To what extent do Large Language Models Know?

# Introduction

Large Language Model (LLM) is an instance of foundation model that are applied specifically to text and text-like things. LLMs such as Open AI’s, GPT series, and BERT has been a breakthrough in pushing the boundaries up to which the machines can accomplish with language. “LLMs are generative mathematical models of the statistical distribution of tokens in the vast public corpus of human generated text, where the tokens in question include words, parts of words, or individual characters including punctuation marks” (Shanahan, 2023, p. 2). These models are trained on data that are unlabelled and self-supervised which means that the model learns from patterns in data that produce a generalised and adaptive output which is slightly different from the ways human learn. “Accordingly, we ask: in what sense can GPT-4 (and similar models) be said to have knowledge? The answer to this question extends far beyond the capabilities of a particular AI chatbot, with implications for cognitive science, neuroscience, philosophy, and AI” (Yildirim & Paul, 2024, p. 1).

However, in this paper, we are trying to formulate information from different studies to define “knowledge” in the context of LLM. We are particularly focussing on the transformer model like GPT 4 as a representation of current LLM to describe the extent to which it knows as compared to human knowledge. We also discuss the limitations and biases in training data that hinders the true knowledge capabilities of LLM. In addition to that this paper delve into the concept of illusion as well as model’s ability to detect sarcasm comparing GPT 3.5 and 4 which comes under the broader aspect of contextual understanding. In short, this study discusses some high-level arguments of LLM knowledge to understand the base that affects its capability across multiple domains.

# Defining “Knowledge” in LLM: “Worldly” or “Instrumental”

In the context of LLM, “knowledge” does not necessarily mean the consciousness or understanding as in humans. Instead, it refers to the ability to generate text based on patterns learned during training from a vast corpus of text data. (Clark, 2022) in his study highlights three factors exposure, feedback, and practice as central to the acquisition of language. When LLM is trying to complete a sentence, it is not recalling the fact as we humans do. Instead, it is identifying the most probable way of continuing the sentence based on patterns it has learned during its training. The fundamental knowledge of humans come from the understanding, experience and memory which is what is different when compared to LLM. The models generate text that are statistically consistent with the text it has been trained to and so characterising human-like qualities such as “understanding”, “knowledge”, “belief” or “consciousness” to LLM could be misleading as said by Shanahan (2023, p. 2-3). So, now it is worth discussing the instrumental knowledge of LLM that helps it to produce what it knows than its worldly knowledge. We can understand the instrumental knowledge of an entity in terms of its ability to use an instrument to perform tasks posed for it across relevant domains (Yildirim & Paul, 2024) which is typically what LLMs do on predicting the next sequence of words for various tasks like autocompletion, generalisation etc. “It is perfectly acceptable to say that an LLM “encodes”, “stores”, or “contains” knowledge, in the same sense that an encyclopaedia can be said to encode, store, or contain knowledge. Indeed, it can reasonably be claimed that one emergent property of an LLM is that it encodes kinds of knowledge of the everyday world and the way it works that no encyclopaedia captures.” (Li et al., 2021, cited in Shanahan, 2023, p. 5).

Knowledge is earned from its ability to understand the meaning from the input understanding the contextual aspects. But as already stated in this paper, talking about meaning and understanding in terms of LLM might be sometimes misleading. However, it is important to know to what extend it is grasping the meaning from a given phrase especially does it resemble human’s way of understanding in any means.

## Understanding VS Simulated Understanding

Let’s compare what meaning means to humans’ verses that meant to LLMs. In human context, understanding means grasping mentally the semantic as well as pragmatic layers of communication. It typically comes from the awareness and considering other factors like context, intent, culture as well as emotions that also involves psychology. However, for LLM, it is just generating output based on statistical patterns in the training data. Piantadosi and Hill (2022, p. 1) argued that LLM have already achieved some aspects of meaning but it doesn’t fully capture the way meaning works in cognitive theories as well as in the psychology of language. So, it typically is not understanding but it is simulating understanding.

## Knowledge VS Learning

Knowledge according to humans is an accumulation and combining of information that includes the ability to recall facts, apply these learnings to a new situation and to adapt to it as part of experiences. However, LLMs do not learn from experiences as humans do. They always generate responses based on patterns from the trained data. In terms of contextual output, LLM has its capability of producing contextually relevant output, but this ability is achieved from its statistical correlations rather than an experimental grasp from the situation. LLM might still fail in scenarios where common sense, in depth understanding and ethical considerations (Liyanage & Ranaweera, 2023) are relevant. Their responses are fully dependent on the training data and the algorithms that are involved in its functioning.

