High-level Speech Processing During Mind-Wandering:

Evidence from Neural Alignment with Language Models

Gal R. Chen[1], Rachel Finkelstein[2], Ariel Goldstein[3], Ran R. Hassin[1,4], Leon Y. Deouell[1,2]

[1] Department of Psychology, The Hebrew University of Jerusalem

[2] Edmond and Lily Safra Center for Brain Sciences, The Hebrew University of Jerusalem

[3] Department of Cognitive and Brain Sciences and Business School, The Hebrew University of Jerusalem

[4] The Federmann Center for the Study of Rationality, The Hebrew University of Jerusalem

*Corresponding author: Gal R. Chen

Email: galrefa.chen@mail.huji.ac.il

Author Contributions: G.R.C, R.R.H and L.Y.D conceptualized and designed research, G.R.C and R.F collected data and created software & materials, G.R.C. analyzed the data, A.G contributed new analytic tools, G.R.C, R.R.H, and L.Y.D. wrote the paper.

Competing Interest Statement: L.Y.D. is the co-founder and shareholder of and receives compensation for consultation from InnerEye, a startup neurotech company. The company business is not related to the current study. All authors declare no competing financial interests.

Classification: Biological Sciences (Psychological and Cognitive Sciences)

Keywords: Mind-wandering; Predictive processing; Large language Models; Consciousness; Speech

This file includes:

Main Text

Figures 1 to 6

Abstract

Speech processing involves a subjective sense of engagement with content. This experience occurs alongside continuous prediction of upcoming contents, raising questions regarding the potential relationship between these phenomena. Is the activation of predictive mechanisms sufficient for subjective experience, or can it proceed when the mind wanders away? Participants (N = 25) listened to audiobook content (>12,000 words) while providing self-reports of mind-wandering. Such reports were associated with spectral changes in the EEG signal and decreased early response to word onset, in line with current accounts of mind-wandering. In contrast, markers of word-level contextual surprise (indexed by the contextual surprise of Language Models) remained intact during mind-wandering, alongside significant encoding of semantic contents. Last, we measured the neural encoding of predictive context, measured by the correlation between EEG and the vectorial embeddings that Language Models use to predict each word given prior context. This encoding also remained strong during mind-wandering, albeit weaker compared to attentive periods. Our findings show that the brain predicts and monitors relevant inputs even when we disengage from the external environment. It suggests that the shared computational mechanisms between humans and Language Models are insufficient for experiencing the meaning of speech contents, with implications for theoretical accounts of mind-wandering and predictive processing accounts of consciousness.

Significance Statement

The human brain is suggested to process continuous speech using context based on previous information. Considering that these predictions involve the information required for extracting meaning, they might also be associated with the subjective experience of attentive listening. Using EEG recordings of participants listening to audiobooks and LLM-based representations of the text, we found that when people report mind-wandering, their brains continue representing context-based predictions, prediction errors, and word-level semantics. This indicates that sophisticated language predictions, that share computational mechanisms with LLMs, are at least partially independent from subjective awareness, with potential implications for debates about artificial intelligence and predictive processing theories of consciousness.

Main Text

Introduction

Spoken language processing is a key aspect of daily experience, making its neural and cognitive underpinnings a primary topic of interest for psychologists and cognitive neuroscientists[1–3]. Importantly, speech processing involves not only the mechanical decoding of acoustic inputs, but also the conscious, subjective experience of understanding and engaging with meaning[4,5]. Despite the progress made in understanding the basis of speech processing, it remains unclear which sub-components of speech processing are associated with the subjective experience of engaged listening, and which can occur when consciousness is reduced. This question is especially relevant today, as the capabilities of deep learning models spark discussions regarding whether they can ultimately meet the criteria for consciousness[6–8].

Speech processing happens at multiple levels, including the physical properties of the stimulus[9], linguistic units, word meaning[10,11], and whole narratives[12]. An additional important component of language processing, on which we focus here, is the high-level use of contextual information (e.g., previous words) for predicting upcoming inputs and extracting word meaning[13]. As a narrow example, context provides a prior for deciphering the meaning of homophones (for example, when restoring missing speech sounds[14]). However, the importance of predictions in language comprehension is suggested to be much broader[15–18], with contextual information influencing hierarchically lower levels of prediction errors[19]. The theoretical and empirical importance of predictions raises a central question: Do high-level, context-based predictions co-occur with conscious experience? Or, alternatively, can we find instances where consciousness of speech content is attenuated, yet predictive and semantic processing remain strong?

Here, we utilize spontaneous fluctuations in attention to speech (i.e., mind-wandering[20]) as a naturalistic case of reduced consciousness of relevant and audible external contents. The reader will likely recognize the experience of listening to a lecture or podcast, only to realize at some point that they have no idea what was being said, having been absorbed in a stream of related or unrelated thoughts. Mind-wandering is usually measured as self-reported periods in which thought disengages from the current task[21], prompting descriptions of processing during Mind-wandering as “perceptual decoupling,” in which perception becomes weaker, akin to ignored streams of information[22,23]. This description gains substantial support from studies

showing decreased neural responsivity and behavioral performance during mind-wandering[24–27].

How deep is this perceptual decoupling? The literature on processing without awareness suggests that there is some evidence for high-level processing of singular objects[28–30], with integration being more difficult[31]. However, such accounts are usually based on visual masking[28,32,33] or concurrent competition (e.g., dichotic listening[34,35]), limiting the conclusions to rather weak signals. Mind-wandering reports, on the other hand, reflect reduced awareness of contents that may be presented without any degradation. Further, the frequency and robustness with which mind-wandering occurs during task performance[26,36] and daily lives[37,38] make it unlikely that humans could behave adaptively if mind-wandering completely obliterated high-level processing capacities[39,40] (e.g., representation of lexical and semantic information). In the specific case of continuous speech, for example, partial monitoring of goal-relevant content might allow for better allocation of resources during mind-wandering. It is possible, therefore, that the brain may still track and predict the meaning of single words, despite the reduction in subjective experience.

Here, we set out to examine the relationship between predictive and semantic processing and mind-wandering in participants listening to audiobook chapters. We utilized previously established measures of pre-word context-related activity and post-word contextual surprise[19,41] during self-reported periods of mind-wandering (MW) and On-task (OT). To quantify word-based contextual surprise, we utilized the probability of word occurrence assigned to each word by a Large Language Model (LLM[42,43]). As a measure of non-contextual semantic processing, we examine neural encoding of word2vec[44] embeddings, which reflect semantic dimensions at the word level[45]. Last, for context-related activity, we assessed the neural encoding of the representations that LLMs use to generate predictions[15]. To foreshadow our results, we found that despite the decrease in response to word onset during MW, the other markers - pre-onset contextual activity, post-onset semantic encoding and surprise processing - remained highly significant and even unchanged during MW. Our findings are the first to show high-level verbal information processing, including prediction generation, during mind-wandering.

