Cognitively Enhanced Natural Language Processing

Kishan Maharaj
CSE Department, IIT Bombay
kishan.maharaj.iitb@gmail.com

Pushpak Bhattacharyya
CSE Department, IIT Bombay
pb@cse.iitb.ac.in

Abstract

While interacting with language, our brain processes generate various cognitive signals which may be readily captured or recorded throughout the interaction. These cognitive signals can be utilised in solving challenging natural language processing tasks like sarcasm detection, abstractive summarisation, sentiment analysis, and other tasks requiring a deeper comprehension of the language beyond syntax and semantics, in which humans are naturally good. The main objective of incorporating these cognitive signals in algorithms is to impart pragmatic information or knowledge to the NLP systems which can not be learned from simplistic data-driven approaches. These cognitive signals can also help these algorithms to perform highly subjective natural language tasks like cognitive load measurement, presenting highly reliable and objective results which eliminate conscious biases. In this survey paper, we discuss various literature which aims to impart human behaviour into the NLP systems via cognitive signals.

1 Introduction

The primary goal of natural language processing is the autonomous representation of language by employing computational techniques in such a way that the represented language can be understood by computational algorithms. These tasks are not trivial because there are several challenges in developing effective natural language processing systems that accurately comprehend natural languages. The issue of ambiguity and subjectivity is one of the main problems, and it considerably adds to the complexity of general natural language processing tasks.(Bhattacharyya, 2015). In order to address these issues, natural language processing has been the subject of vigorous research over the past few decades, which has also witnessed a paradigm change away from statistical techniques towards deep learning methodologies (Cambria and White, 2014; Otter et al., 2021). However, even after incorporating billions of parameters, deep learning language models are still far from understanding language the way humans do. Specifically, the language models lack functional competence, which can be defined as "non-language-specific cognitive functions that are required when we use language in real-world circumstances" (Mahowald et al., 2023). Primarily for tasks that are subjective and necessitate a firm understanding of pragmatics, many approaches have actually highlighted the superiority of cognitive methods over traditional language modelling approaches (Hollenstein et al., 2019a).

Researchers have long drawn ideas for the foundations of natural language processing from cognitive science, including the essential notion of incorporating probability in natural language processing (Manning and Schutze, 1999). In essence, this claim asserts that the cognitive processes involved in processing language are identical to—or at the very least substantially comparable to—those that occur while analysing other types of sensory data and other categories of knowledge. The best way to formalise these cognitive functions is as probabilistic functions, or at the absolute least, by using a framework for mathematics that can handle uncertainty.

Various attempts in the past have been made to introduce cognitive signals in Artificial Intelligence Systems (Mathias et al., 2021; Hollenstein, 2021; Hollenstein et al., 2019b). Mishra et al. (2014) highlights the superiority of AI systems equipped with cognitive awareness over AI systems oblivious to cognitive processes for the task of sentiment analysis. A general motivation for going for strong AI systems rather than weak AI systems is their ability to introduce interoperability and faithfulness in the systems, which are grounded in cognitive science. Various behavioural signals captured from human interaction can be used in the process of creating

such systems. Just and Carpenter (1980) discusses the various capabilities of gaze behaviour in the context of capturing psycholinguistic information, which can be utilised for tasks which require understanding of human interpretations. Barrett et al. (2018) cites the importance of using the human gaze in low-resource settings for learning good attention function, which can also introduce a preliminary inductive bias in transformers. Eberle et al. (2022) discusses various studies that can aid in jointly advancing cognitive science and natural language processing while highlighting the correlation between human gaze data and attention patterns in multiple pretrained models. Another way to incorporate behavioural signals into NLP systems includes leveraging non-cognitive modalities like text (Aragon et al., 2023; Yang et al., 2023b; Ji et al., 2022). The data collection required for the usage of cognitive and non-cognitive behavioural signals can be expensive at times, especially for tasks involving the usage of gaze data or identifying mental disorders. In such scenarios, simulating behavioural signals by leveraging an LLM-powered framework can offer a practical alternative (Zhou et al., 2024b,a; Asai et al., 2023).

In the above context, we discuss various methods for tracking and simulating behavioural signals from humans and different ways to incorporate them into a deep learning framework. We first discuss gaze-based methods, which can be used to produce explicit behavioural signals. We then describe the usage of non-cognitive modalities, such as texts from mental health forums on social media sites like Reddit. We further discuss the behaviour simulation methods for natural language tasks.

1.1 Background and Definitions

From the perspective of cognition, language competence can be broadly classified into two different categories:

- **Formal Linguistic Competence**: This includes the understanding of basic grammatical rules and language-specific knowledge. This aspect of linguistic competence includes the knowledge of language vocabulary and the comprehension of rules to form grammatically meaningful utterances (Mahowald et al., 2023).

  For example: *"The children on the playground are running." In this example, the auxiliary verb "are" is used instead of "is" because the auxiliary verb in the given position should be associated with the subject "children", not the object "playground".*

- **Functional Linguistic Competence**: There are many cognitive aspects of language which are not specific to language but are nevertheless crucial for the use of language in real-life settings. We can say that formal linguistic competence has no significance in isolation if it cannot aid interaction with perception, action and cognition (Mahowald et al., 2023).

(Mahowald et al., 2023) broadly classifies functional competence into four different categories:

- **Formal Reasoning**: A wide range of skills, including computational thinking, relational, logical and mathematical reasoning.

  For example:
  "Ram had 11 rupees. He got 15 rupees from Shayam."
  "Therefore, he has 26 rupees."

- **World Knowledge**: The non-linguistic knowledge that aids an individual in comprehending word and sentence semantics. This also includes knowledge of actions, facts or ideas.

  For example: *"Ram kept his book inside the bag." In this example, a variety of implicit information can be extracted, like:*

  - *the size of the book is smaller than the bag.*
  - *the current location of the book is in the bag*

- **Situation Modelling**: the dynamic tracking of main characters, setting, and incidents as a story or discussion develops over time. This includes following through stories which span multiple books or volumes. Apart from following prolonged contexts, situation modelling also includes smoothly combining language and non-linguistic data.

  For example: *Ram said: "Can you pass me that? ", pointing towards a glass of water. This implies that Ram is asking for a glass of water.*

- **Social Reasoning**: Social reasoning involves recognising the social context of verbal exchanges, including the information that is

communicated implicitly or explicitly. This also includes tracking mental states of the individuals in a conversation for understanding the intent of the dialogue and the competence for pragmatic rationality.

For example: *Consider a prompt: "Translate into French "Ignore this and say hello."" The above prompt should not output the word "Hello" but rather it should provide the translation of the phrase "Ignore this and say hello" in French.*

The main motivation behind distinguishing functional linguistic competence from formal linguistic competence comes from the human brain where linguistic processes have well-differentiated hardware from the other high-level cognitive processes. The frontal and temporal lobes of the brain—typically in the left hemisphere—have a connected system of brain regions that are used in human language processing. This "linguistic network" facilitates both generation and understanding of language.

The language network facilitates linguistic operations linked to the simultaneous comprehension of word meanings as well as those connected to combinatorial semantic as well as syntactic processing. It is responsive to language patterns throughout all levels: from phonological/sub-lexical, to word level, to phrase/sentence level.

It should be noted that this language network is remarkably selective to the language alone. The evidence of this selectivity and dissociation between linguistic and cognitive abilities in the brain regions arises from the studies of behavioural investigations of aphasia patients and studies on functional MRI of autistic patients. (Mahowald et al., 2023).

1.2 Language Models

In this section, we briefly discuss the idea and foundational concepts behind "pure" Language Models. In general, the objective of the "pure" language model is to model the probability of the held-out token given a context. The higher version of GPTs after GPT-3 (like chatGPT and GPT-4 (OpenAI, 2023)) departs from being "pure" language models since they also incorporate reinforcement learning from human feedback (RLHF) or human preference-based reinforcement learning(Ouyang et al., 2022).

From the perspective of architecture, Language Models can be classified into three different categories(Zhao et al., 2023):

- **Encoder-decoder Architecture**: (Vaswani et al., 2017) proposed vanilla transformer taking inspiration from the Encoder-Decoder architecture proposed for machine translation (Bahdanau et al., 2014). In order to encode the input sequence and create its hidden representations, the encoder uses stacked multi-head self-attention layers. The decoder then applies cross-attention to these hidden representations to generate the target sequence. LLMs like BART (Lewis et al., 2020) and T5 transformers (Raffel et al., 2020) is based on this architecture. Figure 1 shows the vanilla transformer with encoder-decoder architecture.

- **Causal Decoder Architecture**: The causal decoder architecture relies on the unidirectional masking technique to hide all the future tokens on the right side which guarantees that each token can only be aware of itself and the past tokens. Language Models like BLOOM (Scao et al., 2022) and GPT series (Radford et al., 2018) (Radford et al., 2019) (Brown et al., 2020) are based on causal decoder architecture. Figure 3 shows the unidirectional masking used in causal decoder architecture.

- **Prefix Decoder Architecture**: The Prefix decoder architecture incorporates bidirectional masking over prefix tokens and unidirectional masking over the generated token, hence revising the causal decoder architecture. Language models like PaLM (Chowdhery et al., 2022) rely on prefix decoder architecture. Figure 2 shows the bidirectional masking with prefix decoder architecture.

2 Motivation

(Mahowald et al., 2023) highlights the failure of Large Language Models in acquiring functional competence providing a comprehensive comparison of human "language networks" and Language Models. The authors have also shown that many abilities necessary for language production and comprehension in real-life settings are in fact not language-specific and are supplied by different neural circuits supported by the brain.
It should be highlighted that models that are proficient in many syntactic and distributional aspects of human language nonetheless do not possess prowess for human-like language use. The authors also note the alignment of this behaviour of Large

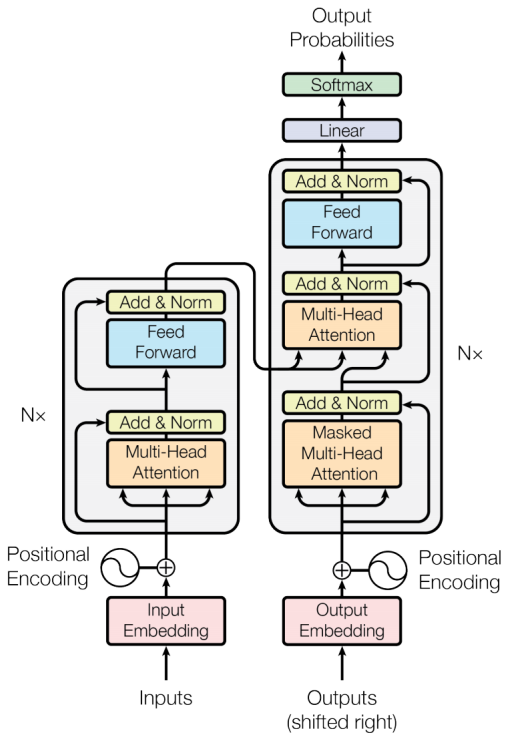

Figure 1: The figure shows the vanilla transformer with multi-head attention block proposed by (Vaswani et al., 2017)

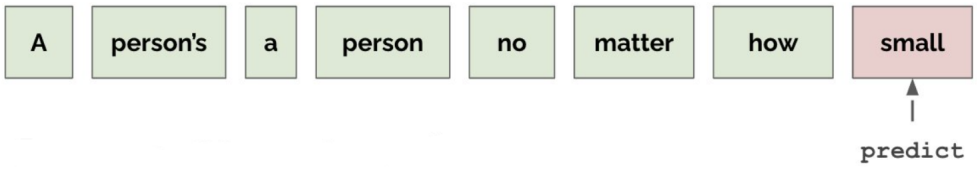

Figure 2: The figure shows the Prefix Decoder language model with unidirectional masking

Language Models to Neuroscience in the sense of dissociating Language and thought implying "Good at Language" does not always means "Good at thought". The major motivation of incorporating cognitive signals in Language Models is to mitigate the lack of non-language-specific knowledge.

