The Flashcard Sorter

Applicability of the Chinese Room Argument to Large Language Models

Johannes Brinz – University of Osnabrück

Preprint – February 2025

Abstract: Does the Chinese Room Argument (CRA) apply to large language models (LLMs)? The thought experiment at the center of the CRA is tailored to Good Old-Fashioned Artificial Intelligence (GOFAI) systems. However, natural language processing has made significant progress, especially with the emergence of LLMs in recent years. LLMs differ from GOFAI systems in their design; they operate on *vectors* rather than symbols and do not follow a program but instead *learn* to map inputs to outputs. Consequently, some have suggested that the CRA is no longer relevant in discussions surrounding artificial language understanding. Contrary to these authors, I argue that *if the CRA successfully demonstrates that implementing a symbolic computation is not sufficient for language understanding, then it also shows that implementing an LLM is not sufficient for language understanding*. At the core of my argument lies a thought experiment called “the flashcard sorter”.

Keywords: Chinese Room Argument, Language Understanding, Artificial Understanding, AI, Large Language Models

Introduction

This paper raises the question of whether the Chinese Room Argument (CRA) applies to large language models (LLMs). The CRA was first articulated in 1980, and the thought experiment at its center is tailored to the specific systems of that time, referred to as Good Old-Fashioned Artificial Intelligence (GOFAI) systems. GOFAI systems were syntactic engines that manipulated interpretable symbols according to a predetermined program. In analogy, the Chinese room thought experiment presents a scenario in which a person shuffles symbols following rules stated in a book. However, natural language processing has made significant progress, especially with the emergence of LLMs in recent years. LLMs differ from GOFAI systems in their design; they operate on *vectors* rather than symbols and do not follow a program but instead *learn* to map inputs to outputs. Consequently, some (e.g. Havlík 2023, Grindrod 2024, Vaidya 2024) have suggested that the CRA is no longer relevant in modern discussions surrounding artificial language understanding. Contrary to these authors, I argue that *if the CRA successfully demonstrates that implementing a symbolic computation is not sufficient for language understanding, then it also shows that implementing an LLM is not sufficient for language understanding.* Importantly, I do not claim that the CRA—either in its original or updated version—is sound. My aim is to suggest that the CRA is not rendered irrelevant simply because GOFAI systems have been replaced by LLMs. I begin by presenting a thought experiment called “the flashcard sorter”, which resembles the Chinese room thought experiment but is tailored to LLMs rather than GOFAI systems. I then discuss the analogy between the flashcard sorter and modern LLMs, along with four objections to my claim that the CRA applies to LLMs. I conclude by presenting a modified version of the CRA.

The Flashcard Sorter

Imagine the following scenario: You come across a call for participants in a research study at a reputable university. The study involves sorting flashcards that feature various different pictograms. The researchers tell you that they are interested in how individuals from various backgrounds approach the sorting task, regardless of their familiarity with the specific pictograms. You decide to volunteer convinced that you are contributing to an intriguing scientific project. Every new participant is required to undergo a two-phase trainee program to learn how to sort the flashcards correctly.

On the first day, you receive a stack of flashcards, a parcel containing several flashcard boxes and two books. Along with your equipment, you find a letter outlining the proceedings for the first trainee phase: Each day, you receive a parcel with several flashcard boxes and remove all of them except for one. Based on the rules in the MODEL book and depending on the boxes you took from the parcel, you are told to sort flashcards from the stack into one *new* flashcard box. Once you're done, the boxes you took from the parcel are no longer needed. After every workday you are required to update the rules in the MODEL book in preparation for the next day. For this you use the second rule book which is labeled "TRAINING". You compare the new flashcard box to the one remaining in the parcel. Based on this comparison, the TRAINING book provides you with precise instructions on how to update the rules in the MODEL book. The next day you repeat the same procedure: You receive a parcel with pictogram flashcard boxes, sort flashcards in a new box according to the MODEL rules, compare the sorted box to the remaining one, and update the MODEL rules according to the TRAINING book.

In phase two, you receive instructions that from now on you take *all* the boxes from the parcel and begin sorting flashcards into *multiple* new boxes, again following rules in the MODEL book. Once sorted, you send back the new flashcard boxes. Also, you are no longer supposed to compare your

sorting result to a remaining box. Instead, every morning you receive *one additional flashcard.* Depending on that card, the rules in the TRAINING book tell you how to update the MODEL rules. Every morning, you update the MODEL according to the TRAINING book and the symbol on the extra card, and get to work.