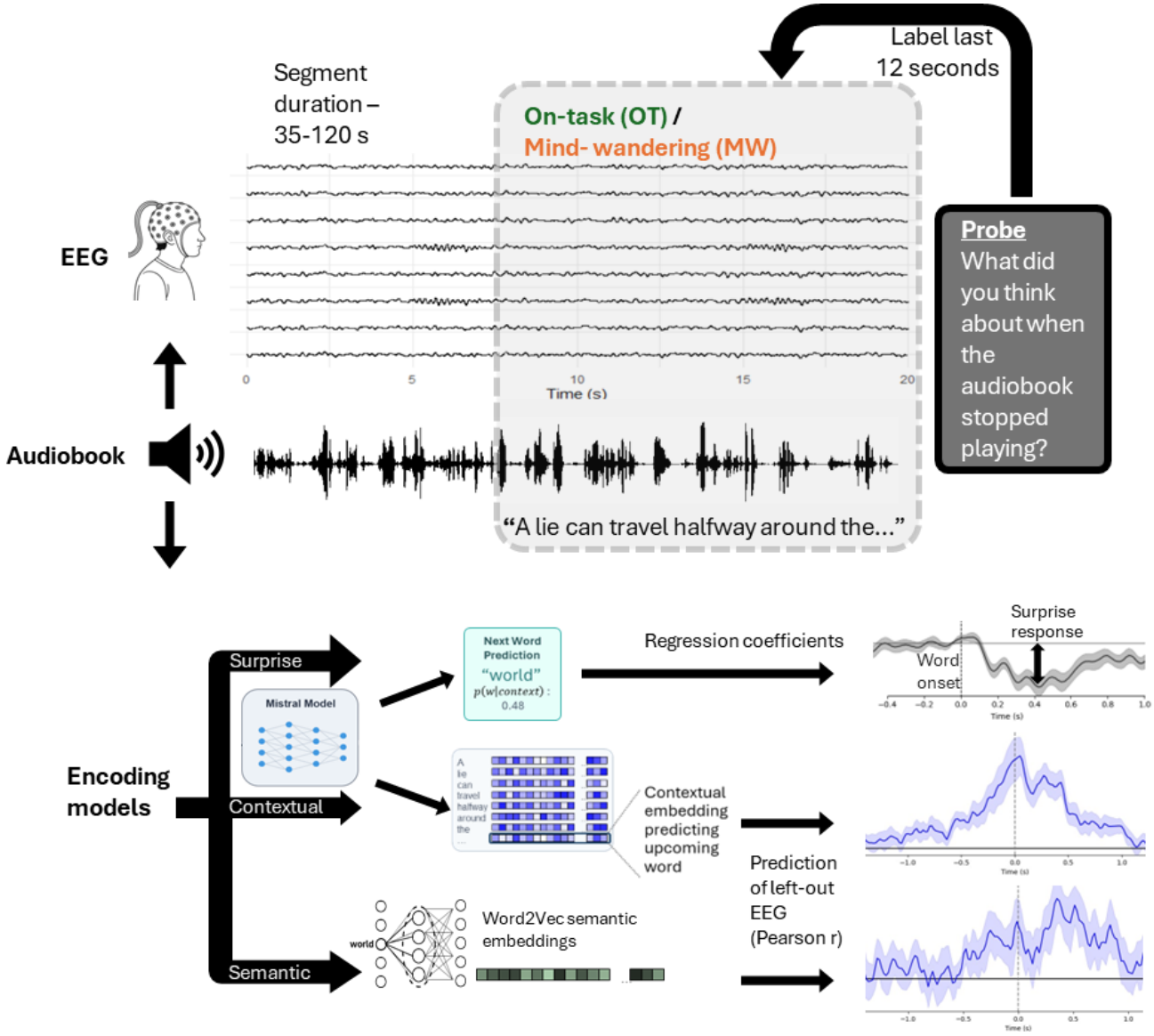

Figure 1. Experimental procedure and analysis approach. EEG was recorded from 25 subjects listening to six audiobook chapters, segmented into 136 parts varying in duration. The last 12 seconds of each segment were labeled according to a report in a post-segment probe. For each word played, we modeled three types of responses: change in response to surprising words (based on LLM), encoding of predictive context (based on LLM), and encoding of semantic embedding (based Word2Vec). Each model was fitted separately. For surprise response, we focused on time-resolved coefficients estimating responses (separating MW/OT periods), and for semantic & contextual encoding we tested left-out EEG data labeled as MW/OT (see Methods).

Results

Our participants listened to 130 minutes of Hebrew audiobooks, comprised of six chapters, with random pauses in which they were asked to choose descriptions that best fits their experience. Using validated methods for experience sampling[36], the choice "I was attentive to the audiobook", was labeled as "on-task" (OT), while the rest of the options reflected different types of mind-wandering (all labeled MW, see Table S1). The MW-associated options were chosen in

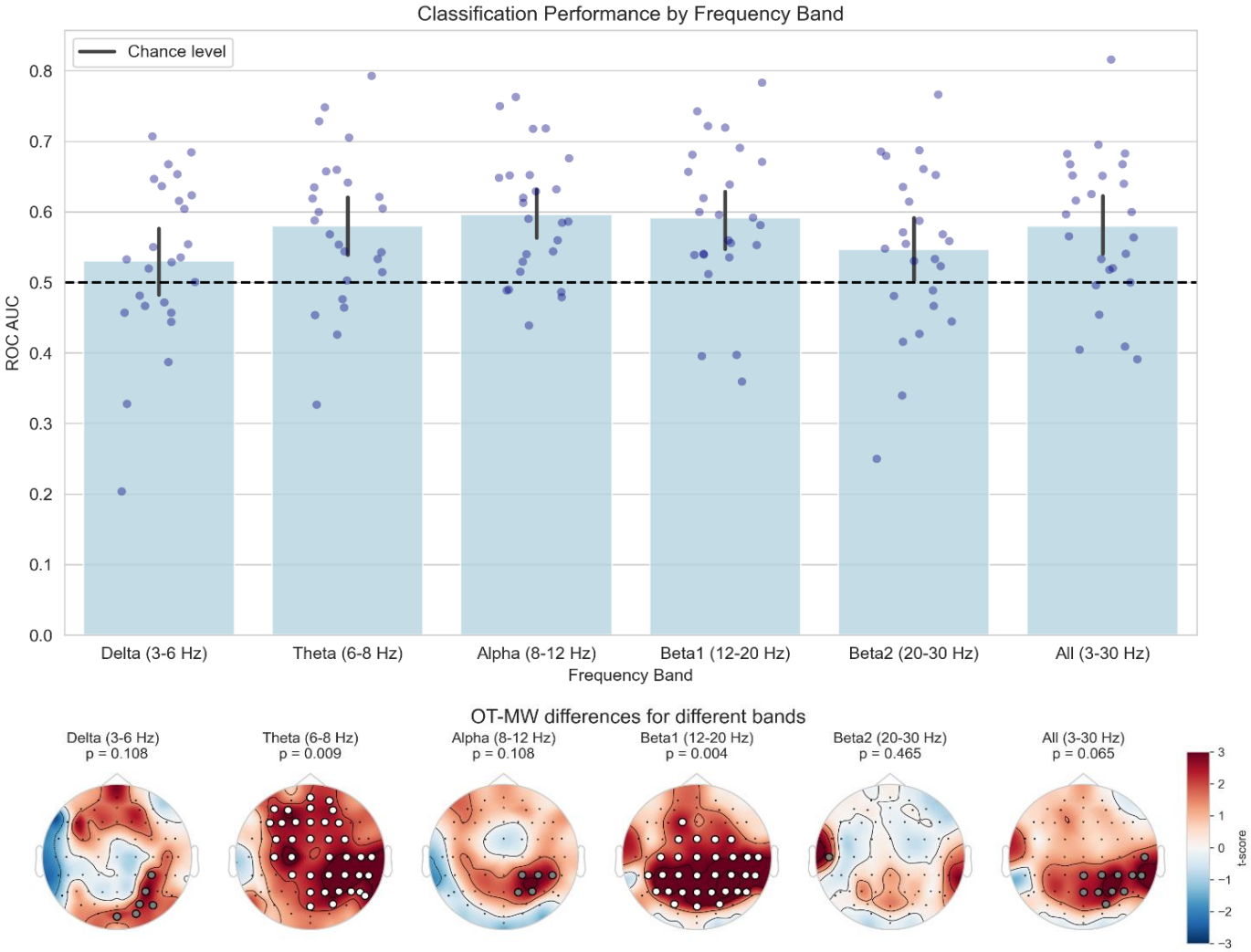

Figure 2. Mind-wandering classification. Top panel shows the results of single-epoch (12 seconds) classification based on the average power in the epochs. The classification was based on a 10-fold cross-validation, with the average power in each band across all 64 channels. Error bars denote 95% confidence interval. Bottom panel shows the average difference in power between OT and MW reports, split between different bands. White circles denote channels belong to a significant (corrected $p < .05$) cluster after spatial cluster-based testing. Grey circles show the maximum cluster (uncorrected $p < .05$) for non-significant tests.

28.1% of the time (SD = 10.0%) and were tightly associated with post-listening ratings of interest at the audiobook level (Figure S1). We labeled the 12-second period prior to report as MW/OT (for similar approaches see review by Kam et al.[27]). We begin by validating that MW reports reflect different states, as evidenced by changes in EEG spectra. Then, we move to examine semantic encoding, which is required for creating high-level predictions. The last two sections include an analysis of evoked responses to LLM-based surprisal and the encoding of pre-word context-related activity, both of which directly assess predictive processing during MW.

Mind-wandering is Associated with Spectral Changes

We first tested whether the EEG signal includes any information regarding MW reports, to ensure their reliability. Therefore, we examined whether mind-wandering reports can be predicted by the spectral properties of the EEG, as shown in previous studies[27,46,47]. We split the Power Spectrum Density (PSD) in each data segment (MW/OT segment) to familiar EEG bands (delta: 3-6Hz, theta: 6-8Hz, alpha: 8-12 Hz, low beta: 12-20Hz, and high beta: 20-30 Hz, and the average broadband power between 3-30). A cluster-based spatial permutation analysis of the power difference MW and OT (per band) revealed significant clusters of stronger power during the OT condition in the theta and low beta regime (Figure 2, lower panel). Beyond group differences, we were able to use the whole topography to classify MW in all bands except for the delta and high beta band (all Bonferroni corrected $p < 0.01$, see Figure 2, upper panel, and Table S3). While discussing the specific role of periodic activity in MW is outside the scope of this study, we note that both theta and beta bands are typically associated with sustained attention and executive control[48,49].

Having validated that EEG activity is sensitive enough to predict subjective reports, we now proceed to examine word-level markers of semantic processing and contextual predictions during OT and MW periods.

