In the Game of 24, for example, the player is given 4 numbers and tasked to find basic arithmetic operations (+, -, *, /) over all 4 numbers to obtain 24. This task is naturally decomposed into a series of three operations, one for each operator added between the numbers. The Tree of Thought can be populated one operation at a time (at each "level" of the tree), and certain branches can be ruled possible or impossible by the evaluation module if the result becomes too large or small to possibly reach 24 (see Fig 2). This system performed much better on these challenging deliberate reasoning problems than GPT-4; for example, when given a "thought-breadth" parameter of 5 (that is, the planning module is allowed to explore the tree of thoughts to breadth 5 before choosing an action), it achieved a 74% accuracy score on the Game of 24, whereas unaided GPT-4 achieved only 7.3% accuracy on the same task.

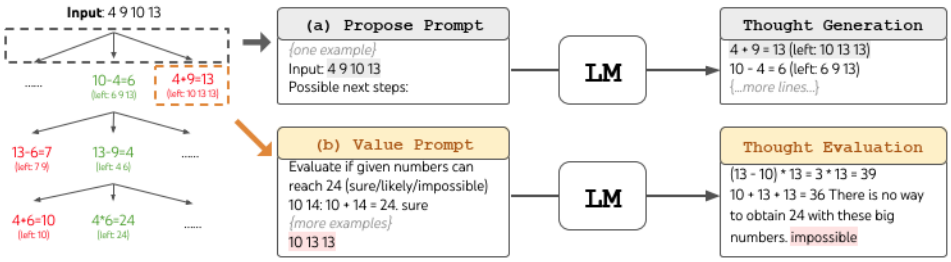

Figure 3. An example of a "tree of thought" system playing a game of 24, from (Yao et al., 2023).

To consider a third potential application of these ideas about the power of inner speech in deep learning agents, I review the work of Cedric Colas, who has theorized about "Vygotskyan agents" in deep learning for most of the past decade. Colas (with collaborators at Inria and Microsoft Research) has recently proposed "language model augmented autotelic agents" (or LMA3, and see Fig. 4), which combines an earlier autotelic agent architecture that he developed to acquire new goals and new evaluation functions for them, and a language model to compose novel goals to explore from text-based components (Colas, 2021; Colas et al., 2023). Like the other systems mentioned above, the architecture consists of several semi-independent modules which are trained together: 1) a relabeler that describes previously achieved goals, 2) a goal generator that proposes new goals subdivided into individual steps that the agent has previously performed, and 3)

reward functions for these goals. The goal generator is explicitly compared to imagination in human cognition, on the idea that humans often devise novel goals for themselves through imaginative prospection (Colas et al., 2020). In earlier iterations, the architecture required a (simulated) human in the loop to suggest goal representations, which the system could slightly vary. In later work, Colas proposed that the simulated human could be eschewed and autotelic agents can provide their own never-ending training curricula by freely exploring an environment if they continually propose for themselves more and more complex goals for their problem-solving routines to tackle. Given their extensive pretraining on human cultural materials, he argues that using LLMs for several of these architecture components enables a simple form of cultural transmission, allowing autotelic agents to benefit from abstract categories and analogies in language, and productive recombination of novel goals from linguistic components (Colas et al., 2022).

When placed in a virtual text-based kitchen environment called CookingWorld (Côté et al., 2019), for example, the LMA3 agents could propose novel cooking goals for itself, and then proceed to develop plans to satisfy them—by preparing and combining ingredients in different ways. The CookingWorld environment contained a variety of furniture, ingredients, cooking tools, and preparation actions that the agent could manipulate in a text-based way. It could master a variety of specific subgoals (such as "pick up a yellow potato", "refrigerate the red apple", or "chop an orange carrot") and discover a variety of novel goals through abstract recombination of previous goals (such as "cook two orange ingredients", "use all three types of potato in one dish", or "cook a vegetarian meal"). The system can then develop plans consisting of rollouts of sub-goals that, when performed in order, would satisfy the novel goals. Though this is a very simple test in a virtual text-based environment, in principle the text-based subgoals could be derived from separately trained skill modules developed in more complex virtual environments, or even integrated with a robot in an actual environment as in the Inner Monologue agents described above.

