UNIVERSITY OF CAMBRIDGE

Multimodal Tuning with Human Cognition Data as Prompts

Yingjia Wan

Supervisor: Dr. Ivan Vulić

Department of Theoretical and Applied Linguistics
University of Cambridge

This dissertation is submitted for the degree of
Master of Philosophy

Murray Edwards College June 2023 (Word Count: 17271)

Abstract

Human cognitive signals reflect humans' attention distribution and neural activation regarding different parts of the input, which are crucial in understanding the mechanism behind their language processing behaviour. Computational linguistics research aims to optimize language models to achieve a human-like level of performance in natural language processing tasks, ideally in an accountable fashion. This renders integrating human cognitive signals into language models an intriguing research area to optimize their downstream task performances in an accountable fashion. Previous works exploring how cognition data could enhance natural language processing (NLP) tasks bore limitations such as weak accuracy increase, heavy engineering bias, and limited generalizability of conclusions drawn from experiments on outdated models. This thesis addresses these issues by introducing a novel approach that leverages prompt-based fine-tuning. In particular, two methods were proposed: (1) inspired by 'hard prompting', Method 1 uses gaze and electroencephalography (EEG) features as discrete prompt tokens to modify model behaviour during training; (2) drawing on 'soft prompting', Method 2 designs a multi-modal prompting framework called 'CogMAP' (**Cog**nition **M**apping **A**nd **P**rompting), which employs these cognition features as multidimensional prompting vectors projected into the continuous embedding space of language models. Task results on ternary sentiment classification were consistently superior when incorporating either gaze or EEG data as prompts in both methods $p < 0.001$, across encoder-only BERT-based models and decoder-only GPT-2-based models. This study signifies a leap in cognition-inspired NLP research, addressing existing limitations while providing a new robust and effective paradigm for future investigations of bridging the gap between human cognition and artificial language processing to improve the performance *and* understanding of language models.

Table of contents

1

Introduction

While large language models have achieved impressive growth of capability in a wide range of tasks, they have not yet surpassed humen across numerous downstream language processing tasks (Fitzsimmons et al., 2014; He et al., 2021). Specifically, natural language processing (NLP) research faces challenges in understanding context, task-generic performance, and overfitting to the bias presented in the *limited* training data. Therefore, it remains an active area of research and a challenge for the field to address these limitations, exploring directions such as incorporating external knowledge, leveraging multi-modal data, etc. (Tamkin et al., 2021), in order to close the performance gap between language models and human performance across various language processing tasks.

Cognition signals can reflect human processing efforts and cognitive load on each word, and could therefore provide extra information on the contextual dependency and predictability of a word among its neighbours. Based on the background, cognition signals provided by human readers, a special category of multi-modal data outside speech, vision or text, are promising in enhancing language models in an accountable approach, by providing insight into how humans process natural language input, as proposed by a field of **cognition-inspired research** (also interchangeably called cognition-enhanced/ motivated/integrated research)(Barrett et al., 2016; Ding et al., 2022; Mishra et al., 2017; Ren and Xiong, 2021; Sood et al., 2020b). Such human-oriented cognition data range from eye-tracking signals, neurophysiological measurements including electroencephalography (EEG), functional magnetic resonance imaging (fMRI) data and heart rate, to user-generated feedback on a macro level. Among them, eye-tracking data (also called gaze data) captures overt attention, i.e., where and how long individuals focus their gaze while reading texts (Rayner, 1998), while EEG captures brain activity related to text processing via voltage fluctuations detected on the scalp, which includes information about cognitive processes such as both covert attention (i.e., processing efforts assigned to a point not manifested on eye movements) and overt attention, memory encoding and retrieval, comprehension, and emotional responses (Shoka et al., 2019). Due to the high temporal resolution nature of

the two technologies, they can capture fine-grained word-level cognition features that are suitable to be leveraged by language models.

Why can cognition signals potentially facilitate the task-specific performance of language models in training? Prior related works lacked a systematic theoretical contextualization of the motivation underlying their empirical exploration of the topic (Hollenstein et al., 2020a, 2019, 2018a; Sood et al., 2020b). Only Ding et al. (2022) self-defined a 'cognitive theory' from language acquisition research (Scarborough et al., 2009) which is not an accurate representation of the actual framework. Therefore, this thesis trace back to several fundamental psycholinguistics theories to systematically established the rationale of cognition-inspired research, including embodied cognition, selective attention theory, etc. (elaborated in Related Works).

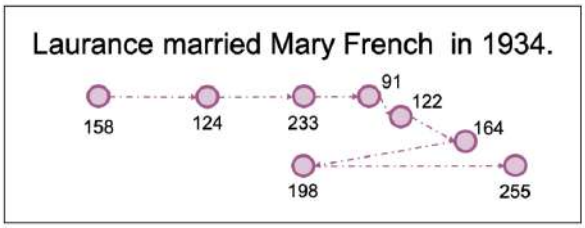


Fig. 1.1 Scanpath visualizaton of gaze points measured by eye-tracking technology on a sentence. The values aside each point denote the fixation duration on the gaze points in milliseconds. A fixation is the period of time where the gaze of a reader is maintained on a single location. In this example, for the word 'Mary' and 'French', number of fixations = 2 while for the rest of the words number of fixations = 1.

In brief, cognition signals can provide additional information about the corresponding text sequence on both token level and sequence level, which can potentially facilitate for language models across task performances. Such information can be categorized as three aspects: (1) syntactic and semantic information about different tokens (Barrett and Hollenstein, 2020; Barrett and Søgaard, 2015; Clifton Jr et al., 2007); (2) attention distribution which indicates the relative importance of the words and their contextual interdependency for effcitient processing and prediction (Duggan and Payne, 2011); (3) in the case of sentiment classification in this study, eye movements and neuron activation detected by EEG can further reveal processing patterns exclusively related to emotional response.

It is therefore tempting to explore whether we can exploit the information provided by human cognitive signals to improve the model performances and generalize the ability over tasks.

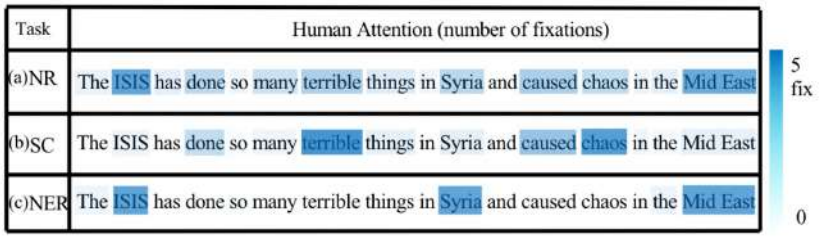


Fig. 1.2 Heatmap visualizaton of number of fixations, one of the variables in eye-tracking measures, in different task paradigms for participants. NR (Normal Reading) refers to human reading without tasks.

Combining with the above-mentioned motivation, the increasing affordability of eye-tracking and neuroimaging technology has also contributed to the feasibility of cognition-inspired NLP: in recent years there have flourished multiple eye-tracking and EEG corpora that collected data from human participants reading samples from well-established NLP task-specific benchmark datasets as text materials (Cop et al., 2017; Hollenstein et al., 2020b; Sui et al., 2022). Such cognition corpora were mostly designed for NLP research and composed of two parts: (1) textual data with (e.g., ZuCo (Hollenstein et al., 2018b)) or without (e.g., GECO (Cop et al., 2017)) NLP task annotations, and (2) word-level human cognition signals measured simultaneously in a normal or task-specific reading paradigm. Hence, this line of research can utilize the numerous features in eye-tracking and EEG signals like number of fixations, fixation durations, voltage frequency of each electronode, etc. to establish a multi-dimensional feature vector for a word, where each dimension contains the information about a certain variable.

Recent cognitively-inspired NLP research adopted two major approaches to incorporating cognition signals into language models, firstly by augmenting word embeddings directly with cognition features as additional embeddings in non-pre-trained neural networks (Hollenstein et al., 2019) e.g., Bi-LSTMs (Hochreiter and Schmidhuber, 1997), secondly by regularizing the attention weights in the model as an auxiliary training task (McGuire and Tomuro, 2021). However, their approach of direct finetuning on the text material with weighted sampling from cognition signals induces not so striking performance boost (Hollenstein et al., 2020c; McGuire and Tomuro, 2021).

Some scholars, therefore, tried to strengthen the positive impact of cognition features by adding preprocessing steps before training, e.g., extract self-categorized linguistic features filtered by word-level cognition signal values (Ding et al., 2022), or de-noising the cognition signals by adding modality alignment algorithms before training the LM's attention (Ren and Xiong, 2021). In these works, multiple neural networks (mostly Bi-LSTMs) were utilized for various intermediate purposes to achieve the final training results.

As a result, such works suffered from heavily biased feature engineering, implementation complexity due to multiple training tasks and complex architecture design, and as a result, low generasability to more task paradigms and the larger state-of-the-art language models.

Considering both the weak enhancement results and the methodological deficits in previous cognition-inspired NLP research, the current study aims to achieve a much more effective yet straightforward approach to integrating cognition signals into state-of-the-art pre-trained language models including both BERT-based encoder-only and GPT-2-based decoder-only models, which remains a blank territory at the current stage.

Current Study This thesis proposes a more robust integration framework by treating human cognition recorded by EEG and eye-tracking as a new modality of data, and introduced prompt-based fine-tuning from prompting research, an increasingly mainstream model deployment paradigm in recent years, to cognition-inspired research for the first time. In particular, this thesis has introduced two methodologies for incorporating cognitive data as prompts:

Method 1, inspired by the 'hard prompting' method that employs discrete tokens as prompts, flattens each word-level gaze feature and synthesized EEG feature to a single dimension. These are treated as unique tokens, and collectively form a sequence of special tokens. Each sequence represents the cognitive processing information for a word that has been tokenized in the cognitive corpora.

Method 2, taking inspiration from the continuous 'soft' prompts in the form of input vectors, treats gaze and EEG features for each sequence as sequences of multidimensional vectors. Every feature in the gaze data and EEG data contributes to one of the dimensions. These vectors, along with the corresponding text input, are imported to a new multimodal framework known as 'CogMAP' (**Cog**nition **M**apping **A**nd **P**rompting) designed by this study. The CogMAP framework maps the dimension of the cognitive vectors to match the input dimension of the backbone language model, inserting them before each corresponding sentence in the input as soft prompts. These prompts are then trained for sentiment classification tasks. Given the task nature and the backbone models to be elaborated on later, the parameters of both the language model and the projection layer are updated during training.

Overall, this thesis offered exclusive contributions to renewing the methodological paradigms of cognition-inspired research by providing the following findings:

1. The prompt-based finetuning approach is proved feasible and effective to be adapted to cognition-enhanced NLP research by integrating cognition data as multimodal prompt input, achieving an impressive performance boost in ternary sentiment classification without the necessity of manually placing supervision on each model's

attention layers. This approach can be further adapted to two implementation methods:

   (a) Cognition features can be introduced as effective prompts to state-of-the-art pre-trained models by concatenating text input with numerical cognition signals either as discrete special tokens (Method 1).

   (b) Cognition features can be treated as multidimensional vectors and projected as continuous prompts in the pre-trained LM's embedding space (Method 2).

2. This study is conducted on a wide range of advanced state-of-the-art language models, including auto-regressive decoder-only and bidirectional encoder-only pre-trained attention-based models, in order to expand the ablation validity and generalizability of the methodology. This is the first time cognition-enhanced research achieved effective enhancement on decoder-only pre-trained models (as previous works mostly focused on simple neural networks, transformers, or BERT as backbone models).

3. Both EEG and gaze (i.e., eye-tracking) signals can provide *unique*, positive guidance to the language models' classification performance, although the degree of this influence can vary. Future research is encouraged in exploring the fine-grained preprocessing methods in denoising EEG data to optimise the leveraging of cognition signals data on advanced language models.

4. The new CogMAP framework also shed inspiration on multi-modal NLP research about grounding language models to other modalities, more precisely, effective multi-modal data incorporating strategies for language models in small-data-size settings.

Significance The significance of the current study unfolds across four key dimensions. First, it addresses the methodological limitations and weaknesses of previous research in utilizing cognitive data to advance NLP research. These include heavy engineering bias in processing cognitive signals, reliance on outdated backbone models, and a lack of generalisability across diverse tasks and model types.

Second, this research pioneers the exploration of cognition-prompting. By treating cognitive data as prompt input that is concatenated with corresponding texts, the study exploits the long-range contextual learning ability of state-of-the-art pre-trained language models. This innovative approach allows the model to capture the meta-cognitive associations between the text and cognitive prompts, pushing the boundaries of our understanding in this area and achieving a leap of performance enhancement compared to previous works (Barrett et al., 2016; Hollenstein et al., 2019; McGuire and Tomuro, 2021).

Third, the study delves into the detailed investigation of the soft-cognition-prompting method within the continuous embedding space of a language model. This provides a platform for a range of future research opportunities, including the application of cognition-prompted language models for instruction tuning, modular tuning, and the integration of numerical reasoning with cognitive-inspired research.

In essence, this study not only addresses existing limitations but also paves the way for groundbreaking advances in the interdisciplinary field of cognition-inspired NLP research. By harnessing the power of cognitive data and state-of-the-art models, this research offers exciting potential for enhancing the performance and understanding of language models.

2

Related Works

2.1 Human Cognition Signals in NLP

The human sentence processing behaviour can be measured via technologies such as eye-tracking, electroencephalogram (EEG), functional magnetic resonance imaging (fMRI), etc., which reveal the reader's varying cognitive processing efforts on each token during different stages of reading processing. In addition, psycholinguistic experiments have also demonstrated that such cognitive processing signals can effectively reflect the relevant textual information of a token in reading comprehension that encompasses word-level features in syntax, semantics, pragmatics, logic, etc. (Rayner et al., 2004; Weiss and Mueller, 2003). This roughly established the feasibility and justification of cognitively-enhanced NLP research. In particular, eye-tracking and EEG data have been popular options to integrate into language models due to their high temporal resolutions, capturing real-time bodily reactions to each perceptual span (i.e., the visual contextual window focusing on the target word).

The section will introduce eye-tracking and EEG in detail as the most representative of cognitive signals, then explain the mechanisms which support them to improve the sentiment classification performance of NLP models separately, and finally describe ZuCo (Hollenstein et al., 2018b), the corpus used in the current study, clarifying its provided cognition features to be incorporated in the current experiments.

2.1.1 Eye-Tracking

What is it?

Among cognitive processing signals, eye-tracking data (also interchangeably called gaze data) provides online measurements with millisecond accuracy on the fixation, saccades, and word skipping behaviour of a reader's eye movements in a close-to-natural reading

experimental set-up (Rayner, 1977). These signals capture the eye movements and fixations that occur when an individual is engaged in a task, such as reading or image processing. The tracking of these ocular activities enables researchers to gauge the points of focus, the sequence of focus shifts, and the time spent on each focus point (Rayner, 1998).

Based on the **embodied cognition theory** which fundamentally posits that cognitive processing can be reflected by physiological and sensor-motor activities such as eye movements, eye-tracking signals can provide unique insights into overt attention, i.e., the amount of attention or relative importance a human reader assigned to certain words in a sequence, as reflected from eye movements. For instance, longer fixation durations often correlate with increased cognitive processing. Conversely, rapid saccades (quick eye movements) between words might indicate a smoother reading flow. This makes these signals a potent tool for studying cognitive load, attention distribution, and processing patterns in psycholinguistics research (Degen et al., 2021; Kaiser, 2013).

How does it facilitate NLP models?

The mechanisms that the incorporation of cognition signals into language models can enhance the processing of NLP models can be divided into three levels: (1) indicating selective attention distribution to the task-related parts of a sequence, (2) conveying contextual linguistic information that in general enhanced the in-context learning ability and semantic embedding space of the language model, and (3) the analogical correlation between the human cognitive attention and the model attention in ubiquitous transformer-based LMs.

Firstly, according to the **selective attention theory** from cognitive psychology (Johnston and Dark, 1986), due to the limited capacity of working memory, humans selectively pay attention to the most 'important' information when processing a text (i.e., the partial information that contributes most to the correct understanding of the texts or to the optimal completion of the language processing tasks). Therefore, by indicating processing efforts and attention distribution, the eye movement features ultimately reflect the hidden semantic characteristics of each word in a sentence that are closely associated with the reading objective. As shown in the case of a number of fixations in Fig1.2, such word-level feature values are highly contingent on the variation of task paradigms and whether the semantics of the targeted word is associated with the task objective. Even when reading the same sentence, there are considerably more fixations on sentiment-related and attitude-related words 'terrible' and 'chaos' than the others in the task-specific reading paradigms for (b) sentiment classification. In contrast, the fixation focus shifted towards named entities in the (c) NER paradigm.