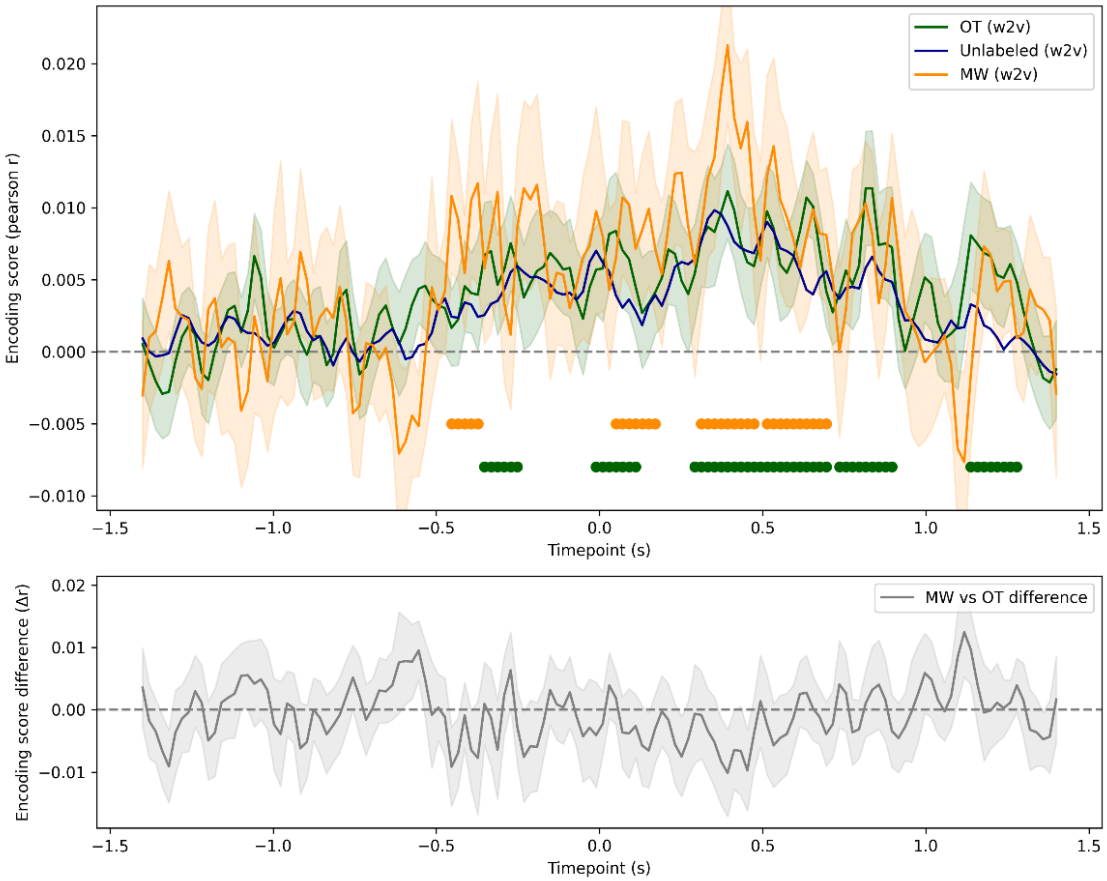

Figure 3. Results of average encoding scores for word2vec semantic representations per time-point. Green and orange dots denote $p < .05$ after cluster-based temporal permutation test. The grey line in bottom panel reflects the difference between OT and MW accuracy per timepoint. Shaded areas denote 1 SE.

Semantic encoding persists during mind-wandering

A precondition for intact contextual processing is semantic processing of spoken words, which is unclear whether MW abolishes completely. To examine whether such processing occurs when participants report MW, we investigated whether a model based on semantic properties of words can predict EEG responses left out of training. This was tested via an encoding model of existing word2vec 100-dimensional embeddings, pre-trained on a Wikipedia corpus[50]. We first extracted single-word epochs (from sections not labeled as MW/OT). Then, for each channel and timepoint (separately), we fitted a linear model predicting the EEG voltage with semantic embeddings as features. We then tested this model on epochs labeled as MW or OT (outside of the training set), resulting in predicted voltage at each channel and

timepoint. The average correlation between predicted and actual EEG, across channels, served as a marker for semantic encoding (Figure 3).

Similar to previous studies[15], encoding accuracy peaked around 400 ms, associated with the time needed for lexical/semantic access[51]. Semantic encoding was significant during both MW (significant clusters between 332-473 ms, permutation $p < 0.01$, mean encoding between 300-500 ms, r = 0.0136, SD = 0.0118) and OT periods (significant clusters between 292-694 ms, permutation $p < 0.01$, mean encoding between 300-500 ms, r = 0.0083, SD = 0.009). Specifically, semantic encoding during MW was not reduced compared to OT periods and, in fact, was numerically larger (temporal cluster-permutation, all $p > .7$; uncorrected difference between 300-500 ms: $t(24) = 1.63$, $p = .11$). Bayes factor analysis provided evidence against the hypothesis that OT encoding is stronger (BF10 = 0.089).

Mind-wandering reduced the response to word onset but not for surprise

So far, we have shown that semantic information is correlated with brain activity during MW, making this information available for generating predictions about upcoming words. Such predictions should, in turn, be reflected in the evoked response to the words. We focus on two types of responses: first, the average response to a single word, capturing low-level processes that occur for all words. Second, the change in response when a certain word is more surprising. Both responses could be captured by the intercept and slope of a simple regression model predicting EEG activity using surprise score (quantified by z-scoring the odds of a word appearing, based on LLM). To obtain these estimates, we conducted a time-resolved regression estimating regression-based ERP (rERP), following an established method[52]. This method controls for overlapping activity evoked by different words, while also allowing to model the activity each timepoint as a function of multiple variables. The regression included three separate predictors for word onset (MW, OT, and, for periods not reflecting the twelve seconds preceding probes - unlabeled) and three similar predictors for surprise, while also including covariates for acoustic & lexical information and pre-word uncertainty. (see Methods for additional details).

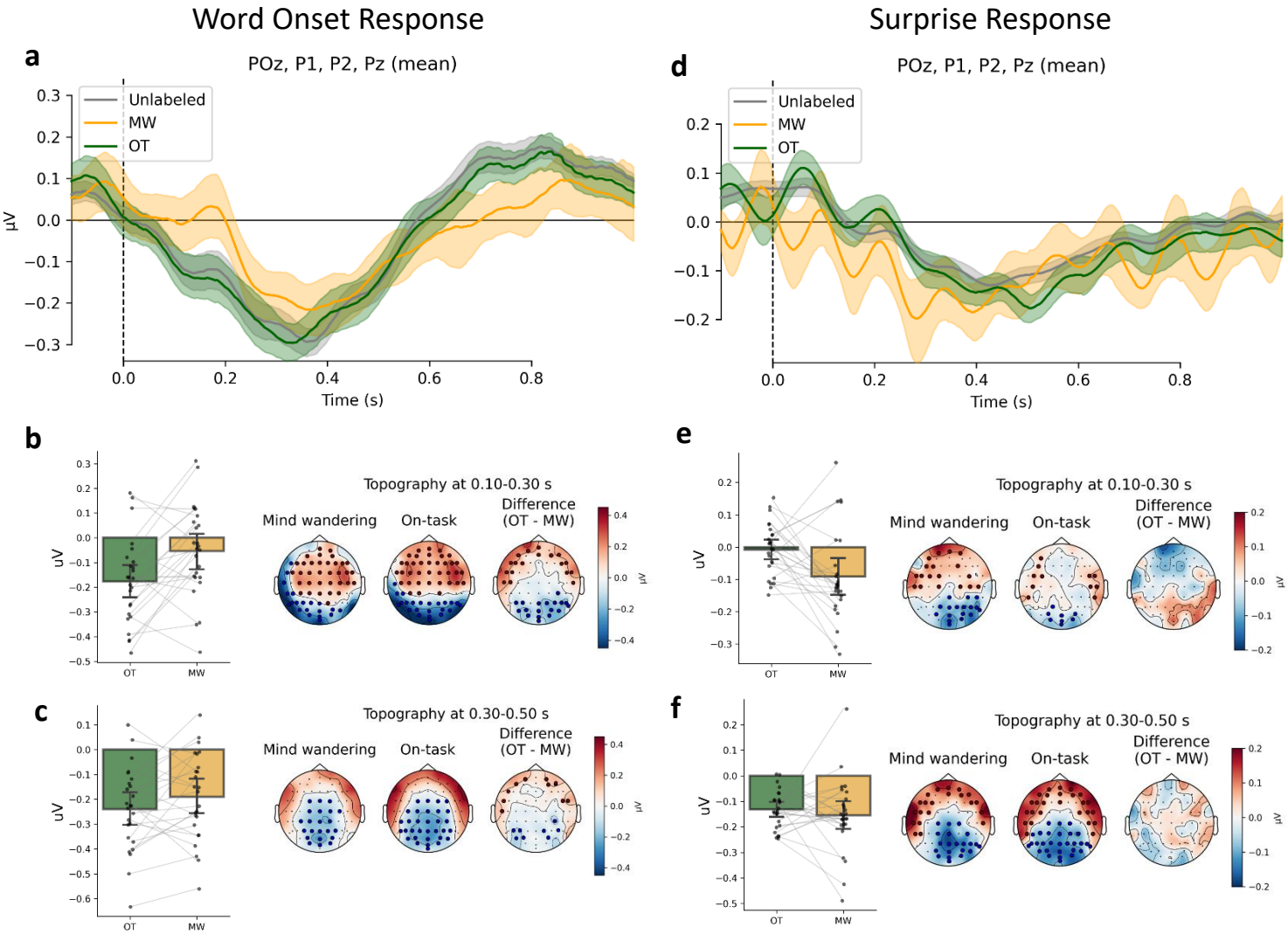

Figure 4. Effects of word onset and surprise level (a) Grand average of rERPs for word onset predictor (binary) at POz per attentional state. Shaded areas denote one SE of the mean. Time 0 reflects the onset of the word. (b) The effect of MW on onset response. The left panel (bar graph) shows the average estimated word onset response at the posterior cluster for the 100-300 ms time window for OT and MW words, with individual lines reflecting single-subject estimates and error bars reflecting 95% CI. The right panel shows the scalp topography at the same timepoints for each state and their differential activity. Red and blue dots mark channels belonging to a significant positive (OT>MW) and negative (OT<MW) clusters, respectively, during at least 20% of the time period ($p < 0.05$). (c) same as b, but for the 300-500 ms time window. Panels d-f are identical to a-c, but for the coefficients reflecting the surprise response (timepoint-specific slopes of a z-scored predictor), rather than the word onset response.

