THESIS

USING NATURAL LANGUAGE PROCESSING TO CHARACTERIZE THE PHENOMENOLOGY OF DEJA VU

Submitted by

Sarah N. Horne

Department of Psychology

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Spring 2026

Master's Committee:

Advisor: Anne Cleary

Matthew Rhodes
Nathaniel Blanchard
Joshua Prasad

ABSTRACT

USING NATURAL LANGUAGE PROCESSING TO CHARACTERIZE THE PHENOMENOLOGY OF DÉJÀ VU

Around two decades ago, cognitive psychologists began empirically studying déjà vu—the odd sensation that one is re-experiencing something from the past while being certain of the situation's novelty. Prior work, investigating déjà vu's subjective phenomenology by soliciting retrospective reports in a survey, uncovered déjà vu was relatively positive in its emotionality, and involved language pertaining to places, consistent with past survey research suggesting that scenes are the most common elicitor of déjà vu. However, there are well-documented limitations to relying on retrospective memories of an experience for studying it. The present thesis aimed to instead examine language that occurred around the moment that a déjà vu experience happened. Natural language processing (NLP) was performed on transcribed spoken language that occurred in a Think Aloud protocol during a Virtual Reality implementation of a well-established method for studying déjà vu—the Virtual Tour Method. Machine learning (ML) classification models were trained to distinguish déjà vu vs. non-déjà vu reports and found that déjà vu was characterized by increased filler words and recollective confabulation (i.e. recollection of incorrect episodic details). In line with prior work, trials where participants reported déjà vu were more positive, but that was not an important feature in the ML models predictions. Additionally, spatial features of the environment were not influential to the models' predictions. These results help characterize more of the déjà vu experience and open a new direction for exploration into the relationship between internal attention and déjà vu.

TABLE OF CONTENTS

CHAPTER 1 - INTRODUCTION

Déjà Vu

For most, occasional déjà vu, the feeling of familiarity for a novel experience, is a standard part of the human experience. Surveys consistently show that the majority of individuals report experiencing déjà vu (Brown, 2003; Cleary & Brown, 2022). Yet despite being something most people have experienced, the phenomenon attracted little intellectual attention until the mid-1900s, and even then, it was primarily survey research. Previously cast aside by scientists for its paranormal associations or lack of observable behavioral correlate (e.g., Cleary et al., 2020), today, researchers are beginning to understand how this understudied phenomenon can reveal details of metacognitive and attentional processes (Cleary & Brown, 2022).

A comprehensive definition of déjà vu is difficult to come by. The phenomenon is characterized by a conflicting feeling of familiarity accompanying a sense that the present situation is a novel scene or experience. Déjà vu has been frequently associated with other similar subjective metamemorial experiences such as jamais vu (the feeling of unwarranted novelty for a familiar situation), déjà vécu (having lived an exact situation before), and déjà rêvé (having dreamed something before; Cleary & Brown, 2022). Some argue that these are separate indices of the same 'déjà vu' experience (Perrin et al., 2023), yet some research should be done to understand the similarities and differences amongst this family of phenomena.

Research supports the idea that familiarity-detection plays a key role in the feeling of deja vu (Cleary et al., 2009, 2012, 2021, Cleary & Brown, 2022, Okada et al., 2024), but the exact mechanism of its involvement is not yet fully understood. One theory, known as the Gestalt familiarity hypothesis, posits that déjà vu arises from an overall sense of familiarity for the spatial organization of a scene (Cleary et al., 2009, 2012, 2021; Huebert et al., 2025, Okada et

al., 2024). On the other hand, the elemental familiarity hypothesis suggest that déjà vu is a result of familiarity for either a single object or multiple objects in the environment (Cleary & Brown, 2022). Multiple studies using experiments derived from recognition memory paradigms have supported the former hypothesis, though at least one study reported in Cleary and Brown (2022) supports the latter as well; importantly, they are not mutually exclusive hypotheses, and it is likely that both a familiar overall Gestalt and familiar elements in the environment can contribute.

A History of the Study of Déjà Vu

As mentioned, one key barrier to déjà vu's entry into the scientific literature was the difficulty of empirically studying it. In particular, the rise of Behaviorism in the early 1900s saw an emphasis on observable measurable behaviors alongside a dismissal of theories regarding unobservable underlying subjective cognitive phenomena. Even after the cognitive revolution in the 1950s and 1960s, which saw renewed interest in studying cognitive phenomena, cognitive psychologists largely avoided the topic of déjà vu (Cleary et al., 2020). Instead, most of the early work on déjà vu was based on retrospective reports and extensive surveys (Brown, 2003; Cleary & Brown, 2022). Experimental approaches to its study did not take flight until after Brown's (2003) review on the topic (Cleary et al., 2020; Cleary & Brown, 2022).

Prior to 1994, when Sno and colleagues developed the first survey dedicated to the déjà vu experience, most data were collected from surveys that had few items that mentioned déjà vu (Sno et al., 1994). In his 2004 book, Brown recounts how in these prior surveys and analyses, the déjà vu item's positioning could have impacted people's reports and understanding of the déjà vu experience itself. For example, the National Opinion Research Center (NORC) at the University of Chicago published a report that included a survey item on déjà vu, but it was embedded within

questions on clairvoyance, Extra Sensory Perception (ESP), contact with the dead and out of body experiences (NORC, 1984, as cited in Brown, 2003). Clearly, if déjà vu is accompanied by paranormal items in these surveys, individuals' biases for or against paranormal activities may influence responses to déjà vu items.

Sno and colleagues (1994) developed the Inventory for Déjà vu Experiences Assessment (IDEA) which assesses individuals' subjective experiences with déjà vu. The inventory contains 23 questions, split into two sections. The first section, containing 8 items, asks about other phenomena related to déjà vu, such as jamais vu (Sno et al., 1994). Critics have noted that of these 8 items, there are still paranormal related items, such as "do you consider yourself someone with paranormal qualities," which could have considerable priming effects for the remainder of the survey (Cleary & Brown, 2022). Thus, some researchers have opted to only use the single item about déjà vu from the first part of the survey to avoid such effects (Okada et al., 2024) The second half of the Sno et al. (1994) survey asks individuals, who have experienced déjà vu, questions about the specifics of their experiences, such as the circumstances in which deja vu arises, degree of familiarity, duration, as well as the time of day. These items have helped to quantify the déjà vu experience globally as the survey have been widely used and validated across different languages (Adachi et al., 2001; Mumoli et al., 2017).

The assessment's development was a key step toward an experimental study of déjà vu, as it led to a number of survey findings over many years that were summarized by Brown (2003) in bringing the study of déjà vu into the mainstream scientific realm. In his summary, Brown also reviewed several century-old theories of déjà vu and presented ways that testable hypotheses could be derived and studied in laboratory settings.

The next major step came when Cleary et al. (2009) developed a novel scene-based paradigm to test the hypothesis, borne out of the Gestalt familiarity theory, that a similar overall Gestalt to an unidentified scene previously experienced will increase the probability that a person will report an experience of déjà vu (relative to a dissimilar overall Gestalt). Cleary et al. (2009) sought to extend the recognition without cued recall (RWCR) phenomenon to scene-based instead of verbal stimuli (e.g., Cleary, 2004; Ryals & Cleary, 2012). In studies of RCWR, participants are shown a set of stimuli during a study phase, then during a subsequent test phase, participants are shown another set of stimuli with varying overlap to studied cues. Participants consistently show familiarity detection for studied but unrecalled items. This paradigm was then extended for the stimuli to be scenes, or locations- a common elicitor of déjà vu according to retrospective reports (Cleary & Brown, 2022; Venkatesha et al., 2025). Cleary et al. (2009) also examined whether Gestalt similarity would not only increase perceived familiarity during recall failure but also reports of déjà vu. To test this, Cleary et al. (2009) developed spatially similar black and white line drawings of various scenes (Figure 1). In four study-test blocks altogether, during a study session, participants saw 15 scenes, then viewed 30 scenes during the test session. Of these 30 scenes, half were spatially similar to studied scenes and half were not. Following

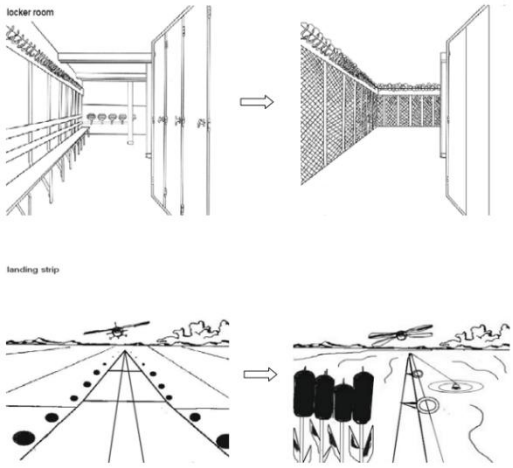

Figure 1: Example scenes from Cleary et al. (2009)

each test scene, participants were asked to recall a similar scene, rate the scenes familiarity, report if they've experienced déjà vu, and finally were given another chance to recall a similar scene. As hypothesized, the results showed that participants reported higher familiarity when in déjà vu like states. Participants were more likely to report deja vu for scenes which had been experimentally familiarized (i.e. the scene is more familiar within the context of the experiment, as different scenes could have different levels of global familiarity; Cleary et al., 2009). This study became the first of its kind to develop an experimental design that elicited déjà vu in the laboratory setting.

Cleary and colleagues would continue to develop this paradigm further to include immersive three-dimensional scenes using virtual reality (VR; Cleary et al., 2012, Okada et al., 2024). Through these experiments, researchers have been able to show that déjà vu, elicited in a laboratory setting, can be driven by familiar stimuli in the environment. The spatial design of this study corroborates the Gestalt familiarity-based hypotheses on déjà vu (See Figure 2).

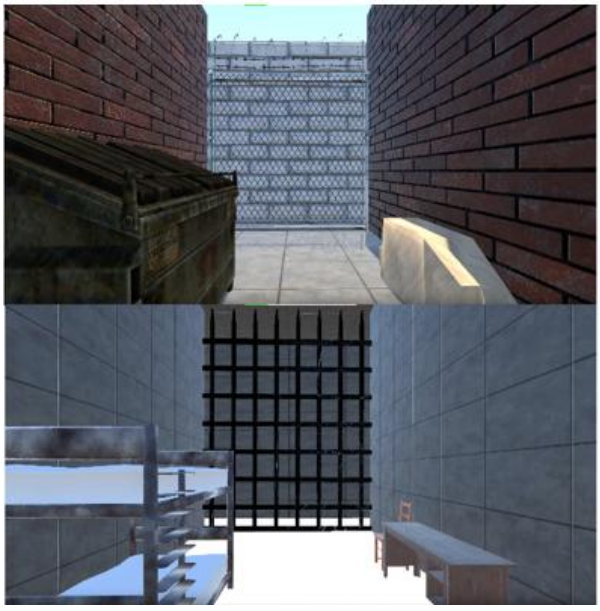

Figure 2: Example Scenes 'Alleyway' and 'Jail Cell'

Phenomenology of Deja Vu

With an established laboratory-based déjà vu paradigm, more recent work has begun to examine the phenomenological characteristics of déjà vu. One experience often associated with déjà vu is the feeling of prediction or knowing what would happen next. Cleary and Claxton (2018) empirically tested this using virtual tours that ended in either a left or right turn. During the test phase of the experiment, when the scene stopped just short of the turn, participants were asked if they had a feeling of knowing which way to turn next and to predict which way the scene would turn. When reporting déjà vu, participants were not significantly better at predicting the next turn, although they *thought* they did (Cleary & Claxton, 2018). Cleary et al. (2018) theorized that familiarity-detection may be driving this aspect of the déjà vu experience. In other words, the feeling of familiarity was leading to the increase in the likelihood of thinking they would know what happens next. To test this idea, Cleary et al. (2018) asked participants to rate their feeling of familiarity on a scale of 0 (no familiarity) to 10 (high familiarity), as opposed to a binary yes/no question. They found that when déjà vu was reported, participants reported higher familiarity when also reporting a feeling of prediction (i.e. knowing which way the scene will turn; Cleary et al., 2018).

