nature portfolio

Peer Review File

High-level Prediction of Continuous Speech During Mind-Wandering

Corresponding Author: Mr Gal Chen

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
This paper uses participant self-reports about mind-wandering during an audiobook listening task to study the necessity of attention for different linguistic processes, especially lexical retrieval and next-word prediction. I think the substance of this paper is exciting and I support eventual publication, but I have a couple technical and presentational concerns about the manuscript as written.

1. (I'll freely acknowledge that this point is a bit pedantic but I thought it was still important enough to mention.) There's a nuance to the cluster-based tests used throughout the paper: the null hypothesis being rejected is that there are no clusters (anywhere), which means that the inference supported by a positive test is that there is a cluster (somewhere). The test by design only allows you to claim that there is an effect, not that the effect falls into a specific spectral band, brain area, or time interval (see Sassenhagen & Draschkow 19 for discussion). This means that, if you use such tests, in principle you should not say that there was a significant effect in X region between time Y and time Z, but rather that the cluster, which happened to be in X region between time Y and time Z, was significant. I acknowledge that this is both awkward to state and unfortunate when the goal is to make inferences not only about whether an effect exists but also where/when it occurs. But these are intrinsic drawbacks to using this kind of test. I think the testing results should be phrased more carefully and there should be an acknowledgment of the difference between what was tested (existence of an effect) and what is claimed (existence of an effect in a specific time interval, brain location, or spectral band).

2. The authors chose to represent the prediction signal using a non-standard negative log-odds (which they call "surprise") rather than the standard negative log probability (Shannon information or "surprisal") used almost exclusively in related work. Their justification for doing so seems thin, namely, subjective researcher intuition that probabilities seem more similar in the center of the probability interval than at the extremes (lines 168-171). I don't share this intuition about high probabilities (0.9 and 0.99 don't seem that different to me), and in any case there is now substantial theory that has been developed around surprisal as a linking hypothesis between word probabilities and measures of cognitive effort (Norris 06; Levy 08; Smith & Levy 08, 13; Hoover et al 23), which is lacking for log-odds. So I think the authors need to either better justify this decision or show that their results are robust to it, or ideally both. In any case, they do sometimes incorrectly call their measure "surprisal" (e.g., line 251), so those cases at minimum should be fixed.

3. Although I find the experiments and analyses to be innovative and convincingly support the core claim that much of language processing is conserved during mind wandering, I think the writing and framing of the paper could be tightened up. In particular, there are several places where the authors use terms that have unclear meaning and/or discuss questions raised by unstated background assumptions. Here are some specifics:

- "It also suggests that subjective experience is more tightly associated with accumulating information from multiple timepoints rather than processing any single object" (lines 41-43): In context, this seems to come out of nowhere and the meaning is unclear. What contrast is being drawn by "accumulating information from multiple timepoints rather than processing any single object"? What is a "single object"? And how do the results referred to by "it" support this conclusion? To be clear, I'm not asking the authors to explain the answers to me -- I think I have inferred them from the paper as a whole. I'm saying that, at the point where this statement is made, readers don't have enough background to interpret it.

- The authors use "consciousness" in contexts when I think they should instead use "attention". For example, in saying that

speech processing "can occur when consciousness is reduced" (lines 55-56), it sounds like means reduced consciousness in the sense of e.g., sleepy or anaesthetized, which isn't the intended meaning. I find it strange to talk about shifts in attention as reductions in consciousness -- the person is fully conscious, they're just attending to something else. Similar usages occur on line 69 and elsewhere. This terminological choice also seems to have the substantive effect of inviting conflations with distinct concepts. For example, while using "consciousness" to refer to attention, the authors then try to connect their study to ongoing debates about AI consciousness in the sense of subjective experience, which seem as far as I can tell to be irrelevant in substance to what was studied here. It's just the terminological overlap that invites the comparison in the first place.

