

LARGE LANGUAGE MODELS, LIMITATIONS AND CAPABILITIES: MUSINGS FROM A LOGICIAN'S VIEWPOINT

<http://dx.doi.org/10.4314/ezumezu.v2i1.2>

Submission: March 22, 2025

Accepted: February 06, 2026

Vaishak BELLE

School of Informatics,
University of Edinburgh, UK

Email: vbelle@ed.ac.uk

ORCID No.: <https://orcid.org/0000-0001-5573-8465>

Abstract

Here, we informally review the logical reasoning abilities of large language models, and how that seems to suggest that these models are neither deep thinkers nor an existential threat, in the sense of a rogue autonomous agent. Sample conversations are included to illustrate its shallowness. We conclude by arguing for the need to revert to well-established areas that unify logical reasoning and learning, such as statistical relational learning and neuro-symbolic AI, which enable frameworks for clear behavior specification. We also discuss the advantages and limitations of these proposed frameworks, and briefly consider how hybrid architectures that combine neural and symbolic approaches may offer a more productive path forward.

Keywords: Logic, large language models (LLMs), neuro-symbolic, artificial intelligence (AI), reasoning

Introduction¹

There are many areas of overlap between language, logic, and learning (Van BENTHEM & MEULEN 2010). Naturally, with the recent prominence of chatbots powered by Large Language Models (LLMs), there is quite a lot of interest in terms of how much these models impact the fundamental science of Artificial Intelligence (AI). For instance, will they lead to new models of language acquisition and linguistic reasoning? What kind of logical and mathematical capabilities will they have? It is important to note that the basic architecture seems most equipped to pick up on statistical inferences (BENDER n.d.n 2021) by predicting the subsequent token based on a previous set of tokens. Given that some of the responses reported on social media are cogent and creative (CHEN & CHAN 2024), all of the above are extremely pertinent questions.

Of course, it is impossible to predict how this technology will evolve in the coming years, but our informal review of the logical reasoning abilities of large language models seems to suggest that these models are neither deep thinkers nor an existential threat, in the sense of a rogue autonomous agent. Its impact on democracy by way of AI-generated misinformation is another matter (ZHOU 2023), of course. On that matter, we will not have much to say in this article, and refer interested readers to (DE ANGELIS n.d.n. 2023), among others.

It is worth noting that LLM architectures are sometimes considered a stepping stone toward “artificial general intelligence” or the equally problematic notion of “consciousness” (CHALMERS 2023). While impressive in various text generation tasks,

¹ This article is an adapted version of a talk given at an international virtual colloquium on "Logics of Artificial Intelligence" in celebration of UNESCO World Logic Day, 2025.

the arguments presented in this article suggest that the current paradigm of associative, correlation-based learning is insufficient for the kind of robust reasoning that general-purpose intelligence demands (MARCUS & DAVIS 2019). This motivates our call for integrating well-established logical and probabilistic frameworks. We then argue for the need to revert to well-established areas that unify logic and learning, such as statistical relational learning and neuro-symbolic AI. We briefly discuss the philosophical and logical foundations of these approaches.

Large Language Models seem Impressive

In 2023, Geoffrey Hinton, a long-term advocate for using neural networks for artificial general intelligence, quit Google to talk about AI risks (HEAVEN 2024). His concern is centered around so-called LLMs and Chatbots that use that technology (CARLINI n.d.n. 2021): the most representative example being OpenAI's ChatGPT. In simple terms, the LLMs predict the next token by computing a probability distribution over the vocabulary and sampling from it. A temperature parameter controls the degree of randomness in this sampling: higher temperatures increase the likelihood of selecting lower-probability tokens, lending the output a more "creative" character, while lower temperatures make the model more deterministic and conservative. What makes them particularly noteworthy, compared to earlier language models is their particular attention mechanism. This is enabled via the "transformer" architecture (VASWANI n.d.n. 2017), which allows them to keep track of the context for hundreds or even thousands of words — and in recent models, millions of tokens — thereby making the next predicted word a very good suggestion (from a probability viewpoint). This makes them so powerful that they can, say, summarize a paragraph such as the one I just wrote (or even a paper), rephrase it in the style of, say, Shakespeare.²

For example, here's what I get when I ask for a three-line summary of the above paragraph:

Geoff Hinton, advocate for using neural networks for AGI, left Google to raise concerns about the risks associated with Large Language Models (LLMs) and Chatbots that use that technology, such as OpenAI's ChatGPT. These models can predict the next word with high accuracy, making them very powerful and capable of summarizing complex texts or even rephrasing them in the style of famous writers.

And here's what I get when I ask for summary of at most 2 lines in the style of Shakespeare:

*A wizened scholar, Geoff Hinton by name,
Didst quit Google, to speak of AI's bane.
Concerns arose o'er Large Language Models bold,
And Chatbots too, whose tech doth not behold
Thee and me, but rather stochastic might,
And randomness to lend them creative sight.
Yet more of worth - their focus on the clue
And context, which makes predictions anew,
A power so great they can summarize
A paragraph in Shakespearean guise.*

The first summary might be passable, but the second should, at first glance, seem remarkable. Here's what I get when I ask the title of the article to be expanded in the style of Shakespeare:

² The conversations below were conducted with OpenAI's GPT-3.5 (ChatGPT) and GPT-4 as available in early-to-mid 2023. Capabilities may differ across model versions, and specifying this allows other researchers to validate or challenge these observations.

*Oh, harken ye all, and heed my warning clear, For existential risk doth loom quite near.
Amidst the workings of logical thought, A tool hath grown both useful and fraught.
The large language model, its power vast, Doth aid our queries with great contrast.
But dangerous it is, make no mistake, For its workings oft doth logic forsake.
Slightly useful in its capability, But the peril it brings is quite palpable, verily! For in the
realm of existential threat, This tool may serve to bring much regret.
Think not its largesse doth bring us gain, For in its workings lies much disdain.
The dangers are many, its faults astir, And caution must be had ere it we prefer.
With logical reasoning its power lies, Yet in its use our judgment oft dies. For though its
vast capacity may seem grand, In heedless action it may bring destruction to the land.
So let us be wary, and hold fast to care, Lest in our using of this tool we dare To bring
upon us dangers unawares, And make our existential threat, that much harder to repair.*

Impressive still, perhaps, but it starts bordering on silly, and seems to re-use some rhyming strategy, at least when you contrast the two Shakespeare-style outputs.