Building on this, Huebert et al. (2025) implemented different degrees of experimental familiarization into the classic déjà vu paradigm. When participants were previously presented with a scene with a similar spatial layout, they were significantly more likely to report a sense of knowing what would happen next. Moreso, when participants had seen the spatially similar scene three times (increased experimental familiarity), they were more likely to report a feeling of prediction than when they had only seen the previous scene once (Huebert et al., 2025). Taken

together, these studies have begun to reveal how the feeling of prediction (and actual predictive abilities) is associated with the déjà vu experience.

When amidst a déjà vu state, individuals will often try to recall the source of the déjà vu or explain why it is occurring. Using the same paradigm, McNeely-White and Cleary (2023) investigated how déjà vu reports are associated with curiosity and information-seeking behaviors. Using Cleary et al's (2012) paradigm, participants were given a limited number of trials in which they could choose to spend more time trying to recall the answer, or have the answer revealed to them. Their results indicated that when reporting déjà vu, participants used more resources to discover the correct answer (McNeely-White & Cleary, 2023). The authors hypothesized that metacognitive judgements (possibly of familiarity) may be driving these aspects of déjà vu. In other words, when a metacognitive signal indicates the possibility of a relevant memory, the participants may spend more time and energy towards retrieving said memory. This explanation is supported by the trend of increased commission errors, frequently seen in RCWR-based paradigms (Huebert et al., 2022, McNeely-White & Cleary, 2023, Okada et al., 2024). In these studies, when a participant detects familiarity, either through subjective report or déjà vu report, they are more likely to produce a commission error (incorrect answer) than an omission error (no answer). This could be interpreted as instances of healthy recollective confabulation (henceforth referred to as confabulation for brevity; Coltheart, 2017)

Overall, with the development of the Virtual Tour paradigm, researchers have begun to investigate the characteristics, qualities and phenomenology of déjà vu, informed by prior survey research. This work informs our understanding of déjà vu, and how it may reveal more about the basic functions of our cognitive systems, including metacognitive processes. Further work will

continue to investigate how subjective parts of the déjà vu experience can be manifested and manipulated in the laboratory.

Déjà Vu as a Form of Involuntary Thought

A potentially productive new direction for the study of déjà vu is to explore it from the perspective that it is a kind of involuntary or spontaneous thought. The scientific literature on spontaneous thought incorporates a wide variety of phenomena, including mind wandering, involuntary autobiographical memories (IAM), and unexpected thought (e.g., Poulus et al., 2023, Steadman et al., 2025). Déjà vu could be considered a form of spontaneous or involuntary thought in that it occurs in life without intention. Although eliciting déjà vu in experimental paradigms might make exploration of déjà vu in these paradigms seem voluntary, the feeling of déjà vu itself is still involuntary for participants, even when it occurs during a memory retrieval task. This is because a person cannot *intend* to feel déjà vu—it happens without a person's volitional control. Thus, although we now know that the likelihood of déjà vu can be increased with certain experimental manipulations (e.g., Cleary et al., 2012; Cleary & Claxton, 2018; Okada et al., 2024), déjà vu should still be considered a type of involuntary thought experience.

A key interest when investigating involuntary thought is what drives the mind inward. Although the phrase "involuntary thought" might seem to imply that there may not be a cue that prompts the thought to occur, the term simply denotes the lack of conscious cognitive control. In fact, research suggests that involuntary thought is often driven by cues, even when those cues cannot be consciously identified by the individual (e.g., Poulos et al., 2023). As far as how and why a cue in the environment can send the mind inward, in a novel theory, Cleary et al. (2023a) propose that cue familiarity-detection could be a driving force that 'flips' attention inward toward memory search. Cleary et al. (2023a) note that most of the attention research has focused

on what pulls attention outwards, but not what pushes it inwards. As déjà vu is thought to involve familiarity-detection as a component mechanism (e.g., Cleary et al., 2021; Huebert et al., 2025), understanding the causes and mechanisms of déjà vu can potentially further illuminate how attention is modulated to make switches back and forth between an external and an internal orientation. Indeed, research on déjà vu has found indicators that it is associated with likely proxies of internally oriented memory search effort, including increased commission errors and use of limited resources to resolve conflict (McNeely-White & Cleary, 2023; Okada et al., 2024). These findings suggest that the déjà vu experience itself might be a modulator of attention, directing it inward toward memory.

A recent theoretical paper from Barzykowski and Moulin (2023) was the first to formally discuss déjà vu as a form of involuntary thought. They proposed the novel theory that IAM and déjà vu are two related outcomes of a common set of involuntary memory mechanisms. Specifically, they proposed that IAM occurs when a specific memory of a past experience involuntarily comes to mind in response to a cue or set of cues in the environment, while déjà vu occurs in response to a cue or set of cues in the environment but without the retrieval of specific content (see also Cleary et al., 2023a). They argue that déjà vu and IAM lie on different ends of a continuum. While a novel theory, their theory does not account for the phenomenological differences between the two types of subjective memory experiences, which seem more distinct from one another than just in terms of the strength of the experience on a continuum. The two subjective experiences differ greatly in their frequency and subjective qualities, such as surprise, oddness and salience (Andonovski & Michaelian, 2025). Nonetheless, their theory emphasized the value of investigating how involuntary memory experiences are related to each other. Is it possible that déjà vu, IAM, and other involuntary thought experiences result from common

underlying mechanisms with differing subjective metacognitive experiences (e.g., Schwartz & Cleary, 2025)? If so, what differentiates the subjective experiences? How can the study of these distinct phenomena, in tandem, influence our understanding of memory and metacognition more broadly? Addressing these questions is a goal of the present study. Before turning to how the present study will examine what characterizes déjà vu phenomenology and subjective experience, it is important to first consider how researchers go about studying involuntary forms of thought more generally.

Empirical Methods for Studying Involuntary Thought

As mentioned, although familiar scenes can lead to increased reporting of déjà vu, gaps still exist regarding the phenomenology and temporal unfolding of the déjà vu experience within current laboratory paradigms. Currently, déjà vu paradigms use quantitative self-reports within an experimental design, typically a design that is aimed at boosting the probability of reporting a déjà vu experience based on theoretical grounds (e.g., Cleary et al., 2012; Okada et al., 2024). When aiming to capture involuntary thoughts, researchers recognize the need for a more open-ended, linguistically rich response from participants that allows for an assessment and analysis of what people are actively thinking about at any given moment so that researchers can examine how thoughts unfold over time and what qualities tend to characterize different types of thought.

One commonly used method of studying involuntary thought is to probe participants at various points throughout tasks and ask about their current thoughts. This is known as the 'probe-caught' method, often used in both laboratory and naturalistic environments (see Weinstein, 2018, for a review). One study by Killingsworth and Gilbert (2010), which probed participants using a smartphone app throughout the day, found that mind wandering occurred in almost half of their time points sampled. After analyzing the emotions behind each sample, they

concluded that people thinking about other things 'typically makes them unhappy,' and 'a wandering mind is an unhappy mind,' (Killingsworth & Gilbert, 2010). While their study cannot infer causality, their approach of probing during naturalistic activity demonstrates how these methods could be used in studying involuntary thought. Of course, this method comes with limitations, such as the limitation that the frequency of probes can itself potentially moderate how frequently mind wandering occurs during the paradigm, or the possibility that one cannot fully and accurately verbalize what was occurring internally in the moment that the probe appeared.

Another common method is a 'Think Aloud' protocol, whereby participants are asked to speak aloud all their thoughts during a task (see Ericsson & Simons, 1998). Think Aloud studies can be used to analyze the structure of thoughts and how thoughts flow from one topic to the next. For example, Sripada and Taxali (2020) employed a Think Aloud protocol to analyze patterns underlying streams of consciousness thought. They found that participants 'exhibited a clump-and-jump pattern with a block of semantically related thoughts about a topic followed by a jump to a new topic, in a repeating pattern,' (Sripada & Taxali, 2020, p. 5).

The Think Aloud method has similar limitations to the probe-caught method. Primarily, participants are likely not reporting all conscious thought. It is reasonable to assume that when speaking aloud, participants cannot verbalize every thought in their conscious mind. In addition, Think Aloud studies require extensive labor to transcribe and process more free-form thoughts, as opposed to probe methods which may have a more structured, predictable response. As such, these studies often involve lower sample sizes than other approaches like the probe-caught method. Unlike the probe-caught method, Think Aloud experiments also must take place in a laboratory setting, which limits the contexts in which it can be used. However, compared to

traditional yes-no responses or ratings measures in laboratory settings, these methods capture more richness of overall thoughts of the participants, as well as the temporal unfolding of their thoughts over the course of the experiment.

Using Natural Language Processing (NLP) to Study Involuntary Thought

NLP Algorithms and Methodologies

Natural Language Processing (NLP) as a field has grown exponentially as researchers continue to harness various models and algorithms to characterize human behavior from textual data (Feuerriegel et al., 2025). These algorithms take in large amounts of data and output complex analyses beyond what a human can accomplish with variance testing or regression. These Machine Learning (ML) algorithms have been particularly powerful in psychological research, in analyzing both written and verbal data (Cero et al., 2024, Feuerriegel et al., 2025). Multiple different algorithms are available for researchers to use; one popular form of NLP is sentiment analysis, which identifies whether a text is more positively or negatively charged. Various methods of sentiment analysis have been used in prior work and are reviewed. In addition to sentiment, ML models can be used to identify patterns in large data sets and classify data along specified dimensions. These models are referred to as 'classification algorithms' and have been used previously to detect or discern internal cognitive states based on eye movement patterns (Castillon et al., 2024, Chartier et al., 2025; Kuvar et al., 2023a, Hutt et al., 2023). Several of these algorithms, and their incorporation of the sentiment analysis results, will be discussed in turn.

Lexicon-Based Sentiment Analysis. Lexicon-based sentiment analyses are popular among researchers due to their ease of interpretation, implementation, and replicability (Cero et al., 2024). In this approach, large dictionaries are used to quantify the sentiment contained in

each part of the text. A variety of these dictionaries exist, but return different output values, and have different strengths respectively. For example, a popular lexicon, VADER (Valence Aware Dictionary and sEntiment Reasoner), is specifically trained from social media texts and returns a single value between -1 (extremely negative) and +1 (extremely positive; Hutto & Gilbert, 2014). Another lexicon, such as NRC (developed by and named after the National Research Council Canada), returns two values, one indicating positive association, and one indicating negative, while also calculating scores for other emotions (i.e. anger, disgust, joy, surprise; Mohammed et al., 2010)

The choice of dictionary depends not only on the type of text to analyze, but also the desired result. Some lexicons are better suited to handle negation (i.e. 'not happy' would be identified as a negative), and others may give more detailed outputs (or different scaling). The present work will employ several of these lexicons (Vader, NRC and TextBlob) to validate each other's results. Based on prior research that used retrospective survey reports with NLP (e.g., Steadman et al., 2025; Venkatesha et al., 2025), one specific a priori hypothesis that I will investigate with these lexicon-based methods is the hypothesis that scenes during which déjà vu was experienced will be associated with more positively-valanced language than scenes during which déjà vu was not experienced.

Machine Learning (ML) Based Sentiment Analysis. Although I will not be using ML based sentiment analysis (MLBSA), it is important to note it as another option for sentiment analysis. In MLBSA, a model has been trained on prior data and is able to discern sentiment of new data on the basis of that training; however, it is unclear what exactly in that prior training data is leading to its classification. Further analyses are needed to detect what could be leading to the model's sentiment decisions. Whereas lexicon-based approaches result in simple quantitative

values per unit of text, depending on the dictionary, such as a valence score (positive or negative) on a scale of -1 to 1, or a Likert-scale, MLBSA approaches often try to classify parts of texts as positive or negative dichotomously with some kind of confidence measure. These ML models do not examine *how* negative or positive a given text is, only its *predicted* overall valence.

