COSTS AND BENEFITS OF INTEGRATING INFORMATION SEQUENCES

by
Touran (Shima) Rahimi Moghaddam

A dissertation submitted to Johns Hopkins University
in conformity with the requirements
for the degree of Doctor of Philosophy

Baltimore, Maryland
January 2023

Abstract:

Information from the world unfolds over time, and to navigate the everyday world and make future predictions, our brain needs to integrate information over time. For instance, when having a conversation with someone, our brain needs to accumulate information about words and sentences to comprehend the ongoing discussion and respond appropriately. However, ubiquitous accumulation of information can cause interference, especially if we end up combining unrelated information. For instance, the topic of conversation may change from one sentence to the next, in which case combining information from consecutive sentences could cause interference and confusion.

These examples demonstrate that integrating information over time is sometimes necessary for successful comprehension and prediction, but it should not be performed indiscriminately. How then should temporal integration mechanisms be implemented, especially in constrained brain-like learning architectures? What kinds of temporal integration and separation mechanisms are employed by contemporary machine learning models? And how do these integration and separation processes compare against what we observe in human behavior?

In this thesis, we examined the costs and benefits of integrating and separating information sequences in humans and machines. In the first two projects we focused on *learning* and tested the performance of biologically-plausible temporal integration mechanisms in neural networks; we characterized the efficacy of these systems in learning categories from a sequence of examples, and investigated how their internal representations are altered by how they integrate information over time. In two further projects we focused on *online comprehension and prediction*, in the setting of humans reading natural language sequences, and we contrasted our findings with neural network models that predict and generate natural language sequences. We tested how online comprehension and subsequent memory

are affected by interruptions in the text that humans are reading. Finally, we tested how neural language models respond to the insertion of incongruent information into a broader coherent text, and we compared these findings against our observations of how humans handle interruptions while reading.

Altogether, these studies identify mechanisms by which humans and machines can exploit temporal continuity in the environment, in the service of learning about, understanding and predicting our dynamic world.

Doctoral Committee:

Dr. Christopher J. Honey (Advisor and Primary Reader)

Dr. Susan Courtney (Secondary Reader)

Dr. Chaz Firestone

Dr. Brenda Rapp

Dr. Rhodri Cusack

Acknowledgements

"Life is a unique scene of art

Everyone sings their song and depart

The scene is eternal

Blessed are the songs that will be remembered"

- Jaleh Esfahani (Persian Poet)

I would like to express my deepest gratitude to my dear Ph.D. advisor, Dr. Christopher Honey, for his unwavering support throughout my Ph.D. I joined Chris's lab as an engineer, quick to ask the "how" questions and wanting to find solutions. Chris taught me to think like a scientist, to think deeply about problems, to ask the "why" questions, and to question the potential solutions. His generous help extended beyond the scope of my thesis, as he was willing to help me pursue my career goals regardless of whether they were in line with his own interests. I am also grateful for all of our scientific and brainstorming discussions along with his counsel and encouragement during the challenging times of my Ph.D. journey. His dedication and commitment to my success have been instrumental in the completion of my thesis.

I am sincerely grateful to my thesis committee members, Dr. Susan Courtney, Dr. Chaz Firestone, Dr. Rhodri Cusack, and Dr. Brenda Rapp for kindly helping me to complete my dissertation. Dr. Courtney's invaluable advice and guidance has been especially helpful for the behavioral study in Chapter 4. Her insight and knowledge helped in improving the design of the study and sharpening our research questions. Dr. Firestone's valuable insight strengthened our study in Chapter 5, in particular, in contrasting the performance of humans and neural language models. I am

thankful to all my thesis committee for their patience, expertise, and support in completing my dissertations.

I am thankful to my lab-mates from the Honey lab and Chen lab, who have become my friends over the years. Their feedback and input have been invaluable to the quality of my research, and our conversations have made my PhD journey a much more enjoyable one. I am grateful to Dr. Janice Chen for her kind support during my PhD journey, especially in helping me achieve my career goals. Additionally, I could not have completed my thesis without the hard work of my research assistants, Frank, who helped me with the study in Chapter 2, and Jinhan, who was my assistant in the study in Chapter 5. I am also thankful to Buddhika, a former Postdoc in the lab, whose contribution to the early development of the behavioral study in Chapter 4 was essential. Finally, I would like to express my appreciation to Xian and Savannah, who were both very helpful in structuring the research paradigm in Chapter 4.

And, finally, I want to thank my family and friends. I am immensely thankful to my parents, Maman and Baba, for their unconditional love and support, and all the sacrifices they have made in their lives that enabled me to be where I am today. Although, due to unfortunate circumstances, I have been unable to see them for a few years, our phone conversations have always left me feeling inspired. Talking to my mom has brightened my days and reminded me of the beauty in life, and talking to my dad has warmed my heart and empowered me with the confidence that no issue is insurmountable. I would also want to thank my lovely brother and sister, Farzad and Forough, who have always encouraged me. Their kind words have given me so much hope. Finally, I want to thank my friends who made Baltimore feel like home and added color to my life during the pursuit of the Ph.D.

Dedication

I dedicate this dissertation to my love, my best friend, my life partner, my wonderful husband, Ehsan.

For all those times you stood by me,

For all the joy you brought to my life,

For every dream you made come true,

For all the love I found in you,

I'll be forever thankful,

My world is a better place because of you.

- From a song by songwriter Diane Warren

Table of Contents

List of Figures

Chapter 1: Introduction

1.1. Costs and Benefits of Integrating Information over Time

Imagine you recently landed a new job, and on your first day in the office, you join a group of new colleagues for an introductory lunch. In this situation, your ability to learn from this experience and to comprehend it in real time will naturally benefit from being able to accumulate information over time.

First, as new colleagues introduce themselves to you at the table, you must learn their names and associate them with the faces you see around you at the table. However, the faces that you see are not distributed randomly over time: instead, you will see multiple samples of one individual's face, as they move their faces and heads while introducing themselves, before moving to the next person (**Figure 1-1**). While learning the facial features from a single individual, it could be beneficial to integrate and combine the multiple samples you receive of their facial features. However, you should not indiscriminately seek to combine any visual samples that are nearby in time as part of the same category: when moving from one colleague to the next, you will see a new face with its own unique features. So, it is beneficial to employ some form temporal integration of information, but you should also temporally separate information when moving from one person to the next. Thus, for the purposes of learning, it seems we need a flexible combination of mechanisms for integrating and separating consecutive samples from the environment (**Figure 1-1**).

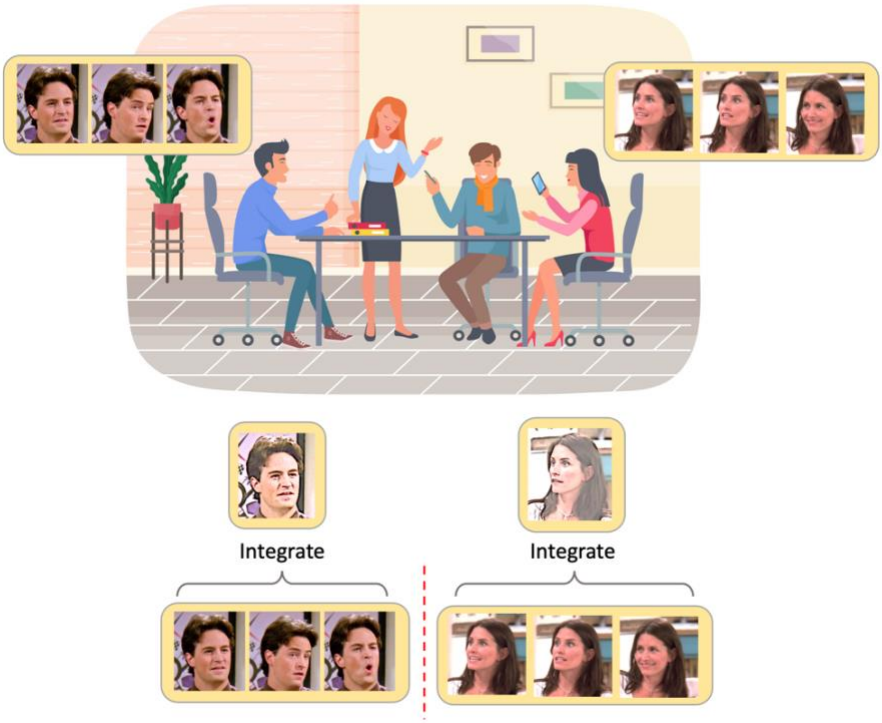
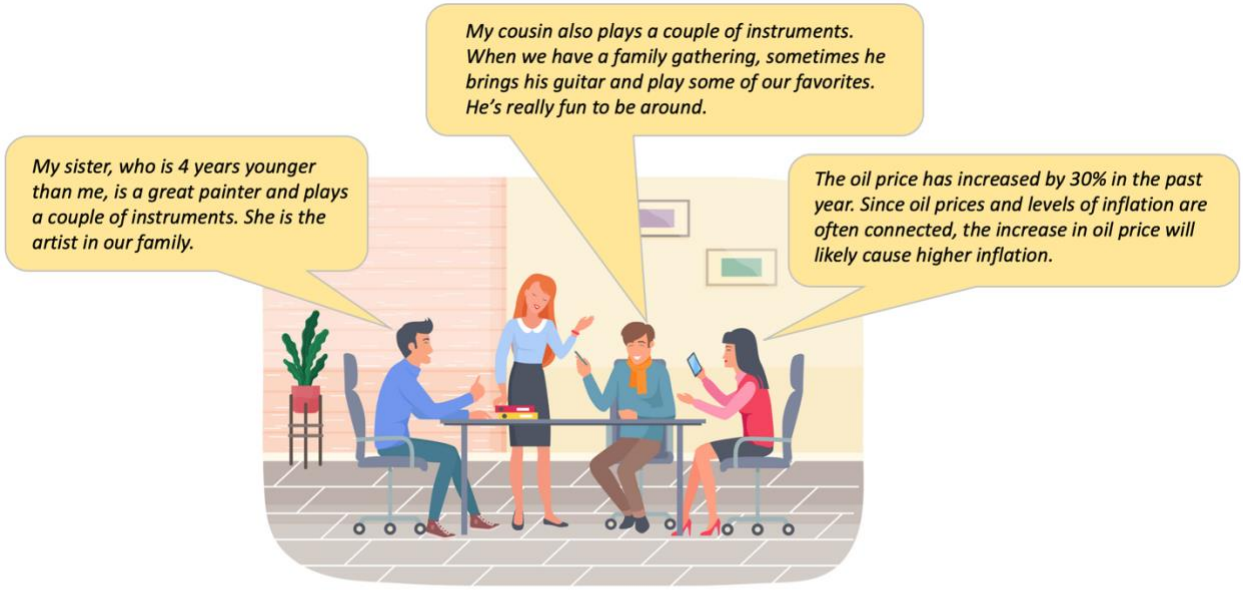


Figure 1-1. Temporally correlated data in the real world accompanied by sudden changes.
Example of a situation where we meet new people and learn about their names and associate them with their faces that we see. The examples that we see from each person's face are usually temporally correlated as we see multiple examples of on person's face consecutively. But this autocorrelation can be suddenly disrupted when moving from one person to the next.

Of course, temporal integration does not only affect how we learn, but also how we comprehend and predict the environment online. For example, during the lunch conversation, you may encounter a situation that requires you to handle a *transient* change in the information stream. For instance, the conversation at the table may concern the musical abilities of each of the colleagues and their families. As the conversation proceeds, you gradually integrate sequences of information about the musical abilities of your colleagues and their families, and you load mental schemas about musical education and music genres. But then, a colleague may notice a political news headline on their smartphone and interrupt the conversation to read the political headline aloud to the group. Less than a minute later, the conversation returns to the topic of families and music, but

now you require cognitive processes that can identify and store the incongruent information (the news headline) and help you to resume integrating information on the broader theme (of musical families), rather than starting the integration process all over again (**Figure 1-2**). Here again, we can benefit if we are able to maintain and integrate information over time, but we also require mechanisms that enable us to separately handle incongruous information, before resuming the integration process.



My sister, who is 4 years younger than me, is a great painter and plays a couple of instruments. She is the artist in our family. My cousin also plays a couple of instruments. The oil price has increased by 30% in the past year. Since oil prices and levels of inflation are often connected, the increase in oil price will likely cause higher inflation. When we have a family gathering, sometimes he brings his guitar and play some of our favorites. He's really fun to be around.

Figure 1-2. Temporally correlated linguistic data in the real world accompanied by transient changes. Example of a situation where we suddenly hear an interrupting piece of information during a conversation. The interrupting information (news about oil prices) can be thought of as a transient change in an otherwise temporally correlated data (conversation about family members).

The scenarios described above motivate a series of research questions: What are the consequences of sudden changes in the information stream for our learning and comprehension? How might biologically plausible mechanisms be implemented to enable us to learn from temporally

correlated input in the real world? How do humans (and how do contemporary machine learning models) deal with sudden changes in their information stream, and then resume the broader integration process? To address these questions, this thesis examines the costs and benefits of integration and separation of temporal information for learning and comprehension in humans and machines.

The first two projects in this thesis (Chapter 2 and Chapter 3) focus on temporal integration for the purposes of *learning*. We designed and tested biologically plausible integration mechanisms for neural networks performing a learning task, and we compared their efficacy for learning incrementally from a sequence of examples, as well as how they alter the internal representations that are learned.

The next two projects in this thesis (Chapter 4 and Chapter 5) focus on temporal integration for the purpose of *online comprehension, prediction and memory*. We quantified human behavior and the output of contemporary machine learning models designed to predict and generate language sequences. Specifically, we focused on the settings of incongruencies and interruptions, and how humans and machine learning models are affected by them. We also compared the integration and separation processes observed in machine language processing against the online comprehension and memory processes in humans, under the same conditions.

In the remainder of this introductory Chapter, I summarize prior works, provide an overview of the current state of knowledge in each topic, and identify the gaps in the literature.

1.2. Learning Incrementally from Temporally Structured Data

Imagine that all the events of your life are captured, second-by-second, and then replayed to you. Except they are not played in their original order: the seconds of life are scrambled, so that a moment

of reaching down to pick up your keys at age 30 is followed by an instant of you opening your mouth to cry at age 3. Much of the classical theoretical work in learning treats the learning problem in this way, as if the events in the world are a homogenous pool from which we can sample randomly. Of course, we do not learn from and experience the world in that way, and this a key reason why temporal integration is so important.

The world is structured temporally in many ways, but one of the most basic is that it is correlated in time: one moment tends, on average, to resemble the other moments nearby in time. In this section, I review prior research on learning from temporally correlated data for (i) category learning and (ii) representation learning.

1.2.1. Effects of Temporally Smooth Data on Category Learning

In the biological world, category learning usually takes place in time, as the learner is exposed to exemplars presented one after the other. For instance, when meeting a group of new people, we learn and remember their faces by seeing multiple samples of each face from different angles. But crucially, such real-world information is usually correlated across nearby points in time. So, for instance, we would usually see multiple angles (examples) of one face, consecutively (**Figure 1-1**). How does this temporal correlation across the exemplars influence category learning? How can a system exploit this temporal correlation for higher efficiency when learning categories?

1.2.1.1. Effects of Sampling Strategies on Incremental Learning

The ordering of training examples affects the speed and quality of learning in artificial neural networks. For example, learning can be sped by presenting "easier" examples earlier, and then gradually increasing difficulty (Bengio et al. 2009; Kumar, Packer, and Koller 2010; Elman 1993). Similarly,

learning can be more efficient if training data is organized so that the magnitude of weight updates increases over training samples (Gao and Jojic 2017).

However, the natural world often presents training data in temporally autocorrelated manner rather than ordering training items based on item difficulty or proximity to category boundaries. We, therefore, here are interested in exploring the effects of ordering similar training items nearby in time. With the data ordered in this way, we then seek to identify mechanisms that can accelerate learning across multiple levels of temporal autocorrelation, adapting to what is present in the data. This ability to adapt to the properties of the data is important in real-world settings, where a learner may lack control over the training order, or prior knowledge of item difficulty is unavailable.

1.2.1.2. Potential Costs and Benefits of Training with Smooth Data

Temporally correlated data may slow learning in feedforward neural networks. If consecutive items are similar, then the gradients induced by them will be related, especially early in training. If we consider the average of the gradients induced by the whole training set as the "ideal" gradient, then subsets of similar samples provide a higher variance (i.e., noisier) estimate of this ideal.

Smoothness in data may also slow learning due to *catastrophic forgetting* (Robert M. French 1999). Suppose that, for smoother training, we sample multiple times from a category before moving to another category. This means that the next presentation of each category will be, on average, farther apart from its previous appearance. This increased distance could lead to greater forgetting for that category, thus slowing learning overall.

On the other hand, smoother training data might also benefit learning. For example, there may be some category-diagnostic features that will not reliably be extracted by a learning algorithm unless

multiple weight updates occur for that feature nearby in time; smoother training data would be more liable to present such features nearby in time.

1.2.1.3. Biological Constraints as Inductive Biases for Category Learning

Neural networks are thought to provide a model of how biological brains perform categorization, which involves learning to classify objects into different categories (Grill-Spector and Weiner 2013; Yamins and DiCarlo 2016; Spoerer, McClure, and Kriegeskorte 2017; Nayebi et al. 2018). However, biological brains have a peculiar combination of properties. On the one hand, neural circuits are replete with anatomical recurrence and exhibit substantial autocorrelation in their dynamics (Honey et al. 2012; S. M. Smith et al. 2013; Kietzmann et al. 2019; Bright et al. 2020; Raut, Snyder, and Raichle 2020; Cocchi et al. 2017). On the other hand, it is not thought feasible for biological brains to implement backpropagation-through-time (BPTT), which is generally necessary for efficiently training artificial neural networks with recurrent connections (Lillicrap and Santoro 2019). How might recurrent dynamics in the brains achieve learning without propagating gradients through time? Particularly, based on the observation that real-world data is also often temporally autocorrelated, in what ways might brains exploit the temporal structure of their input data without propagating gradients through time?

There is ubiquitous correlation across cortical dynamics which can influence item-by-item incremental learning. Cortical dynamics exhibit autocorrelation on the scale of milliseconds to seconds, so that correlation in consecutive internal states is unavoidable (Honey et al. 2012; Murray et al. 2014; Bright et al. 2020; U. Hasson et al. 2008). Primate brains exhibit both local and long-range recurrence in their architectures, and their dynamics are autocorrelated over many seconds (Honey et

al. 2012; S. M. Smith et al. 2013; Chaudhuri et al. 2015; Bright et al. 2020; Murray et al. 2014; Raut, Snyder, and Raichle 2020).

However, cortical states are not always correlated over time: neural circuits can identify "event-boundaries" in the data sequence and can sharply shift their state accordingly (DuBrow and Davachi 2016; Baldassano et al. 2018). This shift appears to be associated with "resetting" of context representations, which has been understood as a form of memory gating (Chien and Honey 2020) so that neural circuits can combine information from related training samples, and avoid interference from unrelated samples.

In Chapter 2, we will explore the possibility that these two neural properties – autocorrelation and context-resetting – serve as an inductive bias for cortical learning. First, if the training data are temporally autocorrelated, then it may be advantageous to bias internal dynamics to be autocorrelated as well, as this effectively acts as an accurate prior over properties of the environment. For instance, data sampled from a slowly- changing environment may contain important features that are stable over time, which can be better extracted by mixing current input with a memory of recent input (e.g., mixing different exemplars of someone's face). Second, by "resetting" local memory states at boundaries between events, brains can reduce interference between irrelevant prior information and current input (e.g., separating exemplars of one person's face from another) (**Figure 1-3**).

Figure 1-3. Examples of temporally correlated training data in the real world.
Information that we see nearby in time are usually similar (correlated), for instance, we see multiple examples from one person's face during a conversation. This autocorrelation is also accompanied with abrupt changes in the information stream such as when we move from one person to another and see a new face with its own unique features.

1.2.2. Learning Representations from Multiscale Data Sequences

In the real world, we may need to learn from data that vary on both slow and fast timescales. For instance, when having a conversation with a new person, we see many examples of the same face from different angles nearby in time, and their facial features vary at different timescales: the features around a person's mouth change quickly, while their face's outline changes more slowly (**Figure 1-4**). When trained using training data that possesses multiple scales of structure, how can a learning system generate internal representations that reflect the multi-timescale properties of the data?

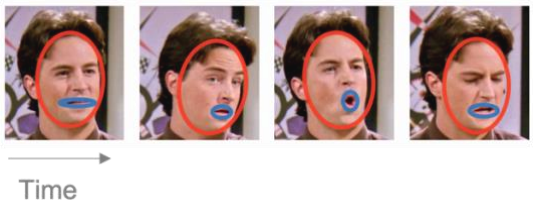


Figure 1-4. Example of real-world data with multiple timescales.
Facial features vary at different timescales: the features around a person's mouth (blue) change quickly, while their face's outline (red) changes more slowly.

1.2.2.1. Learning from Multi-Timescale Data in Artificial and Biological Neural Networks

Artificial recurrent neural networks (RNNs) are powerful models for learning from multi-timescale sequences, and their internal representations reflect temporal properties of their training data (Strobelt et al. 2018; Ming et al. 2018). Evidence of recurrence in cortical circuits has motivated RNN models of sensory and cognitive processes in cortical circuits (Honey et al. 2012; S. M. Smith et al. 2013; Murray et al. 2014; Bright et al. 2020; Raut, Snyder, and Raichle 2020; Kozachkov et al. 2022). However, for learning from temporal information, RNNs employ backpropagation through time (BPTT), which is thought to be implausible for biological systems (Lillicrap and Santoro 2019). How then can biological systems learn from multi-timescale sequences similar to RNNs, without propagating gradient in time?

Evidence from hierarchical processing in cortical circuits shows that neural circuits closer to the sensory cortex shift their state more rapidly than circuits further away from sensory areas (Murray et al. 2014; Chaudhuri et al. 2015; Chien and Honey 2020). Roughly speaking, this means that a sensory region of the neocortex will exhibit relatively small correlation in its state between time t and $t+1$, while a higher order association area will exhibit a relatively large correlation in state between time t and $t+1$. Similarly, sudden changes in state will happen more frequently in sensory areas, and less frequently in higher order association cortices. Assuming that sudden shifts in cortical state reflect the "resetting" of a context representation stored in each brain region, then shifting cortical states at multiple timescales would mean that the context representations are reset at different times.

In Chapter 3, we will explore the hypothesis that multiscale resetting enables cortical circuits to take advantage of multi-timescale data to learn internal representations that reflect temporal

properties of the information. However, the functional efficacy of the multiscale resetting mechanism in learning multiscale representations from the data is still unexplored.

1.3. Effects of Transient Incongruencies on Language Processing in Humans and Neural Language Models

In this section, I review prior works related to how language processing in humans and neural language models is affected by transient incongruencies (interruptions).

1.3.1. Effects of Interruptions on Human Language Comprehension and Memory

Interruptions are commonplace in real-world language processing. Recall the scenario we sketched above, in which a new employee is having lunch with colleagues, and then their conversation about musical families is interrupted by a news item (**Figure 1-2**). After experiencing the interruption (hearing the news), the conversation may then pick up where it left off (discussion about musical families), but we may have difficulty reinstating the original processing of the topic that was in our mind prior to the interruption. Indeed, in some cases, we may never be able to re-instate the same mental processes that preceded an interruption. Therefore, it is practically important to determine which kinds of interruptions are most disruptive to language and discourse comprehension. Moreover, if we can understand which interruptions hinder mental context the most, we may gain better insight into how and what we are holding in mind as we try to navigate interruptions in everyday situations.

1.3.1.1. What Makes an Interruption Disruptive?

The disruptiveness of an interruption is likely to vary with its similarity to the primary task. In this respect, we can distinguish between interruptions that have "cognitive similarity" and interruptions that have "content similarity". For example, if your primary task is reading a narrative, and the interruption also involves reading, we describe these tasks as "cognitively similar", regardless of the content that is being read in each case. On the other hand, if the primary task is reading a story, and the interrupting task involves performing arithmetic or visuospatial reasoning, then these tasks are cognitively dissimilar, because they likely involve different kinds of operations over different kinds of mental representations.

Early investigations of interruption in reading argued that interruptions were harmful if they erased a short-term memory trace of the verbatim content of the primary reading text. Glanzer et al. (Glanzer, Dorfman, and Kaplan 1981; Glanzer, Fischer, and Dorfman 1984) introduced a reading interruption paradigm in which participants would read sentences, self-paced, but would occasionally be interrupted by other tasks, such as arithmetic or reading unrelated materials. They first found that reading was slowed when interleaved with interrupting arithmetic problems (Glanzer, Dorfman, and Kaplan 1981) and then showed that reading of the primary text was slowed when unrelated reading materials were interleaved (Glanzer, Fischer, and Dorfman 1984). They argued, that because the deleterious effects of interruption were present both with task-switching (reading interleaved with arithmetic) and without task-switching (reading interleaved with reading), the primary driver of interruption cannot be task-switching. They argued that the reading slow-downs occurred because the interrupting task (regardless of type) erased or interfered with verbatim textual information stored in short-term memory.

1.3.1.2. Role of Content-Similarity on Disruptiveness of Interruptions

Ledoux and Gordon (Ledoux and Gordon 2006) returned to the question of the type of interruptions by comparing the effects of stylistically similar and dissimilar interruptions. They proposed that more similar interruptions were more likely to cause interference with the memory representations associated with the primary task, leading to slower reading. They tested this hypothesis in a self-paced reading task, in which the primary text and the interrupting text could be of the same or different style, narrative or expository. Narrative texts were more similar (semantically and stylistically) other narrative texts, while expository texts were more similar (semantically and stylistically) to other expository texts. Ledoux & Gordon found that the disruption (measured by increased reading time following the interruption) was greater when the interrupting text was of the same style as the primary text.

1.3.1.3. Role of Cognitive-Similarity on Disruptiveness of Interruptions

The disruption caused by encountering an interruption with a similar content to the main task may come from several kinds of interference between the interruption and the primary task:

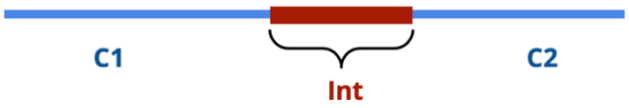


Figure 1-5. Schematic of information flow during interrupted reading.
C1 represents the primary text before an interruption and C2 represents the primary text following an interruption.

I. During encoding of the interrupting materials, there may be interference in working memory, because working memory may contain a combination of both the text preceding the interruption (C1 in **Figure 1-5**) and the interruption material ("Int" in **Figure 1-5**). In this

case, if the two sources of information cannot be kept separate in working memory, the interruption effectively adds noise to the WM representation of the preceding text (C1).

II. Upon resuming the primary text (start of C2), some narrative information from before the interruption (C1) often needs to be retrieved from long-term memory (Ericsson and Kintsch 1995). There may be interference in the retrieval process as the interrupting material (Int) is also present in working memory and in long-term memory.

III. Upon resuming the primary text (start of C2), information about the narrative (C2) needs to be extracted from the text, stored, and manipulated in working memory. However, some of the material from the interruption phase may still be in working memory. Thus, there may be interference in the working-memory between Int and C2 materials.

In addition to content-similarity, the similarities (overlap) in the cognitive processes between the primary task and the interruptions may also enhance the interfering effects of interruptions. For instance, one may expect memory-based interruptions to posit a higher difficulty on narrative processing than math-related interruptions. Rubin et al. (Rubin, Schrauf, and Greenberg 2003) found that cognitive processes involved in autobiographical memory and narrative are more similar than those involved in arithmetic and narrative. Furthermore, the neural pathways that process memory are tightly intertwined with the neural circuits that process narratives (Spreng, Mar, and Kim 2008). Moreover, in an ECoG study, Foster et al. (Foster, Dastjerdi, and Parvizi 2012) found that the default mode network in the brain displays opposing responses during memory and numerical processing. These findings suggest that, in comparison to numerical processing, memory processing is more closely related to narrative processing, both in terms of cognitive processes and the content. Although Glanzer et al. (Glanzer, Fischer, and Dorfman 1984) de-emphasized the role played by task-switching, they did not directly compare the reading time consequences of (e.g.) arithmetic-task and reading-task

interruptions. Therefore, it is still not known how interruption effects vary according to the degrees of similarity of the cognitive processes involved in the primary task and the interrupting task. In Chapter 4, we will return to this question in the context of reading interruptions.

1.3.1.4. Evaluation Methods for Measuring Disruptiveness of Interruptions

The disruptiveness of an interruption during reading has primarily been quantified in two ways: first, via its effects on reading time and second, via its effects on memory accuracy. Specifically, the most common method for quantifying disruption is to measure the reading time of the first sentence immediately following the interruption, with longer reading times interpreted as greater disruption (Ledoux and Gordon 2006). A less common method of quantifying cognitive disruption is to measure what participants remember from the reading task (Foroughi et al. 2015).

Participants are generally slower when resuming a task after interruptions, an effect known as resumption lag. Two findings are consistently obtained across different human studies that investigate the effects of incongruencies on reading comprehension: (1) There is an increase in reading time following the interruptions; (2) The increase in reading time is higher for incongruencies with a longer duration (Monk, Trafton, and Boehm-Davis 2008; Foroughi, Werner, Barragán, et al. 2016). The increase in reading time is interpreted as a higher difficulty in processing the information when resuming reading after the interruptions. For instance, Ledoux and Gordon showed that participants were slower to resume reading after encountering an interruption with a higher similarity to the primary reading task, suggesting a higher disruption caused by the more similar interruptions (Ledoux and Gordon 2006). Despite the consistent observation of resumption lag following, the processes that cause this delay and its dependence on the type of interruption (e.g., arithmetic problems versus memory-oriented problems) remain uncertain.

In contrast to the consistently observed resumption lag, the effects of interruptions on subsequent memory have varied widely. In their classic work introducing the reading interruption paradigm, Glanzer et al. (Glanzer, Dorfman, and Kaplan 1981; Glanzer, Fischer, and Dorfman 1984) employed many different types of interruption tasks during reading (e.g. unrelated reading, counting, arithmetic problems), but consistently observed that subjects' comprehension was unaffected by interruptions. However, Foroughi et al. (Foroughi et al. 2015) later revealed that when the memory questions were not simple recognition questions, but demanded further (perhaps inferential) processing of the text, then interruptions were found to affect memory and comprehension performance.

Interruptions can also be thought of as "event boundaries" in an ongoing perceptual stream, and there are inconsistent findings of how such "event-boundaries" affect memory. On the one hand, arousal at event-boundaries can be beneficial for immediate memory recognition following the event boundaries. Swallow et al. (Swallow, Zacks, and Abrams 2009) showed that event boundaries in perception, improve memory encoding. Similarly, Zheng et al. (Zheng et al. 2022) found that boundaries boost immediate recognition. However, other studies have indicated detrimental effect of event boundaries on memory. When information spanned across boundaries, the memory for information was declined (Lawrence and Peterson 2016; Radvansky and Copeland 2010). Furthermore, Payne et al. (Payne et al. 2007) showed that arousal impaired the consolidation of surrounding memory. Finally, Ben-Yakov et al. (Ben-Yakov, Smith, and Henson 2021) demonstrated that surprise did not retroactively modulate memory of preceding events, neither when participants were tested immediately after the event videos nor when they were tested 24 hours later. Altogether, the literature does not present a consistent picture of how interruptions should influence delayed recall of the information immediately before and after interruptions.

1.3.1.3. Handling Interruptions using Working Memory and Long-Term Memory

Induced interruptions have been used as an experimental technique to understand human memory mechanisms. In traditional memory theories, two qualitatively different memory stores have been identified. First there is a transient short-term memory (STM) with very limited capacity, and this term is often used interchangeably with working memory (WM) (Cowan 2017). Second, there is long-term memory (LTM) with very large capacity, where successful storage required significant time (Malmberg, Raaijmakers, and Shiffrin 2019) .

A standard two-store framework does not easily account for the behavioral results observed in studies of reading interruptions. For instance, encountering reading interruptions (e.g. arithmetic problems inserted in the middle of a text) is found to have little effect on reading comprehension (Glanzer, Fischer, and Dorfman 1984). In the view of Ericsson and Kintsch (Ericsson and Kintsch 1995), this result cannot be explained under conventional models of WM, because if an activity is interrupted, information in temporary storage (such as WM) will be lost and cannot be later resumed. On the other hand, long-term memory is thought to be too slow to mediate the retrieval of linguistic content on demand during reading. Motivated in part by this observation, Ericsson & Kintsch proposed a new model of memory, as we elaborate below.

Ericsson and Kintsch (Ericsson and Kintsch 1995) introduced the concept of long-term working memory (LTWM). Their model challenged the view that only limited-capacity WM is used for maintaining information in reading of short passages. First, they noted that a WM system with limited capacity cannot account for greatly expanded working memory capacity of skilled performers (i.e., experts) in domains such as chess and reading. Second, they argued that if an activity is interrupted, information in temporary storage (i.e., WM) will be lost and cannot be later recovered.

But they also argued that LTM was too slow to be used as storage in real-time reading settings. Therefore, the LTWM model proposed additional control processes added to LTM to facilitate rapid encoding and retrieval of information. A key piece of evidence for the LTWM model was the aforementioned findings from Glanzer et al, (Glanzer, Dorfman, and Kaplan 1981; Glanzer, Fischer, and Dorfman 1984) and Ledoux et al. (Ledoux and Gordon 2006) that interruptions had little effect on participants' ability to recognize and recall material from the passages.

Foroughi et al. (Foroughi et al. 2015) found that interruptions during reading led to a decline in textual comprehension, and on this basis challenged the empirical validity of the LTWM model. They suggest that the transient portion of working memory is essential for text comprehension, and that there is no need to posit additional LTM processes (Foroughi, Werner, Barragán, et al. 2016). Foroughi et al. suggested that the maintenance of information in a short-term store (i.e., working memory) was necessary because interruptions were indeed affecting this information, and that these effects could be detected by text comprehension that went beyond simple recognition testing. Furthermore, they found that adding a time-out period before the interruptions led to a recovery of memory performance, which they interpreted as providing working memory systems with enough time to finish processing the preceding text, so that there was nothing to be disrupted (Foroughi et al. 2015; Foroughi, Werner, Barragán, et al. 2016). Ericsson and Kintsch, however, maintain that Foroughi et al.'s findings do not challenge the LTWM theory, as the theory is not about whether disruptions to comprehension ever occur, but rather regarding the magnitude of the effects (Delaney and Ericsson 2016).

The literature reviewed above suggests that our knowledge about reading interruptions is limited by two factors. First, there is the fact that WM-based and LTWM-based models are quite abstract and do not make sharply differing prediction. But the second, and possibly more fundamental

limitation is the variability in empirical finings across studies, which have used many different primary texts and interruption tasks over the decades. Therefore, in Chapter 4 we will revisit the fundamental questions about memory interruptions while, employing many manipulations within a single consistent paradigm, and employing a large sample of participants in each condition.