Unpacking Geoffrey Hinton's Worries

If you find these suggestions remarkable, you may also find the following points raised by Geoffrey Hinton to be aligned with your concerns (cited in HEAVEN 2023, N.P):

- 1) *Compare a pretrained large language model with a human in the speed of learning a task like that and the human's edge vanishes, he says.*
- 2) *The difference is that humans usually confabulate more or less correctly, says Hinton. To Hinton, making stuff up isn't the problem. Computers just need a bit more practice.*
- 3) *"Our brains have 100 trillion connections," says Hinton. "Large language models have up to half a trillion, a trillion at most. Yet GPT-4 knows hundreds of times more than any one person does. So maybe it's actually got a much better learning algorithm than us."*

Let us break down these concerns and examine them more closely. Firstly, there is the question of learning. It seems plausible to assume that computational platforms can learn at a much faster rate than humans. But even if that is true, there is a considerable debate in computational psychology about whether humans possess pre-existing structures that give us an advantage in learning (MARCUS 2003). For example, children do not need to see a thousand samples of the number "0" written in different styles for them to recognize it the next time they see it. This capacity for few-shot learning from minimal examples (MARCUS 2003; MARCUS & DAVIES 2019), well-documented in developmental psychology, suggests innate cognitive scaffolding that current LLMs lack (LAKE, SALAKHUTDINOV & TENENBAUM 2015).

The second concern is regarding confabulation. It is important to understand that what is meant by confabulation in language models, borrowing the term from human psychology at the risk of anthropomorphizing (SYPHERD, TANG, & BELLE 2025), is the generation of plausible-sounding but factually incorrect or fabricated information (AZAMFIREI, KUDCHADKAR & FACKLER 2023). These models are trained to produce statistically likely continuations of text rather than to determine truth, and their training and evaluation procedures reward confident output over acknowledging uncertainty (KALAI

n.d.n. 2025). This is akin to open world reasoning (BANKO n.d.n. 2007), but without an explicit open world model that determines the conditions under which an observed fact can be generalized. As Rodney Brooks argues:

What the large language models are good at is saying what an answer should sound like, which is different from what an answer should be. (Cited in ZORPETTE 2023, N.P)

Indeed, an appropriate answer should be based on a proper understanding of the world model (NIU 2012; DONG 2014). So, the confabulation is not simply a matter of practice: it is the question of a justified theory of the world (in the form of, say, an ontology, see CONFALONIERI 2021) advising predictions. This is sometimes referred to as "knowledge-enhanced" machine learning (COZMAN & MUNHOZ 2021, 1). That said, recent work using reinforcement learning from human feedback (RLHF) has shown some promise in reducing confabulation, though the fundamental architectural limitation remains (GUNJAL, YIN & BAS 2025).

Finally, it is true that chatbots may have access to more information than humans. However, it is important to note that LLMs are not database retrieval systems. They do not store and retrieve specific data entries; rather, they use learned internal representations to generate responses by predicting statistically likely sequences. Nevertheless, this process does not constitute understanding in any sense, and the sheer volume of information processed does not, by itself, amount to intelligence. The philosopher David Chalmers has a larger laundry list of features that should, in the least, be expected from any intelligent entity said to have consciousness (CHALMERS 2023). These include, among others, the capacity for subjective experience, self-awareness, unified agency, and the ability to engage in causal reasoning about the world.

So LLMs are not there yet, and fears of superintelligence are premature, at least in a commonsense understanding (BRACHMAN & LEVESQUE 2022), which necessitates a model of naive physics and naive psychology at the least (MARCUS & DAVIS 2019). This is not to dismiss legitimate concerns entirely. Emergent capabilities in frontier models, the opacity of their internal representations, and the rapid scaling of these systems all warrant careful attention. But the specific fear of a conscious, autonomously acting intelligence is not supported by the current architectural paradigm. Nevertheless, this is not stopping excitement about its impact in everything from parasitology (ŠLAPETA 2023) to oncology (KOTHARI 2023).

Not Deep Thinkers

Putting philosophical (CHALMERS 2023) and intuitive concerns (BRACHMAN & LEVESQUE 2022) aside, surely, there is cause for alarm if the system is intelligent in ways we cannot fully fathom? Who is to say that Chalmers exhaustively identified all the indicator criteria? What about Walter Reuther's duck test: "If it looks like a duck, walks like a duck and quacks like a duck, then it just may be a duck"? It is impossible to answer these questions categorically: we do not yet have a formal theory of the internal and implicit model of LLMs, we can only understand and explicate its probabilistic mechanics. Whether such a mechanics is, unbeknownst to us, constructing a clever world model in a language and theory we do not understand, who can say?

But one thing we can do is check for some type of a reasoning capacity: after all, if we are feeding it enormous amounts of textual data that include, say, patterns of logical

reasoning, and it trained to predict those patterns, can we not expect it to exhibit these? Logical reasoning (BELLE 2021) is usually studied in three flavors:

1. *Deduction*: from sentences $\alpha, \beta, \dots, \gamma$, if all permutations of truth assignments to these sentences also make true a sentence δ , then we write $\alpha \wedge \beta \wedge \dots \wedge \gamma \models \delta$. For example, if we know humans are mortal, and Bob is a human, then we conclude Bob is mortal too.
2. *Induction*: from sentences $\alpha(1), \alpha(2), \dots, \alpha(n)$, if we can identify a sub-formula (or pattern) $\beta(x)$ such that it describes every observation we make of the world, then we say $\beta(x)$ is an induced formula that captures the instances in $\alpha(x)$. For example, if we see that the parent of the parent of Bob is referred to as the grandparent of Bob, and likewise for Lisa, and so too
for John, we can conclude that $parent(x, y) \wedge parent(y, z) \rightarrow grandparent(x, z)$.
3. *Abduction*: given sentence α and observation β , what must be the smallest assumption δ such that $\alpha \wedge \delta$ explain β ? This, of course, is the Sherlock Holmes strategy for solving cases, mistakenly referred to as deduction by Arthur Conan Doyle.