Generally, MLBSA can be more difficult to interpret in terms of what linguistic aspects of a series of text are contributing to the sentiment score. For example, utilizing the ML-based sentiment analysis model, LLaMA (Large Language Model Meta AI), to categorize sentiment, Venkatesha et al. (2025) analyzed retrospective reports of déjà vu and other involuntary thought types such as IAMs. The model concluded that for déjà vu, the word 'walking' was positively valanced, while for IAMs, the same word was negatively valanced. This example highlights the key differences between lexicon and ML approaches to sentiment analysis. In Lexicon approaches, walking is seen as an independent word, and all sentiment scores for words in a text are pooled together for an average score. In contrast, in the ML-based approach, it is unclear what in particular is causing 'walking' to be interpreted as more positive or negative in a given context. The model simply tells us that within the context of déjà vu retrospective reports, the word 'walking' was more positively weighted.

Feature-Based ML Classification Algorithms. Classification models are powerful tools that allow researchers to identify patterns in massive data sets. Feature-based supervised models allow researchers to test specific hypotheses about the impact of particular features, such as pupil dilation or micro saccades measured from eye gaze recordings, on a predicted outcome, like a psychological construct or behavior, such as mind-wandering (e.g., Kuvar et al., 2023b, Hutt et al., 2023). In a supervised approach, data sets include a 'ground truth' (e.g., whether mind-wandering was or was not reported) that the classification model is trying to accurately predict

by recognizing patterns in large amounts of data. Alternatively, unsupervised methods take a clustering approach in trying to group unlabeled data points and identify commonalities. Unsupervised methods could be beneficial when conducting exploratory data analysis, or when there are few a priori hypotheses. In the present thesis, specific hypotheses have been generated from prior work to be analyzed in a supervised approach.

As a specific example of a supervised method, prior work by Castillon and colleagues (2024) used a static version of Okada et al's (2024) VR paradigm in a head mounted VR system, equipped with eye-tracking, to detect feelings of familiarity from eye gaze patterns. They used classification models to identify reports of familiarity from calculated eye gaze features. Castillon et al. used a guided model search to systematically iterate through multiple classification models, optimizing a specific outcome statistic with each iteration. Their model search was set to include six different classification models: Naïve Bayes, K Nearest Neighbor, AdaBoost, Random Forest, SVC, and Logistic Regression. Their resulting, best performing model could reliability discern at above-chance levels (Cohen's $\kappa = .18$) whether a person had found a given VR scene familiar or not (indicated by a VR controller button-press while immersed in the scene) based on their eye-gaze patterns that were recorded within the VR system. A follow-up study held the button-press constant to ensure that the ML classification model was not picking up on eye gaze differences associated with the button-press; in this study, when examining only cases in the which the button was pressed (indicating a subjective sense of familiarity on the part of the participant), the ML classification algorithm was able to differentiate between scenes that had the same spatial layout as a studied scene and scenes that did not, based solely on the recorded eye gaze patterns (Chartier et al., 2024). Both studies utilized feature-based ML approaches to classify instances of familiarity, specifically using a

guided model search to best identify the highest performing ML model. The same methodology of a guided model search will be used in the present study, incorporating linguistic features in place of eye gaze features.

One difficulty of feature-based classification algorithms concerns the ease of interpreting outcomes. When using simple classification models, like decision trees, it is clear what features are most significant (i.e., what features were most frequently used to make classification decisions). However, other more complex classification models are less clear about how they reach their predictions. One method of dealing with this is the SHAP method. Using SHAP (SHapley Additive exPlanations; Lundburg & Lee, 2017), features are analyzed for how their various values contribute to a model's class prediction. SHAP is model agnostic and can be used and compared across a variety of ML classification models. Previously, Hutt and colleagues utilized SHAP to analyze eye tracking features to detect mind wandering. (Hutt et al., 2023). This approach greatly increases the interpretability of the classification models and indicates some directionality of effect.

Turning to the use of language that was spoken while immersed in the VR scenes rather than eye gaze, a similar ML based classification approach, with SHAP analysis, can be taken to investigate how various linguistic features (e.g., sentiment), contribute to ML's ability to separate déjà vu states from non-déjà vu states on the basis of spoken language alone (that does not include "give away" words like the phrase "déjà vu" itself). This approach can include testing a priori hypotheses (such as the hypothesis that déjà vu reports should be more positively valanced than non-déjà vu reports) as well as taking an exploratory approach to identifying new types of features that the ML algorithm may be using to make the classifications but that were

not hypothesized beforehand, and attempting to understand what their use by the ML may indicate about déjà vu phenomenology.

NLP Studies of Involuntary Thought

Poulos et al. (2023) used NLP to study how three types of involuntary thought: IAM, unexpected thought (UT), and ruminative though, differ from one another. Participants were asked to describe instances of each involuntary thought and appraise them on a set of dimensions including spontaneity, morality, emotion and vividness. ML classification algorithms were employed to classify each type of involuntary thought based on participant's appraisals (Poulos et al., 2023). They found that while the involuntary thought types are not completely separable, they could be relatively accurately classified based on appraisal dimensions.

More recently, following both from Poulos et al.'s (2023) NLP investigation of different types of involuntary thoughts and from Barzykowski and Moulin's (2023) proposal that déjà vu and IAM may exist along a continuum, Steadman et al. (2025) used NLP to examine how déjà vu might be distinct phenomenologically from IAM or UT. Steadman et al. asked participants to write about an IAM that happened to them, a UT and a déjà vu experience. They used lexicon-based sentiment analysis, employing both TextBlob and Vader. From this analysis, Steadman et al. (2025) found that déjà vu was characterized by positive valence and was more positive than either IAM or UTs. IAMs were also scored as slightly positively valanced, while UTs were scored as slightly negative or neutral (depending on the lexicon). These results reflect literature on familiarity which states that familiarity is associated with a positive 'warm glow' (Monin, 2003). This work demonstrated the usage of lexicon-based sentiment analysis for involuntary thoughts when participants are retrospectively reporting about their experiences and

demonstrates the use of multiple lexicon approaches (TextBlob and Vader) for seeking convergence and arriving at their conclusions.

Venkatesha et al. (2025) followed up on this study using ML-based sentiment analysis and an exploratory approach of Term-Frequency-Inverse Document Frequency (TF-IDF), which compares word frequency within a classification group to word frequency across the entire data set. They found results that largely aligned with those of Steadman et al. (2025) in that déjà vu tended to be experienced as more positive, or at least as less negative, than either IAM or UTs. Additionally, TF-IDF revealed that classification by the algorithm of the text description as being an instance of retrospective déjà vu (as opposed to IAM or UT) were more positive, and this positivity was driven by words related to spatial, place-oriented, or navigational descriptors, in line with prior survey data that has shown that déjà vu is most commonly elicited by scenes or places (Cleary et al., 2020, Cleary & Brown, 2022). The TF-IDF approach is beneficial is illuminating the black box of ML based sentiment analysis.

While the studies by Steadman et al. (2025) and Venkatesha et al. (2025) demonstrate the potential of NLP approaches for investigating the phenomenological qualities of various types of involuntary thought, including déjà vu, a drawback to these studies is that they were retrospective in nature. That is, they relied on people's recollections of a past time that they had experienced each of these involuntary thought types and, as Aitken and O'Connor (2020) point out, relying on retrospective reports for studying déjà vu carries with it the potential for memory errors and biases, or even memory distortions, pertaining to what is remembered.

To better understand the phenomenology of déjà vu using NLP, it would be ideal to capture a person's language surrounding a déjà vu experience as it happens. The present study sought to do so by analyzing verbalizations that participants made during a Think Aloud protocol

while they were touring scenes in the Virtual Tour paradigm within Virtual Reality (VR). Specifically, participants engaged in a variant of the VR-based Virtual Tour paradigm reported in Experiment 3 of Okada et al. (2024). In this paradigm, participants tour a number of different VR scenes by being pulled on rails through each scene (which is analogous to being pulled along on a Disney World ride like "It's a Small World."). Participants tour 16 unique scenes in a study phase and then tour 32 unique scenes in a test phase, completing two separate study-test phases altogether for a total of 32 study phase scenes and 64 test phase scenes. In the test phase, all of the scenes are novel, but half of them share the same spatial layout as a scene from the study phase and half do not. As reported by Okada et al. (2024, Experiment 3), participants are more likely to report déjà vu among test scenes that share a spatial layout with an unrecalled studied scene than among test scenes that do not. Moreover, when people report experiencing déjà vu, they have a heightened likelihood of experiencing a concurrent feeling of knowing what the direction of the next turn on the tour should be. They also exhibit a greater likelihood of making a commission error (an incorrect attempt at recalling a potentially similar scene from the study list that could explain the déjà vu, as opposed to not attempting recall).

My lab recently ran a variant of Okada et al.'s (2024) VR Experiment 3 in which a Think Aloud protocol was incorporated into the procedure. Thirty-two participants were instructed to do all of their thinking out loud throughout the experiment, while their verbalizations were recorded using a microphone. These voice recordings were later manually transcribed by multiple people and edited by an additional transcriber.

The purpose of the present thesis was to apply NLP algorithms to the transcribed Think Aloud dataset in order to test hypotheses about the characteristics and phenomenology of déjà vu when the language surrounding the déjà vu experience has been captured in real-time.

Importantly, instead of comparing language regarding déjà vu to other involuntary thought types, which was done in past retrospective research (Steadman et al., 2025; Venkatesha et al., 2025), the present thesis used NLP to compare language occurring during déjà vu reports with language occurring during non-déjà vu reports as people tour scenes in VR. Additionally, the present study took an exploratory approach (e.g., Yanai & Lercher, 2025) in order to identify features of language that are used to classify instances of déjà vu but that may not yet have been previously theorized about or investigated in studies of déjà vu, thus potentially yielding new insights into possible déjà vu phenomenology.

The Present Study

The present study sought to more richly characterize the déjà vu experience by using NLP to explore linguistic properties associated with experiencing déjà vu versus not experiencing déjà vu, as well as déjà vu's emotionality, and other subjective qualities. Past research that has performed sentiment analysis on retrospective memories of déjà vu obtained using a survey approach has suggested that, compared to other forms of involuntary thought, déjà vu is generally experienced as more positive (e.g., Steadman et al., 2025; Venkatesha et al., 2025). This research has also suggested that déjà vu is characterized by words denoting places (e.g., spaces or scenes), which is consistent with past survey research suggesting that the most common elicitor of déjà vu is places (Cleary & Brown, 2022). However, a drawback to this past work is that it relied on retrospective reports of past déjà vu experiences, and no study to date has examined linguistic patterns taking place the moment that déjà vu is experienced. This was the aim of the present study—to capture linguistic patterns occurring in the moments surrounding the déjà vu experience as it happens. To achieve this, a Think Aloud protocol was incorporated

into the Virtual Reality (VR) based virtual tour paradigm developed and used by Okada et al. (2024).

The present analysis approach aimed to apply NLP to an existing data set that was collected using an adapted version of Okada et al's (2024) scenes, along with a Think Aloud protocol. During the experiment, participants toured through various scenes and were probed about déjà vu. During the tours, participants were instructed to verbalize all their thoughts aloud. Subsequent data analysis revealed the valence and other important features of the verbal reports. Harnessing both lexicon-based sentiment analysis and feature-based classification models, I evaluated the most important features that define the phenomenology of déjà vu.

I hypothesized, in line with prior work (Steadmen et al., 2025, Venkatesha et al., 2025), that déjà vu would be more positively valanced than non-déjà vu states. Further, I combined these sentiment features with other coded features to test a set of theoretically grounded a priori hypotheses about the collected data set. A team of trained research assistants manually double coded these features blindly (See Appendix A for a full list of features). I inputted these features from sentiment analysis and manual coding into ML classification models to best predict participants' subjective reports of déjà vu. My second hypothesis was that based on these features, classification models would be able to detect the difference between déjà vu and non déjà vu states, during recall failure. Finally, I also hypothesized that déjà vu states would be distinguished from non-déjà vu states by more positive sentiment, filler words and instances of describing spatial features of the scene during encoding. The results from this study are integral in replicating prior work on involuntary thoughts, and on further defining an understudied phenomenon.