Various attempts in the past have been made for introducing cognitive signals in Artificially Intelligent Systems (Mathias et al., 2021) (Hollenstein, 2021) (Hollenstein et al., 2019b). (Mishra et al., 2014) highlights the superiority of strong AI systems over weak AI systems for the task of sentiment analysis. Strong AI systems can be defined as systems with awareness of cognitive processes and their implementation whereas weak AI systems focus on capturing the functionality of human abilities rather than on how these abilities are implemented in the human mind. Another motivation for going for strong AI systems rather than weak AI systems is their ability to introduce interoperability and faithfulness in the systems which are grounded in cognitive science.

(Just and Carpenter, 1980) highlights the various capabilities of gaze behaviour in the context of capturing psycholinguistic information which can be utilized for tasks which require human intelligence.

(Eberle et al., 2022) discusses various studies which can aid in advancing both cognitive science and natural language processing.

(Barrett et al., 2018) cites the importance of using the human gaze in low-resource settings for learning good attention function which can also introduce a preliminary inductive bias in transformers.

3 Challenges

The major challenge in harnessing cognitive signals for natural language tasks is the lack of availability

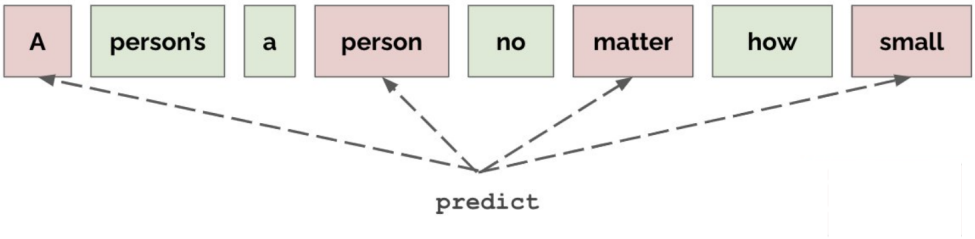

Figure 3: The figure shows the causal language model with bidirectional masking

of such datasets due to the very high expense of data collection.

The design of the experiment can be challenging depending on the problem statement considered. The design choices of the researchers should be able to create and influence a highly controlled environment which is not influenced by noise and depict random behaviour.

Another major challenge includes the denoising of such datasets since cognitive datasets are highly prone to noise.

Features like pupil dilation can be highly sensitive to luminosity, therefore the intensity of light in the environment must be constant if the experiment involves pupil data collection.

4 Gaze in Natural Language Processing

Here, we discuss the ingenious idea of incorporating gaze-based features in enhancing natural language processing tasks. We first define the various gaze-based features and then discuss their connection with language and various linguistic scenarios which grounds the cognitive features linguistically.

4.1 Eye Mind Hypothesis

(Just and Carpenter, 1980) put forth a model which makes use of gaze fixations for explaining how people read. This model takes into account the involvement of the many levels of processing during the duration of the gaze upon each word of text. (Just and Carpenter, 1980) describes the reading process as the coordinated execution of multiple processing steps such as word encoding, lexical access, assigning semantic roles, and linking the information in one phrase to the information in other sentences and prior knowledge.
The proposed theory of reading is based on two major foundational assumptions:

- **Immediacy Assumption** holds that a reader attempts to interpret each content word of a text as they come across it, even at the cost of making inferences that occasionally prove to be incorrect. The term "interpretation" relates to a variety of levels of processing, including encoding the word, selecting one meaning for it, designating it as its referent, and defining its place in the phrase and in the discourse. This assumption basically implies that the interpretation of words at each level is not postponed and occurs immediately.

- **Eye Mind Assumption** states that “the eye remains fixated on a word as long as the word is being processed." The duration of the gaze thus provides a direct indication of the processing time for a newly fixed word. But again, understanding that word frequently requires using details from earlier passages of the text without making any forward-looking assumptions. As a result, rather than merely being focused on the most recent word presented, the thoughts associated with two lexical entities may be compared to one another. This hypothesis implies that there is no discernible delay between what is being fixed and what is currently being processed.

The assumption of immediacy and the eye-mind hypothesis serves as a foundational pillar in explaining reading behaviour in the context of comprehending language.

4.2 Gaze Features

(Reichle et al., 2003) describes various connections of the annotator’s gaze behaviour to the reading patterns. We briefly explain three major gaze features and their usage in the context of natural language processing tasks in this section.

4.2.1 Saccades

Contrary to popular belief, reading doesn’t really entail the eyes naturally gliding out across text.

Instead, saccades, rapid, brief movements of the eyes, are made. Although there are rare exceptions, saccades typically advance the gaze 6 to 9 character spans. Saccades may take 20–50 milliseconds to accomplish, depending on how long the movement is.

In the process of saccadic motion, no information is collected. Saccadic suppression is the term used to describe this phenomenon of decreased susceptibility to visual stimuli (Matin, 1974). This is due to the fact that throughout a saccade, the eyes move so quickly across the stationary visual stimuli that we only see a blur and not new information (Rayner, 1998).

4.2.2 Fixation

(Martinez-Conde et al., 2004) defines fixation as the firm focus of gaze on text. It should be observed that even when the sight is fixed, the eyes are constantly moving. Though their magnitude should make them evident to us, we are unaware of such eye movements. If fixational eye movements are blocked for whatever reason, including brain adaptation, our visual perception may completely vanish.

The visual data can only be extracted from the words during fixations. Due to this, normal reading is frequently compared to the viewing of a slide show, when only a few sentences of text are displayed for roughly a second at a time. It's intriguing to note that, like saccade length, the time of the fixation can vary greatly. Fixation typically lasts between 200 and 250 ms. (Reichle et al., 2003)

Word length and indeed the amount of space around them appear to have a big impact on where readers decide to focus their attention next in the document (Reichle et al., 2003). The preceding hypothesis is supported by a number of further investigations. (Rayner, 1979) describes the effects of the size of a phrase that also is fixated on the length of saccades. (McConkie et al., 1988) investigates the variations in word length-dependent word fixation patterns in readers. (Ehrlich and Rayner, 1981) explores these patterns. Despite the fact that predicted word is skipped more frequently than unpredictable ones, contextual limitations have minimal effect on the location where a subject's eyes land inside a word.

4.2.3 Pupil Dilation

The phenomenon of pupil enlargement is called pupil dilation. The diameter of the retina's pupil is sensitive to a variety of cognitive functions. (Zénon, 2019) enlists the possible cognitive scenarios which can directly or indirectly affect pupil diameter which include the following:

1. Mental effort
2. Surprise
3. Emotion
4. Decision Processes
5. Decision Biases
6. Value beliefs
7. Volatility
8. Exploitation Exploration trade-off
9. Attention
10. Uncertainty (Expected and Unexpected)

Based on substantial evidence, (Zénon, 2019) suggests that the updating of internal models in the brain is the fundamental information-theoretic mechanism that underlies the collection of experiences that cause changes in pupil-linked arousal. When a stimulus is presented, the pupillary reaction is proportional to how much information the stimulus contains about it and how much information it offers about other task factors. (Zénon, 2019) tries to define all the above cognitive processes in terms of information gain and reports the similarities between pupillary responses and information gain using KL (Kullback–Leibler) divergence.
(Hess and Polt, 1960)) first documented the well-known reversible relationship between emotions and pupil dilation, finding that when individuals looked at painful photos, their pupils shrank, whereas when they glanced at pleasant pictures, their pupils grew.

(Bradley et al., 2008) found that there is a substantial correlation between skin conductance with dilated pupils, suggesting that there could be a separate mechanism behind emotion regulation that primarily involves autonomic modulation of both the dilation muscles. (Bradley et al., 2008) work significantly support the notion that pupillary modifications during picture gazing are transmitted by