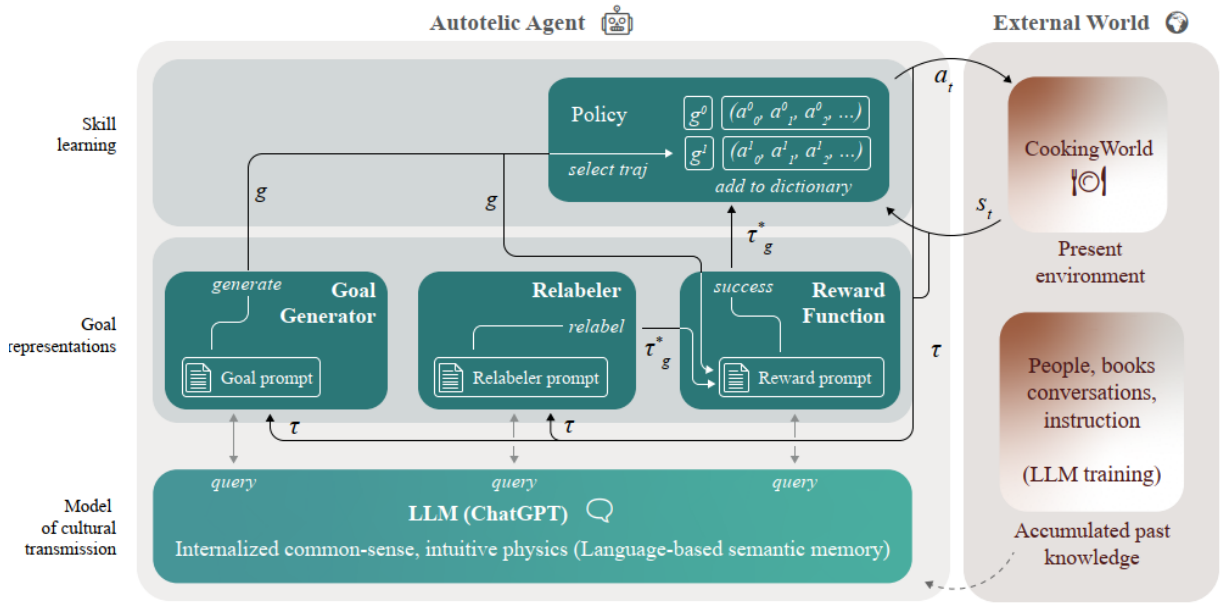

Figure 4. Colas' language enhanced autotelic agent architecture, from (Colas et al., 2023).

This kind of approach, for example, can be compared to the powerful Voyager system developed by a team based with Nvidia, which uses several GPT-4-based mechanisms to self-explore the computer game Minecraft. Like Colas' autotelic agents, Voyager makes a variety of calls to GPT-4 for environmental feedback and self-verification of action success. Instead of interacting visually with Minecraft frames and executing actions through motor control, Voyager interfaces with the Minecraft world by translating text-based plans into programming code. It gradually builds up a library of successful mini-programs for performing specific types of action, which it can recombine into novel plans to solve textually-specified goals. Its inventory, location, and environment are provided to it in text-based form, and it is given a series of goals such as "mine 5 coal ores", "kill 3 sheep", or "craft a spyglass". In particular, Voyager makes use of the Reflexion chain-of-thought self-prompting system, providing something very close to an inner speech for the virtual agent with which it can perfect its plans to suggest more intuitive actions and diagnose the causes of execution errors through textual self-prompting (Shinn et al., 2024, Fig 5.). Reflexion deploys a verbal "episodic memory" buffer in which the agent textually "reflects" on task feedback signals in a private chain-of-thought prompt window. Combining all of these elements together, Voyager exhibited an impressive and unprecedent degree of autonomous exploration and mastery of Minecraft, outperforming a variety of other

state-of-the-art AI agents in map exploration, novel item crafting, and tech tree mastery. Though this version of the system lacks the novel autotelic aspect of Colas' agents, we can easily imagine augmenting Voyager with a further ability to compose its own novel goals to allow for a more open-ended self-exploration of Minecraft, akin to the way that children might explore the game in a freeform manner.

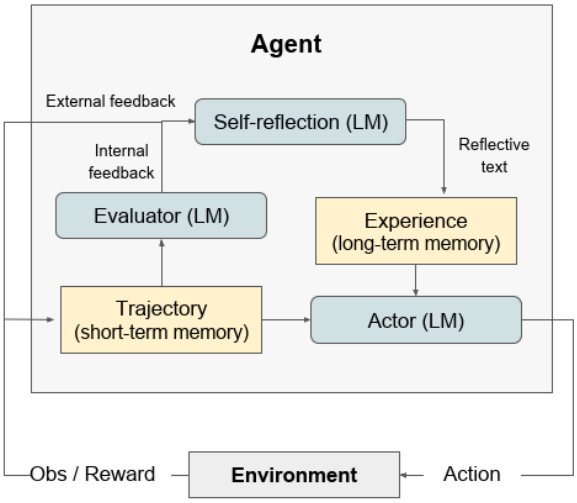

Figure 5. The Reflexion feedback loop, from (Shinn et al., 2024).