Our central focus here is how people tackle a primary task with occasional interruptions, but this setting may share several features with the performance of two concurrent tasks. Specifically, there is longstanding demonstration of which kinds of working memory processes interfere in "dual task" settings, when participants perform two tasks at the same time. In the dual task, participants must divide their attention between both tasks and switch between the associated memory processes for each one (Treisman 1964; Broadbent 1966; Larsen and Baddeley 2003; Logie et al. 2004). In contrast, our focus is on how participants switch from reading (the primary task) to an interruption (a secondary task), and how they resume the primary reading task following each interruption. Importantly, the consecutive interruption tasks are not related to one another – each interruption can be performed individually, without drawing on knowledge of content or responses in the previous interruption. Therefore, participants in our interruption experiments only need to maintain and access long-term information from the primary task. Nonetheless, the process of reinstating memories and goals may be similar between dual-task paradigms and interruption paradigms. Though we do not explore the relationship between these paradigms here, it is a promising direction for future work.

1.3.2. How Neural Language Models Separate Relevant from Irrelevant Information in Prior Context

Having discussed how human temporal integration is affected by interruptions, it is instructive to consider how artificial systems behave when they confront the same problem. Neural language models

(NLMs) have recently achieved impressive performance in predicting linguistic sequences (Khandelwal et al. 2018; Subramanian et al. 2020). To do so, these computational models must determine which aspects of prior context are important for future prediction. In the terminology of cognitive psychology, we might say that these models, when exposed to a string of words, must decide which information should be stored or maintained in "memory" and then later "retrieved" in the service of predicting the next word. More generally, we could say that these language models are continually integrating information over time as they extract the meaning of the input and update their representations of the context. What kinds of integration mechanisms are employed by these contemporary NLMs, both those with feedforward and recurrent architectures? To what extent are they able to separate incongruent information so that it does not affect their predictions? Finally, how do these integration and separation processes compare with human behavior? In this Chapter, we begin to address these questions, by testing language models in situations where in which the language sequence contains a mixture of both relevant and irrelevant (incongruent) information.

NLMs are designed and trained to make reliable language predictions when dealing with consistent data streams, however real-world language data often includes passages of incongruity. Similarly, when we process language data in the real world, we may experience incoherencies, like a brief change of topic during a conversation. For instance, in the middle of a conversation with a group of friends about our family members, one of our friends might suddenly notice a news headline on his cell phone and read it aloud to the group. In this case, the linguist information sequence that we receive includes a segment of unrelated information (**Figure 1-6**). How do contemporary NLMs process such incongruities in the language context that they are using to make predictions? How might the architectural characteristics of NLMs modulate their ability to access information (both relevant and irrelevant) in prior context? Finally, what are the similarities and differences between how language

models solve this problem (of incongruent interrupting text) and how humans handle interruptions in in linguistic context? Below, I review previous work that has explored aspects of these questions, before identifying the gap in the literature that I will address.

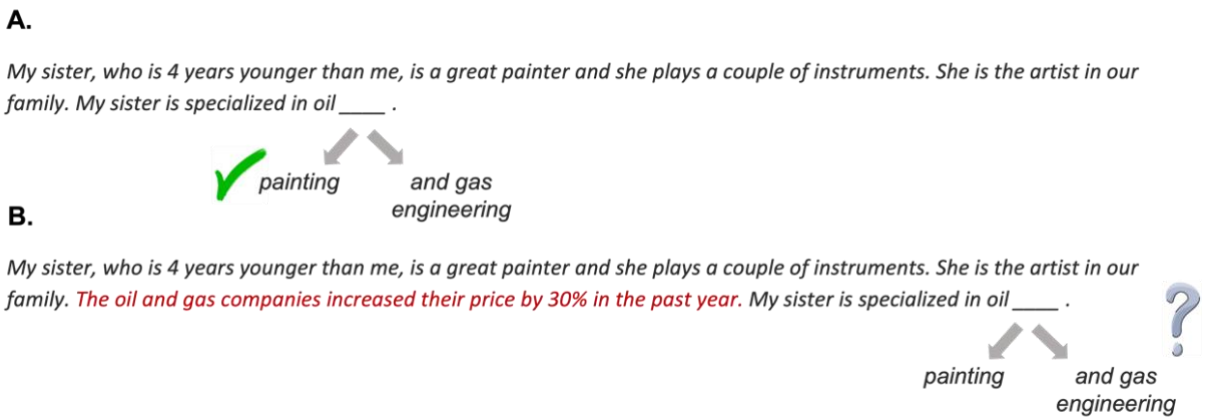


Figure 1-6. Examples of how a language model's predictions may be influenced by an intact versus an incongruent prior context.
A) In this example context is intact, and so the language models can easily choose the correct word from possible options. B) In this case, the incongruent context which may prevent the language model from making an accurate word prediction.

1.3.2.1. The Impact of Architectural Constraints on Maintenance and Retrieval of Information in NLMs

Maintenance and retrieval capabilities of NLMs may depend on their architectural constraints. In this Chapter, we will test the performance of two language models with very different architectures, (i) a long short-term memory (LSTM) model (Schijndel and Linzen 2018; Hochreiter and Schmidhuber 1997) and (ii) GPT-2 Transformer language model (Vaswani et al. 2017; Radford et al. 2020). The LSTM is a recurrent deep learning model that incorporates the past by reusing the information from previous time steps and through dedicated memory cells (Schijndel, Mueller, and Linzen 2019) (see **Figure 1-7** A). The GPT-2 model (GPT stands for "Generative Pretrained Transformer") is a

large Transformer-based language model (LM) trained with the objective of predicting the next word, given all the previous words in a fixed-length context (Radford et al. 2020). A Transformer model is a deep learning model that employs an attention mechanism to differentiate the significance of each token in the input sequence (Vaswani et al. 2017) (see **Figure 1-7** B). The attention mechanism in Transformers determines how each token (word) in a sequence is influenced by all other words in the sequence.

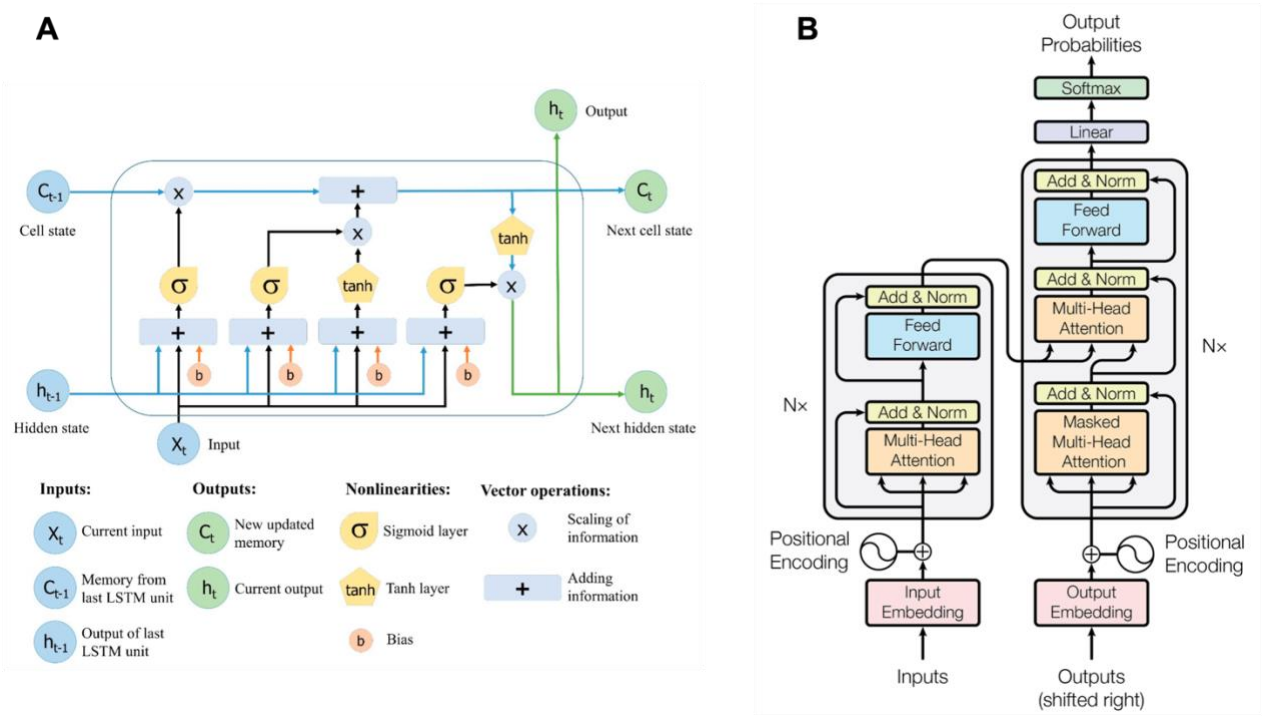


Figure 1-7. Schematic of an LSTM and a Transformer neural language model' architectures.
A) Schematic of an LSTM (memory-based) language model. (This schematic image is borrowed from (Le et al. 2019)).
B) Schematic of a Transformer (attention-based) language model. (This schematic image is taken from (Vaswani et al. 2017)).

1.3.2.2. Accessibility to Prior Context in Different NLMs

In what ways do architectural constraints affect access to prior context in LSTM? In theory and based on its architectural properties, LSTM has the ability to maintain pieces of information from the prior

context in dedicated memory cells for many hundreds of tokens, and possibly longer (Khandelwal et al. 2018; Schijndel, Mueller, and Linzen 2019). This characteristic enables LSTM to learn long-range dependencies. However, LSTM's architectural constraints pose a "limited-capacity" challenge. In order for the LSTM architecture to have access to prior context, it must carry the information forward in its internal states. Then, at the next time step, it must update its inner context based on the current input. Thus, at each time step, the model must make a decision of how to update a limited-dimension representation of context. Crucially, if the model decides not to incorporate a piece of information into its context representation, there is no way for it to later access that token. As a result, LSTM is under the pressure to forget some part of the information to free up space to maintain information that is most relevant to the upcoming prediction. Therefore, what happens in practice is that the access to prior information is proximity-dependent: on average, recent tokens exhibit a greater influence over the prediction of the next token (**Figure 1-8**). Khandelwal et al. (Khandelwal et al. 2018) described this property of LSTM models by saying that their context representation is "sharp nearby, fuzzy faraway".

Because it employs a Transformer architecture, GPT-2 does not face the same capacity pressures that an LSTM does. The GPT-2 model has direct access to a long but fixed-length window of the prior information (Vaswani et al. 2017). We can think of it as having perfect memory of a finite length of the past. In contemporary LM architectures, Transformers typically have access to hundreds of prior tokens, but sometimes thousands in larger models. Even with perfect memory, however, GPT-2 still faces the challenge of identifying what pieces of information are relevant for making predictions about the future token (**Figure 1-8**).

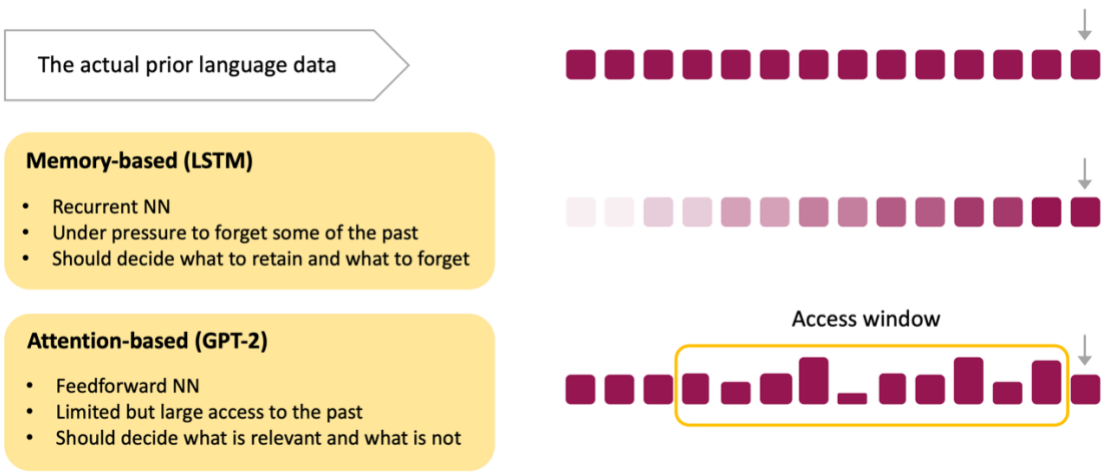


Figure 1-8. Schematics illustrating the accessibility to prior context in different neural language models. Neural language models use a sequence of linguistic data as their prior context to make prediction about the upcoming word, but LSTM (middle) and GPT-2 (bottom) have access to different types of information within their prior context.

1.3.2.3. Similarities and Differences of NLMs to Language Processing in Humans.

Humans and neural language models have some shared capabilities: both have the capability to parse language data, can extract abstract information from it, and can generate natural-seeming sequences of words (Khandelwal et al. 2018; Radford et al. 2020; Linzen, Dupoux, and Goldberg 2016). In doing so, it has been argued that humans and neural LMs share two computational commonalities: both are involved in continuous next-word prediction; and, both represent words as a function of the prior context (Goldstein 2021).

In addition to studying their computational commonalities, we can also assess the "behavior" of humans and LMs using similar criteria (Schijndel and Linzen 2018). In humans, the time taken to read a word or sentence is commonly taken as an index of how easily they are processing linguistic information, or how well the written text accords with their linguistic knowledge and expectations (Monk, Trafton, and Boehm-Davis 2008). For example, Smith and Levy (N. J. Smith and Levy 2013) showed that human reading time for words decreases with higher word probability. Therefore, in

many contexts a lower reading time (RT) reflects less surprise while reading. In NLMs, to measure a model's level of "surprise" while processing a piece of language data, we can calculate their "perplexity" (Khandelwal et al. 2018). A low perplexity indicates that a model is not surprised by the linguistic information, i.e., that the word or words are consistent with its internal model. Data from Schijndel and Linzen (Schijndel and Linzen 2018) support the correspondence of NLMs' surprisal with humans' reading time. They showed that LMs' surprisal matched with humans' self-paced reading times in a Natural Stories corpus. From this perspective, then, LMs' perplexity and humans' reading time reflect similar measures in how easily each of them processes a particular string of language input.

1.3.2.4. Memory Mechanisms in Humans and NLMs

NLMs with distinct architectural constraints and memory capabilities may approximate humans' language processing in different ways. While LSTM are forced to selectively maintain the most relevant information for subsequent predictions, GPT-2 benefits from a "perfect memory" of a fixed-length window of prior input. This distinction may be why LSTMs seem to represent gist-like semantic properties in their context representations, while GPT-2 makes predictions based on verbatim properties of the prior context (Armeni, Honey, and Linzen 2022). In humans, it has long been thought that linguistic tasks (such as reading and next-word prediction) are supported by multiple mnemonic and contextual representations (Kintsch and van Dijk 1978), including verbatim maintenance of prior text, and more gist-like representation of semantics and situation models. Therefore, perhaps aspects of human language processing that rely more on abstracted semantic information, might behave closer to a recurrent model such as our LSTM. In contrast, the aspects of human language processing that depend more on verbatim properties of the prior context, might behave closer to a Transformer, such as GPT-2. That said, the size of the models and their training

corpora and loss functions can dramatically impact how readily they extract abstracted situational and semantic features (Shin et al. 2022).

At this moment, despite interest in augmenting neural network models with dedicated memory systems, there is no large language model that incorporates multiple human-like memory mechanisms. Humans employ multiple distinct memory systems to generate, store and retrieve context representations that make available rich semantic and situational properties, while also maintaining high-resolution surface information over the short term. This multi-memory system, allowing access to short- and long-term data, may render humans (relatively) robust to the irrelevant and incongruent items in a stream of input.

1.3.2.5. Effects of Context Incoherencies on Language Processing in Humans and NLMs

Both humans and NLMs may be sensitive to incongruent items when processing an information stream, but the consequences are likely different across humans and NLMs. In humans, increased similarities between the interruptions and the primary text generate greater processing difficulties (Ledoux and Gordon 2006). We reviewed the prior related works on effects of content-similarity of interruptions on their disruptiveness, earlier in this Chapter (section 1.3.1.2.). In contrast to humans, NLMs may suffer most from the deleterious effects of incongruent information when the incongruous information is moderately similar to the primary text. Consider, three ways in which an interruption's similarity could modulate its interfering effects. First, when the prior context contains incongruencies that are very dissimilar from the primary text, then LMs may suffer from interference between relevant and irrelevant information and demonstrate a reduction in their prediction accuracy. However, if the information is highly incongruous, the model may also ignore it if it cannot be incorporated with the broader context of the text. Second, when the interrupting information is moderately similar to the

primary text, the predictions of the NLMs may be most corrupted. In this "sweet spot", the incongruency is similar enough to be mistaken for relevant material, but dissimilar enough to cause an erroneous prediction. Finally, when the interrupting information is highly similar to the primary text (e.g. information about rifles inserted into a story about handguns) then this additional piece of information might even facilitate NLMs predictions for some context-dependent words, as it can help to reinforce the appropriate context. Altogether, we may expect a U-shape relationship between similarity and disruptiveness of interruption: the most disruptive effects on NLMs may appear at an intermediate level of similarity to the primary text.

The length of the incongruencies (i.e., interrupting information) may also affect their influence on human and NLM language processing. Two findings are consistently obtained in human studies of incongruencies in reading comprehension: (1) There is an increase in reading time following the interruptions; (2) The increase in reading time is higher for incongruencies with a longer duration (Foroughi, Werner, McKendrick, et al. 2016; Monk, Trafton, and Boehm-Davis 2008). Thus, if we use perplexity to quantify NLM surprise, then NLMs should demonstrate a similar pattern to humans in processing information after incongruency, exhibiting: (1) increased perplexity following the insertion of inconsistent or incongruous material; and (2) further increased perplexity when the incongruous material contains more words.

The literature reviewed in this section motivates us to directly test how NLMs handle incongruencies in their input streams. In Chapter 5, we test both a recurrent architecture (LSTM) and a feedforward architecture (Transformer) by exposing them to multiple types of incongruent input, and we characterize their language prediction performance, how this compares with the measures of human reading times obtained in Chapter 4.

1.3.2.6. Achieving human-level language processing capabilities in language models

Finally, when studying an LM from and comparing its performance to humans, we should remember that "human-level performance" does not necessarily reflect "human-level competence" (Firestone 2020). Drawing on Chomsky (Chomsky 1965), Firestone defines these terms as the following: competence refers to what a system knows, whereas performance refers to what a system does. The distinction between performance and competence can arise from multiple factors: First, demonstrating the same level of performance may not reflect the same level of knowledge. Therefore, even if an LM exhibits similar performance to humans in a particular task, it does not mean that the LM has the same language knowledge as humans. Second, LMs and humans might perform similarly in a specific task, but they might not *fail* in the same way. Failing in different ways does not mean that humans and machines have completely different capacities and representations. It could be that humans and machines fail differently despite having similar competence but because they have different constraints on processing the input they receive. And, finally, humans and LMs have different constraints that may influence their performance. To facilitate fair comparisons humans and LMs, one method could be constraining one system to become more similar to the other, while another method could be using species-specific tasks to evaluate each system (Firestone 2020).

Chapter 2: Effects of Temporal Integration and Separation on Category Learning

2.1. Background and Motivation

In the biological world, category learning usually takes place in time, as the learner is exposed to exemplars presented one after the other. These exemplars, however, are not independent from each other, and are usually similar nearby in time. For instance, when meeting a group of new people, we see many examples of the same face from different angles consecutively before moving to the next person. What are the consequences of exposure to such temporally correlated data for learning? How could a learning system with brain-like architectural constraints, exploit the temporal structure in the data for more efficient learning?

We hypothesized that a category-learning system may be able to take advantage of temporal structure in training data by selectively integrating information from related samples, while separating information from unrelated samples. Indeed, there are simple brain- inspired mechanisms that may enable learning systems to benefit from combining related information, while avoiding interference from unrelated information. Our focus on simple integration and separation mechanisms is motivated by the fact that (i) our brain dynamics appear to blur information over time, as there is ubiquitous autocorrelation in cortical dynamics (Honey et al. 2012; S. M. Smith et al. 2013; Chaudhuri et al. 2015; Bright et al. 2020; Murray et al. 2014; Raut, Snyder, and Raichle 2020), while (ii) neural circuits' sometimes shift their state suddenly at "event boundaries", and this appears to be associated with "resetting" of context representations (DuBrow and Davachi 2016; Baldassano et al. 2018; Chien and Honey 2020).

2.1.1. Research Questions

To date, we lack a normative model of how the architectural constraints in biological neural networks (autocorrelation and boundary resetting) influence category-learning, in a realistic setting with temporal correlation in training data. Therefore, in this Chapter, we set out to examine the efficacy of these brain-inspired mechanisms in boosting category learning in neural networks. We investigated the following two brain-inspired mechanisms: (i) local linear recurrence in internal representations; (ii) a gating mechanism that resets internal representation at transitions between categories. In particular, we asked the following questions:

- Can brain-inspired recurrence and gating mechanisms enable neural networks to exploit temporally correlated data for higher category learning efficiency, without using BPTT? We hypothesized that networks equipped with two brain-inspired mechanisms — local linear recurrence and gating — could exploit temporal autocorrelation in data for more efficient category learning.

Significance

In theory, the existence of autocorrelated dynamics in the brain presents a challenge, because it is unclear whether (and under what conditions) learning is facilitated by blending past and present representational states. By quantitatively addressing this question, we can better understand how learning can proceed in neural circuits with unavoidable correlation in their dynamics. We focused on computational consequences of leaky integrator and boundary-resetting mechanisms for category-learning, in terms of the speed and efficiency of learning. Our hypothesis was that networks equipped with two brain-inspired mechanisms — local linear recurrence and boundary-resetting — could exploit temporal autocorrelation in data for more efficient category-learning.

2.2. Methods

We built neural networks with and without two brain-inspired constrains.

2.2.1. Brain-Inspired Constraints

Leaky memory (local linear recurrence): We added leaky memory to the internal representations (hidden units) by linearly mixing them across consecutive time points. Hidden unit activations were updated according to the following function:

$$H(n) = \alpha H(n-1) + (1-\alpha)\, ReLU(W_{IH} I(n)) \qquad \text{Eq. (1)}$$

where H(n) is the state of the hidden units for trial n, I(n) is the state of the input units for trial n, α is a leak parameter, W_{IH} are the connections from the input layer to the hidden layer, and ReLU is a rectified linear activation. We set $\alpha = 0.5$ in these experiments.

Memory Gating: To reduce the interference between items from different categories in the leaky memory, we employed a gating mechanism to reset memory at the transitions between categories. Therefore, if sample n was drawn from a category other than the category of sample n−1, then we set $\alpha = 0$ in Eq. (1) on that trial n (**Figure 2-2** C).

2.2.2. Manipulating Autocorrelation in Data

We manipulated the amount of autocorrelation in data by varying the number of consecutive samples drawn from the same category. We began each training session by generating a random "category order", which was a permutation of the numbers from 1 to N (e.g., the ordering in **Figure 2-1** B). The same category order was used for all conditions in that training session.

To sample with minimum autocorrelation (maximum interleaving), we sampled exactly one exemplar from each category, before sampling from the next category in the category order (1 repeat) (**Figure 2-1** B). This condition is called "minimum autocorrelation" because all consecutive items were from different categories, and there were not more examples from a category until all other categories were sampled. We increased autocorrelation by increasing the number of consecutive samples drawn from each category (3 repeats and 5 repeats in **Figure 2-1** B). Finally, we also used the standard random sampling method, in which items were sampled at random, without replacement, from the training set (**Figure 2-1** B). The training set was identical across all conditions, as was the order in which samples were drawn from within a category (**Figure 2-1** B).

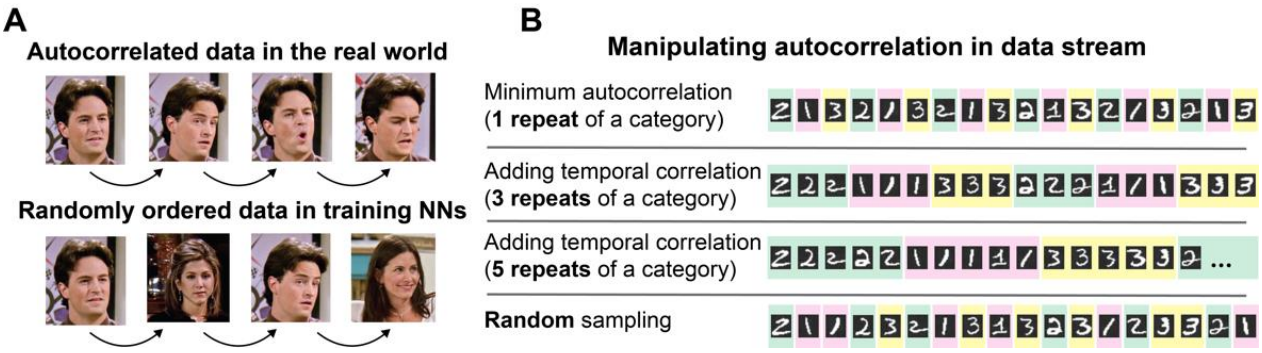


Figure 2-1. Temporal autocorrelation in real-world data and in neural network training.
A) Top: temporally autocorrelated data in the real world. Bottom: randomly ordered data as the common practice in training neural networks. B) Manipulating autocorrelation levels in training data using the ordering of training samples. Colored rectangles indicate the amount of autocorrelation induced by repeating a category.

2.2.3. Training Neural Networks

Dataset. We tested MNIST, Fashion-MNIST, and synthetic datasets containing low category overlap (LeCun, Kavukcuoglu, and Farabet 2010; Xiao, Rasul, and Vollgraf 2017)

Learning rule. For the learning rule, we used backpropagation, however the gradient computation did not account for the fact that the neurons were leaky. Therefore, the update rule in [leaky memory +

reset] model is different from the common update rule in recurrent models (e.g., LSTM). LSTM uses backpropagation through time (BPTT), which is implausible for biological settings. In learning with BPTT, the same neurons must store and retrieve their entire activation history (Lillicrap et al. 2020). In contrast, in the [leaky memory + reset] model, neurons only use local information from their most recent history. Therefore, it is computationally much simpler because it does not require maintaining the whole history and computing the gradient relative to all that history.

Objective function. We used backpropagation with both mean squared error (MSE) and cross-entropy (CE) loss functions. The results reported here are using MSE, primarily for the ease of comparison with later reconstruction error measures in this manuscript. However, the same pattern was observed using CE loss. Also, it has been shown MSE loss provides comparable performance to commonly utilized classification models with CE loss function (Illing, Gerstner, and Brea 2019). To test incremental learning, we employed stochastic gradient descent (SGD), updating weights for each training sample.

Optimization, initialization, and activation function. We tested the model both with and without RMSprop optimization, along with Xavier initialization. We applied ReLU to hidden units and Softmax or Sigmoid to the output units.

Hyperparameters. For MNIST and Fashion-MNIST, we used a 3-layer fully connected network with (784, 392, 10) dimensions and a learning rate of 0.01. The learning rate was not tuned for a specific condition. We used the same learning rate across all conditions; only autocorrelation varied across conditions. To compensate for the potential advantage of a specific set of hyperparameters for a specific condition, we ran 5 runs, each with a different random weight initialization, and reported the

averaged results. When RMSprop was implemented, β1 and β2 were set to 0.9 and 0.99, respectively (Ruder 2016).

2.3. Results

We found that autocorrelated (blocked) training data slowed incremental learning in memoryless networks (**Figure 2-2** A). Moreover, in these networks, minimum autocorrelation (maximum interleaving) yielded more efficient learning than random sampling (**Figure 2-2** A). These observations generalized across all tested datasets and across MSE and CE loss, with and without RMSprop optimization.

In contrast to the detrimental effects of autocorrelation in memoryless learners, autocorrelation in training data increased learning efficiency in learners with leaky memory, as shown in **Figure 2-2** B. Moreover, adding a gating mechanism to the leaky memory units further increased their learning (**Figure 2-2** C). In learners with leaky memory and gating, all levels of autocorrelation significantly outperformed random sampling and sampling with minimum autocorrelation (1 repeat) (**Figure 2-2** C). These findings generalized across MNIST, Fashion-MNIST, and synthetic datasets.

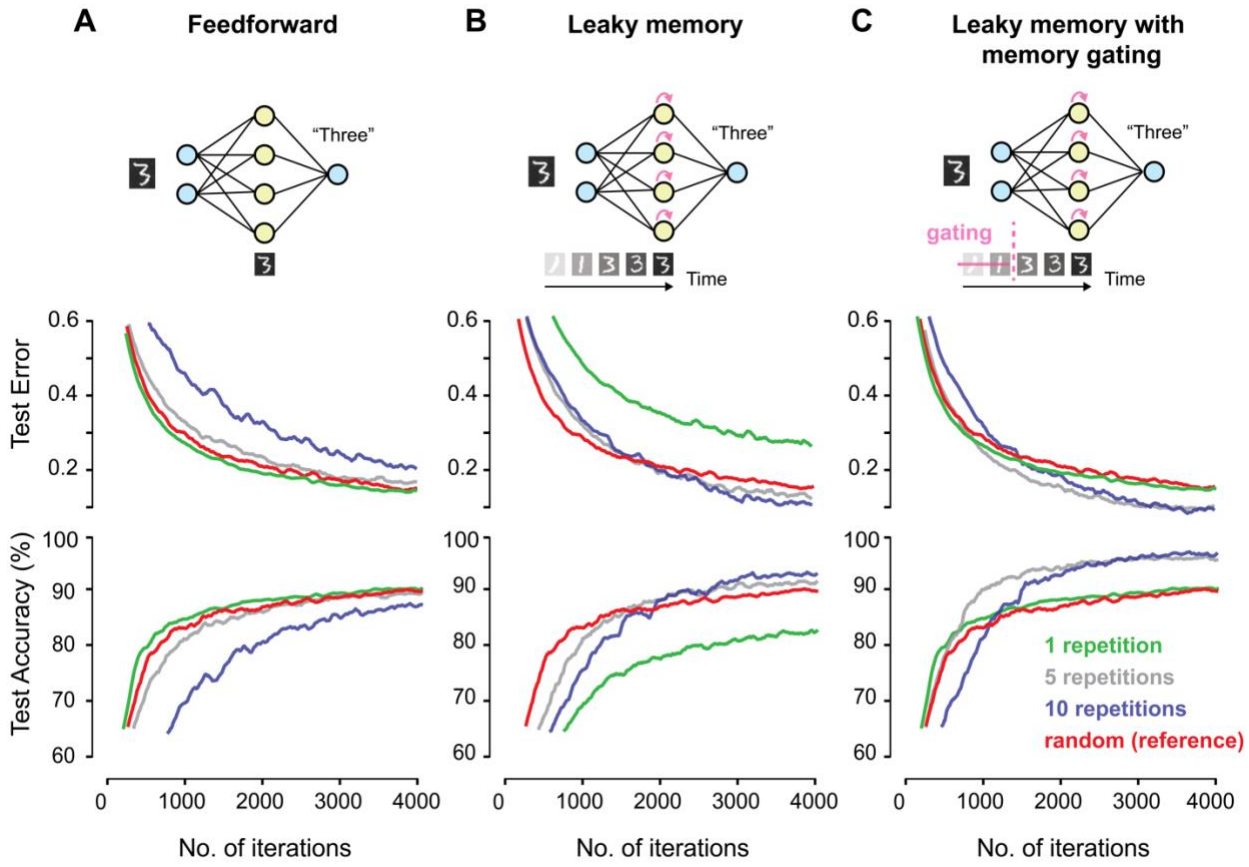


Figure 2-2. Neural architectures for classifying temporally correlated data.
A) Test error (MSE loss) and test accuracy in SGD training of a feedforward neural network (MNIST data) across different autocorrelation levels. B) The same as (A) but for a neural network with leaky memory in internal representations. C) The same as (A) and (B) but for a neural network with leaky memory and gating mechanisms. The performance of the feedforward neural network with random sampling (shown in red) is used as an identical reference in all 3 plots so it can be used for comparing performance across models. [Curves in this figure have been averaged over 5 runs with different initialization and were further smoothed using a 100-iteration moving average.]

Why does averaging current and prior states produce more efficient learning from sequentially correlated data streams?

So far, we showed that when data sampled at a given moment shares category-relevant features with recent samples, learners with leaky memory were able to exploit this property for more efficient category learning. Importantly, the resetting mechanism prevented the mixing of hidden representations from samples of different categories, allowing the system to benefit most from the

correlation in data, while not suffering from between-category interference. However, it is still not clear why averaging current and prior states produces more efficient learning from autocorrelated data.

To further investigate why combining consecutive states produces more efficient learning from autocorrelated data, we looked for situations in which combining consecutive states is not helpful. Our hypothesis was that averaging across consecutive members of the same category instantiates an assumption (an inductive bias) that consecutive samples from the data stream share task-relevant local features. When this assumption is satisfied, averaging (smoothing) across consecutive samples increases the proportion of variance in the hidden units, that is associated with category-diagnostic features. This hypothesis predicts that if consecutive items in the data stream do not share any local features, then the benefits of leaky memory will be eliminated. To test this hypothesis, we trained our model on a data structure in which the consecutive items do not share local features.

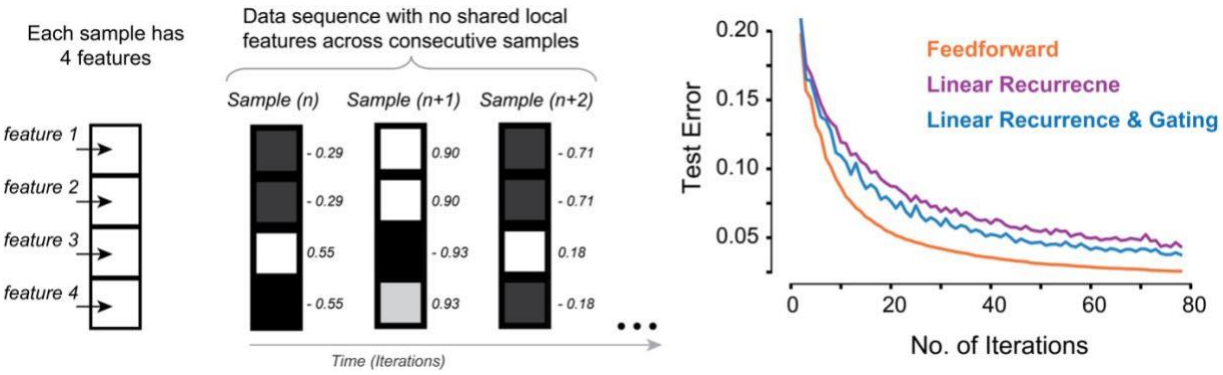


Figure 2-3. Conditions under which recurrence and gating mechanisms impede category learning.
Left: Synthesized data stream with no shared local features across consecutive samples. Feature i in sample (n) is orthogonal to feature i in sample (n+1). The numbers on the right side of each sample show the feature values. Right: Test error for feedforward memoryless model, model with linear recurrence, and model with linear recurrence and gating.