After these two training phases, you no longer need the TRAINING book and only have to sort flashcards into boxes according to the MODEL. So you do. Every morning you receive a parcel with flashcard boxes, sort them according to the MODEL rules, and send them back.

Large Language Models

Some readers might have noticed that the activities of the flashcard sorter are reminiscent of the processes in a LLM. Let me spell out the analogies explicitly. LLMs do not represent words with simple symbols, but rather with high dimensional vectors in a, so called, embedding space. Each word gets assigned a specific vector, such that words that are similar in meaning are represented with vectors that are close to each other. *The flashcard boxes in the thought experiment are analogous to embedding vectors*, as they represent words. Every flashcard is analogous to an entrey in the vector, and the number of flashcards per box, is analogous to the dimensions of the embedding space. LLMs take prompts as input, i.e. list of words, often in form of questions or instructions. *The parcels that the flashcard sorter receives are analogous to prompts*, in that they contain representations of words and are used as input. LLMs *learn* the mapping between inputs/prompts and outputs/responses, rather following a predetermined program. Such mappings are called models and consist of vast numbers of parameters that are iteratively updated during training. *The MODEL rule-book is analogous to the model of LLMs*. Most LLMs undergo a two-tier training process. First, they are pre-trained on vast amounts of data usually drawn from the internet. During pre-training the LLM is prompted with a text

snippet that has the last token removed, e.g. “A loud bang can be an emotional”, and is supposed to predict the correct next word, e.g. “trigger”. It then compares how close its prediction was compared to the actual last token, and then updates its model parameters accordingly. Since this does not require external supervision, this process is sometimes called “self-supervised” learning. *The first phase of the trainee program is analogous to the pre-training procedure of LLMs.* Additionally, most LLMs undergo a second training phase called “fine-tuning”, during which their responses are evaluated by human interlocutors. Based on whether the feedback from these humans is positive or negative, the model parameters are adjusted accordingly. *The second phase of the trainee program is analogous to LLM fine-tuning, and the single card that the flashcard sorter receives during phase two of the trainee program is analogous to the grade from human feedback.* Exactly how the model parameters of a LLM are to be updated during pre-training and fine-tuning is determined by the training algorithm. *The rules in the TRAINING book are analogous to the training algorithm of LLMs.* Whereas LLMs usually run on large server infrastructures, different hardware options are available. Some users implement LLMs locally on their computers. *Similarly, the flashcard sorter implements the computation of an advanced LLM.*

The Flashcard Sorter II: Accusations of the Police

Your routine as the flashcard sorter goes on for a couple of months until, eventually, the police ring your doorbell and ask you to come to the precinct with them. With a clear conscience, you comply. To your great surprise, you learn that you are facing charges as a member of a terrorist organization. They tell you that the flashcard boxes that you were processing represented words and that you were giving detailed descriptions on how to build a bomb. Completely dazzled, you ask how this is possible, and they explain that the pictograms are actually ancient Egyptian numbers. The approximately 12,000

cards per box were used to encode the meaning of a word in a high-dimensional vector space. The parcels you received in the morning are referred to as a “prompt,” while the results you sent back are called the “response.” They claim that you received questions about, and gave answers to, how to build a bomb.

Naturally, you defend yourself by pointing out that all you were doing was shuffling meaningless symbols according to certain rules and that you had no idea what those boxes meant. However, the police present some evidence that is supposed to prove that you indeed understood the meaning of your responses. Your responses were not only indistinguishable from those of an expert in bomb-making, the police point out that boxes which represent similar words like “trigger” and “detonator” contain cards with similar numbers, whereas words with very different meanings are represented by boxes with very different number constellations. Also, the police claim that different positions of the flashcards represent various features of the world, and the number on each corresponding flashcard indicates how strongly that feature is present in the respective referent. For example, words with male referents like “Andreas Baader” were represented with boxes that have high numbers on the 12^{th} flashcard, whereas words with female referents like “Ulrike Meinhof” have cards with small numbers at this position. Furthermore, the embedding space spanned by your flashcard box representations has been shown to be structurally similar to aspects of the real world, as it captures relationships and contexts found in actual data. Consequently, the police assert that you possessed a “world model”. Drawing on literature regarding language understanding (Lyre 2024), the police argue that, therefore, you must have understood the meaning of the boxes.