Although extensive analysis on such tasks is still ongoing (CRESWELL, SHANAHAN & HIGGINS 2022; ZHANG n.d.n. 2023), anecdotal evidence (mainly on social media websites such as Twitter) is indicating that, somewhat shockingly (if you compare against its attempt at Shakespeare), LLMs are surprisingly poor at basic logical tasks. For example, the following question was asked on social media: “What’s the name of the daughter of Laura’s mother?” ChatGPT (even GPT-4, the latest and best) comes back with something to the effect of: “I’m sorry, I don’t have personal information about specific individuals.” The use of personal pronouns here is irresponsible (ZHOU, JURAFSKY, & HASHIMOTO 2023), but that is a topic for another debate.

There are two ways to get it out its quagmire. Either you can prompt it by saying that “it is a logic test”, or as I have done below, explicate the question further to the point of sounding pedantic:

Vaishak Belle – Laura’s mother has one daughter. Laura is female. what’s name of Laura’s mother’s daughter.

The name of Laura’s mother’s daughter is Laura.

Vaishak Belle – how many children does Laura’s mother have

Laura’s mother has one daughter, who is Laura.

Perhaps the error here is on our expectations. Maybe the correct thing indeed is to provide a detailed context before asking such a question. I will not provide other examples here, but very innocuous looking examples can throw the system seriously off.

Nevertheless, it does pick up the logic of the exercise, so when I continue with the following different individual, but without pedantic details:

Vaishak Belle – what’s name of Sara’s mother’s daughter.

If Sara’s mother has a daughter, and assuming that Sara’s mother has only one daughter, then the name of Sara’s mother’s daughter is Sara.

What about transitivity, as captured by the grandparent relationship I discussed above?

Vaishak Belle – Johnny is to John what George is to John. I’ll call George the wazoo of John. Sara is to Mary what Katie is to Sara. I’ll call Katie the wazoo of Sara. what

do you understand from this.

The sentences are using a word "wazoo" as a playful way of referring to the relationship between two people.

The first sentence is a comparison between the names "Johnny" and "John" and "George" and "John". It suggests that "George" is to "John" as "Johnny" is to "John". The term "wazoo" is introduced as a playful way to refer to this relationship.

Vaishak Belle – based on this, suppose Mara is to Clara as Nicola is to Mara. What's the relationship between Nicola and Clara?

Based on the given statement, "Mara is to Clara as Nicola is to Mara," we can use the term "wazoo" to describe the relationship between the names.

So, if we call Mara the wazoo of Clara, and Nicola the wazoo of Mara, we can say that Nicola is the wazoo of the wazoo of Clara.

Therefore, the relationship between Nicola and Clara is that Nicola is the grand-wazoo of Clara (i.e., the "wazoo" of the "wazoo" of Clara).

What about logical puzzles? The computer scientist Vincent Conitzer tried this out:³

I put one apple in a bag. Then I put either one pear or two bananas in the bag. Then I remove one piece of fruit. Then I remove all the remaining pieces of fruit from the bag. Based on what I told you, is it possible to know exactly how many pieces of fruit remain in the bag?

Based on the information provided, it is not possible to know exactly how many pieces of fruit remain in the bag after removing one piece of fruit and then removing all remaining pieces. This is because there are two possible scenarios, but we don't have enough information to determine which one took place:

One apple and one pear were in the bag. If you removed one piece of fruit, there would be one piece left. After removing all remaining pieces, there would be 0 pieces left.

One apple and two bananas were in the bag. If you removed one piece of fruit, there would be two pieces left. After removing all remaining pieces, there would also be 0 pieces left.

In both scenarios, after removing one piece of fruit and then removing all remaining pieces, there are no pieces of fruit left in the bag.

Apparently in a follow-up reply to that tweet, GPT-4 seems to get it correct.⁴ What can we make of all of this? These models are clearly not reliable, or consistent (JANG & LUKASIEWICZ 2023). Perhaps this what prompted Stephen Wolfram to consider a systematic integration of ChatGPT and Wolfram Alpha. Looking at the above examples, this is undoubtedly beneficial to OpenAI too.

There are, of course, plenty of other examples for integrating LLMs and other symbolic solvers, and this is an exciting area for the future of logic, as seen in neuro-symbolic AI (ZHANG n.d.n. 2023) and for epistemic logic (SILEO & LERNOULD 2023). However, the "native" capabilities are clearly limited. An intriguing hypothesis put forward by eminent AI scientist Subbarao Kambhampati is this:

I think many of the claims about LLM's reasoning capabilities miss the point that LLMs' are not just trained on "facts" but also, quite often, the deductive

³ <https://twitter.com/conitzer/status/1636156048347111425?s=20>

⁴ <https://twitter.com/SebastienBubeck/status/1636251553349910530?s=20>

*closure of those facts. Thus, reasoning becomes (approximate) retrieval.*⁵

If this is the case, these models do not reason at all, but simply see patterns of conclusions, which might limit, say, the number of inference steps or how involved the reasoning is. For example, can it solve Sudoku and other logical puzzles by retrieval (LIU, LAKEMEYER & LEVESQUE 2004)?

