Experimental Proposal - Investigating Robustness of In-Context Connections in the Lexicon

Alan Chen

CLPS0800

May 2024

Experimental Proposal - Investigating Robustness of In-Context Connections in the Lexicon

Introduction

State-of-the-art large language models (LLMs) have recently developed a surprising behavior called "in-context learning" (ICL) (Dong et al., 2022). ICL is a phenomenon in which, from just a few examples, a LLM can learn how to perform unfamiliar tasks *without* updating its weights. The primary force underlying ICL is an emergent-with-scale module called the induction head (Elhage et al., 2021). Induction heads activate when they are prompted with the pattern of tokens "$A, B, \ldots, A$" and asked to predict the next token. They then recall the previous instance of A and copy over exactly what came after it (B), modeling primitive inductive reasoning.

Incredibly, induction heads develop suddenly through a training process solely based on predicting a probability distribution over *the next token*. This phenomenon urges one to consider whether computation can approximate the development of human language capabilities, as LLMs have acquired reasoning and recall capabilities surpassing experts in some domains through this simple training process done at scale (Achiam et al., 2023). There has been a rise of work in interpreting LLM reasoning and drawing connections with human cognition, such as reverse engineering neural networks and researching the connection between the temporal memory recall and induction heads (Räuker, Ho, Casper, & Hadfield-Menell, 2023; Binz & Schulz, 2023; Li, 2023).

One unique capability of humans is the ability to engage in context-based language interactions. Questions involving this capability include: what forces underlie human contextual recall? Are they linguistic? What linguistic factors and systems determine how long humans preserve previous contexts in their heads? How do these answers differ from current LLMs and could they be computationally modeled?

The answers to the above questions could relate to the formation of context-specific connections in the lexicon. Humans are believed to develop semantic and phonological connections between words over time that are activated bidirectionally upon stimuli (Meyer & Schvaneveldt, 1971), analogous to the weights in a language model. However, context-specific

connections between words independent of semantic or phonological relationships could also form. In contrast to being limited to connections between adjacent tokens like an induction head, humans can build complex mental models depicting the relations between subjects in a context, as discovered through memory tasks (MacDonald & Just, 1989; Glenberg, Meyer, & Lindem, 1987). However, less is known about the deeper properties of contextual connections. For example, are they stored in the lexicon and/or are they sensitive to linguistic features? Semantic and phonological connections are relatively permanent - are contextual connections permanent or temporary?

In this report, an experiment based on a series of lexical decision tasks is proposed to analyze the location and *robustness* of contextual connections. Two forms of robustness are discussed - the first relates to memory and the second has implications for interactions between memory systems and linguistic processing.

1. **Robustness against context shifts** - At what strength do contextual connections persist across new contexts? Do they persist at all? Competing hypotheses would pose that the strengths of contextual connections either decay or persist over progressive new contexts. Based on the results of (Glenberg et al., 1987), one might expect the former to be true, as the study in that paper observes the priming effect in a memory task decaying even with just the addition of one extra filler sentence between the target relationship and the memory probe.

2. **Robustness against linguistic complexity** - How is the strength of a contextual connection related to the complexity of the corresponding relationship between the words? How does this effect depend on the type of complexity considered? Which linguistic systems are involved? The alternative hypotheses are that the strength of contextual connections could decrease, remain the same, or increase when the complexity of the relationships increases. Different forms of linguistic complexity could cause the same or differing effects in the connection strength. Furthermore, it's possible that the strength of contextual connections involves multiple linguistic systems interacting. Alternatively, it could be dictated purely by one system's ability to do its job correctly on the input stimuli.

The proposed experiments are based on two similar context-augmented lexical decision tasks that control and vary important factors in a base task. At the start of the base task, participants are aurally exposed to a context window describing a scenario with multiple relations between objects. After this exposure, participants will engage in a lexical decision task, similar to previous studies, where they must identify whether words that appear in sequence are real words. Priming effects between words, as quantified by a speedup in recognition time, will be measured and compared.