Furthermore, the confiscated MODEL book clearly contains your handwriting, proving that you did not only follow strict rules but that you rather were trained to translate prompts into responses. In trainee phase one, you learned which boxes usually complete a given prompt. The police could even recover

the cards that you used in training during phase two, linking them to known terrorist organizations, and proving that they contained information on how well your responses matched questions regarding certain bomb mechanisms. Following the thoughts in Coelho Mollo and Millière (2023) as well as Lyre (2024), they argue that this link to actual bomb-makers provides at least an indirect knowledge of the workings of actual bombs.

Moreover, the police followed the rules in the MODEL book and realized that you updated the vector representation of a word depending on the words that have preceded it in the prompt, thus incorporating the context. For example, you updated the box representing “trigger” according to the context in which it appeared so that it closely resembled the box representing “detonator” rather than that for “stimulus” or “incitement.” This enables you to handle homographs as well as polysemes. For example, you updated the embedding of the word “trigger” depending on whether it was part of the sentence “Use a radio detonator as the...” or “A loud bang can be an emotional...”. This allegedly proves that you knew you were talking about triggers for bombs rather than emotional triggers, since, according to some experts (Piantadosi & Hill 2022, p. 5), meaning arises from the relationship between words.

Lastly, they even present proof that some of the boxes you received during training did not only represent words but also pictures of bombs and schematic diagrams of detonation circuits. Therefore, words like “detonator”, “explosive”, etc. were in some sense grounded in sensory input. Following the thoughts in Vaydia (2024), the police take this as evidence that you understood these boxes referred to detonators, and explosives, since you learned to associate them with sensory images.

Some of the well-intentioned police officers even believe that you did not fully understand what the prompts and your responses meant. However, they insist that you must at least have some weaker form of (referential) semantic competence, i.e. “the ability to connect words and sentences to objects, events,

and relations in the real world" (Millière and Buckner 2024). In their eyes, this is the only way to explain your remarkable linguistic abilities in the context of bomb-making. Still, you would rightly defend yourself by insisting that you did not understand the prompts and responses, and therefore did not know that you were providing instructions on bomb-making.

Some Objections

In this section I address objections to the applicability of the CRA to LLMs. Critics argue that to determine whether a system truly understands, it is not enough to look at the correct input-output relationship; one must examine the underlying mechanisms. In the case of the Chinese room, this examination reveals a lack of cognitive plausibility, which explains why Searle does not understand Chinese in the original thought experiment. Therefore, the CRA does not prove that computationalism is incorrect; it only shows that the type of computation implemented is not of the right kind. Authors focus on three cognitive features that the processes in the Chinese room lack, but that LLMs are supposed to have in virtue of their different structure.

(1) *Subsymbolism*: Whereas GOFAI systems represent words with single symbols, LLMs use word embeddings, i.e. high-dimensional vectors that represent words. This makes LLMs subsymbolic. Arguments that the CRA does not apply to subsymbolic systems have been put forward by proponents of the connectionist paradigm already shortly after Searle (1980). Prominent examples include Clark (1989), Churchland and Churchland (1990), as well as Harnad (1990). An explicit formulation of this critique can be found in Chalmers (1992): "Because the levels of syntax and semantics are distinct, [subsymbolic] systems are safe from the [Chinese room] argument" (Chalmers 1992, p. 17).

(2) *Machine Learning:* Unlike GOFAI systems that strictly follow a predetermined program, LLMs learn to map prompts to responses by being trained on data. Authors have claimed for some time now

that the CRA becomes obsolete in the context of systems that *learn* their linguistic capacities, rather than following a program (Sharkey and Ziemke 2001, Grindrod 2024).