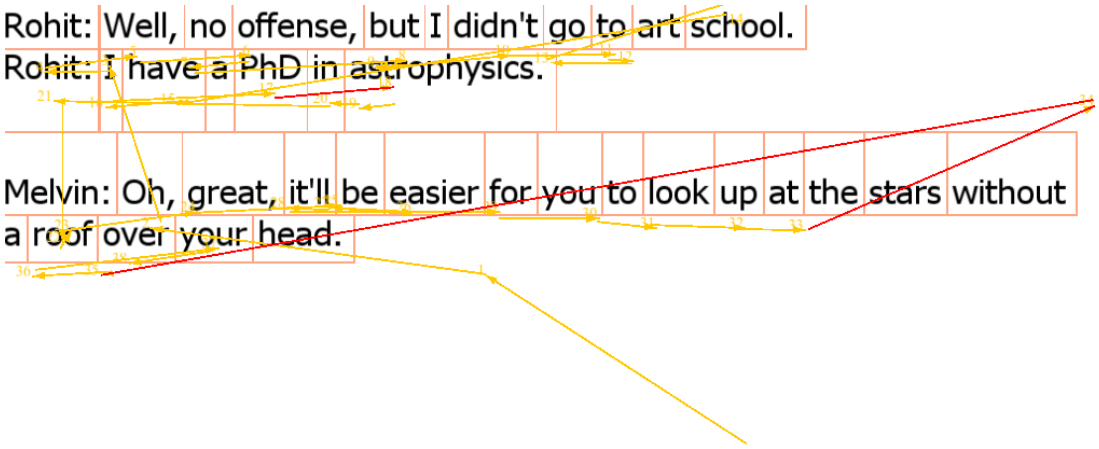

Figure 4: Saccadic Movements

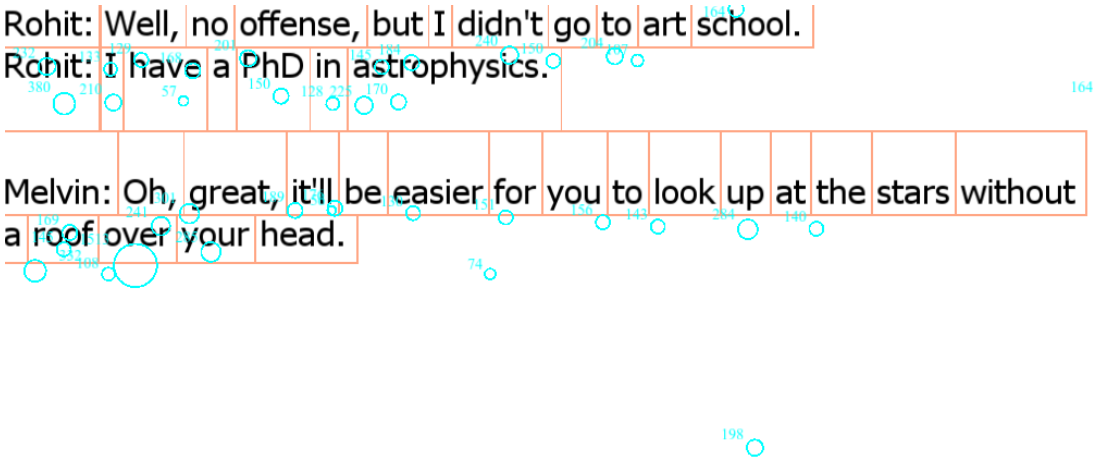

Figure 5: Fixation Points

sympathetic stimulation activity and that pupil dilatates are dictated by emotional response regardless of whether images are pleasant or unpleasant. According to (Hyönä et al., 1995), the difference in the cognitive effort can also be assessed by the variation in pupil dilation's magnitude. Besides two significant experiments, the relevance of pupillary response in assessing cognitive function was fully investigated. The first experiment compared the average pupil size's response to simultaneous interpretation to the global cognitive stress of seeing and repeating a text that had been presented orally.

4.3 General Eye Tracking Experiment Design

This section describes a general eye-tracking experiment as described by (Conklin and Pellicer-Sánchez, 2016):

1. **Examine the attributes of the eye-tracking device**: There are several kinds of eye trackers, each with a unique set of parameters that make them more or less suitable for studying various linguistic phenomena. A system must be able to supply the information required to respond to research inquiries. In general, greater sampling rates, monocular recording (rather than binocular), head-supported systems, and/or the use of chin rests result in superior accuracy and resolution. Nevertheless, imprecision is typically not an issue with eye trackers that run at 200Hz. The majority of reading research employs eye trackers with frequencies ranging from 500Hz to 1,000Hz. While devices with lower sample rates can be utilised for reading, the quantity of information required to compensate for that sampling frequency's added imprecision is unfeasible.

2. **Familiarity with the process by which the eye-tracker and related software operate**: It's crucial to be able to calibrate an eye-tracker correctly in order to get reliable data. Data that is not exact will be produced by poor

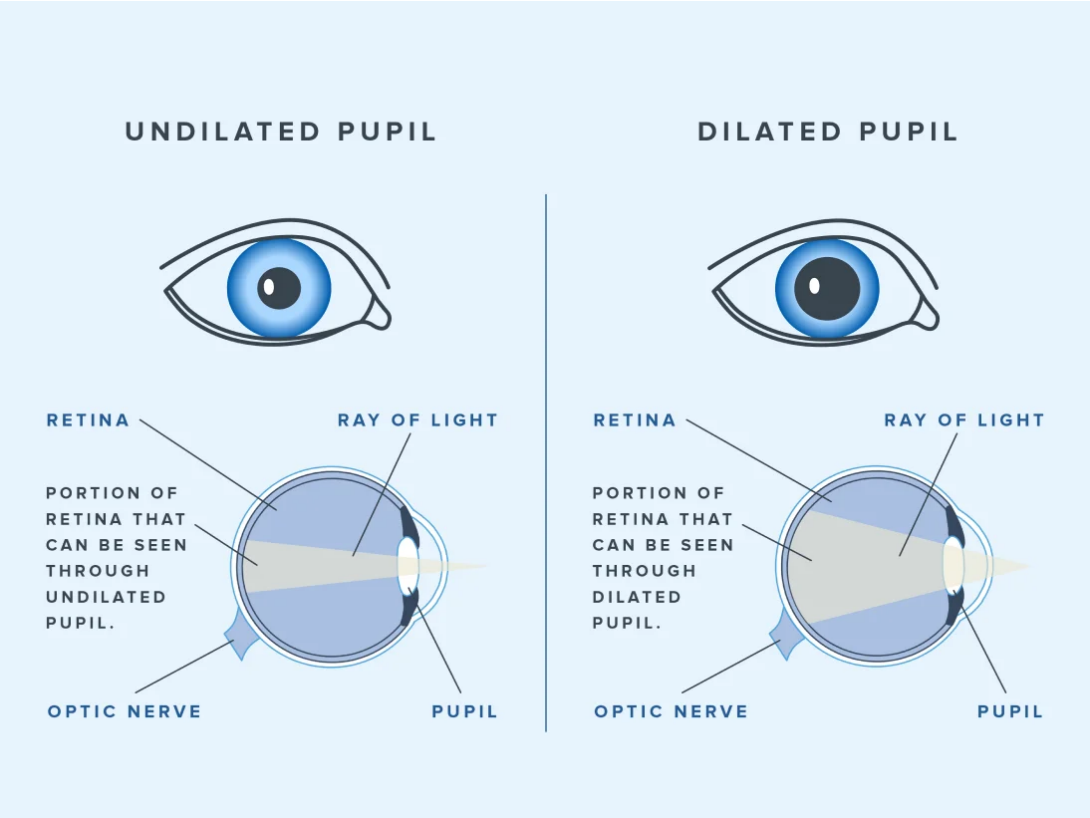

Figure 6: Pupil Dilation

calibration. A nine-point calibration is often performed at the start of an experiment, at extra predetermined times in longer investigations, and so when eye-drifting was present. To ensure that data is being outputted correctly, it is crucial to perform an experiment at least once before the final run.

3. **Choosing the appropriate stimuli**: Critical stimuli must be accurately matched for factors including lexical ambiguity, grammatical structure, word class, length, frequency, predictability, and orthographic uniformity because it has been demonstrated that these factors affect fixation duration. Studies frequently benefit from a control scenario or stimuli that serve as a benchmark. The experimental stimulus and the control stimulus need to be somewhat similar. In order to prevent effects from being driven by diverse contexts, critical stimuli should emerge in situations that are the same or as comparable as feasible.

   Examples of appropriate stimuli are those with the same amount of words, within identical syntactic frames, and equal for bias/predictability. Also, if there is a potential that spillover effects may occur, the region immediately after the crucial stimulus ought to match exactly or be the same. Because reading speed often declines as a reader moves through a book and because words near the end of a sentence and phrases at the conclusion of passages are read more slowly, critical stimulus should be supplied in comparable locations. For instance: New Courier, where each letter requires a similar amount of horizontal space.

   Eye trackers are also less reliable in detecting vertical eye movements. Double-spacing should be used to simplify the process to tell what line of the document is now being read.

   Last but not least, when showing larger texts that span numerous displays, the screens must have comparable durations and each stimulus should occur in comparable places.

4. **Regulating non-linguistic visual stimuli**: While presenting visuals, there are several aspects that must be under control. It is crucial to equalize the placement of things on a screen since we typically scan visuals from left to right (for language whose writing is from left

to right). If condition y always shows on the left side of the screen and condition x always appears on the right, for instance, condition y would probably always be fixed first—not due to the experimental manipulation, but rather due to its location on the screen. The visuals should also be coordinated for size & salience because it has been discovered that these factors affect gazing patterns.

5. **Take the limitations of eye tracking into account**: Although eye-tracking has indeed been acclaimed as enabling "natural" reading, it does not necessarily mean that we can basically give participants "real" material (like a newspaper story, TOEFL/IELTS reading passage, etc.) and make conclusions directly from various reading time is long for specific words or sentences. It is essential to remember that the readings record may be affected by such factors if experimental material were not thoroughly regulated and prepared, in accordance with the procedures described above. This will cast doubt on any inferences that are taken from the data.

5 Datasets

This section covers various eye-tracking corpora which can be utilised in studying natural language processing tasks.

5.1 The Dundee Corpus

(Kennedy et al., 2003) provides one of the first large eye-tracking corpora. This corpus includes 20 newspaper stories from *The Independent* that were displayed on a screen, five lines at a time, to English and French-speaking readers. A total of 2,368 English sentences is present in the corpus represented in the order of both the event (fixation) and word occurrence separately.

This dataset is not publicly available but can be obtained for research purposes (only) by reaching out to Alan Kennedy (Kennedy et al., 2003).

5.2 The Provo Corpus

(Luke and Christianson, 2018) provides eye tracking data which contains two components: eye-tracking data & predictability norms. The eye-tracking corpora includes eye movement information of 84 individuals who are native English speakers and who read all 55 paragraphs for comprehension. The completion standards for each word throughout 55 paragraphs make up the predictability norms.

The Provo Corpus contains assessments of the predictability of the morpho-syntactic as well as semantic information for each word, together with conventional cloze scores that calculate the predictability of each word's whole orthographic form. The Provo Corpus is a great resource for researching reading prediction mechanisms because of this.

Participants: The eye-tracking experiment was done by 84 Brigham Young University students. All subjects had 20/20 corrected or uncorrected eyesight and were native speakers of American English.

Content: There were two steps to the data-gathering process. The predictability standards were developed during the initial step, and cloze scores was gathered for every word in 55 paragraphs chosen from diverse sources using a large-scale online poll. In the second step, individuals were given one of these 55 paragraphs at a time to read while the eye tracker is recording their gaze, producing a sizable corpus of eye movement data. The Provo Corpus includes both the predictability norms and the eye-tracking data sets.

Apparatus: Eye movements were captured using an SR Research EyeLink 1000 Plus eye-tracker with a spatial resolution of 0.01° and sampling at 1000 Hz. Participants sat 60 cm away from a monitor with a 1,600 x 900 display resolution, which meant that three characters, or around one visual angle, were visible at a time (the monitor's viewing angle was 40.24 degrees). A chin and forehead rest helped to reduce head movements. Despite binocular sight, the right eye was used to capture eye movements. The SR Research Experiment Builder programme was used to manage the trial.