We found that when consecutive items in the data stream did not show autocorrelation, the advantage of maintaining a leaky memory in internal states was eliminated, and leaky memory, with or without boundary-reset, always learned more slowly than feedforward models (**Figure 2-3**).

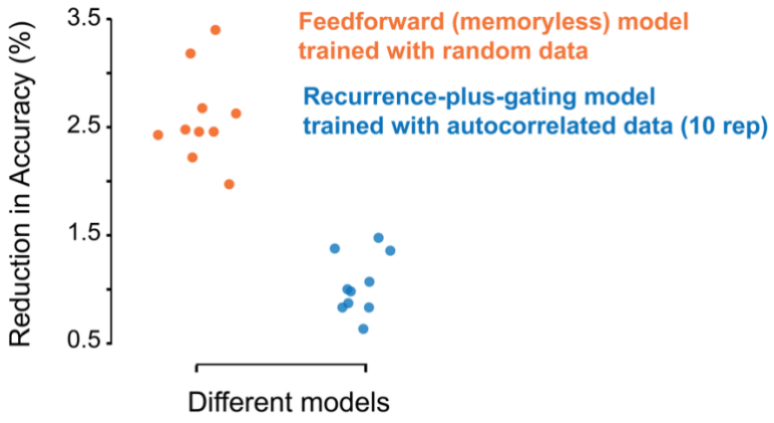


Figure 2-4. Change in classification accuracy when recurrent and feedforward models are tested on noisy data. Y-axis shows the percentage of reduction in classification accuracy when tested on noisy data. X-axis shows different models. Each dot shows the reduction in classification accuracy for each single run with a different random initialization.

Additionally, biological systems might employ this simple linear recurrence because, while it does not impair learning from random data streams, it can reduce the noise in the presence of temporal autocorrelation (**Figure 2-4**). This theory was supported by our results showing that compared to the memoryless model, model with recurrence and gating was less susceptible to misclassify noisy data. Furthermore, linear recurrence can be used in biological systems to increase the accuracy of categorization decisions (reduce misclassification) when multiple members of the same category are observed sequentially (**Figure 2-5**). This would increase the certainty of the classification with each sample.

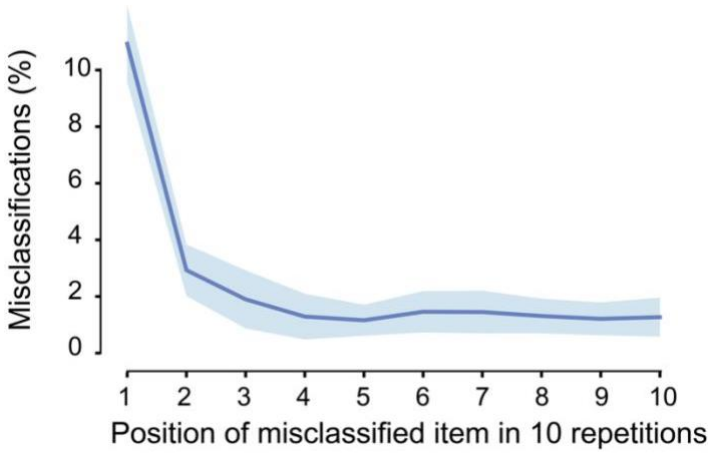


Figure 2-5. Percentage of misclassifications for each position within autocorrelated samples.
Y-axis shows the percentage of misclassification. X-axis shows the position of the tested iteration in 10 repetitions of sampling from one category. (MNIST dataset)

How does the model equipped with leaky memory and gating compare with a more flexible recurrent model trained with backpropagating gradients through time?

Our leaky memory models were trained without propagating gradients backward through time. Backpropagation-through-time (BPTT) is the common learning rule implemented in training recurrent neural networks. However, it is not thought feasible for biological brains to BPTT because it would require neural circuits to have access to all of their history of their internal states (Lillicrap and Santoro 2019). Having said that, it is important to compare the performance of our model to the performance of a more flexible model that can directly learn from task-relevant temporal structure.

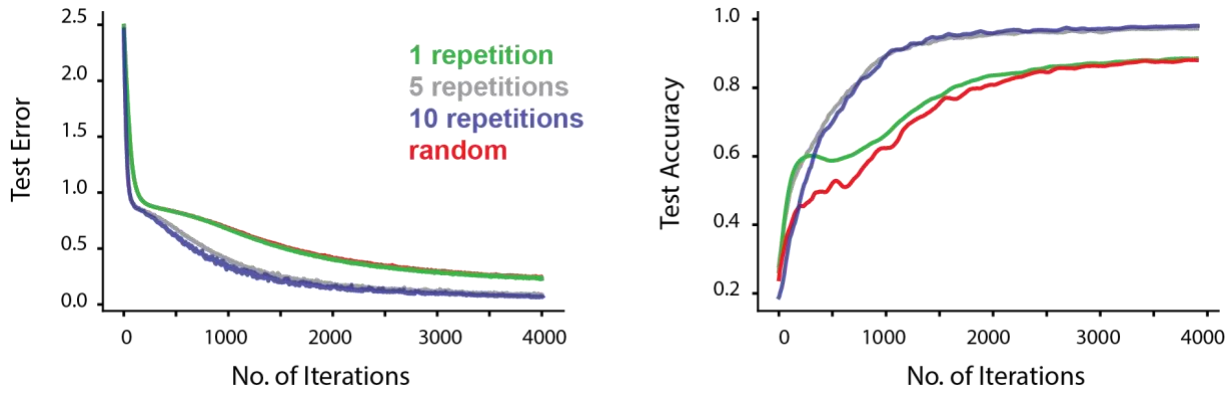


Figure 2-6. Effects of autocorrelation in data on learning in LSTM.
Left) LSTM test error for different amounts of autocorrelation in training data. Right) LSTM test classification accuracy for different amounts of autocorrelation in training data.

We used Long Short-Term Memory (LSTM) model, as a flexible recurrent model trained with BPTT. We then evaluated LSTM's performance using the following criteria: (1) How efficiency of LSTM model is influenced when learning from temporally correlated data; (2) For a specific level of temporal correlation, which model, LSTM vs our [Leaky memory + boundary-reset] model, shows higher learning efficiency. (3) How do these two models perform when generalizing their learning to a sequence of samples different from what they have been trained on.

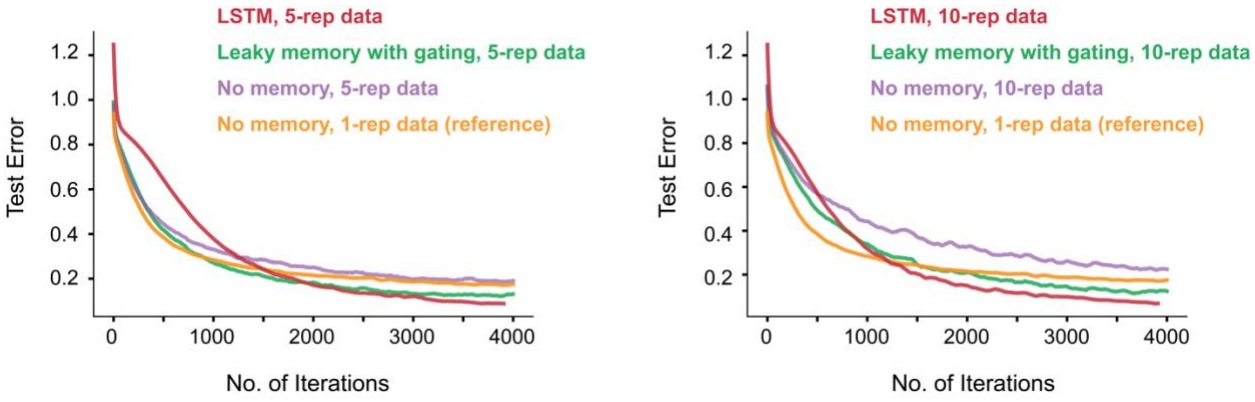


Figure 2-7. Comparing LSTM to models with and without leaky memory and memory gating mechanisms.
Left) Test error for classification with autocorrelation level equal to 5-repetitions (5 consecutive sampling) of each category. Right) Test error for classification with autocorrelation level equal to 10-repetitions of each category. In both

plots, we also show [no-memory with 1-repetition], that is "minimum autocorrelation" (the orange curve), to be used as a reference.

For the first analysis, we found that, similar to our leaky memory model, the LSTM also benefited from autocorrelated data and will produce higher learning efficiency for higher amounts of correlation in data (**Figure 2-6**). In the second analysis, regarding comparing the performance of these two models on the same level of temporal correlation, as expected, we found that the LSTM would demonstrate better results (**Figure 2-7**). We expected this result because the LSTM has a much more flexible architecture and it is trained with BPTT, so the gradient updates are mathematically optimized for the task. For the third analysis, we found that the advantage of the LSTM trained with BPTT was not preserved when the models are tested out-of-domain (**Figure 2-8**). We expected this pattern because the LSTM trained with BPTT can be calibrated to the specific structure of the training data stream and therefore will be more susceptible to suffer from testing on new data sequences.

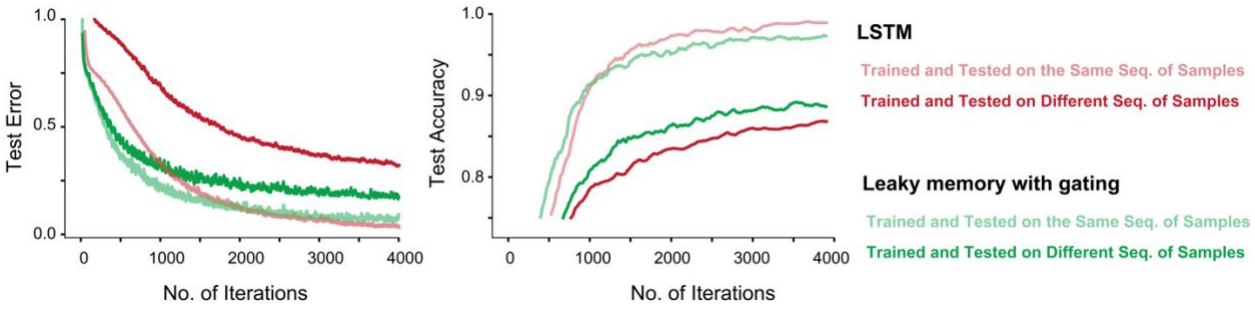


Figure 2-8. Generalization of LSTM and the model with leaky memory and gating to data streams with different temporal structures.
Left) Test error for LSTM and [leaky memory + reset] model, trained and tested on the same sequence of samples or on a different sequence of samples. Right) Test accuracy for LSTM and [leaky memory + reset] model, trained and tested on the same sequence of samples or on a different sequence of samples.

How does the averaging of current and prior activation states (leaky memory) compare to the averaging of gradients (mini-batching)?

The linear recurrence in our leaky memory models effectively generates a weighted average of activations over time. Mini-batching, on the other hand, averages gradients over time (Li et al. 2014). In mini-batching, instead of updating the weights after n single training samples (n updates for n samples), the weights are updated once for each batch of n sample (1 update for n samples). This technique enhances the computational efficiency of training a deep learning model as it reduces the total number of weight update computations.

Leaky memory and mini-batching mechanisms differ in at least two ways. First, implementing leaky-memory and mini-batching affect performance in different ways as a function of the amount of category repetition (autocorrelation in data) (Figure 2-9). We found that mini-batching with varying batch sizes and high or low autocorrelation in the data yielded similar accuracy levels. In contrast, the leaky memory model benefited more from higher levels of autocorrelation in data. Second, the effects of leaky-memory on learning performance can be reversed (switching from improved to worsened performance) when the training data within a category contain non-overlapping features (Figure 2-3). Though we have not tested this directly, an analogous effect seems unlikely for mini-batching. Altogether, leaky memory and mini-batching serve different purposes. Mini-batching is more efficient when the goal is to minimize the number of weight updates, regardless of the number of training samples. However, if there is a limited number of training samples and they are temporally autocorrelated, then leaky memory may produce a higher accuracy by exploiting the temporal correlation in the data.

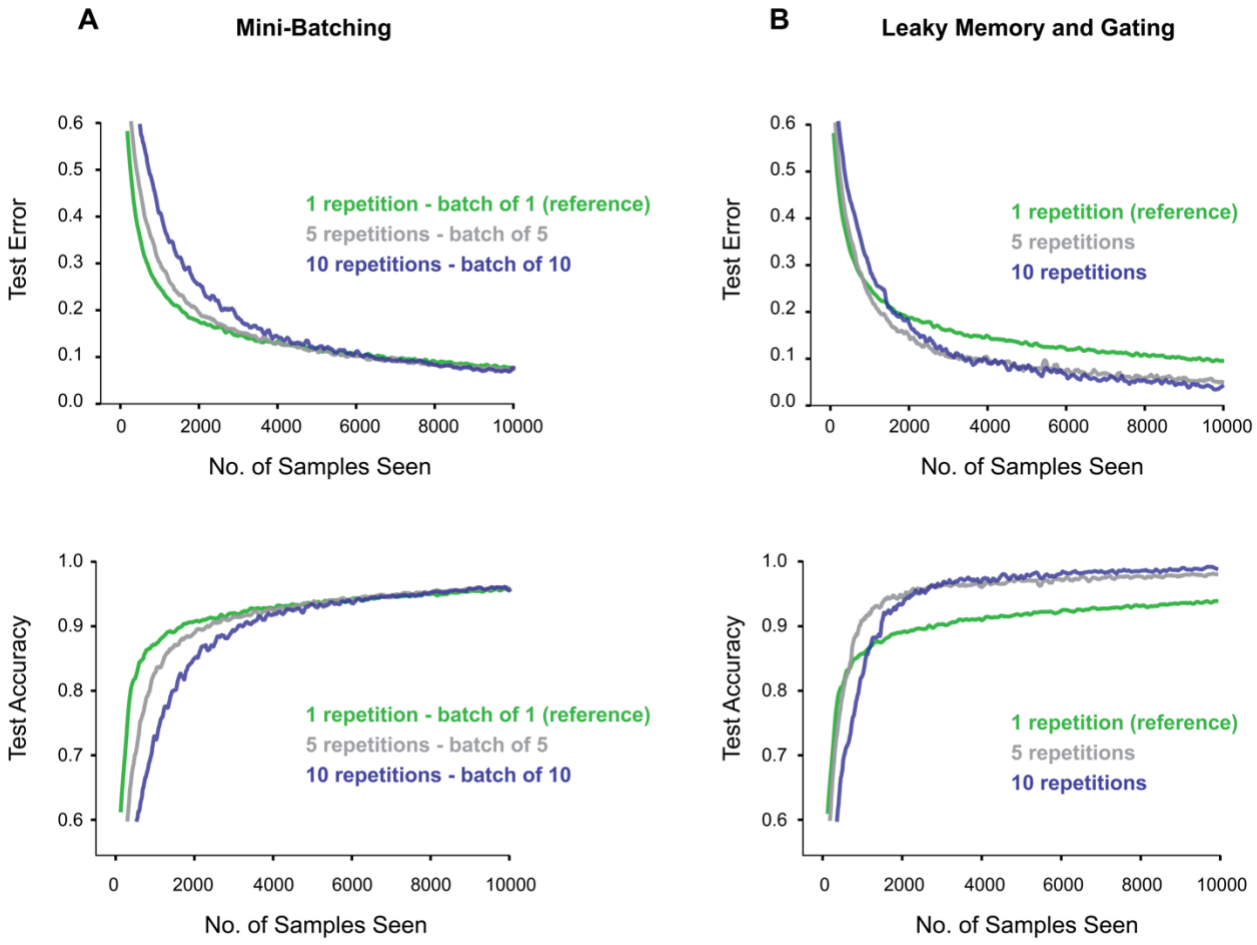


Figure 2-9. Comparing the leaky memory approach against mini-batching.
Mini-batching and the model with memory and gating models were affected in a qualitatively different manner by increasing the number of category repetitions. Left) Test error and test accuracy of mini-batch training on MNIST data. Right) Test error and test accuracy of the model with leaky memory and gating on MNIST data. Both models had the dimension of (784, 392, 10), learning rate of 0.01, and the optimization method of SGD.

2.4. Discussion and Conclusions

In this Chapter, we examined the efficacy of two brain-inspired mechanisms for incremental learning from temporally structured data. These mechanisms were motivated by the fact that (i) our brain dynamics appear to blur information over time, as there is ubiquitous autocorrelation in cortical dynamics, while (ii) neural circuits' sometimes shift their state suddenly at "event boundaries", and this appears to be associated with "resetting" of context representations. Therefore, we introduced

temporal integration into our models via the leaky memory mechanism and we introduced temporal separation via a memory gating mechanism. We hypothesized that, if there is a slowly-varying temporal structure in real-world training data (Dong and Atick 1995), then these leaky memory and memory gating mechanisms would enable neural networks to exploit temporal autocorrelation in real-world information streams, without using backpropagation through time (BPTT).

We found that, when training data were temporally autocorrelated, this slowed category learning in feedforward neural networks, but it accelerated learning in systems with leaky memory (**Figure 2-2**). Moreover, adding a simple memory gating mechanism to the leaky memory networks enabled them to flexibly adapt to the autocorrelation level, so that they could benefit from repeating structure while not suffering from the interference of unrelated prior information.

While prior work has focused on the implications of slow neural dynamics for online comprehension (Uri Hasson, Chen, and Honey 2015) as well as working memory and decision making (Soltani et al. 2021), here we emphasized the implications of slow intrinsic dynamics for learning. At first glance, the presence of ubiquitous autocorrelation presents a problem for learning representations of individual items in the world: the information about one stimulus is mixed with information about stimuli that came before, and it is unclear how to resolve this interference without propagating gradients through time. However, here we demonstrated a beneficial effect: when the properties of the training data themselves are autocorrelated in time, then an intrinsic autocorrelation in neural dynamics can provide an inductive bias which accelerates learning. Moreover, some of the unwanted interference between unrelated items can be reduced if neural circuits are additionally equipped with a resetting mechanism.

We highlight two main directions for future extension of these results. First, it would be interesting to examine the efficacy of leaky memory and gating mechanisms in more sophisticated category-learning architectures such as convolutional neural networks trained to perform video segmentation (Wang and Hu 2021). Second, future work should test whether these results generalize to larger architectures and more realistic datasets, and should include a broader search of the hyperparameter space. We expect that the present results do have some generality, because we used simple architectures, made few domain-specific assumptions, and demonstrated the phenomena across multiple datasets. We expect the method to work best for datasets in which important or diagnostic data features persist over consecutive samples.

In sum, in Chapter 2, we identified simple brain-inspired mechanisms which enabled neural networks to learn categories more efficiently when the training data are temporally correlated, and to do so without propagating gradients in time.

Chapter 3: Effects of Temporal Integration and Separation on Learning Multi-Timescale Representations

3.1. Background and Motivation

Our results from Chapter 2 indicate that brain-like integration mechanism accompanied by separation through boundary-resetting can enable a neural network to learn more rapidly from temporally correlated data. However, in the real world, we may need to learn from data with multiple levels of autocorrelation (Dong and Atick 1995; O'Doherty and Cusack 2022). For instance, when having a face-to-face conversation: the features around a person's mouth change quickly, while their face's outline changes more slowly (**Figure 3-1** A). Moreover, there are no pre-defined labels to support the learning of representations in this setting. What are the consequences of learning representations from information with multiple timescales? How can a learning mechanism with brain-like architectural constraints exploit multiple timescales of data to learn multi-timescale representations?

Evidence from research in cortical circuits suggests that our neural processes can integrate information at multiple time-scales (Bernacchia et al. 2011; Ulanovsky et al. 2004; Honey et al. 2012; Murray et al. 2014; Bright et al. 2020). Circuits in the sensory cortex exhibit rapid state-shifts, and they do so more frequently than circuits in the higher-order association cortex (DuBrow and Davachi 2016; Chien and Honey 2020). Assuming that these state-shifts reflect information-separation through boundary-resetting processes, how does the frequency of resetting affect the representations that are learned by each circuit?

We propose that temporal integration mechanisms augmented by memory-gating ability may be an ingredient of how neural circuits can separately represent fast- and slow-changing features in its input. Additionally, we proposed that these mechanisms can work even in setting such as an everyday conversation, where there are no pre-defined category labels, and representations must be learned without supervision.

In this Chapter, we use the same brain-inspired integration and separation mechanisms that was used in Chapter 2: leaky memory and memory gating (boundary-resetting). However, instead of focusing on the efficacy of category learning, we investigate how the representations that are learned (without supervision) are altered when boundary-resetting occurs more or less frequently. In particular, in this Chapter, we test the hypothesis that more infrequently resetting circuits will come to learn representations of more slowly varying features of the sensory environment. This hypothesis was informally proposed by Honey et al. (Christopher J. Honey, Ehren L. Newman 2017), but has not been quantitatively tested.

3.1.1. Research Questions

It remains unclear how neural circuits (without employing BPTT) can separately represent quickly changing features (e.g., a speaking mouth) and slowly changing feature (e.g., head position), when input data vary on both fast and slow timescales. In this project we addressed the following questions:

- When a neural network learns to represent features of multi-scale training data, how are these represented altered when the neural networks are equipped with brain-inspired integration and multiscale boundary-resetting mechanisms?

- Will the subsets of a neural network that reset less frequently come to represent slow-varying features of the data?

Significance

Answering these questions would improve our understanding of how architectural constraints in neural circuits may bias the nature of their internal representations, when learning from information with multiple timescales. We hypothesized that neural networks equipped with multi-scale (i.e., fast and slow) leaky memory and boundary-resetting could exploit multiscale autocorrelation in data to learn more "meaningful" representations, by separating their representations of structures that vary on fast and slow timescales.

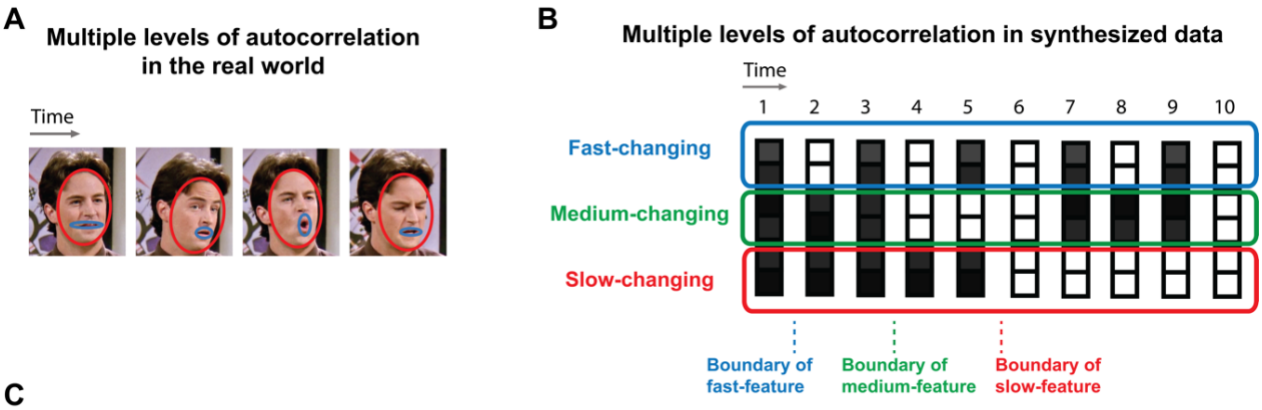
3.2. Methods

3.2.1. Multi-Timescale Autocorrelation in Training Data

Dataset. To test the ability of our networks to learn temporal representations, we synthesized simplified training datasets in which we could control the temporal autocorrelation structure. The input to the model at each time point consisted of 3, and each subcomponent had two elements. Each subcomponent was generated to express a different level of autocorrelation over time: for example, subcomponents changed feature-category every 1, 3, and 5 iterations, respectively (**Figure 3-1** B). The individual features sampled at each time were generated as the sum of (i) an underlying binary state variable (which would switch every 1, 3 or 5 iterations for "fast", "medium" and "slow" scales respectively) and (ii) uniformly distributed noise. As a result, the model was provided with features that varied at 3 timescales: fast, medium, and slow (**Figure 3-1** B).

3.2.2. Autoencoder Architectures with Brain-Inspired Constraints

Brain-inspired constraints. We used the same brain-inspired mechanisms outline in Chapter 2, here, for unsupervised learning models: local linear recurrence and gating mechanisms.



Autoencoder Models	Memory Coefficients
Feedforward (memoryless)	$\alpha_1 = \alpha_2 = \alpha_3 = 0$
Uniform Leaky Memory (LM)	$\alpha_1 = \alpha_2 = \alpha_3 \neq 0$
Multiscale LM	$\alpha_1 < \alpha_2 < \alpha_3$
Unifrom LM with Multiscale Gating	$\alpha_1 = \alpha_2 = \alpha_3 \neq 0$
Multiscale LM with Multiscale Gating	$\alpha_1 < \alpha_2 < \alpha_3$

Figure 3-1. Unsupervised learning from data with autocorrelation on multiple timescales.
A) Example of multiple levels of autocorrelation in samples from the real world: the mouth shape changes more quickly than face-shape, and so has shorter autocorrelation time. B) Multiple levels of autocorrelation in synthesized data: first 2 elements change every item, representing the fast-changing feature; second two elements change every 3 items, representing a medium-changing feature; and third two elements change every 5 items, representing a slow-changing feature. X-axis shows time, each 6-by-1 item is one sample. C) 5 Different AE models: feedforward, uniform leaky memory (LM), multiscale LM, uniform LM with multiscale gating, and multiscale LM with multiscale gating. α1, α2, and α3 indicate the memory coefficient in the hidden representations (see Eq. 1).

Local Linear Recurrence (leaky memory): We implemented two types of linear recurrence: uniform and multiscale. In uniform linear recurrence, the time constant was constant across the nodes in the hidden layer. In multiscale linear recurrence, we varied the time constants across the nodes in the hidden layer, gradually increasing α from node 1 to node 3 (**Figure 3-1** C).

Multiscale Memory Gating: We used a multi-timescale gating mechanism that was sensitive to the three levels of temporal structure in the input stream. For each timescale in the input, the gating mechanism would use information from current and previous input to decide to reset memory when the change passed a threshold (see Eq. 2) (Chien and Honey 2020). Therefore, for the data stream in **Figure 3-1** B, the internal representation in node 1 reset when there was a shift in the fast-changing property of the input stream; the internal representation in node 2 reset when there was a shift in the medium-changing property of the input stream; and, the internal representation in node 3 reset when there was a shift in the slow-changing property of the input stream.

Architectures. We used the same brain-inspired mechanisms for unsupervised learning models: leaky memory and gating mechanisms. To evaluate the effectiveness of the added mechanisms, we compared 5 types of autoencoder (AE) models (See **Figure 3-1** C): i) Feedforward AE; ii) AE with leaky memory in internal representations; iii) AE with multi-scale leaky memory in internal representations, inspired by evidence showing that levels of processing in the brain can integrate information at different time-scales (Honey et al. 2012; Murray et al. 2014; Bright et al. 2020) and that multiple time-scales are present even within a single circuit (Bernacchia et al. 2011; Ulanovsky et al. 2004) ; iv) AE with leaky memory in internal representations and boundary-sensitive gating, motivated by the evidence showing that processing in cortical circuits are sensitive to event-boundaries and these boundaries can shift learned representations (DuBrow and Davachi 2016; Chien and Honey 2020); and (v) AE with multi-scale leaky memory in internal representations and boundary-sensitive gating. We used a multi-timescale gating mechanism that was sensitive to multi-timescale changes in the input stream. For each timescale in the input, the gating mechanism would use information from current and previous input to decide to reset memory when the change passed a threshold based on the Eq. (2). Here, we used the comparison between the difference and the average of the following input items

as the resetting criterion, but other sorts of computations are also possible. Our implemented method is consistent with neurophysiological studies that demonstrate a sudden shift in memory representations in the face of a surprise in the input stimuli (DuBrow and Davachi 2016; Chien and Honey 2020). This bio-plausible boundary-detection mechanism resets the memory when the difference between the consecutive inputs is larger than their average. For instance, the memory of the hidden node with long memory will be reset based on the amount of change in the slow-changing feature of the input. In Eq.(2), t represents the iteration number during training, I_t is the current state, and I_{t-1} is the previous state.

$$| I_t - I_{t-1}| > | (I_t + I_{t-1})/2 | \qquad \text{Eq. (2)}$$

3.2.3. Training Autoencoder Networks

Learning algorithm, optimization, and initialization. Similar to the classification task studied in Chapter 2, we used backpropagation to train the networks, but again we ensured that the gradient computation did not account for the fact that the neurons were leaky. In other words, we trained the leaky-memory networks as if they were feedforward networks. We used MSE loss, both with and without RMSprop optimization method, and Xavier initialization (Tieleman and Hinton 2012; Glorot and Bengio 2010). We applied ReLU and Sigmoid as activation functions for hidden and output units, respectively.

Hyperparameters. The networks were 3-layer, fully connected autoencoders with (6, 3, 6) dimension. Learning rate was 0.005. In cases where RMSprop was implemented, the beta-1 and beta-2 were set to 0.9 and 0.99. For linear recurrence in internal representations, we tested a range of memory coefficients (α in Eq. 1). For uniform linear recurrence, we tested 3 sets of memory coefficients:

model1: $\alpha 1 = \alpha 2 = \alpha 3 = 0.3$; model2: $\alpha 1 = \alpha 2 = \alpha 3 = 0.6$; model3: $\alpha 1 = \alpha 2 = \alpha 3 = 0.9$. For multiscale linear recurrence, we tested the following combination of memory coefficients: Model 1: $\alpha 1 = 0$, $\alpha 2 = 0.3$, $\alpha 3 = 0.6$; Model 2: $\alpha 1 = 0$, $\alpha 2 = 0.3$, $\alpha 3 = 0.9$; Model 3: $\alpha 1 = 0$, $\alpha 2 = 0.6$, $\alpha 3 = 0.9$.

3.2.4. Evaluation Methods for Autoencoder Networks

Measure of multiscale representation-learning. We measured the network's ability to "un-mix" the timescales of its input. By un-mixing, we mean learning representations that selectively track distinct latent sources that generated features within each training sample. In particular, we tested whether different nodes in the internal representations of the network would track different temporal features of the data stream: one node would track the fast-changing features of the data, a second node would track the medium-changing features of the data, and the third node would track the slow-changing features of the data.

To this end, using squared of Pearson correlation, we measured the similarity between the internal representations (node 1, node 2, node 3) and the data features (fast, medium, and slowly changing) (Figure 3-3 A). In the models with multi-timescale linear recurrence, node 1 has no memory, node 2 has short memory, and node 3 has long memory. For AE models that do not have a multiscale memory, when referring to no-memory, short-memory, and long-memory nodes, we are referring to the hidden nodes in those positions (Figure 3-3 A).

In the similarity matrix, the diagonal elements show the similarity of hidden representation to their corresponding timescale of data feature (similarity of fast-feature and no-memory hidden node, similarity of medium-feature and shot-memory hidden node, and similarity of slow-feature and long-memory node). Off-diagonal elements of similarity matrix show the similarity of hidden representation

to their non-matching timescale of data feature (e.g., similarity of fast-feature and long-memory hidden node) (Figure 3-3 A and B).

We then quantified the "timescale-selectivity" — e.g., whether the slow-changing feature was more correlated with long-memory node than other nodes (no-memory and short-memory nodes) (Figure 3-3 B). We therefore calculated timescale-selectivity as the average of the difference between the on-diagonal and the off-diagonal elements of the similarity matrix (Figure 3-3 B).

Measure of learning efficiency. Learning efficiency was measured using the reconstruction error of the test data, computed as the MSE across all 3 subcomponents of each data sample.

3.3. Results

We first confirmed that all the AE models could learn to reconstruct the input. The most efficient architectures (architectures with lowest test error) were the networks with recurrence and multiscale gating AE, and the feedforward memoryless AE (**Figure 3-2**).

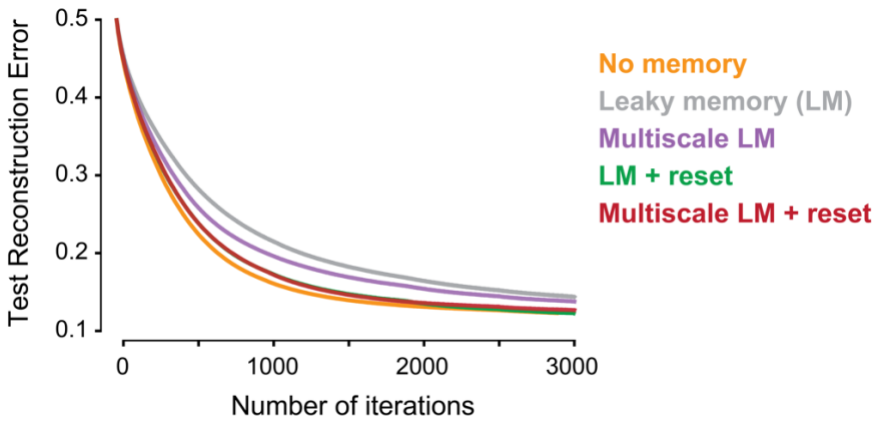


Figure 3-2. Reconstruction test error (MSE loss, for individual items) during training across 5 different AE models.
All the curves in this plot have been averaged over 50 runs with different random initialization. (The [LM + reset] model, shown in green, mostly overlaps with [multiscale LM + reset] model.)

Networks equipped with both recurrence and multiscale gating could successfully un-mix the data, learning separated representations for the quickly and slowly changing features within the data stream (See "AE with Memory and Multiscale Gating" in **Figure 3-3** C). The individual hidden state units in these AE models were selectively more correlated with their corresponding data features (e.g., the slow-changing feature was more correlated with the long-memory node than with the other nodes; **Figure 3-3** C). These findings generalized across different synthesized datasets, different learning rates, and different recurrence (memory) coefficients.