CHAPTER 2 - METHOD

Participants

Data (n = 32) were collected from undergraduate students, enrolled in psychology courses, at Colorado State University in the spring semester of 2024. Participants were compensated with course credit. Based on the power analysis conducted by Okada et al. (2024), the necessary sample size is about 20 participants to detect the same effects that they did pertaining to déjà vu reports in VR. Thus, we aimed for a similar sample size, although we aspire to examine different effects that have no exact precedent, as the primary measure in the present study is transcribed spoken language data, which has never before been examined in the context of the VR virtual tour paradigm. Thus, we aimed for a sample size that would allow us to replicate the effects found by Okada et al. (2024) using their protocol while additionally collecting Think Aloud data within that protocol for the first time. During the experiment, two participants were sent home for motion sickness. Eight participants were excluded for not talking during more than 50% of scenes, and three participants were excluded for software errors. Thus, there were 19 valid participants for data analysis.

Materials

The experiment was created within the Unity Game Engine and displayed using the HTC VIVE Pro head-mounted display. The scenes updated from Okada et al. (2024), were utilized, which included 64 pairs of spatially similar scenes, totaling 128 scenes. These pairs were divided into two different blocks and scenes within each block were displayed randomly. As with the black and white scenes from Cleary et al. (2009), the layouts of the two scenes were spatially identical (See Figure 2 for examples). These 3-D scenes were displayed as virtual tours which pulled participants through the scene as if they were “on rails”, or a slow rollercoaster. In the

study session, the tour ended with a turn towards either the left or right corner of the scene. Four versions of the experiment were created to counterbalance the scene stimuli based on their critical turn and whether a given test scene corresponded to a spatially similar study scene.

Procedure

This study was conducted in person. After obtaining consent, participants were instructed that once set in VR, they should relay all thoughts aloud. Participants sat in a chair during the duration of the experiment with a handheld remote to respond to prompts.

Once set in VR, participants were instructed that they would see a series of video tours, and they were to watch each scene and try to remember it. Participants were then pulled through 16 tours of scenes. Each tour ended with a turn either left or right within the scene. During the scene, the participant heard an audio recording saying the name of the scene (e.g. "Alleyway, this is an Alleyway"). Following the study phase, participants were given the instructions for the test phase:

> "*You will view short virtual tours of scenes again, but this time, they will be new ones not seen in the previous phase. Some of the virtual tours will resemble scenes from the study phase and others will not. After you view a tour, you will be asked several questions about it.*
>
> -
>
> *First, you will be asked if the test scene prompted you to feel a sense of déjà vu (the feeling of having experienced something before without knowing why and despite knowing that the current situation is new). You will indicate Yes or No for this question.*
>
> -

Following this, you will be asked if you have a sense of knowing where to turn next. You will indicate Yes or No for this question. You may just have a feeling about this without knowing why. You will also be asked, regardless of your previous answer, to indicate (even if it's just a guess) whether the scene will proceed in a Left or Right turn.

-

Next, you will be asked if the test scene feels familiar to you. You will indicate Yes or No for this question. Please proceed to the next page.

-

Finally, you will be asked if the current scene reminded you of a particular earlier-presented scene from the first phase. If viewing the test scene triggered a memory of a similar-looking scene from earlier, please indicate the name of that earlier-viewed scene when prompted. Even if the test scene did not remind you of a specific earlier-viewed scene, go ahead and take a guess when you see this prompt. Please note that there will be no sound during the test phase."

Then, participants were brought through 32 scenes which stopped just short of the ending turn. Participants were then asked the questions described, following each test scene.

Following the completion of the first study/test block, participants were set up into the second identical block of the experiment, with different scenes. Prior to the second block, participants were instructed that the next block would have the same procedure with brand new scenes. They were instructed to only refer to scenes from this block moving forward and reminded to speak aloud all their thoughts during the study.

At the conclusion of the VR experiment, participants were given both a demographics questionnaire and the IDEA (Sno et al., 1994). Finally, participants were debriefed and explained the purpose of the study.

Data Preprocessing

The first step in processing the Think Aloud data was manually transcribing audio files from the VR sessions. This labor-intensive process is one of the biggest limitations of Think Aloud studies. Each audio file was reviewed twice to ensure accuracy and minimize human error. To fast track this, we implemented OpenAI's Whisper large-v3 model to replace the initial phase of audio transcription. This meant that only one human rater would be needed to edit the transcript for automation errors. Whisper is a transformer-based speech recognition model, which helps it do well even with accented or noisy speech. While prior studies have reported successfully using this software (Mildner & Tamir, 2024), the AI was widely unreliable for our data and was only implemented for two transcriptions. The unreliability could be due to audio quality issues as some sessions were recorded from an in-room microphone, and others were recorded from the VR headset. For the remainder of the recordings, multiple graduate research assistants were employed to transcribe audio files, then I reviewed the audio transcription files for accuracy.

Following transcription, data coding took place to quantify various features of the linguistic data to input into various ML classification models. Some features were automatically coded, including word count, sentiment, and number of filler words. On the other hand, manual coding was necessary for some features of interest, including elaboration or whether participants mentioned physical objects in the environment (for a full list, see Appendix A). Manual coding was performed by undergraduate research assistants, using Cohen's kappa to determine interrater

reliability. Discrepancies between raters were satisfied by a third independent rater. In instances where the third rater disagreed with both raters, I made the final decision.

Planned Analysis

The first set of analyses were lexicon-based sentiment analysis using three different dictionaries. The first two, Vader and TextBlob, have been used previously to characterize déjà vu experiences and allow for direct comparison to prior research (Steadman et al., 2025). The Vader lexicon results in four scores: the proportion of text with positive sentiment, negative sentiment, or neutrality, as well as an aggregated sentiment score from -1 (extremely negative) to + 1 (extremely positive; Hutto & Gilbert, 2014). TextBlob also returns a comparable aggregated sentiment score as well as a score of subjectivity (Loria, 2018). Finally, the NRC emotion lexicon was also used. This lexicon not only returns sentiment scores, but also scores for eight emotions: anger, anticipation, disgust, fear, joy, sadness, surprise, and trust (Mohammed et al., 2010).

When employing lexicons, a significant amount of data preprocessing was needed, beyond transcriptions. The first step was lowercasing all words, correcting misspellings, and expanding contractions. All filler words (umm, uhhh, hm) were standardized (um, uh, hm) so they may be accurately tallied. Next, common stop words, words that don't add meaning to a sentence (e.g. a, of, the) were removed. Once counted, I removed filler words. And finally, words were stemmed, to remove their endings (i.e. 'walking' becomes 'walk'). These steps ensure that the analysis looked solely at meaningful words in a sentence and not meaningless grammar constructs.

In this approach to sentiment analysis, pre-defined lexicons have scored tens of thousands of English words. Each of these words was compared against words in the study's data sets to

aggregate a score for each unit of text, representing its sentiment. The resulting scores from sentiment analysis were used as sentiment features in our classification algorithms. Based on the resulting scores, each trial was assigned a sentiment category, either positive, negative, or neutral, based on the agreement of at least two lexicons. In the case where all three lexicons predicted a different category, the compound scores were averaged and then the trial was categorized accordingly.

Then, I created the final data sets, combining sentiment features, manually coded features, and automatically coded features (See Appendix A for a full list). I trained and tested the data on a multitude of ML classifiers. Hyperopt, a guided model search, can be used to cycle through different ML models, testing a set space of hyperparameters for each model (Bergstra et al., 2013). Naïve Bayes, K Nearest Neighbors, AdaBoost, Random Forest, SVC and Logistic Regression were used in line with prior work using Hyperopt (Castillon et al., 2024, Chartier et al., 2025). While computationally extensive, this approach removes researcher bias when choosing a model for analysis and automatically compares accuracy across models. This model search sought to classify déjà vu reports from non déjà vu reports in the instances of recall failure over 300 iterations, optimizing the average Cohen's kappa.

Three different model searches were planned a priori to investigate which features lead the model to predict déjà vu report. First, a Retrieval model was trained over all features from the test phase (i.e. verbal reports during the test phase). Next, a Retrieval Failure model was trained over all features from the test phase during instances of recall failure. (See Table 1 for each model and its feature sets). Finally, an Encoding model was trained over all features from the encoding phase (i.e. verbal reports during the study phase) to classify later déjà vu from non-déjà vu states.

Table 1

Trial Training Set and Predicted Label of Guided Model Searches

	Trials	Included Trials	Predicted Value
Retrieval Model	Test	All	Déjà vu Report
Retrieval Failure Model	Test	Recall Failure	Déjà vu Report
Encoding Model	Study	All	Déjà vu Report

Each model was trained using data from all but one participant (the Leave-One-Participant-Out-Cross-Validation procedure). Then, the model was tested on that left out participant. For each left out participant, Cohen's kappa was calculated, ranging from -1 (completely imperfect) to +1 (compSletely perfect), with zero representing chance accuracy. For each iteration of each model, each participant was treated as the 'left out' participant and kappa was calculated, then an average kappa was calculated across participants. The distribution of Cohen's kappa was analyzed for outliers. This Leave-One-Participant-Out-Cross-Validation procedure addresses concerns of generalizability, ensuring that I am accurately measuring our psychological concept, not noise, and that the model did not overfit to its training data.

From our guided model search, I am able to identify which models do the best at classifying verbal reports as déjà vu versus not. Another important analysis examined which features are driving the model toward this accuracy. Using SHAP (Lundberg & Lee, 2017), I can identify which features are primarily influencing the models' predictions.

A key limitation to hyperparameter optimization is overfitting, thus SHAP was used on both the best and median performing models, seeking consistent evidence of feature importance.

SHAP breaks down how different feature values influence ML outcomes in a model agnostic approach, allowing for comparison across different ML classification models.

From SHAP, more simple statistical analyses, like regression and variance testing, were used in conjunction with it to understand which features are indicative of déjà vu and how that particular features increase or decreases the probability of reporting déjà vu. Through these methods, I am better be able to capture the phenomenological experience of déjà vu.

CHAPTER 3 - RESULTS

Replication of Okada et al. (2024)

First, I replicated the main trends from prior work as a manipulation check. This ensures the added Think Aloud protocol did not affect the established effects of the paradigm. During recall failure, the participants is truly faced with novelty (due to the lack of identification) and with familiarity (due to the spatially similar stimuli) leading to the feeling of déjà vu. I performed a series of two-sided paired samples t-tests, on trials in which the participants were unable to recall a previously viewed scene. When participants had previously viewed a similar scene, they were significantly more likely to report déjà vu (M = 0.54, SD = 0.23), than when in a completely novel layout (M = 0.40, SD = 0.17), $t(18)$ = -4.67, SE = 0.03, p <0.001, d = 0.68). I also found, in line with Okada et al. (2024), that participants' commission error rates were significantly higher for scenes with reported déjà vu (M = 0.56, SD = 0.23) than no report (M = 0.16, SD = 0.22), $t(18)$ = -6.74, SE = 0.06, $p < 0.001$, d = 1.77.

Sentiment Analysis

Next, I conducted lexicon-based sentiment analysis on all trials with three different lexicons-VADER, TextBlob and NRC. The resulting scores were used to categorize trials for inclusion in later classification models, as described in the methods. To analyze the agreement amongst lexicons, I first examined the correlation amongst all trials (Table 2) for each lexicon. The lexicons showed moderate, and consistently positive agreement. Each lexicon's polar predictions (highly negative and highly positive) were manually analyzed to reveal no major domain issues. The variation between lexicons occurs across intensity of valence, not polarity.

Table 2

Correlations among Lexicons used for Sentiment Analysis across all Retrieval trials.

	Correlation
VADER - Textblob	0.30
Textblob - NRC	0.44
NRC - Vader	0.53

Variation across lexicons is less when analyzing how the lexicons categorized trials. For all retrieval trials, at least two lexicons agreed on 96.0% of trials. As planned, the fifty trials where all three lexicons disagreed, their composite scores were averaged, then the trial was categorized.