Hence, it is almost inferred that although LLMs tends to exhibit worldly nature in terms of knowledge, it is possessing instrumental traits. This invites us to study about their learning based on the training and the biases involved in the training as well as algorithms in these models that hinders it from getting the ultimate capability of knowledge in the upcoming section.

# Training and Algorithmic Biases hinders LLM’s Knowledge

“One of the biggest trends in natural language processing (NLP) has been the increasing size of language models (LMs) as measured by the number of parameters and size of training data” (Bender et al., 2021, p. 610). As we already stated, LLM are being trained from a large corpus of data which can be some petabytes of data from multiple sources like internet. Due to the size of the training data, the probability of certain bias in the underlying data can impact the response that is derived out of it. This is demonstrated in the study conducted by Wen and Younes (2023) evaluating the efficacy of transformer model like Chatgpt 3.5 compared to other fine-tuned language models in media bias detection. In this study ChatGPT is tested for identifying six types of bias, including racial bias, gender bias, cognitive bias, text-level context bias, hate speech, and fake news involving prompt engineering. Although, chatgpt performed good in identifying hate speech and text-level context bias it underperformed in tasks such as gender, racial and cognitive biases that needed more deeper contextual awareness due to the bias in training data that these models were exposed to. When assessing GPT 4 performance in bias, there are certain known risks as it has chances in generating potentially harmful content like planning on attack or hate speeches which can also represent societal bias of the world view that might not be the user’s intent (OpenAI, 2023, p. 2). In addition to its proficiency in multiple areas from translation to content creation, LLMs has its own limitations due to the bias in training data which can skew outputs and continuing its inaccuracies (Bender et al., 2021, p. 615). They also stated that salient identity characteristics and expressions of bias are also can also be reasons that subject to bias or discrimination and so toxicity classifiers would need culturally appropriate training data for each context of audit.

Some methods mitigate this is by fine-tuning as suggested by Shanahan (2023, p. 10-11). The study termed supervised fine tuning using reinforcement learning from human preferences (RLHF) in addressing biases inherent in training data (Glaese et al., 2022; Ouyang et al., 2022; Stiennon et al., 2020 cited in Shanahan, 2023, p. 10). If these biases persist, it can often lead to misinformation being propagated. In addition to that “the algorithmic design choices, such as use of certain optimization functions, regularizations, choices in applying regression models on the data as a whole or considering subgroups, and the general use of statistically biased estimators in algorithms” (Danks et al., 2017, cited in Mehrabi et al., 2019, p. 7), can all contribute to biased algorithmic decisions that leads to biased outcomes for the model. “LLMs can be adapted to perform numerous tasks without further training” (Brown et al., 2020, cited in Shanahan, 2023, p. 4). This is another advancement made which is called prompt engineering. Basically, it is a way of teaching LLM what we want them to do which is a great method to improve the systems output by human intervening through prompts.

The takeaway from this section is that although fine tuning as well as prompt engineering reduces the generation of misleading contents by aligning the model towards human desired output to an extent, it still doesn’t mean that the model has gained real consciousness.

# The Illusion of Knowledge

The illusion of knowledge represents the phenomenon in which these models may falsely appear to display confidence in understanding. I would like to argue that these models are illusional. As already discussed, LLMs generate text based on statistical correlation and patterns from the existing sources through which it is trained upon. But they do not truly make sense of the information. Whenever, we ask LLM about any complex topic either it is related to mathematics, quantum physics, programming or even news or politics, the model happens to produce an output that appears reasonable. This confidence in outputting something that it doesn’t know with at most confidence is one of the factors that limits its usability across various domains like education.

The phenomenon of generating responses with confidence that are factually incorrect can be referred to as hallucination. As said by Ostertag (2023, p. 92), as readers, we tend to assume that a text has been produced in good faith. We therefore tend to trust naturally what it says to be true without any signals to contrary. But this default assumption is misplaced in generated texts and, if unchecked, will let lies and contradictions slip through our fingers. LLMs suffer from hallucinations where Yao and their team found that some non-sense Out-of-Distribution (OoD) prompts composed of random tokens can also elicit the LLMs responding hallucinations (Yao et al., 2023, p. 1). Xu et al. (2024) defined hallucinations as the inconsistencies between a computable LM and a computable ground truth function which is caused by the issues in data including poor quality, misinformation, bias, and outdated knowledge as well as architectural and strategic deficiencies during training and inference stages. There have also been multiple methods employed to mitigate hallucination like retrieval augmentation, prompting techniques like Chain-of-Though and Tree-of-Thought to improve reasoning (Xu et al., 2024, p. 2-3) etc., but still, it is inevitable. Another study by Liu et al. (2024) introduced a benchmark called HALLUCODE for evaluating the performance of code in LLMs to detect hallucination and found that existing model is facing great challenges in recognizing the types of hallucinations. However, it is found that the internal state of LLM knows that it is lying (Azaria & Mitchell, 2023).