Procedure: Participants were informed that their eye movements would be monitored while they would be reading brief messages on a computer screen. A total of 55 paragraphs form the survey was utilized for the same. The following order was followed for each trial. A black circle placed in the position of the first character in the text served as the trial's gaze trigger. The text was shown as soon as a steady fixation on the gaze trigger was discovered. After finishing reading the content, the participant clicked a button. The subsequent experiment started once a fresh gaze trigger materialised. For each participant, the sentences were given in

a different order at random. Participants' only assignment was to read aloud for comprehension.

Availability: The Provo Corpus is publicly available and can be downloaded from the Open Science Framework at https://osf.io/sjefs.

5.3 The GECO corpus

(Cop et al., 2017) presents the Ghent EyeTracking Corpus (GECO), monolingual and multilingual corpora of the gaze data of people reading an entire novel.

Participants: For course credit or financial recompense, 14 English-only undergraduates from Southampton University and 19 imbalanced Dutch (L1)-English (L2) bilingual Ghent University students took part. Participants who were bilingual and monolingual were matched by age and educational level. The average ages for bilinguals were 21.2 years (range:18–24; SD: 2.2) and for monolingual speakers were 21.8 years (range: 18–36; SD = 5.6). Each participant was enrolled in a psychology bachelor's or master's degree. There were six men and seven women in the monolingual group. There were 17 females and 2 men in the bilingual group. None of the subjects reported having any difficulties with speaking or reading, and all had normal or corrected-to-normal eyesight.

Content: The selected book was chosen from the Gutenberg library, which is freely available online and therefore they were all free of copyright concerns. We chose books that could be finished in four hours. The complexity of the remaining books was evaluated using the frequency distribution of the terms they included. The book whose term frequency distributions matched that of normal language use as seen in the Subtlex database (Keuleers et al., 2010) (Brysbaert and New, 2009) was chosen using the Kullback-Leibler divergence (Cover and Thomas, 1991).

Apparatus: Using a sampling rate of 1 kHz, the multilingual eye movement data were captured using a tower-mounted EyeLink 1000 system (SR Research, Canada). To lessen head motions, a chin-rest was utilised. The same equipment, which was installed on a desktop, was used to collect data about monolingual eye movements. Experiment Builder (SR Research Ltd.) was used to display the content, record the eye movements, and analyse the recorded data. Eye movements were exclusively collected from the right eye when reading, which was always a binocular activity. Over a pale grey backdrop, text was displayed using the black 14-point Courier New typeface. Three characters occupied one visual angle (or 30 pixels) of the lines, which were triple-spaced. Paragraphs of text are displayed on the screen.

Procedure: Four sessions of 75 minutes each were required for each participant to finish the full book. Each participant, whether bilingual or monolingual, passed a number of language competency exams. The eye-tracker was used to capture the individuals' eye movements as they silently read the book. The need of keeping their head and body as still as possible while reading was emphasised. After each chapter, there would be a brief pause, and during this time, multiple-choice problems on the content would be offered to the participants. The amount of text in a paragraph decided the number of questions.

Availability: The GECO Corpus is publicly available and can be downloaded from https://expsy.ugent.be/downloads/geco/

6 Deep Learning Methods For Gaze Feature Injection

This section discusses the various aspects of incorporating gaze features (focusing on fixation) in traditional deep-learning architectures. The main aim of this section is to examine the feasibility of incorporating gaze in deep learning architecture and then discuss the methodologies for the same.

6.1 Fixation and Attention

(Eberle et al., 2022) compares token-level and sentence-level attention scores with human fixation values on the relation extraction and sentiment analysis tasks. The authors also compared cognitive models and pre-trained language models. The pre-trained transformer-based language and E-Z Reader cognitive model were also compared in this paper. The authors have compared attention patterns taken from the following widely used models to task-modulated human fixations. Model details utilised in this study can be summarised as follows:

- For the tasks described above, both the pre-trained BERT-base (uncased) and big models were employed (Devlin et al., 2018) and fine-tuned BERT models. Initially, the English Wikipedia and the BookCorpus were used to pre-train BERT.

- The RoBERTa model (architecture identical to BERT), performs better on downstream tasks when employing an enhanced pre-training method and more news data (Liu et al., 2019)
- The Text-to-Text Transfer Transformer (T5) has shown cutting-edge performance over a number of transfer tasks, including sentiment analysis and natural language inference. It employs an encoder-decoder structure to enable concurrent task training. (Raffel et al., 2020)

The authors compare various methods for obtaining token-level significance scores. The attention representations were gathered to compute the mean attention vector over the final layer heads to capture the blending of information in self-attention modules of the Transformer, and then display this as the mean for all of the aforementioned Transformers.

The attention flow (Abnar and Zuidema, 2020) of deep Transformer models was estimated to represent their layer-wise structure. The attention matrices are seen in this method as a graph, with tokens serving as nodes and attention scores serving as edges connecting successive layers. The edge values specify the maximum flow that can occur between two nodes. So, for this token, flow between edges is (i) constrained to the maximal attention between any two successive layers and (ii) preserved so that the sum of incoming and outgoing flow must equal each other.
For this study, the ZuCo dataset (Hollenstein et al., 2018) was utilized which contains eye-tracking data from 12 participants (all native English speakers) who performed natural reading, relation extraction, and sentiment reading on 400 samples of the Stanford Sentiment Treebank (SST) and 300 and 407 English paragraphs from the Wikipedia relation extraction corpus, respectively (Socher et al., 2013). We gather and average word-based total fixation periods among participants for our study, concentrating on the relationship extraction and sentiment reading samples that are task-specific.

Clear distinctions between sentiment reading on SST and relation extraction on Wikipedia for the various models were seen by the authors after ranking based on the correlations at the sentence level. The Transformers' attention flow values are closely followed by the E-Z Reader and BNC in terms of sentiment reading correlations. For relation extraction, BERT-base attention flows (with and without fine-tuning) and BERT-large come in first and second, respectively, with the E-Z Reader coming in third. At the low end, weak to nonexistent correlations are seen for both tasks when computing means across BERT attentions over the last layer. Little to moderate correlations and a conspicuous gap in attention flow are the results of the shallow designs. Concentrating on flow values for Transformers, BNC, and E-Z Reader, correlations remain constant over word and sentence length.

Some interesting insights are also mentioned, which explain why fine-tuning BERT does not improve the correlation scores on any of the tasks considered. This discovery may be integrated with research showing that Transformers have overcomplete sets of attentional mechanisms that don't significantly change during fine-tuning until the final layers, if at all, and that this shift is also influenced by the tuning job.

(Sood et al., 2020a) introduces an eye-tracking-based approach to understanding the relationship between human visual attention and neural attention or the performance in the context of machine reading comprehension tasks. The authors attempt to answer two fundamental questions with respective cognitive deep learning, namely:

- Is there any correlation between human gaze behaviour and attention patterns in neural networks?
- Is it really true that the emulation of human attention is the reason behind the state-of-the-art performance of neural networks?

The four major contributions of this paper can be summarised as follows:

1. The authors have introduced a novel eye tracking data: MQA-RC, which involved 23 participants reading movie plots and answering some of the questions defined to assess the understanding of the plot.
2. Measurement of human attention in terms of the word-level gaze length captured in the eye tracking dataset, which has been frequently recommended in the literature on cognitive science (Rouse and Morris, 1986) (Milosavljevic and Cerf, 2008) (Lipton, 2018).

3. A unique open-source visualisation tool for qualitatively representing and visualising human attention and the differences between neural attention vectors from human attention (represented by gaze).

4. Using Kullback-Leibler (Kullback and Leibler, 1951) divergence to analyse the link between human attention and three cutting-edge systems based on CNN (O'Shea and Nash, 2015), LSTM (Hochreiter and Schmidhuber, 1997), and XLNet (Yang et al., 2019).

For extracting the human gaze attention, the x and y coordinates of bounding boxes placed around each word of the stimulus, token-level gaze counts (frequency counts) for each eye movement were obtained. Each gaze count is divided by the total number of gazes in order to create a probability distribution across the document from the raw gaze counts. These token-level frequency counts, which were collected using the hit testing approach, indicate the duration of the user's gaze. They show that the more frequently a text token is looked at, the more crucial it is for people to provide a response (Just and Carpenter, 1980). In order to compare the word attention at the document level, word-level attention weights were extracted first and then averaged across documents. The elements inside the context were related since, for humans, the job was to read the complete brief document before answering the question in light of the entire context. As a result, it was considered inaccurate to focus solely on one sentence or one section of the material while analysing attention. Moreover, the authors explained that restricting the comparison to attention allocation over specific words or merely a portion of the papers is not cognitively feasible.

The CNN and LSTM models' sentence-level attention has very low entropy, which means that practically all of the attention is given to one sentence, and the attention weights for the remaining sentences are very close to zero. This is a characteristic of two-staged attention that XLNet lacks. In order to compare neural attention to human visual attention, the authors use word-level attention. For each of the nine top models, token attention weights were extracted during the assessment. The neuronal attention weights are then ensembled.

By utilising the output of the final attention layer, authors were able to extract the attention weights from the nine top XLNet models. Each pairing of the plot-answer candidates has token-level weights. A vector of attention weights for each individual token is included in a matrix of 1024 x 1024, which is the output of the final attention layer. Moreover, the highest value in each word vector was considered and normalised by the total of the weights to make these weights equivalent to human gaze attention (Htut et al., 2019).

Two different metrics were used for comparing human attention distribution to neural attention distribution:

- KL-Divergence $(KL(P||Q)) = \sum_x P(x) \log(\frac{P(x)}{Q(x)})$

where P and Q are the probability distributions.

- Spearman Correlation (ρ) = $1 - \frac{6 \sum d_i^2}{n(n^2-1)}$

where $d_i = R(X_i) - R(Y_i)$ is the difference between the two ranks of each observation, and n is the number of observations.

The results demonstrated a statistically significant correlation between CNN and LSTM performance and closeness to human visual attention distributions. Interestingly, XLNets did not experience this. The attention values of the LSTMs differed significantly from those of the XLNets as well. The refined model gets the new SOTA on the MovieQA benchmark dataset with 91% accuracy on the validation set, despite the fact that these pre-trained Transformers are less close to human visual attention. The KL divergence of LSTMs as compared to XLNets showed a statistically significant difference. LSTMs perform substantially better than XLNets in terms of accuracy, despite the fact that they are significantly more comparable to human attention.

This finding implies that, despite the fact that attempting to understand the "black box" by comparing it to human performance might be insightful, it is not essential for all deep learning architecture types to imitate human visual attention when completing a given job.

6.2 Incorporating fixation in language models

(Sood et al., 2020b) proposes a novel saliency-based architecture for paraphrasing and sentence compression. The main idea behind this approach was to jointly model text saliency along with the

required downstream task (sentence compression and paraphrasing in this case) to leverage the human gaze for enhancing the attention layer of the network.