To test my first hypothesis, that déjà vu reports would be more positively classified that non-déjà vu reports, I calculated the proportion of trials that were categorized as negative, neutral and positive for positive and negative déjà vu reports, respectively. As hypothesized, when participants reported déjà vu, their language was significantly more positive (M = .64, SD = .24) compared to when not reporting déjà vu (M = .52, SD = .24), $t(19) = 2.473$, p = .02, d = 0.28.

Retrieval Models

The first guided model search was done across the retrieval features. Each participant had an average of 63.3 trials included (Range = 62-64). Some software errors caused a single repeated scene in the test phase for some participants; only first instance of the scene was included for those trials. A total of 1202 trials were included for the model search. The best performing model had an average kappa of .20, and the median model had an average kappa of .15 (See Table 3). The histogram of kappas for each is visualized in Figure 4. The histogram

indicates that the average kappa is being skewed by two outliers. Also, six participants, almost a third of the sample, have a very low kappa ($\kappa < .1$).

Table 3

Retrieval Model Search Results

	Model	Average κ	Average Accuracy	Average *F1*
Best	AdaBoost	.20	.62	.62
Median	AdaBoost	.15	.61	.61

Figure 4

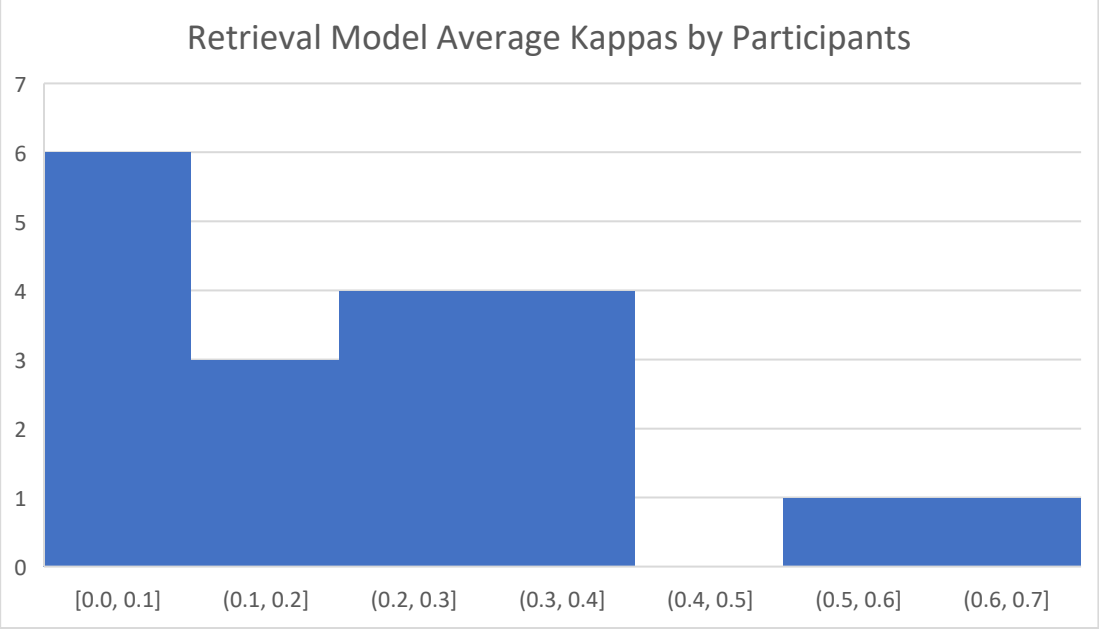

Figure 3: Histogram of individual participant Cohen's Kappas from guided Retrieval model search.

SHAP Analysis

Next, I performed SHAP analysis on both the median and best performing models to see how features are influencing the models' decisions throughout the model search (Figure 5).

Figure 5

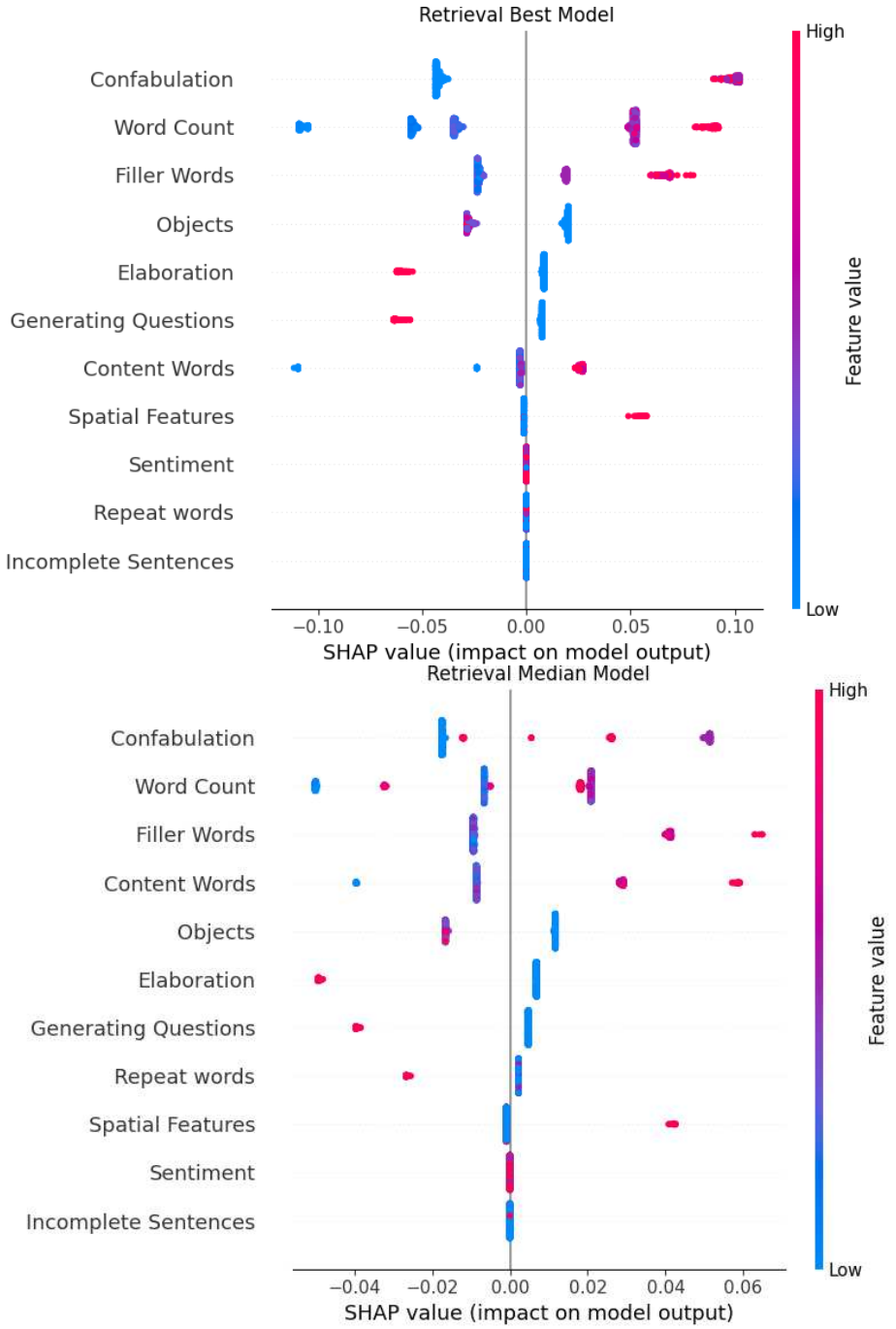

Figure 4: SHAP analysis on two Retrieval models from guided model search, best and median performing models. For specific hyperparameters of all models, see Appendix A.

The best and median models largely agreed on which features were important. The most influential features were confabulation, total word count, number of filler words, and mentioning objects in the environment. The visualization of this analysis revealed that mentioning more objects in the environment was indicative of no déjà vu report. The other top three features all positively influenced the model to predict a déjà vu report.

Retrieval Failure Models

The next planned search I performed was across all retrieval trials in which the participant was unable to recall the previously studied scene. On average, 43.5 trials were included per participant (Range = 37-57), for a total of 827 trials (See Table 4). The best performing model was an AdaBoost model with an average kappa of .39 and the median model had an average kappa of .33 (See Figure 5 for histogram of participant level kappas).

Table 4

Retrieval Failure Model Search Results

	Model	Average κ	Average Accuracy	Average *F1*
Best	AdaBoost	.39	.74	.74
Median	AdaBoost	.33	.71	.71

Figure 6

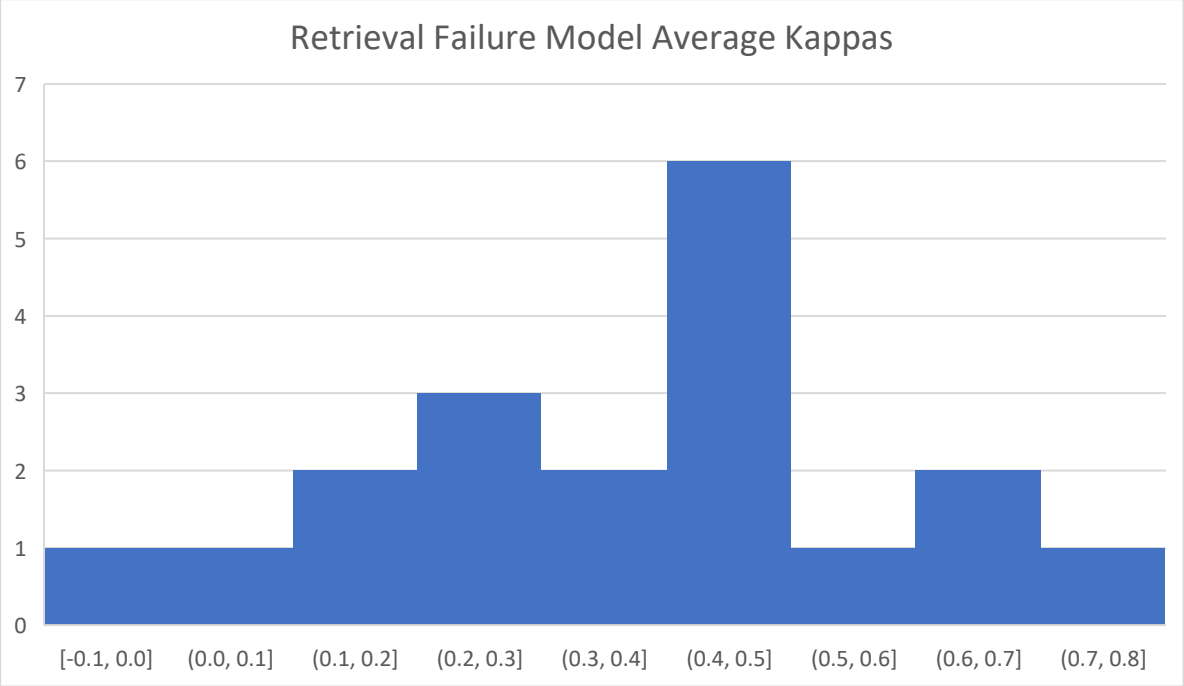

Figure 5: Histogram of participant Cohen's kappa from guided Retrieval Failure model search's best performing model.

SHAP Analysis

Once again, I performed SHAP analysis to understand which features were contributing to the models' predictions (See Figure 7). Similar to the retrieval model, the most important features during retrieval failure were total word count, confabulation, objects, and elaboration. The same trends emerge as objects, along with elaboration, seem to contribute towards non-déjà vu reports. Once again, increased total word count and instances of confabulation led the model to predict a positive instance of déjà vu report.