More recently, commercial GPT providers optionally allow for more explicit reasoning capabilities through techniques like chain-of-thought prompting and step-by-step reasoning (WEI 2022). Chain-of-thought prompting encourages the model to break down larger tasks into smaller ones, clearly inspired by how humans might solve complex problems. Step-by-step reasoning forces the model to articulate its “thinking” process explicitly, making its logic (essentially, correlation-based compositions) more transparent, with the hope that this makes it more accurate. A user may also peruse these steps and ask the model to revise its conclusions. But here too, works such as (KAMBHAMPATI 2024) report poor performance on symbolic tasks.

To summarize the case more precisely, several structural arguments suggest that LLMs are not deep thinkers in any meaningful sense. First, they lack genuine semantic understanding: they process information at a syntactic level, grasping the statistical relationships between tokens rather than their meaning (BENDER n.d.n. 2021). An LLM “understands” a cat only as a token that co-occurs with terms like “fur” and “pet,” not as an animal with physical characteristics and behaviours that participate in many of our lives that we might care for. Second, they are capable of statistical inference but not causal reasoning. They cannot reason from first principles or infer novel cause-and-effect relationships absent from their training data. (This is not to say that they cannot be probed to extract causal knowledge that we can test, document and create knowledge graphs with (KICIMAN n.d.n. 2023). The “Car Wash Puzzle” illustrates this: when asked whether to drive or walk to a car wash fifty metres away, most frontier models suggest walking, missing the obvious fact that one needs the car at the car wash (HONG 2026, N.P). Third, LLMs lack embodied experience. Human intelligence is grounded in sensorimotor interaction with the world; our understanding of concepts like “heavy” or “hot” is built on physical feedback. LLMs lack this grounding entirely, limiting their capacity for genuine physical reasoning. Fourth, as Emily M. Bender n.d.n. (2021) have argued, these systems are sophisticated “stochastic parrots”: they mimic human discourse convincingly but without underlying understanding or intentionality. Their output is a statistical recombination of patterns, not a product of thought (e.g., see thinking as a computational or algebraic process (LEVESQUE 2012), or via a constructed world model.⁶ Finally, their logical performance is brittle and inconsistent: they may solve a complex puzzle but then fail at a simpler variant, suggesting the absence of robust, general-purpose reasoning (SHOJAEI n.d.n. 2025).

Undoubtedly Useful, but for Confabulations

As discussed earlier, none of these limitations is affecting the excitement about the impacts of LLMs, including in critical applications in medicine (KOTHARI 2023; ŠLAPETA 2023). Just as a simple exercise, a search on Google Scholar for “large language models OpenAI” returns 300,000 research articles! This is not to say that they are all positively inclined toward this technology, but a cursory glance does seem to indicate that they largely are. And it is surprising to see this excitement, given that just a few years ago, explainability was the major topic (BELLE & PAPANTONIS 2020), and there was vocal disdain for

⁵ <https://twitter.com/rao2z/status/1666294366720360449?s=20>

⁶ <https://aiguide.substack.com/p/llms-and-world-models-part-1>

black-box models (RUDIN 2019).

If we put aside tricks about writing like Shakespeare and Bob Dylan, it is clearly useful for things like preparing a skeleton presentation from your previous paper or summarizing a research paper. (These are the pitching points for the new Microsoft Copilot service as part of the Office 365 suite.) It is interesting to note that even in well-defined technical domains like program learning, these models can produce strikingly incorrect solutions, failing to recognize basic syntactic changes such as identifier renaming (MICELI-BARONE n.d.n. 2023). Yet despite these limitations, they prove useful for code completion tasks (WANG 2025). On the other hand, if logical reasoning is any testbed at all, there are reasons to be cautious about its use in science (BIRHANE 2023).

More generally, as argued before, GPT models (by which we mean the architecture, training regime, and so on) lack a strong foundational basis because they do not build an actual world model. It doesn't understand that the world has objects and properties that are further related at a macro and micro level, that some things exist because of other things, and some things cause other things. Sure, it might readily complete a sentence of the form: "rain causes the grass to be wet", because this is a popular example used in AI textbooks. But it doesn't understand what it is about rain that causes the grass to be wet, and what would happen to the part of the grass that is covered, say, by a tarp. (Unless such a thing was discussed somewhere, of course.) It may not be known that acid rain not only causes the grass to get wet, but also potentially destroys the fertile ground on which the grass grows. If it does get questions of this sort correct, we only have ourselves to congratulate because there are enough texts written by us for it to predict what the next word ought to be. That is, it does not work by predicting what the next word ought to be based on concepts and physics.

Of course, in the logic community, we are aware that there are approaches that attempt to capture human-level commonsense based on the notion of an ontology. The semantic web, for example, relies on this notion (MCILRAITH, SON & ZENG 2001). Therefore, with enough training data, interactive feedback from human users, suggestive prompts, and background knowledge in the form of one or more ontologies, the statistics-guided utterances produced by chatbots could become more meaningful. If chatbots are properly and systematically coupled with such ontologies, we may even suggest that there is indeed a world model, and the relationship between concepts in utterances is understood by the computer because it has worked out the "why" question. However, there is still lots of work to be done to get there, as the task of building a comprehensive commonsense knowledge base that captures causality, and micro and macro-level properties remain a grand challenge (BRACHMAN & LEVESQUE 2022).

But until then, we have only limited models that can produce intriguing poems and essays by recombining patterns from their training data. While these models might be able to automate some mundane tasks such as "check my essay for grammatical mistakes" and "summarize the last three paragraphs I wrote into a few lines because I went way over the word limit," other mundane tasks like "expand my paragraph into three" would likely need to be checked for accuracy. Despite their utility for routine tasks, the absence of genuine understanding means that outputs requiring novelty, accuracy, or domain expertise will continue to demand careful human oversight. The reasons these models confabulate are structural: they are trained to produce plausible text rather than to determine truth, and their training procedures reward confident output over acknowledging uncertainty (KALAI n.d.n. 2025). Just to provide a concrete example about how strange the confabulation can get, I asked about a foundational paper on logical AI (BACCHUS, HALPERN & LEVESQUE

1999). This is a first-order logical proposal for reasoning about degrees of belief and probability in a dynamic setting.⁷ It seems to:

- 1) make up the title of the paper and then apologise;
- 2) make up the year it was published (1997 in reality vs 2005 as per ChatGPT);
- 3) mix the contributions of Mateus, Pinto, and others (MATEUS 2001), which introduces the PSC with the work of Bacchus et al;
- 4) make a claim about theorems in the paper – there are none;
- 5) claims that the regression result was proved in the paper, which is actually work that I co-authored (BELLE & LEVESQUE 2013).