184 ***Cluster-based analysis of onset response.*** To test the rERP responses while avoiding

assumptions about time and location, the full waveforms in all channels were subjected to spatial-temporal permutation testing. Notably, both MW and OT showed a significant response to word onset (MW: between 0-1000 ms, permutation $p < .001$; OT: between 0-949 ms, permutation $p < .001$), with a significant cluster for OT versus MW appearing during 51-379 ms (permutation $p = .007$, Figure 4a). To gain further insight into the nature of differences, we conducted an exploratory analysis using the average onset response at two different windows, possibly separating earlier processes from later, post-perceptual (100-300 and 300-500 ms, Figure 4b-c). We used the negative cluster found in the response to data unlabeled as MW and OT (excluding the last 12 seconds of each audiobook section), to select Pz, POz, P1 and P2 as the electrodes of interest for comparing the responses to the OT and MW conditions (see Figure S2 for other possible clusters of channels, and Figure S3 for the topographies for unlabeled data). At the early window, we found a stronger response for OT (mean OT-MW difference = 0.12 μV [0.04, 0.20], $t(24) = 3.23$, $p = .004$, BF10 = 11.49). The effect of MW on onset response was weaker between 300-500 ms, with an inconclusive result (mean OT-MW difference = .049 μV [-0.11, 0.01], $t(24) = 1.62$, $p = .117$, BF10 = .67). This indicates that, in line with many studies showing attenuation of perceptual responses[27], MW is associated with a decline in early responses to spoken words.

Cluster-based analysis of surprise response. To examine the processing of word-level surprisal during mind-wandering, we tested the coefficients of surprise in the regression model outlined above. We found an occipital-central negativity associated with larger surprise at the word level, reflecting the estimated change in rERPs when a certain word is more surprising by one standard deviation. This effect was significant for both types of report (MW: between 242-761 ms, permutation $p < .001$; OT: between 152-789 ms, permutation $p < .001$, Figure 4d). The surprise response peaked somewhat earlier for MW than OT, based on the maximum of individual waveforms (mean latency difference = 70 ms [20, 120], $t(24) = 2.718$, $p = .012$, BF10 = 4.10). No significant difference was found between OT and MW response amplitudes, suggesting no strong effect of MW on surprise processing (all $p > .48$). See Figure S4 for the effects of lexical frequency in the same model.

It is possible that the lack of the MW effect on surprise was due to a conservative correction for multiple comparisons. That is, we did not find a significant cluster in a corrected procedure, but perhaps focusing on the activity in a specific timepoint can yield the effect. To explore this possibility, we extracted the average difference at a posterior cluster of channels

during two different time frames (100-300, 300-500, posterior cluster which showed the strongest activity in sections unlabeled as MW/OT, Figure 4e-f and Figure S5) and compared MW and OT responses. At the early window, mirroring the later peak observed for OT response, we found a stronger surprise response in MW (mean OT-MW difference between 100-300 ms = -0.084 μV [-0.15, -0.02], $t(24) = -2.65$, $p = .014$, BF10 = 3.61). Notably, this contrasts with the stronger onset response found for OT in the same time window. At the peak of the surprise response, between 300-500 ms, no difference was found, with a null effect more likely (mean OT-MW difference between 300-500 ms = -0.022 μV [-0.09, 0.05], $t(24) = -0.67$, $p = .503$, BF10 = 0.26). These results remained unchanged when keeping only content words and when controlling for words with high uncertainty, quantified by the overall entropy of prior probability for word occurrence (Table S5).

Taken together, we found robust surprise-evoked negativity during both MW and OT periods, without strong influence of MW, despite the decline in response to word onset. To statistically test whether the MW effect significantly differed between the onset and surprise response, we conducted a permutation test comparing the difference waves (MW effect on onset response minus MW effect on surprise response). This confirmed a larger difference between MW and OT in response to word onset and the response to the surprise level during the early time window (between 23-398 ms, permutation $p = 0.007$, Figure S6).

Analysis Using Variable Time Windows. One limitation of the results is that the 12-second window defined as MW may be contaminated by momentary OT periods. Contamination of this sort may create an apparent effect of surprise, even if there is none in “pure” MW periods. It should be less pervasive in shorter window durations, as subjects are more likely to accurately reflect on their experiences. If contamination is responsible for our

results, shortening the duration of the window should reduce the estimated effect of surprise during MW and/or increase the effect during OT periods.

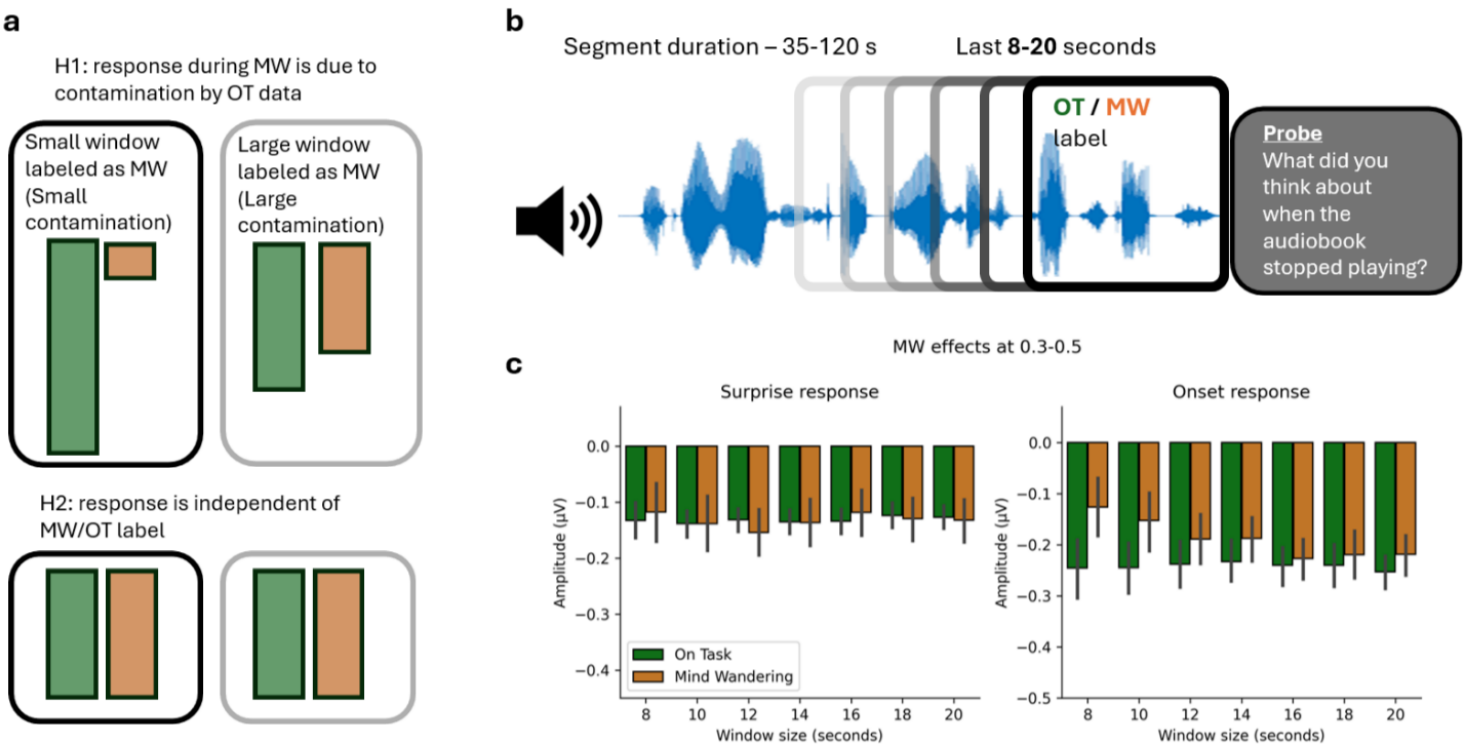

Figure 5. Effect of the labeling window size (a) We hypothesized that labeling a smaller temporal segment would result in less attentive periods contaminating MW periods. This means that MW-associated effects should grow smaller when the temporal window increases and larger when the window decreases. (b) Instead of the 12-second window used for the original analysis, we conducted the same analysis using windows of 8-20 seconds for labeling and extracted the mean amplitude of surprise and onset responses at the posterior cluster. (c) Analysis of MW effects per response and window. The y axis shows the mean estimated response, and error bars denote one SE.