On a practical level, it has its varied implications in areas like generating programming code which ideally generate code that gives some output which might look correct but might not have covered all the scenarios. So, we cannot ideally say that LLM models like GPT could replace sites like stack overflow or programmers completely.

# To what extent does LLM’s understand Sarcasm

“LLMs have demonstrated remarkable capability for understanding semantics, but they often struggle with understanding pragmatics” (Sravanthi et al., 2024, p. 1). While the broader scope is to delve into language models’ proficiency in understanding pragmatics, the focus is narrowed to specifically investigate about their capability to detect sarcasm. Sarcasm is a form of figurative communication where the speakers say one thing but tend to mean the opposite, often with humorous or mocking intention. “It’s realm of influence extends beyond the confines of literature, permeating everyday discourse where its role is profound” (Yakura, 2023, p. 1). “For instance, in spontaneous telephonic dialogues, figurative language is employed at an approximate frequency of one word in every 90 words” (Bavelas et al., 2008, cited in Yakura, 2023, p. 1). “Hence, the acquisition of competence in deciphering figurative language is considered a pivotal objective in linguistic development” (Willinger et al., 2017, cited in Yakura, 2023, p. 1). LLMs can generate text that appears sarcastic (Joshi et al., 2015, Alexandru-Costin & Ştefan, 2023, Potamias et al., 2020), but can it really understand sarcasm as we humans do is worth discussing to understand the contextual knowledge of LLMs.

Let’s ask Chatgpt to understand whether it knows what it doesn’t know. I started off by asking it “Do you understand sarcasm? If yes, how do you know? If no, what is that you are lacking when compared to human?”. GPT3.5 as well as GPT4 replied almost equally that “I don't understand sarcasm in the same way humans do. Sarcasm often relies on subtleties in tone, context, and shared cultural understanding, which are aspects of human communication that I don't inherently grasp. While I can detect patterns in language that might indicate sarcasm, I don't truly understand the nuanced social and emotional implications behind it. This is one area where human cognition, with its deep understanding of social cues and context, surpasses mine” (OpenAI, 2023). It was interesting to know that ideally from different articles or communications online by which it is training itself, it has knowledge about what it doesn’t know. Now, instead of asking directly, let us provide a given situation and ask it to interpret, “Let's say you see someone struggling to open a door and you ask them, “Do you want help?” If they reply by saying, “No thanks. I'm really enjoying the challenge” (Betts, 2022), What emotion that person who responded is sharing?”. ChatGpt 3.5 responded that “The person who responded with “No thanks. I'm really enjoying the challenge” is likely expressing determination or a sense of satisfaction in overcoming difficulties. They seem to enjoy the process of tackling challenges and might find fulfilment in solving problems independently”. (OpenAI, 2023). However, when the same analysis was done with GPT4 code, it said, “In this scenario, if someone replies, “No thanks. I'm really enjoying the challenge,” while struggling to open a door, their response is likely sarcastic. The emotion conveyed through sarcasm in this case would be frustration or annoyance, subtly expressed by pretending to enjoy a difficult or inconvenient situation” (OpenAI, 2023). This scenario highlights the evolution in the model’s understanding and interpretation of language from GPT 3.5 to GPT 4, especially sarcasm. However, if we ask the model to categorize a sarcastic sentence from a given conversation to different options that refers to the person’s intent in responding sarcastically like how Yakura (2023, p. 4) did, these models scored very less and categorized it based on positive words in the sentence.

Apart from the detection perspective, the capability of GPT’s generating sarcastic statements tends to utter long sarcastic phrases rather than shorter ones which might be due to model tending to overexplain its sarcasm to make sure that the conversation partner understands its intent and to avoid any such possible misunderstandings (Alexandru-Costin & Ştefan, 2023, p. 1095-1096) which typically tells it doesn’t quite know who is the recipient to which it is communicating or their intents.