Apart from the direct association between the fixation results indicating selective attention distribution and the task label, the eye movement data can also indicate complex contextual linguistic factors of the words in the sequence (Clifton Jr et al., 2007; Demberg and Keller, 2008). As a result, such linguistic information reflected in the gaze data can provide additional contextual cues to the language model outside the semantic information provided by the text input.

As established by psycholinguistics research, these linguistic features include morphological and syntactic complexity features at early or later stages of language processing that can be studied separately. In terms of morphological features, for instance, the word frequency effect (Rayner, 1977) posits that word length, frequency and predictability from context affect fixation duration and counts. As another well-known example, readers are more likely to fixate on open-class words (Carpenter and Just, 1983); In terms of syntactic features for later stages of sentence processing, a prime illustration is from Barrett and Søgaard (2015) who found that most syntactic categories can be reliably predicted from eye movements of readers, represented by 10 derived variables such as fixation probability on the target word, whether the previous word is fixated, whether next word is fixated, number of fixations, first fixation duration, and so on. This study supplemented fixation information about the preceding and subsequent words due to the consideration of the spillover effect and preview effect in eye movements, which was considered as a guideline for more efficient utilization and aggregation of word-level gaze features in subsequent cognition-inspired NLP studies (Barrett and Hollenstein, 2020). Since then, Barrett et al. (2016); Hollenstein et al. (2019); Hollenstein and Zhang (2019) adapted such derived categories of gaze features as integration sources for early models like CNN, Bi-LSTMM, etc. Along with theoretical connections, these studies served as empirical evidence that justifies the motivation of developing cognition-enhanced NLP research (which will be elaborated later in Section 2.2).

Lastly, gaze data can potentially help regulate the attention weights or the gradient-based saliency of the prediction output of language models, based on the approximation between them. Attention NLP-related neuroscience studies have conducted comparative analogies between the LM attention mechanism and the human attention extracted from the cognition corpora. For instance, Hao et al. (2021) fitted the fixation durations of each word from eye-tracking data by a mixed-effect analysis model to the attention weights in LSTMs (Hochreiter and Schmidhuber, 1997) and transformers (Vaswani et al., 2017), concluding that such correlation is robust and independent of model perplexity across different neural networks. Sood et al. (2020a) as well as Bensemann et al. (2022) delved deeper into pretrained language models and found that the association strength differ in attention layers (Bahdanau et al., 2014): earlier layers (esp. first layer) of BERT (Devlin et al., 2018) shows the highest correlations consistently, although notable exceptions occurred in multilingual

encoders (Hollenstein et al., 2021). Hollenstein and Beinborn (2021); Ikhwantri et al. (2023)

The approximation between gaze features shaped the possibility to explore whether gaze features can optimize when integrated into language models. McGuire and Tomuro (2021) examined an increased similarity between human attention represented by fixation durations and EEG features after using them to supervise BERT's attention layers in relation classification task-specific fine-tuning, which explains why the cognition-integrated model can predict more correctly where the baseline failed. Ding et al. (2022), using a different training approach in integrating cognition data, also drew a similar conclusion in attention similarity analysis.

Therefore, based on the three aspects of potential facilitation, in the context of natural language processing (NLP), by integrating these eye-tracking signals with NLP tasks, cognition-inspired research aims to build models that capture and leverage these reading patterns to improve their performance.

2.1.2 EEG

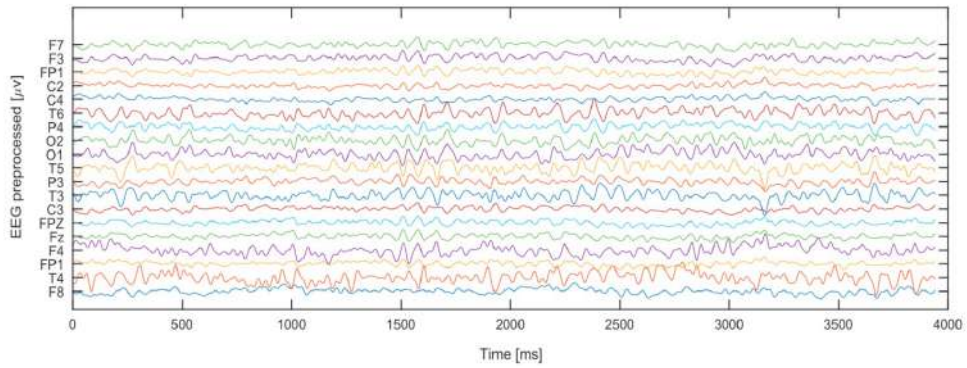


Fig. 2.1 A subset visualization of the preprocessed EEG data during the sentence, from ZuCo (Hollenstein et al., 2018b). Electrodes matching the 10–20 systems (F7-F8) were chosen for feature demonstration; for plotting purposes, data were bandpass-filtered (0.5–30 Hz). Such fluctuation information within a time frame is extracted and mapped to the fixation time window of words shown from eye-tracking data, rendering word-level EEG features shown in Table 2.1.

What is it?

Electroencephalography (EEG) measures potential fluctuations caused by the activity of neurons in the cerebral cortex, as demonstrated in Fig 2.1. The recorded signals reflect the

collective activity of neurons close to the electrode, enabling the measurement of real-time brain activity across different regions. Hence, EEG can exclusively capture the processing efforts over a sequence beyond eye movements (i.e., covert attention).

Within the NLP field, EEG signals can be used to understand the neurological underpinnings of language comprehension and production. For example, certain patterns of brain activity can correspond to the processing of semantic or syntactic information in a sentence. By associating these patterns with NLP tasks, we can create more cognitively congruent models that mirror human-like language processing (Hollenstein et al., 2020b; Shoka et al., 2019). Furthermore, EEG signals can help detect cognitive load or engagement level, which could be used to adjust the complexity level of generated text or to customize learning materials dynamically.

How does it facilitate NLP models?

Studies exploring EEG data for NLP development are more limited, the first of which was in 2019 (Hollenstein et al., 2019), since its corpus collection is more expensive and time-consuming. However, because EEG and eye-tracking use the same temporal resolution with non-invasive technologies (Sereno and Rayner, 2003), the three aspects of the mechanisms for eye-tracking features supporting NLP models in downstream tasks should also apply to EEG.

EEG data can also reveal the underlying linguistic properties of the text input. Dambacher and Kliegl (2007) found that longer fixation duration correlates with larger N400 amplitude effects. N400 is part of the normal brain response to words and other meaningful stimuli (Kutas and Federmeier, 2000). Effects of word predictability on eye movements and EEG co-registration have also been studied in serialized word representation and in natural reading (Dimigen et al., 2011). In terms of syntactic categories, EEG also can reflect differences in processing verbs and nouns, concrete nouns and abstract nouns, as well as common nouns and proper nouns (Weiss and Mueller, 2003).

2.1.3 Cognition Corpora for NLP Research

The primary dataset utilized in this study is the Zurich Cognitive Language Processing Corpus (ZuCo), as detailed by Hollenstein et al. (2018b). As the first openly accessible corpus featuring concurrent eye-tracking and EEG recordings during natural sentence reading, the ZuCo corpus encompasses recordings from 12 adult native English speakers, each engaging with roughly two paradigms of 1,100 English sentences. The choice of datasets in this paper is based on the established rigorousness of the ZuCo datasets by empirical literature that supports the dataset as an accurate and representative corpus of eye-tracking and EEG data to denote human processing (Ding et al., 2022; Hollenstein

et al., 2020a, 2018a; McGuire and Tomuro, 2021; Ren and Xiong, 2021). However, the methodology framework in this study can be easily scaled up to more cognition datasets that are growing in recent five years.

For the task of sentiment classification, ZuCo includes a reading paradigm that contains 400 positive, negative and neutral sentences (8138 tokens) from the Stanford Sentiment Treebank (Socher et al., 2013), to analyze the elicitation of emotions and opinions during reading. Gaze data and EEG data are recorded simultaneously on readers in a natural reading setting with the hidden goal of comprehending and identifying the sentiment of the text.

To ensure effective measurements, a control condition was interleaved with the experimental condition where the subjects were to rate the quality of the described movies in 47 of the 400 sentences. The average response accuracy compared to the ground-truth labels of the Stanford Sentiment Treebank is 79.53%.

Gaze features Eye position and pupil size were recorded with an infrared video-based eye tracker (EyeLink 1000 Plus) at a sampling rate of 500 Hz and an instrumental spatial resolution of 0.01°. Five basic eye-tracking features were extracted and provided as word-level variables:

1. Gaze Duration (GD), which accounts for the aggregate of all fixations on the present word during the initial read-through before the eye progresses beyond it;

2. Total Reading Time (TRT), which encapsulates the combined duration of all fixations on the target word, including regressions to it;

3. First Fixation Duration (FFD), representing the duration of the initial fixation on the target word;

4. Go-Past Time (GPT), the cumulative duration of all fixations before moving rightward beyond the current word, including movements to antecedent words that originated from the ongoing word.

5. Number of Fixations (nFix), which refers to how many times the reader has fixated on the target word during reading, including regressions to it.

From a psycholinguistics perspective, each of the five gaze variables should render exclusive yet complementary information about different processing stages and linguistic features regarding the target word in the sentence. GD and FFD provide insights into the initial stages of lexical access and word recognition, shedding light on the complexity of the word or its fit within the sentence's context "to a human's eye"; TRT includes all processing efforts over a word and is hence an indicator of overall cognitive processing

Table 2.1 Example of word-level cognition data for a text sample. The demonstrated sample contains a sequence of 23 words in the example based on the sequence length, and thus 23 arrays of cognition signals. Each array encompasses 5 eye-tracking feature values and 105 EEG feature values.

Text sequence	'Presents a good case while failing to provide a reason for us to care beyond the very basic dictums of human decency.'										
Sequence Label	NEUTRAL										
Word-level cognition data	**word**	**word_index**	**Gaze features**					**EEG electrodes**			
			FFD	**GD**	**GPT**	**TRT**	**nFix**	**F7**	**FP1**	**...(102 omitted columns)...**	**P3**
	Presents	1	120	110	130	600	5	0.12	0.15	...	0.22
	a	2	100	100	120	580	4	0.11	0.14	...	0.20
	...	...(omitted rows)...									
	decency.	23	115	115	125	590	4.5	0.13	0.16	...	0.23

effort for the word; GPT reflects the difficulty in integrating the word into the sentence context for syntactic and semantic processing; nFix is an integer value for the frequency of a reader's gaze returning to the word, indicating the reader's need for re-analysis or confirmation and heavily influencing the GD, TRT, and GPT values. Together, these gaze metrics offer a multifaceted view of a reader's cognitive engagement with a word during sentence reading, reflecting a variety of cognitive processes including lexical access, syntactic parsing, and semantic integration.

EEG features Because EEG signals were recorded simultaneously with eye-tracking, ZuCo provides available word-level EEG signals over 128 electrodes by offline referencing and aligning the temporal stamps of the two signals. The recorded EEG signals consist of voltage fluctuations generated by ionic currents within the neurons of the brain. Preprocessing steps in ZuCo included filtering to remove noise, correcting for eye movement and blinks, segmenting the continuous EEG signal into epochs related to specific events (e.g., reading a word), etc (Hollenstein et al., 2018b). The electrical activity detected by a given electrode reflects the cumulative activity of millions of neurons near that electrode, which are normalized together with eye-tracking data in most cognition-inspired research.

There are altogether 128 EEG electrodes employed in ZuCo datasets. After basic preprocessing and erasing sections unrelated to language processing, there remain 105 useful electrodes, rendering 105 EEG features available for incorporation in this study (Fig 2.1).

It is crucial for the current study to address the term "word-level features" in the context of EEG and eye-tracking data. In the ZuCo corpus, this term refers to the splitting of sequences into separate words, each of which is associated with specific eye-tracking and EEG feature values. These values symbolize cognitive processing effort or neural activity pertinent to each word. However, such word splitting in the original dataset does *not* align with the tokenization configurations of a language model. This is a critical consideration when integrating cognitive data into language models to establish associations, ensuring

that the data aligns accurately with the model's structure and requirements. The distinction between the word splitting and LM tokenization will be revisited in the methodology and results sections to elucidate its potential impacts on training outcomes in some experimental scenarios.

2.2 Empirical Approaches to Incorporating Eye-tracking and EEG Corpora

The majority of cognition-enhanced research focuses on the most common two types of cognition data: eye-tracking and EEG, with the exploration of the latter relatively limited due to its higher cost for collection and lower signal-to-noise ratio which requires expertise to perform preprocessing (Hollenstein et al., 2020a; Ren and Xiong, 2021).

In early works, eye-tracking signals have been used in machine learning approaches to a range of NLP tasks, such as part-of-speech tagging (Barrett et al., 2016), multi-word expression extraction (Rohanian et al., 2017), syntactic category prediction (Barrett and Søgaard, 2015). Later when developing neural models, there are mainly two approaches to incorporating cognition data into language models in the existing research. One is directly augmenting the word embeddings with eye tracking or EEG features as additional layers, while the other is leveraging the cognition signals to supervise neural attention to approximate the online human attention when reading the same texts. The two approaches are broken down in the following subsections.

2.2.1 Augmenting Embedding Layers

The first approach has been popular when the predominant language model to employ for NLP research is a neural network like a transformer. The simple architecture of such neural networks enables the feasibility to practice a direct combination of the word embeddings with the cognition features by augmenting the embedding dimensions. For example, Mishra et al. (2017) augment linguistic features used for sentiment analysis and sarcasm detection (e.g., from WordNet) with eye-tracking features; Hollenstein et al. (2019) extracted the Glove word embeddings to augment the embedding layers with both eye-tracking and EEG features, and trained a Bi-LSTM on named entity recognition, sentiment analysis, and relationship classification. Similar studies were also conducted on other sequence classification tasks including grammatical error detection (Barrett et al., 2018), readability prediction (Gonzalez-Garduno and Søgaard, 2018).

However, this approach inevitably depends on a simple neural network structure as the backbone model, and is therefore fatally restrained along with the development of large prestrained language models as the state-of-the-art models for task-specific deployment.

2.2.2 Cognition Signals as Attention Supervision

Another trend is using cognition signals as the supervision label to selectively enhance or constrain the model attention. For example, McGuire and Tomuro (2021) utilized this strategy in a multitask learning paradigm, training BERT on relation classification while setting an auxiliary training task on supervising BERT's attention weights with the gaze features from ZuCo. However, they failed to observe a significant increase in classification F1 scores.

Later works explored designing modular frameworks in order to better supervise the model's attention (Barrett et al., 2018; Sood et al., 2020b,a; Takmaz et al., 2020). Sood et al. (2020) built a hybrid text saliency model by combining a Bi-LSTM with a transformer. CogAlign (Ren and Xiong, 2021) adopted two separate encoders for two modalities of input and a shared encoder (i.e., three interconnected Bi-LSTMs), together with a special modality discriminator and text-aware attention for alignment. They used adversarial learning to train the model. CogBERT (Ding et al., 2022) paired cognition features with linguistic features (e.g., content words) extracted using the SpaCY tool and filtered out statistically insignificant linguistic features by referencing human attention in the eye-tracking data. Muttenthaler et al. (2020) leverage EEG features to regularize attention on relation extraction.

2.2.3 Limitations of Previous Cognition-Integrated NLP Research

Having reviewed the related works in leveraging eye-tracking and EEG data when training language models, this section summarized several limitations in their methodologies that this study aims to address using the novel prompt-based tuning paradigm.

Heavy human-biased engineering in feature extraction

The majority of NLP studies leveraging human gaze signals from reading use a selective range of manually-defined features for task-specific training rather than the basic word-level features provided by the cognition corpora. However, there lack of consistency and concrete reasons for their practices of signal pre-processing and feature extraction. For example, Barrett et al. (2016) derived 22 gaze features for a part-of-speech tagging task, from the basic fives gaze variables in ZuCo which allegedly encompass both early and late measures of cognitive processing from a psycholinguistic perspective (e.g., re-read

probability), as well as context features calculated from gaze variables of the surrounding words (e.g., w+1 fixation probability, w-1 fixation duration). Hollenstein and Zhang (2019) modified the features to 17 for named entity recognition, but maintained the 5 basic gaze features for the relation classification and sentiment analysis task paradigm. Strzyz et al. (2019) opted for 12 features for dependency parsing. On the other hand, some works obtained a single value out of the five gaze features as a proxy to represent the overall attention to a word: Barrett et al. (2018) used the mean fixation duration by dividing total reading time (TRT) by the number of fixations (nFix), and McGuire and Tomuro (2021) chose TRT itself. As for EEG features, Hollenstein and Zhang (2019); Ren and Xiong (2021); Sood et al. (2020b) grouped the EEG values on all electrodes to 8 common frequency bands; McGuire and Tomuro (2021) obtained a scalar per word by averaging over all EEG electrodes, while Muttenthaler et al. (2020) took the maximum value.