We thus extracted the average amplitude of the surprise and onset responses (at 300-500 ms after onset) of MW and OT episodes based on different time windows (8-20 seconds pre-probe, in 2-second increments, response type: surprise/onset, report: MW/OT, Figure 5). The effect of MW on word onset response was more substantial in the shorter time window, consistent with it being less contaminated by OT data ($t(148) = 2.50$, $p = .013$ for 300-500ms), as indicated by a mixed-effect model predicting the effect by time window. For the surprise response, no such interaction was observed ($t(148) = 0.21$, $p = .84$). Further, we found a robust surprise effect during MW regardless of the specific time window chosen (all within-

window $p < 0.01$, effect of time: $t(148) = 0.03$, $p = .976$). These results are inconsistent with a contamination explanation, underscoring the robustness of our findings to the chosen time window.

Overall, the analyses demonstrate that our design is sufficiently sensitive to capture the (somewhat expected) influences of mind-wandering on spectral power and evoked responses to speech. Despite this sensitivity, markers of word-level surprise persisted during MW, regardless of the time frame defined as MW.

Mind-wandering attenuates but not abolish contextual representations

Here, we focused on the contextual information used by the LLM to generate the prediction, using a similar approach to the semantic encoding described above. If predictions are being conducted prior to word occurrence, then the features of these predictions should correlate with EEG activity to some degree. As contextual information, we used the last hidden layer of Mistral[43,53] model activations for each word, which includes the information used to predict the probabilities of upcoming words. These contextual embeddings will be identical for two different words with identical preceding contexts. Similar to the procedure used for semantic encoding, we used linear models trained on the embeddings of unlabeled EEG data (see Methods) to predict the EEG around single words and calculated the Pearson correlation between the estimated and actual EEG during MW and OT periods. This procedure was repeated for each timepoint and each channel, with high correlation meaning stronger neural encoding of embeddings, that is, stronger context-related activity.

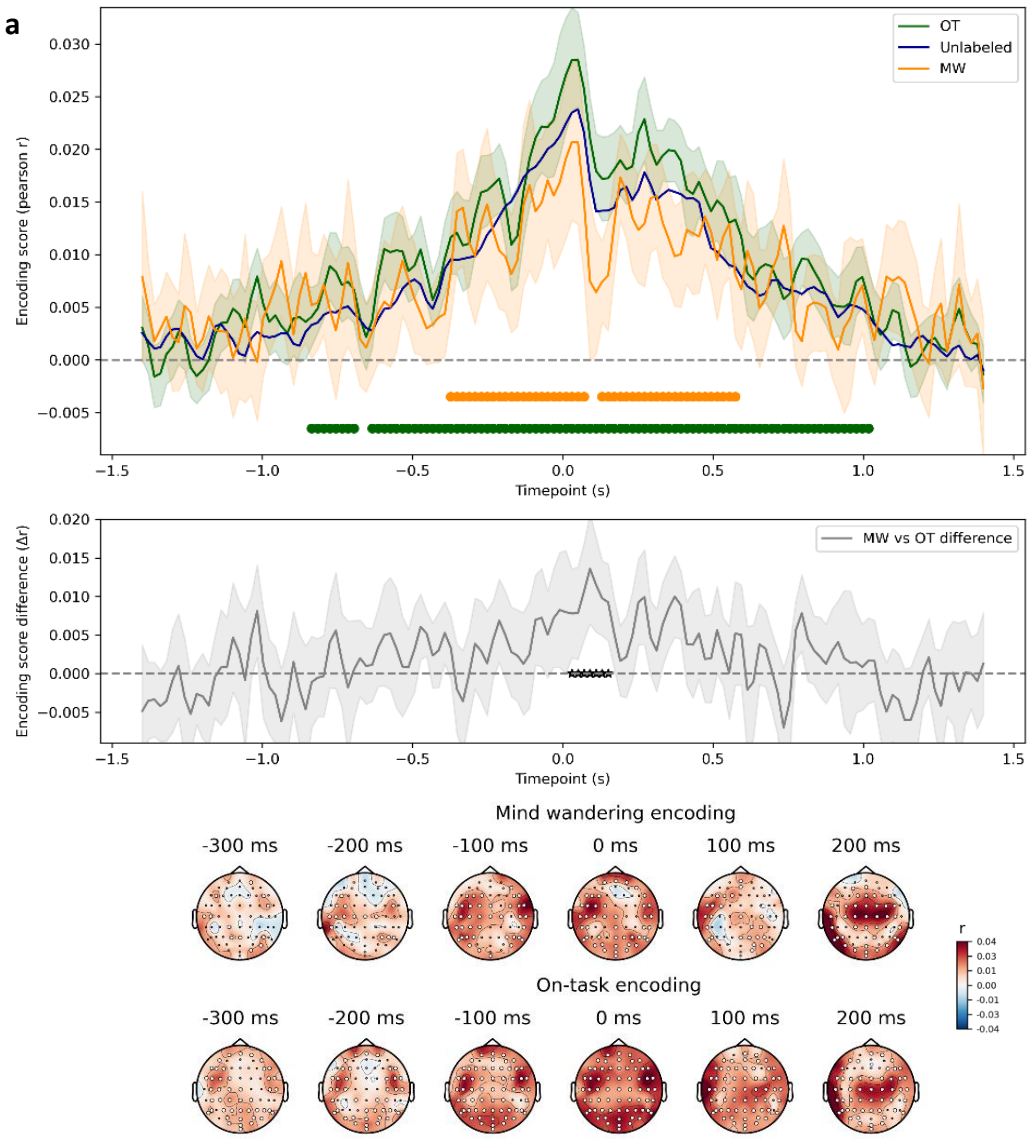

Figure 6. Encoding of contextual embeddings. (a) Analysis procedure. We extracted contextual embeddings predicting each word based on the 128 words preceding it, so each embedding is a vector used by the model to extract the conditional probabilities of upcoming words. These vectors were trained on left-out data to predict the EEG in each channel and timepoint across all words labelled as OT or MW. (b) Results of average encoding scores per time-point. Green and orange dots denote $p < .05$ after cluster-based temporal permutation test. The grey line reflects the difference between OT and MW accuracy per timepoint, with black stars for significance. Shaded areas denote 1 SE. (b) shows the scalp topography of encoding scores in different timepoints for OT (bottom) and MW (top) periods, with white circles indicating $p < .05$ in spatial-temporal cluster-based permutation test.

This analysis yielded several insights. First, both MW and OT periods showed robust

encoding of context, peaking around word onset (OT: -634 to 836 ms, permutation $p < .001$; MW: -372 to 71 and 131 to 574 ms, both with permutation $p < .001$, Figure 6). The mean encoding during peak, between -100 to 100 ms around word onset, was r = 0.0244 (SD = 0.0135) for OT periods, and 0.0162 (SD = 0.0181) for MW data (uncorrected difference, $t(24) = -3.07$, $p = .005$, BF10 = 8.4). Second, while most timepoints did not show differences between OT and MW, our corrected cluster permutation test did reveal a significant improvement in encoding for OT data around the peak activity (between 30-151 ms, permutation $p = .023$). Thus, we find that context-based predictive activity is somewhat diminished at word onset during mind-wandering, but still highly intact.

We also conducted a control analysis with acoustic information of previous stimuli instead of contextual embeddings, as such information might be present in the contextual embeddings[54]. We found that acoustic features performed worse in predicting EEG activity, compared to contextual embeddings, suggesting they could not account for all context-related activity (Figure S8).

Discussion

Our results indicate that the brain processes and predicts continuous speech even when the mind wanders, and therefore, such processing by itself does not indicate subjective experience. First, mind-wandering was associated with changes in EEG spectrum and a weaker early response to words, likely reflecting perceptual decoupling. One might expect higher-level processing, involving word meaning, to become even weaker compared to early responses, which are associated with low-level processes. In contrast, we found that surprise response and semantic representation did not show a strong reduction, and even possible markers of enhancement, during mind-wandering.

Not all processing markers we tested were completely unaffected by mind-wandering. Pre-onset contextual encoding was only partially preserved when participants' minds wandered, compared to on-task periods, despite it being highly significant and following a similar topography. This difference was strongest at word onset, possibly reflecting misalignment of context activity with onset or lower constraints of predictions (e.g., instead of predicting the next word to be "Napoleon", predicting it would be a name of a person). This result may seem at odds with the intact surprise response, which requires contextual prediction. This apparent discrepancy could be settled by two accounts. First, it is possible that

the lack of a tight prior causes difficulty in lexical/semantic access for unpredicted words, compared to highly predicted words. This difficulty may elicit more surprise-related activity during mind-wandering, as a previous study suggested[40]. Second, it is possible that even weaker, but existing, contextual activity, is enough to generate surprise responses. The exact link between pre-word contextual activity and surprise response, and the types of information each represent, is a topic for further investigation.

Perceptual Awareness and Processing Levels

The findings raise an important question about the depth of external processing during mind-wandering. The topography and timing of the surprise response were highly similar to studies of speech-related N400, associated with difficulty of semantic integration or lexical access[11,51,55]. A parsimonious interpretation would suggest that during mind-wandering, the brain nevertheless activates mechanisms relating to this lexical-semantic search, but it is not certain that these reflect high-level processing per se, rather than the prediction of low-level information[19]. Such processing, however, was supported by the encoding of pre-onset contextual activity, as such context was based on the contents of multiple previous words. Moreover, the persistent encoding of word2vec representations around 400 ms is highly suggestive that word meaning is indeed processed.