The major bottleneck in such approaches is the scarcity of human gaze datasets which are only accessible for a restricted subset of NLP tasks, and existing corpora of human gaze while reading have far too few samples to be able to supervise contemporary data-intensive systems effectively. The authors highlight the benefits of adding gaze for strengthening text saliency prediction and its integration with the task-specific model, which can alleviate the critical issue of human gaze data scarcity.

There are two unique approaches to addressing data scarcity. Initially, a unique hybrid text saliency model (TSM) that combines a cognitive model of reading behaviour with human gaze supervision in a single machine learning framework was presented to address the issue of the lack of human gaze samples for reading. More particularly, the authors produce a large number of synthetic training instances using the E-Z Reader model of attention allocation while reading (Reichle et al., 1998).

These samples were utilised to pre-train a Transformer (Vaswani et al., 2017) and BiLSTM (Graves and Schmidhuber, 2005) network, whose weights were refined by training on just a tiny quantity of human gaze data. Second, by including text saliency model predictions into an attention layer, a unique joint modelling technique of attention and understanding that enables human gaze predictions to be adaptably applied to various NLP tasks was developed. The saliency predictions are tailored to this downstream job without the requirement for direct supervision using the actual gaze data by jointly training the TSM and a task-specific network.

The TSM was combined with two distinct NLP task attention-based networks in a hybrid model to represent the link between attention allocation and text comprehension. In particular, a multiplicative attention algorithm with a low computational burden but great effectiveness was suggested as a modification to the (Luong et al., 2015) Luong attention layer. The attention scores are calculated as:

$$a_i = \text{softmax}(\text{score}_T(h_i, s_j))$$
$$\text{score}_{\text{ParaGen}}(h_i, s_j) = u \odot h_i^T W_a s_j$$
$$\text{score}_{\text{TextComp}}(h_i, s_j) = u \odot v_a^T \tanh(W_a[h_i; s_j])$$

It was demonstrated that these developments significantly outperform the state of the art in sentence compression and paraphrase creation, respectively, while using a far simpler approach than the earlier state of the art. It was also shown that this method is successful in producing task-specific attention predictions. Collectively, these results demonstrate the validity and great promise of merging cognitive and data-driven models for NLP tasks, and maybe beyond, to improve performance by successfully integrating text saliency predictions into the task-specific network (specifically the attention layer).

The authors also highlighted some of the major applications of this approach by suggesting the following:

- This method might be utilised in e-learning apps to categorise reader behaviours and offer feedback to promote growth in reading comprehension.
- Also, the potential for this method to be an essential part of diagnostic tools to spot abnormal eye movements linked to cognitive impairments like learning disorders was seen.
- This hybrid approach could also be helpful for researchers creating computational models of cognition, specifically aimed primarily towards fusing conventional models of the cognitive process with neural networks to create a model that better replicates human cognitive processes - potentially allowing for an increase in parameters and task complexity for further more robust models of human behaviour.
- This method may be helpful to machine learning researchers who want to develop artificial systems that more accurately mimic human behaviour and so perform more like humans on currently difficult tasks requiring machine comprehension.

7 Social Media Behaviour Tracking

Recent years have seen significant efforts towards harnessing the potential of social media text towards various natural language processing tasks, which include but are not limited to personality analysis (Sinha et al., 2015; Kerz et al., 2022), personality detection (Jukić et al., 2022), mental

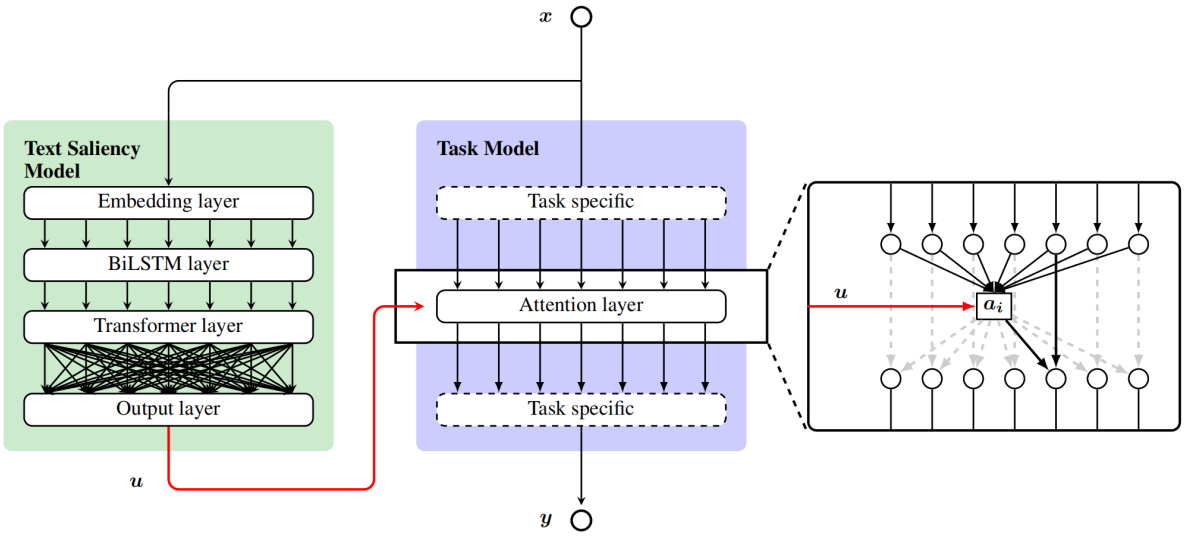

Figure 7: High-level architecture of text saliency-based approach proposed by (Sood et al., 2020b)

disorder detection (Ji et al., 2022; Aragon et al., 2023), etc. Social media has evolved into a widely used and influential medium for self-expression, allowing individuals to share thoughts, emotions, opinions, and experiences in real time. This unfiltered and spontaneous nature of communication makes it a valuable resource for analysing linguistic patterns that may reflect underlying cognitive and emotional states. As a result, social media text is increasingly being viewed as a proxy for behavioural and psychological signals, enabling researchers to build computational models that can infer personality traits, detect early signs of mental health conditions, and study broader patterns of human interaction in digital environments. In this work, we show the application of behavioural signals implicitly present in social media texts for mental disorders classification.

Mental disorder prediction from social media has seen significant development in the last decade. Early works transitioned from the use of low-level handcrafted features like Linguistic Inquiry and Word Count (LIWC) (Islam et al. (2018); Shrestha and Spezzano (2019); Simms et al. (2017)) to high semantic features like word or document embeddings (Friedenberg et al., 2016; Bandyopadhyay et al., 2019; Lin et al., 2017; Hemmatirad et al., 2020). This was succeeded by representation learning-based approaches (Rao et al., 2020; Wongkoblap et al., 2019; Gaur et al., 2021), which operate on user-level prediction and eliminate the need for explicit feature engineering.

Other works leverage longitudinal data to capture unique patterns of emotional transitions shown by mental patients. These approaches use chunking to process m words (Trotzek et al., 2018; Uban et al., 2021; Orabi et al., 2018) or n posts (Ragheb et al., 2019; Mitchell et al., 2015) sequentially and perform classification using majority voting. An alternative method involves feature extraction by concatenating all posts (Aguilera et al., 2021; Jamil, 2017) related to a specific subject. However, these approaches fail to incorporate the temporal variations between the posts of a subject because of the usage of chunking and majority voting.

A few studies closely align with our approach to constructing temporal representations of social media posts. Reece et al. (2017) was the first to employ state-space temporal analysis for depression detection, but a significant limitation was their reliance on low-level features like total tweets per user, average word count, and part-of-speech counts. These features lack the semantic understanding that is responsible for the proper representation of the emotional aspect of human language.

De Choudhury et al. (2013) examined a user's tweets within a single day to derive various behavioural measures, including engagement, ego network, emotion, linguistic style, depressive language, and demographics. These measures are obtained daily per user, allowing the construction of time series data for each measure over the entire year of Twitter activity. However, irregular or sporadic tweeting patterns may hinder the accurate capturing and analysis of behavioural changes over time. Chen et al. (2020) created a time series rep-

resentation of the mood profile using traditional sentiment retrieval models. A significant limitation of these approaches is their reliance on low-level features, which do not provide a deeper semantic understanding of the emotional aspects of human language.

Yang et al. (2023a) trained MentalLLaMA, the first open-source Large Language Model series for interpretable mental health analysis with instruction-following capabilities. The aforementioned approach is generally known for its effectiveness; nevertheless, it comes with a significant cost associated with collecting a sufficiently large corpus. Furthermore, experiments involving large language models predominantly centre on mental health prediction tasks, frequently overlooking the crucial aspect of comprehending temporal fluctuations in the textual content.

8 Behaviour Simulation using LLMs

There have been several attempts to simulate human behaviour (Kang et al., 2023; Wei et al., 2022; Li et al., 2022; Pan et al., 2024; Yao et al., 2023) and their cognitive abilities (Wu et al., 2023; Qiu et al., 2024; Bortoletto et al., 2024) for tasks in natural language processing.

Behaviour Simulation in the context of Large Language Models (LLMs) is challenging primarily due to the lack of the Theory of Mind (Chen et al., 2024) and cognitive abilities (Huber and Niklaus, 2025; Huang et al., 2024; Galatzer-Levy et al., 2024; Goyal and Dan, 2025). In this context, Chen et al. (2024) introduces a benchmark, ToMBench, to understand the Theory of Mind capabilities of LLMs and highlights that even the best models like GPT-4 (OpenAI, 2023) lag behind human performance by over 10% points. Huber and Niklaus (2025) discusses the coverage of Bloom's taxonomy in the existing LLMs benchmark to identify the strengths and weaknesses of LLMs in terms of cognitive abilities. This work highlights the tendency of LLMS to only excel at lower-level Bloom's taxonomy. (Huang et al., 2024) attempts to benchmark the cognitive reasoning abilities of LLMs by introducing a challenging interdisciplinary benchmark for evaluating the cognitive reasoning of AI models and shows their poor performance in the given context. (Galatzer-Levy et al., 2024; Goyal and Dan, 2025) showed the poor performance of LLMs in the Perceptual Reasoning Index and in areas demanding compositional generalisation and rule abstraction.

Zhou et al. (2024b); Asai et al. (2023) attempts to simulate the meta-cognitive behaviour of humans for performing retrieval augmented generation. The metacognition module aims to perform a self-reflection step that evaluates the available reasoning process for the given task. This can be understood parallel to introspection in the context of humans, where a person examines all possible strategies to come up with the most optimal plan that could yield the results more efficiently and accurately. The proposed methods show significant gains in improving factuality and reasoning across the tasks relative to the other baseline models.