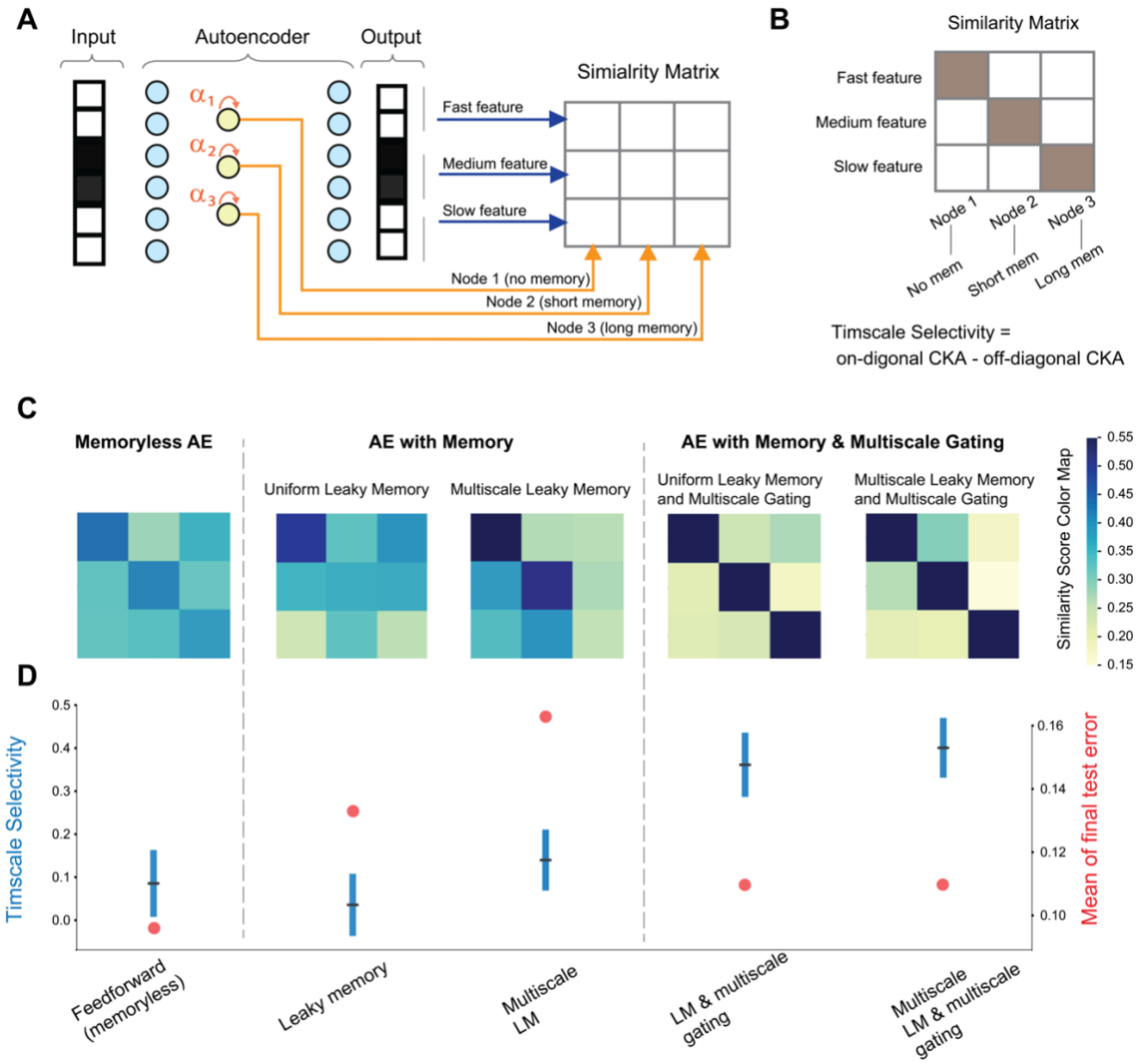


Figure 3-3. Quantifying the similarity between internal representations and data features that vary on fast, medium, and slow timescales.
A) Illustration of which network-units and which training-data elements contribute to each element of the similarity matrix. For instance, the top-left element of the matrix shows the similarity between the fast feature of the data and the no-memory internal representation (node 1). B) Illustration of how timescale selectivity is computed. C) Similarity matrices for 3 groups of models: feedforward AE, AE with memory, and AE with memory and multiscale gating. D) Timescale selectivity and final test error for different AE models. Blue bars show the timescale selectivity. Each bar shows the 95% confidence interval for results from 40 runs with different random initializations. Red dots show the mean final test error across 40 runs. The memory coefficients of the models shown in (C) and (D) are as follows: leaky memory: 0.6; Multiscale LM: 0, 0.3, 0.9; LM & multiscale gating: 0.6; Multiscale LM and multiscale gating: 0, 0.3, 0.9.

We found that models with recurrence and multiscale gating showed significantly higher timescale selectivity (**Figure 3-3** D). Overall, the AE models with recurrence showed higher loss compared to feedforward AE. However, in the models that had both recurrence and multiscale gating, the small increase in the loss was accompanied by a significant increase in timescale-selectivity scores (**Figure 3-3** D).

Why might recurrence and gating mechanisms benefit the learning of timescale-separated representations?

The two autoencoder models that had both recurrence and multiscale gating mechanisms were most successful in learning internal representations that tracked distinct timescales of the input. Slowly (or quickly) varying features were extracted by slowly (or quickly) varying subsets of the network, analogous to a matched filter (see also (Mozer et al. 1992)). Features that change on different timescales may correspond to different levels of structure in the world (Wiskott and Sejnowski 2002). Thus, by adding recurrence and multiscale gating to a simple feedforward AE model, we equipped it with an ability to separate different levels of structure in the environment. Moreover, because intrinsic dynamics vary on multiple scales in the human brain (Honey et al. 2012; S. M. Smith et al. 2013; Raut, Snyder, and Raichle 2020; Murray et al. 2014) this implies that slowly-varying brain circuits may be biased to extract slowly-varying structure from the world (Honey, Newman, and Schapiro 2017).

Why did the feedforward (memoryless) model produce slightly lower reconstruction error than models with recurrence?

There is a cost to using AE models with recurrence (rather than feedforward AE models) to reconstruct data that varies on multiple timescales: the overall test error increases slightly. This most likely occurs because slowly changing internal states are ineffective for reconstructing quickly changing

features. In other words, we attribute this (small) additional error to be problem of reconstructing input from an internal representation that is mismatched to the data. Consistent with this interpretation, we found that if a model's "slow" hidden units (i.e., its medium and long memory units) were correlated with the fast-changing features in the training data, the model's per-feature error was worse (**Figure 3-4**). This slight increase in reconstruction error is, however, accompanied by a significant benefit: learning more temporally interpretable, un-mixed representations of a multi-scale data stream.

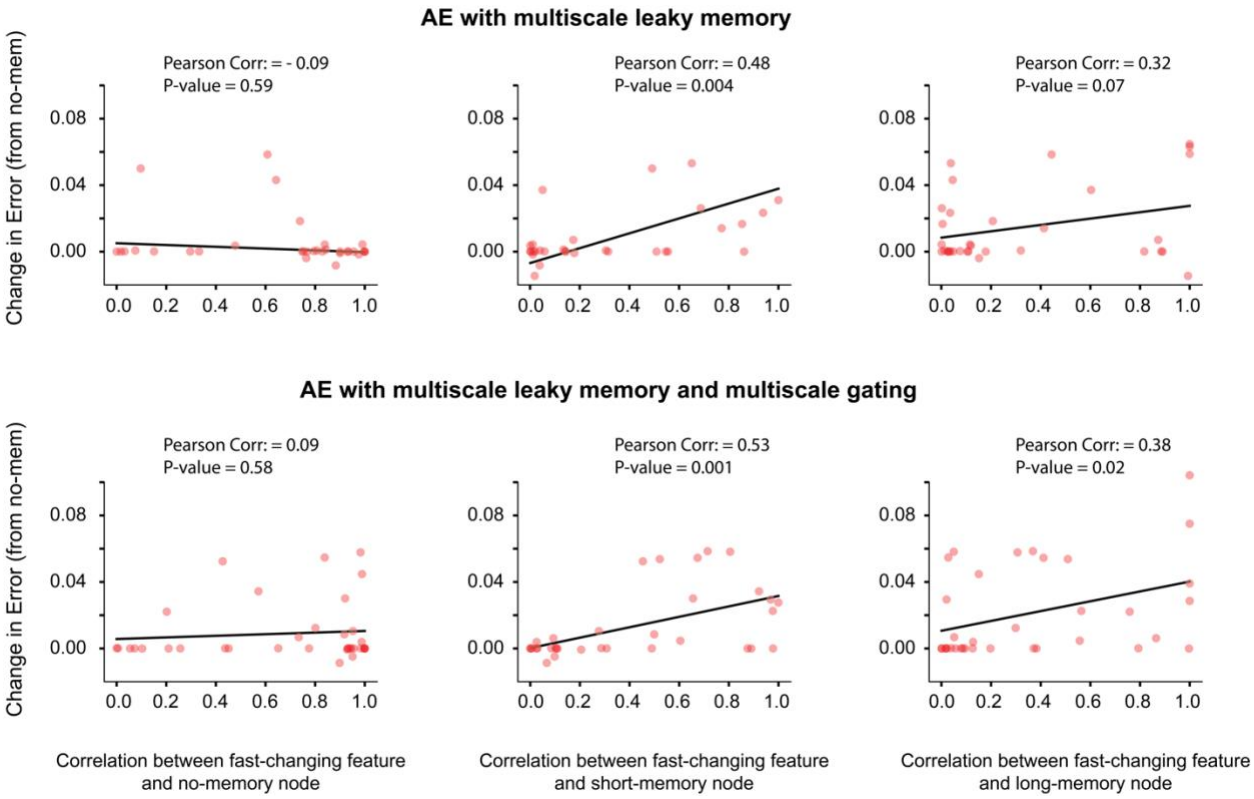


Figure 3-4. Relationship between the correlation of hidden nodes with fast output, and the reconstruction error for fast output.
Top) [Multiscale leaky memory] model: y-axis shows the difference between test reconstruction error for fast feature (e.g. error of reconstructing fast feature in [multiscale leaky memory] model - error of reconstructing fast feature in [no memory] model); x-axis in the left plot shows the correlation of fast output with no-memory hidden nodes, x-axis in the middle plot shows the correlation of fast output with short-memory hidden nodes, x-axis in the right plot shows the correlation of fast output with long-memory hidden nodes. Bottom) Plots show similar results shown at the top row for [multiscale leaky memory + reset] model.

How does the magnitude of memory coefficients interact with efficiency and timescale selectivity of the AE models?

Although equipping AE models with recurrence and gating enabled them to learn temporally interpretable representations, it also slightly increased their overall test error (**Figure 3-2**, **Figure 3-3** D). Indeed, in all AE models, the greater the memory coefficients (i.e., linear self-recurrence), the greater the test error (**Figure 3-5** B). Especially when models lacked memory gating, the higher memory coefficients were associated with inferior reconstruction accuracy and reduced timescale selectivity and lower efficiency (higher test error). But when the AE models had multiscale gating in addition to recurrence, there was only a slight reduction in learning efficiency, and this was accompanied by much greater timescale selectivity (**Figure 3-5** B). Thus, in practical settings, the memory coefficients may be adjusted to trade off timescale selectivity (when we want to learn separated representations for distinct sources) and test error (when we wish to optimize the accuracy of reconstruction of momentary input).

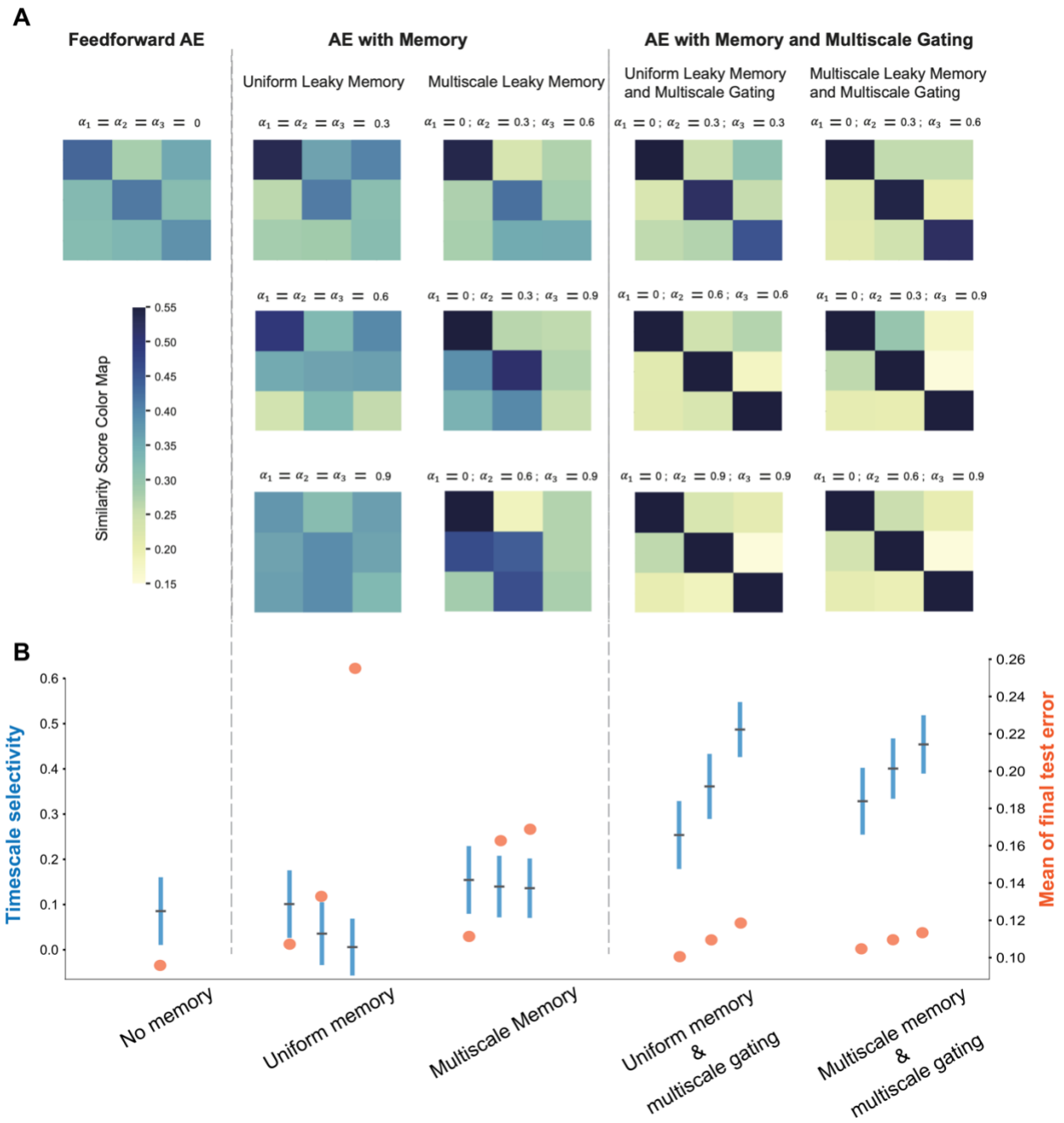


Figure 3-5. Timescale selectivity and final test error for different AE models.
A) Results of similarity matrices for 3 groups of models: feedforward AE, AE with memory, and AE with memory and multiscale gating. Each similarity matrix shows the similarity pattern for a model with a specific set of memory coefficients shown above it. B) The colored bars show the timescale selectivity. Error bars show the 95% confidence interval for 40 runs with different ransom initializations. The gray dots show the average of final test error for 40 runs. For each group of models (e.g., uniform memory, etc.), we tested 3 models with 3 sets of memory coefficients. From left to right, memory coefficients increase gradually. For instance, for uniform memory, memory coeffects of all the internal nodes in the models left model were 0.3, for the middle model were 0.6, and for the right model were 0.9.

3.4. Discussion and Conclusions

In this Chapter, we explored the efficacy of brain-inspired mechanisms to enhance unsupervised learning from data streams that contain both fast and slow temporal structure. Inspired by the observation that different areas of cortex display different rate of state-shifts, we implemented a multiscale memory resetting mechanism: circuits in sensory cortex display frequent, rapid state-shifts, which occur more often than those in the higher-order association cortex. We then asked whether more infrequently resetting circuits would be biased to represent the more slow-varying features of the multi-timescale data.

We found that AE models equipped with recurrence and multi-timescale gating could successfully learn to reconstruct the input, and, moreover, that they learned representations which selectively separated multiple timescales of the data stream (**Figure 3-3**). Specifically, we found that when individual nodes in the network were imbued with slow dynamics (and infrequent resetting), these nodes became more correlated with data features that varied slowly across training examples. If distinct timescales in the data stream reflect distinct data generators, then these neural network nodes can be understood as "un-mixed" representations may provide a more "meaningful" description of the input data (Mitchell 2020; Mahto et al. 2020; Jain et al. 2020).

The local linear recurrent networks we tested exhibited more interpretable internal representations, even though they were trained with a learning rule that did not employ any temporal information. Specifically, all networks were trained incrementally using backpropagation with a loss function that only depended on the immediate state of the network. Architectures with recurrence and gating can thus exploit temporal structure in a way that is computationally simpler and more biologically plausible than backpropagation through time (Lillicrap et al. 2020; Ilya Sutskever 2013).

With respect to biological plausibility, we note that the linear-recurrence-plus-gating system worked well for autoencoder, for which there are simple activation-based learning rules that do not require the propagation of partial derivatives (Lee et al. 2015). Therefore, it is possible that our model could be extended to a setting without any backpropagation at all.

On the computational side, we emphasize that the gradients computed for the recurrent networks were, in a sense "inaccurate", because the update rule was unaware of the recurrent leak connections. Nonetheless, the overall classification accuracy in the recurrent networks was higher than in feedforward nets, for which the gradients should be more accurate. This indicates that training a recurrent neural network as if it is a feedforward network can be effective, and that backpropagation through time may not always be necessary, depending on the dataset.

We propose two main directions for future extension of these results. First, our autoencoder models were trained to reconstruct the present stimuli. Future work could extend our autoencoder architecture to a predictive architecture so that the model would predict the next stimulus rather than the present one (Arora and Cai 2021). Second, it would be interesting to investigate the broader consequences of learning in the context of recurrent internal states: for example, human internal representations of natural sensory input sequences appear to be smooth in time, in contrast to the representations of most feedforward nets (Hénaff, Goris, and Simoncelli 2019). Training neural networks with smooth data and an internal leaky memory could potentially reduce the discrepancy between the representational spaces acquired by artificial neural networks and the "smooth" representational spaces learned by humans.

In sum, in this Chapter, we identified simple mechanisms which enabled neural networks to generate internal representations that separated distinct timescales of the data, without propagating gradients in time.

Chapter 4: Determining How Human Language Processing is Affected by Interruptions

4.1. Background and Motivations

In Chapters 2 and 3, we explored learning biologically-plausible integration mechanisms in neural networks, comparing their effectiveness when it comes to incremental learning from a sequence of examples with different levels of autocorrelation. In this Chapter 4 and the next, we shift our focus away from learning, and toward online comprehension and prediction of linguistic sequences. Specifically, we characterize the behavior of humans and neural language models when processing linguistic data with interrupting incongruencies.

Experiencing moments of incongruencies while performing a task is an inevitable part of our daily experience. For instance, while performing a primary task of reading a research article, we may be distracted by an email notification and quickly read the email, before returning to our primary task. Here, we refer to these incongruent portions of the temporal stream as "interruptions".

As noted in the Introduction, the literature suggests that when humans encounter interruptions while reading, they will be slower when resuming the reading after interruptions, and this resumption lag is longer for longer interruptions (Foroughi et al. 2015; Foroughi, Barragán, and Boehm-Davis 2016; Glanzer, Fischer, and Dorfman 1984). Additionally, some studies have also shown that interruptions can disrupt participants' deeper textual understanding while leaving their superficial comprehension and memory intact (Foroughi et al. 2015). However, some studies posit that interruptions do not disrupt our comprehension due to LTWM capabilities of our cognitive processes (Ericsson and Kintsch 1995). Prior research also focused on the similarity of the

interruption to the primary task and concluded that if reading interrupting content that is similar to the content of the primary reading task, the disruption is more severe (Ledoux and Gordon 2006).

Despite prior research, inconsistencies among the results make it difficult to draw connections between the findings of various studies. First, the empirical findings regarding the disruptive effects of interruptions are inconsistent because of the use of varying designs and interruption tasks, as well as different criteria for measuring them. Next, the proposed WM-based and LTWM-based models that were proposed to explain the observed behavioral data are quite abstract and do not necessarily lead to different predictions. Finally, it is still uncertain whether the disruption caused by having more similar content would also extend to the similarity of the cognitive processes involved in the main task and the interruptions. To address this, in this Chapter, we revisit the question of effects of interruptions on reading using various manipulations in one paradigm and a large sample of participants in each condition.

4.1.1. Research Questions

It remains unclear when reading a narrative: (i) how processing incongruencies with low or high similarity to the reading task (similarity in terms of cognitive processing and in terms of information content) modulates how humans process the narrative; (ii) how text processing is influenced after different types of incongruencies; (ii) how our subsequent memory of information is affected by different types of incongruencies. Hence, in this project, we set out to address the following three research questions.

While reading a narrative,

i. How does encountering different types of interruptions influence our ongoing processing?

ii. How does encountering different types of interruptions modulate our memory of pre-interruption information?

iii. How does encountering different types of interruptions modulate our memory of post-interruption information?

Significance

Our goal is to characterize the kinds of interruptions that produce larger and smaller effects on (1) ongoing processing difficulty in reading and on (2) subsequent memory. Thus, in this Chapter, using behavioral experiments, (1) we operationalize processing difficulty using reading times and comprehension metrics to probe the effects of these interruptions on the ongoing processing of the information; (2) we operationalize subsequent memory using delayed recognition memory of information that precede and follow different types of interruptions.

In the short term, such data can be practically useful by providing some general guidelines for minimizing the deleterious effects of interruption in real-world settings. In the longer term, these data, all collected within a common paradigm, may constrain more concrete process-level models that explicitly characterize WM and LTM updating and interference processes happening at each moment during discursive reading (Budiu and Anderson 2004).

4.2. Research Design and Methods

4.2.1. Primary Task

For the main stimuli, we used a story because it has rich temporal properties and can have short-persistent and long-persistent features. "So Much Water So Close to Home" is an interesting short story (~2300 words), by Raymond Carver, detailing the tensions between a woman and her husband,

who she suspects may have committed a murder. The reading will be self-paced. Participants will read the, story sentence by sentence, with each individual sentence presented centrally. When participants press a key to proceed, the current sentence is erased from the screen and replaced by the next.

4.2.2. Experimental Conditions

When reading the narrative sentence by sentence, participants occasionally encountered interruption trials during which they are asked to perform a secondary task. In a between-subjects design, each participant saw only one type of interruption, and that type of interruption depended on the experimental condition to which they are assigned (**Figure 4-1**).

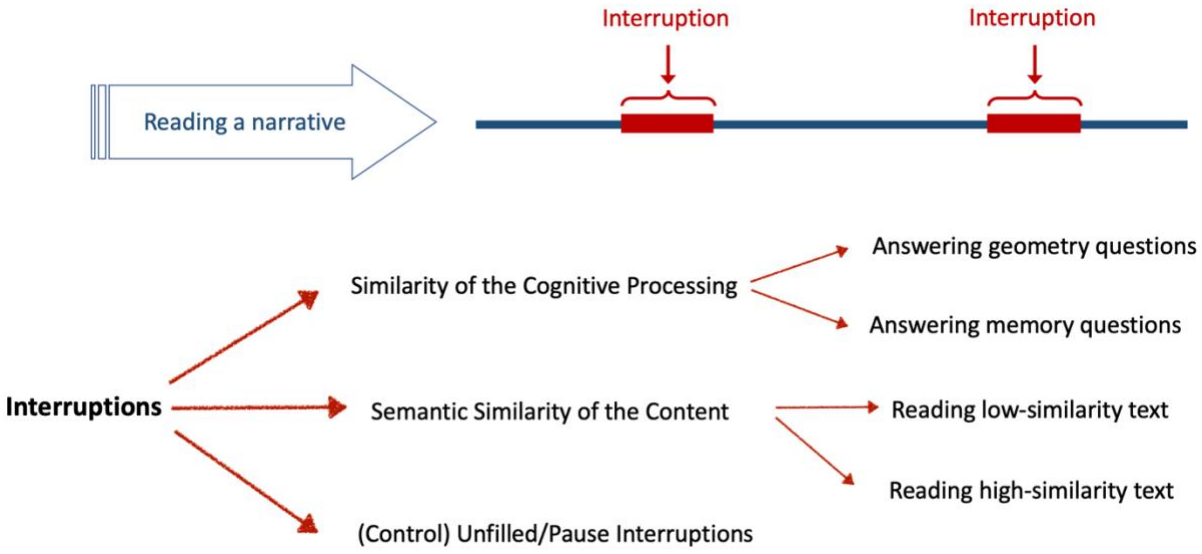


Figure 4-1. Schematic of the information flow in the interrupted reading task.
Top: schematic of encountering interruptions during reading a narrative. Bottom: types of interruptions that we use in this experiment.

In this experiment, we manipulated 2 independent variables: the type of interruptions similarity (*similarity of cognitive processing* and *similarity of content*) and the level of similarity (*low* or *high*). We also had three additional control conditions. First, there was a Pause interruption condition in which the interruptions were not filled with a secondary task, but participants simply waited. Second, we also tested a condition in which participants were not given any information about encountering interruptions. We refer to this condition as un-signaled interruptions. The un-signaled interruption condition can also be considered as one of the sub-conditions that tested the effects of similarity of content. Third, there was a continuous (no-interruption) condition in which participants read the entire story uninterrupted, providing a baseline for the reading time and memory performance in the absence of interruption (**Figure 4-2**).

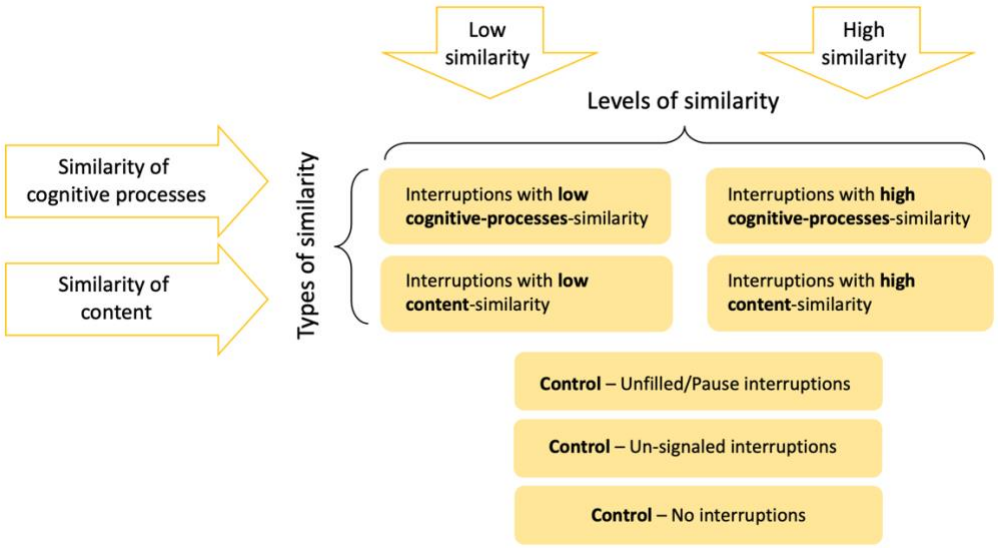


Figure 4-2. Conditions that we study in the behavioral experiment.

Interruption Duration and Frequency.

Duration. All interruption conditions testing the similarity of cognitive processes had a fixed duration and were between 18 to 20 sec long. Interruption conditions testing the effects of similarity of content were either timed (had a fixed duration, 18-20 sec long), or varied depending on participants' reading speed (interruptions involving self-paced reading).

Frequency. The interruption trials appeared after reading every 11 to 13 sentences of the story. Each participant saw a total of 17 interruptions. The location of these 17 interruptions within the narrative was shared across all participants.

4.2.2.1. Interruptions with Low/High Similarity of Cognitive Processing

Participants were randomly assigned to one of the cognitive-similarity interruption conditions and each only experienced one type of interruption. This led to a between-subjects design.

Geometry interruptions (Low similarity of cognitive processing). Participants were presented with an image (an example shown in **Figure 4-3**) and asked to count the number of triangles or squares it contained. They were provided 18 seconds to complete the task. Remaining time was indicated by a horizontal time-bar that shrank in size. After 18 seconds, a two-alternative forced choice (2 AFC) question appeared, asking if the image had exactly X number of the asked shape. Participants had to choose between 'Yes' or 'No'. Following a 300 ms delay, they were presented with the next sentence from the story, and then could resume their reading.

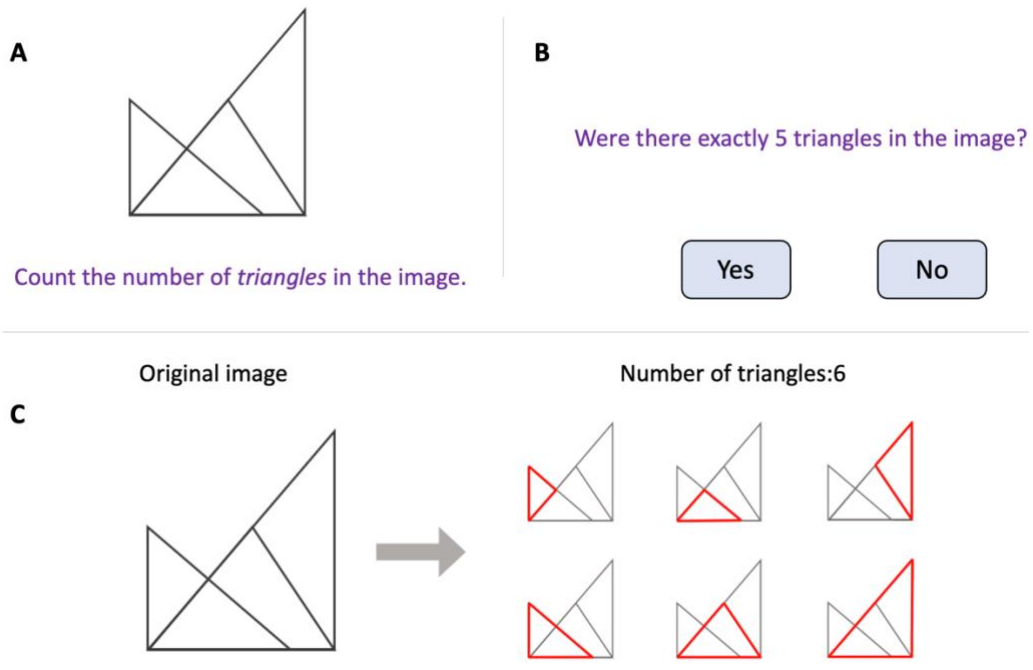


Figure 4-3. Schematic of a Geometry interruption trial.
A) An example of a Geometry trial as it appears in the experiment. B) An example 2 AFC question that follows the interruption trial. C) Demonstration of the correct way of answering the interruption question.

Scenario interruptions (High similarity of cognitive processing). For Scenario interruptions, participants had 18 seconds (shown by the same horizontal time bar) to read a brief scenario (~2-4 sentences) sentence by sentence. Next, they were asked to answer a question about that scenario. These scenarios and questions were chosen from Dodell-Feder et al. (Dodell-Feder et al. 2011), a study in which these scenarios were used to localize brain regions involved in theory of mind.

We divided the Scenario conditions into two sub-conditions: the first sub-condition tested situational comprehension with a 2AFC question, and the second subsection tested verbal working memory with a 2AFC question. Examples of these two types of question are shown below. These two sub-conditions enabled us to ask how different kinds of engagement with the same interrupting material (i.e., situational comprehension versus surface-level reading of the same text) may alter memory and speed in the primary reading task.

An example scenario:

"When Lisa left Jacob, he was deep asleep on the beach. A few minutes later a wave woke him. Seeing Lisa was gone, Jacob decided to go swimming."

I. An example comprehension memory question
    - In this subsection participants were asked a situational or a theory-of-mind question about the scenario.
        - Example for situational question: Did Lisa Wake Jacob?
        - Example for theory-of-mind question: Does Lisa now believe that Jacob is awake?

II. An example verbal working memory question
    - In this subsection participants were asked a verbatim question about the scenario.
        - Example for verbatim question: In the preceding segment, did you see the word "surfing"?

Pause interruptions (Control). Participants in the Pause interruption trial were instructed to wait for ~20 seconds, as indicated by a horizontal time bar that shrinks over time. Once the time bar reached zero, a "Next" button appeared on the screen. The participants clicked this button to resume reading the story. This button click was implemented so that a motor response was required at the end of each Pause interruption trial, similar to the Geometry and Memory interruptions. Additionally, to ensure that the total interruption length in the Pause interruptions was the same as Geometry and Memory interruptions, the *fixed* duration of Pause interruptions was slightly (~2 s) longer than those of Geometry and Memory interruptions (18 s). This adjustment was introduced because, after Pause

interruptions, participants only had to click a Next button, while after Geometry and Memory interruptions, they had to answer a 2AFC question.

Experiment Instructions

Participants were told to read a story sentence by sentence. They were also informed about the Geometry, Memory, or Pause trials that they would encounter during reading. As part of the instructions, they were shown a demonstration trial similar to the interruption task that would take place in the actual experiment. They were asked to make their best effort to answer the questions in the Geometry and Memory trials correctly. However, no explicit information was provided to participants regarding which task - the story or the intermediate trials - was the main task; they were instructed to pay attention to everything that they read and saw.

4.2.2.2. Interruptions with Low/High Similarity of Content

We employed a within-subject design in which each participant would randomly encounter both high-similarity and low-similarity interruption trials while reading the primary narrative. For each participant, the positions of high- and low-similarity interruptions were randomly assigned such that they would see a combination of either 9 high-similarity and 8 low-similarity interruptions or 8 high-similarity and 9 low-similarity interruptions. Then for each interruption position, the interrupting text was randomly selected from a pool of 5 possible options at that position.

To identify interrupting text with low or high semantic similarity to the story, we randomly picked several unrelated sentences from the corpus. We then computed the semantic similarity between the interrupting text and the text from the story preceding the interruption. To calculate similarity scores, we generated embeddings of the interrupting text and the preceding text, and we

then calculated the cosine similarity score between the embeddings. To obtain text embeddings, we used the Universal Sentence Encoder (USE) (Cer et al. 2018). The implemented encoder was the Transformer USE model, which encodes sentences into embedding vectors.

We created a pool of interpretation content such that for each of the 17 interruption-positions in the story, we had 5 semantically similar text and 5 semantically dissimilar text to choose from in the interruption pool. To create the interruption pool, the segments of interrupting text (~ 2-4 sentences) were chosen randomly from the Brown corpus, which contains many categories of text (e.g., fiction, scientific articles, books, and news articles, etc.).

- Sample sentences from the main story:
    - "The others stirred the sand with their shoes, said they didn't feel inclined that way. They pleaded fatigue, the late hour, the fact that the girl wasn't going anywhere."
- A sample interruption content with low similarity to the preceding story sentences:
    - "Much of the available information comes not from the Federal government but from an exchange of experiences among states. Proposals State and local agencies in the vocational education field must be encouraged to adopt a wider outlook on future job opportunities."
- A sample interruption content with high similarity to the preceding story sentences:
    - "At last they concluded that the heavy, full feeling in their stomachs was due to lack of exercise. Walking was the remedy, they decided, but a deck full of chicken coops and pigpens was hardly suitable. Skipping was the alternative."