Further, it has been a common practice to amalgamate linguistic features with gaze features, as this combination has been demonstrated to bolster model performance (e.g., Rohanian et al. (2017) and Yaneva et al. (2018)). Barrett et al. (2016) use word frequency and word length features in combination with eye-tracking features, because the two properties explain much of the variance in fixation duration (Just and Carpenter, 1980; Levy, 2008). This practice has been followed by most of the later literature: In Hollenstein and Zhang (2019) and Barrett et al. (2018), word frequency and word length are auxiliary learning tasks in parallel with the gaze or EEG features. Ding et al. (2022); Ren and Xiong (2021, 2022) focused on a more fine-grained framework to model the relationship between cognitive processing signals and linguistic features, by assigning a cognition-informed 'importance scores' to each linguistic feature. In such bridging frameworks, more specific linguistic features were introduced on the levels of lexical, syntax, and semantics and selected based on the criteria of 'interpretability' and 'extensibility' (e.g., lexical density, complex nominals per clause, sentence length, subject number, object number, discourse connector count). Nevertheless, the classification and selection criteria for these features lack strong justification, appearing to be more a product of empirical trial-and-error exploration that are only applicable to the mode-specific and task-specific experimental setting than of theory-based or evidence-based reasoning, thereby lacking a universally accepted reference point for future research.

As shown, all these previous works rely on heavy feature engineering to ensure the desired effect of integrating cognitive information to LMs (Hollenstein et al., 2020a). However, such human-guided feature engineering can bring a number of potential disadvantages. Firstly, the inherent subjectivity of human-defined features leads to a lack of consistency across different studies. This variability not only hinders the generalization of the research but also constrains the direct comparison and benchmarking of different models; Secondly, the process of manual feature engineering is labor-intensive and time-consuming, since

it requires not only extensive psycholinguistic knowledge but also substantial effort into the careful selection and implementation of these features. Last and most importantly, by depending so heavily on manually engineered features, the methods risked imposing arbitrary boundaries on the model's ability to learn and generalize. These boundaries are based on the current understanding of language and cognition in psycho-linguistics, which, while advanced, remains incomplete. This may lead to models that are overly tailored to specific tasks or datasets, lacking the robustness and flexibility required to handle new, unseen data or to adapt to evolving language use.

Outdated and non-scalable Training Methods

Previous cognition-related NLP research leverage fine-tuning to incorporate cognitive signals, by treating them as either extra layers of augmentable word embeddings (e.g., (Hollenstein and Zhang, 2019)) or guiding supervision signals for attention weight supervision as auxiliary tasks in multitask learning paradigms (e.g., (McGuire and Tomuro, 2021; Ren and Xiong, 2021; Sood et al., 2020b); (McGuire and Tomuro, 2021)). In particular, the former approach of directly augmenting dimensions of word embeddings with dimensions of cognitive features is not replicable on pre-trained LMs like BERT, let alone large LMs with exponentially bigger parameters, that constitute multiple layers of neural network, with each layer importing and forwarding a fixed multi-dimension of attention masks and word embeddings. This is because the simple neural networks like Bi-LSTM adopted by Hollenstein et al. (2019) can be trained from scratch with a self-defined dimension of word embeddings as input, whereas in comparison it is impractical and likely detrimental for the model performance to modify the inner structure of a multilayer *pretrained* model to process word embeddings whose dimensions are not the same as the pretraining configuration. Hence, although there have been reported comparative performance growths from integrating cognition data into the model over the baseline, such pioneering work admittedly lost its relevance with the evolving LMs and NLP research.

Another potential weakness of previous works is concerned with the potential hazard of directly fine-tuning the model attention with human cognition signals. The cognition datasets available for model fine-tuning are relatively insufficient with irrelevant or misleading noise. When placed in front of the enormous capacity of current large language models, this creates a problem of over-fitting in situations of data sparsity, and in turn, hurts the generalizability of cognition-integrated research as the trained model would easily be overfitted to the cognition signals provided together with the textual input. What's more, because of the sensitivity to training data in direct model fine-tuning, supervising model attention with the cognition signals exposes the model to the exaggerated impact of signal noise which extensively damaged the model performance even below baselines. Based

on this methodological concern, it could explain the weak or null results in empirical explorations of cognition-inspired NLP studies, e.g., McGuire and Tomuro (2021) failed to find a significant increase of accuracy by supervising BERT's attention with either gaze or EEG data.

In summary, the old-school training techniques imposed greater and greater challenges to adapt these previous training methods to increasingly large LMs. These challenges are further reinforced considering the relatively high cost of collecting cognition corpora using neuroimaging resources. Therefore, the current field of cognition-inspired research calls for more effective and stable-to-noise approaches to incorporating cognition data for performance gains.

2.3 Prompting Technique in Applying Large LMs to Downstream Tasks

Among empirical methods developed for deploying pre-trained large language models for downstream tasks, prompting has been proposed as an alternative to counter the drawbacks of model fine-tuning in response to the increasing size of large language models. It will be introduced and elaborated in this section why prompting can be significantly advantageous as a new method to integrate cognition data into state-of-the-art pretrained language models.

2.3.1 Introduction

Prompting offers another approach to adapt pre-trained language models to specific tasks by optimize the input. In a general sense, prompting refers to prepending carefully crafted prompts to the task input to guide the pre-trained model to produce the desired output. This makes prompting a versatile as well as efficient tool for various tasks, especially for training settings with small datasets on LMs of a large parameter size.

In their survey of comprehensive development of prompting research, Liu et al. (2023) commented that NLP research is witnessing a significant paradigm shift, transitioning from the "pre-train, fine-tune" method to a new approach of "pre-train, prompt, and predict". In this emergent framework, rather than modifying pre-trained LMs to suit downstream tasks via meticulous objective engineering, these tasks are restructured to resemble those addressed during the original LM training through the use of textual prompts. For instance, in the task of sentiment classification on "I missed the bus today", a prompting choice can be such as "I felt so ...", and then task the LM to complete the sentence with an appropriate emotion-laden word that can serve as intermediate output for probability prediction. Thus,

by choosing the right prompts, we can effectively guide the model's behaviour, enabling the pre-trained LM to predict the expected output. For example, Sun and Lai (2020) utilize keyword-based prompting to guide the sentiment or theme of the sentences generated by the model.

2.3.2 Prompt Types

Two Placements There are two main varieties of prompts: (1) cloze prompts (Petroni et al., 2019; Cui et al., 2021), which require filling in the blanks of a textual string, hence interleaved with the original text input, and (2) prefix prompts (Li and Liang, 2021; Lester et al., 2021), which continue a string prefix before the input.

Two Forms On the other hand, in many cases these template words do not necessarily take the form of natural language tokens (i.e., **hard prompts**) but virtual words which would be embedded in a continuous space later. Further, some prompting methods even generate continuous vectors directly by concatenating the word vectors with the prompt vectors. This format is also termed continuous prompts (or **soft prompts**). Since the answer tokens are optimized directly in the embedding space, research using the soft-prompting method do not make use of the embeddings learned by the LM and instead learns an embedding from scratch for each label. As discussed later, this logic in prompt design provides the feasibility for the current study in formatting cognition features into multidimensional vectors that can be projected into the embedding space of the LM, hence as soft prompts for integration with the text input.

2.3.3 Prompt Design: Discrete vs Continuous

Effective prompting can often require significant trial and error or expert knowledge to craft the most effective prompts. In the search for optimized prompts, prompt engineering has been explored in prior works (Liu et al., 2019; Jiang et al., 2020; Schick and Schutze, 2020). Aligned with the categories of soft prompts and discrete prompts, prompt engineering methods also include searching for virtual tokens or virtual vectors that are conditioned on the original input.

Works on discovering discrete prompts (a.k.a hard prompts) automatically search for templates described in a discrete space, usually corresponding to natural language phrases. Specifically developed methods include prompt mining (Jiang et al., 2020), prompt paraphrasing, and gradient-based search (Wallace et al., 2019a), etc. For instance, AutoPrompt, as proposed by Shin et al. (2020), applied gradient-based searches for a sequence of discrete trigger tokens conditioned on each original input iteratively to

stimulate sentiment detection or factual knowledge extraction from a masked Language Model (LM).

In contrast, because the purpose of prompt construction is to find a method that allows an LM to effectively perform a task, rather than being for human consumption, it is not necessary to limit the prompt to human-interpretable natural language. Therefore, continuous prompts in the form of vectors (a.k.a. soft prompts) are developed that perform prompting directly in the embedding space of the model.

Generally, continuous prompts remove two constraints (Liu et al., 2023): (1) relaxes the constraint that the embeddings of prompts be the embeddings of *natural language words*, opening the possibilities for the integration of multi-modal prompting vectors (as proposed by Method 2 of the study). (2) Remove the restriction that the template is parameterized by the pre-trained LM's parameters but offers modular tuning flexibility. Instead, templates have their own parameters that can be tuned based on training data from the downstream task.

2.3.4 Parameter Updating Strategies in Prompt Research

Due to prompt engineering in prompt-based downstream task learning, there are usually two types of parameters, namely those from (1) pre-trained models and (2) prompts. This leads to multiple parameter updating strategies in prompt-based studies, including tuning-free prompting, fixed-LM prompt tuning, fixed-prompt LM Tuning, prompt+LM tuning, as summarized by Liu et al. (2023).

2.3.5 Advantages: aligning prompts with cognition signals for NLP

The advantages of adopting prompting in NLP over direct model fine-tuning echo with the challenge of cognition-enhanced NLP research reviewed in the last section 2.2 in the following aspects.

Noise-resistant model optimization First, prompting improves task-specific performance by providing explicit instructions or cues, enabling models to focus on relevant information and excel in specific NLP tasks. This makes prompting an ideal approach for integrating cognition signals that also contain inevitable noise, especially in terms of EEG data in a high noise-to-signal rate (Hollenstein et al., 2020a).

Task-generic flexibility Second, it provides flexibility and adaptability by incorporating domain knowledge or external information, allowing models to leverage context and generate more accurate responses. With the right prompt, the language model can be

steered towards a wide array of tasks without having to retrain the model from scratch for each task. This flexibility echoes with the essence of cognition-enhanced NLP studies to ultimately guide the model processing with human processing information without damaging the model performances, and can allow the cognition-prompted model potentially perform well on both text-cognition paired input and purely textual input in various task settings.

Data-efficiecny Furthermore, prompting reduces reliance on large amounts of task-specific data for training, making it a tempting approach in integrating cognition data into NLP, since task-specific textual data annotated with cognition signals is still limited or expensive at present (Hollenstein et al., 2020a; Ribeiro et al., 2023).

3

Present Study: Cognition-Prompt-Based Finetuning

3.1 Motivations: grounding cognition signals as conditional prompts

Based on the assumption from the cognition theory and empirical cognition-enhanced NLP research, gaze information (e.g., fixation durations) and word-level EEG features (from different electrodes mapping the brain) inherently encompass processing patterns of human readers about which parts of a sentence is more critical for performing the natural language processing task. Such processing information has been corroborated by previous cognition-inspired research to be able to provide positive guidance for LMs in performing downstream tasks. However, the challenge of integrating cognition data is to make the LMs learn to associate the cognition signals of humans processing the text input during reading with the textual input that are also given to the LMs for task-specific training, in search for a performance boost. As enumerated in Related Works, the previous methods in cognition-integrated NLP studies, including (a) directly concatenating dimension layers on word vectors and (b) regularizing the attention weights of the LM with numerical human cognition values, suffer from issues like heavy manual bias and low possibility of extending to larger models.

This study proposes that the cognition features can be considered as the equivalent of language data, and therefore, prepended as prefix prompts that are 'cognitively conditioned' on the text inputs. Such conditioning is not updated by traditional prompt engineering algorithms developed in NLP research (Li and Liang, 2021; Shin et al., 2020), but produced by human readers during their intrusive processing of the same text input and readily provided by the cognition corpora before training. Another nuance between the cognition prompts and the traditional text/text embedding prompts is that cognition features as prompts serve as some meta-processing guidance on the importance or certain traits of the word in the exact same position of the later input sequence. In other words, whereas

the traditional prompts are typically semantically related texts to the original input, the 'cognitively condition' prompts in this study are inherently the same sequences as the input on a meta-cognitive level. Consequently, the LM is tasked to learn the word-by-word correspondence between the prompt and the input and use the information to improve the label prediction accuracy.

Hence, the study integrates cognition features as additional 'cognitively conditioned' prompts with the original text inputs, and updates the model parameters on such input, with the aim of training the model to capture the inherent association that the cognition prompt patterns are informative about the processing complexity and relative importance of the specific part among the sequences.

3.2 Study Design

This thesis takes inspiration from the recent NLP approaches to prompt-based finetuning and multimodal learning and develops two different methods for incorporating cognition data that are much more adaptable for state-of-the-art language models. Instead of the traditional methods of directly augmenting dimensions of word embedding or extracting cognition features as target output for supervision, the cognition signals were treated as (1) sequences of cognition special tokens (i.e., analogical to text tokens) in Method 1, or (2) sequences of cognition vectors in Method 2, to be prepended with the textual sources from which the cognition signals were originally recorded human participants reading.

Hence, the cognition data were treated as additional prompts in the input to specify model behaviour, based on the power of prompting as a mainstream training approach from previous research (Liu et al., 2023). Method 1 assumes the pre-trained LM's ability to directly process and leverage numerical values of cognition signals when converted to string tokens; Method 2 converts cognition feature values into multidimensional vectors that can be projected in a 'CogMAP' framework to the embedding space of pre-trained LMs.

In general, this thesis offered exclusive contributions to renewing the methodological paradigms of cognition-inspired research by investigating the following hypothesis:

1. Can cognition features can be introduced as effective prompts to state-of-the-art pre-trained models for sentiment classification task training, by concatenating text input with numerical cognition signals either as discrete special tokens (Method 1)?

2. Can cognition features can be introduced as effective prompts for sentiment classification task training, by training a projection module to map the numerical cognition signals as prompting vectors in the continuous embedding space of the pre-trained language models (Method 2)?

3. Among cognition features, can both EEG and gaze (i.e., eye-tracking) signals can offer *unique* positive guidance to the language models' classification performance? Will the combination of cognition features further increase the training performance by providing potentially more meta-cognitive guiding information for text-processing?

4

Method 1: Inserting Special Tokens of Cognition Features for Prompt-Based Finetuning

Drawing on the hard prompting approaches in recent vision-language models (Koh et al., 2023), Method 1 operates feature incorporation in a straightforward approach by directly converting the numerical values in cognition features into strings. The multiple features of gaze and EEG data were therefore linearized into a sequence of strings and integrated separately or coordinately with the original textual sequences. The added prompts in this study, therefore, do not count as a crafted task-specific instruction, answer example, or a semantically related sentence in traditional prompting research (Liu et al., 2023), but are numerical annotations of the subsequent sentence about the cognitive processing effort on each word in it.

4.1 Input

The study utilizes the ZuCo corpus, which includes eye-tracking and EEG recordings from 12 human participants reading 400 sentences—123 neutral, 137 negative, and 140 positive—from the Stanford Sentiment Treebank (SST) for sentiment classification tasks. These samples, which constitute 4% of the full treebank, were randomly selected by the authors of the ZuCo corpus from the 'very positive', 'very negative', or 'neutral' categories of the SST-5. As of now, the ZuCo corpus is the only available dataset containing cognitive data applicable to supervised learning tasks for sentiment classification. Despite this, there are numerous other multilingual datasets (e.g., Danish, French, Dutch, Chinese, etc.) available that focus on more common NLP tasks, such as named entity recognition, relation classification (Hollenstein et al., 2020b), text summarization (Yi et al., 2020), or visual question answering (Sood et al., 2021), thereby offering considerable scope for future research utilizing the methods outlined in the current study.

4.1.1 Gaze Features

As introduced, ZuCo readily provides 5 eye-tracking variables: number of fixations (nFix), first fixation duration (FFD), total reading time (TRT), gaze duration (GD), and go-past time (GPT). Fixations shorter than 100 ms were excluded in data preprocessing following standard practices of data processing since these are unlikely to reflect language processing (Sereno and Rayner, 2003). To increase the robustness of the signal (Bingel et al., 2016; Hollenstein et al., 2020a), in this study, the eye-tracking features are averaged over all subjects. However, the methods in this thesis allow modular training on multiple participants for the purpose of catering to personalized contextualization in future studies and applications. Notably, five gaze features are created based on the raw eye-tracking variables, without any high-level transformation. Compared to the diverse feature extraction methods mentioned in 2.2.3, this study adopts a much more straightforward feature extraction solution for two reasons: to avoid heavy human-biased engineering in feature extraction from which the early works suffered, and to reduce the sequence of the special tokens to be added before each sentence input. The latter effectively constraints the computational cost and device capability requirements for training, as well as avoids potential contamination of the inputs due to a dis-proportionally long prompt. Such concerns will also become more significant in handling EEG features.