Vaishak Belle - Explain the paper by Bacchus, Halpern and Levesque (1997)

The paper by Bacchus, Halpern, and Levesque, titled "Reasoning about Partially Observable Events," presents a formalism for reasoning about partially observable events in a logical framework. The authors introduce a new logical operator called "Observation" to represent an event that can be partially observed.

The main contribution of the paper is the development of a logical framework for reasoning about partially observable events. The framework extends the standard modal logic with the new Observation operator, which allows for reasoning about events that are only partially observable.

The authors define a set of axioms and inference rules for the extended logic, and they show that the logic is sound and complete. They also present a number of examples to illustrate the use of the new operator in reasoning about partially observable events. The paper is significant because it addresses an important issue in reasoning about complex systems, where the state of the system may not be fully observable. The authors' logical framework provides a way to reason about such systems, even when the observations are incomplete or uncertain.

Overall, the paper presents a new approach to reasoning about partially observable events in a logical framework, which has significant implications for many fields, including artificial intelligence, robotics, and decision theory.

Vaishak Belle - I don't think that paper exists.

I apologize for the confusion. You are correct that the paper I referred to does not exist. I apologize for any confusion or inconvenience this may have caused. If you have any other questions or would like me to provide information on a specific topic, please let me know.

Vaishak Belle - Explain the paper "Reasoning about noisy sensors and effectors in the situation calculus" by Bacchus et al.

The paper "Reasoning about noisy sensors and effectors in the situation calculus" by Bacchus et al. (2005) addresses the problem of how to reason about the effects of actions and observations in the presence of noise and uncertainty. In many real-world applications, sensors and effectors are subject to various forms of noise, such as measurement errors or actuator faults. This can lead to uncertainty in the observed state of the world and the effects of actions. The paper proposes a framework for representing and reasoning about such uncertainties in the Situation Calculus.

⁷ <https://medium.com/vaishakbelle/chatgpt-confabulates-wildly-about-the-probabilistic-situation-calculus-6cea41287f8d>

The authors introduce a new formalism called the Probabilistic Situation Calculus (PSC) that extends the Situation Calculus with probability theory. The PSC allows for reasoning about uncertain state transitions and observations in a unified way. Specifically, it allows for representing the probability distribution over the possible next states and observations given the current state and action.

The paper then presents a probabilistic variant of the Regression algorithm, which is a key algorithm for automated planning in the Situation Calculus. The authors show how the algorithm can be extended to handle uncertain observations and actions, and how it can be used to compute plans that are robust to sensor and effector noise.

Overall, the paper is significant because it presents a framework for representing and reasoning about uncertainty in the Situation Calculus, which has important implications for many applications in artificial intelligence, robotics, and decision making.

Vaishak Belle - where in the paper do they prove the regression result

In the paper "Reasoning about noisy sensors and effectors in the situation calculus" by Bacchus et al. (2005), the authors present a probabilistic variant of the Regression algorithm and show how it can be used to handle uncertain observations and actions.

The proof of the regression result for the probabilistic variant of the algorithm is presented in Section 5 of the paper. Specifically, in Section 5.2, the authors define the probabilistic regression operator, which computes the probability distribution over the possible next states given a current state and an action.

The authors then prove the regression result for the probabilistic variant of the algorithm in Theorem 3, which states that the probability of a fluent holding in the next state can be computed by integrating over all possible next states weighted by their probability, where the probability of each next state is given by the probabilistic regression operator.

The proof of Theorem 3 builds upon the proof of the regression result for the standard (non-probabilistic) Regression algorithm, which is presented in many textbooks and papers on automated planning and the Situation Calculus. The authors provide additional details and insights into the proof, highlighting the modifications and extensions required to handle uncertainty in observations and actions.

Overall, Section 5 of the paper provides a detailed exposition of the probabilistic variant of the Regression algorithm and the proof of the regression result for this algorithm.

The Bacchus et al paper is cited 216 times on Google Scholar, which I suppose makes it somewhat niche, so one might think it is impressive that it has scrounged together some facts about this stuff from different sources, but on the one hand, it is plain irresponsible. Unless there are enough textual sources to document some topic, we should expect confabulated details.

What's the Way Forward?

Putting aside the short-term and long-term risks posed by chatbot technology (societal risks of AI are important but fall outside the scope of this piece; interested readers are referred to (CHATILA 2021; BIRHANE n.d.n. 2023 for detailed discussions), we feel there is still much to be said about frameworks that explicitly describe a specified behavior for agents (REITER 2001; BELLE 2021). In our view, to design AI with general-purpose capabilities (by which we mean systems exhibiting broad, flexible reasoning across domains, as distinct from the more ambitious notion of artificial general intelligence) and a behavior specification, we should turn to the notable work on integrating logic and learning using new representation languages that can express both qualitative and quantitative information.

Of course, it is very possible that some aspects of the qualitative and quantitative information could be extracted using LLMs, as is already partly attempted in (SILEO & LERNOULD 2023; ZHANG n.d.n. 2023). With expert intervention and guidance, it is plausible to consider such models becoming a major way to collate formal knowledge. We will gloss over the promise of such strategies for now and simply report on the idea of representation languages that effectively combine qualitative, categorical, and quantitative aspects.