- The meaning of the authors' use of the term "high-level" is unclear to me. What types of representations would count for them as being "high-level" vs not?

- The claim in line 101 that LLMs and humans "share computational resources" as worded sounds like some sort of cyborg, but I think all that may be meant is that LMs can predict brain activity? If so, the latter wording would be clearer.

- The authors make a terminological distinction between "conceptual" and "semantic", which seems like a false opposition because semantic interpretation requires context. Because these labels map respectively onto word2vec vs Mistral, I think a clearer description of what is actually being studied would be "contextualized" vs "uncontextualized", but then additional motivation would be helpful as to why this is a distinction that matters for their research question.

- I don't understand the sentence on lines 505-507: "Our results carry immediate implications for decoupling accounts of mind-wandering21,22,24, which implies that externally-oriented processes are inactive during such periods"

- Lines 511-512 say "Considering that we do not find a major effect of MW on predictive signals, this postulated mechanism appears to come without any processing cost." I don't see how this follows. Is the assumption that no processing occurs during MW? Isn't that what this paper is arguing against? What is the signature of processing cost that the authors believe is lacking in their results?

4. It would be a helpful sanity check to see whether comprehension question accuracy is correlated with the percentage of time on task, as an additional piece of validation for the self report data.

Minor:

5. Please clarify how subword tokenization was handled when computing probabilities.

6. It would be helpful to match the y/z-axis bounds in Fig 3, to make the effect sizes more comparable.

Reviewer #2

(Remarks to the Author)
Summary

This manuscript investigates the dissociation between high-level predictive speech processing and conscious awareness during mind-wandering. Using EEG recordings from 25 participants listening to continuous audiobook narratives, the authors employed a neural encoding framework to compare neural responses during self-reported mind-wandering (MW) versus on-task (OT) states. Specifically, they examined neural measures of word onset, word-level surprisal, semantic content, and predictive contextual embeddings. The results indicate that while MW is characterized by distinct spectral changes and attenuated early responses to word onsets, neural signatures of word-level surprisal and semantic encoding remain robust. Notably, the encoding of predictive context persists during MW, albeit with reduced strength compared to OT, suggesting an impaired capacity for multi-word integration rather than a complete cessation of prediction. These findings support the authors' conclusion that single-word predictive mechanisms can operate independently of subjective experience, which may instead rely on the accumulation of information across extended temporal windows.

Overall, this is an interesting study that advances our understanding of the neural encoding of continuous speech during mind-wandering. The analysis is methodologically sound, and the manuscript is clearly written and accessible. However, several major concerns and a few minor issues require further consideration.

Major:

1. While the manuscript provides details on experimental procedures and data preprocessing, it critically lacks a comprehensive description of the EEG data acquisition process. To ensure the study's reproducibility and allow for a proper assessment of data quality, the authors must include a dedicated section or paragraph detailing the recording hardware and setup. Specifically, please clarify the make and model of the EEG system, the electrode configuration (e.g., channel count, montage layout, and reference placement), and the sampling parameters. Furthermore, it is essential to state whether electrooculography (EOG) was recorded to monitor eye blinks and movements, and if so, how these signals were integrated into the artifact correction pipeline. Without these fundamental acquisition details, it is difficult to fully evaluate the robustness of the subsequent preprocessing and analysis results.

2. The current results indicate a significant class imbalance in the labeled data, with On-Task (OT) states comprising approximately 72% of probes compared to only 28% for Mind-Wandering (MW) states. It is crucial to clarify whether specific strategies were employed to mitigate this imbalance. For instance, the authors should provide a control analysis demonstrating that the main effects persist when the sample sizes for both conditions are strictly equated (e.g., via downsampling). Without such evidence, it remains uncertain whether the reported neural findings reflect genuine state-dependent mechanisms or are merely artifacts of unequal data quantities.