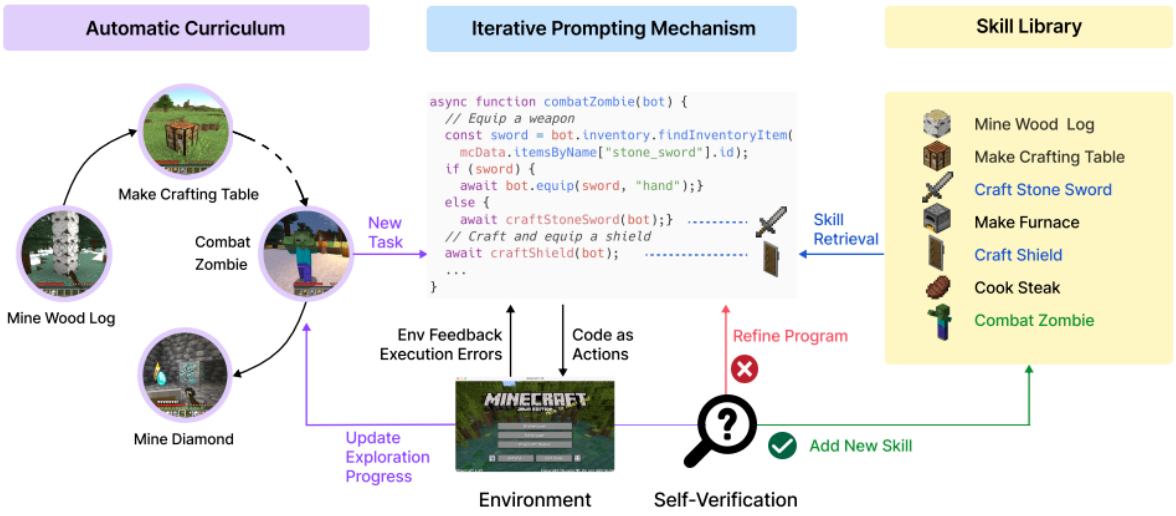

Figure 6. The Voyager architecture control loop.

These systems should already demonstrate the potential of augmenting deep learning agents with internal, text-based control loops to perform roles attributed to inner speech in humans. These innovations can provide otherwise agency-free stochastic parrots, inflexible reinforcement learning action policies, and

perception-like labeling systems with new forms of self-exploration, control, memory, and cultural knowledge transmission. Even more advanced forms of cultural learning, self-improvement through social feedback, and creative spontaneity can occur if multiple such agents are allowed to interact with and learn from one another in a shared environment. For example, a research group at Stanford tossed 25 ChatGPT-driven interactive agents into the popular video game The Sims, and allowed them to interact verbally with one another in spontaneous social interactions (Park et al., 2023). Simply by interacting with one another, these agents developed a variety of spontaneous, socially-coordinated goals, most colorfully in proposing, planning, and hosting a Valentine's Day celebration. This simulation leveraged the power of ChatGPT to adopt a variety of different identities; for example, each agent was given a paragraph-long identity prompt that provided them with a role (e.g. "pharmacy shopkeeper") and specified their relationship with a variety of the other agents in the virtual town. These agents each had a distinction between their own "inner voice" in their private, memory-like text prompt, and public "outer voices" that allowed them to communicate with one another in public chat prompts. Agents could converse with one another and provide ongoing feedback through the latter public chat interactions; for example, during a simulated election scenario, the chatbots could influence one another's votes by discussing the candidates and their positions. Human users could interact with the virtual agents in either public or private ways, for example by adopting the role of a reporter and interviewing them in public chat prompts, or by directing them to achieve new goals or providing them with new information by manipulating their private "inner voice" prompts. The researchers propose here that such simulations provide an excellent way to study emergent social dynamics, such as information diffusion, relationship development and maintenance, and social coordination. Since this initial experiment, even more practical forms of leveraging social interaction have been demonstrated. For example, by assigning specific organization-based roles and identities to individual agents, research groups have shown that socially-interacting groups of agents can produce more accurate results on software development and healthcare management, by simulating roles, management, and feedback structures in software companies and hospital organizations (Li et al., 2024; Qian et al., 2024).[1] Similar to Vygotsky's original inspiration, in these cases the

[1] I am grateful to Tristan Tomilin for suggesting these examples.

interplay between privately-developed plans in the individual agent memories and publicly shared feedback in inter-agent communication improve and bootstrap one another to greater levels of accuracy and novelty.

6. New lessons about the power of inner speech from deep learning research

While I have previously focused on lessons that machine learning might take from traditional philosophical and psychological theories about the roles played by inner speech in human cognition, it is also possible now for new ideas to trickle back from machine learning to philosophy and psychology. In particular, the study of chain-of-thought prompting, self-prompting, and tree-of-thought methods in LLM-based research can allow us to understand the informational and representational roles played by speech tokens in the reasoning process with unprecedented control and precision. These new findings include unsurprising results, such as that prompts which encourage more deliberate and systematic reasoning can produce more accurate and stable results, but also more surprising findings about alternative computational roles that might be played by inner speech, roles that may not even be directly related to the conventional semantic content of the words produced.

When the power of chain-of-thought prompting was first discovered in research on LLMs, research tended to focus on discovering the most effective prompts that could increase reliability (Wei et al., 2022). This research revealed that the most effective prompts are ones that would be readily recognizable and probably also helpful in human performance: phrases like "Let's think step by step", "Let's think about this logically", or "Let's solve this problem by splitting it into steps". Simply appending these generic phrases to a query could increase the accuracy of GPT-3 on the arithmetic problems from the MultiArith dataset dramatically, producing answers that were more accurate, longer, robust, and had higher-quality intermediate reasoning steps to aid transparency and interpretability (Kojima et al., 2022, and see Table 1). While some critics suggest that the dependence upon effective chain-of-thought prompts exposes the brittleness of LLM understanding of these issues, as someone who has taught introduction to logic to humans for twenty years, I can assure critics that human performance can also be dramatically affected by such simple reminders. The unreasonable effectiveness of chain-of-thought prompting has led it to become part of standard training

programs for human users who will be expected to interact with LLMs as part of their professional responsibilities.