Figure 7

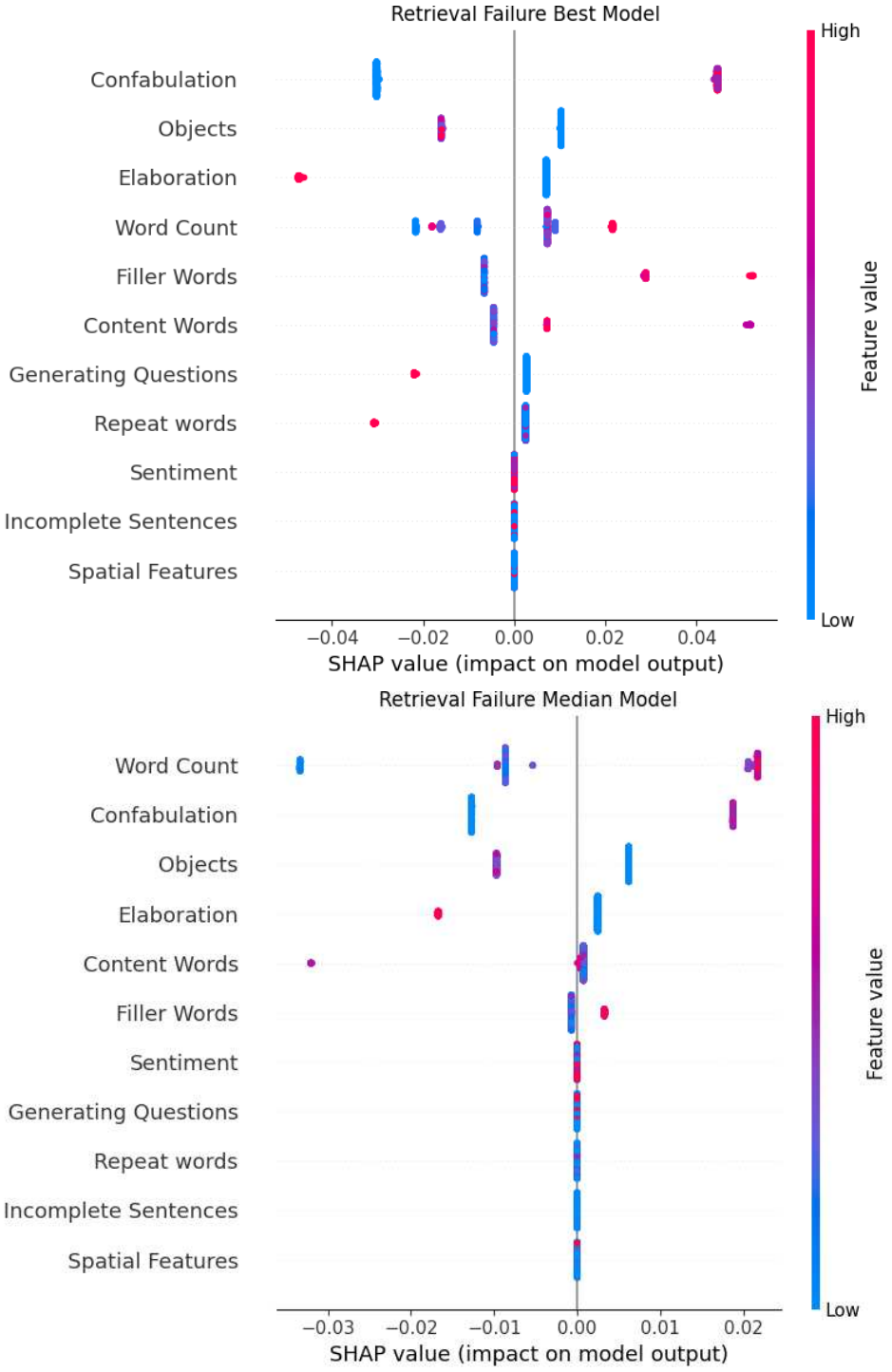

Figure 6: SHAP feature analysis results for best and median performing Retrieval Failure models. See Appendix B for specific hyperparameters of trained models.

Encoding Models

The last guided model search was set across the encoding features. Each participant had an average of 31.9 trials (Range = 31-32) included for analysis across the 2 blocks, resulting in 638 trials for analysis. A software error causes a repeat scene during the test trial which leads to one studied scene's pair not being shown (For a list of features included in each model, see Appendix A). After 300 iterations, the encoding models performed poorly; the best model had an average kappa of .06, and the median model had an average kappa of -.002. (Further scores reported in Table 3; For a full list of hyperparameters for all models, see Appendix B).

Table 3

Encoding Model Search Results

	Model	Average κ	Average Accuracy	Average *F1*
Best	SVC	.05	.47	.50
Median	Random Forest	-.06	.56	.59

Variance Testing of Important Features for Retrieval Models

Following up on the retrieval models' predictions, I performed variance testing on the most prominent features to seek converging evidence of feature importance. I ran a series of two-sided paired samples t-tests for each of the top five features (word count, confabulation, objects, elaboration, and filler words; full results in Table 6) across all retrieval trials, and retrieval failure, separately. For both retrieval trials and retrieval failure trials, there was a significant difference for total word count, number of filler words, and instances of confabulation, or attempting to explain away familiarity incorrectly.

Table 6

Paired Samples T-test for most Important Features across Déjà vu vs. Non-déjà vu Reports

Model	Feature	Déjà vu	Non-Déjà vu	*t*	*df*	*p*	*d*
Retrieval	Confabulation	0.53	0.27	3.24	19	.005	0.74
	Word Count	33.62	26.86	5.51	19	<.001	1.27
	Filler Words	2.49	1.90	3.13	19	.006	0.72
	Objects	0.66	0.69	0.33	19	.743	0.08
	Elaboration	0.10	0.16	1.73	19	.101	0.40
Retrieval Failure	Confabulation	0.85	0.30	5.26	19	<.001	1.21
	Word Count	33.85	26.12	5.80	19	<.001	1.33
	Filler Words	2.40	1.87	2.52	19	.021	0.58
	Objects	0.56	0.68	1.31	19	.206	0.30
	Elaboration	0.09	0.17	1.88	19	.077	0.43

Exploratory Feature Analysis

After identifying confabulation, filler words, total word count, elaboration, and objects as the best performing features, I fit a multitude of models with the best performing hyperparameters from the planned analyses. The models were fit across retrieval and retrieval failure trials separately, with their respective best performing hyperparameters and are reported separately.

Simplified Models

First, I fit models for both trial sets with only the top five features. This approach aims to identify how much of the full models' predictions are being driven solely by the top five factors.

A significantly lower simplified model Kappas would indicate that the lower importance features are still having an impact on the models' predictions. For both models, the top five features performed almost as well as the full feature models, with average kappa of .19 for the retrieval model and .37 for the retrieval failure model (Table 7).

Table 7

Average Kappas of Simplified versus Full Feature Models

Best Model	Top Five Feature	Full Feature
Retrieval	.19	.20
Retrieval Failure	.37	.39

Leave One Feature Out

Next, I used the same approach as Kuvar et al (2023a) to analyze how the model results vary across the inclusion and exclusion of each important feature. I fit the best performing model for both retrieval and retrieval failure trials across all five features individually and with each one excluded (i.e. one model with only filler words and another with the top features other than filler words). This approach aims to show which features are predictive of déjà vu on their own, and how differently the model performs when the feature is excluded. High feature importance would be indicated by a higher kappa when training on the feature along, and a lower kappa when training with the feature left out. The results of all 10 models can be found in Table 8.

Table 8

Average Kappas of Training AdaBoost Models with and without Important Features

	Retrieval Model		Retrieval Failure Model	
Feature	Train with feature	Train with all except feature	Train with feature	Train with all except feature
Confabulation	.16	.12	.37	.35
Filler Words	.12	.18	.07	.35
Word Count	.13	.23	.14	.35
Elaboration	.01	.19	.00	.33
Objects	.00	.19	-.01	.38
All Features	.19	-	.37	-

The models' results indicate that confabulation, word count and filler words were the best performing features on their own. Both models struggled to predict deja vu when the models were trained on only elaboration or number of objects. This was consistent across both the Retrieval and Retrieval Failure models. When elaboration and objects were left out of the Retrieval model, the models still performed well; however, for the retrieval failure model, the model suffered when elaboration was left out (average kappa of .23 compared to the full top feature model's average kappa of .37), and the model improved slightly when objects was left out (average $\kappa = .38$).

As mentioned, when the models were trained on word count alone, the model did fairly well. In the same way, for the Retrieval Failure model, when word count was excluded, the model performed slightly worse (average $\kappa = .35$) compared to the full top feature model (average $\kappa = .37$). But, when word count was excluded in the Retrieval model, the model performed even better (average $\kappa = .23$) than the original full feature model (average $\kappa = .20$).

Then, to visualize the generalizability of this model, I plotted the histogram of participant kappa's for that model to visualize how it differs from the full model (See Figure 8). While there are still two outliers on the right side, the model's predictive ability did become more consistent.

Figure 8

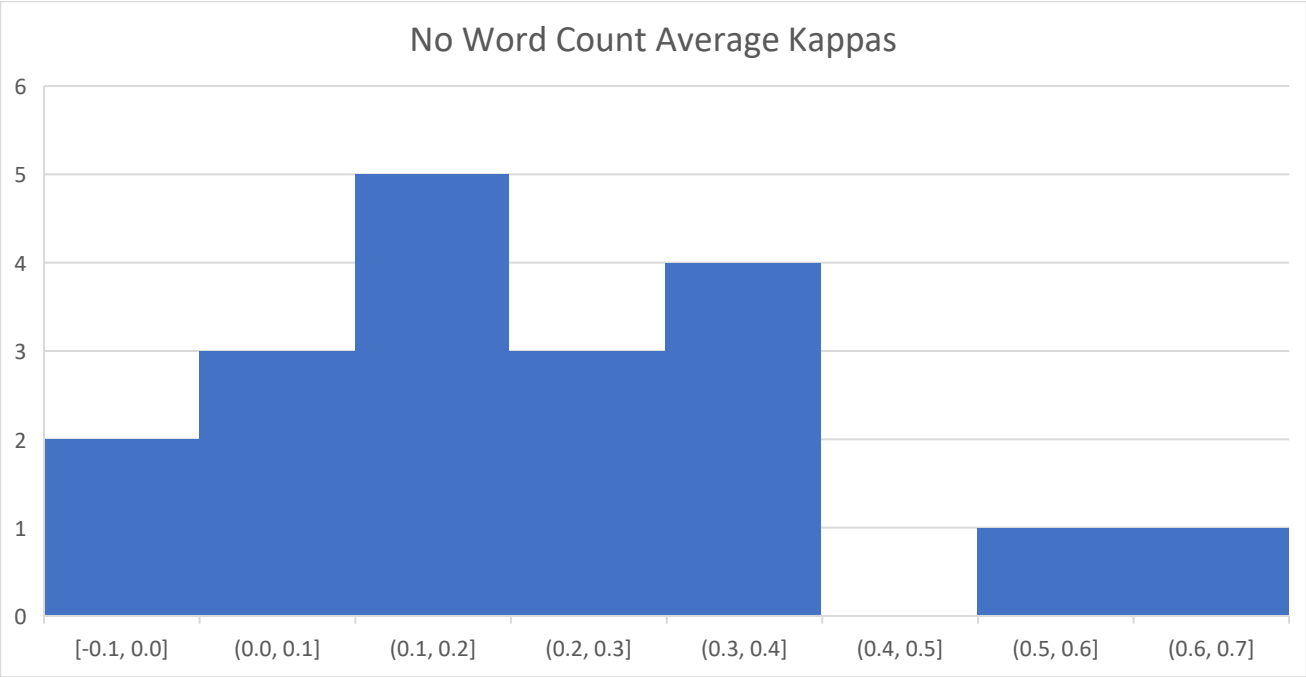

Figure 7: Histogram of individual Kappas for best performing retrieval model from guided model search. Model was trained on top features, excluding word count (i.e. confabulation, filler words, elaboration, and objects)

CHAPTER 4 - DISCUSSION

The present study took a novel approach to examining déjà vu phenomenology. It aimed to describe the verbal reports of people while they are actively experiencing the feeling of déjà vu. This study was the first to implement a think aloud protocol into a paradigm aimed at eliciting déjà vu, presenting a novel opportunity to study the real time reports of participants. I set out to test how different features of verbal reports were indicative of déjà vu states. I employed machine learning classification models and various feature analysis techniques to characterize participants verbal reports in déjà vu versus non déjà vu states.