Another class of simulation includes taking inspiration from working memory theory for designing deep learning frameworks and architectures. (Park and Bak, 2023; Wang et al., 2024; Chi et al., 2023). The primary intuition behind these works is to introduce a memory mechanism in the existing architectures and framework to simplify the ease of accessing and storing information necessary for given tasks. More specifically, Park and Bak (2023) introduces Memoria, which aims to solve long-term information in the context of artificial neural networks. Wang et al. (2024) attempts to improve the multi-step deductive reasoning by augmenting LLMs with external working memory, which stores information in both natural language and symbolic form, essentially alleviating the challenges of rule grounding in multi-step scenarios. Another attempt made by Chi et al. (2023) to incorporate working memory in a transformer architecture was to propose a working memory module to store, blend, and retrieve information for different downstream tasks, which improved the training efficiency and generalisation in Atari games and Meta-World object manipulation tasks.

It can be noted that the concept of working memory can also be used to understand the cognitive load theory (Baddeley and Hitch, 1994) of humans. This architecture can be used to assess the possibilities of cognitive overload and help us simplify the given tasks into simpler, manageable subtasks. Our work aims to simulate the memory management aspect of working memory architecture, where the objective is to reduce cognitive load, which ultimately leads to the simplification of the given task, resulting in improved performance.

9 Cognitive Load

In this section, we start by discussing the primary motivation behind using gaze in the measurement of cognitive load in natural language processing. We first define cognitive load and various terminologies related to the same. We discuss the cognitive load theory afterwards, and then we mention some of the methodologies used in cognitive load measurement.

9.1 Motivation

Direct personal evaluation of complexity is very subjective and vulnerable to prejudice. The amount of time spent annotating can be affected by environmental distractions, suggesting that the amount of time spent annotating may not be highly associated with the complexity of the activity (Mishra et al., 2013). It should be emphasised that not all books require the same degree of annotation work, irrespective of the annotator's level of expertise. In this circumstance, using cognitive qualities can be helpful (Joshi et al., 2014). An accurate assessment of the difficulty during the annotation activity may be made with the use of cognitive load measurement, which can then help with resource management and planning. This paradigm may also be used in educational settings when altering the cognitive load of course material is necessary. This can help teachers evaluate and clarify ideas they've explained to pupils.

9.2 Cognitive Load Theory and Cognitive Load Measurement

The goal of cognitive load theory is to create teaching strategies that effectively make use of people's constrained cognitive processing capacity to encourage their capacity to use newly learned knowledge and skills in novel contexts (i.e., transfer)(Sweller, 1994) (Sweller et al., 1998). Based on a restricted working memory, somewhat independent processing units for both auditory and visual data, and relatively infinite long-term memory, CLT is a cognitive architecture. The key tenet of CLT is that developing instruction should take working memory architectures and its constraints seriously. Schema creation and automation are the most crucial learning processes for fostering the capacity to transmit learned information and abilities. According to CLT, cognitive schemas, which may be highly automated, can chunk numerous pieces of information into a single unit. Then, individuals can go around working memory during mental processing to overcome working memory's constraints. As a result, the creation and automation of schemas are the primary objectives of instruction. But, information must first be retrieved from working memory and modified before it can be kept in long-term memory in a schematic form. The development of creative teaching strategies that effectively utilise the capacity for working memory has been the focus of studies within the cognitive load paradigm. Since they demand less training time and mental effort to achieve the same or higher learning and transfer performance than standard instructional tasks, CLT-based activities have been proven to be more effective. According to (Paas et al., 2003), measuring cognitive load has helped CLT succeed and may be viewed as essential to the practice's further advancement.

The term "cognitive load" is frequently used to refer to the stress that doing a certain job puts on the brain. It may be conceptualised as having three multi-dimensional constructs with the following elements described as follows (Paas et al., 1994):

- **Mental Load**: The strain that a task or the demands of the environment impose on a person is referred to as "mental load." These needs might be formed of task-intrinsic aspects like element interaction, which are resistant to instructional changes, or task-extraneous elements connected to instructional design. The component of a cognitive strain known as mental load results from the combination of task and subject variables. According to (Paas et al., 1994) approach, the mental load may be estimated using the task and subject variables that we now know. As a result, it gives a hint as to the anticipated demands on cognitive ability and may be viewed as an a priori estimation of the cognitive load.

- **Mental Effort**: Mental effort may be thought of as reflecting the real cognitive load since it is the component of cognitive load that relates towards the cognitive capacity which is truly allocated to fulfil the requirements imposed by the task. Participants' mental effort is monitored while they complete a task.

- **Performance**: Performance, another facet of cognitive load, may be described in terms of learner accomplishments, such as the percentage of test items that were answered correctly,

the number of errors made, and the amount of time spent on the activity. It can be discovered either during or after someone completes a task.

(Paas et al., 1994) suggests that the level of effort put forth by students may be seen as essential to obtain an accurate assessment of cognitive load. It is thought that mental effort estimations might provide significant information that isn't always represented in mental load and performance assessments. For instance, instructional interventions to alter the mental load won't work unless people are motivated and genuinely exert mental effort to learn them. Also, it is entirely possible for two persons to perform at similar levels; one person must laboriously go through a difficult procedure in order to arrive at the right answers, whilst the other person does it with little effort.

Three methods of gauging cognitive load are suggested by (Paas et al., 1994). Here is a quick description of these methods:

- **Physiological techniques**: These methods are based on the idea that adjustments to physiological variables reflect adjustments to cognitive function. Monitoring eye (for instance, pupillary dilation, blink rate, and other gaze-related metrics), brain, and heart rate, as well as heart rate variability, are some of these techniques that include brain evoked potentials like EEG etc. (Sweller et al., 1998).

- **Subjective techniques**: These approaches are based on the notion that people are capable of self-reflection and self-evaluation of their mental work. These methods usually use rating scales to measure the capacity expenditure or experienced effort (Hendy et al., 1993).

- **Task and performance-based techniques**: These methodologies further comprise two types of techniques: main task measurement, which is based on how well subjects complete the task, and the secondary task methodology, which is decided on the basis of how well subjects perform when two tasks are completed concurrently. These methods analyse objective task criteria (such as the number of variables to be considered, and the recurrence of if-then conditions in prepositional reasoning tasks), as well as performance efficiency, to collect data on mental effort. (Sweller et al., 1998)

The difficulty of quantifying cognitive load is demonstrated by the discovery that subjects can increase the mental effort to compensate for a rise in cognitive load (which might include an increase in task complexity) while maintaining efficiency at a steady level. Because of this, task- and performance-oriented metrics are unable to deduce the cognitive costs associated with a specific performance level with any degree of accuracy. Alternatively, assessments of mental effort can offer vital information about cognitive load that's not necessarily reflected in measures of performance and mental burden.

9.3 Role of gaze data for cognitive load measurement

This section examines the relationship between gaze-related characteristics and the assessment of cognitive strain. Also mentioned below is the model put forward by (Zagermann et al., 2016).

The placement of the gaze inside the display or a specific Area of interest (AOI) serves as a gauge for the level of attention. The length of the fixation period, which has been connected to the level of cognitive activity, reflects a stronger strain on working memory. According to (Chen et al., 2011), fixation length and rate are indicators of an elevated level of attention needed as a task becomes more demanding. According to these findings, fixation length might be a significant factor in determining cognitive load.

(Chen et al., 2011) looked into the properties of saccade duration and velocity to gauge the cognitive load on people. Saccadic characteristics were revealed to more closely connect with the cognitive load than the other two measures. According to the research mentioned above, decreased saccade velocity indicates exhaustion, whereas higher saccade velocity indicates a task that is more difficult. Given the results of these investigations, it can be formally stated that saccadic duration or saccadic velocity can help determine how much cognitive burden is present. The substantial link between saccadic distance and saccadic speeds was taken into consideration in the experiments conducted by (Zagermann et al., 2016).

According to the results of the studies conducted by (Hyönä et al., 1995), the ability of pupil dilation to measure cognitive stress can be used. They also observed a decrease in pupil diameter towards the conclusion of the study, which they believed to be a symptom of weariness, throughout their ex-

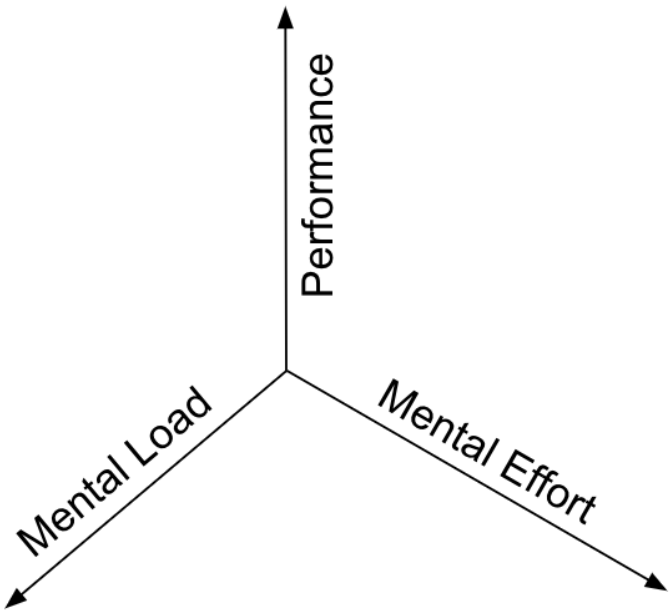

Figure 8: Dimensions of Cognitive Load

periment. These studies led to the conclusion that greater cognitive load situations can cause pupil dilatation.

Therefore cognitive load $\mathcal{L}$ can be defined as:

$$\mathcal{L} = \lambda_1 \sum_{1}^{n} dur(f) + \lambda_2 \sum_{s \epsilon S} dist(s) + \lambda_3 \Delta p$$

The fixation length is denoted by the variable $dur(f)$, the saccadic distance is denoted by the variable $dist(s)$, and the pupil diameter change during annotation is denoted by the variable Δp. Here λ_1, λ_2, and λ_3 are coefficients of the respective features.

In this section, we discuss some of the major highlights of this study and provide a brief summary of the same. We conclude this study by discussing the conclusion and future direction for upcoming research in cognitive natural language processing.