4.1.2 EEG features

The simultaneous recordings of eye-tracking and EEG data from ZuCo enabled the extraction of word-level EEG features from the time-stamped raw data. During preprocessing, 23 out of 128 original electrodes in the outermost circumference (chin and neck) that were used to detect muscular artifacts were removed for subsequent analyses, rendering 105 electrode values (Hollenstein et al., 2018b).

Following McGuire and Tomuro (2021), the 105 electrode values were mapped to first-pass fixation onsets to create 105 fixation-related potentials (FRPs) for each word. Although the variability between subjects is much higher in the EEG signal compared to gaze data, EEG features were averaged across subjects following the established practices. Both eye-tracking and EEG feature values were normalized between 0 and 1 using a Min-Max-Nomralization scaler. Regarding words that only receive saccades (i.e., rapid and ballistic movements of the eyes that abruptly change gaze position) without fixation landed, the raw data yields void values for all gaze features and EEG features. Such features were set as 0 to ensure the corresponding relationship between cognition prompts and actual textual input on the word level. Lastly, all feature values are uniformed as

floats with two decimals before converting recessive zeros and inconsistent patterns in the prompts to minimize potential confusion for the model.

Due to the constraints of the sequence length for training, it is impractical to flatten the 105 features into 105 special tokens and insert them as prompts because the input would exceed the maximum sequence length allowed by LMs, and risks diluting the original text that contains the semantic information for class label prediction. This is admittedly one of the major limitations of method 1 despite its perks of simplicity. As a compromised alternative, this study resorts to dimension-reduction algorithms in machine learning to shrink the number of special tokens denoting EEG features. Specifically, Principal Component Analysis (PCA), a widely used technique for reducing the dimensionality of large datasets, was adopted as an attempt to condense the 105 electrode values down to a more manageable 5 features. It allows the potential to retain the essence of the EEG data while not overwhelming the model's input or diluting the semantic content of the text, thereby preserving the feasibility of including EEG features in the training process.

4.2 Models

For the generalizability of the conclusion, a diversity of mainstream state-of-the-art language models are employed as backbone models for training in Method 1. They include both bidirectional encoder-only models (i.e., BERT, RoBERTa-base, RoBERTa-large) and auto-regressive decoder-only models (i.e., GPT-2 and its larger variants), as shown in 4.1.

Table 4.1 Architecture hyperparameters of the pre-trained backbone LMs (Radford et al., 2019). (GPT-2 refers to the smallest version of GPT-2s when not specified.)

Type	Backbone model	Parameter size	Layers	Dimensions
Encoder-only	BERT	110 M	12	768
	RoBERTa-base	125 M	12	768
	RoBERTa-large	335 M	24	1024
Decoder-only	GPT-2 (small)	117 M	24	1024
	GPT-2 Medium	345 M	36	1280
	GPT-2 Large	762 M	48	1600

To address the encoder-only category, both BERT and RoBERTa-large have been enlisted due to their structural similarities, yet varied scale. These models have gained widespread recognition for their robust performance in tasks involving understanding the context of a sentence, making them suitable for investigating the cognition-prompting approaches in the study. Their transformer-based architecture pre-trained on extensive

volumes of data allow the models to capture long-range dependencies in the input data. This, in turn, positions them advantageously to potentially discern the association between cognition prompts and texts with greater effectiveness, in comparison to previous works employing Bi-LSTMs in ternary sentiment classification tasks on the same set of data (Barrett et al., 2018; Hollenstein et al., 2019).

GPT-2 and its variants as the decoder-only models were also included in the experiment implementation. Method 1 starts with the basic smallest version of GPT-2, a publicly accessible 124M-parameter autoregressive LM trained on the Pile dataset (Gao et al., 2020). This base model is later expanded to larger-scale variants. Firstly, the structural similarity across GPT-2 and its more advanced variants such as GPT-2-medium (355M), GPT-2-large (774M), GPT-2-XL (1.5B), GPT-J (6B) and other state-of-the-art causal language models like OPT (Zhang et al., 2022), etc., lends a level of convenience and feasibility when scaling the backbone model. While encoder-only models are specifically suitable for classification tasks, the trajectory of advancement in the field of NLP has shifted somewhat towards autoregressive, decoder-only models, especially those with a structure akin to the GPT family. This trend has motivated the inclusion of GPT-2 and its larger variants in the framework design. Secondly, GPT models opened possibilities to extend the proposed approach to a range of more diverse tasks other than classification (i.e., generation tasks) in future studies. In addition, thirdly, the GPT-2 models, despite their pretraining objective, have demonstrated similarly remarkable performance on tasks including sentiment classification when properly fine-tuned with an adequate data volume. (Radford et al., 2019). Therefore, in summary, opting for GPT-2 and its advanced variants aligns with the study's objective to evaluate the efficacy of large language models when grounded on cognitive data in a generalizable setting. Considering a series of factors including scalability, popularity, and task-generic adaptability, the GPT model architecture offers the opportunity to systematically assess the impact of model size in the experiments, as discussed later in the results section.

To deploy GPT-2 (and its variants) for the ternary sentiment classification task, a classification head, i.e., a linear layer followed by a softmax function, is added on top of the transformer's output for the last token in the sequence. This head converts the hidden state of the last token into probability distributions over the sentiment labels. The sentiment label ('positive', 'negative', 'neutral') with the highest probability is then selected as the model's prediction for the input text.

4.3 Experiment Setup

Prompt design: special tokens

Method 1 linearizes 5 word-level gaze features/5 EEG features into 5 text tokens to represent the human processing attention to each 'word'. As introduced in related works, each feature value ideally could denote a unique but complementary aspect of cognitive information: for eye-tracking data that measures overt attention, GD, FFD, GPT, nFix, and TRT denote different stages of processing; for EEG that mostly measures covert attention and neuron activities, the original 105 values represent information extracted from different regions of the brain which are compressed to 5 features. One challenge regarding designing a prompt suitable for the model input is organizing the multiple eye-tracking/EEG features for each word while at the same time not contaminating the text input.

In order to demarcate word-level multi-features among the prompt sequence for model readability, as well as to guide the model to identify and differentiate each feature as prompt input within its fixed vocabulary, five pairs of open and closing special tokens are added to the LM tokenizer to account for five eye-tracking, and similarly five for the experiment on EEG features. After adding these special tokens, the model's embedding matrix also needs to be resized to accommodate the new tokens. Taking GD (i.e., gaze duration) among the eye-tracking features as an example, `[GD]` and `[\GD]` were added as an opening or closing special token that signifies the start/end of the particular feature annotation. As shown in Figure 2, between the pair of special tokens is the value of the GD variable in the ZuCo corpus (after normalization) that sets the context for the numerical value that follows and indicates that the value represents the GD feature. This format clearly isolates between the multiple features within eye-tracking/EEG data for the given word or phrase as well as excludes them from the model processing of the corresponding textual input, allowing it to be easily recognized and utilized by the model tokenizer.

Another necessity of adding special tokens for filling the cognitive features is they offer boundaries that prevent the string-formatted floats from being further tokenized. Both tokenizers of BERT (WordPiece) (Wu et al., 2016) and GPTs (Byte-Pair Encoding or BPE)(Sennrich et al., 2016) are sub-word tokenizers which split the floats into further subunits if not in the tokenizer vocabulary (e.g., '0.56' into ['0','5','6']). In contrast, special tokens that encapsulate these string-formatted features effectively "shields" numerical values from subword tokenization, ensuring that these cognitive measures are processed in a single unit with their specific contexts preserved. This function is essential because the meaningfulness of the cognitive features lies not in their potential sub-parts, but in their entirety as numerical values representing specific cognitive measures. Furthermore, this practice also reduces input complexity. Conventional tokenization could inflate the

number of tokens for string-formatted numbers. With special tokens, each cognitive feature contributes only a single token value to the input sequence, improving efficiency but also methodological feasibility considering the sequence length constraint of Method 1.

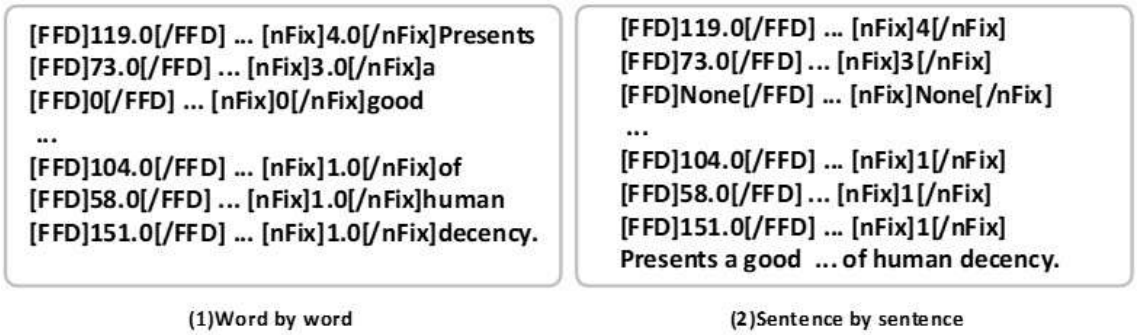


Fig. 4.1 Tokens Arrangement in a sample sequence prompted by gaze features. Each 'word' (split by the ZuCo corpus) carries five word-level gaze features, and hence five pairs of special tokens (i.e., [FFD], ..., [nFix]).
The first prompt-placement paradigm was shown on the left, where the word-level gaze features are inserted before each word within the textual sequence; the second paradigm was shown on the left, where the word-level features are aggregated together as a sequence before concatenating with the textual sequence as a whole.

Prompt placement

The goal of the cognition-integration experiments is to investigate whether the prompt-based finetuning method with raw cognition features as sequences of textual tokens can be spontaneously 'analysed' by state-of-art LMs as useful information to enhance task-specific performance. As shown in Figure 1, two paradigms of prompt placement were explored: (a) inserting the corresponding word-level feature token before each *word* during a sentence (as split by blanks, and aligned with the word tokenization style by ZuCo), or (b) concatenating the corresponding sequence of feature tokens altogether before each complete sentence. This is due to a conundrum in the motivation of pursuing superior model performance: on the one hand, word-by-word prompting may guide the model to better capture the denoting association between cognition features and the corresponding word in the exact same position among two sequences; on the other hand, such method of introducing multiple tokens within a sentence sequence can potentially interfere with the language model's ability to effectively tokenize the text and correctly assign attention masks to the word embeddings. This in turn can disrupt the contextual information associated with the target token, an aspect that is vitally important for attention-based language models. In comparison, the second paradigm resembles traditional tuning more, where feature tokens are grouped together and placed before the complete sentence. This

approach attempts to balance the need for integrating cognitive feature information without compromising the inherent contextual flow of the sentence. This presents a compromise between highlighting cognitive feature-word associations and maintaining the integrity of the sentence structure for effective language model processing.

Ablation experiments of gaze/EEG/combination features were also conducted to further assess the efficacy of these paradigms. By systematically omitting certain feature types and observing the resultant model performance, the ablation analysis aimed to discern the impact of each feature on the overall predictive capability of the model.

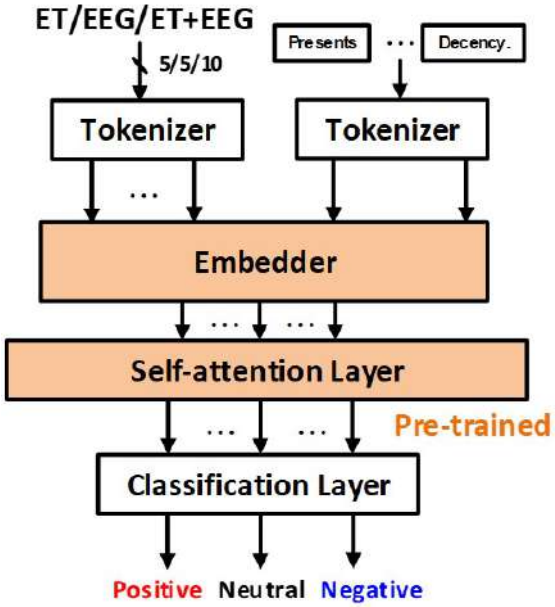


Fig. 4.2 Model Architecture for Method 1 (ET refers to eye-tracking i.e., gaze data).'5/5/10' denotes the number of features carried by ET, EEG, and their combination per 'word' (as split by the ZuCo)

.

4.4 Training

The employed language models were trained on ternary sentiment classification tasks. Given the limited size of the available dataset, we diverge from the conventional 8:1:1 training-development-testing split, and instead, randomly shuffle and divide the data into 50% for training, 10% for development, and the remaining 40% for testing. This atypical configuration prioritizes obtaining a more reliable and rigorous testing performance score by allocating a larger proportion to the testing dataset. The prompting approach inherently allows for effective training on smaller datasets, further justifying this unconventional data split. To enhance the robustness and representation of the model evaluation on the

development dataset, we employ a 10-fold cross-validation method. An early stopping mechanism into our model training protocol. This strategy stopped additional training after a duration of 5 epochs without discernible enhancements on the development set. It offers a balance between the computational expenses of continued training and the prospective benefits derived from supplementary iterations.

Instead of a multi-task learning framework in previous works (Hollenstein et al., 2019; McGuire and Tomuro, 2021; Sood et al., 2020b), the training objective in this study is solely to optimize the prediction of sentiment labels. Due to the flexibility of the prompting approach, we conducted experiments utilizing a variety of backbone models, encompassing bidirectional encoder models (BERT and RoBERTa), as well as autoregressive decoder models (GPT-2, GPT-2-medium, and GPT-2-large), from the Hugging Face Transformer as outlined by Wolf et al. (2020) for implementation. During the training process, the AdamW optimizer (Loshchilov and Hutter, 2019) is utilized, along with a linear learning rate scheduler in accordance with the recommended setup from Hugging Face. All the training experiments were implemented over five randomly chosen identical seeds. The tuned hyperparameters include the number of epochs, batch size, learning rate, and random seeds. Further details regarding these hyperparameters can be found in Appendix 1.

4.5 Advantages

The core of Method 1 revolves around linearizing the arrays of multiple features in eye-tracking/EEG data, originally in the form of float values in the cognition corpus, into strings and concatenate them as prompts before the text input. Despite seeming extremely simple and 'forceful' in operation, it can potentially demonstrate effectiveness in integrating cognitive processing information to enhance model performance due to several reasons.

Firstly, by transforming numerical cognition features into textual format, this method fully exploits the inherent strengths of language models, which are essentially developed to handle and extract patterns from textual data. Empirical examination of LMs has demonstrated that pre-trained language models, though not liable for complex mathematical reasoning tasks, naturally possess a great level of numerical comprehension (Wallace et al., 2019b). Directly incorporating cognitive information in a format these models are primed for, hence, ensures seamless integration and efficient processing while at the same time avoiding damaging the model performance, a potential hazard in previous methods that attempted at altering the inner attention layers of the neural network or the pretrained LM.

Secondly, the conversion process maintains the inherent structure and semantics of the cognition data. For instance, the sequence of EEG features might correspond to the

temporal sequence of cognitive states while reading the text, thereby providing temporal context to the model, which could be critical for understanding the text.

Thirdly, when these transformed cognition features are combined with the original text, it creates a richer and more informative input sequence for the model. This can enable the model to ground textual information within the cognitive context, potentially enhancing comprehension and representation learning.

Lastly, the simplicity of Method 1 may contribute to its effectiveness by avoiding computational complexity and overfitting issues that could occur with more intricate feature integration methods. The uncomplicated incorporation approach is not only computationally attractive but also has fewer parameters, thus reducing the risk of overfitting and potentially making the model more generalizable.

In conclusion, the effectiveness of Method 1 could largely be attributed to its ability to utilize the strengths of language models, retain the meaningful structure of cognitive data, create a richer input sequence, and maintain computational simplicity and robustness against overfitting. Future research should further investigate the impacts of this method on various language tasks and seek to optimize the conversion and integration process for even greater performance enhancements.

5

Method 2: Mapping Multidimensional Cognition Features as Prompting Vectors

Method 2 draws reference from the 'soft prompting' approach by converting the multi-feature gaze and EEG data into multidimensional vectors and trains a mapping network to map their dimensions to that of the language model input. By reshaping the numerical values into a single multidimensional vector to represent the cognitive processing information per word, this method maps the cognition features, another data modality, to the semantic vector space of language models. As a result, cognition prompting was performed directly in the embedding space of the model. Method 2 can effectively address the constraints of Method 1 with a limited number of incorporable cognition features due to the concern of excessive sequence lengths.