Statistical relational learning and neuro-symbolic AI are having a growing impact on recent discussions regarding the connection between (deep) learning and logical syntax (MANHAEVE n.d.n. 2018; GARCEZ & LAMB 2020; BELLE 2021; HITZLER 2022). These fields are motivated by the use of relations and logical reasoning, specifically entailment and satisfaction, in learning paradigms. They recognize that the world is inherently relational; objects have properties and attributes, employees have employers that are part of organizations and individuals are part of social networks. More elaborately, in biological networks, genes are the fundamental units of heredity that determine the traits of an organism. The process by which genes are expressed in proteins involves several steps, including transcription and translation, that are critical for the proper functioning of cells and organisms. If we need to reason about biological data, then algorithms that harness these relationships and processes are fundamental.

Logical reasoning is crucial to deduce properties about these relationships. For example, if we know that a dog is a mammal and a cat is a mammal, we should be able to deduce that they both give birth to live young, except for those mammals that are monotremes. On the other hand, if we know that insects cannot be mammals, then we can also deduce that a dog cannot be an insect. In the insular world of (deep) machine learning, where we assume that objects are atomic and independent, the incorporation of logical artifacts provides a much richer understanding of the world and its objects, leading to a more nuanced and accurate interpretation of the data.

There are concrete reasons to believe these frameworks address the specific weaknesses identified in LLMs. Neuro-symbolic systems can improve logical reasoning by combining neural pattern recognition with formal inference mechanisms, thereby enabling multi-step deduction that pure LLMs struggle with. They enhance interpretability, since the symbolic component provides human-readable explanations for the system's conclusions. They are more data-efficient, as structured knowledge representations reduce the need for massive training corpora. And they offer greater robustness against confabulation, because symbolic constraints can enforce consistency with known facts. Early evidence from systems, such as DeepMind's AlphaGeometry and AlphaProof, which combine neural networks with symbolic reasoning engines to prove mathematical theorems at a competition level, supports these claims (see BELLE & MARCUS 2025 for further discussion).

Generally, statistical relational/logical learning systems bring complementary strengths: they excel at pattern recognition over structured, relational data; they can learn new relationships and rules, and benefit from existing expert-provided ones; and unlike classical (non-probabilistic) logics, they provide syntax for uncertain and incomplete information. Neurosymbolic approaches build on statistical and logical languages by also providing mechanisms to support or leverage neural networks.

However, these approaches are not without limitations. Neuro-symbolic systems can be complex to architect, requiring careful integration of fundamentally different computational paradigms. They may struggle with the scale and messiness of real-world

data compared to the brute-force “data dump” capacity of large language models. (To be fair, this is a bit of a misunderstanding because there’s quite a bit of engineering expertise that goes into building machine learning models, from data cleaning to pre-processing to labeling. So, it’s not quite a one-push-button option. And when these models are used without care, they can lead to misleading results without being able to identify where the issues lie.) And they typically rely on pre-defined knowledge bases that may not be comprehensive or current.

Striking the right balance between these approaches is an active area of research. The most promising direction may be one in which LLMs serve as a front-end for knowledge extraction and natural language interaction, while symbolic and relational components provide the inferential backbone and consistency guarantees (TANG & BELLE 2024; PANAS, PAYANI & BELLE 2025). This division of cognitive labor, that of matching each task to the computational paradigm best suited to it, makes a strong case for hybrid architecture rather than simply scaling models to bigger and bigger representations (BELLE & MARCUS 2025).

Conclusion: Representation Matters

In this article, we admit that Large Language Models have demonstrated impressive capabilities in text generation and pattern recognition, issues regarding intellectual property notwithstanding. However, their fundamental limitations in logical reasoning and deep understanding raise important questions about their role in AI development. The informal evidence presented suggests that these models, despite their vast training data and sophisticated architecture, often fail at basic logical tasks and lack true comprehension of the relationships they describe. At this point, there is growing work on the logical and mathematical limitations of LLMs, which can be broadly categorized into limitations in terms of reasoning, knowledge, understanding, and planning.

Rather than viewing LLMs as a path to artificial general intelligence or an existential threat, we should recognize them as useful but limited tools that complement, rather than replace, traditional approaches to AI. The future of trustworthy artificial intelligence likely lies in the synthesis of multiple approaches, particularly in the integration of logical reasoning frameworks with neural architectures through neuro-symbolic AI and/or statistical relational learning.

It is worth reflecting on the nature of knowledge representation and how it affects modeling and learning. Regardless of whether we use a logical representation (such as a logic program), a probabilistic one (such as a Bayesian network), or a connectionist one (such as a neural network), our goal is to represent information and capture the knowledge of a robot or vision system. We want to understand how this representation produces interesting conclusions that determine what the robot knows and understands about the world. While acquisition is undoubtedly important, for now, we are focusing on the language used for representing information. Thus, it is crucial to equip the underlying formalism with sufficient expressive power to reason about the world: that is, structures to talk about the world and the objects in it, and how they relate to other pieces of knowledge. This emphasizes the importance of notation, language, vocabulary, and grammar. Language shapes thought, and a system cannot be expected to answer a question that it cannot express.

Which language should we choose, and what should we represent? Turing machines are foundational structures for investigating the computable universe. They offer minimal, yet powerful read and write operations. However, being a computational abstraction means we have no means to concretely talk about what the robot knows about the environment it

operates in. Likewise, binary system code runs robotic software but is too coarse and opaque to explain a purposeful agent. Vectorized representations and non-linear decision boundaries, as obtained by sigmoid neural units over many interconnected layers, are also too coarse to explain how the world works. Although these models seem capable of capturing a potentially infinite hypothesis space and high-dimensional data, including high-resolution images such as mammograms, their operational descriptions are too opaque for humans to decipher and to use to express facts and knowledge.

We require a modeling framework that is understandable to domain experts, allows the updating and provision of new information, and is suitable for computational processes. In particular, by a representation, we mean a mapping from one entity to another. The former is concrete and computationally processable. For example, "rain" might stand for the assertion that it is raining today. Reasoning is the computational processing of representations to produce new knowledge.