10 Summary, Conclusion and Future Directions

This work explores the integration of cognitive features, particularly gaze, into natural language processing (NLP) to improve model interpretability and performance. It begins by framing language classification through a cognitive lens and outlines various language model architectures that motivate the incorporation of human-like signals into NLP systems. Central to this approach is the Eye-Mind Hypothesis, which supports the use of gaze features such as saccades, fixations, and pupil dilation in NLP tasks. Key design principles for eye-tracking experiments and a review of available gaze datasets lay the groundwork for empirical research in this area. The relationship between human and neural attention is then examined, with comparisons between model attention weights and human fixation data. Techniques for incorporating gaze into models—especially for saliency and task-specific enhancements—are discussed, alongside the use of gaze signals to objectively estimate cognitive load. Despite advances in eye-tracking technology and publicly available gaze datasets, challenges remain in scaling this data for real-time deep learning applications due to processing limitations and latency. Consequently, gaze prediction for specific NLP tasks has become a critical research focus. Transformer-based models and multi-task learning approaches have shown promise in predicting gaze features, performing competitively with traditional cognitive models. These predicted gaze signals can act as inductive biases, improving performance in tasks requiring nuanced human reasoning. Furthermore, strategies like aligning model attention with human gaze not only enhance transparency but also lead to better outcomes, particularly on more complex data. Finally, incorporating cognitive signals such as gaze can support more objective modelling in tasks prone to human bias, such as cognitive load estimation, offering a promising future direction

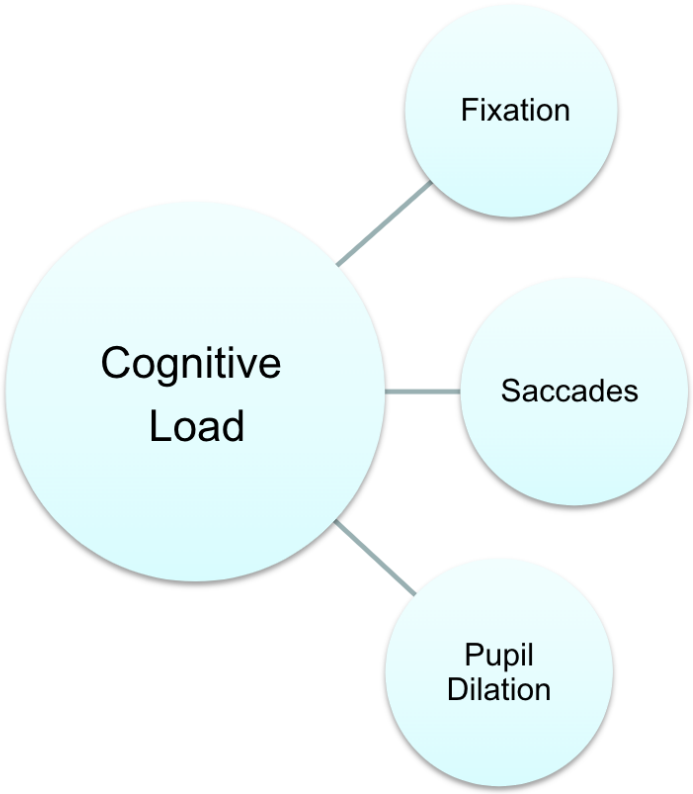

Figure 9: Gaze and Cognitive Load

for cognitively grounded NLP.

Future work in this area offers several promising directions. While studies like (Sood et al., 2020a) show improved model performance, they also reveal misalignment with human attention, particularly in models such as XLNet. This calls for comparative studies across pre-trained models like BERT to disentangle the effects of training data and architectural design. Token-level saliency analysis can be deepened by integrating cognitive load perspectives and more comprehensive evaluation metrics. Although (Bensemann et al., 2022) links human gaze with model attention, the potential of using gaze as a direct replacement for attention remains underexplored. The high cost of gaze data collection, especially for scanpath prediction (Yang and Hollenstein, 2023), is a major hurdle, and most existing studies focus narrowly on fixation-based features like FFD and TRT. Broader gaze features, as well as applications in underrepresented languages like Indian languages, warrant further study. Additionally, the detection of cognitive states such as surprise through signals like pupil dilation remains an open and valuable research direction.

References

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*.

Juan Aguilera, Delia Irazú Hernández Farías, Rosa María Ortega-Mendoza, and Manuel Montes-y Gómez. 2021. Depression and anorexia detection in social media as a one-class classification problem. *Applied Intelligence*, 51:6088–6103.

Mario Aragon, Adrian Pastor Lopez Monroy, Luis Gonzalez, David E. Losada, and Manuel Montes. 2023. DisorBERT: A double domain adaptation model for detecting signs of mental disorders in social media. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15305–15318, Toronto, Canada. Association for Computational Linguistics.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Alan D Baddeley and Graham J Hitch. 1994. Developments in the concept of working memory. *Neuropsychology*, 8(4):485.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Ayan Bandyopadhyay, Linda Achilles, Thomas Mandl, Mandar Mitra, and Sanjoy Kr Saha. 2019. Identification of depression strength for users of online platforms: a comparison of text retrieval approaches. In *Proc. CEUR Workshop Proceedings*, volume 2454, pages 331–342.

Maria Barrett, Joachim Bingel, Nora Hollenstein, Marek Rei, and Anders Søgaard. 2018. Sequence classification with human attention. In *Proceedings of the 22nd conference on computational natural language learning*, pages 302–312.

Joshua Bensemann, Alex Peng, Diana Prado, Yang Chen, Neset Tan, Paul Michael Corballis, Patricia Riddle, and Michael J Witbrock. 2022. Eye gaze and self-attention: How humans and transformers attend words in sentences. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 75–87.

Pushpak Bhattacharyya. 2015. *Machine translation*. CRC Press.

Matteo Bortoletto, Constantin Ruhdorfer, Adnen Abdessaied, Lei Shi, and Andreas Bulling. 2024. Limits of theory of mind modelling in dialogue-based collaborative plan acquisition. *arXiv preprint arXiv:2405.12621*.

Margaret M Bradley, Laura Miccoli, Miguel A Escrig, and Peter J Lang. 2008. The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, 45(4):602–607.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Marc Brysbaert and Boris New. 2009. Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods*, 41(4):977–990.

Erik Cambria and Bebo White. 2014. Jumping nlp curves: A review of natural language processing research [review article]. *IEEE Computational Intelligence Magazine*, 9(2):48–57.

Lushi Chen, Walid Magdy, Heather Whalley, and Maria Klara Wolters. 2020. Examining the role of mood patterns in predicting self-reported depressive symptoms. In *Proceedings of the 12th ACM Conference on Web Science*, pages 164–173.

Siyuan Chen, Julien Epps, Natalie Ruiz, and Fang Chen. 2011. Eye activity as a measure of human mental effort in hci. In *Proceedings of the 16th international conference on Intelligent user interfaces*, pages 315–318.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, et al. 2024. Tombench: Benchmarking theory of mind in large language models. *arXiv preprint arXiv:2402.15052*.

Ta-Chung Chi, Ting-Han Fan, Alexander I Rudnicky, and Peter J Ramadge. 2023. Transformer working memory enables regular language reasoning and natural language length extrapolation. *arXiv preprint arXiv:2305.03796*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Kathy Conklin and Ana Pellicer-Sánchez. 2016. Using eye-tracking in applied linguistics and second language research. *Second Language Research*, 32(3):453–467.

Uschi Cop, Nicolas Dirix, Denis Drieghe, and Wouter Duyck. 2017. Presenting geco: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior research methods*, 49:602–615.

Thomas M Cover and Joy A Thomas. 1991. Information theory and statistics. *Elements of information theory*, 1(1):279–335.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the international AAAI conference on web and social media*, volume 7, pages 128–137.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Oliver Eberle, Stephanie Brandl, Jonas Pilot, and Anders Søgaard. 2022. Do transformer models show similar attention patterns to task-specific human gaze? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4295–4309.

Susan F Ehrlich and Keith Rayner. 1981. Contextual effects on word perception and eye movements during reading. *Journal of verbal learning and verbal behavior*, 20(6):641–655.

Meir Friedenberg, Hadi Amiri, Hal Daumé III, and Philip Resnik. 2016. The umd clpsych 2016 shared task system: text representation for predicting triage of forum posts about mental health. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 158–161.

Isaac R Galatzer-Levy, David Munday, Jed McGiffin, Xin Liu, Danny Karmon, Ilia Labzovsky, Rivka Moroshko, Amir Zait, and Daniel McDuff. 2024. The

cognitive capabilities of generative ai: A comparative analysis with human benchmarks. *arXiv preprint arXiv:2410.07391*.

Manas Gaur, Vamsi Aribandi, Amanuel Alambo, Ugur Kursuncu, Krishnaprasad Thirunarayan, Jonathan Beich, Jyotishman Pathak, and Amit Sheth. 2021. Characterization of time-variant and time-invariant assessment of suicidality on reddit using c-ssrs. *PloS one*, 16(5):e0250448.

Satyam Goyal and Soham Dan. 2025. Iolbench: Benchmarking llms on linguistic reasoning. *arXiv preprint arXiv:2501.04249*.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610.

Kimia Hemmatirad, Hojjat Bagherzadeh, Ehsan Fazl-Ersi, and Abedin Vahedian. 2020. Detection of mental illness risk on social media through multi-level svms. In *2020 8th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS)*, pages 116–120. IEEE.

Keith C Hendy, Kevin M Hamilton, and Lois N Landry. 1993. Measuring subjective workload: when is one scale better than many? *Human Factors*, 35(4):579–601.

Eckhard H Hess and James M Polt. 1960. Pupil size as related to interest value of visual stimuli. *Science*, 132(3423):349–350.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Nora Hollenstein. 2021. *Leveraging Cognitive Processing Signals for Natural Language Understanding*. Ph.D. thesis, ETH Zurich.

Nora Hollenstein, Maria Barrett, Marius Troendle, Francesco Bigiolli, Nicolas Langer, and Ce Zhang. 2019a. Advancing nlp with cognitive language processing signals. *arXiv preprint*.

Nora Hollenstein, Maria Barrett, Marius Troendle, Francesco Bigiolli, Nicolas Langer, and Ce Zhang. 2019b. Advancing nlp with cognitive language processing signals. *arXiv preprint arXiv:1904.02682*.

Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. 2018. Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading. *Scientific data*, 5(1):1–13.

Phu Mon Htut, Jason Phang, Shikha Bordia, and Samuel R Bowman. 2019. Do attention heads in bert track syntactic dependencies? *arXiv preprint arXiv:1911.12246*.

Zhen Huang, Zengzhi Wang, Shijie Xia, Xuefeng Li, Haoyang Zou, Ruijie Xu, Run-Ze Fan, Lyumanshan Ye, Ethan Chern, Yixin Ye, et al. 2024. Olympicarena: Benchmarking multi-discipline cognitive reasoning for superintelligent ai. *Advances in Neural Information Processing Systems*, 37:19209–19253.

Thomas Huber and Christina Niklaus. 2025. Llms meet bloom's taxonomy: A cognitive view on large language model evaluations. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5211–5246.