5.1 CogMAP Architecture

Method 2 developed a unique framework termed '**Cog**nition **M**apping **A**nd **P**rompting' (CogMAP), which aims to leverage the power of large-scale pre-trained language models, and learns a small mapping network to convert n-dimensional cognition features into i-dimensional embeddings that can be concatenated before text token embeddings as prompts, where i refers to the token embedding in the vector space of the LM, as illustrated in Figure 2. This is the first time that the prompting approach, one of the thriving model deployment trends in NLP, has been introduced to cognition-enhanced research.

The CogMAP framework is composed of an LM tokenizer, a projection layer that handles modality alignment by performing dimensionality projection, and lastly, a transformer-based language model that trains on the concatenated embedding. The design of the projection module is partially borrowed from Mañas et al. (2023) who implemented a projection method for multi-modality alignment that learns a lightweight mapping between the representation spaces of a pre-trained language model and a vision-language model

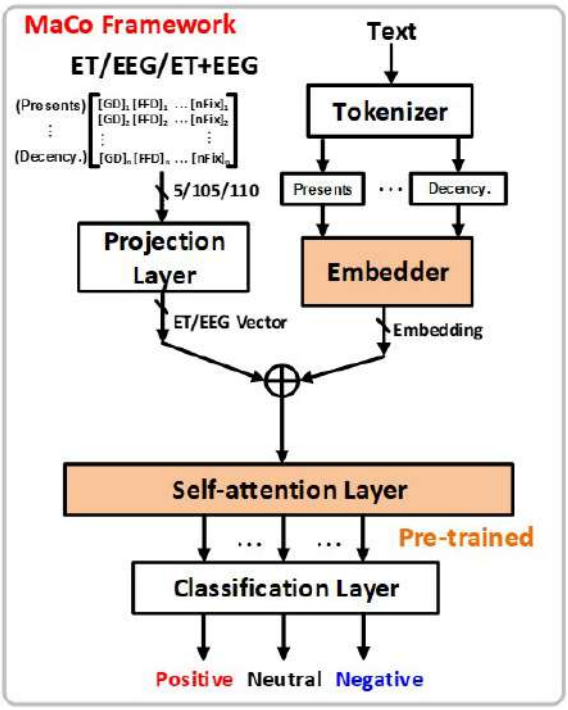


Fig. 5.1 Overall Architecture of CogMAP.

to leverage their strong generalization capabilities. In the context of this study, since cognition corpora can provide an aggregation of variables that resemble multidimensional vectors, cognition features already serve as 'encoded' cross-modal embedding (similar to image embedding encoded by a vision-language model). Hence, the main objective of the CogMAP architecture is to introduce the cognition vectors into the representation spaces of a pre-trained language model in order to process and comprehend the multimodal input.

In general, the CogMAP framework accepts a pair of texts and the corresponding cognition features (gaze/EEG/combination). The cognition features are first forwarded as multidimensional vectors to a projection layer for dimensionality projection, then concatenated together with the tokenized text embeddings to create the input that can be learned by the pretrained language model. Below is a detailed breakdown of its structure.

Pre-trained language models The LMs employed by CogMAP in Method 2 as backbone models are decoder-only pretrained models (i.e., GPT-2, GPT-2-medium, GPT-2-large), leaving the design of adapting encoder-only models for future work.

Before training, the text input is first segmented and padded into a fixed sequence of discrete tokens by the LM's tokenizer. In particular, when implementing training on decoder-only models, sequences are left-padded adhering to the established norms for training a GPT model for classification tasks, thereby designating the final token as the target for predicting the output. Then, each token is subsequently transformed into a

continuous embedding of size D_l (e.g., in GPT-2-small's case, D_l = 768) by the LM's token embedder. Subsequently, the sequence of token embeddings is then introduced to the self-attention layers within the language model's transformer block, which employs causal attention. At this point: rather than directly utilizing the language model to produce hidden states corresponding to each text token or embedding for probability distribution and prediction, the CogMAP concatenates the cognition embedding sequence (which has already been mapped into the same dimension of D_l) with text embeddings representing the same sequence. This process merges the two modalities of input into a singular sequence. The attention masks for both types of embeddings are also concatenated to signal to the language model their respective positions of attention. Ultimately, the language model processes this concatenated input (in the form of embeddings in this study), outputting a sequence of hidden states corresponding to each token that are utilized to predict sentiment labels.

Projection Layer In terms of cognition input, each gaze/EEG vector represents the cognitive processing information of one word in the textual sequence - as readily split by the ZuCo corpus and therefore requiring no tokenization. However, note that such word splitting is defined by the corpus and not aligned with the tokenization of the LM of the same text sequence. Since all transformer models require the input, regardless of the format, to be of the same length, such multidimensional vectors are also left-padded using a specially added padding vector (set as the vector of -1s) by CogMAP with the maximum sequence length being the maximum number of word-level cognition vectors among sequences (split by ZuCo) before forwarded to the projection layer.

The projection layer takes as input the fixed-length sequence of original multidimensional vectors representing the complete information of cognition features corresponding to the sequence of text, and performs a linear projection. It is therefore a simple mapping network of a linear projection layer that updates its weight matrix. The weight parameters of the projection layer are randomly initialized by default using the Kaiming Uniform initialization method from PyTorch, and biases are set to zero. Two additional hyperparameters were concatenated as constants with the projection layer parameters to improve the quality of projection: output length (i.e., the maximum number of cognition vectors in a sequence in padding, 41 for the dataset of this study) and input dimensions (e.g., = 768 in the case of GPT-2 and its variants). After dimensionality projection, the projection layer forwards the sequence of transformed 'cognition embedding' to LM for concatenation with the corresponding sequence of token embedding.

5.2 Input

The preprocessing of eye-tracking and EEG features from the ZuCo corpus in Method 2 mirrors the approach used in Method 1. However, Method 2 alleviates the limitations regarding the quantity of cognitive features, thereby maintaining the original 105 EEG features per word without necessitating dimensionality reduction via PCA. For the purpose of ablation studies, experiments were conducted using eye-tracking features, EEG features, and a combination of both as input prompts. With the array of variables available in the ZuCo dataset, each cognitive vector—prior to dimensionality projection—consists of 5 dimensions for gaze vectors, 105 dimensions for EEG vectors, and 110 dimensions when gaze and EEG vectors are combined.

The sequence of multi-dimensional vectors is input into the multi-modal CogMAP framework, where it is merged with the original text input to form 'prompting vectors' for the classification encoder of language models to process. Specifically, the handling of cognition vectors is bifurcated into two facets: dimensionality-wise and sequence-wise. In terms of dimensionality, the mapping module in the CogMAP framework aligns the modality of cognitive data with texts, as elaborated in 5.1; As for sequence-wise handling, akin to Method 1, the mapped cognition vectors are padded and inserted before the text token embeddings in the language model's inner vector space, accompanied by a combination of their attention masks.

Essentially, in Method 2, a sequence of multi-modal embeddings — encapsulating the full cognitive feature information (gaze, EEG, or a combination of both) of its subsequent textual input sequence — is concatenated prior to the original textual sequence. Through mapping the cognition embedding to the vector space of the language model, it is hypothesized that during training using the input added with cognition-vector prompts, the LM will pick up the association between texts and its corresponding cognitive guiding information indicated by the continuous values in the preceding gaze/EEG/both embeddings. Hence, based on the cognition informatory theory, this is expected to inspire superior performances in downstream task deployment.

5.3 Experiments

The training task and experiment setting maintains the same as Method 1. However, due to practicality considerations based on empirical observation of Method 1's results (i.e., 6.1.1), only the second prompting paradigm is implemented, concatenating the whole sequence of cognition prompts before texts, rather than interleaving the word-level fragments of cognition features into text sequences.

In addition, Method 2 doesn't require the addition of special tokens to alter the language model tokenizer's configuration. This is because cognition prompting is performed in the continuous multidimensional embedding space of the pre-trained language model. Compared to Method 1 where the string sequences of cognition prompts need to be marked as special token values because such numerical input does not fall within the pretrained language model's fixed vocabulary of discrete tokens, Method 2 offers an alternative with less manual operation to the inner structure of the LM. Instead, through updating the parameters of the projection layer as well as the language model during training, while the language encoder learns the sentiment classification task, the projection layer in CogMAP also learns to adjust the position of the cognition prompt embeddings closer and closer to the embeddings representing the corresponding output sentiment label in the continuous semnatic vector space within the language model at the same time. Hence, the finetune CogMAP framework can learn to

5.4 Training

Similar to Method 1, prompt-based fine-tuning was adopted, except in the form of continuous soft prompts in the vector space. Specifically, CogMAP is trained with a simple fine-tuning objective on cognition-text pairs (Lamb et al., 2016), treating cognition input as multimodal pseudo-dynamic prompts, i.e., we minimized the cross-entropy loss (i.e., negative log-likelihood) of the sentiment labels for each sequence i in batches of size N with sequence-level prediction y_i, under the LM conditioned on the corresponding prompted input. Rather than LM-fixed tuning, this study chose to train both the projection layer (i.e., linear projection layer) from scratch and the pretrained LM, because decoder-only pre-trained models like GPT-2 do not have the zero-shot capability of performing classification tasks. The training process is also illustrated in Figure 2.

The training objective of CogMAP is defined as

$$\mathscr{L}_{SC}(\theta_m, \theta_l) == -\frac{1}{N}\sum^{N}\sum_{i=1}^{3} y_i \log p_\theta(y_i|T_i, C_i)$$

where θ_m and θ_l refer to the learnable parameters of the projection layer and the LM, T_i and T_c are the text and cognition input. y_i is the ground-truth sentiment class while $p_\theta(y_i|T_i, C_i)$ is the Softmax probability of the predicted sentiment label for the i-th input under the parameter set $\theta = \{\theta_l, \theta_m\}$, During training, θ were updated to output the the predicted label with the highest probability compared with the ground-truth sentiment y_i.

The modular framework is designed to be flexible for future adaptation to a diversity of tasks. Due to the nature of the employed models and the task, the parameters of both the

mapping network and the LM are updated. However, similar to MAPL, this framework can be trained modularly by only updating the mapping network but keeping the LM frozen for other deployments. This opens possibilities for future studies investigating few-shot learning or LM-freezed prompt tuning on LM for more tasks on the cognition-integrated input in the future.

5.5 Advantages

Method 2 not only retains the advantages of Method 1 but also remedies its shortcomings by mapping the cognition features into a higher-dimensional space. By simply connecting an additional linear projection layer before the language encoder, it renders the cognition-prompting approach more flexible and clean. To elaborate, Method 2 offers the following three key advantages.

Firstly, it removes the restrictions associated with the number of cognition features that can be integrated. By converting the multi-feature gaze and EEG data into multidimensional vectors, it bypasses the excessive sequence length issues encountered in Method 1. This allows the preservation and use of the full set of 105 EEG features per word, effectively leveraging the richness of the cognitive processing information.

Secondly, Method 2 enables a more explicit word-by-word association within the prompted input, thus creating an optimized learning environment for the language models. The sequence of multimodal embeddings, which is of the same length as the number of words in the original text input, is concatenated before the original text sequence. This approach provides a comprehensive representation of cognitive feature information (gaze, EEG, or combination) associated with each word. Consequently, this can aid the language model in discerning the cognitive-feature-to-word association, facilitating a deeper and better comprehension of cognitive influences on language processing.

Thirdly, Method 2 eliminates the constraint observed in both Method 1 and previous methods wherein the input is solely parameterized by the pre-existing parameters of the language model. Instead, input features are endowed with parameters based on the architecture of the model framework, specifically in CogMAP, the mapping network interconnected with the language model. This setup provides the flexibility for modular tuning, which means that these parameters can be separately optimized to better adapt to the cognitive feature inputs, thereby providing a more efficient and adaptive model for processing and mapping cognitive data. This advantage underscores the ability of Method 2 to evolve and optimize according to the specificities of the cognitive data at hand, further enhancing the performance and adaptability of the language model.

6

Results and Discussions

After training, the final models were evaluated on the held-out test dataset of 160 samples. The splitting ratio of 40% of the total 400 datasets ensures the reliability and validity of test accuracy, sacrificing the relative size of training data which is less of a concern considering the nature and perks of using prompt-based fine-tuning that requires fewer data in nature in this study. The results are reported after averaging over five random seeds. 6.1 and 6.2 show the evaluation results in Method 1 and Method 2 on a range of backbone LMs including encoder-only models (i.e., BERT, RoBERTa) and decoder-only models (i.e., various variants of GPT-2).

When assessing the outcomes of the current study, it's crucial to underscore that the principal metric hinges on contrasting cognition-prompt-enhanced models against baseline models trained exclusively on textual data. This is guided by our primary research question revolving around whether human cognitive processing signals can enrich language models, thereby bolstering their performance on specific NLP tasks, via a novel method of prompt-based fine-tuning. In contrast, comparisons with prior research in cognition-inspired NLP, which predominantly utilized more basic language models (generally confined to Bi-LSTMs and BERT, without delving into more advanced models), may not yield substantial insights. This is because the current study is geared towards integrating a cognitive corpus into state-of-the-art language models of increasing sophistication and scale. Thus, our focus is primarily on advancements within this specific context, rather than benchmarking against older, potentially less-relevant research.

The performance metric across all experiments is the overall accuracy score. In line with the statistical analysis approach employed in cognition-inspired research (Hollenstein et al., 2019; McGuire and Tomuro, 2021), the two-sided Pitman's permutation tests (Dror et al., 2018) were conducted on final accuracy measures to establish statistical significance. This was applied to comparisons between six different configurations across two methods and their corresponding baselines for each backbone model.

Furthermore, the Bonferroni correction was also implemented as a conservative statistical technique to account for multiple hypotheses. This adjustment rejects the global null hypothesis if $p < \alpha/N$, where N signifies the total number of hypotheses (Dror et al., 2017). In the scenario of the current study, N equals 6, representing the combination of the two methods and three configurations (EEG, gaze, and EEG+gaze). The values are marked with asterisks in the result tables where the improvements were statistically significant over the text baselines (** for $p < \alpha = 0.001$, * for $p < \alpha = 0.01$), even after applying the Bonferroni correction. Despite the limited data, the findings of consistent significance can suggest that incorporating cognitive features into NLP systems through prompt-based fine-tuning is not only effective but also consistently generalizable.

6.1 Method 1

6.1.1 Word-Level Interleaved Cognition Prompts are Sub-Optimal

As previously outlined in Section 4.3, Method 1 adopted two distinct paradigms of prompt placement. The primary, relatively unconventional paradigm that involved introducing word-level cognition prompts prior to each 'word' (as segregated directly by the ZuCo corpus), generated initial results that fell short of expectations. Notably, classification accuracy for BERT (41.80%), RoBERTa-base (43.22%), and RoBERTa-large (51.87%) demonstrated a considerable downturn when contrasted with text-baseline outcomes.

This decline is hypothesized to stem from the disruptive influence that the introduction of cognition prompts had on the inherent contextual dependency among tokens within transformer-based language models. The fragmentation, brought on by the inclusion of 'word-level' prompts, appeared to negatively impact these models' capacity to process comprehensive text sequences. Despite theoretical proposals suggesting that pretrained language models could discern the distinctive role of cognitive prompts as special tokens and learn to compartmentalize them during attention assignment, the empirical observations pointed out that such a prompting paradigm results in an unnatural sequence flow that hinders the model's ability to correctly comprehend and interpret textual context.

Upon a more in-depth examination, it was observed that a potential issue contributing to the problematic prompted input was the misalignment between the sub-word tokenization, carried out by language model tokenizers, and the position of word-level prompt insertion, determined by ZuCo's predefined splitting of word-level cognitive data. Consequently, it becomes challenging for the language model to learn the corresponding association between text fragments (tokenized as subword tokens) and special tokens of word-level cognition prompts, let alone strategically allocate attention scores.

In summation, the exploration of fragmentary cognition prompts interlaced with text in Method 1 highlights the necessity of preserving the natural flow and continuity of sentences for optimal performance in future cognition-prompting research. Given the less than satisfactory preliminary outcomes, the examination of the first prompt placement paradigm was discontinued to conserve computational resources for further exploration and fine-grained analyses of more promising settings and models.

6.1.2 Sentence-level Cognition Prompts

Table 2 shows the overall accuracy, as well as weighted precision, recall, and F1 scores of the cognition-prompted models compared with baselines which were pre-trained backbone LMs finetuned solely on text inputs. For better visualization, the accuracy results were extracted and drawn in 4.2 too.