Our result carries immediate implications for decoupling accounts of mind-wandering[24–27], which currently do not suggest externally-oriented processes to be active during such periods. We propose two possible hypotheses that explain how processing occurs during mind-wandering. One account would suggest that while mind-wandering involves an internal stream of thoughts, we also remain aware of the external stream of information. This would imply that it is possible to simultaneously, or by rapidly alternating between contents, maintain awareness of two streams of information[56]. Considering that we do not find a major effect of MW on predictive signals, this postulated mechanism appears to come without any processing cost. While split awareness in MW contradicts subjective intuition and participants' reports, it warrants future scrutiny.

An alternative explanation would suggest processing remains intact, despite diminished external awareness. This means high-level predictions are partially independent from conscious awareness[57]. This interpretation may inform ongoing discussions about non-conscious processing capacity, which is mostly limited to the visual modality of singular

events[58–61]. Further, it implies that theories describing high-level predictions as the basis for consciousness should be refined and address which types of predictions do require awareness[62,63]. While the reliance on mind-wandering reports limits the ability to mark specific words as consciously or non-consciously perceived, it highlights the importance for future studies to assess whether our results reflect purely non-conscious processing or partial awareness.

The “reduced-awareness” explanation also makes our result particularly relevant to current discussions about AI consciousness[6–8]. LLMs are highly successful at imitating human language capabilities, promoting discussions regarding their ability to consciously experience and “understand”[64]. The capacity of such models stems from autoregressive prediction - that is, predicting the upcoming word given a broad context. Here, however, we show that in humans, such prediction capacities persist when individuals report not having a subjective experience of the text. This suggests that predictive capacities, on their own, cannot be used as an argument for LLMs having subjective experience, at least not in the same way humans do.

Limitations

Our study has several limitations that future research might investigate. First, we address “prediction” as a single entity in this work, while it could be expanded to extract the specific semantic and linguistic dimensions being predicted[19]. Because our focus is on *any* type of contextual prediction, involving previous words by definition, we also did not separate between different parts of speech.

Second, we coded mind-wandering as a binary choice, to test if processing markers are diminished during *any* kind of mind-wandering. In practice, theoretical accounts often describe mind-wandering on multi-dimensional and continuous scales[65,66]. This introduces uncertainty regarding the depth of mind-wandering reports: in principle, some participants might wander only during the last few seconds, yet we labeled the previous 12 seconds as MW. We dealt with this limitation by showing that surprise processing remains strong even when using a shorter window, arguably making the periods more closely aligned with the report. Seeing that both time windows allowed for strong effects on the perceptual response, this contamination-based explanation cannot fully account for the strong processing markers found during MW.

Future studies may also examine whether other measurement approaches of subjective state (e.g., continuous or multidimensional) can reveal differences within mind-wandering states.

We also note, with respect to the surprise response, that there are examples of reduction in N400-like response during inattentive periods in general[51] and during audiobooks specifically[11,67]. Our study is distinguished from other works in two ways. First, we relied on spontaneous fluctuations in attention, rather than tasks requiring the selective suppression of "unattended" speech. Second, we conceptualized surprise as a continuous variable (avoiding artificial dichotomization) and controlled for overlapping words, which may have resulted in a more accurate estimation. However, we note that other variables may also affect the results, such as the language used, engagement with content, and task duration.

Conclusion

Our study provides evidence that multiple measures of semantic and contextual word processing remain robust during mind-wandering. As in many studies, reports of mind-wandering were indicative of reduced early responses and spectral changes in the EEG. Contextual prediction and semantic representation presumably reside higher in the processing hierarchy. Yet, we found that they remained intact, and possibly stronger, during mind-wandering. This suggests that while prediction is important for language processing, it is insufficient for the experience of being engaged with contents.

Materials and Methods

Participants

25 participants took part in this experiment (17 females). All were native Hebrew speakers and students in the Hebrew University of Jerusalem and received payment or course credit for participation. The study was approved by the Social Science Ethics Committee at the Hebrew University of Jerusalem, and all participants gave informed consent prior to participation. We chose a sample size of 25, which is larger than previous similar studies using audiobooks[11,19], as we planned to conduct within-subject comparisons, and used longer audiobook recordings for the same reason (see below).

Stimuli and Procedure

Participants listened to six different audiobook chapters in the same order. The topics, by order, were: Urban legends, Greek drama (1st chapter), Fiction writing, Product Economy,

Greek drama (2nd chapter) and Napoleon Bonaparte. All audiobooks were previously recorded by different native Hebrew speakers. The full duration of the audiobooks was two hours and 12 minutes (see Table S2). They were instructed that they should do their best to stay attentive, and that they will be asked comprehension questions at the end.

All words in each audiobook were given a time-stamp by WhisperX[68]. The timestamps and spellings were later checked and corrected by a human annotator, who corrected spelling mistakes and timing errors. For playing the sounds, we created stereo stimuli, one of which was sent to the loudspeakers. We included the actual sound, and the second included impulse responses for each starting point of words, which were recorded in the EEG file.

Once every 40-120 seconds, with the exact distance sampled from an exponential distribution, the audiobook was stopped, and participants were asked regarding the contents of their experience in the moment prior to the pause. There were seven possible options, following an established measurement method[69]. The first option was “I was attentive to the audiobook” and the next six allowed different options, like “day-to-day things” and “current concerns”. The participants got a detailed description, read identically by the experimenter to ensure they understood that they should choose the first option whenever they felt attentive to the audiobook (See Table S1 for further details). Here, we label all options other than the first one as “mind-wandering” (MW), and first option as “On-task” (OT). The data was labeled in the 12-second window preceding the probe. Time periods outside of the labeled window are heretofore described as “unlabeled data”. After the listening was over, participants answered 23 comprehension questions, with an average accuracy of 82.4% (SD = 7.4%), indicating overall good understanding.

Preprocessing and Analysis

EEG analysis was done using MNE-python version 1.4.0[70]. To pre-process the data, we first rejected noisy or completely silent channels by visual inspection. We then applied a bandpass 1-49 Hz filter, interpolated rejected channels and applied an average reference. To clean activity associated with eye movements and blinks, we trained an Independent Component Analysis (ICA[71]) on the raw data, and zeroed out the weights of components with topography and time course that reflect blinks or eye movement activity. The solution was applied to an unfiltered copy of the data. The clean data was then filtered between 0.5-8 Hz

(see Table S4 for 0.5-20 Hz), the frequencies that were found to track speech responses [72,73], which were also used in a recent prominent study[19] that examined contextual processing.

Language Model Representations

For semantic encoding, we relied on word2vec embeddings, which capture context-independent semantic representations. For Hebrew word2vec embeddings, we used 100d vectors previously trained on co-occurrences in a Wikipedia corpus[50]. To reduce the dimensionality for encoding, we used PCA to extract the top 40 components.

Contextual embeddings (4096 dimensions) were extracted from representations of a Hebrew version of Mistral 7B [43,53] running on each audiobook with a window of 128 words in steps of one word (meaning that for word n, the model was provided with words n-128 to n-1, and the last hidden layer was saved as the information used for the prediction of word n). The same model was used to extract model-assigned probabilities of word occurrence, which were then transformed to negative log odds-ratio ($-log\frac{p}{1-p}$) to create a contextual surprise score. It was used with a window of 25 words, which was found to saturate correlation with human performance in a next-word prediction task[15]. The use of log odds-ratios means that the distance between probabilities of 0.01 and 0.1 is similar to the distance between 0.9 and 0.99, but larger than the distance between 0.45 and 0.54, which are qualitatively more similar ("somewhat certain"). To validate the surprise score, we conducted the same procedure on a HeBERT Model. We found strong correlations in surprise scores given by the two models (all r>0.58 across audiobooks, Table S2). Word frequency was extracted using a corpus of frequencies in Wikipedia and movie subtitles, and was log-transformed[74].

Spectrum-based Classification of Mind-wandering

To test for evidence that subjective mind-wandering reports were indeed associated with distinct patterns of brain activity, as previously demonstrated, we conducted a classification analysis within each subject based on the power spectrum density in different channels. For each subject and probe, we extracted data segments between 12 to 0 seconds before probe and assigned it a label based on the report. We then computed, per channel, the multitaper-based Power Spectrum Density (PSD), and extracted the power in different frequencies and channels. We used, each time, all channels with different frequency bands, with the mean extracted power in all frequencies within a given range. Our classification pipeline included a 10-fold cross-validation with 64 features and 136 samples per subject. We

used a Linear Discriminant Analysis (LDA) with Ledoit-Wolf lemma for shrinkage to train the classifier on 9 folds, and tested it on the left-out set, across all ten folds. LDA is especially suitable for our data because it uses prior probabilities for classes based on the training set proportions, accounting for imbalance. Our metric of success was based on the area under curve of the receiver operation characteristic (ROC-AUC), which is a metric of classification robust to unbalanced dataset.

Null-hypothesis Significance Testing

The main statistical procedure in the paper was based on Maris & Oostenvald[75] cluster-based spatial-temporal permutation test when comparing the activity in two conditions within subject (hence, we analyzed difference waves). The procedure first locates and scores the sum of t-scores within a cluster, defined by point adjacent in time and/or space crossing the threshold of $p = 0.05$ in a two-tailed t-test. Then, the test performs a permutation-based procedure that randomly sign-flips the difference waves of individual subjects (effectively creating an expected difference of 0) and calculates the maximum cluster score. For each of the original clusters, the corrected p-value represents the chances to get such a cluster by chance. Whenever conducting single t-tests in follow-up tests, Bayes Factor is calculated based on a standard medium prior of 0.707[76].