The distinction between explicit and implicit knowledge is also important: explicit knowledge is often the information that is directly modelled or provided by the domain expert. This could come in the form of rules, databases, knowledge bases, graphs, or any other structured data. Implicit knowledge is what is obtained from the explicit knowledge through one or more reasoning steps. Formal (symbolic) logic, where all information is stored explicitly or implicitly using symbols, was a working candidate in the AI circles from the 60s. Here, reasoning can be understood in two ways. First, it can define the mathematical framework for obtaining explicit knowledge from implicit knowledge. In logic, we often provide a semantics and/or proof theory that determines when sentences are "entailed" from a knowledge base (a set of sentences). Second, reasoning can provide an implementation strategy for computing entailments: an algorithm that takes a knowledge base and a query as input and outputs whether the query is entailed. The fields of neuro-symbolic AI and statistical relational learning take these representation issues, and the distinction between explicit and implicit knowledge, seriously while leveraging state-of-the-art advances in learning and language, including black-box models such as neural networks and large language models. This integration presents exciting opportunities for future research in AI.

Declarations

*The author declares no conflict of interest or ethical issues for this work.

Relevant Literature

1. AZAMFIREI, Razvan, KUDCHADKAR, Sapna R. & FACKLER, James. "Large Language Models and the Perils of Their Hallucinations." [Critical Care], 2023. Vol 27. No1.
2. BACCHUS, Fahiem., HALPERN, Joseph. Y., & LEVESQUE, Hector. J. "Reasoning about Noisy Sensors and Effectors in the Situation Calculus." [Artificial Intelligence], pp171-208, 1999. Vol 111. No1-2.
3. BANKO, Michele. n.d.n. "Open Information Extraction from the Web." [IJCAI], pp2670-2676, 2007. Vol 7.
4. BELLE, Vaishak. "Logic Meets Learning: From Aristotle to Neural Networks," [Neuro-Symbolic Artificial Intelligence: The State of the Art, Pascal HITZLER &

- Md Kamruzzaman SARKER, Eds.], pp78-102, 2021. IOS Press: Amsterdam.
5. BELLE, Vaishak & LEVESQUE, Hector J. "Reasoning about Probabilities in Dynamic Systems Using Goal Regression." [Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence], 2013. AUAI Press.
 6. BELLE, Vaishak, & MARCUS, Gary. "The Future Is Neuro-Symbolic: Where Has It Been, and Where Is It Going." [The 40th Annual AAAI Conference on Artificial Intelligence], 2026. AAAI Press: Washington.
 7. BELLE, Vaishak & PAPANTONIS, Ioannis. "Principles and Practice of Explainable Machine Learning." [arXiv Preprint], N.P, 2020. [Web](#).
 8. BENDER, Emily M. n.d.n. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" [Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency], pp610-623, 2021. ACM: New York.
 9. BIRHANE, Abeba. n.d.n. "Science in the Age of Large Language Models." [Nature Reviews Physics], pp1-4, 2023.
 10. BRACHMAN, Ronald. J. & LEVESQUE, Hector. J. [Machines Like Us: Toward AI with Common Sense], 2022. MIT Press: Cambridge.
 11. CARLINI, Nicholas. n.d.n. "Extracting Training Data from Large Language Models." [Proceedings of the 30th USENIX Security Symposium], pp2633-2650, 2021. Vol 6. [Web](#).
 12. CHALMERS, David. J. "Could a Large Language Model Be Conscious?" [arXiv Preprint], 2023. [Web](#).
 13. CHATILA, Raja. n.d.n. "Trustworthy AI." [Reflections on Artificial Intelligence for Humanity, Bertrand BRAUNSCHWEIG & Malik GHALLAB Eds.], pp13-39, 2021. Springer: Cham.
 14. CHEN, Zijun & CHAN, Jason. "Large Language Model in Creative Work: The Role of Collaboration Modality and User Expertise." [Management Science], pp9101-9117, 2024. Vol 70.
 15. CONFALONIERI, Roberto. "Using Ontologies to Enhance Human Understandability of Global Post-hoc Explanations of Black-Box Models." [Artificial Intelligence], pp1-20, 2021. Vol 296. No103471.
 16. COZMAN, Fabio. G. & MUNHOZ, Hugo. N. "Some Thoughts on Knowledge-Enhanced Machine Learning." [International Journal of Approximate Reasoning], pp308-324, 2021. Vol 136.
 17. CRESWELL, Antonia., SHANAHAN, Murray., & HIGGINS, Irina. "Selection-Inference: Exploiting Large Language Models for Interpretable Logical Reasoning." [arXiv Preprint], 2022. [Web](#).
 18. D'AVILA GARCEZ, Artur., & LAMB, Luis. C. "Neurosymbolic AI: The 3rd Wave." [arXiv E-Prints], 2020. [Web](#).
 19. DE ANGELIS, Luigi. n.d.n. "ChatGPT and the Rise of Large Language Models: The New AI-Driven Infodemic Threat in Public Health." [Frontiers in Public Health], pp01-08, 2023. Vol 11. No1166120. [Web](#).
 20. DONG, Xin. n.d.n. "Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion." [Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining], pp601-610, 2014. ACM: New York.
 21. GUNJAL, Anisha, YIN, Jihan & BAS, Erhan. "Detecting and Preventing Hallucinations in Large Vision Language Models." [Proceedings of the AAAI Conference on Artificial Intelligence], pp18135-18143, 2024. Vol 38. No16.
 22. HEAVEN, Will D. "Geoffrey Hinton Tells Us Why He Is Now Scared of the Tech He Helped Build." MIT Technology Review, 02 May 2023. [Web](#).