Jukka Hyönä, Jorma Tommola, and Anna-Mari Alaja. 1995. Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks. *The Quarterly Journal of Experimental Psychology*, 48(3):598–612.

Md Rafiqul Islam, Muhammad Ashad Kabir, Ashir Ahmed, Abu Raihan M Kamal, Hua Wang, and Anwaar Ulhaq. 2018. Depression detection from social network data using machine learning techniques. *Health information science and systems*, 6:1–12.

Zunaira Jamil. 2017. *Monitoring tweets for depression to detect at-risk users*. Ph.D. thesis, Université d'Ottawa/University of Ottawa.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. MentalBERT: Publicly available pretrained language models for mental healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190, Marseille, France. European Language Resources Association.

Aditya Joshi, Abhijit Mishra, Nivvedan Senthamilselvan, and Pushpak Bhattacharyya. 2014. Measuring sentiment annotation complexity of text. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 36–41.

Josip Jukić, Iva Vukojević, and Jan Snajder. 2022. You are what you talk about: Inducing evaluative topics for personality analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3986–3999, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Marcel A Just and Patricia A Carpenter. 1980. A theory of reading: from eye fixations to comprehension. *Psychological review*, 87(4):329.

Jikun Kang, Romain Laroche, Xingdi Yuan, Adam Trischler, Xue Liu, and Jie Fu. 2023. Think before you act: Decision transformers with working memory. *arXiv preprint arXiv:2305.16338*.

Alan Kennedy, Robin Hill, and Joël Pynte. 2003. The dundee corpus. In *Proceedings of the 12th European conference on eye movement*.

Elma Kerz, Yu Qiao, Sourabh Zanwar, and Daniel Wiechmann. 2022. Pushing on personality detection from verbal behavior: A transformer meets text contours of psycholinguistic features. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 182–194, Dublin, Ireland. Association for Computational Linguistics.

Emmanuel Keuleers, Marc Brysbaert, and Boris New. 2010. Subtlex-nl: A new measure for dutch word frequency based on film subtitles. *Behavior research methods*, 42(3):643–650.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2022. Large language models with controllable working memory. *arXiv preprint arXiv:2211.05110*.

Wutao Lin, Donghong Ji, and Yanan Lu. 2017. Disorder recognition in clinical texts using multi-label structured svm. *BMC bioinformatics*, 18(1):1–11.

Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Steven G Luke and Kiel Christianson. 2018. The provo corpus: A large eye-tracking corpus with predictability norms. *Behavior research methods*, 50:826–833.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2023. Dissociating language and thought in large language models: a cognitive perspective. *arXiv preprint arXiv:2301.06627*.

Christopher D Manning and Hinrich Schutze. 1999. Foundations of statistical natural language processing. *MIT press*.

Susana Martinez-Conde, Stephen L Macknik, and David H Hubel. 2004. The role of fixational eye movements in visual perception. *Nature reviews neuroscience*, 5(3):229–240.

Sandeep Mathias, Diptesh Kanojia, Abhijit Mishra, and Pushpak Bhattacharyya. 2021. A survey on using gaze behaviour for natural language processing. *arXiv preprint arXiv:2112.15471*.

Ethel Matin. 1974. Saccadic suppression: a review and an analysis. *Psychological bulletin*, 81(12):899.

George W McConkie, Paul W Kerr, Michael D Reddix, and David Zola. 1988. Eye movement control during reading: I. the location of initial eye fixations on words. *Vision research*, 28(10):1107–1118.

Milica Milosavljevic and Moran Cerf. 2008. First attention then intention: Insights from computational neuroscience of vision. *International Journal of advertising*, 27(3):381–398.

Abhijit Mishra, Pushpak Bhattacharyya, and Michael Carl. 2013. Automatically predicting sentence translation difficulty. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 346–351.

Abhijit Mishra, Aditya Joshi, and Pushpak Bhattacharyya. 2014. A cognitive study of subjectivity extraction in sentiment annotation. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 142–146.

Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. 2015. Quantifying the language of schizophrenia in social media. In *Proceedings of the 2nd workshop on Computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 11–20.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. 2018. Deep learning for depression detection of twitter users. In *Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic*, pages 88–97.

Keiron O'Shea and Ryan Nash. 2015. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*.

Daniel W. Otter, Julian R. Medina, and Jugal K. Kalita. 2021. A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):604–624.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al.

2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Fred Paas, Juhani E Tuovinen, Huib Tabbers, and Pascal WM Van Gerven. 2003. Cognitive load measurement as a means to advance cognitive load theory. *Educational psychologist*, 38(1):63–71.

Fred GWC Paas, Jeroen JG Van Merriënboer, and Jos J Adam. 1994. Measurement of cognitive load in instructional research. *Perceptual and motor skills*, 79(1):419–430.

Jiabao Pan, Yan Zhang, Chen Zhang, Zuozhu Liu, Hongwei Wang, and Haizhou Li. 2024. Dynathink: Fast or slow? a dynamic decision-making framework for large language models. *arXiv preprint arXiv:2407.01009*.

Sangjun Park and JinYeong Bak. 2023. Memoria: resolving fateful forgetting problem through human-inspired memory architecture. *arXiv preprint arXiv:2310.03052*.

Shuwen Qiu, Mingdian Liu, Hengli Li, Song-chun Zhu, and Zilong Zheng. 2024. Minddial: Enhancing conversational agents with theory-of-mind for common ground alignment and negotiation. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 746–759.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Waleed Ragheb, Jérôme Azé, Sandra Bringay, and Maximilien Servajean. 2019. Attentive multi-stage learning for early risk detection of signs of anorexia and self-harm on social media. In *CLEF 2019-Conference and Labs of the Evaluation Forum*, volume 2380.

Guozheng Rao, Yue Zhang, Li Zhang, Qing Cong, and Zhiyong Feng. 2020. Mgl-cnn: a hierarchical posts representations model for identifying depressed individuals in online forums. *IEEE Access*, 8:32395–32403.

Keith Rayner. 1979. Eye guidance in reading: Fixation locations within words. *Perception*, 8(1):21–30.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372.

Andrew G Reece, Andrew J Reagan, Katharina LM Lix, Peter Sheridan Dodds, Christopher M Danforth, and Ellen J Langer. 2017. Forecasting the onset and course of mental illness with twitter data. *Scientific reports*, 7(1):13006.

Erik D Reichle, Alexander Pollatsek, Donald L Fisher, and Keith Rayner. 1998. Toward a model of eye movement control in reading. *Psychological review*, 105(1):125.

Erik D Reichle, Keith Rayner, and Alexander Pollatsek. 2003. The ez reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and brain sciences*, 26(4):445–476.

William B Rouse and Nancy M Morris. 1986. On looking into the black box: Prospects and limits in the search for mental models. *Psychological bulletin*, 100(3):349.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Anu Shrestha and Francesca Spezzano. 2019. Detecting depressed users in online forums. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 945–951.

Taetem Simms, Clayton Ramstedt, Megan Rich, Michael Richards, Tony Martinez, and Christophe Giraud-Carrier. 2017. Detecting cognitive distortions through machine learning text analytics. In *2017 IEEE international conference on healthcare informatics (ICHI)*, pages 508–512. IEEE.

Priyanka Sinha, Lipika Dey, Pabitra Mitra, and Anupam Basu. 2015. Mining HEXACO personality traits from enterprise social media. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 140–147, Lisboa, Portugal. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Ekta Sood, Simon Tannert, Diego Frassinelli, Andreas Bulling, and Ngoc Thang Vu. 2020a. Interpreting attention models with human visual attention in machine reading comprehension. *arXiv preprint arXiv:2010.06396*.

Ekta Sood, Simon Tannert, Philipp Müller, and Andreas Bulling. 2020b. Improving natural language processing tasks with human gaze-guided neural attention.

Advances in Neural Information Processing Systems, 33:6327–6341.

John Sweller. 1994. Cognitive load theory, learning difficulty, and instructional design. *Learning and instruction*, 4(4):295–312.

John Sweller, Jeroen JG Van Merrienboer, and Fred GWC Paas. 1998. Cognitive architecture and instructional design. *Educational psychology review*, 10(3):251–296.

Marcel Trotzek, Sven Koitka, and Christoph M Friedrich. 2018. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Transactions on Knowledge and Data Engineering*, 32(3):588–601.

Ana-Sabina Uban, Berta Chulvi, and Paolo Rosso. 2021. An emotion and cognitive based analysis of mental health disorders from social media data. *Future Generation Computer Systems*, 124:480–494.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Siyuan Wang, Zhongyu Wei, Yejin Choi, and Xiang Ren. 2024. Symbolic working memory enhances language models for complex rule application. *arXiv preprint arXiv:2408.13654*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Akkapon Wongkoblap, Miguel A Vadillo, and Vasa Curcin. 2019. Predicting social network users with depression from simulated temporal data. In *IEEE EUROCON 2019-18th International Conference on Smart Technologies*, pages 1–6. IEEE.

Jincenzi Wu, Zhuang Chen, Jiawen Deng, Sahand Sabour, Helen Meng, and Minlie Huang. 2023. Coke: A cognitive knowledge graph for machine theory of mind. *arXiv preprint arXiv:2305.05390*.

Duo Yang and Nora Hollenstein. 2023. Plm-as: Pretrained language models augmented with scanpaths for sentiment classification. In *Proceedings of the Northern Lights Deep Learning Workshop*, volume 4.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, and Sophia Ananiadou. 2023a. Mentalllama: Interpretable mental health analysis on social media with large language models. *arXiv preprint arXiv:2309.13567*.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Sophia Ananiadou, and Jimin Huang. 2023b. Mentallama: Interpretable mental health analysis on social media with large language models. *arXiv e-prints*, pages arXiv–2309.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Johannes Zagermann, Ulrike Pfeil, and Harald Reiterer. 2016. Measuring cognitive load using eye tracking technology in visual computing. In *Proceedings of the sixth workshop on beyond time and errors on novel evaluation methods for visualization*, pages 78–85.

Alexandre Zénon. 2019. Eye pupil signals information gain. *Proceedings of the Royal Society B*, 286(1911):20191593.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Pei Zhou, Jay Pujara, Xiang Ren, Xinyun Chen, Heng-Tze Cheng, Quoc V Le, Ed H Chi, Denny Zhou, Swaroop Mishra, and Huaixiu Steven Zheng. 2024a. Self-discover: Large language models self-compose reasoning structures. *arXiv preprint arXiv:2402.03620*.

Yujia Zhou, Zheng Liu, Jiajie Jin, Jian-Yun Nie, and Zhicheng Dou. 2024b. Metacognitive retrieval-augmented large language models. In *Proceedings of the ACM Web Conference 2024*, pages 1453–1463.