3. Regarding the Encoding Analyses (lines 315-318, and 433-436), I have a concern about the data distribution used for the training procedure. The authors utilized unlabeled data as the training set, while using WM and OT data as test sets. Since the OT state was reported disproportionately more often than the WM state during the experiment, the unlabeled training data is presumably dominated by OT-related neural patterns. Consequently, the trained encoding models might be biased towards capturing OT-specific features, potentially compromising their validity when predicting WM states. Could the authors clarify how they addressed this potential bias?

4. The current analyses rely heavily on several critical empirical parameters, s, e.g., PCA dimensionality (line 159), context window size (lines 161-162), and EEG data segments (lines 180, 206, 243-244), that may influence the feature extraction and modeling outcomes. I recommend that the authors should explicitly justify these parameter selections based on prior literature or theoretical grounds.

5. The authors' finding that high-level word processing remains robust during mind-wandering (MW) states aligns well with prior work, notably Har-shai Yahav & Zion Golumbic (eLife, 2021), which demonstrated neural processing of words and phrases in task-irrelevant continuous speech. The authors should expand their discussion in light of these findings to clarify whether the linguistic processing observed during MW reflects a similar mechanism to that of actively ignored speech, or if distinct neural signatures differentiate internal distraction from external inattention.

Minor:

Some typos need to be edited and corrected.
e.g., "context-based predictions of *a single word" in line 64;
"between *word response" in line 247;
"reflect *a reduction" in line 528.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
I'm happy with the authors' responses and revisions and have no further concerns.

Reviewer #2

(Remarks to the Author)
All my previous concerns have been adequately resolved. I am satisfied with the changes made and believe the paper now meets the standards for publication.

Response to referees

We thank the editor and reviewers for their thoughtful comments on our manuscript. We reply to the comments one by one below, and hope that the revised manuscript addresses them adequately.

Editorial requests

First, we highlight here several changes made following the Editorial Request Table:

1. Data availability statements and code availability statements were added.

2. All abbreviations for “mind-wandering” and “on-task” (MW/OT) in the main text were replaced with the full term.

3. The last sentence of the introduction, presenting the gist of the results, was removed.

4. We added a statement regarding preregistration.

5. Information regarding participants’ gender is now clearly reported.

6. All effect sizes for t-tests reported now appear (with confidence interval). Whenever tests are summarised, a reference to a supplemental table was added. Since for permutation tests, no additional information can be added beyond p-value (at least not in a way that can avoid double-dipping that inflates effect sizes), we complemented them with analyses on the average effect during time windows of interest.

7. A sensitivity power analysis, based on a simulation that considers the number of within-participant measurements, is added in p.6.

8. All tests comparing MW vs OT effects, which previously included a Bayes Factor analysis, now also include a direct interpretation that describes findings as providing evidence for effect / evidence supporting the null.

Reviewer #1 (Remarks to the Author):

This paper uses participant self-reports about mind-wandering during an audiobook listening task to study the necessity of attention for different linguistic processes, especially lexical retrieval and next-word prediction. I think the substance of this paper is exciting and I support eventual publication, but I have a couple technical and presentational concerns about the manuscript as written.

1. (I'll freely acknowledge that this point is a bit pedantic but I thought it was still important enough to mention.) There's a nuance to the cluster-based tests used throughout the paper: the null hypothesis being rejected is that there are no clusters (anywhere), which means that the inference supported by a positive test is that there is a cluster (somewhere). The test by design only allows you to claim that there is an effect, not that the effect falls into a specific spectral band, brain area, or time interval (see Sassenhagen & Draschkow 19 for discussion). This means that, if you use such tests, in principle you should not say that there was a significant effect in X region between time Y and time Z, but rather that the cluster, which happened to be in X region between time Y and time Z, was significant. I acknowledge that this is both awkward to state and unfortunate when the goal is to make inferences not only about whether an effect exists but also where/when it occurs. But these are intrinsic drawbacks to using this kind of test. I think the testing results should be phrased more carefully and there should be an acknowledgment of the difference between what was tested (existence of an effect) and what is claimed (existence of an effect in a specific time interval, brain location, or spectral band).