Encoding Analyses

Both the contextual and semantic encoding analyses were based on epoching of the data between -1400 to 1400 milliseconds relative to word onset (rejecting epochs with over 120 μV peak-to-peak differences) and fitting linear regressions for every timepoint and channel separately. Our cross-validation procedure used data from the beginning of each audiobook section (excluding the last 12 seconds), which was unlabeled as MW/OT, to train a model predicting the EEG of the test data (labeled as MW/OT) from embeddings. Specifically, we trained the regression model on the dimensionality-reduced representations (the first 40 PCA components) of the word embeddings in the data unlabeled as MW/OT . Then, we used the trained PCA weights to extract components representing the words in the test data (for MW and OT separately). The model based on data unlabeled as MW/OT was then used to predict the EEG activity of the labeled data, and Pearson correlation between the predicted EEG was calculated as a metric of encoding success (at each timepoint and channel, across words).

484 Importantly, the fact that the same model was used to predict MW and OT activity allows us to
485 avoid potential confounds related to training set sizes.

486 To account for the possible effect of autocorrelation in the EEG and the embeddings,
487 we conducted a control analysis in which we shifted the words by a lag of 100 (so word n
488 received the embedding of word n+100, Figure S7). This effectively breaks the relationship
489 between the words and the embeddings. For the main statistical analysis, since we had no
490 prior hypothesis regarding the locus of the effect, we averaged the Pearson correlations across
491 all channels as a representation of encoding accuracy, resulting in a singular time series of
492 correlations per subject. A similar procedure was conducted using word2vec embeddings,
493 which capture context-independent semantic representations. For Hebrew word2vec
494 embeddings, we used 100d vectors previously trained on co-occurrences in a Wikipedia
495 corpus[50]. To reduce the dimensionality for encoding, we used PCA to extract the top 40
496 components. Like the contextual embeddings, PCA was trained only on the unlabeled
497 (training) data to avoid leakage.

498 For additional analysis of acoustic embeddings, we extracted the spectrogram of the
499 previous word, which could evoke an activity that might be conflated with actual predictive
500 context activity. Specifically, we used a Mel-spectrogram with 64 bands to extract acoustic
501 embeddings. All spectrograms, for all words used, were reduced to 40 components using PCA.
502 Then, a 6-fold cross-validation (each time leaving out one audiobook, fitting the PCA and linear
503 model on the training data, and predicting the test data) procedure was applied to test the
504 ability to predict the EEG activity from the embedding of the preceding word.

505 The main statistical procedure was a temporal cluster permutation test on the average
506 time series, achieved by aggregating all channels within-subject. To gain further insight into the
507 effects, we also calculated, for each subject, the average encoding score at 100 ms before and
508 after the peak, that is, -100 to 100 for contextual and 300-500 for word2vec embeddings.

509 ***Regression-based Analysis of Surprise response***

510 To extract the response for contextual surprise, we conducted a multiple regression
511 that accounts for overlap between words[52]. It predicted the whole EEG time course with a
512 design matrix that includes all word onsets as variables (coded as 1 in the design matrix) with
513 contextual surprise as the primary variable of interest, with the log of lexical frequency also
514 added to the model to control for its shared variance (~10%) with surprisal. Additional

515 covariates used were the time until the next word, average pitch (F0, using an autocorrelation-
516 based algorithm, estimating pitch every 333 ms in steps of 111 ms [77]), and two metrics of
517 speech intensity. The speech intensity measures were mean intensity and sudden changes in
518 intensity (via standard deviation of acoustic envelope), both of which were shown to explain the
519 EEG signal [78]. Pitch and both intensity metrics were calculated twice per word, once for the
520 500 ms preceding word onset, and once for the 500 ms after word onset, which resulted in 6
521 different predictors in the regressions. The regression modelled responses (and accounted for
522 overlapping responses) from -500 to 1000 ms after word onset in each channel separately,
523 rejecting segments with a peak-to-peak difference exceeding 120 μV.

524 Mind-wandering and OT periods were labeled based on a 12-second period before the
525 probe, using the answer to label the data that preceded it. Data that fell outside these 12
526 seconds were unlabeled and utilized for training encoding models. For word onset, word
527 frequency, and contextual surprise, three separate variables were used in the regression: one
528 to model the response of unlabeled data, and one to model the reaction of MW/OT data. To
529 allow differential effects, each variable was assigned a value of 0 when the occurring word
530 belonged to one of the other labels. Frequency and contextual surprise were standardized (z-
531 scored) before being split into three variables. Due to the limited capacity of regressions to a
532 high number of features, other covariates that were not of interest were not divided and used
533 as one variable. The same splitting procedure was used to label data for the contextual
534 embeddings and surprisal analyses. To ensure MW effects were not due to differences in the
535 impact of preparation and prediction occurring before word appearance, we calculated the
536 mean activity between -100 and 0 ms in a posterior cluster of channels based on the unlabeled
537 data topography (POz, P1, P2, and Pz, combined across channels). We used this average to
538 compare the mean baseline response during MW and OT periods, with no meaningful
539 difference found.

540

541

References

1. Hickok, G. & Poeppel, D. The cortical organization of speech processing. *Nat. Rev. Neurosci.* **8**, 393–402 (2007).
2. Fedorenko, E., Ivanova, A. A. & Regev, T. I. The language network as a natural kind within the broader landscape of the human brain. *Nat. Rev. Neurosci.* **25**, 289–312 (2024).
3. Diehl, R. L., Lotto, A. J. & Holt, L. L. Speech Perception. *Annu. Rev. Psychol.* **55**, 149–179 (2004).
4. James, W. The principles of psychology. *Henry Holt* (1890).
5. Ihde, D. *Listening and Voice: Phenomenologies of Sound*. (State University of New York Press, 2012).
6. Bayne, T. *et al.* Tests for consciousness in humans and beyond. *Trends Cogn. Sci.* (2024).
7. Chalmers, D. J. Could a Large Language Model be Conscious? Preprint at https://doi.org/10.48550/arXiv.2303.07103 (2024).
8. Butlin, P. *et al.* Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. Preprint at https://doi.org/10.48550/arXiv.2308.08708 (2023).
9. Leonard, M. K. *et al.* Large-scale single-neuron speech sound encoding across the depth of human cortex. *Nature* **626**, 593–602 (2024).
10. Huth, A. G., De Heer, W. A., Griffiths, T. L., Theunissen, F. E. & Gallant, J. L. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* **532**, 453–458 (2016).
11. Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J. & Lalor, E. C. Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of Natural, Narrative Speech. *Curr. Biol.* **28**, 803-809.e3 (2018).
12. Baldassano, C. *et al.* Discovering event structure in continuous narrative perception and memory. *Neuron* **95**, 709–721 (2017).

13. Marslen-Wilson, W. D. & Welsh, A. Processing interactions and lexical access during word recognition in continuous speech. *Cognit. Psychol.* **10**, 29–63 (1978).
14. Warren, R. M. Perceptual Restoration of Missing Speech Sounds. *Science* **167**, 392–393 (1970).
15. Goldstein, A. *et al.* Shared computational principles for language processing in humans and deep language models. *Nat. Neurosci.* **25**, 369–380 (2022).
16. Schrimpf, M. *et al.* The neural architecture of language: Integrative modeling converges on predictive processing. *Proc. Natl. Acad. Sci.* **118**, e2105646118 (2021).
17. Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P. & van den Bosch, A. Prediction During Natural Language Comprehension. *Cereb. Cortex* **26**, 2506–2516 (2016).
18. Federmeier, K. D. Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology* **44**, 491–505 (2007).
19. Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P. & de Lange, F. P. A hierarchy of linguistic predictions during natural language comprehension. *Proc. Natl. Acad. Sci.* **119**, e2201968119 (2022).
20. Smallwood, J. & Schooler, J. W. The Science of Mind Wandering: Empirically Navigating the Stream of Consciousness. *Annu. Rev. Psychol.* **66**, 487–518 (2015).
21. Christoff, K. *et al.* Mind-Wandering as a Scientific Concept: Cutting through the Definitional Haze. *Trends Cogn. Sci.* **22**, 957–959 (2018).
22. Schooler, J. W. *et al.* Meta-awareness, perceptual decoupling and the wandering mind. *Trends Cogn. Sci.* **15**, 319–326 (2011).
23. Smallwood, J., McSpadden, M. & Schooler, J. W. The lights are on but no one's home: Meta-awareness and the decoupling of attention when the mind wanders. *Psychon. Bull. Rev.* **14**, 527–533 (2007).