Model	Text-baseline				Gaze				EEG				Gaze + EEG			
	A	P	R	F1	A	P	R	F1 A	P	R	F1	A	P	R	F1	
BERT	70.0	70.8	70.0	70.0	**74.8****	75.3	75.5	75.4	72.9**	72.5	72.7	72.6	71.2	71.7	71.9	71.8
RoBERTa-base	71.1	71.8	71.1	71.2	**76.6****	77.1	77.3	77.2	74.9**	75.4	75.6	75.5	70.1	70.6	70.8	70.7
RoBERTa-large	75.9	75.6	75.9	75.5	**79.4****	80.0	80.2	80.1	79.1**	79.7	79.9	79.8	70.1	70.6	70.8	70.7
GPT-2	44.0	57.5	44.0	44.4	**45.6***	46.2	46.4	46.3	45.1	45.5	45.7	45.6	41.2	41.7	41.9	41.8
GPT-2-medium	48.5	50.3	48.5	47.3	**52.1****	49.6	49.8	49.7	51.3**	51.9	52.1	52.0	42.4	42.9	43.1	43.0
GPT-2-large	47.3	46.5	47.3	43.0	**53.0****	50.5	53.8	50.6	50.2**	50.8	51.0	50.9	42.0	42.5	42.6	42.6

Table 6.1 Overall Accuracy (A), Weighted Precision (P), recall (R) and F1-score (F1) for the ternary sentiment classification tasks augmented by gaze features, EEG features, and the combination of both as discrete hard prompts. Experiments were run on encoder-only models and decoder-only models for ablation purposes. The significance of accuracy as the metric is indicated with the asterisks: * = p<0.01, ** = p<0.001 (Bonferroni method). The largest accuracy values are in bold.

In general, the hard prompting method of inserting cognition features as sequences of discrete special tokens achieved consistently superior classification accuracy when evaluated in comparison with text baselines. However, a comparative distinction should be drawn between BERT-based models and GPT-based models due to their variation in absolute accuracy scores and performance under the influence of cognition-prompt-based finetuning. The following subsections describe and discuss the results individually in detail categorizations for BERT-based models and GPT-based models to shed light on their distinctive performance characteristics.

Encoder-only vs Decoder-only Models Broadly inspecting 6.1, there is a consistent statistically significant performance boost across BERT-based encoder-only models and GPT2-based decoder-only models, with the inclusion of gaze and EEG features compared to the text baseline. For example, RoBERTa-large, for instance, goes from a baseline

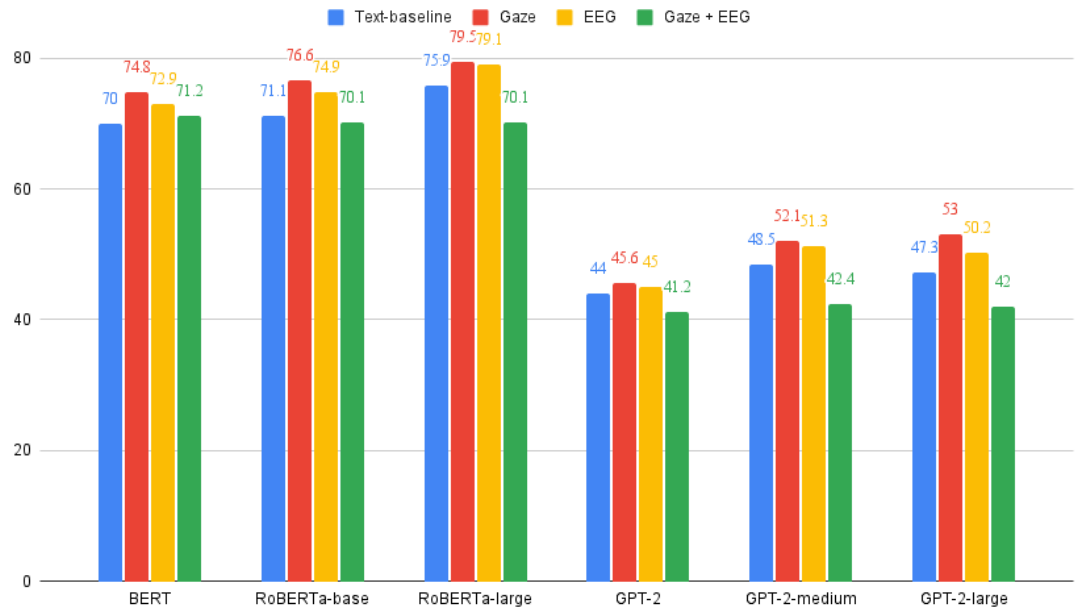


Fig. 6.1 Graph of accuracy scores across models extracted from 6.1

accuracy of 75.9% up to 79.4% and 79.1% with the inclusion of gaze and EEG features, respectively. This demonstrates that these cognitive features significantly enhance the model's ability to accurately perform sentiment classification tasks.

Nonetheless, a discernible performance gap emerges between GPT2-based models and BERT-based models during the experimental implementations. In particular, GPT-2 and its variants illustrate a comparatively weaker performance increase when prompted with gaze or EEG data and also lag behind BERT-based models in text-input baseline experiments (i.e., 70.0-75.9 vs 44.0-48.5). This discrepancy may be primarily attributed to the nature of GPT-2s as pre-trained language decoders designed for generation tasks rather than classification tasks. Although a classification head allows for finetuning such models for sentiment classification, it necessitates substantial data for proper finetuning to ensure acceptable task-specific performance. This highlights the pivotal role of corpus size and underscores the need for the collection of more NLP-task-related cognition signal corpora in future cognition-inspired research.

Gaze vs EEG Delving into a detailed comparison between different types of cognitive inputs, the models' performance exhibited more significant improvement when gaze features were incorporated compared to EEG features. This finding was consistent across model categories, as accuracy scores with gaze features consistently surpassed those with EEG features. To substantiate this observation, paired t-score analyses were conducted comparing gaze and EEG columns across models, revealing a statistically significant

difference between the classification performances ($p < \alpha = 0.01$). This suggests that EEG information, when introduced as discrete prompts in Method 1, offers less effective guidance for sentiment class prediction than gaze information. This conclusion aligns with previous research which enriched language models with cognitive data for various downstream tasks (Hollenstein et al., 2019; Hollenstein and Zhang, 2019; McGuire and Tomuro, 2021; Ren and Xiong, 2021).

This study proposes two potential, non-mutually exclusive reasons that may underpin this observed nuance between gaze and EEG data, irrespective of the model types used. Firstly, the gap may stem from the inherent capacity of gaze and EEG features to convey the processing information that can be utilized by the models. As described in Chapter 2, EEG has the capability to capture covert attention, implying processing efforts towards peripheral vision or internal mental representations, without any related eye movements. In contrast, eye-tracking explicitly monitors overt attention, reflecting the subject's direct visual focus.

Secondly, the difference could be a result of the low signal-to-noise ratio inherent to EEG data and the additional preprocessing steps required to handle the EEG features that could potentially obscure the underlying neural activation patterns in the brain. Specifically, after excluding noisy electrodes, Method 1 involves a manual dimensionality reduction of the 105 raw electrode values through Principal Component Analysis, as detailed in Chapter 4. This technique fits the sequence of EEG features into the text input as discrete special-token prompts within the constraint of the LM's maximum sequence length. However, EEG data requires meticulous preprocessing and noise deduction, which is contingent on different psycholinguistic testing paradigms, suggesting that the PCA technique used in the current study may not be optimal (Shoka et al., 2019; Winkler et al., 2011).

Despite these considerations, the significance across models empowered by EEG and gaze feature prompts demonstrates that Method 1 is an effective novel approach to integrating either type of cognition data into state-of-the-art language models, underpinning the fundamental first and second hypotheses of this thesis.

Single Feature Type vs Combined Features Contrary to the third hypothesis, models incorporating single-type gaze or EEG features consistently outperformed those combining both types of cognitive data, as shown in the last column of Table 6.1. Furthermore, the amalgamation of gaze and EEG features in Method 1 failed to exceed the text baseline in terms of accuracy and F1 scores for most of the underlying models.

To juxtapose the finding with past cognition-informed NLP research, there has been conflicting evidence supporting a superior or inferior performance when integrating a combination of gaze and EEG features compared to introducing solely gaze or EEG data (Hollenstein and Zhang, 2019; McGuire and Tomuro, 2021; Sartakhti et al., 2021).

Hollenstein et al. (2019) provided two plausible explanations for the nuances in such feature ablations: firstly, the combination of gaze and EEG features decreases the signal-to-noise ratio even more than for only one type of cognitive data; secondly, another interpretation is that the eye-tracking and EEG signals contain information that is overly similar rather than complementary. Thus, the combination does not improve yield better results.

On top of the two explanations above, this thesis discusses a unique interpretation concerning the experiment configuration that potentially harms the accuracy of gaze+EEG-prompted models: The underperformance of models using combined gaze and EEG prompts in Method 1 likely stems not only from the interaction of these cognitive features, but also from the constraints imposed by sequence length. As a significant limitation of Method 1, the need to flatten multidimensional, word-level gaze and EEG features into one-dimensional text tokens for hard prompting invariably results in extensive cognitive prompt sequences. This has two adverse effects: (a) it dilutes attention towards the actual text input, and (b) more detrimentally, it risks exceeding the maximum sequence length configuration of a tokenizer.

Given the maximum sequence lengths accepted by GPT-2-based models (length = 1024) and BERT-based models (length = 512) against each word-level gaze+EEG feature corresponding to ten (5+5) special tokens, it can be inferred that a portion of dataset samples was likely truncated from the text part situated at the end of the prompted sequence. This could lead to a negative impact on accuracy scores derived from prompting-based finetuning. This calls for Method 2 evaluated in the following section which avoids the constraints thoroughly by performing multidimensional prompting in the continuous embedding space of the LM.

Scaling up Backbone Model Lastly, unsurprisingly an increase in the size of the models, moving from BERT to RoBERTa-base and further to RoBERTa-large, leads to improved performance. This trend suggests that larger models may be better equipped to exploit the supplementary cognitive features for sentiment classification tasks. This observation corroborates one of the rationales of this study, which aims to extend cognition-inspired research to more advanced, large-scale, state-of-the-art language models.

Concerning the decoder-only models, although they also display performance enhancements with the addition of cognitive features, these improvements are less conspicuous. For instance, the standard GPT-2 model sees its baseline accuracy rise from 44.0% to 45.6% with the addition of gaze features and to 45.1% with the incorporation of EEG features. This subtle increase accentuates the necessity of cultivating more high-quality cognition corpora specifically designed for the adaptation of Natural Language Processing tasks.

6.2 Method 2

Method 2 focuses on decoder-only models whose text baselines yielded a considerably low performance regardless of the model size. Nevertheless, while an insufficient amount of data was shown to be unable to tune the GPT-2-based models to achieve an accuracy comparable to the encoder-only BERT-based models, Method 2 aims to examine whether employing a continuous soft prompting approach within the CogMAP framework could guide the language model (LM) and the projection layer to capture the relationship between cognitive prompting vectors, text embeddings, and the output labels, ultimately improving the classification performance.

As presented in Table 6.2 and illustrated in Figure 6.2, the results clearly demonstrate a significant increase in accuracy scores compared to the baseline models across all experimental settings. Notably, when incorporating gaze features as cognition prompting vectors, the performance surpasses that of the other two configurations of cognitive features for all variants of GPT-2. The accuracy significantly rises from 44% to 66.3% in GPT-2 (small), with a p-value of less than 0.001, indicating strong statistical significance.

These findings highlight the effectiveness of the CogMAP framework in leveraging gaze features as continuous soft prompts to enhance the classification performance of decoder-only models. The extent of enhancement surpassed previous works on other baselines ranging from null to %4 of accuracy growth on the same corpus (Barrett et al., 2018; Hollenstein et al., 2019; Mishra et al., 2017; Ren and Xiong, 2021). The significant improvements achieved in accuracy scores demonstrate the ability of the framework to guide the LM in effectively utilizing the relationship between cognitive prompts, text embeddings, and target labels.

Model	**Text-baseline**				**Gaze**				**EEG**				**Gaze + EEG**			
	Acc	**Prec**	**Rec**	**F1**	**Acc**	**Prec**	**Rec**	**F1**	**Acc**	**Prec**	**Rec**	**F1**	**Acc**	**Prec**	**Rec**	**F1**
GPT-2	44.0	57.5	44.0	44.4	63.0**	66.3	63.0	62.9	62.0**	64.4	62.0	62.5	60.5**	62.9	60.5	60.9
GPT-2-medium	48.5	50.3	48.5	47.3	51.5**	54.3	51.5	50.8	54.0**	56.0	54.0	53.7	59.0**	61.3	59.0	58.4
GPT-2-large	47.3	46.5	47.3	43.0	55.5**	57.8	55.5	55.7	57.0**	57.6	57.0	56.0	55.0**	53.8	55.0	53.4

Table 6.2 Results of the CogMAP framework: precision (P), recall (R) and F1-score (F1) for the ternary sentiment classification tasks augmented by gaze features, EEG features, and the combination of both as continuous soft prompts. The significance of accuracy as the metric is indicated with the asterisks: * = p<0.01, ** = p<0.001 (Bonferroni method).

While both Method 1 and Method 2 have successfully demonstrated the effectiveness of the prompt-based finetuning approach for integrating multimodal cognition data proposed in this study, there are some aspects of the results that render distinctive patterns worth discussing, as elaborated below.

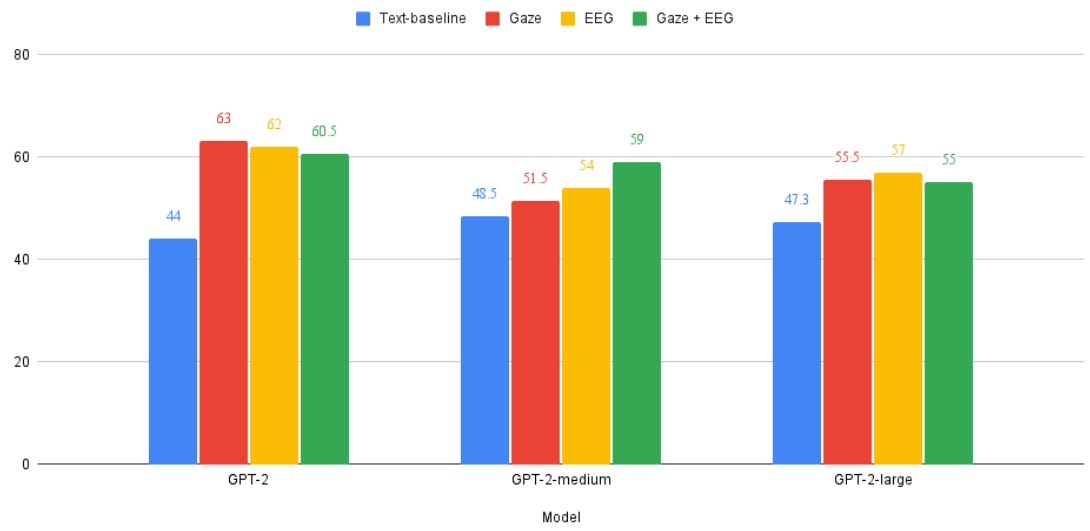


Fig. 6.2 Graph of accuracy scores across models extracted from 6.2

6.2.1 Single Feature Type vs Combined Features

In examining the contrasting evaluation of Methods 1 and 2 in the results table, there was observed a distinctive difference in performance when combining gaze and EEG features to generate prompting input. Method 1 demonstrates a decline in downstream performance in sentiment classification, ostensibly due to the excessive length of sequences that surpass the tokenizer's constraint. In contrast, Method 2, which circumvents this limitation by conducting prompt concatenation in the high-dimensional embedding space of the Language Model (LM), exhibits a significant enhancement in performance compared to the baselines ($p < 0.001$).

Moreover, the comparative analysis of the three experimental settings (gaze/EEG/gaze+EEG) reveals that the amalgamation of cognitive features does not consistently outperform single-feature prompting conditions, except in the case of GPT-medium models. These mixed outcomes are consistent with previous research, which has reported conflicting evidence regarding the efficacy of combining two types of cognitive features. This leads to intriguing questions: how can EEG signals be optimally preprocessed and more effectively denoised for NLP tasks? Furthermore, how can we amalgamate EEG and eye-tracking data (and potentially other cognitive processing signals or incidental data as per Plank, 2016) to enhance downstream NLP applications more effectively?

As we scrutinize the influence of model size, it is intriguing to note that escalating the size of backbone models from GPT-2 to GPT-2-large does not confer any improvements in test accuracy scores over smaller-sized variants across all three gaze/EEG/combination conditions. Two possible interpretations for this result are proposed.

Firstly, the unexpected decrease in accuracy as model size increases may be related to the changing input dimensions of the LM (i.e., 1024, 1280, 1600 as previously listed in Table 4.1). These changes affect the requirements for mapping cognitive vectors into higher-dimensional vectors in alignment with the token embedding of the LM. Given that the dimensions of original cognitive vectors remain constant, a larger model essentially implies a more complex task when projecting these vectors to a larger dimension, thereby complicating the update process for the weights of the projection layers. As a result, larger GPT variants may have a relatively more distorted projection module compared to smaller models with a lower input dimension.