Although the main goal of this condition was to assess the influence of semantic similarity, we also wanted to know whether this effect was modulated by whether the interrupting material was (i)

timed (so that participants could not control its duration) and (ii) signaled (so that participants knew that they were about to encounter interrupting material). Therefore, independently of the other experimental factors, we divided the semantic-similarity interruption condition into three sub-conditions: (1) timed interruptions; (2) self-paced signaled interruptions; and (3) self-paced un-signaled interruptions (control). Each participant was assigned to one of these three sub-conditions.

Timed semantic interruptions. In timed semantic interruptions, participants had 18 seconds to read the interrupting text. The remained time was shown by a shrinking time bar. Once the duration was over, in 4 out of 17 interruptions, participants were asked to answer 2AFC questions about whether they saw a particular word in the interruption (verbal working memory question). The occasional verbatim questions were asked to ensure that participants read the interrupting content. In the rest of the interruptions (13 out of 17), at the end of the interruption duration (18 s) they would be presented with the rest of the story.

Self-paced signaled semantic interruptions. In the self-paced signaled interruptions, participants read the interrupting text self-paced, one sentence at a time, similar to reading the rest of the story. To signal the participants about encountering an interruption trial, the interruption text was written in a different format (purple and italic) than the primary text (black). Identical to the timed semantic interruptions, 4 out of 17 interruptions were followed by a 2 AFC verbal working memory questions.

Self-paced un-signaled semantic interruptions (Control). We also tested a self-paced un-signaled interrupted reading condition. In this condition, participants were not given any information about the story and the occasional unrelated text, so that it was possible they might now be aware that they were processing interrupting material at all. When presented with instructions for this task, they were

simply asked to read some text, sentence by sentence, and alerted that their memory would later be tested on what they read. The story and the interruptions all were written in the same format (black).

This condition enables us to separate out which interruption effects arise from the interruption trials being marked as a clearly distinct task. Moreover, the data from this condition are especially appropriate for comparison between human interruptions (this Chapter) and language model interruptions (Chapter 5). Since the language models we test in Chapter 5 are not provided with any meta-cognitive cues about task states or the occurrence of interruptions, the un-signaled interruption condition enables us to measure human reading times under conditions more like those faced by the language model, for which interruptions are also "un-signaled".

Experiment Instructions.

The participants were told that they would read a story sentence by sentence. In the timed condition and the self-paced signaled condition, they were informed about encountering occasional unrelated text. They were told to read all the text - the main story and the occasional unrelated text - carefully as their memory would later be tested on everything they read. In the self-paced un-signaled interruptions condition, they were not given any information about the story and the occasional unrelated text. They were only instructed to read some text carefully because their memory would later be tested on everything they read.

4.2.3. Dependent Variables

We measured two main dependent variables: 1) reading time and 2) delayed recognition memory (**Figure 4-4**). Reading time was used to operationalize effects of interruptions on processing difficulty. Delayed recognition memory was used to operationalized effects of interruptions on memory encoding and consolidation.

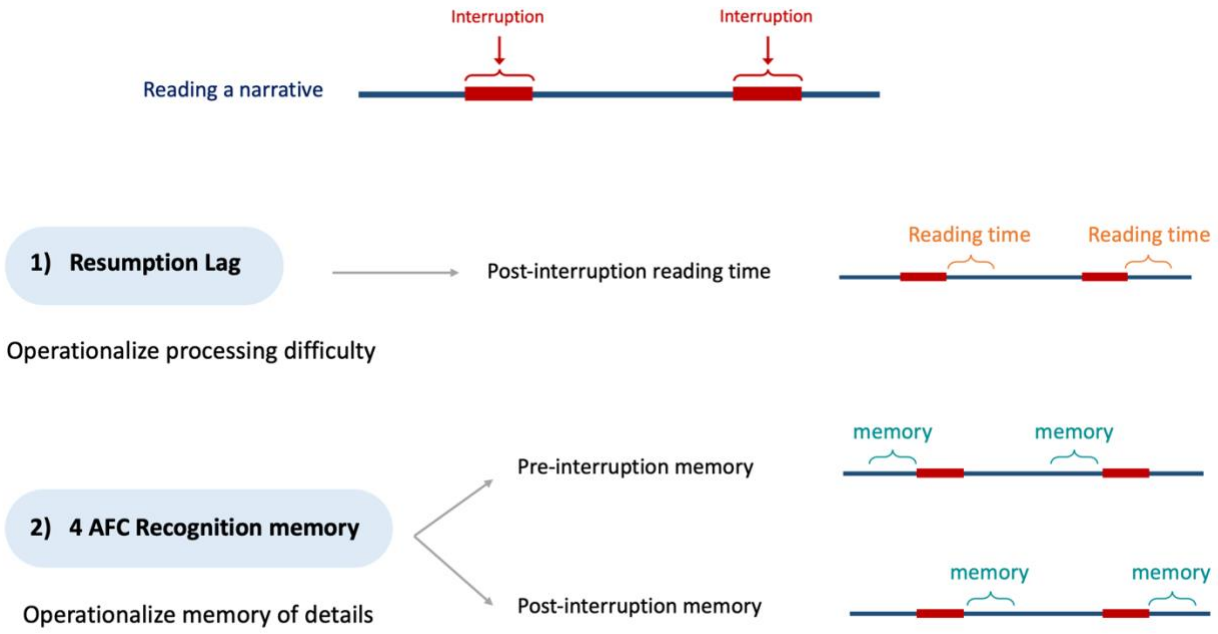


Figure 4-4. Schematic illustrating the two main dependent variables collected in this study.
1) Resumption lag: we measured the change in the reading time for the first sentence following the interruptions relative to sentences preceding the interruptions. 2) We tested recognition memory regarding the verbal content of sentences in the story that immediately preceded or immediately followed the interruptions.

Recognition memory was tested after participants had completed reading the entire story. On each memory trial, we showed a section from the story with one highlighted sentence. Within the highlighted sentence, a single word or short phrase was missing. Participants were given 4 option to select the word or phrase that they saw the in the actual story. The 4 options in the recognition memory test were balanced in piloting data to ensure that they were approximately exchangeable (i.e., they were equally likely to be chosen by a participant who had not read the story).

- An example recognition memory question:

  "My husband eats with a good appetite. But I don't think he's really hungry. He chews, arms on the table, and stares at something across the room. He looks at me and looks away. He wipes his mouth on the napkin. He shrugs, and goes on eating. "What are you staring at me

for?" he says. "What is it?" he says and lays down his fork. "Was I staring?" I say, and shake my head. The telephone rings. "Don't answer it" he says. "It might be your mother," I say. "Watch and see" he says. I pick up the receiver and listen. My husband _____."

- ❑ stops eating
- ❑ stops chewing
- ❑ goes on eating
- ❑ goes on chewing

4.2.4. Participants

We recruited participants online through the Prolific platform. Participants were 18 to 65 years old, native English speakers, and all located in United States. A total of 1176 participants were recruited across all conditions. Demographic data of our participants revealed that 526 identified as female, 624 as male, and 26 as non-binary or with unknown gender.

Exclusion criteria. To ensure that our participants remained attentive throughout the experiment, we excluded non-attentive participants by measuring their reading time in correlation to the length of the sentence. We assumed that if participants were attentively reading the story as instructed, there should be a positive correlation between their reading time and the length of the sentence. We excluded any participants who displayed a correlation less than 0.2 between their reading time and the sentence length, excluding sentences immediately following interruptions. The length of a sentence was measured as the number of characters it contained.

4.2.5. Statistical Analyses

To investigate the statistical significance of the results, we used permutation test as described below.

Testing whether the means of two conditions are different. The null hypothesis is that the data points in each condition are drawn from identical distributions, and therefore, the difference between their means is zero. For comparing the difference between two groups with *n1* and *n2* number of data points, we first calculated the true difference between the group means. We then combined the data points from both groups into a pool of all data points. We then repeated 10,000 sampling of *n1* examples from the pool without replacement. For each of 10,000 sampling, we calculated the mean of the *n1* examples as the average of group 1. Using the mean of pool of combined data, we calculated the mean of group 2, for each of 10,000 repetitions. We then calculated the difference between the mean of group 1 and mean of group 2, for each sampling repetition. This gave us a distribution of 10,000 differences between the mean of group1 and mean of group 2. We then compared the true difference between the mean of group 1 and the mean of group 2 against the distribution of 10,000 values for difference of the means. Finally, p-value was calculated as the frequency of true difference of the means being lower than the values in the distribution, divided by 10,000. Here, p-value represents the probability of obtaining the observed difference between the means (true difference), assuming the null hypothesis is true.

4.3. Results

4.3.1. Reading Time

All interruptions (Geometry, Pause, verbal WM, and situation comprehension) led to a significant increase in reading time following the interruptions. **Figure 4-5** shows the change in RT per character for the sentence following the interruptions relative to the sentences preceding the interruptions,

normalized by subtracting the change in RT per character in intact (uninterrupted) condition. To gain a sense of what RT per character means in terms of the actual time, an average sentence has 50 characters in it. Hence, a 20-unit increase in RT per character would mean 1 second increase in the reading time of an average sentence (20*50 / 1000 (ms) = 1s).

We did not observe a difference between two sub-categories of memory interruptions: Scenarios plus verbal working memory questions, and Scenarios plus comprehension questions. Therefore, for the statistical analysis in this section, I refer to the combination of two sub-categories of memory interruptions as Scenario interruptions. We found that change in RT per character for Scenario interruptions (verbal WM and comprehension combined) was significantly higher than those in Geometry interruptions (permutation test for difference between Scenario and Geometry, Δ(RT per character) = 5.57, p-value = 0.05). Change in RT per character for Pause interruptions was higher than Geometry interruptions but the difference did not reach the significance level (Δ(RT per character) = 2.27, p-value = 0.19). Scenario interruptions also showed a higher increase in RT per character compared to Pause interruptions, but there was not a significant difference between Pause and Scenario interruptions (Δ(RT per character) = 1.03, p-value = 0.384).

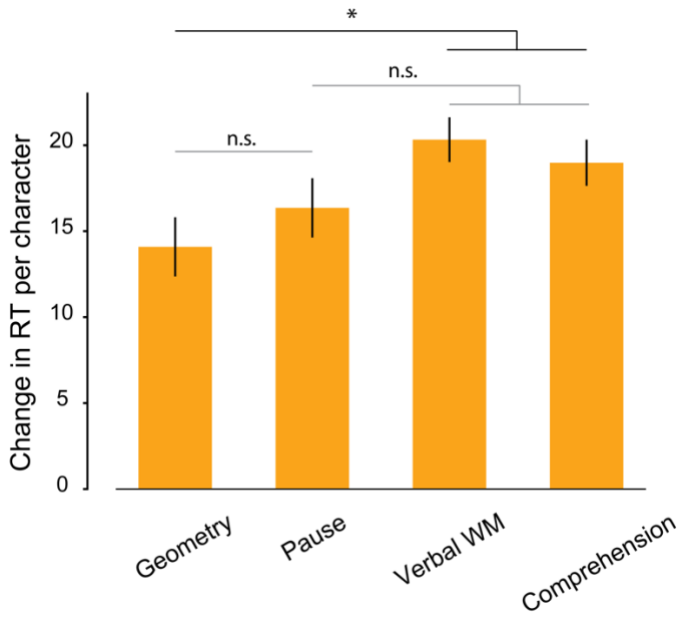


Figure 4-5. Change in RT per character for the first sentence following the interruptions with high and low similarity of cognitive processes to the reading task.
For each interruption, we calculated the change in RT per character by calculating the difference between the RT per character of the first sentence subtracted by the average of RT per character of a few sentences preceding that interruption. Error bars show standard error of the mean. (n.s. = not statistically significantly different (p-value>0.05); * = p-value < 0.05)

We did not find any difference in the change in RT per character following low-similarity versus high-similarity interruptions. There was no reliable difference in any of the 3 sub-conditions of the semantic-similarity interruptions (**Figure 4-6**) (e.g. in Timed sub-condition, for low- vs. high-similarity: (Δ(RT per character) = 3.56, p-value = 0.258).

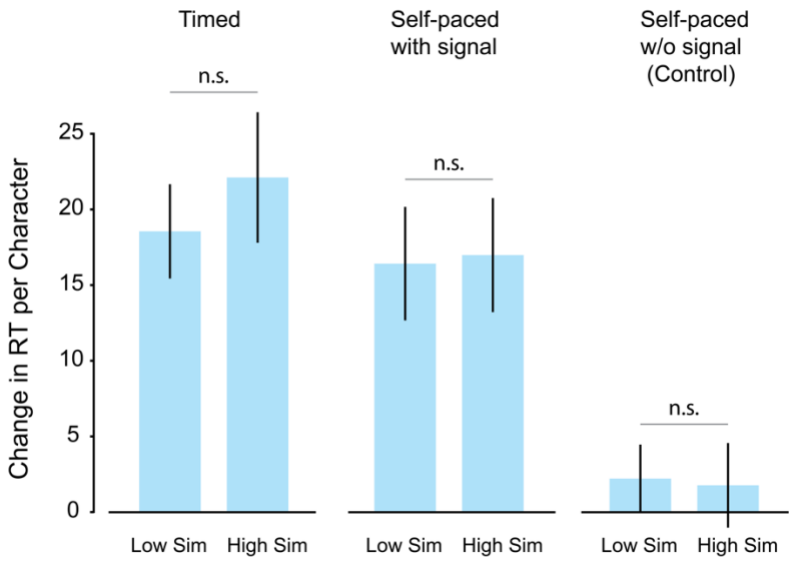


Figure 4-6. Change in RT per character for the first sentence following the interruptions with high and low similarity of content to the reading task.
Results shows the change in RT per character in 3 sub-conditions (Timed, Self-paced w signal, Self-paced w/o signal) testing the effects of content similarity. Error bars show standard error of the mean. (n.s. = not statistically significantly different (p-value>0.05))

We also compared the change in RT per character for content-similarity interruptions to the cognitive-similarity interruptions (**Figure 4-7**). The overall change in RT per character for timed and self-paced-signaled (**Figure 4-7**, right) was comparable to those in Scenario interruptions (**Figure 4-7**, left) interruptions (mean RT per character for Self-Paced with Signal = 15.5, mean RT per character for Verbal WM = 20.3). However, the self-paced interruptions without signal were clearly different from other conditions, as they exhibited negligible change in RT per character after the interruptions which was not different from zero (e.g., bootstrap test, Δ(RT per character for low-similarity interruptions in Self-Paced w/o Signal) = 2.2, p-value = 0.84).

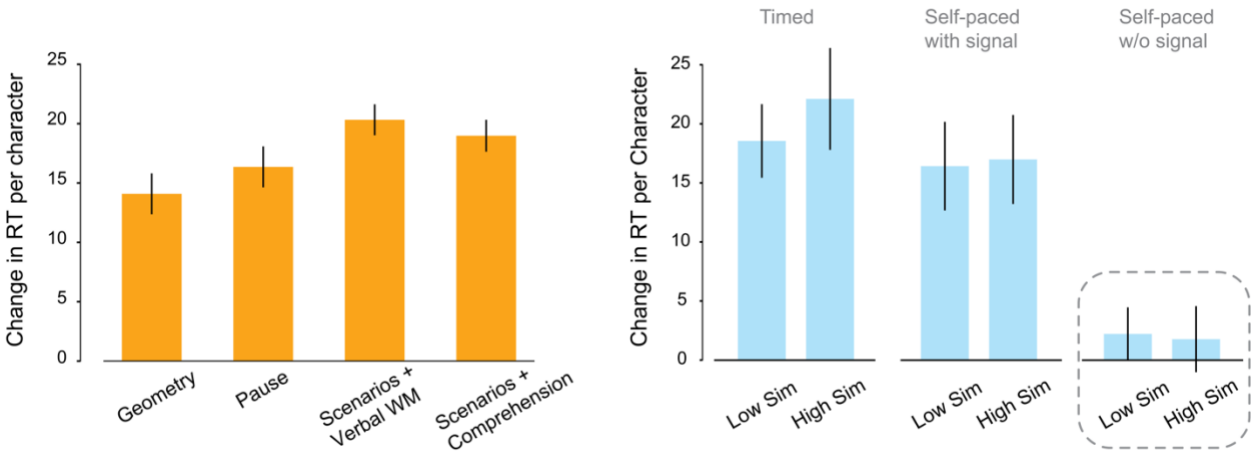


Figure 4-7. Comparing the change in RT per character following interruptions in cognitive-similarity conditions and content-similarity conditions.
Left) Change in RT per character for the first post-interruption sentence in interruption conditions for testing the similarity of cognitive processing. Right) Change in RT per character for first post-interruption sentence in interruption conditions for testing the similarity of interruption content.

We were surprised that we observed near-zero resumption effect in the reaction times when the interruption was un-signaled, and therefore checked whether participants were sensitive to the content of the interruption at all. To do so, we examined the reading times *during* (rather than after) the interruption (**Figure 4-8**). We found that in both of the self-paced interruption conditions, there was a significantly increased RT per character for low-similarity interruptions compared to high-similarity interruptions (Δ(RT per character for low- versus high-similarity in signaled condition) = 36.68, p-value < 0.001; (Δ(RT per character for low- versus high-similarity in un-signaled condition) = 29.93, p-value < 0.001).

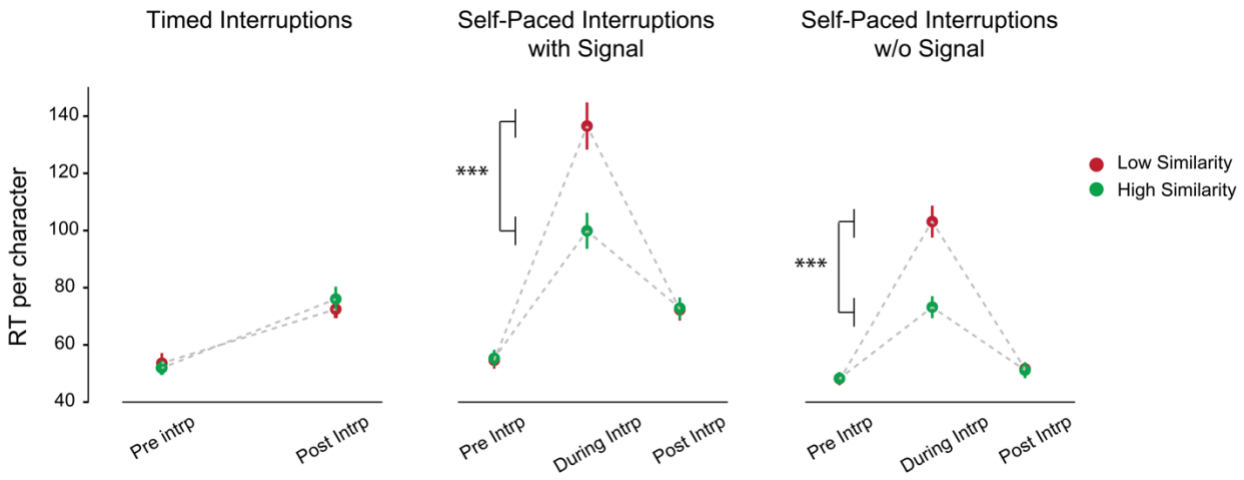


Figure 4-8. RT per character before, during, and after interruptions in content-similarity interruptions. Left) RT per character before and after the interruptions for timed interruptions. Middle) RT per character before, during, and after the interruptions for self-paced interruptions with signal. Right) RT per character before, during, and after the interruptions for self-paced un-signaled interruptions. Error bars are bootstrapped standard errors of the mean. (*** = p-value < 0.001)

4.3.2. Recognition Memory

We calculated how memory was altered by different types of interruption, by computing the change in 4AFC recognition performance relative to the Intact (uninterrupted) condition. In the Intact condition, the accuracy of information at pre-interruption positions was 49.3%, and the accuracy of information at post-interruption positions was 51.7%. The change in memory accuracy was computed both for information in the sentences that preceded an interruption and in the sentences that followed an interruption (**Figure 4-9**).

Pause interruptions significantly increased memory accuracy of preceding interruptions (permutation test for Pause vs. Intact: ΔAcc = 5.8%, p-value = 0.001). Recognition memory of information following verbal WM interruptions was also enhanced (permutation test for Verbal WM vs. Intact: ΔAcc = 3.3%, p-value = 0.003). Conversely, Geometry interruptions numerically decreased recognition memory for the sentences that preceded an interruption (permutation test for Geometry

vs. Intact: ΔAcc = -2.7%, p-value = 0.11), though the effect was not statistically significant. We did not observe a reliable difference in pre-interruption recognition memory when participants encountered comprehension interruptions (permutation test for Comprehension vs. Intact: ΔAcc = 0.6%, p-value = 0.31). Pre-interruption recognition memory in verbal WM interruptions was significantly higher than comprehension interruptions (permutation test, ΔAcc = 2.7%, p-value = 0.033). The difference between Pause and verbal WM conditions was not statistically significant (permutation test, ΔAcc = 2.5%, p-value = 0.14). In terms of effects of similarity of cognitive processes on recognition memory, we found that memory of pre-interruption information for Geometry interruptions (low cognitive similarity) was significantly lower than Scenario interruptions (high cognitive similarity) (permutation test, ΔAcc = 6.8%, p-value < 0.001).

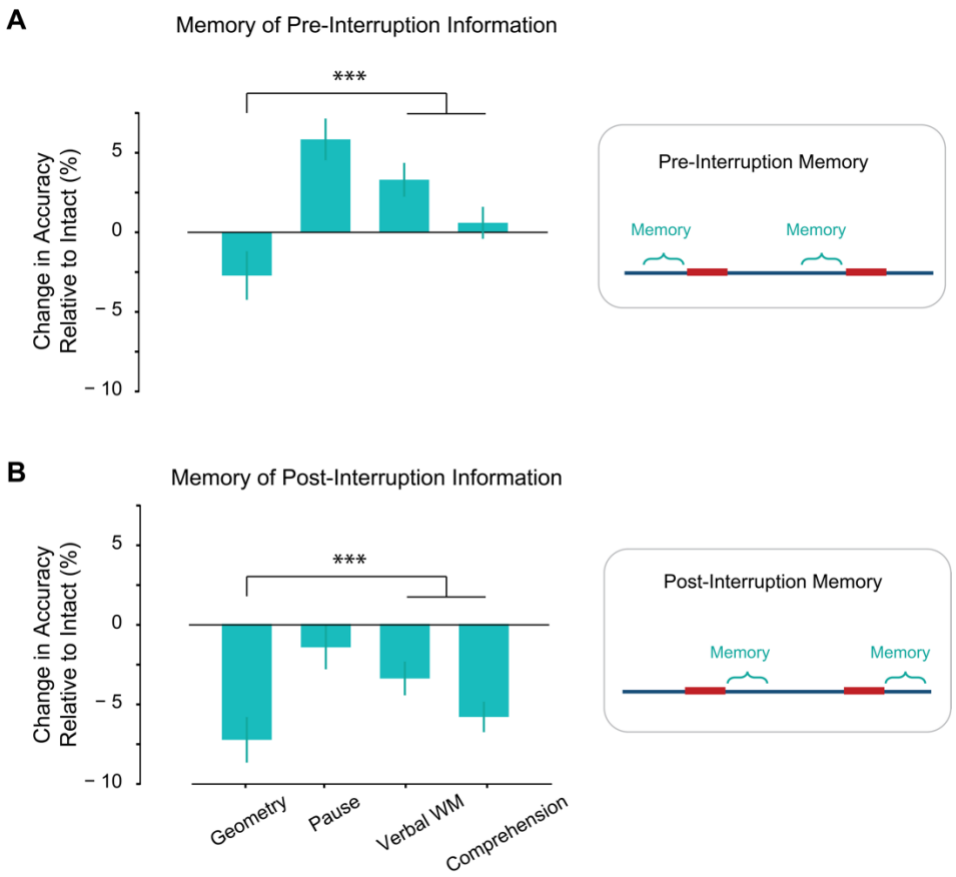


Figure 4-9. Change in recognition memory of information pre- and post-interruptions relative to intact, in cognitive-similarity interruption conditions.
Change in recognition memory for pre- and post-interruption information in conditions testing the effects of cognitive similarity of the interruptions relative to the intact (uninterrupted) condition. A) Change in pre-interruption memory. B) Change in post-interruption memory. (*** = p-value < 0.001)

We next examined the effects on memory for sentences that followed an interruption. Recognition memory of information following interruptions decreased relative to the Intact baseline in Geometry and Scenario interruptions but not in Pause interruptions (Permutation test: Pause vs. Intact: ΔAcc = -1.5%, p = 0.23; Geometry vs. Intact: ΔAcc = -7.2%, p < 0.001; Verbal WM vs Intact: ΔAcc = -3.4%, p = 0.002; Comprehension vs. Intact: ΔAcc = -5.8%, p < 0.001). Post-interruption recognition memory in comprehension interruptions was lower than verbal WM, but the effect was only marginally significant (ΔAcc = 2.4%, p-value = 0.049). Post-interruption recognition memory

was numerically lower in Geometry than comprehension interruptions (ΔAcc = 1.4%, p-value = 0.09) and verbal WM interruptions (ΔAcc = 3.8%, p-value = 0.066). Finally, we found that memory of post-interruption information for Geometry interruptions (low cognitive similarity) was significantly lower than Scenario interruptions (high cognitive similarity) (permutation test, ΔAcc = 2.7%, p-value < 0.001).

Neither the high- nor low-similarity interruptions appeared to substantively affect recognition memory of information preceding the interruptions (absolute changes in 2AFC accuracy all < 2% (**Figure 4-10** A). However, both the high- and low-similarity interruptions worsened recognition memory for material following the interruptions (**Figure 4-10** B). These reductions in memory were similar for high- and low-similarity interruptions, except in the control un-signaled condition, where we observed marginally larger memory-loss in the low-similarity condition (permutation test, low-vs-high similarity, ΔAcc = -4.6%, p-value = 0.06).

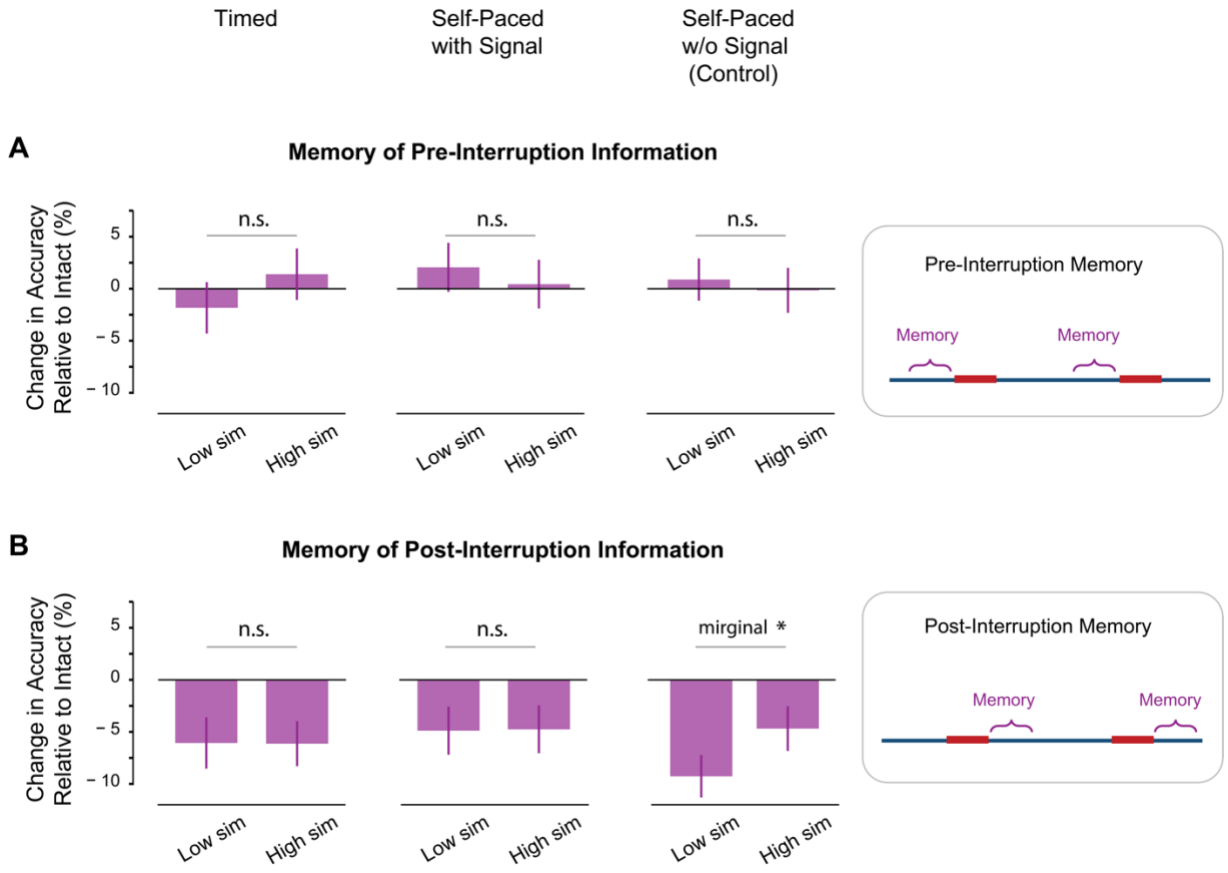


Figure 4-10. Change in recognition memory of information pre- and post-interruptions relative to intact, in content-similarity interruption conditions.
Change in recognition memory of pre-interruption (A) and post-interruption (B) information for low- and high-similarity of interruptions content. Error bars show standard error of the mean. (n.s. = not statistically significantly different (p-value>0.05); * = p-value < 0.05)

Finally, we compared the pre- and post-interruption memory accuracy for the content-similarity interruptions (high and low similarity reading) to the cognitive-similarity interruptions (Geometry, Scenarios, and so forth) (**Figure 4-11**). The change in recognition memory for content-similarity interruptions were analogous to those of comprehension interruptions: no difference in pre-interruption memory and significant decline in post-interruption memory. We did not find any difference between recognition memory data of the 3 sub-conditions of content-similarity

interruptions (e.g., ΔAcc(between self-paced conditions with versus without signal) = 2.4%, p-value = 0.185).

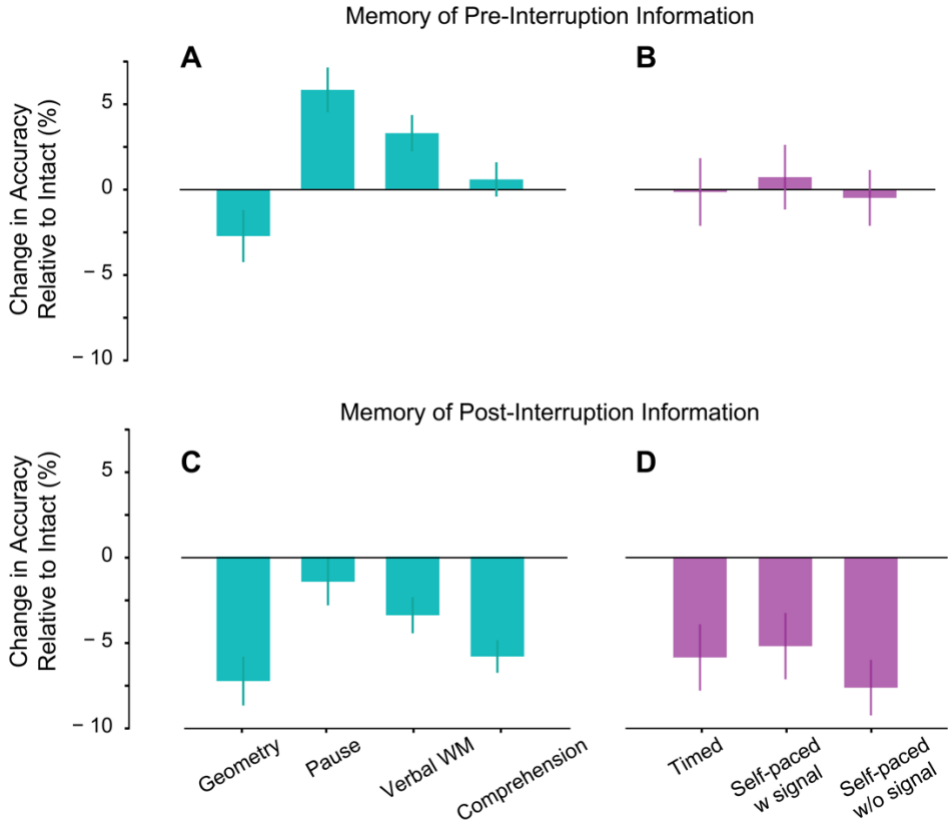


Figure 4-11. Comparing the overall change in recognition memory between conditions testing the effects of cognitive similarity and conditions testing the effects of content similarity.
(A and C) show the recognition memory from cognitive-similarity interruption conditions. (B and D) show the recognition memory from content-similarity interruption conditions.

4.3.3. Additional Analyses

4.3.3.1. Accuracy in Interruption Trials

Interruptions questions were designed to match in their level of difficulty. To that end, we had tested the interruption trials, when not inserted within the story. We balanced the difficulty level such that accuracy in Geometry questions, verbal working memory questions, and comprehension questions

were 87% (± 0.3 %). When the interruptions were inserted inside the story, we found the accuracy of Geometry interruptions dropped significantly (p-value < 0.001). On the other hand, the accuracy in comprehension interruptions increase significantly (p-value < 0.001). There was not a difference between the accuracy in verbal WM questions when tested alone and when inserted within the story (p-value = 0.24) (**Figure 4-12**).

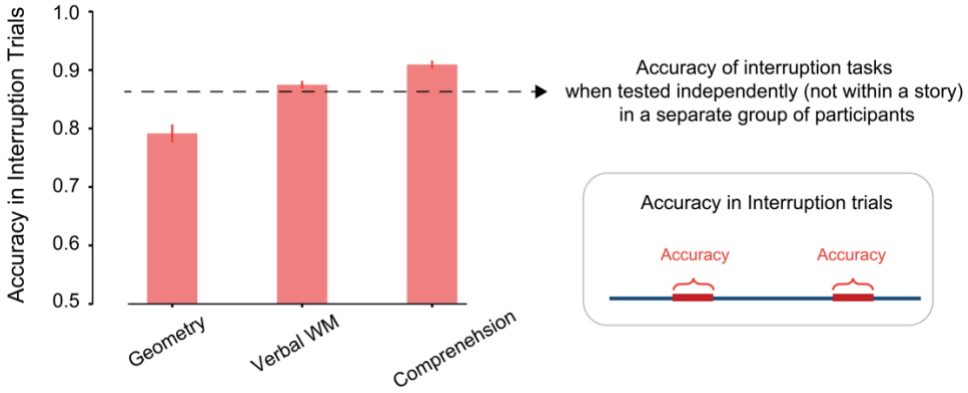


Figure 4-12. Accuracy in interruption trials for Geometry, verbal WM, and comprehension interruptions. Dashed line (87% accuracy) shows the baseline accuracy in interruption trials when tested alone, not within the story. Chance level is at 50% (2 AFC questions). Error bars show standard error of the mean.