My first hypothesis was that, in line with prior work with retrospective reports, deja vu states would be more positively valanced than non déjà vu states. This pattern was found, and was statistically significant, however my follow-up analyses revealed that sentiment was not an important predictor of déjà vu states in any model, and when more simplified models were run without the sentiment feature, the models' performance barely decreased, indicating the feature didn't contribute significantly to the models' predictions. One could also argue that the difference in positivity between déjà vu and non-déjà vu reports coming from sentiment analysis could be because of perceived task relevance of déjà vu and demand characteristics. If a participant perceives that a positive déjà vu report is a good thing, in terms of the experiment, they may speak more positively when experiencing déjà vu. That said, if the present sentiment analysis results were merely the result of such demand characteristics, it is difficult to explain why Steadman et al. (2025) also found déjà vu to be relatively positive in their comparing of retrospective reports of past déjà vu vs. involuntary autobiographical memory vs. unexpected thought experiences. In that study, there would be no reason why the association between déjà vu and relative positivity would be due to demand characteristics, as people were simply reflecting

on their memories of different involuntary thought types. Thus, the current study's alignment with this prior work does suggest converging evidence for the idea that relative positivity is a characteristic of the déjà vu experience.

My second hypothesis was that I would be able to consistently distinguish between déjà vu and non déjà vu states during retrieval failure, which was supported. The resulting Cohen's kappas, for retrieval models, were comparable to prior work on involuntary thought (Chartier et al., 2025, Castillon et al., 2025). However, the features extracted from the study trials, in the encoding model, were unable to distinguish between the two states. No prior work has examined whether characteristics of cognitive processing at encoding would impact later déjà vu states. For example, some participants attempt to recite the names of scenes as they progress through the study scenes while others state what it reminds them of from day-to-day life. Although the present study did not find any significant encoding factors to be predictive of later déjà vu reports, examining processes at encoding remains an opportunity for further exploration in future research, such as whether study strategies at encoding make déjà vu more or less likely. Future experimental manipulations could seek to examine the impact of specific encoding strategies.

My final hypothesis was that déjà vu states would be characterized by increased filler words, more positive sentiment, and referring the spatial features at encoding. This hypothesis was partially supported in that filler words were one of the top two features in distinguishing déjà vu from non déjà vu states. However, the other two hypothesized important features were not supported. Although déjà vu states were more positively valanced than non-deja vu states, the models did not indicate this as an important feature in their decisions. Also, spatial features were not important to any of the retrieval models, and all encoding features failed to predict déjà vu from non déjà vu states. Participants did sometimes refer to the layout of the scene, but it is not

clear that that led to them reporting déjà vu. If the participant is not aware that spatial familiarity can lead to an increase probability of déjà vu, as most lay people do not, then their response to the spatial manipulation may be largely unconscious and thus not spoken about. This is a key limitation of the Think Aloud protocol—it only captures what participants are consciously aware of and able to articulate out loud. Researchers should not assume that spoken words capture all thought, and certainly not any unconscious thoughts and perceptions. While Think Aloud can capture some aspects of cognition, that probe-based methods cannot, it should not be assumed to provide the full picture of one's thought processes.

Word count was also a consistently important feature in the models' predictions; however, it is fairly correlated with filler words (r = .58). When the best retrieval model was trained on all top features except word count (i.e. confabulation, filler words, elaboration, and objects), the model performed better than the full feature model (average κ = .23), but the same trend was not observed for the retrieval failure model. A further look at the histogram of participants level kappa's reveals that the model performs consistently better without the total word count feature. This suggests that during instances of recall success, the word count feature decreases the accuracy of the model. With the high correlation between the filler words and word count features, the importance of the filler words feature also captures the importance of the total word count, but without a possible distraction from some high word count instances. Further, increased word count could also be a result of increased confabulations. If so, it would make sense that the variability in word count is represented within the other features, leading to better model performance with a simplified model.

There is an established link between increased cognitive effort and increased filler word usage (Christenfeld, 1994). From the attentional perspective, filler words could indicate a switch

towards internal attention and particularly towards memory search. I theorize that participants are using increased filler words to fill pauses in speech, as memory is searched, as a result of unresolved familiarity. Further implications for the relationship between internal attention and déjà vu will be discussed later.

My models also all pointed to confabulation being an important feature of déjà vu states, which was not a priori hypothesized about specifically. Confabulation was operationally defined as attempting to explain away familiarity by mentioning any episodic memory (within or outside the experiment, other than the correct scene pair). According to the SHAP analysis, a high number of confabulatory remarks was highly indicative of reporting déjà vu. Within the experiment, some participants would report both autobiographical memories and scene memories in the same trial, while others responded to the recall prompt with autobiographical memories. All these instances were counted as confabulation, hypothesized to be potentially explaining away unresolved familiarity. Moulin (2018) noted that familiarity-driven confabulation often occurs in cases of seizure-induced déjà vu, whereby the patient will attempt to explain away the spurious familiarity sensation in a confabulatory way, such as by claiming that a new TV show seems familiar because it is a re-run of something already viewed (even when that is impossible). Moulin (2018) suggested that perhaps familiarity-driven confabulation is a feature of everyday memory even in non-clinical situations—the sensation of familiarity in everyday situations may drive people to try to explain away the familiarity. To examine this idea, Huebert et al. (2022) increased experimental familiarity among test cues and found that, during instances of cued recall failure, reports of illusory recollective details increased with increasing cue familiarity, possibly reflecting the need to explain way familiarity through an assumption of access to recollective detail even when no such recollective detail is available.

Surprisingly, mentioning specific objects within the environment was another feature at the top of each SHAP analysis—specifically in predicting non-déjà vu reports. When a trial had a high number of objects mentioned, the models predicted there would be no déjà vu report. One interpretation of this is that this represents instances of external attention, while confabulation and filler words are instances of internal attention. Thus, when the participant is focused on the environment, they are not expending much effort to search for their memory and are instead externally oriented, referencing multiple objects in the environment. The 'familiarity flip of attention' framework proposed by Cleary et al. (2023b, 2025) aligns with this idea that déjà vu would be more characterized by internal attention when the environment is detected to be familiar. However, very importantly, the difference in the number of objects mentioned for participants when reporting déjà vu versus not reporting déjà vu was not statistically significant, according to a paired samples t-test. Thus, further work is needed to replicate these patterns with new data and increase sample size. While the current work was in line with previous power analyses for the déjà vu analyses, so the present sample may be too small to detect small effects. Future research with increased sample sizes would greatly increase power and lower the risk of model overfitting, allowing for a more fine-grained replication.

Like the number of objects, instances of elaboration was also an important factor in most models, but there was no significant difference between déjà vu and no déjà vu. Interestingly, these tended to have high overlap with instances of confabulation, but high values of elaboration were attributed to the no déjà vu report group. Confabulation was defined as recalling incorrect details from episodic memory (either within or outside of the experiment). Elaboration was described as attributing aspects of the environment to something outside of VR. These features were coded separately, and raters were blind to what other features were being rated. Thus,

reporting ‘this seems familiar like my elementary school’ would be both an instance of confabulation and elaboration. One interpretation is that confabulation is more autobiographical, while elaboration is not always, or that this is simply overfitting as the inferential statistics revealed no statistical difference. This speaks to the limitation of feature-based ML algorithms, particularly with manually coded features. The way the researcher defines and explains each feature could have an impact on ratings. Future work should seek to replicate the trend of elaboration being predictive of no déjà vu report, to ensure it is not merely an artifact of an overfit model.

Future Directions

Future research could explicitly test hypotheses born out of the present work. Previously, Long and Kuhl (2021) implemented a memory paradigm where, at test, participants were instructed on whether a stimulus would be New or Old. This New or Old cue would indicate how the participant should cognitively engage with the task, focusing on encoding it if it was New or on retrieving a previously viewed item if it was Old. This approach is theorized to differentially create an internal and external attention orientation for participants (Long & Kuhl, 2021). A similar approach could be taken with the present study’s scenes, to examine how verbalizations about scenes change according to the participant’s attentional orientation. If recollective confabulation is associated with internal attention, then when in the external orientation, participants should be less likely to show confabulation (and more likely when in the internal orientation). Furthermore, if elaboration and attention to objects are associated with external attention, then being in the external orientation should lead to greater elaboration and attention to objects; conversely, being in an internal orientation should lead to less elaboration and less focus on objects.

Conclusion

The current analytic approach presented evidence of a positive relationship between filler words, confabulation and the experience of déjà vu. I theorize that these are indicative of the internal attention orientation of déjà vu, which warrants further investigation and understanding. In terms of déjà vu phenomenology, sentiment analyses suggested an association between déjà vu reports and relative positivity; however, positive valence was not a top contributing factor to ML models' ability to discriminate déjà vu from non-déjà vu reports.

The mechanism of déjà vu continues to present opportunities for researchers to study attention mechanisms, and more recently internal attention. The present work also presents a novel analysis approach which seeks convergent evidence from the combination of inferential statistics and explainable ML classification models to predict the characteristics of an involuntary thought type.

REFERENCES

Adachi, N., Adachi, T., Kimura, M., Akanuma, N., & Kato, M. (2001). Development of the Japanese version of the Inventory of Déjà vu Experiences Assessment (IDEA). *Seishin Igaku (Clinical Psychiatry), 43*(11), 1223–1231.

Aitken, C. B., & O'Connor, A. R. (2020). Converging on an understanding of the déjà vu experience. In Cleary, A.M. & Schwartz, B.L. (Eds.) *Memory Quirks* (pp. 288-305). Routledge.

Andonovski, N., & Michaelian, K. (2023). Accounting for the strangeness, infrequency, and suddenness of déjà vu. *Behavioral and Brain Sciences*, *46*, e358. https://doi.org/10.1017/S0140525X23000237.

Barzykowski K, Moulin C. J. A. (2023) Are involuntary autobiographical memory and déjà vu natural products of memory retrieval? Behavioral and Brain Sciences 46, e356: 1–67. http://dx.doi.org/10.1017/S0140525X22002035

Bergstra, J., Yamins, D., & Cox, D., (2013). Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. *Proceedings of the 30th International Conference on Machine Learning*, in *Proceedings of Machine Learning Research* 28(1), 115-123 https://proceedings.mlr.press/v28/bergstra13.html

Brown, A. S. (2003). A review of the déjà vu experience. *Psychological Bulletin, 129*(3), 394–413. https://doi.org/10.1037/0033-2909.129.3.394

Castillon, I., Chartier, T., Venkatesha, V., Okada, N. S., Davis, A., Cleary, A. M., & Blanchard, N. (2024). Automatically identifying the human sense of familiarity using eye gaze

features. *Lecture Notes in Computer Science*, 291–310. https://doi.org/10.1007/978-3-031-60405-8_19

Cero, I., Luo, J., & Falligant, J. M. (2024). Lexicon-based sentiment analysis in Behavioral Research. *Perspectives on Behavior Science*, *47*(1), 283–310. https://doi.org/10.1007/s40614-023-00394-x

Chartier, T., Castillon, I., Venkatesha, V., Cleary, A. M., & Blanchard, N. T., (2024) Using Eye Gaze to Differentiate Internal Feelings of Familiarity in Virtual Reality Environments: Challenges and Opportunities. *Annual Review of Cybertherapy and Telemedicine*, 96

Christenfeld, N. (1994). Options and UMS. *Journal of Language and Social Psychology*, *13*(2), 192-199. https://doi.org/10.1177/0261927X94132005

Cleary, A. M. (2004). Orthography, phonology, and meaning: Word features that give rise to feelings of familiarity in recognition. *Psychonomic Bulletin & Review, 11*, 446- 451. https://doi.org/10.3758/BF03196593

Cleary, A.M., Ryals, A.J. & Nomi, J.S. (2009) Can déjà vu result from similarity to a prior experience? Support for the similarity hypothesis of déjà vu. *Psychonomic Bulletin & Review* 16, 1082–1088. https://doi.org/10.3758/PBR.16.6.1082

Cleary, A. M., Brown, A. S., Sawyer, B. D., Nomi, J. S., Ajoku, A. C., & Ryals, A. J. (2012). Familiarity from the configuration of objects in 3-dimensional space and its relation to déjà vu: A virtual reality investigation. *Consciousness and Cognition, 21(2), 969–975.* https://doi.org/10.1016/j.concog.2011.12.010

Cleary, A.M. & Claxton, A.B. (2018). Déjà vu: An illusion of prediction. Psychological Science, 29, 635-644. https://doi.org/10.1080/09658211.2018.1503686

Cleary, A.M., Huebert, A.M., & McNeely-White, K.L. (2020). The déjà vu phenomenon's entry into the realm of science. In Cleary, A.M. & Schwartz, B.L. (Eds.). Memory Quirks: The Study of Odd Phenomena in Memory. Routledge. (pp. 271-287).