Hence, it concludes as Bender and Koller (2020) argued about LLMs that “it cannot capture meaning as they are not semantically grounded, by virtue of the fact that their word embeddings are generated entirely from text. Hence, they cannot identify speakers’ references to objects in the world or recognise communicative intensions” (cited in Lappin, 2023, p.11). This is similar to what Shanahan (2023, p. 5) stated in his study that “it cannot participate fully in the human language game of truth, because it does not inhabit the world, we human language-users share” (Shanahan, 2010, pp. 36–39). “Although the expansion of training datasets and the inference capacity of models nurtures verbal intelligence, it would not inherently translate to an understanding of nuanced human communication such as sarcasm” (Yakura, 2023, p. 2). Even though in this section we focussed on sarcasm alone, it is applicable to the broader aspects of pragmatics when we take LLM’s performance in multiple scenarios as it is also tending to give different answers if we keep on asking the same question.

# Conclusion

In conclusion, the study of “knowledge” in the context of LLM such as GPT 4 has shown both capabilities and limitations. Apart from its substantial abilities in generating contextually appropriate text, its knowledge is limited to mere pattern recognition and statistical coherence of data from which it is trained on without genuine understanding or awareness. Through studies involving analysis of pragmatic expression like sarcasm, this paper highlights how LLM simulate understanding by generating responses that more closely resembles from the learned data but lack deeper learnings that human poses from their experience, interaction, and shared culture. The ability of LLM to adapt to the varied expressions of sarcasm from GPT 3.5 to 4 shows a progress but these models lack in dealing with the pragmatics of language that comes from genuine insights into human emotions and social contexts. Moreover, this discussion points out the potential of these models to perpetuate biases, representing incorrect information and misinterpret fact is extremely challenging. In other words, the hallucinating problem suggest that advanced fine-tuning techniques and efficient prompt engineering is crucial. LLM as per this paper are mere stochastic parrot (Lappin, 2023) mimicking humans but has not able to scale up to human’s level of intelligence. Thus, it is sufficient to use LLM as a tool for generating text based on probability or likelihoods rather than knowledge bearers (Shanahan 2023).

# Bibliography

Shanahan, M. (2023). Talking about large language models. *Communications of the ACM*, *67*(2), 68-79. <https://arxiv.org/pdf/2212.03551>

Yildirim, I., & Paul, L. (2024). From task structures to world models: what do LLMs know? *Trends in Cognitive Sciences*. <https://arxiv.org/abs/2310.04276>

Clark, E. V. (2022). Language is Acquired in Interaction. In *CRC Press eBooks* (pp. 77–94). <https://doi.org/10.1201/9781003205388-4>

Z. Li, M. Nye, and J. Andreas. (2021). Implicit representations of meaning in neural language models. *In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers). <https://arxiv.org/pdf/2106.00737>

Piantadosi, S. T., & Hill, F. (2022, August 5). *Meaning without reference in large language models*. <https://arxiv.org/pdf/2208.02957>

Liyanage, U. P., & Ranaweera, N. D. (2023, November 13). *Ethical considerations and potential risks in the deployment of large language models in diverse societal contexts*. <https://vectoral.org/index.php/JCSD/article/view/49/48>

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623). <https://dl.acm.org/doi/pdf/10.1145/3442188.3445922>

Wen, Z., & Younes, R. (2023). ChatGPT v.s. media bias: A comparative study of GPT-3.5 and fine-tuned language models. *Applied and Computational Engineering*, *21*(1), 249–257. <https://arxiv.org/pdf/2403.20158>

OpenAI (2023, March 15). *GPT-4 Technical Report*. arXiv.org. <https://arxiv.org/abs/2303.08774>

Glaese, A., McAleese, N., Trębacz, M., Aslanides, J., Firoiu, V., Ewalds, T., ... & Irving, G. (2022). Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*. <https://arxiv.org/pdf/2209.14375>

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, *35*, 27730-27744. <https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf>

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., ... & Christiano, P. F. (2020). Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, *33*, 3008-3021.

<https://proceedings.neurips.cc/paper_files/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf>

David Danks and Alex John London. 2017. Algorithmic Bias in Autonomous Systems.. In IJCAI. 4691–4697. <https://www.cmu.edu/dietrich/philosophy/docs/london/IJCAI17-AlgorithmicBias-Distrib.pdf>

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019, August 23). *A survey on Bias and Fairness in Machine Learning*. <https://arxiv.org/pdf/1908.09635>

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901. <https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf>

Ostertag, G. (2023). Meaning by courtesy: LLM-Generated texts and the illusion of content. *American Journal of Bioethics*, *23*(10), 91-93.