4.3.3.2. Transportation and Lingering Scores

We used Green and Brock's transportation questionnaire to measure participants' immersion level while reading the story (Green & Brock Narrative Transportation Scale, 2000). Participants in Pause interruptions reported a slightly higher transportation compared to intact condition. Also, Geometry and Scenario interruptions showed lower transportation scores. Although there was a numerical difference, we did not find a significant difference between the transportation score of different types of cognitive-similarity interruptions compared to intact (e.g. Δ(Transportation for Intact versus Geometry) = 0.11, p-value = 0.16) (**Figure 4-13** A). The biggest difference in transportation score

among different types of cognitive-similarity interruptions was observed between Pause and Geometry interruptions (Δ(Transportation) = 0.19, p-value = 0.051).

Transportation score in content-similarity interruptions were significantly lower than intact (e.g. (Δ(Transportation for intact versus self-paced with signal) = 0.25, p-value = 0.019). There was not any difference between the transportation scores among the 3 sub-conditions of content-similarity interruptions (**Figure 4-13** B).

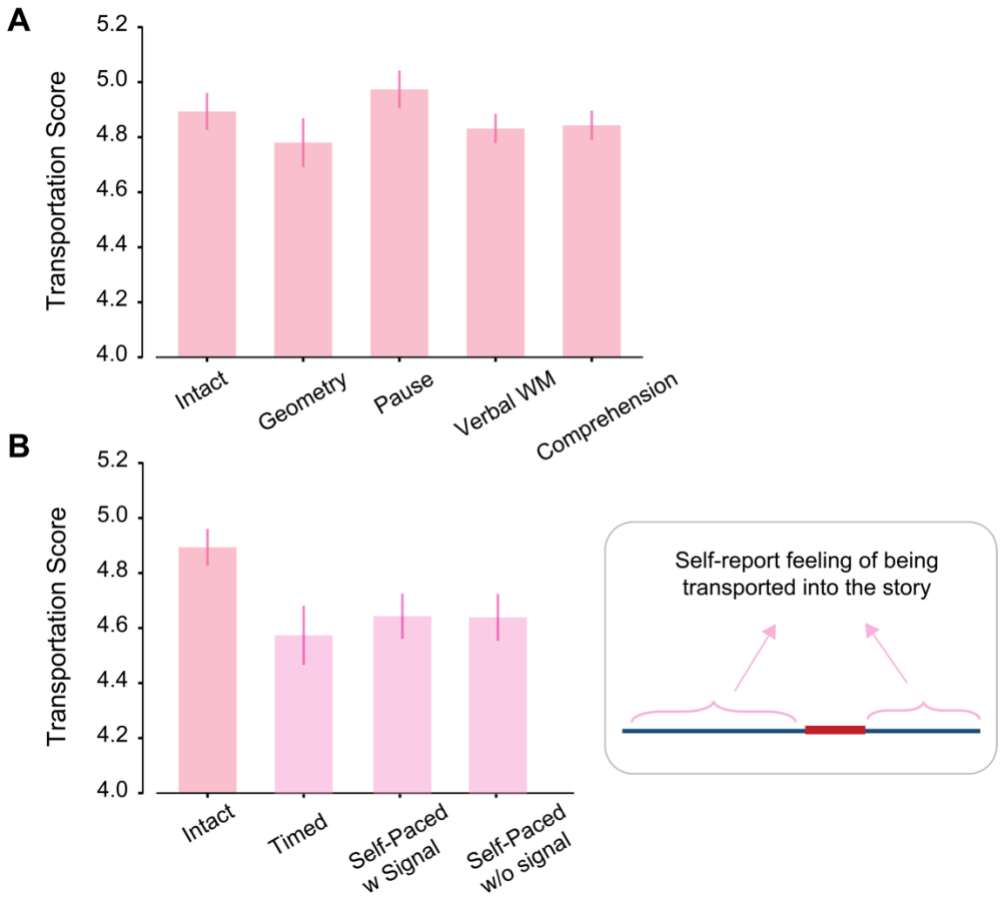


Figure 4-13. Self-reported transportation scores for different interruption conditions.
Score range was between 1 (not at all transported) to 7 (very much transported). Transportation score at 4 shows the neutral level. Error bars show standard error of the mean.

To measure whether the story was lingering in mind during the cognitive-similarity interruptions, we asked participants if they were intentionally or unintentionally reflecting on the story during interruptions. Self-report lingering was significantly lower in Geometry interruptions compared to Pause and Scenario interruptions (p-value for Pause versus Geometry = 0.009; p-value for Geometry versus verbal WM = 0.03; p-value for Geometry versus comprehension = 0.08). We did not observe any difference in lingering between the Pause and Scenario conditions (**Figure 4-14**).

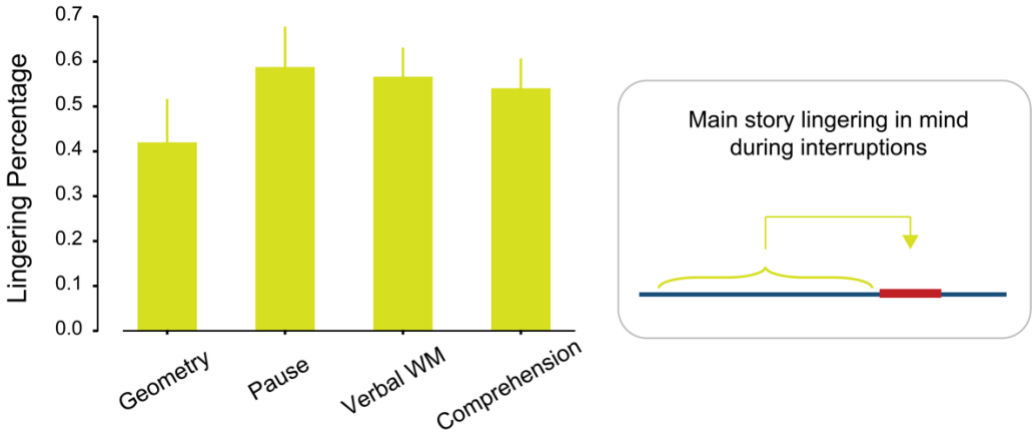


Figure 4-14. Self-reported lingering of the primary narrative during performance of interruption tasks. Percentage of participants who reported experiencing intentional or unintentional lingering of the main story during the interruptions. Error bars show 95% confidence interval of 10,000 bootstrapping with replacement.

4.4. Discussion and Conclusions

In this Chapter, we investigated how encountering different types of reading interruptions influence our processing and subsequent memory of a narrative. We investigated two main axes along which interruptions may vary. Along the first axis, we tested interruptions that similar or dissimilar to reading in terms of the cognitive processes they involve. Along the second axis, we tested interruptions whose internal semantic content was similar or dissimilar to the semantic content of the narrative.

Similarity of cognitive processing was an important factor in determining the effects of the interruptions. Among these interruption conditions, we found that participants were, in general, slower to read the primary text after encountering an interruption involving more similar cognitive processes (Scenario conditions, **Figure 4-5**) relative to more dissimilar cognitive processes (Geometry condition). On the other hand, the more cognitively dissimilar interruptions (Geometry) decreased recognition memory for sentences both before and after the interruption (**Figure 4-9**). Indeed, Geometry interruptions were the only type that negatively impacted the consolidation of preceding information. When participants encountered Scenario interruptions, thought to be more cognitively similar to reading, the interruption either improved the consolidation of prior memory (verbal WM condition) or did not change it (comprehension condition) (**Figure 4-9**). Encountering Pause interruptions, also, significantly enhanced recognition memory for preceding information, consistent with the findings that pauses at event boundaries aid memory consolidation (Ben-Yakov, Eshel, and Dudai 2013). Recognition memory for material following the interruptions was impaired in all conditions except Pause interruptions, with Geometry interruptions, the most dissimilar, generating the largest decrease (**Figure 4-9**).

In contrast to the similarity of cognitive processing, the similarity of semantic content in the interruptions had a much smaller effect on reading times and memory. We did not find any difference between the resumption reading time following low- versus high-similarity interrupting content (**Figure 4-6**). Also, the degree of content similarity did not affect consolidation of pre-interruption memory (**Figure 4-10** A). Consistently, when participants were given a signal about the occurrence of interruptions, there was no difference between consolidation of information following low- versus high-similarity interrupting content. Only when participants did not have any signal about the

occasional unrelated content, they showed worse consolidation following less similar interrupting content (**Figure 4-10** B).

The results from the content-similarity interruptions (high and low semantic similarity) were consistent with our observations when varying cognitive process similarity (Geometry, Pause, and Scenario interruptions). In terms of cognitive similarity, reading unrelated text regardless of its content, is very similar to reading a narrative. Both the resumption reading time and the memory consolidation findings were consistent between sub-conditions of content-similarity interruptions (especially timed and self-paced with signal) and highly similar cognitive interruptions (comprehension and verbal WM) (see **Figure 4-7** and **Figure 4-11**).

Our data suggest that when investigating the disruptive nature of interruptions, one needs to differentiate between the resumption lag and the memory effects. For instance, resuming reading after Geometry interruptions was faster relative to Pause and memory interruptions. However, the quicker resumption in this case does not represent a less disruptive interruption because Geometry showed a much higher negative impact on memory consolidation. Therefore, when comparing different types of interruptions on their negative impact for an ongoing task, we must determine what aspects of the performance are most important for the task. Negative impacts on reading time versus memory might draw a contrary conclusion on disruptiveness of an interruption.

In contrast to the literature which indicated that effects on textual memory were negligible (Glanzer, Dorfman, and Kaplan 1981; Glanzer, Fischer, and Dorfman 1984; Ledoux and Gordon 2006), we observed reliable effects of interruption on a detailed recognition test for the contents of the sentences in the text. Foroughi et al. (Foroughi et al. 2015) had already shown that deeper probes

of textual understanding could reveal interruption effects; now, we have found that even more surface-level verbatim memory for the text is affected by interruption.

What potential mechanisms underlie the memory effects and resumption delays induced by interruptions? It could be that when the main task and the interruptions engage similar cognitive processes, there is a higher interference between the ongoing processes which in turn makes it harder to resume the main task after the interruption as reflected in higher resumption lag. On the other hand, when the main task and the interrupting task engage similar cognitive processes (e.g., parsing a sentence or constructing a situation model), our cognitive processes seem to be capable of separately process unrelated information with minimal interference. One possibility is that our cognitive processes can identify the change in the information stream, potentially create an event or situational representations, and process unrelated information separately. This theory was supported by reading time differentiating low- and high-similarity interrupting content, but equally quickly resuming afterward (**Figure 4-8**).

Our results also shed lights on potential underlying mechanisms modulating the effects of cognitive similarity and content similarity on memory. It could be that engaging less similar cognitive processes reduces the memory by lowering the possibility of lingering. In fact, in Pause interruptions, participants reported the highest lingering during the interruptions and showed the highest increase in pre-interruption memory (see **Figure 4-14** and **Figure 4-9** A). However, different types of interruptions might reduce the post-interruption memory for different reasons. Geometry interruptions might reduce post-memory because of interference caused by using less related cognitive processes, whereas comprehension interruptions could diminish post-memory because of the continuation of ongoing processing (lingering) in high-similarity cognitive processes. Further

investigation is needed to explore the effects of interruptions on subsequent memory in future research.

The findings of this research project are subject to limitations that may be addressed in future work. To begin with, our method of measuring sematic similarity between interrupting texts and the primary text (Universal Sentence Encoder) may primarily involve lexical information, rather than deeper situational information conveyed by the sentence (Cer et al. 2018). Thus, further research is necessary to differentiate the effects of token-level information from high-level semantics contained in a sentence. A second limitation is that this project employed one text, a short story, as the main stimulus. To further investigate the effects of semantic properties of the interruptions on reading, future studies should evaluate and compare the effects of semantic interruptions on a wider variety of primary texts in different styles. Lastly, our experimental conditions provided evidence for situations in which the interruption and the primary task have similar cognitive processes, however, we only had one condition (Geometry interruptions) that involved cognitive processes that were very different from the main task. Future research should further explore a wider selection of interrupting tasks with dissimilar cognitive processes to reading. Finally, although this study reports new memory effects and reading time effects arising from interruption, it has not yet established how they are related. For example, when participants exhibit a longer-than-usual resumption delay on a specific trial, does this also mean that they will be more successful in avoiding memory interference on that trial? Further work, involving across-individual and across-trial correlations, will be required to address this important question.

In summary, in Chapter 4, we investigated how encountering various types of interruptions while reading influences readers' ongoing processing and retention of the information. Our findings can be used in the short term to provide general guidelines for reducing the negative consequences of

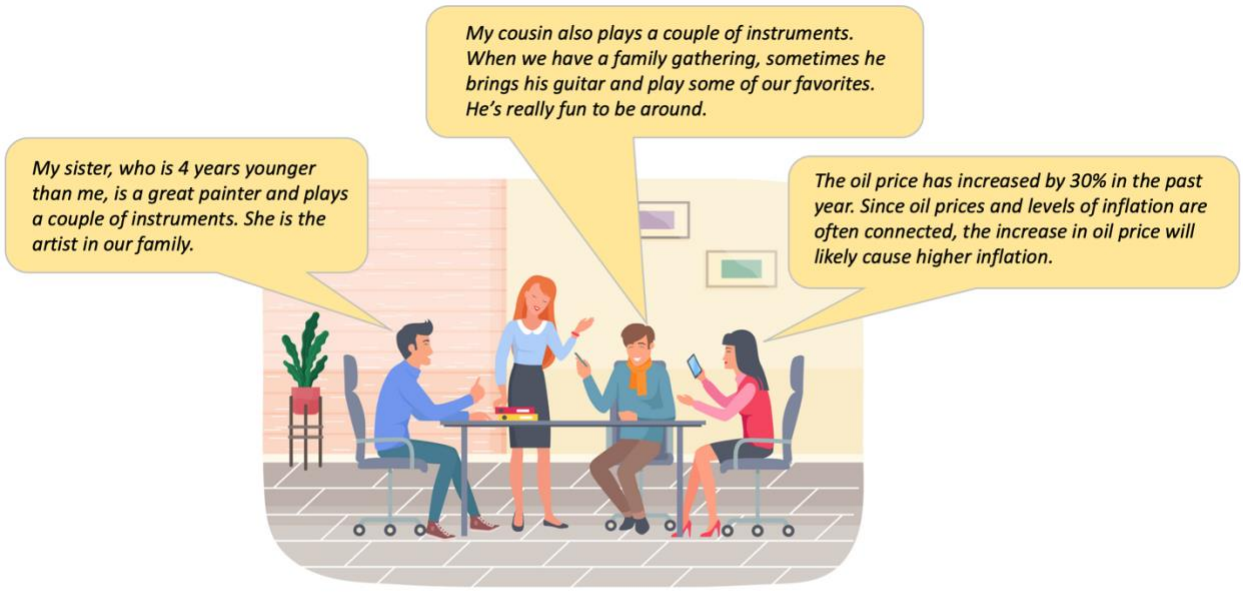
interruption in real-world situations. In the long run, these results can help develop more complex process-level models that describe the updating and interference processes that occur during reading in working memory and long-term memory.

Chapter 5: Determining How Processing in Neural Language Models is Affected by Interruptions

5.1. Background and Motivation

In the previous Chapter, we looked at how interruptions while reading can affect humans' understanding and subsequent memory of the information. In this Chapter, we explore how contemporary neural language models process linguistic data that contains incongruencies.

Neural Language Models (NLMs) are powerful and flexible models supporting robust language predictions when processing coherent data streams. However, when we process language data in the real world, we may experience incongruencies, such as a sudden change in topic from one sentence to the next (**Figure 5-1**). When humans encounter interruptions in language processing, the content and length of the interruptions affect their comprehension performance (Ledoux and Gordon 2006; Foroughi et al. 2015a). This effect in language comprehension is reflected in longer reading times following the incongruent information. NLMs and human brain share some computational principles as they process natural language: both involve in continuous predictions, and both use prior context to make predictions (Goldstein 2021). Therefore, it is possible that NLMs could be influenced similarly by processing incongruent information, or that they could use similar approaches to mitigate the costs of interruption.



My sister, who is 4 years younger than me, is a great painter and plays a couple of instruments. She is the artist in our family. My cousin also plays a couple of instruments. The oil price has increased by 30% in the past year. Since oil prices and levels of inflation are often connected, the increase in oil price will likely cause higher inflation. When we have a family gathering, sometimes he brings his guitar and play some of our favorites. He's really fun to be around.

Figure 5-1. Example of a situation with a transient change in a temporally coherent information stream.

However, it remains unclear how NLMs integrate and separate information at moments of incongruency in prior context and how their predictions are influenced by inconsistent prior context. We hypothesized that the predictions generated by NLMs will be especially affected when their prior contexts are contaminated by incongruent information that (i) is similar to the main information sequence, rather than distinct and (ii) has a long duration. The impairment of NLM prediction quality manifest as an increase in their perplexity scores for the veridical word sequences.

In this Chapter, we test two kinds of NLMs architectures: a recurrent NLM (LSTM, (Hochreiter and Schmidhuber 1997; Schijndel and Linzen 2018)) and an attention-based NLM (GPT-2, (Vaswani et al. 2017; Radford et al. 2020). And, we have two main motivations: First, we want to understand whether and when these powerful models are successfully able to integrate information

over time, despite the insertion of interrupting incongruent material; Second, we aim to compare the performance of these models against humans facing the same interruptions (from Chapter 4), to determine whether the models can predict which kinds of interruptions and incongruencies humans will struggle with.

Our central question, here, is not whether NLMs have the same performance and the same representational capabilities as humans. We want to explore how robust NLMs are to different kinds of interruptions and how this varies as a function of architecture. However, we do also want to test performance of the LMs against humans. Therefore, in Chapter 4, we exposed human participants to high-similarity and low-similarity interruptions in a setting where they did not have any knowledge that there was any interrupting content. Our goal in doing so was to put humans and LMs on a more equal footing, facilitating the comparison between them. If a particular LM is disrupted by the same kinds of interruptions that humans are, then that LM could provide a useful practical predictor of when and where humans will be especially impacted by interruption (even if the LM does not have the same language competence as a human).

5.1.1. Research Questions

It is uncertain how contemporary language models (both recurrent LSTMs and feedforward Transformer models) are impacted by incongruencies in the prior context in their input stream. Therefore, in this Chapter we address the following research questions:

- How does the insertion of incongruent text in an otherwise coherent text modulate the predictions of language models? How do these effects depend on the length, location, and semantic confusability of the inserted text? What role NLMs' architecture play in the

magnitude of this impact? And finally, as a practical matter, can language models enable us to predict which passages of incongruous text humans will find difficult to process?

Significance

Language comprehension is one of the most complex human cognitive feats, and neural language models (NLMs) are (currently) the most successful machine language models for parsing, abstracting, and generating human language data (Kelly and Reitter 2018). Despite the purported computational commonalities between NLMs and the human brain and cognition (e.g., (Goldstein et al. 2022)), it remains unknown whether they exploit incongruent prior information similarly.

Studying these models in a naturalistic setting, similar to the linguistic data humans typically encounter in the real world, is important for at least two reasons. First, we can better understand how NLM performance is impacted by incongruities in prior context. It would be interesting to see how systems with the ability to understand and generate language data handle and process incongruent information. Also, given that these models are widely used in various real-life applications (e.g., email services, question answering chatbots, etc.), understanding them allows us to be aware of their limitations. Second reason) We can compare how interruptions influence language processing in NLMs versus humans. Understanding NLMs can provide some new insights in how we can study human language processing.

5.2. Methods

5.2.1. Language Models

We tested two pretrained language models: the GPT-2 model (Radford et al. 2020; HuggingFace (Wolf et al. 2020)) and an LSTM model (Hochreiter and Schmidhuber 1997; Schijndel, Mueller, and Linzen 2019).

5.2.1.1. GPT-2 Transformer LM.

GPT-2 is a large Transformer-based LM with the objective of predicting the next word, given all the previous words in a fixed-length context (Radford et al. 2020). Here, we used "GPT2-large" model which has 36 attention layers with 1.5 billion parameters. For computing word-by-word perplexity, we used a maximum context size of 1024 tokens. We implement GPT-2 using the HuggingFace library.

5.2.1.2. Long Short-Term Memory LM.

LSTM is a recurrent deep learning model that incorporates the past by reusing the information from previous time steps and through dedicated memory cells. The model that we implemented here has 2 LSTM layers with 1600 units in each hidden layer (Schijndel, Mueller, and Linzen 2019).

5.2.2. Experiment Design

5.2.2.1. Stimuli

Primary text

We tested two main primary texts: One primary text was an American Psychological Association (APA) article (~2500 words), titled "Can this marriage be saved?", which is about research on factors that influence the success of a marriage (Miller, 2013). Another primary text was a short story (~2300

words) by Raymond Carver, titled "So much water, so close to home", which narrates the tension between a couple. The same story was used in the behavioral experiment in Chapter 4.

The first primary text (APA article) was used to investigate the effects of incongruent prior context on the predictions of NLMs. Testing the second primary text (the narrative) had two objectives: (1) to replicate the results with a different primary stimulus; and (2) to compare the performance of NLMs and human participants under the same conditions (Human results are from Chapter 4). By testing NLMs and human participants under the same conditions, we mean the using same primary text, the same interrupting material, and evaluating performance for the same target sentences.

Interrupting text

To assess how NLMs perform with incongruities in prior context, we generated the incongruent context by randomly selecting unrelated sentences from Brown Corpus and incorporating them into the prior context. Brown Corpus is a collection of modern American English, comprised of approximately one million words, sourced from a diverse range of areas (e.g., fiction, science, technology, reviews, religions, hobbies, and so on (Brown Corpus, 2022).

5.2.2.2. Operationalization

One of our research objectives in Chapters 4 and 5 was to understand of how interrupting information affects the online processing of language models and humans. To measure the processing difficulty in humans, in Chapter 5, we used the reading time per character, relying on the common assumption that the longer it takes to read a sentence, the more difficult it is to process. To investigate how

interrupting material impacts language models' capacity to process prior context, here in Chapter 4, we used perplexity measures.

Perplexity measures. We evaluated the models using perplexity measures. Perplexity (PPL) is a common evaluation method for assessing LM's performance. PPL is the exponentiated average negative log-likelihood of a sequence. Intuitively, if a model assigns a high probability to the test data, it means that it is not perplexed (not surprised) by the test data, suggesting it is a good language model. So, a lower perplexity would identify a better model.

If we have a tokenized sequence $X = (x_0, x_1, \dots, x_t)$, then the perplexity of X is defined as Eq. 5-1

$$PPL(X) = \exp\left(-\frac{1}{t} \sum_{i}^{t} \log p_\theta \left(x_i \mid x_{<i}\right) \right) \quad \text{Eq. 5-1}$$

In Eq. 5-1, $\log p_\theta \left(x_i \mid x_{<i}\right)$ shows the log-likelihood of the model for the i-th token (x_i) given the prior context ($x_{<i}$).

5.2.2.3. Experimental Paradigm

To investigate how NLMs' prediction is influenced by having incongruency, we followed these steps: 1) Given the intact primary text, we chose a random sentence and use it as the target sentence; 2) We then calculated the perplexity of the NLM for the target sentence, when the prior context is the intact context in the original text preceding the target sentence (**Figure 5-2Figure 5-2**); 3) To create the incongruent prior information, we picked a sentence randomly from the Brown corpus and insert it before the target sentence; 4) Then we calculated the perplexity of NLM in processing target sentence, when the prior data that it uses for its prediction has both the original prior context and the

added unrelated sentence (**Figure 5-2**); 5) Finally, we compared the perplexity (PPL) of the language model with intact prior text and incongruent prior text.

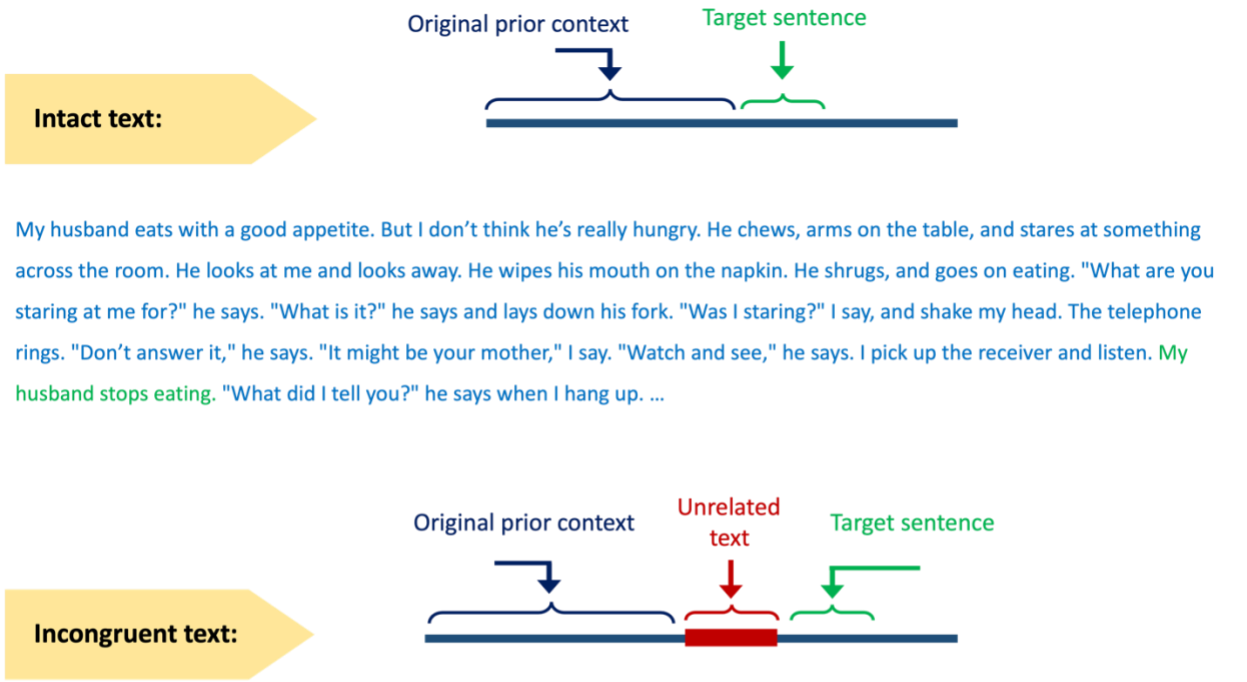


Figure 5-2. Examples of intact and incongruent prior context.
Top) Intact text, where the prior context has the original (uninterrupted) information; Bottom) Incongruent text, where the prior context consists of an incongruent price of information (the unrelated sentence inserted at the end of the original context.)

Change in PPL of target sentence after adding unrelated sentence to the prior was calculated as follows:

$$Change\ in\ PPL\ = \frac{Incongruent\ PP\ (\text{target sentence})}{Base\ PPL\ (\text{target sentence})} - \frac{Intact\ PP\ (\text{target sentence})}{Base\ PPL\ (\text{target sentence})}$$

Where Base PPL, Incongruent PPL, and Intact PPL are defined as follows:

Base Perplexity (target sentence): PPL (target sentence)

Intact Perplexity (target sentence): PPL (target sentence | Prior context)

Incongruent Perplexity (target sentence): PPL (target sentence | Prior context + Unrelated sentence)

5.2.2.4. Experimental Conditions

What does make an incongruency more disruptive in language models? We first started with the simplest condition and assess how models' perplexity for the target sentence change when they have 1 sentence of incongruency preceding the target sentence. We then examined the impact of three moderating factors: (i) length of the incongruent information; (ii) semantic similarity of the incongruent information to the target sentence text; and (iii) distance of the incongruent information from the target sentence (**Figure 5-3**).

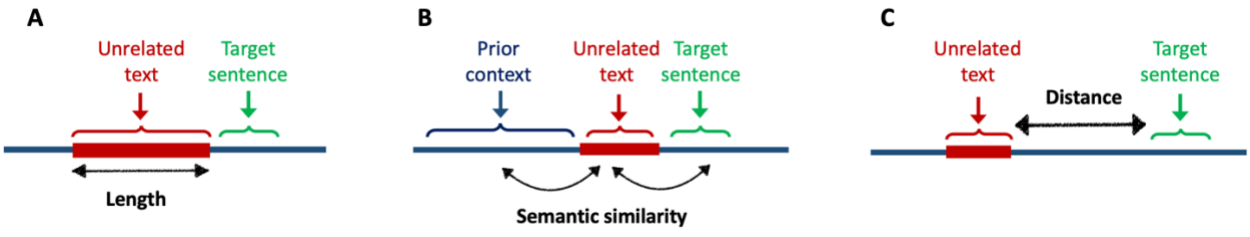


Figure 5-3. Three potential moderating factors in modulating the disruptive effects of interruptions.

Length of the interruptions. To investigate how length of the incongruent material modulate the NLMs' performance, we tested the models with both a set of 1-sentence interruption and a set of 4-sentence interruptions. We measured the change in perplexity of the target sentence in both cases (see **Figure 5-5**).

Semantic similarity of the interruptions. In this analysis, we examined how the semantic properties of incongruent material would modulate models' confusability.

We first tested the effects of semantic similarity of the interruptions to the target sentence, on processing the target sentence. Using the same semantic similarity methods outlined in Chapter 4, we calculated the semantic similarity between the interrupting text and the target sentence. The similarity score was calculated, as before, as the cosine similarity of the embeddings obtained from the Transformer-based USE model (Cer et al. 2018).

Based on the similarity score of a wide range of randomly chosen sentences, we divided the similarity scores into 6 bins (lowest similarity centered around -0.2; highest similarity bin centered around +0.3, **Figure 5-4**).

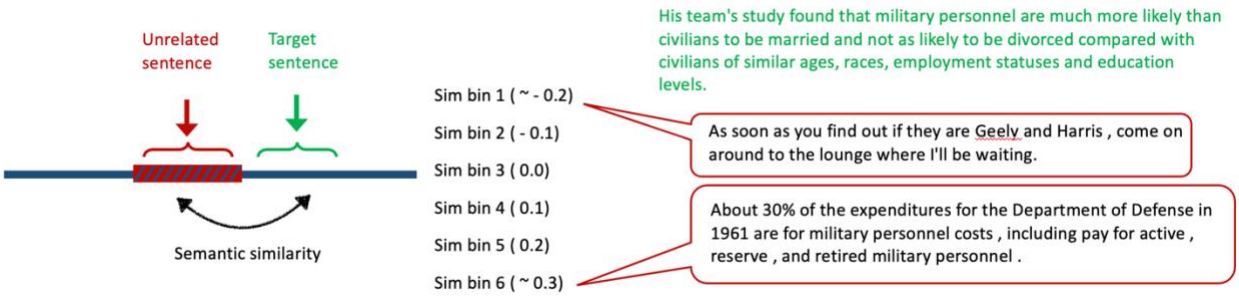


Figure 5-4. An example of a target sentence and two unrelated sentences with low and high similarity scores.

Additionally, to facilitate a close comparison between effects of semantic similarity on processing in both NLMs and humans, we tested them in the same conditions described in Chapter 4. In this condition, the primary text was a narrative, and the semantic similarity was calculated based on the similarity between the interruption material and a few sentences preceding the interruption (see **Figure 5-7**).

Distance of target sentence from the interruptions. To assess how NLMs recover from interruptions as they recede further into the past, we measured perplexity while varying the distance of the incongruent material to the target sentence. To that end, for a given incongruent context, we tested the (change in) perplexity of the first sentence following the interruption (lowest distance), the second sentence following the interruption, the third sentence following the interruption, and so on (see **Figure 5-10**).

5.3. Results

5.3.1. Effects of Interruptions Length

We found that the insertion of even one unrelated sentence to the prior context significantly increased the perplexity both the LSTM model and GPT-2 when processing the target sentence (One-sample t-test: LSTM: t-statistic=8.15, p-value < 0.001; GPT-2: t-statistic=23.23, p-value < 0.001). Additionally, the detrimental effect of interruptions was significantly higher for longer interruptions compared to shorter interruptions (Paired t-test: LSTM: t-statistic=13.99, p-value < 0.001; GPT-2: t-statistic=13.03, p-value < 0.001) (**Figure 5-5**).

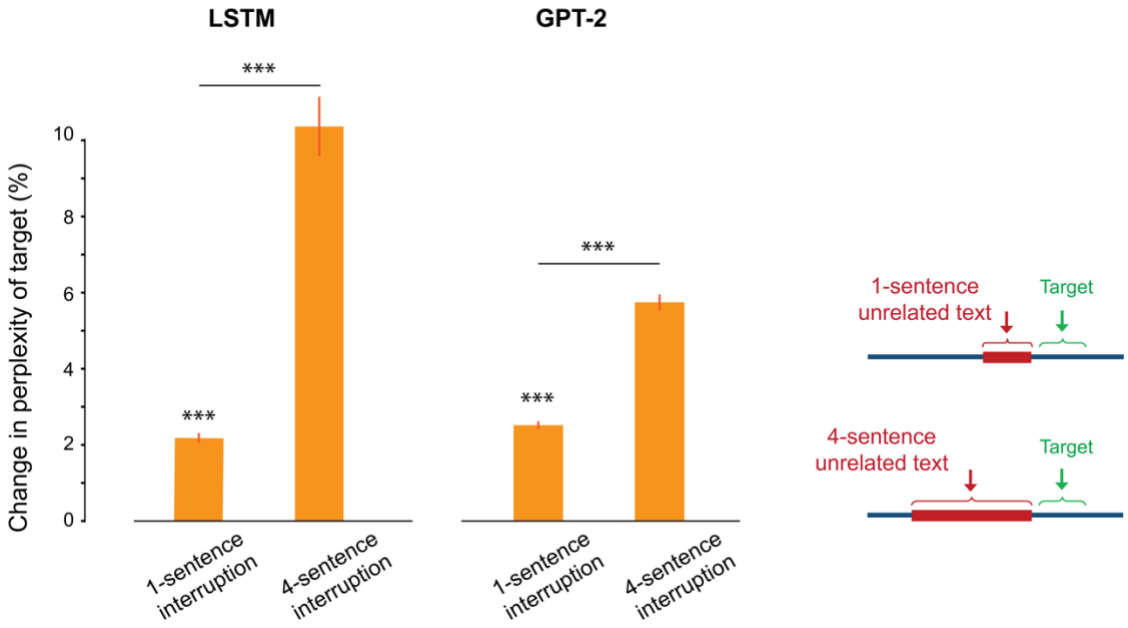


Figure 5-5. Change in perplexity of NLMs with incongruent prior context relative to intact prior context. Right) Schematic of a 1-sentence and 4-sentence interruption prior to target sentence. Left) Percentage change in perplexity of LSTM and GPT when processing a target sentence, given incongruities in prior context due to a 1-sentence or 4-sentence interruption. Each bar shows the data from 600 datapoints: there were 100 randomly selected target sentences followed by randomly interruptions from 6 levels of similarity (100*6). Error bars show standard error of the mean, after excluding outliers that were more than 3 times the interquartile range from the median. The primary text was the APA article. (*** = p-value < 0.001)

5.3.2. Effects of the Semantic Similarity of Interruptions

For the LSTM language model, we observed a negative correlation between the similarity score and the magnitude of change in models' perplexity. Interruptions with low similarity to the target sentence increased the models' perplexity (**Figure 5-6** A). Surprisingly, however high-similarity interruptions showed a different impact, in some cases even lowered the models' perplexity. The negative correlation between similarity and change in perplexity remained, and was even numerically larger, for longer interruptions (**Figure 5-6** C). Using a linear least-squares regression, for 1-sentence interruptions, we found Pearson correlation coefficient of -0.11 (slope = -0.35, significantly different from zero, p-value < 0.001), whereas for 4-sentence interruptions, Pearson correlation coefficient was -0.19 (slope = -0.96, significantly different from zero, p-value < 0.001).