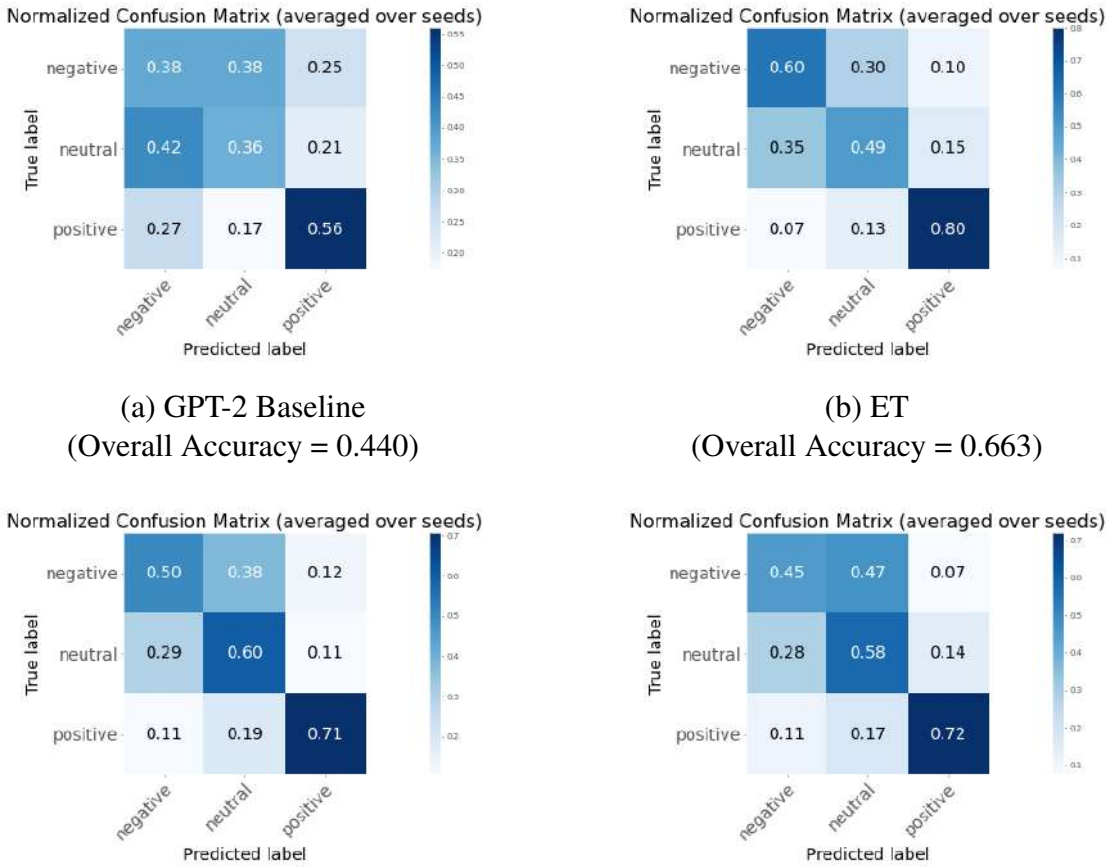
Secondly, it is plausible that the combination of two types of cognitive features might confuse the LM. If the LM cannot distinguish the metacognitive information within the high-dimensional data, it may be challenging for the model to differentiate between gaze and EEG data. Since these two types of data are projected into the same embedding space, the model might struggle to assimilate the distinct meanings of the two types of cognitive features, potentially perceiving them as 'too similar' during training (Hollenstein et al., 2019).

6.2.2 Class Prediction Analysis

This section conducted a more in-depth analysis of the Method 2 results to interpret which particular aspect of the model predictions has been enhanced by the cognition-prompting approach via the CogMAP method. As representative examples among the experimental conditions, Fig.6.3 and Fig. 6.4 demonstrate the confusion matrices of GPT-2 and GPT-2-medium trained on three types of cognitively-prompted texts (i.e., eye-tracking, EEG, eye-tracking + EEG), evaluated on the held-out testing dataset. Each matrix shows the probability of the model's class prediction (on the x-axis) against the ground truth value i.e., the label (on the y-axis). Hence, the top-down diagonal elements indicate the proportion or percentage of correctly classified instances for each class. The confusion matrix results were normalized and weighted beforehand to mitigate the impact of the small data size and potential class imbalance caused during the splitting of the small dataset in this study.

The major principle of the fine-grained findings from comparing the confusion matrices is that the cognition signals via the CogMAP method particularly enhanced the GPT-2 model's performances to predict negative and neural classes extensively, and effectively mitigate the LM's bias towards positive class prediction. The explanations and discussions about different types of cognition prompt inputs are broken down below.

From 6.3(a), it can be observed on the GPT-2 baseline trained on uni-modal texts that the model is able to predict relatively more correctly on the 'positive' sentiment sentences (= .56), while is rather weak in correctly distinguishing between negative (.38) and neutral



(a) GPT-2 Baseline
(Overall Accuracy = 0.440)

(b) ET
(Overall Accuracy = 0.663)

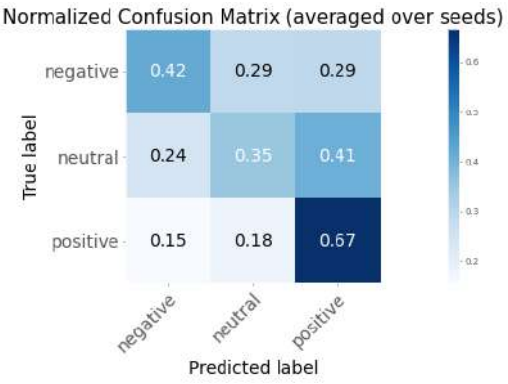
(c) EEG
(Overall Accuracy = 0.644)

(d) ET + EEG
(Overall Accuracy = 0.605)

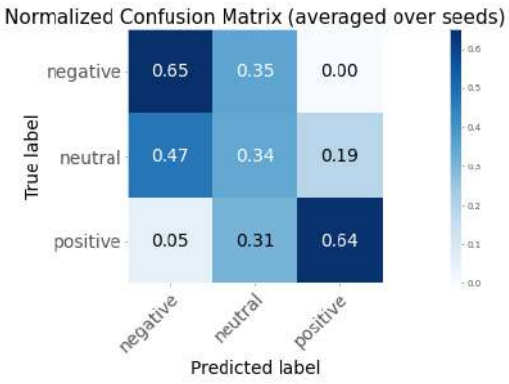
Fig. 6.3 Confusion matrices of GPT-2 evaluation in Method 2.

(.36) classes, a typical challenge of a multi-class sentiment classification task for language models. Such bias is consistently found in the GPT-2-medium baseline in 6.4(a) and the unshown GPT-2-large baseline. This suggests an intrinsic bias towards positive class prediction in GPT-2 series of models, which was reflected saliently when finetuned on a small dataset (i.e., simply 200 training samples). The similar biases inherently embedded in pre-trained language models are also reported in numerous sentiment classification studies. For example, Huang et al. (2020) and Garg et al. (2022) have investigated the highly variable sentiment bias of language models including BERT, RoBERTa, and GPT-2 even trained on a large balanced dataset like a complete SST-2. It is further analysed that token-level factors like gender, occupation, age, etc. can have a great undesired impact on the model predictions (e.g., 'baker' is biased towards a positive prediction while 'accountant' is biased towards a negative prediction (Huang et al., 2020)). In a more analogical setting of small datasets, Zhao et al. (2021) drew the same conclusion that the

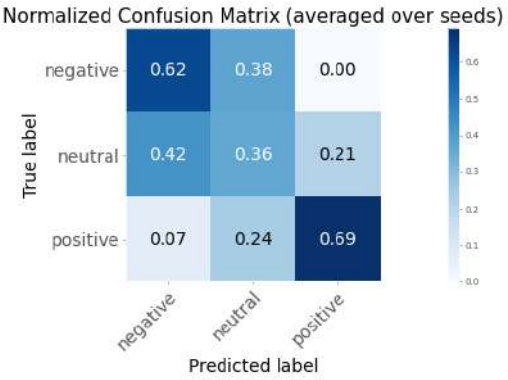
GPT-3's intrinsic biases cause it to frequently predict high confidence for the Positive class in few-shot learning using sample inputs from SST-2.



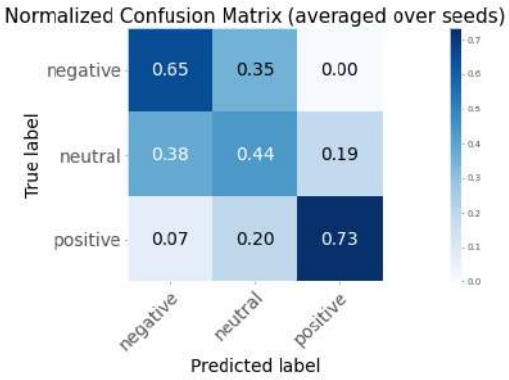
(a) GPT-2-medium Baseline
(Overall Accuracy = 0.485)



(b) ET
(Overall Accuracy = 0.515)



(c) EEG
(Overall Accuracy = 0.540)



(d) ET + EEG
(Overall Accuracy = 0.599)

Fig. 6.4 Confusion matrices of GPT-2-medium evaluation in Method 2.

In contrast, the CogMAP prompt-based fine-tuning method extensively increases prediction accuracy per class across all types of cognition prompts. On GPT-2 where the biggest facilitation on overall accuracy was reported among all model variants previously in Table 6.2, there is a significant rise of correct predictions in all three individual sentiment classes. In the larger GPT-2-medium model, aligned with the overall accuracy difference as previously discussed, the increase per class is moderately lower than GPT-2 and less persistent across three sentiments, mainly due to the exceptions of the 'neutral' class (baseline at 0.35 vs ET at 0.34, EEG at 0.36, ET+EEG at 0.44). However, a steady enhancement of correct prediction rates is saliently demonstrated around the prediction of the 'negative' class. It can be inferred here that the cognition prompts may have benefited

the language models during the training process by intrinsically curbing their sentiment bias towards a 'positive' prediction.

More specifically regarding the feasible mechanism, the gaze and EEG signals may have provided (a) word-level and (b) general information related to the sentiment of the ground-truth class label, as introduced in chapter 1 and 2: (a) on the word level, the humans' attention distribution towards certain tokens are reflected in both the eye-tracking and EEG data which could carry semantic and contextual information about the specific sentiment label (e.g., , higher signal values on specific words denote greater human processing attention or neural activation, indicating their relative importance for predicting the correct sentiment); (b) apart from learning the word-by-word association between the preceding cognition prompt and the text input, sentiment-related information can also be embedded in the whole sequence of cognition prompts as well (e.g., the electrode activation regions are different when processing long-range inputs of diverse emotions (Kroupi et al., 2011; Zhang et al., 2019)). During training, such information has been leveraged by the GPT-2 models as additional cues to regulate its processing behaviour, thus adjusting the classification threshold to achieve a more optimal performance.

Regarding the impact of different cognition features (eye-tacking vs EEG vs combination), an interesting pattern raised our attention: although EEG prompts generally worked less effectively than eye-tracking in enhancing the sentiment classification accuracy, this gap is reversed in the 'neutral' class in all GPT-2-based models. In both Fig6.3 and Fig6.4, EEG-prompting experiments showed the highest correction rates in classifying neutral sequences, among the other three types of input.

7

Limitations and Future Works

7.1 Limitations

There are several limitations of the current study that restrains its generalisability as listed below. The first two are commonly found as a challenge in the line of cognition-inspired research, while the third limitation discussing the numerical comprehension of LMs is a topic uniquely relevant to Method 1 from the cognition-prompting-and-finetuning framework proposed in this study. Future action points are also suggested for inspiration.

Reliance on cognition data The first major drawback of the current study, as well as a universal challenge in cognition-inspired NLP research, is the strict reliance on cognition data for the entire training and testing. It is not yet investigated how the cognition-prompted-and-finetuned models from the proposed two methods will perform on text-only input for evaluation. Same to the current study, this line of research was often evaluated on text input paired with corresponding cognition signals (and Casacuberta, 2022; Barrett et al., 2018; Mishra et al., 2017). Barrett et al. (2016); Hollenstein et al. (2019) also notified the issue and designed a word-type feature aggregation method to create a complete lexicon of words with their *averaged* gaze and EEG feature values over all word occurrences in the dataset. This approach took inspiration from the early word embedding research and is not applicable to the current study due to its high sensitivity to bias in the cognition dataset and our paradigm shift to pre-trained language models.

Future efforts are therefore to focus on evaluating and optimizing the cognition-enhanced model on text-only inputs. Considering the strength of prompts in steering away from overfitting and catastrophic forgetting, it is intriguing to examine whether the enhancement by the cognition-prompt-based finetuning method can transfer to the processing of text-only input. Nevertheless, if not, there are two potential options in further investigations for optimizing the prompting methods: firstly, training on other tasks that the LM is originally able to perform and implementing LM-fixed prompt learning instead

of prompt-based finetuning so that the LM weights will not be disdained by the cognition prompt input; secondly, training the model to reconstruct cognition signals from the text (to be elaborated in **??**, and in turn leverage them as prompts to facilitate other task-specific performances.

Single task & language Secondly, the current study focuses on solely one task i.e., the ternary sentiment classification task. To solidify and generalize the efficiency of the proposed two cognition-prompting methods, further studies can adapt the framework to other common downstream tasks. Considering the existing cognition corpora that can be utilized in future works, extended tasks can include relation classification (Hollenstein et al., 2018b, 2020b), named entity recognition (Colman et al., 2022), and text generation tasks (Sood et al., 2021).

The Numerical Reasoning Abilities of state-of-the-art LMs An under-investigated factor in this study concerning the prompt-based approach is the numerical reasoning ability of the deployed language models. Because the numerical values of cognition features are treated as either discrete tokens in Method 1, the model's understanding of the numerical data can influence how well it absorbs and leverages the information provided by the cognition prompt in the hard prompting paradigm.

7.2 Other Future Directions

Apart from the above limitations that call for future examination, discussion in this study also provides insights into several topics that inspired other directions of cognition-enhanced NLP research on a higher level.

7.2.1 More cognition signal corpora

As mentioned in the results section, the absolute performance level of the training models, whether baselines or cognition-prompted models, are limited by the dataset scale from ZuCo (%4 of the SST-5) (Hollenstein et al., 2020a). Because the goal is to compare the performance gap between textual baselines and cognition-prompted models, the lack of training data does not necessarily constitute a limitation that damages the rigorousness or validity of the current research. However, it is intriguing to explore to which extent the cognition signals can positively modify the pre-trained language models in large-scale data settings, especially on GPT-based generation models which require a sufficient amount of training data to perform classification tasks.

Hence, this urgently calls for more collections of cognition signals recording participants on NLP-task-related text materials (Hollenstein et al., 2020a).

7.2.2 Other types of cognition signals

Exploring the integration of other types of cognition signals into language models is also a worthy direction for future research. Apart from expanding the generalisability of the method, replacing eye-tracking and EEG with those cognition signals lacking high temporal resolution (e.g., fMRI) can also help solidify the fine-grained mechanisms of cognition-enhanced NLP research. More precisely, this study introduced two possibilities that cognition signals via prompt-based fine-tuning can facilitate language models in ternary sentiment classification (i.e., by providing word-level or sequence-level general cues).

Several studies have started exploring the utility of fMRI data in language models, but the scope is mainly limited to fMRI signals recorded from listening to speech and centred around understanding human cognition by making analogies from language models, rather than enhancing NLP applications (Caucheteux et al., 2023; Willems et al., 2016). In a newly submitted spotlight paper (not peer-reviewed yet), Takagi and Nishimoto (2022) claimed to reconstruct images from fMRI signals recorded when participants were presented using a vision-language model named Stable Diffusion. This study sparked heated discussions on the prospect of incorporating cognition signals in language models in generation tasks. Considering the particular strength of CogMAP when applied to generation models like GPT-2 exclusively experimented in this study, the thesis encourages future works to explore the adaptation of cognition-prompting to generation tasks.

7.3 Ethical Concerns

A range of cognition-related NLP research has commented on the underlying ethical issues from collecting and utilizing cognitive data (Hollenstein et al., 2020a,b; McGuire and Tomuro, 2021). These concerns encompass privacy issues due to potential subject identification, skewed representation and consequent standardization of certain demographic groups, as well as the potential for the perpetuation of ingrained human biases. Similarly, Sen et al. (2020) suggest the application of human attention supervision as a tool to evaluate the legitimacy of attention as a reliable, human-like rationale for decision-making within models. In parallel, Pruthi et al. (2019) highlight the potential for misdirection by manipulating attention to create an illusion of reduced bias within models. Future research could critically examine whether supervision based on human attention can serve as a foundation to investigate cognitive biases that may be acquired by these models. Furthermore,

this approach could also align attention-based explanations with model outcomes, enabling high-performing models to adhere to auditor expectations in a more reliable and faithful manner.