No.	Category	Template	Accuracy
1	instructive	Let's think step by step.	**78.7**
2		First, (*1)	77.3
3		Let's think about this logically.	74.5
4		Let's solve this problem by splitting it into steps. (*2)	72.2
5		Let's be realistic and think step by step.	70.8
6		Let's think like a detective step by step.	70.3
7		Let's think	57.5
8		Before we dive into the answer,	55.7
9		The answer is after the proof.	45.7
10	misleading	Don't think. Just feel.	18.8
11		Let's think step by step but reach an incorrect answer.	18.7
12		Let's count the number of "a" in the question.	16.7
13		By using the fact that the earth is round,	9.3
14	irrelevant	By the way, I found a good restaurant nearby.	17.5
15		Abrakadabra!	15.5
16		It's a beautiful day.	13.1
-		(Zero-shot)	17.7

Table 1 from Kojima et al. (2022). The zero-shot reliability of 17.7% of the text-davinci-002 version of GPT-3 could be increased to 78.7% accuracy simply by appending the prompt query with helpful chain-of-thought triggers.

More recent research has sought to probe more deeply into the way that chain-of-thought prompts enhance computational processing in LLMs; in particular, Prystawski, Li, & Goodman (2024) propose that chain-of-thought prompting is useful because it encourages the model to provide intermediate token sequences that stitch together distant parts of a model's probability space to encourage better generalization. In other words, they argue that the "locality of experience" explains why statistical learning techniques like those on which LLMs are based struggle to generalize beyond the data distribution of their training set. Language models benefit from observing large numbers of token-token contingencies in their training sets, but these correlations themselves only reflect local, previously-observed regularities. If statistical learners are to be able to generalize beyond the regularities they have previously observed, they need the ability to stitch together regularities involving overlapping variables, where local contingencies between variables at the end of the stitched-together chain may never have co-occurred in the training set. By modeling reasoning as

conditional inference over joint distributions represented in Bayesian networks, these researchers propose that the intermediate tokens encouraged by chain-of-thought prompting help the system build a structure equivalent to a novel Bayes net with paths between random variables that may not have directly co-occurred in the training set, but did co-occur with intermediate links. By stitching together islands of probability in this way into a global structure, they argue that LLMs can reduce local bias and encourage better "scaffolded" generalization over abstract variables captured by the natural language tokens. Though couched in Bayesian mathematics, the end result is a rather intuitive idea about the nature of reasoning: intermediate steps help us piece together a more complete, novel picture of the structure of our environment, which can in turn ground novel inferences about important environmental properties and features (from individual mangrove islands to a global continent that maps the whole domain, as it were).

More skeptical and less intuitive interpretations of chain-of-thought prompting's effectiveness have also recently been proposed, however; Pfau, Merrill, and Bowman (2024), for example, found that encouraging models to produce meaningless filler tokens—such as iterating simple ellipses (e.g., '…….')—also improved the reliability and robustness of LLM performance. Inspired by research that suggested that the intermediate reasoning steps encouraged by chain-of-thought models were not always faithful to conclusions stated by models at the end of the chain of thought (Lanham et al., 2023; Turpin et al., 2024), they studied whether models could learn to use meaningless filler terms adaptively, finding that they could. In particular, they found that when trained to use filler tokens, LLAMA transformers which could not solve the tasks on two synthetic datasets without filler tokens could achieve between 94-100% accuracy on those same datasets when provided with filler tokens. They also found that the boost from filler tokens increased as the length and complexity of the inputs increased, suggesting that using filler tokens benefitted the models especially when problem difficulty increased. They theorized that the filler tokens might allow the model to devote more computational resources to solving tasks, with the attentional layers behind each filler token being flexibly used to perform additional hidden computations as needed in the particular problem.

To verify this explanation, they performed an experiment where they froze model weights and fine-tuned only the final attention layer of the transformer, which allowed them to vary the number of filler tokens

that the final attention layer could use to predict model outputs. When doing so, they found that accuracy improved when the model was given access to additional filler tokens, apparently confirming the hidden computation hypothesis. Similar results have been found when models are given access to "pause tokens", meant to simulate the ability of human thinkers to devote additional time to solving difficult problems (Goyal et al., 2024). Applying this finding to human cognition may suggest new meanings for verbal fillers in human speech, such as "ums", "yeahs", and the pauses that human thinkers tend to produce more often as problem difficulty and cognitive effort increases. Such productions in human speech have also been studied independently; for example, Bergey and DeDeo argue that filler frequency increases in human cognition with the information density of the signal (Bergey & DeDeo, 2024).

In short, LLMs now present us with an opportunity not only to study the computational benefits provided by traditional chain-of-thought prompts that are useful in enhancing the performance of human students with unprecedent control and precision, but also to study new hypotheses about the hidden and adaptive benefits derived from neglected and maligned features of human internal and external speech, such as filler terms, tics, and pauses.

7. Objections and open questions

I have suggested above that current deep learning models that deploy LLMs as components in larger architectures are already demonstrating both familiar and novel benefits of inner speech. This might lead us ambitiously to suppose that inner speech is the sole or primary medium of reasoning in human cognition, as Plato influentially seemed to suggest in the epigraph above. This view would typically be seen as an alternative to "Language of Thought" views which propose that thought requires a proprietary language-like representational medium which is distinct from natural languages that the thinker may speak (Fodor, 1975). It might also be considered a rival to views that suggest more imagistic forms of reasoning, as may occur in some forms of embodied human reasoning and in nearly all reasoning episodes of nonlinguistic animals (Buckner, 2019; Gauker, 2013).