In contrast to the LSTM model, GPT-2 was not influenced by the semantic similarity of the interruptions. For both 1-senetence and 4-sentence interruptions, there was no correlation between the similarity score and the magnitude of change in models' perplexity (**Figure 5-6** B and D) (Pearson correlation coefficient for 4-sentence interruptions = 0.03; regression slope not different from zero; p-value = 0.56).

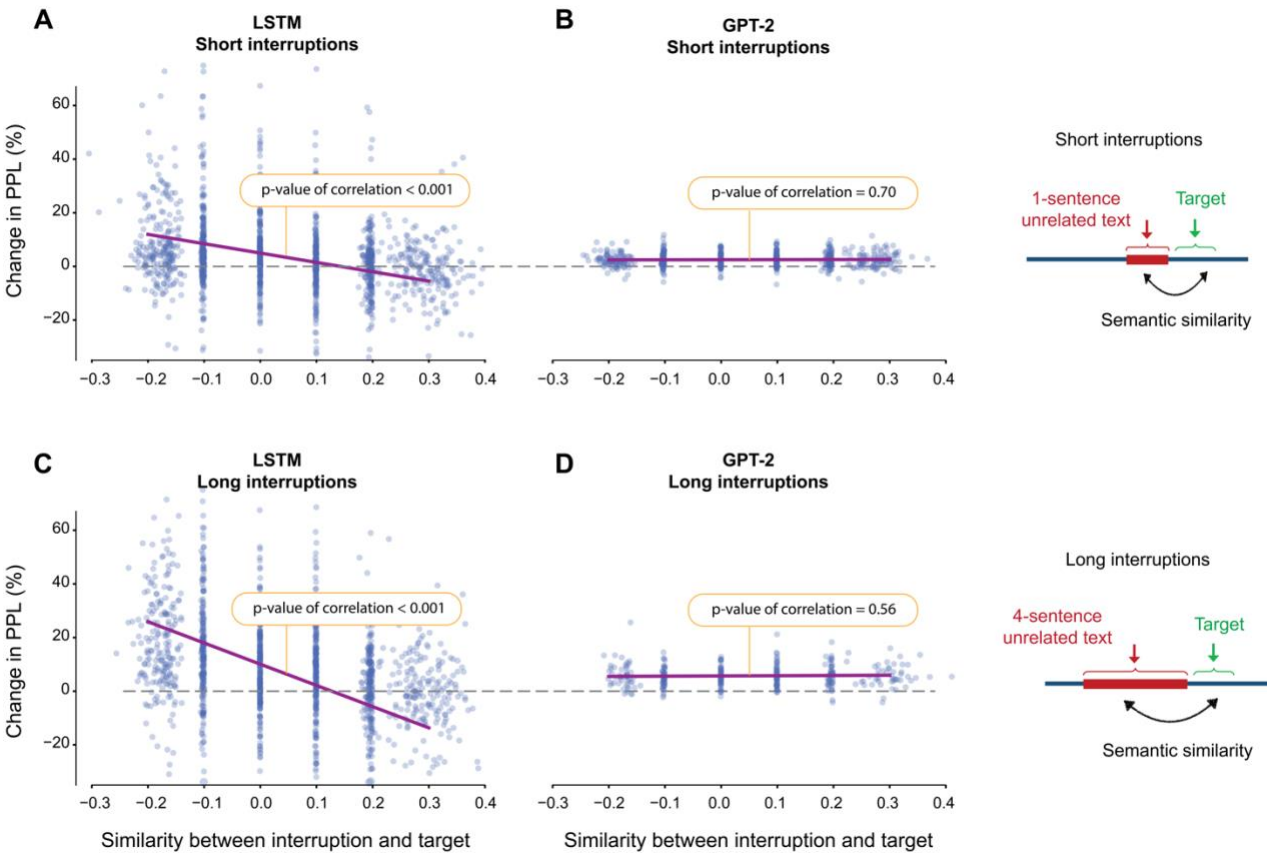


Figure 5-6. Effects of semantic similarity of interruptions to the target sentence on change in perplexity of NLMs when processing the target sentence.
A) Percentage of change in perplexity of LSTM when processing incongruent context composed of one interrupting sentence across a range of similarity scores. B) The same as A, for GPT-2 model. C) Percentage of change in perplexity of LSTM when processing incongruent context composed of 4 interrupting sentences across a range of similarity scores. D) The same C, for GPT-2 model. The primary text used for this result was the APA article.

We confirmed the same pattern on the effects of semantic similarity for our second primary text, the narrative. Again, we observed that the increase in LSTM perplexity was higher for low-similarity interruptions compared to high-similarity interruptions (paired t-test: statistic=7.83, p-value < 0.001) (**Figure 5-7** A). However, for GPT-2, we did not observe any difference between low-similarity and high-similarity interruptions (paired t-test: statistic=0.44, p-value=0.66) (**Figure 5-7** B).

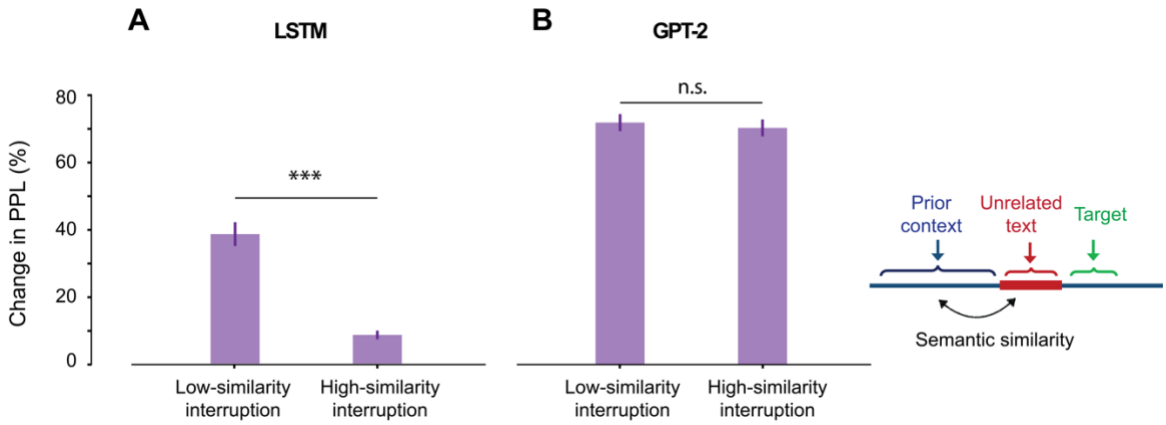


Figure 5-7. Effects of semantic similarity of interruptions to the prior context on change in perplexity of NLMs when processing the target sentence.
A) Percentage of change in perplexity of LSTM when processing incongruent context with low- or high-similarity score. B) The same as A, for GPT-2. We had 85 data points for each bar: total of 17 target sentences with 5 different samples for low-similarity interruptions (17*5= 85) and 5 different samples for high-similarity interruptions. Error bars show standard error of the mean. (n.s. = not statistically significantly different; *** = p-value < 0.001)

To compare the effects of semantic properties of the interruptions on processing in NLMs and humans, we compared how they process information following low- versus high-similarity interruptions (**Figure 5-8**). Recall that the reading time of human participants (following the interruption) was not influenced by whether the interruption was composed of low- or high-similarity material (**Figure 5-8** C).

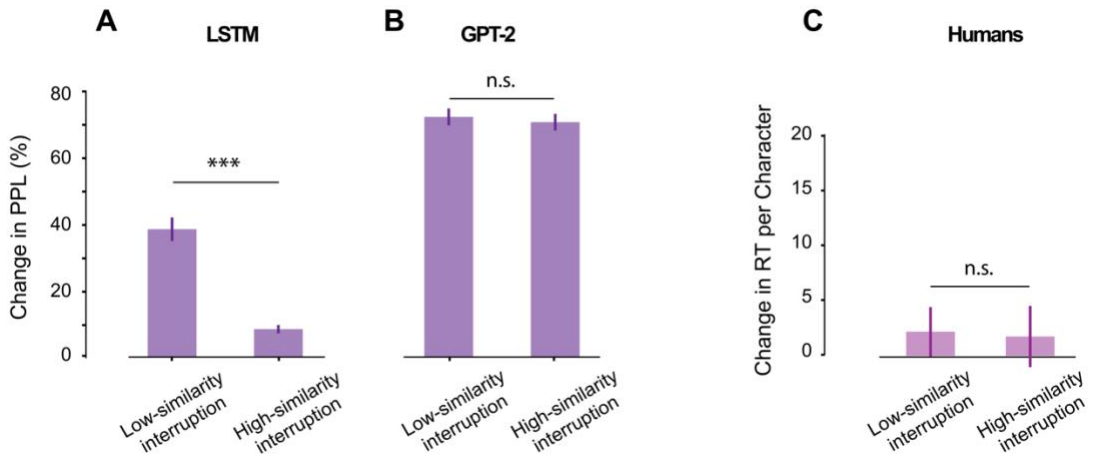


Figure 5-8. Effects of interruptions' semantic characteristics on processing in NLMs compared to humans. A and B) The same results as in Figure 5-7 C) Change in humans' reading time per character when processing a target sentence following a low-similarity or high-similarity interruption. Error bars show standard error of the mean. NLMs and humans were tested on the same primary text (the narrative) and interruption content. (n.s. = not statistically significantly different; *** = p-value < 0.001)

To gain a better understanding of the effects of the semantic properties of interruptions on ongoing processing in NLMs, we compared how LSTM and GPT-2 process information *during* the interruptions, relative to humans (**Figure 5-9**).

We found that humans and LSTM processed interruptions differently, but with some qualitative similarities. When encountering interruptions, LSTM showed a significant increase in perplexity of low-similarity interruptions (paired t-test: statistics=3.14, p-value=0.002), but not for high-similarity interruptions (paired t-test: statistics=0.74, p-value=0.46) (**Figure 5-9**, LSTM panel). Nonetheless, there was a numerical increase in perplexity for both types of interrupting material. This pattern qualitatively matches the human data, where we observed a significant increase in their reading time when encountering both low- and high-similarity interruptions. However, humans returned to their original reading times following both low- and high-similarity interruptions, whereas the LSTM showed an elevated perplexity following low-similarity interruptions (paired t-test: statistics=5.37, p-value<0.001) (**Figure 5-9**).

We found more striking disparities in interruption handling between humans and GPT-2. Although neither humans nor GPT-2 demonstrated any difference between following low- versus high-similarity interruptions (**Figure 5-8** B and C), their behavior when initially encountered interruptions was vastly different. GPT-2 did not demonstrate any increase in perplexity when it encountered either low- or high-similarity interruptions. Humans, however, showed a higher reading time for both interruption types, with a significantly greater increase for low-similarity interruptions (**Figure 5-9**).

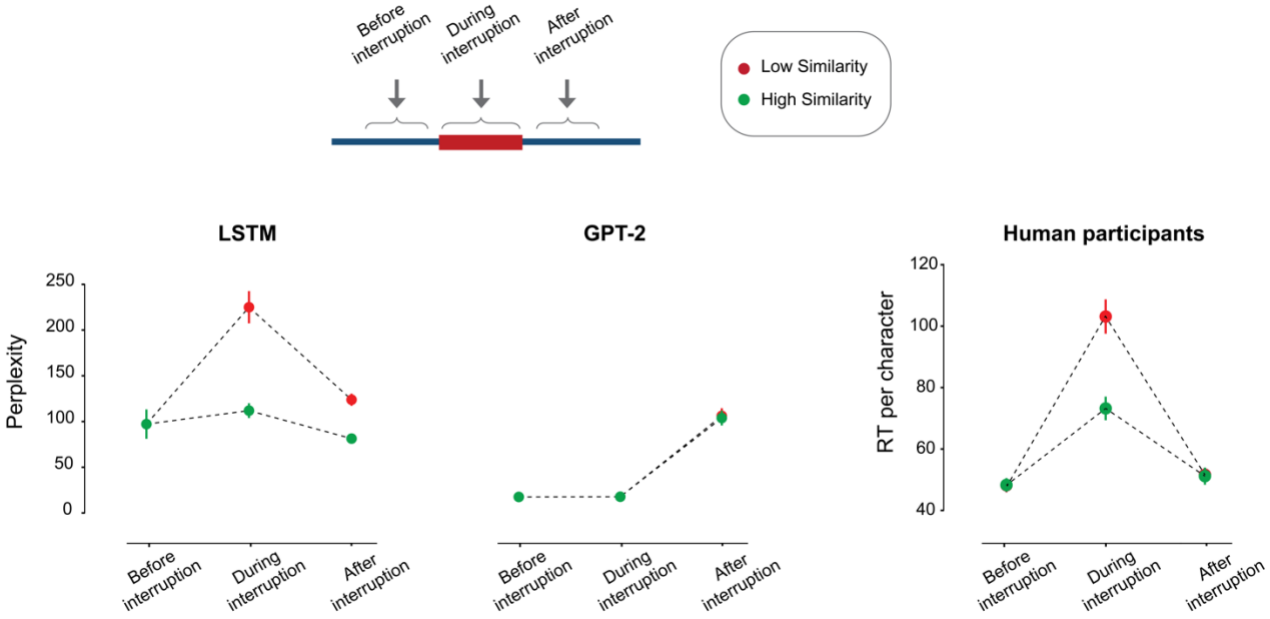


Figure 5-9. Effects of low- versus high-similarity interruptions on processing in LSTM, GPT-2, and human participants during and after interruptions.
Left) Perplexity of LSTM and GPT-2 for the sentence preceding the interruptions (before), first sentence during the interruptions (during), and the first sentence following the interruptions (after). Y-axis shows the raw perplexity of the sentence given its prior context. Right) Reading time per character for human participants when reading the sentence preceding the interruptions, the first sentence during the interruptions, and the first sentence following the interruptions. Error bars show standard error of the mean. LSTM, GPT-2, and human participants were tested on the same primary text (the narrative) and interruption content.

5.3.3. Effects of Distance from Interruptions

We observed a gradual decrease in the detrimental effects of interruptions as the target sentence was shifted further from the incongruent material. This effect was observed for both LSTM and GPT-2, and for both primary texts, the APA article, and the narrative (**Figure 5-10**).

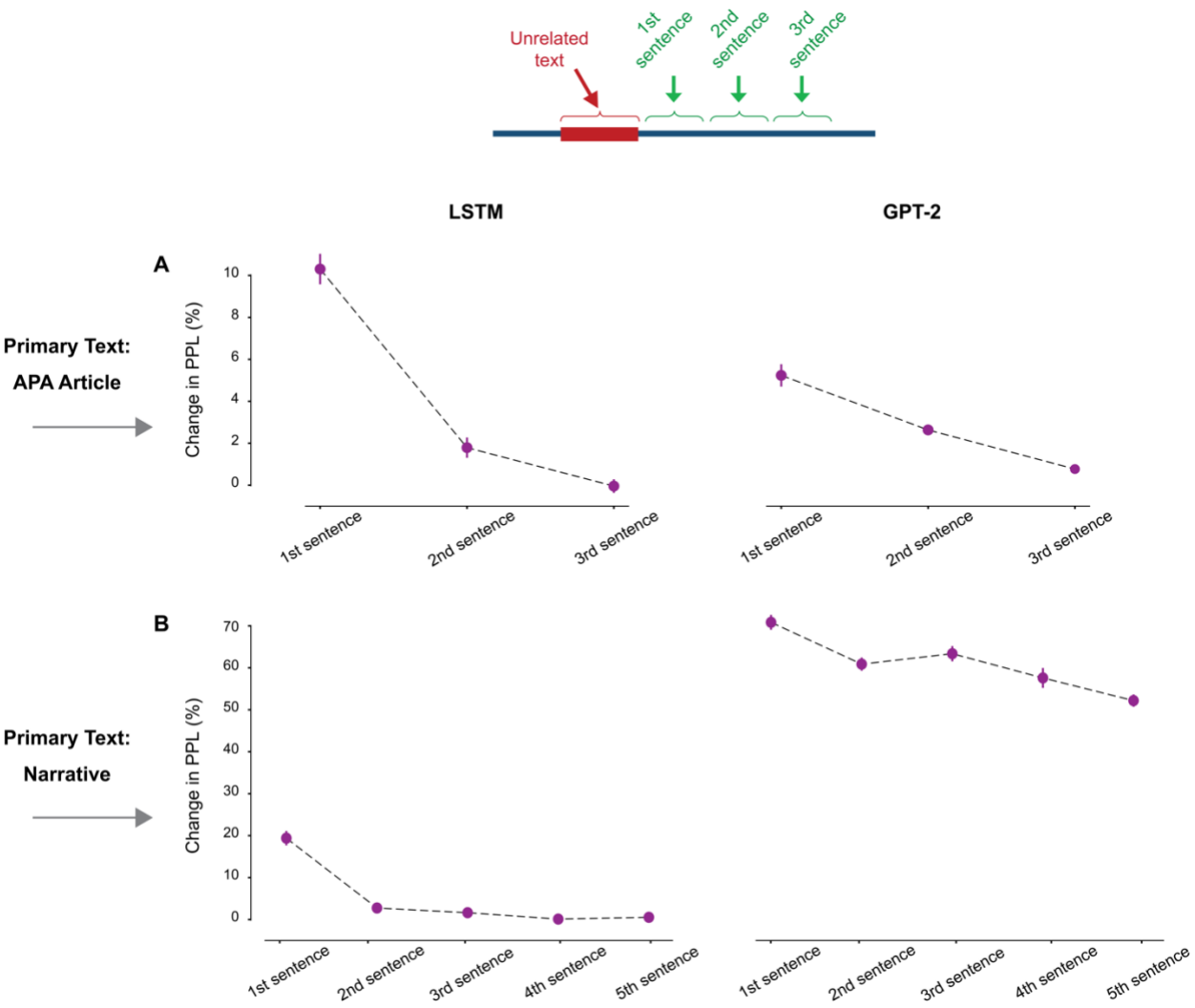


Figure 5-10. Effects of distance from interruptions on perplexity of NLMs.
A) Percentage of change in perplexity of LSTM and GPT-2 when processing the n-th sentence (e.g., first sentence, second sentence, and so on) following the interruptions. The results in (A) are from using the APA article as the primary text. B) The same as (A) but using the narrative as the primary text.

5.4. Discussion and Conclusions

In Chapter 5, we explored how NLMs respond to interrupting moments in language data and how length, location, and semantic confusability of the interruptions, as well as NLMs' architectural constraints, affect their predictions. Our research was designed to determine the impact of these incongruencies on LSTM and GPT-2, which are typically trained to reliably predict language in consistent data streams. We also compared these models with humans in their handling of incongruent text segments.

The presence, length, and distance of interruptions had similar effects on LSTM and GPT-2. Both the LSTM and GPT-2 faced higher processing difficulty when prior context included interruptions: even one unrelated sentence prior to a target sentence significantly increased NLM perplexity (**Figure 5-5**). The disruptive impact was increased for longer interruptions. In both LSTM and GPT-2, longer interruptions induced higher processing difficulty (**Figure 5-5**). Conversely, the disruptive effects of interruptions on both models were decreased when the incongruent material was placed at a greater distance (**Figure 5-10**).

The semantic properties of the interruptions produced dramatically different impacts on the LSTM and GPT-2. The LSTM's predictions were more disrupted by interruptions that were semantically dissimilar from the original context (**Figure 5-6** A and C, **Figure 5-7** A). In contrast, GPT-2's predictions were essentially invariant to the various degrees of semantic similarity of the interruptions (**Figure 5-6** B and D, **Figure 5-7** B). One possible interpretation of this effect is that GPT-2 is able to discount the interrupting material in making its predictions, because it can find more highly matching material elsewhere in its context buffer.

The observed discrepancies between the effects of the semantic properties of interruptions on LSTM's and GPT-2's predictions may stem from the difference in their architectural constraints. The recurrent architecture and limited-size hidden states of LSTM necessitate that the model must identify the most relevant features of prior context for subsequent predictions. Thus, when the model cannot store all details of its prior context, it is incentivized to represent a semantic gist of past information (Armeni, Honey, and Linzen 2022; Khandelwal et al. 2018). This hypothesis was supported by the evidence showing the dependence of LSTM on the semantic characteristics of interruptions. In contrast, we expected GPT-2's predictions to rely more heavily on verbatim properties of the prior context due to its "perfect memory" of the past, which is enabled by a fixed-length window of prior information input.

Despite GPT-2's overall higher capabilities, LSTM was more sensitive to tracking the change in the information sequence. GPT-2's perplexity was disrupted when the incongruent material was in the prior context but not when it processed incongruencies given an intact prior context. On the other hand, LSTM demonstrated a higher processing difficulty when facing low-similarity interruptions even given an intact context. LSTM's ability in detecting low-similarity interruptions could come from maintaining a "leaky" and abstract version of the past, like a moving average of the context history (Figure 5-3).

Our findings revealed disparities in the way humans and NLMs process incongruent material in a mostly consistent context. Humans did not experience an increase in reading time when returning to the original context after interruptions, regardless of whether the interruption was of low or high similarity (see **Figure 5-8** C). This does not mean, however, that humans are not affected by the semantic properties of interruptions; rather, they exhibited a significant difference in response to different similarity levels during the interruptions (**Figure 5-9**). In contrast, neither LSTM nor GPT-

2 could identify the change between the narrative and a sudden high-similarity interruption (see "during interruptions" in **Figure 5-9**). Furthermore, neither of the models could return to the same level of perplexity after interruptions as before (**Figure 5-9**).

There was a qualitative resemblance between how the LSTM and humans processed incongruencies (**Figure 5-9**). Both humans and the LSTM demonstrated a significant increase in processing difficulty when first encountering interruptions. And, in both humans and LSTMS, there was a greater increase when the interrupting material had low similarity to the primary text. Finally, in both humans and the LSTM, the indices of processing difficulty decreased following interruptions when the original context resumed. In contrast to humans and the LSTM model, GPT-2's predictions were hardly altered at all during the interruptions; instead, the interruption only impacted GPT-2's prediction after it was complete, when it altered for the resumed primary text. Why did human reading times and LSTM perplexity shift in an analogous manner? One possibility is that both of these indices track a very local predictability of the information stream, based on the immediate word-frequency and the context of the past several words, as in an n-gram model. This is in contrast to GPT-2, a model whose strength in incorporating long-range context may have reduced the influence of transient shifts (a four-sentence interruption) in the information.

The distinction between processing of interruptions in humans and NLMs could arise from the difference in their memory capabilities. Humans exploit a variety of short-term memory (verbatim and semantic) and long-term memory (semantic and episodic) systems simultaneously. This combination of mechanisms may enable humans to keep track of both abstracted semantic / situational and verbatim properties of the information stream. Additionally, episodic memory capabilities of humans may enable humans' cognitive processes to identify boundaries in the information, create and store events information, and reload it as needed.

What architectural modifications could improve the susceptibility of high-performing language models to interruptions? Despite GPT-2 impressive and human-level performance in many real-world tasks, it did not process incongruent context similarly to humans. Currently, no language model uses a combination of all human-like memory mechanisms. Equipping large-scale language models, such as GPT-2, with short- and long-term memory mechanisms to detect changes in both short- and long-term, could enable them to process semantic properties of information more sensitively. Additionally, providing these models with a capacity to separately store irrelevant material could make them resilient when facing incongruent context.

The work in this Chapter is subject to several limitations which may be addressed by future work. First, because GPT-2 is a larger model trained on a larger dataset than our LSTM, we cannot confidently generalize our findings to feedforward Transformer architectures and recurrent LSTM architectures in general. Future work should employ an LSTM that is more closely matched in terms of training corpus, model size and average prediction performance (see also (Armeni, Honey, and Linzen 2022)). Second, GPT-2's apparent lack of sensitivity to the semantic similarity of interruptions may be due to the characteristics of the method that we used to calculate semantic similarity. In other words, GPT-2 may use semantic similarity information, but the semantic properties that it captures may differ from those captured by our implemented method, Universal Sentence Encoding (Cer et al. 2018), which may prioritize token-level or syntactic similarity. Future research could explore this further by differentiating between semantic properties and particular tokens by using semantically similar interruptions that involve different tokens, as well as semantically dissimilar interruptions that comprise similar tokens. Third, to estimate the generalizability of our findings regarding architectural constraints, future studies should evaluate the effects of interruptions on another and perhaps a higher performing LSTM, as well as another attention-based language model. Fourth, it is also unclear

whether the position of the incongruencies in the primary context affects susceptibility to interruptions. To address this, future research could investigate if inserting disruptors within an event or at the transition between two events would produce equal effects on the models.

This Chapter has explored how processing incongruent contexts affect the functioning of NLMs. Our results shed light on the shortcomings of LSTM and GPT-2 when faced with incongruities in their prior context. Additionally, we explored in what ways LSTM and GPT-2 process incongruent language data differently from humans. The contrast between human and NLMs behavior can be used to suggest modifications to NLMs for greater resilience to disruption – for example, by augmenting them with memory systems. The comparison of humans and NLMs may also further our ability to predict human language processing, as when the perplexity of the LSTM model qualitatively tracked the changes in human reading times during the interruptions.

Chapter 6: General Discussion and Conclusions

What are the implications of temporal correlation in data for learning and comprehension? How can temporally correlated real-world data be leveraged for learning, particularly in brain-inspired architectures? What effects do transient changes in the information stream have on contemporary machine learning models, and how do these effects compare to what is seen in human behavior? In this thesis, I addressed these questions using both computational and empirical methods. As each of the four preceding Chapters contain a more targeted discussion of the specific results, here I review and summarize the broader contributions of this dissertation.

6.1. Summary of Research Contributions

6.1.1. Efficacy of Brain-Inspired Integration and Separation Mechanisms for Category Learning

Cortical circuits exhibit two characteristics that may influence how they learn category categories from a sequence of temporally correlated examples, as we commonly encounter in the real world. First, our brain dynamics appear to combine information over time, as there is ubiquitous autocorrelation in cortical dynamics. Second, neural circuits sometimes shift their context rapidly at "event-boundaries", which appears to be related to resetting of context representations. It was unknown how these neural circuits' characteristics (autocorrelation and boundary resetting) influence category learning from temporally correlated data. Therefore, we examined the efficacy of two brain-inspired mechanisms – leaky memory and memory gating – for incremental category learning from temporally structured data. We found, firstly, that equipping neural networks with these brain-inspired mechanism enables them

to learn more efficiently from temporally autocorrelated data, relative to a conventional feedforward network. Second, we found that when performing representation learning, these biologically plausible models were able to flexibly adapt to the autocorrelation level, so that they could take advantage of repeating structure without suffering from the interference of unrelated prior information.

6.1.2. Efficacy of Brain-Inspired Integration and Separation Mechanisms for Learning Representations from Multi-Timescale Data

In the real world, we may need to learn from data with multiple levels of autocorrelation. Moreover, in real-world settings, there are no pre-defined labels to support the learning. It was unknown how or whether a learning mechanism with brain-like architectural constraints could exploit multiple timescales of data to learn multi-timescale representations. We tested the efficacy of brain-inspired leaky memory and multiscale memory gating mechanisms for unsupervised representation learning. We found that autoencoder models equipped with recurrence and multi-timescale gating could successfully learn to reconstruct the input, and, moreover, that they learned representations which selectively separated multiple timescales of the data stream.

6.1.3. How Human Language Processing is Affected by Interruptions

In the real world, our cognitive processes may experience a transient change in the information stream, such as encountering interruptions, while carrying out a primary task. Experiencing interruptions when reading is unavoidable in our daily life, however, the consequences of encountering different types of interruptions for how we process the information remains unclear. Despite decades of research, there

were conflicting findings on which kinds of interruptions most impacted online reading comprehension and memory. We investigate how encountering many different types of interruptions, all tested within a common paradigm, would influence our processing and subsequent memory of a narrative. We explored two main axes along which interruptions may vary: (i) interruptions that similar or dissimilar to reading in terms of the cognitive processes; (ii) interruptions whose internal semantic content was similar or dissimilar to the semantic content of the narrative. Our data showed that resumption lags were smaller when the interrupting task was cognitively dissimilar from reading (e.g. a simple geometry task), but the similarity of the internal content of the interruption (e.g. its lexical and semantic similarity) had near-zero effect on resumption lags. Strikingly, we did not observe any resumption lag when participants were not told they would be experiencing interruptions. Moreover, we found consistent effects of interruption on detailed textual memory, in contrast to many prior claims that human reading comprehension was largely robust to interruption.

6.1.4. How Processing in Neural Language Models is Affected by Interruptions in the Prior Context

Neural Language Models (NLMs) are powerful and flexible models that can process coherent data stream, leading to robust predictions. However, when dealing with real-world language data, changes in the topic of a sentence from the previous one, can lead to incongruencies in the models input stream. It remained unexplored how NLMs process information in the presence of interrupting incongruent material: whether they can separate out the incongruent material to prevent it from influencing their predictions, and which kinds of incongruencies were most disruptive to their predictions. We tested two NLMs: a recurrent LSTM model and a feedforward attention-based model called GPT-2. We explored how NLMs respond to interrupting moments in language data and how

length, location, and semantic confusability of the interruptions, as well as NLMs' architectural constraints, affect their predictions. Our goal was to explore whether and when these models can effectively combine relevant information over time, despite the insertion of interruptions; and further, to compare the performance of NLMs against humans in order to determine which interruptions and incongruencies humans find difficult to handle.

We observed that the perplexity of the LSTM model qualitatively matched changes in human reading times during incongruent passages of text. Both the LSTM and humans showed a significant increase in processing difficulty when encountering material that was different from the content of the main text. Unlike humans and the LSTM model, GPT-2's predictions remained largely unchanged during the interruptions; it was only after the interruption was over and the primary text resumed that the GPT-2 predictions shifted to accommodate the new context.

6.2. Future Directions

6.2.1. Testing Biologically-Constrained Neural Networks for Learning from Naturalistic Information Sequences

Category Learning

In Chapter 2, we first investigated the influence of autocorrelation in sequence of training examples on category-learning in feedforward neural networks. We then identified brain-inspired integration and separation mechanisms that enabled these models to exploit autocorrelation in data sequences for learning more efficiently. Future research should explore the effects of autocorrelation in data on

category-learning in more complex neural networks models for object classification, such as convolutional neural network (CNN) models.

Consistent with our prior findings, our preliminary data showed that feedforward CNNs also suffer from autocorrelation in exemplar sequences. We ran an exploratory experiment in which for the CNN model, we employed ResNet-18 (pretrained with ImageNet dataset (He et al. 2016)) to categorize Fashion MNIST examples. We found that similar to the simple feedforward neural networks tested in Chapter 2 (**Figure 2-2** A), learning was slower for higher levels of autocorrelation, and fastest when any autocorrelation was avoided. Further research is required to investigate whether equipping CNNs with our biologically plausible integration and separation mechanisms would enable these models to exploit the autocorrelation in training sequences for higher learning efficiency.

Representation Learning

In Chapter 3, we investigated the temporal properties of simple autoencoder models' internal representations by using a synthetic dataset that had multi-timescale autocorrelation. Future research could explore the effect of equipping CNN models with leaky memory and reset mechanisms on their capability to learn interpretable temporal representations from data structures which are more realistic and temporally complicated, such as videos from the real world.

Our preliminary data suggested that when training a CNN with frames of a video dataset with natural and complex temporal structure, internal representations of higher levels of CNN demonstrate multiple timescales. We tested a CNN model with sequential data from frames of a real-world video (**Figure 6-1** A). For the CNN model, we employed ResNet-18 (He et al. 2016). We took the internal representations of the last hidden layer in ResNet-18, which was the output of the average pooling

layer, and analyzed it to see whether it contains multiple timescales (**Figure 6-1** B). In our preliminary analyses, we measured the temporal autocorrelation in the ResNet-18 internal representations (vector of size 1*512), we examined how these hidden representations change over time (over different iterations). To this end, we calculated how fast or slow the autocorrelation of each of 512 features in the internal representations change over time. We called a feature fast-changing if its autocorrelation drops below the threshold with fewer lags (e.g. lag of 1 or 2). In contrast, a feature was called slow-changing if its autocorrelation drops below the threshold with more lags (e.g. lag of 4 or 5). We used an arbitrary threshold equal to 0.35 but using other thresholds is also possible. Using lower threshold allows for getting more distinct timescales. We found that internal representations in higher levels of ResNet demonstrate multiple timescales, some changing faster over time and some changing more slowly (**Figure 6-1** C and D). Therefore, consistent with our prediction, we found that when a CNN model is exposed to real-world information stream that contains multiple levels of slowness, internal representations in its higher levels also reflect multiple timescales.