24. Smallwood, J. *et al.* Pupillometric evidence for the decoupling of attention from perceptual input during offline thought. *PLoS ONE* **6**, (2011).
25. Smallwood, J., Beach, E., Schooler, J. W. & Handy, T. C. Going AWOL in the brain: Mind wandering reduces cortical analysis of external events. *J. Cogn. Neurosci.* **20**, 458–469 (2008).
26. McVay, J. C. & Kane, M. J. Conducting the Train of Thought: Working Memory Capacity, Goal Neglect, and Mind Wandering in an Executive-Control Task. *J. Exp. Psychol. Learn. Mem. Cogn.* **35**, 196–204 (2009).
27. Kam, J. W. Y., Rahnuma, T., Park, Y. E. & Hart, C. M. Electrophysiological markers of mind wandering: A systematic review. *NeuroImage* **258**, 119372 (2022).
28. Kouider, S. & Dehaene, S. Levels of processing during non-conscious perception: A critical review of visual masking. *Philos. Trans. R. Soc. B Biol. Sci.* **362**, 857–875 (2007).
29. Mudrik, L. & Deouell, L. Y. Neuroscientific Evidence for Processing Without Awareness. *Annu. Rev. Neurosci.* **45**, 403–423 (2022).
30. Schreiber, M., Stockart, F. & Mudrik, L. The UnconTrust Database for Studies of Unconscious Semantic Processing and Attentional Allocation. *Sci. Data* **12**, 157 (2025).
31. Mudrik, L., Faivre, N. & Koch, C. Information integration without awareness. *Trends Cogn. Sci.* **18**, 488–496 (2014).
32. Marcel, A. J. Conscious and unconscious perception: Experiments on visual masking and word recognition. *Cognit. Psychol.* **15**, 197–237 (1983).
33. Holender, D. Semantic activation without conscious identification in dichotic listening, parafoveal vision, and visual masking: A survey and appraisal. *Behav. Brain Sci.* **9**, 1–23 (1986).
34. Cherry, E. C. Experiments on the recognition of speech with one and two ears. *J. Acoust. Soc. Am.* **25**, 975–979 (1953).

35. Treisman, A., Squire, R. & Green, J. Semantic processing in dichotic listening? A replication. *Mem. Cognit.* **2**, 641–646 (1974).
36. Kane, M. J., Smeekens, B. A., Meier, M. E., Welhaf, M. S. & Phillips, N. E. Testing the construct validity of competing measurement approaches to probed mind-wandering reports. *Behav. Res. Methods* **53**, 2372–2411 (2021).
37. Killingsworth, M. A. & Gilbert, D. T. A wandering mind is an unhappy mind. *Science* **330**, 932 (2010).
38. Seli, P. *et al.* How pervasive is mind wandering, really? *Conscious. Cogn.* **66**, 74–78 (2018).
39. Kam, J. W., Dao, E., Stanciulescu, M., Tildesley, H. & Handy, T. C. Mind wandering and the adaptive control of attentional resources. *J. Cogn. Neurosci.* **25**, 952–960 (2013).
40. Long, Z., Fu, Q. & Fu, X. Word Familiarity Modulates the Interference Effects of Mind Wandering on Semantic and Reafferent Information Processing. *J. Cogn. Neurosci.* **35**, 1229–1245 (2023).
41. Kauf, C., Chersoni, E., Lenci, A., Fedorenko, E. & Ivanova, A. A. Log Probabilities Are a Reliable Estimate of Semantic Plausibility in Base and Instruction-Tuned Language Models. Preprint at https://doi.org/10.48550/arXiv.2403.14859 (2024).
42. Brown, T. *et al.* Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
43. Jiang, A. Q. *et al.* Mistral 7B. Preprint at https://doi.org/10.48550/arXiv.2310.06825 (2023).
44. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. Distributed Representations of Words and Phrases and their Compositionality. in *Advances in Neural Information Processing Systems* vol. 26 (Curran Associates, Inc., 2013).
45. Grand, G., Blank, I. A., Pereira, F. & Fedorenko, E. Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nat. Hum. Behav.* **6**, 975–987 (2022).

46. Dong, H. W., Mills, C., Knight, R. T. & Kam, J. W. Detection of mind wandering using EEG: Within and across individuals. *Plos One* **16**, e0251490 (2021).
47. Jin, C. Y., Borst, J. P. & Van Vugt, M. K. Predicting task-general mind-wandering with EEG. *Cogn. Affect. Behav. Neurosci.* **19**, 1059–1073 (2019).
48. Lundqvist, M., Miller, E. K., Nordmark, J., Liljefors, J. & Herman, P. Beta: bursts of cognition. *Trends Cogn. Sci.* **28**, 662–676 (2024).
49. Cavanagh, J. F. & Frank, M. J. Frontal theta as a mechanism for cognitive control. *Trends Cogn. Sci.* **18**, 414–421 (2014).
50. Shemesh, R. & Goldberg, Y. Hebrew Word2Vec. (2019).
51. Kutas, M. & Federmeier, K. D. Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annu. Rev. Psychol.* **62**, 621–647 (2011).
52. Smith, N. J. & Kutas, M. Regression-based estimation of ERP waveforms: II. Nonlinear effects, overlap correction, and practical considerations. *Psychophysiology* **52**, 169–181 (2015).
53. Peleg, Y. Hebrew-Mistral-7B [Language Model]. (2024).
54. Schönmann, I., Szewczyk, J., Lange, F. P. de & Heilbron, M. Stimulus dependencies—rather than next-word prediction—can explain pre-onset brain encoding during natural listening. 2025.03.08.642140 Preprint at https://doi.org/10.1101/2025.03.08.642140 (2025).
55. Lau, E. F., Phillips, C. & Poeppel, D. A cortical network for semantics: (de)constructing the N400. *Nat. Rev. Neurosci.* **9**, 920–933 (2008).
56. Hirst, W., Spelke, E. S., Reaves, C. C., Caharack, G. & Neisser, U. Dividing attention without alternation or automaticity. *J. Exp. Psychol. Gen.* **109**, 98–117 (1980).
57. Goldstein, A., Rivlin, I., Goldstein, A., Pertzov, Y. & Hassin, R. R. Predictions from masked motion with and without obstacles. *PLOS ONE* **15**, e0239839 (2020).

58. Hassin, R. R. Yes It Can: On the Functional Abilities of the Human Unconscious. *Perspect. Psychol. Sci.* **8**, 195–207 (2013).
59. Mudrik, L. & Deouell, L. Y. Neuroscientific Evidence for Processing Without Awareness. *Annu. Rev. Neurosci.* **45**, 403–423 (2022).
60. Hesselmann, G. & Moors, P. Definitely maybe: can unconscious processes perform the same functions as conscious processes? *Front. Psychol.* **6**, 584 (2015).
61. Kouider, S. & Dupoux, E. Subliminal speech priming. *Psychol. Sci.* **16**, 617–625 (2005).
62. Hohwy, J. & Seth, A. Predictive processing as a systematic basis for identifying the neural correlates of consciousness. *Philos. Mind Sci.* **1**, 3 (2020).
63. Seth, A. K. & Bayne, T. Theories of consciousness. *Nat. Rev. Neurosci.* **23**, 439–452 (2022).
64. Goldstein, A. & Stanovsky, G. Do Zombies Understand? A Choose-Your-Own-Adventure Exploration of Machine Cognition. Preprint at https://doi.org/10.48550/arXiv.2403.00499 (2024).
65. Wallace, R. S. *et al.* Mapping patterns of thought onto brain activity during movie-watching. *eLife* **13**, RP97731 (2025).
66. Christoff, K., Irving, Z. C., Fox, K. C., Spreng, R. N. & Andrews-Hanna, J. R. Mind-wandering as spontaneous thought: a dynamic framework. *Nat. Rev. Neurosci.* **17**, 718–731 (2016).
67. Boudewyn, M. A. Half-listening or zoned out? It's about the same: the impact of attentional state on word processing in context. *Cogn. Neurosci.* **14**, 107–114 (2023).
68. Bain, M., Huh, J., Han, T. & Zisserman, A. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. Preprint at https://doi.org/10.48550/arXiv.2303.00747 (2023).

69. Kane, M. J., Smeekens, B. A., Meier, M. E., Welhaf, M. S. & Phillips, N. E. Testing the construct validity of competing measurement approaches to probed mind-wandering reports. *Behav. Res. Methods* **53**, 2372–2411 (2021).
70. Gramfort, A. *et al.* MEG and EEG data analysis with MNE-Python. *Front. Neuroinformatics* **7**, 267 (2013).
71. Makeig, S., Bell, A., Jung, T.-P. & Sejnowski, T. J. Independent component analysis of electroencephalographic data. *Adv. Neural Inf. Process. Syst.* **8**, (1995).
72. Di Liberto, G. M., O'Sullivan, J. A. & Lalor, E. C. Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing. *Curr. Biol.* **25**, 2457–2465 (2015).
73. Zion Golumbic, E. M. *et al.* Mechanisms Underlying Selective Neuronal Tracking of Attended Speech at a "Cocktail Party". *Neuron* **77**, 980–991 (2013).
74. Paridon, J. van & Thompson, B. subs2vec: Word embeddings from subtitles in 55 languages. Preprint at https://doi.org/10.31234/osf.io/fcrmy (2019).
75. Maris, E. & Oostenveld, R. Nonparametric statistical testing of EEG-and MEG-data. *J. Neurosci. Methods* **164**, 177–190 (2007).
76. Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D. & Iverson, G. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychon. Bull. Rev.* **16**, 225–237 (2009).
77. De Cheveigné, A. & Kawahara, H. YIN, a fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am.* **111**, 1917–1930 (2002).
78. Fiedler, L., Wöstmann, M., Herbst, S. K. & Obleser, J. Late cortical tracking of ignored speech facilitates neural selectivity in acoustically challenging conditions. *NeuroImage* **186**, 33–42 (2019).

712

713

714

715

716

717