Cleary, A. M., McNeely-White, K. L., Huebert, A. M., & Claxton, A. B. (2021). Déjà vu and the feeling of prediction: An association with familiarity strength. *Memory, 29(7),* 904– 920. https://doi.org/10.1080/09658211.2018.1503686

Cleary, A.M. & Brown, A.S. (2022). The Déjà vu Experience (2nd Edition). Routledge.

Cleary, A. M., Poulos, C., & Mills, C. (2023a). A possible shared underlying mechanism among involuntary autobiographical memory and déjà vu. *The Behavioral and brain sciences*, *46*, e361. https://doi.org/10.1017/S0140525X23000079

Cleary, A. M., Irving, Z. C., & Mills, C. (2023b). What flips attention? Cognitive Science, 47(4), e13274. https://doi.org/10.1111/cogs.13274

Cleary, A. M., McNeely-White, K. L., Neisser, J., Drane, D. L., Liégeois-Chauvel, C., & P Pedersen, N. (2025). Does familiarity-detection flip attention inward? The familiarity-flip-of-attention account of the primacy effect in memory for repetitions. *Memory & cognition*, *53*(5), 1622–1635. https://doi.org/10.3758/s13421-024-01673-x

Coltheart M. (2017). Confabulation and conversation. *Cortex, 87*, 62–68. https://doi.org/10.1016/j.cortex.2016.08.002

Ericsson, K. A., & Simon, H. A. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture, and Activity*, *5*(3), 178-186. https://doi.org/10.1207/s15327884mca0503_3

Feuerriegel, S., Maarouf, A., Bar, D., Geissler, D., Schweisthal, J., Prollochs, N., Robertson, C. E., Rathje, S., Hartmann, J., Mohammed, S. M., Netzer, O., Siegel, A. A., Plank, B., Van

Bavel, J. J., (2025) Using natural language processing to analyze text data in behavioral science. *Nature Reviews, 4,* 96-111. https://doi.org/10.1038/s44159-024-00392-z

Huebert, A.M., McNeely-White, K.L., & Cleary, A.M. (2022). Can cue familiarity during recall failure prompt illusory recollective experience? *Memory & Cognition, 50*, 681 - 695.

Huebert, A. M., Carlaw, B. N., McNeely-White, K. L., & Cleary, A. M. (2025). I want to know why this feels so familiar: Familiarity-detection during recall failure prompts curiosity and information seeking. *Cognition*, *265*, 106286. Advance online publication. https://doi.org/10.1016/j.cognition.2025.106286

Hutt, S., Wong, A., Papoutsaki, A., Baker, R. S., Gold, J. I., & Mills, C. (2024). Webcam-based eye tracking to detect mind wandering and comprehension errors. *Behavior research methods*, *56*(1), 1–17. https://doi.org/10.3758/s13428-022-02040-x

Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of social media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014

Killingsworth, M. A., & Gilbert, D. T. (2010). *A Wandering Mind Is an Unhappy Mind. Science, 330(6006), 932–932.* doi:10.1126/science.1192439

Kuvar, V., Blanchard, N., Colby, A., Allen, L., & Mills, C. (2023a). Automatically detecting task-unrelated thoughts during conversations using keystroke analysis. *User modeling and user-adapted interaction*, *33*(3), 617–641. https://doi.org/10.1007/s11257-022-09340-z

Kuvar, V., Kam, J. W., Hutt, S., & Mills, C. (2023b). Detecting when the mind wanders off task in real-time: an overview and systematic review. In *Proceedings of the 25th international conference on multimodal interaction* (pp. 163-173).

Long, N. M., & Kuhl, B. A. (2021). Cortical Representations of Visual Stimuli Shift Locations with Changes in Memory States. *Current biology: CB*, *31*(5), 1119–1126.e5. https://doi.org/10.1016/j.cub.2021.01.004

Loria, S. (2018). textblob documentation (Release 0.19). https://textblob.rea dthedocs.io/en/dev/

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds), *Advances in Neural Information Processing Systems* (Vol. 30).

McNeely-White, K. L., & Cleary, A. M. (2023). Piquing Curiosity: Déjà vu-Like States Are Associated with Feelings of Curiosity and Information-Seeking Behaviors. *Journal of Intelligence*, *11*(6), 112. https://doi.org/10.3390/jintelligence11060112

Mildner, J. N., & Tamir, D. I. (2024). Why do we think? The dynamics of spontaneous thought reveal its functions. *PNAS nexus*, *3*(6), pgae230. https://doi.org/10.1093/pnasnexus/pgae230

Mohammad, S., & Turney, P. (2010, June). Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text* (pp. 26-34).

Monin B. (2003). The warm glow heuristic: when liking leads to familiarity. *Journal of personality and social psychology*, *85*(6), 1035–1048. https://doi.org/10.1037/0022-3514.85.6.1035

Moulin, C. J. A. (2018). Recollective confabulation. In C. J. A. Moulin, *The cognitive neuropsychology of déjà vu* (pp. 112-134). Routledge.

Mumoli, L., Tripepi, G., Aguglia, U., Augimeri, A., Baggetta, R., Bisulli, F., Bruni, A., Cavalli, S. M., D'Aniello, A., Daniele, O., Di Bonaventura, C., Di Gennaro, G., Fattouch, J., Ferlazzo, E., Ferrari, A., Giallonardo, A., Gasparini, S., Nigro, S., Romigi, A., Sofia, V., Tinuper, P., Grazia Vaccaro, M., Zummo, L., Quattrone, A., Gambardella, A., & Labate, A. (2017). Validation Study of Italian Version of Inventory for Déjà Vu Experiences Assessment (I-IDEA): A Screening Tool to Detect Déjà Vu Phenomenon in Italian Healthy Individuals. *Behavioral sciences (Basel, Switzerland)*, *7*(3), 50. https://doi.org/10.3390/bs7030050

Okada, N. S., McNeely-White, K. L., Cleary, A. M., Carlaw, B. N., Drane, D. L., Parsons, T. D., McMahan, T., Neisser, J., & Pedersen, N. P. (2023). A virtual reality paradigm with dynamic scene stimuli for use in memory research. Behavior Research Methods, 56(7), 6440–6463. https://doi.org/10.3758/s13428-023-02243-w

Perrin, D., Moulin, C. J. A., & Sant'Anna, A. (2023). *Déjà vécu* is not *déjà vu*: An ability view. *Philosophical Psychology*, *37*(8), 2466–2496. https://doi.org/10.1080/09515089.2022.2161357

Poulos, C., Zamani, A., Pillemer, D., Leichtman, M., Christoff, K., & Mills, C. (2023). Investigating the appraisal structure of spontaneous thoughts: evidence for differences among unexpected thought, involuntary autobiographical memories, and ruminative thought. *Psychological research*, *87*(8), 2345–2364. https://doi.org/10.1007/s00426-023-01814-y

Ryals, A. J., & Cleary, A. M. (2012). The recognition without cued recall phenomenon: Support for a feature-matching theory over a partial recollection account. Journal of Memory and Language, 66, 747–762. https://doi.org/10.1016/j.jml.2012.01.002

Schwartz, B. L., & Cleary, A. M. (2025). Tulving's (1989) Doctrine of Concordance Revisited. *Journal of cognition*, *8*(1), 36. https://doi.org/10.5334/joc.447

Sno, H. N., Schalken, H. F., de Jonghe, F., & Koeter, M. W. (1994). The inventory for déjà vu experiences assessment. Development, utility, reliability, and validity. *Journal of nervous and mental disease*, *182*(1), 27-33. https://doi.org/10.1097/00005053-199401000-00006

Sripada, C., & Taxali, A., (2020) Structure in the stream of consciousness: Evidence from a verbalized thought protocol and automated text analytic methods. *Consciousness and Cognition, 85*, 103007, https://doi.org/10.1016/j.concog.2020.103007

Steadman, C., Venkatesha, V., Poulos, C., Cleary, A. M., Blanchard, N., & Mills, C. (2025). Involuntary Thoughts in Older Versus Younger Adults: A Multidisciplinary Approach to Investigating Déjà Vu, Involuntary Autobiographical Memories, and Unexpected Thoughts. *Technology, Mind, and Behavior*, *6*(2). https://doi.org/10.1037/tmb0000150

Weinstein, Y. (2018). Mind-wandering, how do I measure thee with probes? Let me count the ways. *Behavior Research Methods*, *50*(2), 642-661. https://doi.org/10.3758/s13428-017-0891-9

Venkatesha, V., Poulos, M. C., Steadman, C., Mills, C., Cleary, A. M., & Blanchard, N. (2025). A Linguistic Analysis of Spontaneous Thoughts: Investigating Experiences of Deja Vu, Unexpected Thoughts, and Involuntary Autobiographical Memories. *arXiv preprint arXiv:2507.04439*.

Yanai, I., & Lercher, M. J. (2025). Openness guides discovery. *Nature Biotechnology*, *43*(5), 667–668. https://doi.org/10.1038/s41587-025-02635-7

Appendix A

Manually and Automatically Coded Features for ML Models

Feature	Trial Type	Models[a]	Coding	Operational Definition
Word Count	Encoding, Test	E, R, RF	Automatic	Total words
Scene Name	Encoding	E	Automatic	Mention of scene name
Filler Words	Encoding, Test	E, R, RF	Automatic	Total filler words (ex. uh, um, hmm)
Content Words	Encoding, Test	E, R, RF	Automatic	Words remaining after removing filler words and stop words (ex. a, the, and)
Sentiment Features	Encoding, Test	E, R, RF	Automatic	Quantitative results from sentiment analysis
Repeated Words	Encoding, Test	R, RF	Manual, 3rd Rater Automatic[b]	Multiple words repeated immediately (i.e. stuttering, restarting sentence)
Involuntary Autobiographical Memories	Encoding	E	Manual	Referencing autobiographical memories (a form of elaboration)
Confabulation	Test	R, RF	Manual	Explaining familiarity through autobiographical memories or incorrect prior scene (i.e. not a correct answer)
Elaboration	Encoding, Test	E, R, RF	Manual	Connecting aspects of the scene to things outside of VR (ex. this looks like a TV show)
Generating Questions	Encoding, Test	E, R, RF	Manual	Proposing questions about details of scenes
Specific Objects	Encoding, Test	E, R, RF	Manual	Referencing specific objects by name in the scene
Spatial Features of Objects	Encoding, Test	E, R, RF	Manual	Referencing size or location of specific objects in the scene
Incomplete Sentences	Encoding, Test	R, RF	Manual	Sentences that do not end naturally but make a topic jump to another sentence.

[a]E = Encoding, R = Retrieval, RF = Retrieval Failure

[b]Programming script developed to count duplicate words, dyads and triads.

Appendix B: Best and Median Model Hyperparameters

Feature Set	Iteration	Classification Model	Hyperparameters
Encoding	Best	SVC	C: 1000 gamma: 1 kernel: sigmoid
	Median	Random Forest	Criterion: gini Max features: log2 n estimators: 341
Retrieval	Best	AdaBoost	n estimators: 63 learning rate: 0.5081632653061224
	Median	AdaBoost	n estimators: 30 learning rate: 1.6510204081632653
Retrieval Failure	Best	AdaBoost	n estimators: 30 learning rate: 1.7326530612244897
	Median	AdaBoost	n estimators: 30 learning rate: 1.8959183673469386