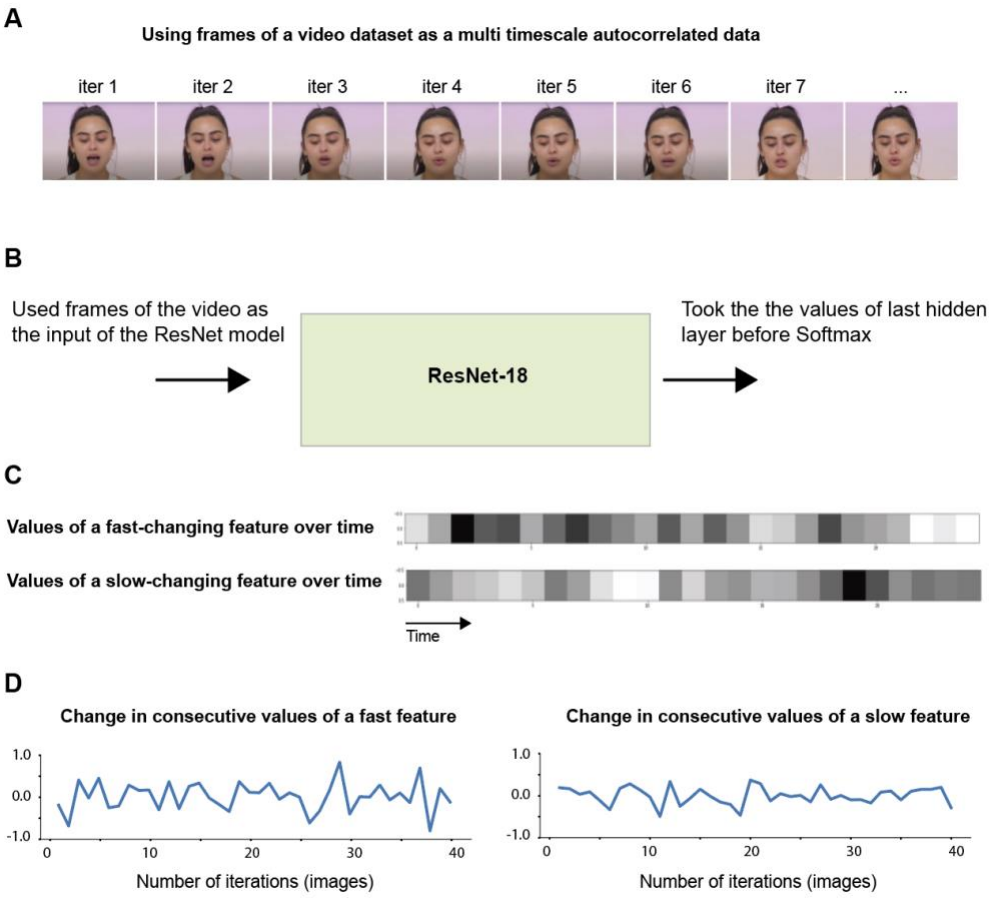


Figure 6-1. Real-world video with multiple levels of autocorrelation.
(A) Frames of a real-world video used as the input to the model shown in section B. (B) Input and output of the pretrained ResNet-18. Input: frames of the video shown in A; Output: values of ResNet-18 last hidden layer before Softmax. (C) A sample fast-changing feature and a sample slow-changing feature of the values taken from ResNet model. (D) Change in the consecutive values of a fast (left) vs slow (right) feature.

Future research could explore whether equipping CNN models with leaky memory and reset mechanisms enables them to learn interpretable temporal representations from temporally complicated data structures, such as videos from the real world.

6.2.2. Effects of Lingering during Interruptions on Subsequent Memory

The results of Chapter 4 indicated that participants' memories of the information preceding the interruptions varied depending on the type of interruption. When some material was followed by a Pause interruption, recognition memory for that material was significantly enhanced, whereas it marginally declined when the reading material was followed by Geometry interruptions (**Figure 6-2** A). Moreover, participants from Pause interruptions reported the highest level of lingering of the story in their minds during the interruptions, while those in the Geometry interruption condition reported the lowest (**Figure 6-2** B).

Figure 6-2. Memory of pre-interruption information versus lingering during interruptions.
A) Change in accuracy of recognition memory of information preceding interruptions relative to Intact condition. B) Percentage of Participants who reported lingering on the main story during interruptions.

If lower subsequent memory is associated with lower lingering, individuals with lower lingering may demonstrate lower memory. Therefore, a promising line of future inquiry would be interesting to explore the individual differences in participants' memory performance and their lingering scores.

A more general possibility is that engaging in an interrupting activity that requires different cognitive processes than reading reduces the chances of lingering of information that precede the interruption, thus disrupting the encoding of that information. Here, we use the term "lingering" to describe the act of reflecting, either consciously or unconsciously, on information that has been previously read. If lower lingering during Geometry interruptions is the cause of the decreased recognition memory, then modulating the lingering level should also modulate the memory performance. One might test this hypothesis by asking participants in the Pause condition to intentionally linger on the preceding material or to avoid lingering on it. We expect that those who actively linger will show an improvement in memory encoding, while those who actively avoid lingering will show a decrease in memory.

6.2.3. Effects of Semantic Properties of Information on Processes in NLMs and Humans

In Chapters 4 and 5, we tested the semantic similarity of interruptions to the main text within a limited range of similarity scores between -0.2 to 0.3. In humans, we found that when encountering high-similarity interruptions, participants were able to identify the change in the information flow reflect in an increase in their reading time. Additionally, GPT-2 did not show any difference between processing low- versus high-similarity interruptions. One possibility is that our current high-similarity interruptions (~0.3 cosine similarity) are still sufficiently distinct from the main text that both humans

and GPT-2 can recognize that they are distinct and then process them separately without difficulty or confusion. In contrast to humans and GPT-2, the LSTM not only was not negatively affected by semantically similar interruption, it actually benefited from these interruptions (as reflected in decrease in perplexity).

Future studies should explore a broader array of semantic similarity, including interruption content that has a higher level of similarity to the primary text. In an extreme case of similarity, the interruptions would be so semantically similar to the primary context that they would seem to be part of the main context. Humans and GPT-2 might also benefit from such interrupting material, although it may not be considered an "interruption" anymore.

We ran an exploratory experiment to test the effects of extremely similar interruptions on LSTM and GPT-2. To create extremely similar interruptions, we utilized an Transformer language model called PEGASUS (Zhang et al. 2020), which is fine-tuned for paraphrasing, to automatically paraphrase the target sentence from the main text, and implemented it using HuggingFace library (Wolf et al. 2020). These interruptions had a similarity score of around 0.85, which is much higher than the scores previously tested (~0.3).

Our initial findings indicate that GPT-2, similar to the LSTM, may benefit from the inserted interrupting context if the interruptions are very similar to the main text. GPT-2's perplexity dropped drastically (up to 80% reduction) when the interruptions were paraphrases of the target sentence. Interestingly, LSTM's perplexity was only slightly lower than that of the randomly selected interruptions (up to 30% reduction), despite the fact that the target sentence and the paraphrased sentence contained a lot of the same tokens. The small change in LSTM's response to similar interruptions (~0.3) versus extremely similar interruptions (~0.85) supports what prior research

suggested – that LSTM relies on the gist of the information rather than specific tokens (Armeni, Honey, and Linzen 2022).

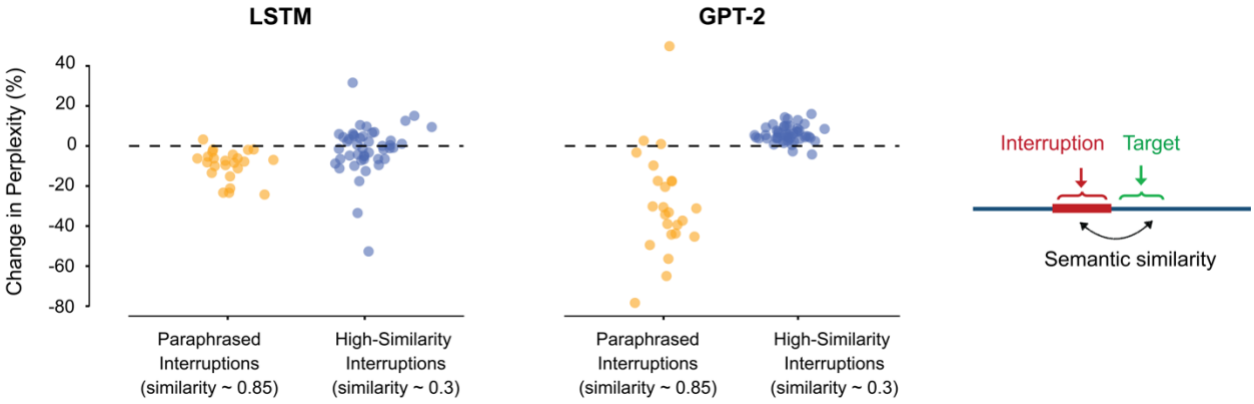


Figure 6-3. Effects of extremely similar interruptions on LSTM and GPT-2.
Left) Change in perplexity of LSTM. Blue points show data for interruptions that are randomly selected from corpus with similarity score around 0.3. Yellow points show data for interruptions that are paraphrased of the target sentence. Middle) Change in perplexity of GPT-2 for the same interruptions. Right) Schematic of how similarity is captured in current analysis. Each dot shows data from one target sentence. The primary text used in this experiment was the APA article.

The presence of repeated tokens could explain the drastic drop in GPT-2's perplexity when processing the target sentence following a paraphrase. To further our understanding of the effects of semantic similarity on how humans and NLMs process incongruencies, future studies should investigate other paraphrasing methods that convey the same information without using the same tokens. Additionally, they should design interruptions that contain overlapping tokens whose contextual meaning differs across the two instances. By examining the proposed conditions in both humans and NLMs, we can gain better insights into how each model utilizes the semantics of data and how this is distinct from individual tokens.

Additionally, future research should examine other semantic-capturing methods for designing semantically similar and dissimilar material. In our current studies, we extracted semantic properties of data using USE method which may only captures a subset of mostly lexical similarity (Cer et al.

2018). Further research is needed to investigate other semantic-capturing methods, possibly utilizing a more cutting-edge language model, to capture deeper aspects of semantic properties conveyed by a sentence such as situational information rather than just lexical information.

6.3. Conclusions

In this thesis, I pursued four lines of investigation into the costs and benefits of integration and separation for learning and comprehension in humans and machines. The first two studies focused on learning and tested the efficacy of brain-inspired constraints for exploiting temporal correlation in data for higher learning efficiency and for learning temporally interpretable representations. The third and fourth studies examined the performance of modern pretrained neural language models and human behavior when processing information with transient moments of incongruencies.

In summary, our findings have identified mechanisms by which both humans and machines can take advantage of temporal continuity in the environment for learning about, comprehending and predicting our dynamic world. I hope that the methods and results in this dissertation will be useful for future studies investigating learning and comprehension from real-world temporally structured data in both artificial and biological neural networks.

References

Armeni, Kristijan, Christopher Honey, and Tal Linzen. 2022. "Characterizing Verbatim Short-Term Memory in Neural Language Models." http://arxiv.org/abs/2210.13569.

Arora, Tushar, and Ming Bo Cai. 2021. "Learning To Perceive Objects By Prediction," no. Svrhm.

Baldassano, Christopher, Janice Chen, Asieh Zadbood, Jonathan W Pillow, and Kenneth A Norman. 2018. "Discovering Event Structure in Continuous Narrative Perception and Memory" 95 (3): 709–21. https://doi.org/10.1016/j.neuron.2017.06.041.Discovering.

Ben-Yakov, Aya, Neetai Eshel, and Yadin Dudai. 2013. "Hippocampal Immediate Poststimulus Activity in the Encoding of Consecutive Naturalistic Episodes." *Journal of Experimental Psychology: General* 142 (4): 1255–63. https://doi.org/10.1037/a0033558.

Ben-Yakov, Aya, Verity Smith, and Richard Henson. 2021. "The Limited Reach of Surprise: Evidence against Effects of Surprise on Memory for Preceding Elements of an Event." *Psychonomic Bulletin and Review*. https://doi.org/10.3758/s13423-021-01954-5.

Bengio, Yoshua, Jerome Louradour, Ronan Collobert, and Jason Weston. 2009. "Curriculum Learning." *Technical Report* 60 (1): 22–23. https://doi.org/10.7547/87507315-60-1-22.

Bernacchia, Alberto, Hyojung Seo, Daeyeol Lee, and Xiao Jing Wang. 2011. "A Reservoir of Time Constants for Memory Traces in Cortical Neurons." *Nature Neuroscience* 14 (3): 366–72. https://doi.org/10.1038/nn.2752.

Bright, Ian M., Miriam L.R. Meister, Nathanael A. Cruzado, Zoran Tiganj, Elizabeth A. Buffalo, and Marc W. Howard. 2020. "A Temporal Record of the Past with a Spectrum of Time Constants in the Monkey Entorhinal Cortex." *Proceedings of the National Academy of Sciences of the United States of America* 117 (33): 20274–83. https://doi.org/10.1073/PNAS.1917197117.

Broadbent, DE. 1966. "Perception and Communication." *Applied Psychology Unit of the Medical Research Council, Cambridge* 8 (6): 264–69. https://doi.org/10.1108/eb015727.

Budiu, Raluca, and John R. Anderson. 2004. "Interpretation-Based Processing: A Unified Theory of Semantic Sentence Comprehension." *Cognitive Science* 28 (1): 1–44. https://doi.org/10.1016/j.cogsci.2003.10.001.

Cer, Daniel, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, et al. 2018. "Universal Sentence Encoder for English." *EMNLP 2018 - Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Proceedings*, 169–74. https://doi.org/10.18653/v1/d18-2029.

Chaudhuri, Rishidev, Kenneth Knoblauch, Marie Alice Gariel, Henry Kennedy, and Xiao Jing Wang. 2015. "A Large-Scale Circuit Mechanism for Hierarchical Dynamical Processing in the Primate Cortex." *Neuron* 88 (2): 419–31. https://doi.org/10.1016/j.neuron.2015.09.008.

Chien, Hsiang Yun Sherry, and Christopher J. Honey. 2020. "Constructing and Forgetting Temporal Context in the Human Cerebral Cortex." *Neuron* 106 (4): 675-686.e11. https://doi.org/10.1016/j.neuron.2020.02.013.

Chomsky, Noam. 1965. "Aspects of the Theory of Syntax." *The MIT Press*, 827–30.

Cocchi, Luca, Leonardo L. Gollo, Andrew Zalesky, and Michael Breakspear. 2017. "Criticality in the Brain: A Synthesis of Neurobiology, Models and Cognition." *Progress in Neurobiology* 158: 132–52. https://doi.org/10.1016/j.pneurobio.2017.07.002.

Cowan, Nelson. 2017. "The Many Faces of Working Memory and Short-Term Storage." *Psychonomic Bulletin and Review* 24 (4): 1158–70. https://doi.org/10.3758/s13423-016-1191-6.

Delaney, Peter F., and K. Anders Ericsson. 2016. "Long-Term Working Memory and Transient Storage in Reading Comprehension: What Is the Evidence? Comment on Foroughi, Werner, Barragán, and Boehm-Davis (2015)." *Journal of Experimental Psychology: General* 145 (10): 1406–9. https://doi.org/10.1037/xge0000181.

Dodell-Feder, David, Jorie Koster-Hale, Marina Bedny, and Rebecca Saxe. 2011. "FMRI Item Analysis in a Theory of Mind Task." *NeuroImage* 55 (2): 705–12. https://doi.org/10.1016/j.neuroimage.2010.12.040.

Dong, Dawei, and Joseph Atick. 1995. "Statistics of Natural Time-Varying Images." *Network: Computation in Neural Systems* 6 (3): 345–58. https://doi.org/10.1088/0954-898x/6/3/003.

DuBrow, Sarah, and Lila Davachi. 2016. "Temporal Binding within and across Events." *Neurobiology of Learning and Memory* 134: 107–14. https://doi.org/10.1016/j.nlm.2016.07.011.

Elman, Jeffrey L. 1993. "Learning and Development in Neural Networks: The Importance of Starting Small." *Cognition* 48 (1): 71–99. https://doi.org/10.1016/0010-0277(93)90058-4.

Ericsson, K. Anders, and Walter Kintsch. 1995. "Long-Term Working Memory." https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.20.2523&rep=rep1&type=pdf.

Firestone, Chaz. 2020. "Performance vs. Competence in Human–Machine Comparisons." *Proceedings of the National Academy of Sciences of the United States of America* 117 (43): 26562–71. https://doi.org/10.1073/pnas.1905334117.

Foroughi, Cyrus K., Daniela Barragán, and Deborah A. Boehm-Davis. 2016. "Interrupted Reading and Working Memory Capacity." *Journal of Applied Research in Memory and Cognition* 5 (4): 395–400. https://doi.org/10.1016/j.jarmac.2016.02.002.

Foroughi, Cyrus K., Nicole E. Werner, Daniela Barragán, and Deborah A. Boehm-Davis. 2015. "Interruptions Disrupt Reading Comprehension." *Journal of Experimental Psychology: General* 144 (3): 704–9. https://doi.org/10.1037/xge0000074.

Foroughi, Cyrus K., Nicole E. Werner, Daniela Barragán, and Deborah A. Boehm-Davis. 2016. "Multiple Interpretations of Long-Term Working Memory Theory: Reply to Delaney and Ericsson (2016)." *Journal of Experimental Psychology: General* 145 (10): 1410–11.

https://doi.org/10.1037/xge0000221.

Foroughi, Cyrus K., Nicole E. Werner, Ryan McKendrick, David M. Cades, and Deborah A. Boehm-Davis. 2016. "Individual Differences in Working-Memory Capacity and Task Resumption Following Interruptions." *Journal of Experimental Psychology: Learning Memory and Cognition* 42 (9): 1480–88. https://doi.org/10.1037/xlm0000251.

Foster, Brett L., Mohammad Dastjerdi, and Josef Parvizi. 2012. "Neural Populations in Human Posteromedial Cortex Display Opposing Responses during Memory and Numerical Processing." *Proceedings of the National Academy of Sciences of the United States of America* 109 (38): 15514–19. https://doi.org/10.1073/pnas.1206580109.

Gao, Tianxiang, and Vladimir Jojic. 2017. "Sample Importance in Training Deep Neural Networks." *ICLR 2017*, no. 2011: 1–12.

Glanzer, Murray, David Dorfman, and Barbara Kaplan. 1981. "Short-Term Storage in the Processing of Text." *Journal of Verbal Learning and Verbal Behavior* 20 (6): 656–70. https://doi.org/10.1016/S0022-5371(81)90229-2.

Glanzer, Murray, Beth Fischer, and David Dorfman. 1984. "Short-Term Storage in Reading." *Journal of Verbal Learning and Verbal Behavior* 23 (4): 467–86. https://doi.org/10.1016/S0022-5371(84)90300-1.

Glorot, Xavier, and Yoshua Bengio. 2010. "Understanding the Difficulty of Training Deep Feedforward Neural Networks." *Journal of Machine Learning Research* 9: 249–56.

Goldstein, Ariel. 2021. "Thinking Ahead: Spontaneous next Word Predictions in Context as a Keystone of Language in Humans and Machines," 1–43.

Goldstein, Ariel, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, et al. 2022. "Shared Computational Principles for Language Processing in Humans and Deep Language Models." *Nature Neuroscience* 25 (3): 369–80. https://doi.org/10.1038/s41593-022-01026-4.

Grill-Spector, Kalanit, and Kevin S. Weiner. 2013. "The Functional Architecture of the Ventral Temporal Cortex and Its Role in Categorization." *Nature Reviews Neuroscience* 23 (1): 1–7. https://doi.org/10.1038/nrn3747.The.

Hasson, U., E. Yang, I. Vallines, D. J. Heeger, and N. Rubin. 2008. "A Hierarchy of Temporal Receptive Windows in Human Cortex." *Journal of Neuroscience* 28 (10): 2539–50. https://doi.org/10.1523/JNEUROSCI.5487-07.2008.

Hasson, Uri, Janice Chen, and Christopher J. Honey. 2015. "Hierarchical Process Memory: Memory as an Integral Component of Information Processing." *Trends in Cognitive Sciences* 19 (6): 304–13. https://doi.org/10.1016/j.tics.2015.04.006.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. "Deep Residual Learning for Image Recognition." *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2016-Decem: 770–78. https://doi.org/10.1109/CVPR.2016.90.

Hénaff, Olivier J., Robbe L.T. Goris, and Eero P. Simoncelli. 2019. "Perceptual Straightening of Natural Videos." *Nature Neuroscience* 22 (6): 984–91. https://doi.org/10.1038/s41593-019-0377-4.

Hochreiter, Sepp, and Jurgen Schmidhuber. 1997. "Long Short-Term Memory." *Neural Computation*.

Honey, Christopher J., Ehren L. Newman, and Anna C. Schapiro. 2017. "Switching between Internal and External Modes: Amultiscale Learning Principle." *Network Neuroscience* 104 (51): 20167–72. https://doi.org/10.1073/pnas.0709640104.

Honey, Christopher J., Thomas Thesen, Tobias H. Donner, Lauren J. Silbert, Chad E. Carlson, Orrin Devinsky, Werner K. Doyle, Nava Rubin, David J. Heeger, and Uri Hasson. 2012. "Slow Cortical Dynamics and the Accumulation of Information over Long Timescales." *Neuron* 76 (2): 423–34. https://doi.org/10.1016/j.neuron.2012.08.011.

Illing, Bernd, Wulfram Gerstner, and Johanni Brea. 2019. "Biologically Plausible Deep Learning — But How Far Can We Go with Shallow Networks?" *Neural Networks* 118: 90–101. https://doi.org/10.1016/j.neunet.2019.06.001.

Ilya Sutskever. 2013. "Training Recurrent Neural Networks." *Ph.D Thesis*.

Jain, Shailee, Vy A. Vo, Shivangi Mahto, Amanda LeBel, Javier S. Turek, and Alexander G. Huth. 2020. "Interpretable Multi-Timescale Models for Predicting FMRI Responses to Continuous Natural Speech." *Advances in Neural Information Processing Systems* 2020-Decem (NeurIPS): 1–12.

Kelly, Matthew A., and David Reitter. 2018. "How Language Processing Can Shape a Common Model of Cognition." *Procedia Computer Science* 145: 724–29. https://doi.org/10.1016/j.procs.2018.11.047.

Khandelwal, Urvashi, He He, Peng Qi, and Dan Jurafsky. 2018. "Sharp Nearby, Fuzzy Far Away: How Neural Language Models Use Context." *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)* 1: 284–94. https://doi.org/10.18653/v1/p18-1027.

Kietzmann, Tim C., Courtney J. Spoerer, Lynn K.A. Sörensen, Radoslaw M. Cichy, Olaf Hauk, and Nikolaus Kriegeskorte. 2019. "Recurrence Is Required to Capture the Representational Dynamics of the Human Visual System." *Proceedings of the National Academy of Sciences of the United States of America* 116 (43): 21854–63. https://doi.org/10.1073/pnas.1905544116.

Kintsch, Walter, and Teun A. van Dijk. 1978. "Toward a Model of Text Comprehension and Production." *Psychological Review* 85 (5): 363–94. https://doi.org/10.1037/0033-295X.85.5.363.

Kozachkov, Leo, John Tauber, Mikael Lundqvist, Scott L Brincat, Jean-Jacques Slotine, and Earl K Miller. 2022. "Robust Working Memory through Short-Term Synaptic Plasticity." *BioRxiv*, 2022.01.09.475558. https://www.biorxiv.org/content/10.1101/2022.01.09.475558v1%0Ahttps://www.biorxiv.org/content/10.1101/2022.01.09.475558v1.abstract.

Kumar, M. Pawan, Benjamin Packer, and Daphne Koller. 2010. "Self-Paced Learning for Latent

Variable Models." *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010, NIPS 2010*, 1–9.

Larsen, Janet D., and Alan Baddeley. 2003. "Disruption of Verbal STM by Irrelevant Speech, Articulatory Suppression, and Manual Tapping: Do They Have a Common Source?" *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology* 56 (8): 1249–68. https://doi.org/10.1080/02724980244000765.

Lawrence, Zachary, and Daniel Peterson. 2016. "Mentally Walking through Doorways Causes Forgetting: The Location Updating Effect and Imagination." *Memory* 24 (1): 12–20. https://doi.org/10.1080/09658211.2014.980429.

Le, Xuan Hien, Hung Viet Ho, Giha Lee, and Sungho Jung. 2019. "Application of Long Short-Term Memory (LSTM) Neural Network for Flood Forecasting." *Water (Switzerland)* 11 (7). https://doi.org/10.3390/w11071387.

LeCun, Yann, Koray Kavukcuoglu, and Clément Farabet. 2010. "Convolutional Networks and Applications in Vision." *ISCAS 2010 - 2010 IEEE International Symposium on Circuits and Systems: Nano-Bio Circuit Fabrics and Systems*, no. May: 253–56. https://doi.org/10.1109/ISCAS.2010.5537907.

Ledoux, Kerry, and Peter C. Gordon. 2006. "Interruption-Similarity Effects during Discourse Processing." *Memory* 14 (7): 789–803. https://doi.org/10.1080/09658210600679915.

Lee, Dong Hyun, Saizheng Zhang, Asja Fischer, and Yoshua Bengio. 2015. "Difference Target Propagation." *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 9284: 498–515. https://doi.org/10.1007/978-3-319-23528-8_31.

Li, Mu, Tong Zhang, Yuqiang Chen, and Alexander J. Smola. 2014. "Efficient Mini-Batch Training for Stochastic Optimization." *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 661–70. https://doi.org/10.1145/2623330.2623612.

Lillicrap, Timothy P., Adam Santoro, Luke Marris, Colin J. Akerman, and Geoffrey Hinton. 2020. "Backpropagation and the Brain." *Nature Reviews Neuroscience* 21 (6): 335–46. https://doi.org/10.1038/s41583-020-0277-3.

Lillicrap, Timothy P, and Adam Santoro. 2019. "Backpropagation through Time and the Brain." *Current Opinion in Neurobiology* 55: 82–89. https://doi.org/10.1016/J.CONB.2019.01.011.

Linzen, Tal, Emmanuel Dupoux, and Yoav Goldberg. 2016. "Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies." *ACL* 4 (1990): 521–35.

Logie, Robert H., Gianna Cocchini, Sergio Della Sala, and Alan D. Baddeley. 2004. "Is There a Specific Executive Capacity for Dual Task Coordination? Evidence from Alzheimer's Disease." *Neuropsychology* 18 (3): 504–13. https://doi.org/10.1037/0894-4105.18.3.504.

Mahto, Shivangi, Vy A. Vo, Javier S. Turek, and Alexander G. Huth. 2020. "Multi-Timescale Representation Learning in LSTM Language Models," 1–19. http://arxiv.org/abs/2009.12727.

Malmberg, Kenneth J., Jeroen G.W. Raaijmakers, and Richard M. Shiffrin. 2019. "50 Years of Research Sparked by Atkinson and Shiffrin (1968)." *Memory and Cognition* 47 (4): 561–74. https://doi.org/10.3758/s13421-019-00896-7.

Ming, Yao, Shaozu Cao, Ruixiang Zhang, Zhen Li, Yuanzhe Chen, Yangqiu Song, and Huamin Qu. 2018. "Understanding Hidden Memories of Recurrent Neural Networks." *2017 IEEE Conference on Visual Analytics Science and Technology, VAST 2017 - Proceedings*, 13–24. https://doi.org/10.1109/VAST.2017.8585721.

Mitchell, Melanie. 2020. "On Crashing the Barrier of Meaning in Artificial Intelligence." *AI Magazine* 41 (2): 86–92. https://doi.org/10.1609/aimag.v41i2.5259.

Monk, Christopher A., J. Gregory Trafton, and Deborah A. Boehm-Davis. 2008. "The Effect of Interruption Duration and Demand on Resuming Suspended Goals." *Journal of Experimental Psychology: Applied* 14 (4): 299–313. https://doi.org/10.1037/a0014402.

Mozer, M, R Lippmann, J Moody, and D Touretsky. 1992. "Induction of Multiscale Temporal Structure." *Advances in Neural Information Processing Systems 4*, 275–82.

Murray, John D., Alberto Bernacchia, David J. Freedman, Ranulfo Romo, Jonathan D. Wallis, Xinying Cai, Camillo Padoa-Schioppa, et al. 2014. "A Hierarchy of Intrinsic Timescales across Primate Cortex." *Nature Neuroscience* 17 (12): 1661–63. https://doi.org/10.1038/nn.3862.

Nayebi, Aran, Daniel Bear, Jonas Kubilius, Kohitij Kar, Surya Ganguli, David Sussillo, James J. DiCarlo, and Daniel L.K. Yamins. 2018. "Task-Driven Convolutional Recurrent Models of the Visual System." *Advances in Neural Information Processing Systems* 2018-Decem (Nips): 5290–5301.

O'Doherty, Cliona, and Rhodri Cusack. 2022. "Objects or Context? Learning From Temporal Regularities in Continuous Visual Experience With an Infant-Inspired DNN." https://doi.org/10.32470/ccn.2022.1093-0.

Payne, Jessica D., Eric D. Jackson, Siobhan Hoscheidt, Lee Ryan, W. Jake Jacobs, and Lynn Nadel. 2007. "Stress Administered Prior to Encoding Impairs Neutral but Enhances Emotional Long-Term Episodic Memories." *Learning and Memory* 14 (12): 861–68. https://doi.org/10.1101/lm.743507.

Radford, Alec, * Jeffrey Wu, Rewon Child 1, David Luan, ** Dario Amodei, and Ilya Sutskever. 2020. "Language Models Are Unsupervised Multitask Learners." *OpenAI Blog* 1 (May): 1–7. https://github.com/codelucas/newspaper.

Radvansky, Gabriel A., and David E. Copeland. 2010. "Reading Times and the Detection of Event Shift Processing." *Journal of Experimental Psychology: Learning Memory and Cognition* 36 (1): 210–16. https://doi.org/10.1037/a0017258.

Raut, Ryan V., Abraham Z. Snyder, and Marcus E. Raichle. 2020. "Hierarchical Dynamics as a Macroscopic Organizing Principle of the Human Brain." *Proceedings of the National Academy of Sciences of the United States of America* 117 (34): 20890–97. https://doi.org/10.1073/pnas.2003383117.

Robert M. French. 1999. "Catastrophic Forgetting in Connectionists Networks." *Trends in Cognitive Sciences* 3 (4): 128–35.

Rubin, David C., Robert W. Schrauf, and Daniel L. Greenberg. 2003. "Belief and Recollection of Autobiographical Memories." *Memory and Cognition* 31 (6): 887–901. https://doi.org/10.3758/BF03196443.

Ruder, Sebastian. 2016. "An Overview of Gradient Descent Optimization Algorithms." http://arxiv.org/abs/1609.04747.

Schijndel, Marten Van, and Tal Linzen. 2018. "A Neural Model of Adaptation in Reading." *ACL*, no. 1.

Schijndel, Marten Van, Aaron Mueller, and Tal Linzen. 2019. "Quantity Doesn ' t Buy Quality Syntax with Neural Language Models," 5831–37.

Shin, Seongjin, Sang Woo Lee, Hwijeen Ahn, Sungdong Kim, Hyoung Seok Kim, Boseop Kim, Kyunghyun Cho, et al. 2022. "On the Effect of Pretraining Corpora on In-Context Learning by a Large-Scale Language Model." *NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, 5168–86. https://doi.org/10.18653/v1/2022.naacl-main.380.

Smith, Nathaniel J., and Roger Levy. 2013. "The Effect of Word Predictability on Reading Time Is Logarithmic." *Cognition* 128 (3): 302–19. https://doi.org/10.1016/j.cognition.2013.02.013.

Smith, Stephen M., Diego Vidaurre, Christian F. Beckmann, Matthew F. Glasser, Mark Jenkinson, Karla L. Miller, Thomas E. Nichols, et al. 2013. "Functional Connectomics from Resting-State FMRI." *Trends in Cognitive Sciences* 17 (12): 666–82. https://doi.org/10.1016/j.tics.2013.09.016.

Soltani, Alireza, John D. Murray, Hyojung Seo, and Daeyeol Lee. 2021. "Timescales of Cognition in the Brain." *Current Opinion in Behavioral Sciences* 41: 30–37. https://doi.org/10.1016/j.cobeha.2021.03.003.

Spoerer, Courtney J, Patrick McClure, and Nikolaus Kriegeskorte. 2017. "Recurrent Convolutional Neural Networks: A Better Model of Biological Object Recognition." *Frontiers in Psychology* 8: 1551. https://doi.org/10.3389/fpsyg.2017.01551.

Spreng, R Nathan, Raymond A Mar, and Alice S N Kim. 2008. "The Common Neural Basis of Autobiographical Memory, Prospection, Navigation, Theory of Mind, and the Default Mode: A Quantitative Meta-Analysis," 489–510.

Strobelt, Hendrik, Sebastian Gehrmann, Hanspeter Pfister, and Alexander M. Rush. 2018. "LSTMVis: A Tool for Visual Analysis of Hidden State Dynamics in Recurrent Neural Networks." *IEEE Transactions on Visualization and Computer Graphics* 24 (1): 667–76. https://doi.org/10.1109/TVCG.2017.2744158.

Subramanian, Sandeep, Jonathan Pilault, Raymond Li, and Chris Pal. 2020. "On Extractive and Abstractive Neural Document Summarization with Transformer Language Models," 9308–19. https://doi.org/10.18653/v1/2020.emnlp-main.748.

Swallow, Khena M., Jeffrey M. Zacks, and Richard A. Abrams. 2009. "Event Boundaries in Perception Affect Memory Encoding and Updating." *Journal of Experimental Psychology: General* 138 (2): 236–57. https://doi.org/10.1037/a0015631.

Tieleman, T., and Geoffrey Hinton. 2012. "Divide the Gradient by a Running Average of Its Recent Magnitude." *COURSERA: Neural Networks for Machine Learning - RMSprop Lecture* 04 (05): 107–11. https://doi.org/10.4236/jcc.2016.45016.

Treisman, Anne. 1964. "Monitoring and Storage of Irrelevant Messages in Selective Attention." *Journal of Verbal Learning and Verbal Behavior* 3 (6): 449–59. https://doi.org/10.1016/S0022-5371(64)80015-3.

Ulanovsky, Nachum, Liora Las, Dina Farkas, and Israel Nelken. 2004. "Multiple Time Scales of Adaptation in Auditory Cortex Neurons." *Journal of Neuroscience* 24 (46): 10440–53. https://doi.org/10.1523/JNEUROSCI.1905-04.2004.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." *Advances in Neural Information Processing Systems* 2017-Decem (Nips): 5999–6009.

Wang, Jianfeng, and Xiaolin Hu. 2021. "Convolutional Neural Networks with Gated Recurrent Connections." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–16. https://doi.org/10.1109/TPAMI.2021.3054614.

Wiskott, Laurenz, and Terrence J Sejnowski. 2002. "Slow Feature Analysis: Unsupervised Learning of Invariances." *Neural Computation* 14 (4): 715–70. https://doi.org/10.1162/089976602317318938.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, et al. 2020. "Transformers: State-of-the-Art Natural Language Processing," 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6.

Xiao, Han, Kashif Rasul, and Roland Vollgraf. 2017. "Fashion-MNIST: A Novel Image Dataset for Benchmarking Machine Learning Algorithms," 1–6. http://arxiv.org/abs/1708.07747.

Yamins, Daniel L K, and James J. DiCarlo. 2016. "Using Goal-Driven Deep Learning Models to Understand Sensory Cortex." *Nature Neuroscience* 19 (3): 356–65. https://doi.org/10.1038/nn.4244.

Zhang, Jingqing, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. "PEGASUS: Pre-Training with Extracted Gap-Sentences for Abstractive Summarization." *37th International Conference on Machine Learning, ICML 2020* PartF16814: 11265–76.

Zheng, Jie, Andrea G.P. Schjetnan, Mar Yebra, Bernard A. Gomes, Clayton P. Mosher, Suneil K. Kalia, Taufik A. Valiante, Adam N. Mamelak, Gabriel Kreiman, and Ueli Rutishauser. 2022. "Neurons Detect Cognitive Boundaries to Structure Episodic Memories in Humans." *Nature Neuroscience* 25 (3): 358–68. https://doi.org/10.1038/s41593-022-01020-w.