8

Conclusion

In conclusion, the integration of human cognitive signals as prompts in language models presents a promising approach to enhancing their performance on downstream NLP tasks. This thesis introduced two novel methodologies, Method 1 and Method 2, for incorporating cognition data into language models and evaluated their effectiveness on the task of ternary sentiment classification.

Method 1 explored the use of word-level and sentence-level prompts for introducing gaze and EEG features into language models. The results showed that word-level prompts disrupted the natural flow of sentences and hindered the models' ability to comprehend and interpret textual context. On the other hand, sentence-level prompts consistently improved the performance of both encoder-only BERT-based models and decoder-only GPT-2-based models. Gaze features were found to be more effective than EEG features as prompts, and the combination of both types of features did not yield better results than using a single feature type.

Method 2 employed a continuous soft prompting approach within the CogMAP framework to integrate gaze and EEG features into decoder-only GPT-2-based models. The results demonstrated a significant increase in accuracy compared to the text baselines across all experimental settings. Gaze features as soft prompts outperformed EEG features, and the combination of both types of features did not consistently outperform single-feature prompting conditions.

These findings highlight the effectiveness of the prompt-based fine-tuning approach in leveraging cognition data to enhance the performance of language models. The results also suggest the need for further research in optimizing the preprocessing and utilization of EEG signals, as well as exploring effective ways to combine multiple types of cognitive features.

Overall, this thesis contributes to the field of cognition-inspired NLP research by addressing limitations in current methodologies and providing a new framework for integrating cognitive signals into language models. It opens up new possibilities for

bridging the gap between human cognition and artificial language processing, improving the performance and understanding of language models.

References

, S. C.-P. and Casacuberta, F. (2022). Few-shot regularization to tackle catastrophic forgetting in multilingual machine translation. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 188–199, Orlando, USA. Association for Machine Translation in the Americas.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate.

Barrett, M., Bingel, J., Hollenstein, N., Rei, M., and Søgaard, A. (2018). Sequence classification with human attention. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, page 302–312, Brussels, Belgium. Association for Computational Linguistics.

Barrett, M., Bingel, J., Keller, F., and Søgaard, A. (2016). Weakly supervised part-of-speech tagging using eye-tracking data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics: Short Papers*, volume 579, page 584.

Barrett, M. and Hollenstein, N. (2020). Sequence labelling and sequence classification with gaze: Novel uses of eye-tracking data for natural language processing. *Language and Linguistics Compass*, 14(11):1–16.

Barrett, M. and Søgaard, A. (2015). Reading behavior predicts syntactic categories. In *Proceedings of the nineteenth conference on computational natural language learning*, pages 345–349.

Bensemann, J., Peng, A., Benavides-Prado, D., Chen, Y., Tan, N., Corballis, P. M., Riddle, P., and Witbrock, M. (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, Dublin, Ireland. Association for Computational Linguistics.

Bingel, J., Barrett, M., and Søgaard, A. (2016). Extracting token-level signals of syntactic processing from fmri-with an application to pos induction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 747–755.

Caucheteux, C., Gramfort, A., and King, J.-R. (2023). Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature human behaviour*, 7(3):430–441.

Clifton Jr, C., Staub, A., and Rayner, K. (2007). Eye movements in reading words and sentences. *Eye movements*, pages 341–371.

Colman, T., Fonteyne, M., Daems, J., Dirix, N., and Macken, L. (2022). GECO-MT: The ghent eye-tracking corpus of machine translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 29–38, Marseille, France. European Language Resources Association.

Cop, U., Dirix, N., Drieghe, D., and Duyck, W. (2017). Presenting geco: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, 49(2):602–615.

Degen, J., Kursat, L., and Leigh, D. D. (2021). Seeing is believing: testing an explicit linking assumption for visual world eye-tracking in psycholinguistics. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43.

Demberg, V. and Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.

Ding, X., Chen, B., Du, L., Qin, B., and Liu, T. (2022). Cogbert: Cognition-guided pre-trained language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, page 3210–3225, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Duggan, G. B. and Payne, S. J. (2011). Skim reading by satisficing: evidence from eye tracking. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1141–1150.

Fitzsimmons, G., Weal, M. J., and Drieghe, D. (2014). Skim reading: An adaptive strategy for reading on the web. In *Proceedings of the 2014 ACM Conference on Web Science*, WebSci '14, page 211–219, New York, NY, USA. Association for Computing Machinery.

Garg, A., Srivastava, D., Xu, Z., and Huang, L. (2022). Identifying and measuring token-level sentiment bias in pre-trained language models with prompts. *arXiv preprint arXiv:2204.07289*.

Gonzalez-Garduno, A. and Søgaard, A. (2018). Learning to predict readability using eye-movement data from natives and learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Hao, Y., Dong, L., Bao, H., Xu, K., and Wei, F. (2021). Learning to sample replacements for ELECTRA pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4495–4506, Online. Association for Computational Linguistics.

He, P., Liu, X., Gao, J., and Chen, W. (2021). {DEBERTA}: {DECODING}-{enhanced} {bert} {with} {disentangled} {attention}. In *International Conference on Learning Representations*.

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

Hollenstein, N., Barrett, M., and Beinborn, L. (2020a). Towards best practices for leveraging human language processing signals for natural language processing. In *Proceedings of the Second Workshop on Linguistic and Neurocognitive Resources*, page 15–27, Marseille, France. European Language Resources Association.

Hollenstein, N., Barrett, M., Troendle, M., Bigiolli, F., Langer, N., and Zhang, C. (2019). Advancing nlp with cognitive language processing signals. (arXiv:1904.02682). arXiv:1904.02682 [cs].

Hollenstein, N. and Beinborn, L. (2021). Relative importance in sentence processing. *CoRR*, abs/2106.03471.

Hollenstein, N., Pirovano, F., Zhang, C., Jäger, L., and Beinborn, L. (2021). Multilingual language models predict human reading behavior. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 106–123, Online. Association for Computational Linguistics.

Hollenstein, N., Rotsztejn, J., Troendle, M., Pedroni, A., Zhang, C., and Langer, N. (2018a). Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading. *Scientific Data*, 5(1):180291.

Hollenstein, N., Rotsztejn, J., Troendle, M., Pedroni, A., Zhang, C., and Langer, N. (2018b). Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading. *Scientific data*, 5(1):1–13.

Hollenstein, N., Troendle, M., Zhang, C., and Langer, N. (2020b). ZuCo 2.0: A dataset of physiological recordings during natural reading and annotation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 138–146, Marseille, France. European Language Resources Association.

Hollenstein, N., Troendle, M., Zhang, C., and Langer, N. (2020c). Zuco 2.0: A dataset of physiological recordings during natural reading and annotation. (arXiv:1912.00903). arXiv:1912.00903 [cs].

Hollenstein, N. and Zhang, C. (2019). Entity recognition at first sight: Improving ner with eye movement information. (arXiv:1902.10068). arXiv:1902.10068 [cs].

Huang, P., Zhang, H., Jiang, R., Stanforth, R., Welbl, J., Rae, J., Maini, V., Yogatama, D., and Kohli, P. (2020). Reducing sentiment bias in language models via counterfactual evaluation. In Cohn, T., He, Y., and Liu, Y., editors, *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 65–83. Association for Computational Linguistics.

Ikhwantri, F., Putra, J. W. G., Yamada, H., and Tokunaga, T. (2023). Looking deep in the eyes: Investigating interpretation methods for neural models on reading tasks using human eye-movement behaviour. *Information Processing Management*, 60(2):103195.

Johnston, W. A. and Dark, V. J. (1986). Selective attention. *Annual review of psychology*, 37(1):43–75.

Kaiser, E. (2013). Experimental paradigms in psycholinguistics. *Research methods in linguistics*, pages 135–168.

Koh, J. Y., Salakhutdinov, R., and Fried, D. (2023). Grounding language models to images for multimodal inputs and outputs.

Kroupi, E., Yazdani, A., and Ebrahimi, T. (2011). Eeg correlates of different emotional states elicited during watching music videos. In *Affective Computing and Intelligent Interaction: Fourth International Conference, ACII 2011, Memphis, TN, USA, October 9–12, 2011, Proceedings, Part II*, pages 457–466. Springer.

Li, X. L. and Liang, P. (2021). Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Mañas, O., Rodriguez, P., Ahmadi, S., Nematzadeh, A., Goyal, Y., and Agrawal, A. (2023). Mapl: Parameter-efficient adaptation of unimodal pre-trained models for vision-language few-shot prompting.

McGuire, E. and Tomuro, N. (2021). Relation classification with cognitive attention supervision. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 222–232, Online. Association for Computational Linguistics.

Mishra, A., Kanojia, D., Nagar, S., Dey, K., and Bhattacharyya, P. (2017). Leveraging cognitive features for sentiment analysis. *CoRR*, abs/1701.05581.

Muttenthaler, L., Hollenstein, N., and Barrett, M. (2020). Human brain activity for machine attention.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners.

Rayner, K. (1977). Visual attention in reading: Eye movements reflect cognitive processes. *Memory & cognition*, 5(4):443–448.

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

Rayner, K., Warren, T., Juhasz, B. J., and Liversedge, S. P. (2004). The effect of plausibility on eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(6):1290.

Ren, Y. and Xiong, D. (2021). Cogalign: Learning to align textual neural representations to cognitive language processing signals. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, page 3758–3769, Online. Association for Computational Linguistics.

Ren, Y. and Xiong, D. (2022). Bridging between cognitive processing signals and linguistic features via a unified attentional network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):49–58.

Ribeiro, T., Brandl, S., Søgaard, A., and Hollenstein, N. (2023). Webqamgaze: A multilingual webcam eye-tracking-while-reading dataset. *arXiv preprint arXiv:2303.17876*.

Sartakhti, M. S., Etezadi, R., and Shamsfard, M. (2021). Improving persian relation extraction models by data augmentation. In *Proceedings of the Second International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2021) co-located with ICNLSP 2021*, pages 32–37, Trento, Italy. Association for Computational Linguistics.

Scarborough, H., Fletcher-Campbell, F., Soler, J., and Reid, G. (2009). Connecting early language and literacy to later reading (dis)abilities: Evidence, theory, and practice. *Approaching Difficulties in Literacy Development: Assessment, Pedagogy, and Programmes*, pages 23–39.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Shin, T., Razeghi, Y., Logan IV, R. L., Wallace, E., and Singh, S. (2020). Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.

Shoka, A., Dessouky, M., El-Sherbeny, A., and El-Sayed, A. (2019). Literature review on eeg preprocessing, feature extraction, and classifications techniques. *Menoufia J. Electron. Eng. Res*, 28(1):292–299.

Sood, E., Kögel, F., Strohm, F., Dhar, P., and Bulling, A. (2021). Vqa-mhug: A gaze dataset to study multimodal neural attention in visual question answering. *arXiv preprint arXiv:2109.13116*.

Sood, E., Tannert, S., Frassinelli, D., Bulling, A., and Vu, N. T. (2020a). Interpreting attention models with human visual attention in machine reading comprehension. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, page 12–25, Online. Association for Computational Linguistics.

Sood, E., Tannert, S., Mueller, P., and Bulling, A. (2020b). Improving natural language processing tasks with human gaze-guided neural attention. In *Advances in Neural Information Processing Systems*, volume 33, page 6327–6341. Curran Associates, Inc.

Strzyz, M., Vilares, D., and Gómez-Rodríguez, C. (2019). Towards making a dependency parser see. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1500–1506. Association for Computational Linguistics.

Sui, L., Dirix, N., Woumans, E., and Duyck, W. (2022). Geco-cn: Ghent eye-tracking corpus of sentence reading for chinese-english bilinguals. *Behavior Research Methods*, pages 1–21.

Takagi, Y. and Nishimoto, S. (2022). High-resolution image reconstruction with latent diffusion models from human brain activity. *bioRxiv*.

Tamkin, A., Brundage, M., Clark, J., and Ganguli, D. (2021). Understanding the capabilities, limitations, and societal impact of large language models. *arXiv preprint arXiv:2102.02503*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

Wallace, E., Feng, S., Kandpal, N., Gardner, M., and Singh, S. (2019a). Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2153–2162, Hong Kong, China. Association for Computational Linguistics.

Wallace, E., Wang, Y., Li, S., Singh, S., and Gardner, M. (2019b). Do nlp models know numbers? probing numeracy in embeddings. *arXiv preprint arXiv:1909.07940*.

Weiss, S. and Mueller, H. M. (2003). The contribution of eeg coherence to the investigation of language. *Brain and language*, 85(2):325–343.

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., and Van den Bosch, A. (2016). Prediction during natural language comprehension. *Cerebral Cortex*, 26(6):2506–2516.

Winkler, I., Haufe, S., and Tangermann, M. (2011). Automatic classification of artifactual ica-components for artifact removal in eeg signals. *Behavioral and brain functions*, 7:1–15.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Yi, K., Guo, Y., Jiang, W., Wang, Z., and Sun, L. (2020). A dataset for exploring gaze behaviors in text summarization. In *Proceedings of the 11th International Conference on Multimedia Systems*, MMSys '20, Istanbul, Turkey. ACM.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., and Zettlemoyer, L. (2022). Opt: Open pre-trained transformer language models.

Zhang, T., Zheng, W., Cui, Z., Zong, Y., and Li, Y. (2019). Spatial–temporal recurrent neural network for emotion recognition. *IEEE Transactions on Cybernetics*, 49(3):839–847.

Zhao, Z., Wallace, E., Feng, S., Klein, D., and Singh, S. (2021). Calibrate before use: Improving few-shot performance of language models. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

List of figures

List of tables

A

Model Variants and Hyperparameter Configuration

Table A.1 The list of deployed language models in the study, with training hyperparameters (lr: learning rate; bs: batch size; epoch: training epochs). Experiments were run over 5 seeds [16, 17, 18, 19, 20]. The max sequence lengths are set as default.

	Text-baseline			Gaze			EEG			Gaze + EEG		
Model	lr	bs	epoch	lr	bs	epoch	lr	bs	epoch	lr	bs	epoch
Method 1												
BERT	3e-5	16	30	1e-4	16	30	1e-4	16	30	1e-4	16	30
RoBERTa-base	1e-5	16	30	1e-5	16	30	1e-5	16	30	1e-5	16	30
RoBERTa-large	1e-5	16	30	1e-5	16	30	1e-5	16	30	1e-5	16	30
GPT-2	3e-5	16	30	3e-5	16	30	3e-5	16	30	3e-5	16	30
GPT-2-medium	3e-5	16	30	3e-5	16	30	3e-5	16	30	3e-5	16	30
GPT-2-large	1e-5	16	30	1e-5	16	30	1e-5	16	30	1e-5	16	30
Method 2												
GPT-2	3e-5	16	30	3e-5	16	30	5e-5	16	30	5e-5	16	30
GPT-2-medium	3e-5	16	30	3e-5	16	30	5e-5	16	30	5e-5	16	30
GPT-2-large	1e-5	16	30	5e-5	16	30	5e-5	16	30	5e-5	16	30

B

Dataset Details

Table B.1 Descriptive statistics of reading materials (M = mean, SD = standard deviation, R = range). The 400 selected sentences are comprised of 123 neutral, 137 negative and 140 positive sentences.

Task 1 Normal reading (Sentiment)			
Total words	7079		
Word types	3080		
Total sentences	400		
	M	SD	R
Words per sentence	17.7	8.29	3–43
Word length	6.97	2.71	1–26

Table B.2 Details of all subjects in the study from ZuCo (Hollenstein et al., 2018b). These scores are the percentages of correctly answered control questions in the respective task. Reading speed is measured in seconds (with standard deviation in brackets). The vocabulary and language proficiency of the participants was tested with the LexTALE test (Lexical Test for Advanced Learners of English), an unspeeded lexical decision task, which is for intermediate to highly proficient language users.

Subject ID	Age	Gender	LexTale	Reading Speed	Score
ZKW	25	Female	96.25%	6.94	69.57%
ZDN	32	Male	97.50%	3.91	89.13%
ZPH	26	Male	97.50%	4.78	89.13%
ZMG	51	Male	100.00%	4.39	91.30%
ZAB	41	Female	100.00%	4.88	76.09%
ZJN	51	Female	97.50%	8.71	54.34%
ZKH	41	Female	81.25%	5.42	76.09%
ZGW	49	Male	91.25%	6.87	71.74%
ZJS	42	Male	97.50%	4.34	91.30%
ZKB	26	Female	100.00%	5.39	89.13%
ZDM	25	Male	100.00%	4.41	76.09%
ZJM	41	Male	77.50%	6.22	80.43%
Average	38	-	94.69%	5.52	79.53%