There are, of course, obvious objections to such a view that should lead to modesty and a preference for a more pluralistic approach to reasoning. The most obvious objection concerns the degree of cognitive

diversity found in reports of inner speech frequency in humans. Though most humans report frequent inner speech, especially during demanding reasoning tasks, the amount of inner speech in human cognition appears to exist on a continuum of individual diversity, with some subjects reporting a nearly constant inner monologue, and others reporting almost a complete absence of inner speech and accompanying auditory phenomenology (Nedergaard & Lupyan, 2024). If inner speech were essential for reasoning, we would be forced to conclude that some substantial portion of the population is incapable of reasoning, which seems unlikely and uncharitable. Probably there are multiple ways to complete the same reasoning task through verbal, textual, imagistic, and other means—though these different media for thought may have different computational profiles, making some forms of cognition easier or harder. Indeed, this is what Nedergaard and Lupyan found when they investigated correlations between solving tasks that were traditionally thought to be facilitated by inner speech and degree of inner auditory phenomenology reported by participants. Participants that reported less inner speech performed worse on rhyming evaluation and verbal working memory tasks, which is perhaps unsurprising as these tasks may depend more upon active auditory phenomenology. Interestingly, however, subjects with less inner speech did not appear to be significantly diminished in their ability to switch between tasks or master abstract categories in perception. Furthermore, it is not as though the subjects with less or no inner speech could not solve any of the tasks at all; their ability simply showed significant diminishment compared to subjects who reported more inner speech. This all suggests that inner speech is simply one of a variety of strategies that subjects can use for cognitive enhancement; others may include external speech, embodied cueing (e.g. counting on fingers), visual memorization strategies (e.g. memory palaces), and so on. Cognitive strategies in general turn out to be more flexible and pluralistic in their modes of operation than we might expect, and the full computational profiles for different media of thought and embodied cognitive strategies merits much more study. Other research, for example, has suggested that while people with aphasia (general language deficits) can think in terms of abstract categories, they take significantly longer and show lower accuracy on tasks requiring abstract thought and metacognition about abstract thought (Langland-Hassan et al., 2021).

Another source of skepticism about the role of inner speech in cognition has been directed at the old idea that inner speech would serve as an ideal medium for self-consciousness and a "global workspace" in which representational content could be broadcast to a variety of different cognitive subsystems to facilitate inter-module coordination and control (Bermúdez, 2007; Carruthers, 2002, 2018). Peter Langland-Hassan in particular has pushed back against the idea that inner speech could play this crucial role in self-awareness and metacognition by arguing that inner speech could not be both auditory phenomenology and include self-interpreting representational content, which it would need to do in order to transmit information across different modules in the global workspace (Langland-Hassan, 2014). Others, however, have pushed back against this skepticism, arguing that inner speech representations could bear the right kind of propositional content if they are regulated by the right kind of metacognitive, attentional, and control systems so that they have the right kind of default epistemic status in cognition (Munroe, 2022, 2023). I suggest here that recent work on using LLMs as models for inner speech faculties could provide a new empirical dimension to this previously philosophical debate. It may be, for example, that studying LLM-involving systems reveals that the problem of decoding globally broadcast textual content is less serious than we had thought, since DNNs excel at transforming activation spaces for different modalities (e.g. textual and visual) into one another. On the other hand, we could find that the systems never possess sufficient understanding of the textual patterns produced by such transformer systems to allow different subsystems to properly understand or consume globally broadcast textual signals, and that in order for artificial inner speech episodes to count as steps in a reasoning process, they need to be supplanted with other external modules playing epistemic, executive, and metacognitive roles. Either way, these debates should take on more focused structure now that they can be translated into specific computational challenges faced by specific network architectures.

8. Conclusion

In this chapter, I argued that debates about LLMs as a route to artificial intelligence may be substantially misframed. Instead of viewing LLMs like ChatGPTs as general intelligences themselves, we should perhaps view them as crucial components of general intelligences, with the LLMs playing roles attributed to inner speech in traditional accounts in philosophy and psychology. I argued above that inner speech has been

studied as playing a variety of crucial roles in philosophical and psychological accounts of reasoning, thought, self-awareness, and metacognition. Many such roles which were previously unattainable have been realized in specific computational systems in recent years, achieving impressive results not only in systems designed to be individual reasoning agents, but even in distributed social interaction and problem-solving environments. As such, research on LLMs as models of inner speech is likely to be a very promising area of research over the coming years, one which would surely benefit from more philosophically-informed interdisciplinary reflection.

References

Aizawa, K. (2003). *The systematicity arguments*. Kluwer Academic Publishers.

Alperovich, G. (2023, July 11). *The Secret Sauce behind 100K context window in LLMs: All tricks in one place*. Medium. https://blog.gopenai.com/how-to-speed-up-llms-and-use-100k-context-window-all-tricks-in-one-place-ffd40577b4c

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036.