First, the base task will identify whether context windows can alter the lexicon in ways not related to semantics or phonology. If speedup effects are observed in the base task, it would imply that the context window is augmenting the lexical connections, whereas the opposite is true if no priming effects are observed.

The robustness experiments assume that the context windows alter the lexicon and instead study the impact magnitude based on two variables. Briefly, to address *robustness against context shifts*, the variation of the base task includes multiple rounds of the base task with different context windows. The decision task of a specified later round will contain the same target priming pair as the first round, identifying the permanence of the contextual connections. The response speedup from the first time a priming pair appears will be compared with the second time it appears across a varying number of intermediate contexts to evaluate the decay in strength. To address *robustness against complexity*, two forms of linguistic complexity are studied: phonological length and syntactic complexity. Both are computationally tractable and have impacts on initiation times and planning (Ferreira, 1991). The complexities of relations will be varied according to these metrics and the effects on the priming strength will be measured.

Alternative hypotheses can be ruled out by observing the impact of context shifts and varying complexity on the response speedup. A decay or lack thereof in response speedup over additional context shifts implies the decay or lack thereof, respectively, in the strength of contextual connections. Likewise, a decay, lack thereof, or increase in response speedup over increasing phonological length or syntactic complexity (isolated from each other) distinguishes

between the corresponding hypotheses about the robustness to linguistic complexity.

The experiment also proposes answers to additional questions and offers insights into LLM architecture design and limitations. For the first variation, whether the response speedup completely disappears or persists at a low level is intriguing to determine. Empirically, humans are able to remember important bits of previous contexts and adjust quicker than in their initial exposure, whereas LLMs can only adjust quicker if the original context is still within the same fixed-size (measured by number of tokens) context window. If the priming effect completely disappears, this context-remembering capability is unrelated to dormant contextual connections, whereas the opposite is true if the priming effect sticks around. If contextual connections are used by humans to remember previous contexts, LLM researchers could consider developing a context window management system based on preserving the right contextual connections rather than naively storing raw tokens.

For the second variation, the individual speedup effects of phonological length and syntactic complexity can be disentangled, presenting a fine-grained analysis of the involved linguistic modules and robustness of contextual connections to different forms of complexity. From the computational view, current LLMs are particularly sensitive to the *length* complexity of connections, as induction heads only activate when the tokens are adjacent whereas their performance is invariant in the syntax and semantics of the tokens. If humans are less sensitive to increases in length complexity and syntactic complexity in forming strong connections, LLM's sensitivity to length complexity could pose as a limiting factor on their long-context capabilities.

The respective methods sections for the variants discuss more specific logic and implications.

Methods

The base task will consist of a context window with R relationships between pairs of subjects represented by $\{(X_i, Y_i)\}_{i=1}^{R}$. Each relationship will be presented aurally in the form "X_i Z_i Y_i" where Z_i describes the relationship between X_i and Y_i. All relationships will be active i.e. ones where X_i acts on Y_i as opposed to Y_i passively acting on X_i to remove an additional

confounding factor.

The lexical decision task consists of a sequence of words and non-words, where the participant will identify whether the word is real by responding "yes" or "no." Specific words will have their response times measured depending on which group the participant is in. Participants are either in the treatment group, which will see X_i directly followed by Y_i during the task and have their response time for Y_i measured, or the control group, who will only see Y_i at some point during the task and have their response time measured for Y_i as a baseline. If there is no significant difference between the two groups, no priming effect occurs and contextual connections are likely not lexical. If the response time is faster for Y_i in the treatment group, there is evidence that seeing X_i also recalls Y_i. Crucially, the *amount* of speedup quantifies the strength of the connection.