(3) *Sensory data*: The idea of grounding meaning in sensory input to offer a way around the CRA was first posited by Harnad (1990). Harnad's idea is that, in order for a system to acquire language understanding, the symbols it uses must be linked to the real-world objects they represent through sensory input and adequate abstraction processes. Whereas Harnad himself believes that LLMs are ungrounded in this sense (Harnad 2024), other authors disagree (Lyre 2024). Precisely here lies the difference between multimodal LLMs and GOFAI systems: “One might think that Searle’s argument can easily be extended to LLMs […]. However, depending on how an LLM is trained there is symbol grounding in an LLM because semantic items are grounded in images” (Vaidya 2024, p. 12). Havlík even takes the image recognition capacities of neural nets to be “empirical evidence of the invalidity of Searle’s argument against the possibility of strong artificial intelligence” (Havlík 2023, p. 5).

The flashcard sorter thought experiment undermines these objections. (1) The flashcard sorter is itself a subsymbolic system, in that it operates over cards that only represent words when arranged in boxes. (2) Analogous to LLMs, the flashcard sorter undergoes a phase of self-supervised learning followed by a phase of reinforcement learning from human feedback. (3) The flashcard sorter is trained on multi-modal data as well. Some of the boxes they received during training were not representing words but images, audio, and video data. Yet, even though the flashcard sorter scenario addresses all that was supposedly missing in the original Chinese room, according to those who argue that the CRA does not apply to LLMs, you still do not understand how to make bombs simply because you can respond appropriately to relevant prompts.

Conclusion

In the present paper, I formulated a modified version of the Chinese room thought experiment called “the flashcard sorter”, designed as an analogy to how LLMs compute. The modified argument, from now on called the “flashcard sorter argument”, can be stated as follows:

(1) The flashcard sorter implements the computation of an advanced LLM.

(2) The flashcard sorter does not understand the prompts and the responses.

(C) Therefore, implementing the computation of an advanced LLM is not sufficient for understanding.

In this modified version, the CRA applies to LLMs.

Original replies to the CRA, like the systems, robot, and other minds reply (Searle 1980), were not discussed since they are meant to show that the CRA is not sound, and not that it does not apply to LLMs. The flashcard sorter argument inherits all the problems of the CRA.

The flashcard sorter is an extension of the Chinese room thought experiment, in that both portray a system in which a certain computation is implemented which has been claimed to be sufficient for language understanding. Furthermore, in both scenarios there is strong prima facie plausibility to the claim that the person implementing the computation, either Searle or the flashcard sorter, does not understand the respective input and output, despite appearing as a perfectly competent speaker from an outside perspective. In the thought experiment presented in this paper, these intuitions are further reinforced by the fact that the flashcard sorter is genuinely surprised by the appearance of the police, and opens the door with a clear conscience. Furthermore, the flashcard sorter thought experiment goes beyond the Turing test in that it applies to systems that not only converse indistinguishably from native speakers, but also process language in a cognitively sophisticated manner. It shows that even systems

that transform *embedding vectors* rather than symbols, that *learn* to map inputs to outputs, and that incorporate *sensory training data*, are not immune to the CRA.

References

Chalmers, D. J. (1992). Subsymbolic computation and the Chinese room. In *The Symbolic and Connectionist Paradigms: Closing the Gap* (pp. 25-48).

Churchland, P. M., & Churchland, P. S. (1990). Could a machine think? *Scientific American, 262*(1), 32-39.

Clark, A. (1989). *Microcognition: Philosophy, cognitive science, and parallel distributed processing*. MIT Press.

Grindrod, J. (2024). Large language models and linguistic intentionality. *Synthese, 204*(2), 71.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena, 42*(1-3), 335-346.

Harnad, S. (2024). Language writ large: LLMs, ChatGPT, grounding, meaning, and understanding. *arXiv preprint arXiv:2402.02243*.

Havlík, V. (2023). Meaning and understanding in large language models. *arXiv preprint arXiv:2310.17407*.

Lyre, H. (2024). Understanding AI: Semantic grounding in large language models. *arXiv preprint arXiv:2402.10992*.

Millière, R., & Buckner, C. (2024). A philosophical introduction to language models—Part I: Continuity with classic debates. *arXiv preprint arXiv:2401.03910*.

Mollo, D. C., & Millière, R. (2023). The vector grounding problem. *arXiv preprint arXiv:2304.01481*.

Piantadosi, S. T., & Hill, F. (2022). Meaning without reference in large language models. *arXiv preprint arXiv:2208.02957*.

Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(3), 417-424.

Vaidya, A. J. (2024). Can machines have emotions? *AI and Society, 1-16*.