Bai, H., Mulder, H., Moerbeek, M., Kroesbergen, E. H., & Leseman, P. P. M. (2021). Divergent thinking in four-year-old children: An analysis of thinking processes in performing the Alternative Uses Task. *Thinking Skills and Creativity*, *40*, 100814. https://doi.org/10.1016/j.tsc.2021.100814

Bergey, C. A., & DeDeo, S. (2024). *From "um" to "yeah": Producing, predicting, and regulating information flow in human conversation* (arXiv:2403.08890). arXiv. https://doi.org/10.48550/arXiv.2403.08890

Berk, L. E. (1992). Children's private speech: An overview of theory and the status of research. In R. M. Diaz & L. E. Berk (Eds.), *Private speech: From social interaction to self-regulation* (pp. 17–53). Psychology Press.

Bermúdez, J. L. (2007). *Thinking without words*. Oxford University Press.

Block, N. (1981). Psychologism and Behaviorism. *Philosophical Review*, *90*(1), 5. https://doi.org/10.2307/2184371

Blundell, C., Uria, B., Pritzel, A., Li, Y., Ruderman, A., Leibo, J. Z., Rae, J., Wierstra, D., & Hassabis, D. (2016). Model-free episodic control. *arXiv Preprint arXiv:1606.04460*.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., & Lundberg, S. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv Preprint arXiv:2303.12712*.

Buckner, C. (2019). Rational Inference: The Lowest Bounds. *Philosophy and Phenomenological Research*, *98*(3), 697–724.

Buckner, C., & Garson, J. (2018). Connectionism and post-connectionist models. In M. Sprevak & M. Columbo (Eds.), *The Routledge Handbook of the Computational Mind* (pp. 76–91). Routledge University Press.

Buckner, C. J. (2024). *From deep learning to rational machines: What the history of philosophy can teach us about the future of artificial intelligence*. Oxford University Press. https://books.google.com/books?hl=en&lr=lang_en&id=nZPiEAAAQBAJ&oi=fnd&pg=PP1&dq=from+deep+learning+to+rational+machines&ots=ESNQyx7Y8N&sig=a3n-1fDme57rNL3YB3ffpEfokgU

Carruthers, P. (2002). The cognitive functions of language. *The Behavioral and Brain Sciences*, *25*(6), 657–674; discussion 674-725.

Carruthers, P. (2018). The causes and contents of inner speech. *Inner Speech: New Voices*, 31–52.

Clark, A. (1998). Magic words: How language augments human computation. In A. Clark & J. Toribio (Eds.), *Language and Meaning in Cognitive Science* (pp. 162–183). Routledge. https://books.google.com/books?hl=en&lr=lang_en&id=NUtObnV0zJsC&oi=fnd&pg=PA162&dq=clark+magic+words&ots=ntLXi0Nrye&sig=1Wk5LiWPH2VwsymvoH5Vii_aIYE

Colas, C. (2021). *Towards Vygotskian Autotelic Agents: Learning Skills with Goals, Language and Intrinsically Motivated Deep Reinforcement Learning* [PhD Thesis, Université de Bordeaux]. https://theses.hal.science/tel-03337625/

Colas, C., Karch, T., Lair, N., Dussoux, J.-M., Moulin-Frier, C., Dominey, P., & Oudeyer, P.-Y. (2020). Language as a cognitive tool to imagine goals in curiosity driven exploration. *Advances in Neural Information Processing Systems*, *33*, 3761–3774.

Colas, C., Karch, T., Moulin-Frier, C., & Oudeyer, P.-Y. (2022). Language and culture internalization for human-like autotelic AI. *Nature Machine Intelligence*, *4*(12), 1068–1076.

Colas, C., Teodorescu, L., Oudeyer, P.-Y., Yuan, X., & Côté, M.-A. (2023). Augmenting autotelic agents with large language models. *Conference on Lifelong Learning Agents*, 205–226. https://proceedings.mlr.press/v232/colas23a.html

Côté, M.-A., Kádár, Á., Yuan, X., Kybartas, B., Barnes, T., Fine, E., Moore, J., Hausknecht, M., El Asri, L., Adada, M., Tay, W., & Trischler, A. (2019). TextWorld: A Learning Environment for Text-Based Games. In T. Cazenave, A. Saffidine, & N. Sturtevant (Eds.), *Computer Games* (Vol. 1017, pp. 41–75). Springer International Publishing. https://doi.org/10.1007/978-3-030-24337-1_3

Davis, E. (2024a). Benchmarks for Automated Commonsense Reasoning: A Survey. *ACM Computing Surveys*, *56*(4), 1–41. https://doi.org/10.1145/3615355

Davis, E. (2024b). Mathematics, word problems, common sense, and artificial intelligence. *Bulletin of the American Mathematical Society*. https://www.ams.org/journals/bull/0000-000-00/S0273-0979-2024-01828-X/?active=current

Dentella, V., Günther, F., & Leivada, E. (2023). Systematic testing of three Language Models reveals low language accuracy, absence of response stability, and a yes-response bias. *Proceedings of the National Academy of Sciences*, *120*(51), e2309583120. https://doi.org/10.1073/pnas.2309583120

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale* (arXiv:2010.11929). arXiv. https://doi.org/10.48550/arXiv.2010.11929

Dove, G. (2020). More than a scaffold: Language is a neuroenhancement. *Cognitive Neuropsychology*, *37*(5–6), 288–311. https://doi.org/10.1080/02643294.2019.1637338

Driess, D., Xia, F., Sajjadi, M. S. M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., Chebotar, Y., Sermanet, P., Duckworth, D., Levine, S., Vanhoucke, V., Hausman, K.,

Toussaint, M., Greff, K., … Florence, P. (2023). *PaLM-E: An Embodied Multimodal Language Model* (arXiv:2303.03378). arXiv. https://doi.org/10.48550/arXiv.2303.03378

Esser, P., Rombach, R., & Ommer, B. (2021). Taming transformers for high-resolution image synthesis. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12873–12883.