Confounding Factors

Uncontrolled confounding factors can elicit false positive results in lexical decision tasks. In particular, frequency, length, age of acquisition, and semantic category of the subject words are all confounding variables that could affect the response times (Rayner, 1998; Juhasz, 2005). Furthermore, when investigating contextual relationships, the priming effects must be disentangled from semantic and phonological priming effects. One way to remove the influence of these factors is by using artificial words for X_i and Y_i. However, these "Jabberwocky" sentences with artificial subject words do not exhibit the same strength of priming effects as real or plausible sentences - this is likely because they are not in the lexicon prior to the experiment (Fedorenko et al., 2016).

Instead, a set of nouns will be carefully selected from a database constructed by (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012) with similar age-of-acquisition, frequency, and length but minimal semantic and phonological relationships to serve as the main subject words used in the contexts. Contexts will be kept to around $R \approx$ four relationships or eight total subjects to limit the memory burden within each round of the task (Cowan, 2010).

Another nontrivial confounding factor is the presentation order of the subject relationships as seen from (Glenberg et al., 1987), as earlier relationships might experience a decay in strength

only related to their position. Priming times will thus only be compared across pairs that appeared in the same position in the context (for example, all pairs with $i = 3$).

Robustness Against Context Shift

In this variant, the only difference from the base task is the number of times each participant performs the task. Each participant (in both groups) will repeat the entire task with a new context window N times (with N relatively small). The *first recurrence* of a subsequence in the decision task is defined as the first time the subsequence reappears in a later decision task. The recurrent subsequence pattern will either be X_i, Y_i (for the treatment group) or just Y_i (for the control group). The first recurrence time, as measured by "rounds between the first appearance and the first recurrence," will be varied and the response time to Y_i will be measured in all cases. The measured effects determine the permanence of contextual connections in the lexicon. As the first recurrence length increases, the response speedup is expected to return to approximately the baseline level, implying that the brain is demoting the importance of the original connection because more information is coming in yet the connection remains unused. If true, the evidence would support but not confirm the hypothesis that humans have an efficient lexical memory management system that knows when to overwrite old contextual information with presumably more relevant contextual information.

In addition to confirming the decay of strength over new contexts, this experiment will also answer if the priming effect completely disappears or persists at a weaker strength. If the priming effect does not completely disappear, it explains why language users can assimilate easier when engaging with a context for the second time. In particular, this would differ from LLMs, which have fixed context windows that, when exceeded in size, cause the LLM to forget about the overflown tokens. Without a more intelligent context window management system, LLMs may require significantly more compute power than humans to achieve the same contextual recall performance, which could serve as a fundamental limitation with long contexts and raw tokens filling the context window.

Robustness Against Complexity

How does the brain know which connections to prioritize and strengthen? It is likely more sophisticated than a simple context window or recency bias. One such factor it could use is the linguistic complexity of the relationship. To this end, this experiment proposes two forms of absolute complexity - phonological length and syntactic complexity - and studies their relationship to the strength of contextual connections. Absolute complexity is preferred over relative complexity, as (1) a user-independent abstraction is desired to make broader conclusions about the nature of contextual connections and (2) relative complexity is difficult and expensive to quantify in an experimental setting (Miestamo, 2004).

Phonological length (PL) is defined as the number of phonological units. Syntactic complexity (SC) is defined as the number of nodes in the syntactic tree of a relationship (Ferreira, 1991). Importantly, two phrases can have varying SCs but have identical PLs and vice versa. As a simple example, "napkins eat dinosaurs" would be considered a less phonologically and syntactically complex relationship than "napkins paint and drink wine with dinosaurs" whereas "the napkin that was above sank into the dinosaur" and "the napkin captivates the attention of the dinosaur" have similar phonological length but the former is more syntactically complex.

The details of the experiment are similar to the context shift experiment, except that instead of varying the first recurrence time, different participants are given varying Z_i complexities. The response times on Y_i are again measured, giving numerical effect measurements. Firstly, the impact of Z_i's PL on Y_i is measured while holding Z_i's SC constant. This experiment extracts just the phonological length effect, denoted by PLE, independent of varying syntactic complexity. Secondly, Z_i's PL will be varied while holding its PL constant. Similarly, this subexperiment isolates the syntactic complexity effect, denoted by SCE, independent of the PLE.