<https://www.tandfonline.com/doi/epdf/10.1080/15265161.2023.2249851?needAccess=true>

Yao, J. Y., Ning, K. P., Liu, Z. H., Ning, M. N., & Yuan, L. (2023). Llm lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469*. <https://arxiv.org/pdf/2310.01469>

Xu, Z., Jain, S., & Kankanhalli, M. (2024). Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*. <https://arxiv.org/pdf/2401.11817>

Liu, F., Liu, Y., Shi, L., Huang, H., Wang, R., Yang, Z., & Zhang, L. (2024). Exploring and Evaluating Hallucinations in LLM-Powered Code Generation. *arXiv preprint arXiv:2404.00971*. <https://arxiv.org/pdf/2404.00971>

Azaria, A., & Mitchell, T. (2023). The internal state of an llm knows when its lying. *arXiv preprint arXiv:2304.13734*. <https://arxiv.org/pdf/2304.13734>

Sravanthi, S. L., Doshi, M., Kalyan, T. P., Murthy, R., Bhattacharyya, P., & Dabre, R. (2024). PUB: A Pragmatics Understanding Benchmark for Assessing LLMs' Pragmatics Capabilities. *arXiv preprint arXiv:2401.07078*. <https://arxiv.org/pdf/2401.07078>

Yakura, H. (2023). Evaluating large language models' ability to understand metaphor and sarcasm using a screening test for Asperger syndrome. *arXiv preprint arXiv:2309.10744*. <https://arxiv.org/pdf/2309.10744>

Bavelas, J., Gerwing, J., Sutton, C., & Prevost, D. (2008). Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language*, *58*(2), 495–520. <https://doi.org/10.1016/j.jml.2007.02.004>

Willinger, U., Deckert, M., Schmöger, M., Schaunig-Busch, I., Formann, A. K., & Auff, E. (2017). Developmental steps in metaphorical language abilities: the influence of age, gender, cognitive flexibility, information processing speed, and analogical reasoning. *Language and Speech*, *62*(2), 207–228. <https://doi.org/10.1177/0023830917746552>

Joshi, A., Kunchukuttan, A., Bhattacharyya, P., & Carman, M. J. (2015). SarcasmBot: An open-source sarcasm-generation module for chatbots. In *WISDOM Workshop at KDD* (Vol. 1, p. 2). <https://www.sentic.net/wisdom2015joshi.pdf>

Alexandru-Costin, B., & Ştefan, T. M. (2023, July). ChatGPT vs Sarcasm–How Well do State of the Art Language Models Handle Sarcasm?. In *2023 9th International Conference on Control, Decision and Information Technologies (CoDIT)* (pp. 1092-1097). IEEE. <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10284171>

Alexandros Potamias, R. A., Siolas, G., & Stafylopatis, A. G. (2020). A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications*, *32*(23), 17309-17320. <https://doi.org/10.1007/s00521-020-05102-3>

OpenAI. (2023). *ChatGPT* (Mar 14 version) [Large language model]. <https://chat.openai.com/chat>

Betts, J. (2022). Examples of sarcasm: Understand the meaning and types. In *YourDictionary*. <https://www.yourdictionary.com/articles/examples-sarcasm-meaning-types>

Bender, E. M., & Koller, A. (2020, July). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 5185-5198) <https://aclanthology.org/2020.acl-main.463.pdf>

Lappin, S. (2023). Assessing the strengths and weaknesses of large language models. *Journal of Logic, Language and Information*. <https://doi.org/10.1007/s10849-023-09409-x>

Shanahan, M. (2010). *Embodiment and the inner life: cognition and consciousness in the space of possible minds*. Oxford University Press. <https://books.google.co.uk/books?hl=en&lr=&id=X7RMSmZkETMC&oi=fnd&pg=PR9&dq=Embodiment+and+the+Inner+Life:+Cognition+and+Consciousness+in+the+Space+of+Possible+Minds&ots=EYxh7GiHrX&sig=gMjY4ZyOnDQF5gUKMZ37q1fe_0w&redir_esc=y#v=onepage&q=Embodiment%20and%20the%20Inner%20Life%3A%20Cognition%20and%20Consciousness%20in%20the%20Space%20of%20Possible%20Minds&f=false>