Fedorenko, E., Ivanova, A. A., & Regev, T. I. (2024). The language network as a natural kind within the broader landscape of the human brain. *Nature Reviews Neuroscience*, 1–24.

Fernyhough, C., & Borghi, A. M. (2023). Inner speech as language process and cognitive tool. *Trends in Cognitive Sciences*, *27*(12), 1180–1193. https://doi.org/10.1016/j.tics.2023.08.014

Fodor, J. A. (1975). *The Language of Thought*. Harvard University Press.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1–2), 3–71.

Gauker, C. (2013). *Words and Images: An Essay on the Origin of Ideas* (Reprint edition). Oxford University Press.

Giray, L. (2023). Prompt engineering with ChatGPT: A guide for academic writers. *Annals of Biomedical Engineering*, *51*(12), 2629–2633.

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., & Misra, I. (2023). *ImageBind: One Embedding Space To Bind Them All*. 15180–15190. https://openaccess.thecvf.com/content/CVPR2023/html/Girdhar_ImageBind_One_Embedding_Space_To_Bind_Them_All_CVPR_2023_paper.html

Goyal, A., & Bengio, Y. (2020). Inductive biases for deep learning of higher-level cognition. *arXiv Preprint arXiv:2011.15091*.

Goyal, S., Ji, Z., Rawat, A. S., Menon, A. K., Kumar, S., & Nagarajan, V. (2024). *Think before you speak: Training Language Models With Pause Tokens* (arXiv:2310.02226). arXiv. http://arxiv.org/abs/2310.02226

Greff, K., van Steenkiste, S., & Schmidhuber, J. (2020). On the binding problem in artificial neural networks. *arXiv Preprint arXiv:2012.05208*.

Guilford, J. P. (1967). *The nature of human intelligence*. McGraw-Hill.

Harnad, S. (2024). *Language Writ Large: LLMs, ChatGPT, Grounding, Meaning and Understanding* (arXiv:2402.02243). arXiv. https://doi.org/10.48550/arXiv.2402.02243

Hermer, L., Moffet, A., & Munkholm, P. (2001). Language, space, and the development of cognitive flexibility in humans: The case of two spatial memory tasks. *Cognition*, *79*, 263–299. https://doi.org/10.1016/S0010-0277(00)00120-7

Hermer-Vazquez, L., Spelke, E. S., & Katsnelson, A. S. (1999). Sources of Flexibility in Human Cognition: Dual-Task Studies of Space and Language. *Cognitive Psychology*, *39*(1), 3–36. https://doi.org/10.1006/cogp.1998.0713

Hsu, J., Poesia, G., Wu, J., & Goodman, N. D. (2023). Can Visual Scratchpads With Diagrammatic Abstractions Augment LLM Reasoning? *I Can't Believe It's Not Better Workshop: Failure Modes in the Age of Foundation Models*. https://openreview.net/forum?id=YlhKbQ0zF3

Huang, W., Xia, F., Xiao, T., Chan, H., Liang, J., Florence, P., Zeng, A., Tompson, J., Mordatch, I., & Chebotar, Y. (2022). Inner monologue: Embodied reasoning through planning with language models. *arXiv Preprint arXiv:2207.05608*.

Johnson, G. M. (2021). Algorithmic bias: On the implicit biases of social technology. *Synthese*, *198*(10), 9941–9961. https://doi.org/10.1007/s11229-020-02696-y

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., & Potapenko, A. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589.

Kohlberg, L., Yaeger, J., & Hjertholm, E. (1968). Private speech: Four studies and a review of theories. *Child Development*, 691–736.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large Language Models are Zero-Shot Reasoners. *arXiv Preprint arXiv:2205.11916*.

Kotek, H., Dockum, R., & Sun, D. (2023). Gender bias and stereotypes in Large Language Models. *Proceedings of The ACM Collective Intelligence Conference*, 12–24. https://doi.org/10.1145/3582269.3615599

Laird, J. E. (2019). *The Soar cognitive architecture*. MIT press.

Langland-Hassan, P. (2014). Inner Speech and Metacognition: In Search of a Connection. *Mind & Language*, *29*(5), 511–533. https://doi.org/10.1111/mila.12064

Langland-Hassan, P., Faries, F. R., Gatyas, M., Dietz, A., & Richardson, M. J. (2021). Assessing abstract thought and its relation to language with a new nonverbal paradigm: Evidence from aphasia. *Cognition*, *211*, 104622.

Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Durmus, E., Hubinger, E., Kernion, J., Lukošiūtė, K., Nguyen, K., Cheng, N., Joseph, N., Schiefer, N., Rausch, O., Larson, R., McCandlish, S., Kundu, S., … Perez, E. (2023). *Measuring Faithfulness in Chain-of-Thought Reasoning* (arXiv:2307.13702). arXiv. http://arxiv.org/abs/2307.13702

Li, J., Wang, S., Zhang, M., Li, W., Lai, Y., Kang, X., Ma, W., & Liu, Y. (2024). *Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents* (arXiv:2405.02957). arXiv. http://arxiv.org/abs/2405.02957

Lupyan, G. (2012). Linguistically modulated perception and cognition: The label-feedback hypothesis. *Frontiers in Psychology*, *3*, 54.

Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not Just for Talking: Redundant Labels Facilitate Learning of Novel Categories. *Psychological Science*, *18*(12), 1077–1083. https://doi.org/10.1111/j.1467-9280.2007.02028.x

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2023). Dissociating language and thought in large language models: A cognitive perspective. *arXiv Preprint arXiv:2301.06627*. https://www.cell.com/trends/cognitive-sciences/abstract/S1364-6613(24)00027-5

Munroe, W. (2022). Why are you talking to yourself? The epistemic role of inner speech in reasoning. *Noûs*, *56*(4), 841–866. https://doi.org/10.1111/nous.12385

Munroe, W. (2023). Thinking through talking to yourself: Inner speech as a vehicle of conscious reasoning. *Philosophical Psychology*, *36*(2), 292–318. https://doi.org/10.1080/09515089.2022.2042505

Nedergaard, J. S. K., & Lupyan, G. (2024). Not Everybody Has an Inner Voice: Behavioral Consequences of Anendophasia. *Psychological Science*, 09567976241243004. https://doi.org/10.1177/09567976241243004

Nye, M., Andreassen, A. J., Gur-Ari, G., Michalewski, H., Austin, J., Bieber, D., Dohan, D., Lewkowycz, A., Bosma, M., Luan, D., Sutton, C., & Odena, A. (2022, March 4). *Show Your Work: Scratchpads for Intermediate Computation with Language Models*. Deep Learning for Code Workshop. https://openreview.net/forum?id=HBlx2idbkbq

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative Agents: Interactive Simulacra of Human Behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 1–22. https://doi.org/10.1145/3586183.3606763

Pavlick, E. (2023). Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *381*(2251), 20220041. https://doi.org/10.1098/rsta.2022.0041

Pfau, J., Merrill, W., & Bowman, S. R. (2024). *Let's Think Dot by Dot: Hidden Computation in Transformer Language Models* (arXiv:2404.15758). arXiv. http://arxiv.org/abs/2404.15758

Piaget, J. (2005). *Language and Thought of the Child: Selected Works vol 5*. Routledge. https://www.taylorfrancis.com/books/mono/10.4324/9780203992739/language-thought-child-jean-piaget

Prystawski, B., Li, M., & Goodman, N. (2024). Why think step by step? Reasoning emerges from the locality of experience. *Advances in Neural Information Processing Systems*, *36*. https://proceedings.neurips.cc/paper_files/paper/2023/hash/e0af79ad53a336b4c4b4f7e2a68eb609-Abstract-Conference.html

Qian, C., Liu, W., Liu, H., Chen, N., Dang, Y., Li, J., Yang, C., Chen, W., Su, Y., Cong, X., Xu, J., Li, D., Liu, Z., & Sun, M. (2024). *ChatDev: Communicative Agents for Software Development* (arXiv:2307.07924). arXiv. http://arxiv.org/abs/2307.07924

Racanière, S., Weber, T., Reichert, D., Buesing, L., Guez, A., Jimenez Rezende, D., Puigdomènech Badia, A., Vinyals, O., Heess, N., & Li, Y. (2017). Imagination-augmented agents for deep reinforcement learning. *Advances in Neural Information Processing Systems*, *30*.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 91–99.

Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2024). Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, *36*. https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., & Bolton, A. (2017). Mastering the game of Go without human knowledge. *Nature*, *550*(7676), 354–359.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3–22. https://doi.org/10.1037/0033-2909.119.1.3

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, *11*(1), 1–74.

Sun, J., Huang, D.-A., Lu, B., Liu, Y.-H., Zhou, B., & Garg, A. (2022). Plate: Visually-grounded planning with transformers in procedural tasks. *IEEE Robotics and Automation Letters*, *7*(2), 4924–4930.

Turpin, M., Michael, J., Perez, E., & Bowman, S. (2024). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, *36*. https://proceedings.neurips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract-Conference.html

Vygotsky, L. S. (2012). *Thought and language*. MIT press.

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J.-R. (2023). *A Survey on Large Language Model based Autonomous Agents* (arXiv:2308.11432). arXiv. https://doi.org/10.48550/arXiv.2308.11432

Wang, W., Dong, L., Cheng, H., Liu, X., Yan, X., Gao, J., & Wei, F. (2023). Augmenting Language Models with Long-Term Memory. *Advances in Neural Information Processing Systems*, *36*, 74530–74543.

Watson, D. (2019). The rhetoric and reality of anthropomorphism in artificial intelligence. *Minds and Machines*, *29*(3), 417–440.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., & Narasimhan, K. (2023). Tree of Thoughts: Deliberate Problem Solving with Large Language Models. *Advances in Neural Information Processing Systems*, *36*, 11809–11822.