Both subexperiments will distinguish (for the corresponding complexity) whether more linguistically complex relationships have weaker or stronger lexical priming effects. If simpler relationships have stronger priming effects, a reasonable conclusion would be that simpler

relationships form clearer mental models, thus strengthening the embedding in the lexicon of the connection. Alternatively, if more complex relationships had stronger priming effects, the evidence would support that storing more complex linguistic relationships requires a higher mental real estate investment, in turn strengthening the connection in the lexicon between the two subjects.

If SCE is small but PLE is significant, the phonological system is likely the key factor in determining the strength of connections. In the opposite direction, if PLE is small and SCE is significant, it would mean that the contextual connections depend more closely on the parser and corresponding syntactic processing network. When both are small, we can conclude that neither phonological complexity nor syntactic complexity are tightly *individually* related to the strength of contextual connections (though together, they could still have an impact).

Because induction heads in LLMs operate independently of the semantics or syntax of the tokens themselves, they are more sensitive to phonological distance rather than the syntactic structure of the relationship. If SCE is significant, it would exemplify one form of information (syntactic structure) used by humans to represent the context that LLMs do not. Even further, PLE is also small, it would mean that LLMs lack human robustness to length-based complexity, which could be a roadblock with longer and more complex contexts. In that case, developing true mental models of the state space instead of just copying adjacent tokens is vital to human-style reasoning and contextual recall.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... others (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120.

Cowan, N. (2010, February). The magical mystery four: How is working memory capacity limited, and why? *Curr Dir Psychol Sci*, *19*(1), 51–57.

Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., ... Sui, Z. (2022). A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., ... Olah, C. (2021). A mathematical framework for transformer circuits. *Transformer Circuits Thread*. (https://transformer-circuits.pub/2021/framework/index.html)

Fedorenko, E., Scott, T. L., Brunner, P., Coon, W. G., Pritchett, B., Schalk, G., & Kanwisher, N. (2016). Neural correlate of the construction of sentence meaning. *Proceedings of the National Academy of Sciences*, *113*(41), E6256–E6262.

Ferreira, F. (1991). Effects of length and syntactic complexity on initiation times for prepared utterances. *Journal of Memory and Language*, *30*(2), 210-233. Retrieved from `https://www.sciencedirect.com/science/article/pii/0749596X91900044` doi: https://doi.org/10.1016/0749-596X(91)90004-4

Glenberg, A. M., Meyer, M., & Lindem, K. (1987). Mental models contribute to foregrounding during text comprehension. *Journal of Memory and Language*, *26*(1), 69–83.

Juhasz, B. J. (2005, September). Age-of-acquisition effects in word and picture identification. *Psychol Bull*, *131*(5), 684–712.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, *44*, 978–990.

Li, J.-A. (2023, July). *Relating induction heads in transformers to temporal context model in human free recall.* https://apartresearch.com. (Research submission to the Interpretability

research sprint hosted by Apart.)

MacDonald, M. C., & Just, M. A. (1989). Changes in activation levels with negation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(4), 633.

Meyer, D., & Schvaneveldt, R. (1971, 10). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*, 227-34. doi: 10.1037/h0031564

Miestamo, M. (2004). On the feasibility of complexity metrics. In *FinEst Linguistics, Proceedings of the Annual Finnish and Estonian Conference of Linguistics, Tallinn* (pp. 11–26).

Räuker, T., Ho, A., Casper, S., & Hadfield-Menell, D. (2023). Toward transparent ai: A survey on interpreting the inner structures of deep neural networks. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)* (pp. 464–483).

Rayner, K. (1998, November). Eye movements in reading and information processing: 20 years of research. *Psychol Bull*, *124*(3), 372–422.