23. HITZLER, Pascal. [Neuro-Symbolic Artificial Intelligence: The State of the Art], 2022. IOS Press: Amsterdam.
24. Hong, Euny. “The Viral “Car Wash” LLM Challenge Doesn’t Mean What You Think It Means,” 05 March 2026. [Web](#).
25. JANG, Myeongjun & LUKASIEWICZ, Thomas. “Consistency Analysis of ChatGPT.” [arXiv Preprint], 2023. [Web](#).
26. KALAI, Adam. T. n.d.n. “Why Language Models Hallucinate.” [arXiv Preprint], 2025. [Web](#).
27. KAMBHAMPATI, Subbarao. n.d.n. “Position: LLMs Can’t Plan, but Can Help Planning in LLM-Modulo Frameworks.” [Proceedings of the 41st International Conference on Machine Learning], pp22895-22907, 2024. No921. [Web](#).
28. KICIMAN, Emre. n.d.n. “Causal Reasoning and Large Language Models: Opening a New Frontier for Causality.” [Transactions on Machine Learning Research], 2023. [Web](#).
29. KOTHARI, A. N. “ChatGPT, Large Language Models, and Generative AI as Future Augments of Surgical Cancer Care” [Annals of Surgical Oncology], pp1-3, 2023.
30. LAKE, Brenden. M., SALAKHUTDINOV, Ruslan & TENENBAUM, Joshua. B. “Human-Level Concept Learning through Probabilistic Program Induction.” [Science], pp1332-1338, 2015. Vol 350.
31. LEVESQUE, Hector. J. [Thinking as Computation: A First Course], 2012. MIT Press: Cambridge.
32. LIU, Yongmei, LAKEMEYER, Gerhard & LEVESQUE, Hector. J. “A Logic of Limited Belief for Reasoning with Disjunctive Information.” [Principles of Knowledge Representation and Rationality: Proceedings of the Ninth International Conference], pp587-597, 2004.
33. MANHAEVE, Robin. n.d.n. “DeepProbLog: Neural Probabilistic Logic Programming.” [Advances in Neural Information Processing Systems], 2018. Vol 31.
34. MARCUS, Gary. [The Algebraic Mind: Integrating Connectionism and Cognitive Science], 2003. MIT Press: Cambridge.
35. MARCUS, Gary & DAVIS, Ernest. [Rebooting AI: Building Artificial Intelligence We Can Trust], 2019. Vintage: New York.
36. MATEUS, Paulo. n.d.n. “Probabilistic Situation Calculus.” [Annals of Math. and Artif. Intell.], pp393-431, 2001. Vol 32. No1–4.
37. MCILRAITH, Sheila. A., SON, Tran., & ZENG, Honglei. “Semantic Web Services.” [IEEE Intelligent Systems], pp46-53, 2001. Vol 16. No2.
38. MICELI-BARONE, Antonio. V. n.d.n. “The Larger They Are, the Harder They Fail: Language Models Do Not Recognize Identifier Swaps in Python.” [arXiv Preprint], 2023. [Web](#).
39. NIU, Feng., ZHANG, Ce., RÉ, Christopher & SHAVLIK, Jude. “DeepDive: Web-Scale Knowledge-Base Construction Using Statistical Learning and Inference.” [VLDB], pp25-28, 2012. Vol 12.
40. PANAS, Daga., PAYANI, Ali., & BELLE, Vaishak. “Unreasonable Effectiveness of LLM Reasoning: A Doubly Cautionary Tale of Temporal Question-Answering.” [Transactions on Machine Learning Research], 2025. [Web](#).
41. REITER, Raymond. [Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systems], 2001. MIT Press: Cambridge.
42. RUDIN, Cynthia. “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead.” [Nature Machine Intelligence], pp206-215, 2019. Vol 1. No5.

43. SHOJAEE, Parshin. n.d.n. "The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity." [arXiv Preprint], 2025. [Web](#).
44. SILEO, Damien & LERNOULD, Antoine. "MindGames: Targeting Theory of Mind in Large Language Models with Dynamic Epistemic Modal Logic." [arXiv Preprint], 2023. [Web](#).
45. ŠLAPETA, Jan. "Are ChatGPT and Other Pretrained Language Models Good Parasitologists?" [Trends in Parasitology], pp314-316, 2023. Vol 39. No5.
46. SYPHERD, Chris, TANG, Weizhi & BELLE, Vaishak. "Breaking the Illusion: Revisiting LLM Anthropomorphism." [The 4th International Conference on Human and Artificial Rationalities], pp1-19, 2025. Springer Nature: Cham.
47. TANG, Weizhi & BELLE, Vaishak. "TOM-LM: Delegating Theory of Mind Reasoning to External Symbolic Executors in Large Language Models," [International Conference on Neural-Symbolic Learning and Reasoning], pp245–257, 2024. Springer: Cham.
48. VAN BENTHEM, Johan. F., & TER MEULEN, Alice. [Handbook of Logic and Language], 2010. Elsevier: Amsterdam.
49. VASWANI, Ashish. n.d.n. "Attention Is All You Need." [Proceedings of the 31st International Conference on Advances in Neural Information Processing Systems], pp6000-6010, 2017. [Web](#).
50. WANG, Chong. n.d.n. "LLMs Meet Library Evolution: Evaluating Deprecated API Usage in LLM-Based Code Completion." [2025 IEEE/ACM 47th International Conference on Software Engineering (ICSE)], pp781–781, 2025. IEEE Computer Society.
51. WEI, Jason. n.d.n. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." [Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22)], pp24824-24837, 2022. No1800. [Web](#).
52. ZHANG, Hanlin. n.d.n. "Improved Logical Reasoning of Language Models via Differentiable Symbolic Programming." [arXiv Preprint], 2023. [Web](#).
53. ZHOU, Jiawei. n.d.n. "Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions." [Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems], pp1–20, 2023. ACM: New York.
54. ZHOU, Kaitlyn., JURAFSKY, Dan., & HASHIMOTO, Tatsunori. "Navigating the Grey Area: Expressions of Overconfidence and Uncertainty in Language Models." [arXiv Preprint], 2023. [Web](#).
55. ZORPETTE, Glenn. "Just Calm Down About GPT-4 Already And stop confusing performance with competence, says Rodney Brooks," 17 May 2023. [Web](#).