We appreciate this comment and do not see it as pedantic at all. We are well aware of the Sassenhagen & Draschkow argument about what can and cannot be deduced. Our focus was, as the reviewer noted, on whether an effect exists , rather than on a precise onset or offset of the effect (which was not tested or discussed). Following the reviewer's comment we now explicitly acknowledge these limitations of cluster testing in the methods section (p. 10):

> *"Cluster-based permutations provide evidence that a significant cluster existed in the window tested while keeping the family-wise error rate below 5%. However, the procedure does not ensure that any single point can be marked as significant by itself. Therefore, the exact extent of the clusters should we treated with due caution (Sassenhagen & Draschkow ,2019)"*

And after the first report of a significant cluster (p.15):

> *"Note that due to the nature of the cluster permutation procedure, the exact onset and offset here and in the rest of the results are not guaranteed, and should be treated with caution (Sassenhagen & Draschkow ,2019)"*

We also changed the language of the reporting to be in line with the limitations. While we previously reported "significant clusters between...", which aligns with the

reviewer's interpretation, at some points we omitted the specific mention of "cluster" and just wrote "between XX-XX ms..." – which is not fully accurate out of context. Therefore, we added "significant cluster" before all these mentions (p.15-20 and p.23), such that no point from a significant cluster is stated as significant by itself.

While the exact onset and offset are not guaranteed by the procedure, we do believe that the general spatial and temporal locations of a significant cluster are of interest (without attaching too much significance to their onset and offset times), and therefore report them. To provide additional information, we also tested the average amplitude in two specific time windows (100-300 ms and 300-500 ms) while acknowledging the exploratory nature of this test. We believe that together, it allows the reader to get a good grasp of when and where, approximately, effects were found.

2. The authors chose to represent the prediction signal using a non-standard negative log-odds (which they call "surprise") rather than the standard negative log probability (Shannon information or "surprisal") used almost exclusively in related work. Their justification for doing so seems thin, namely, subjective researcher intuition that probabilities seem more similar in the center of the probability interval than at the extremes (lines 168-171). I don't share this intuition about high probabilities (0.9 and 0.99 don't seem that different to me), and in any case there is now substantial theory that has been developed around surprisal as a linking hypothesis between word probabilities and measures of cognitive effort (Norris 06; Levy 08; Smith & Levy 08, 13; Hoover et al 23), which is lacking for log-odds. So I think the authors need to either better justify this decision or show that their results are robust to it, or ideally both. In any case, they do sometimes incorrectly call their measure "surprisal" (e.g., line 251), so those cases at minimum should be fixed.

We are grateful for this comment, which resulted in a substantial improvement of the manuscript. Our initial motivation to use log-odds was that it conforms to the measure that is linearly modelled in logistic regressions. However, upon reading this comment and the literature it cites, we agree that negative-log probability is more suitable, and more theoretically grounded.

Accordingly, all analyses in the paper using "surprisal" now refer to negative log probability (-log(p)), including the additional down-sampling analysis proposed by reviewer 2 (see below). The change allows our results to be connected to a large theoretical and empirical literature while remaining loyal to the traditional definition of surprisal. We are thankful for the reviewer's suggestion.

Notably, the new analysis showed near-identical patterns of results. This almost identical result is to be expected, considering the high correlation between negative log odds and -log(p), which defines the degree to which this variable switch should

affect the regression (Pearson r = 0.977). This is because the number of words which are actually highly probable is very low (i.e., 1.87% above p = 0.9).

3. Although I find the experiments and analyses to be innovative and convincingly support the core claim that much of language processing is conserved during mind wandering, I think the writing and framing of the paper could be tightened up. In particular, there are several places where the authors use terms that have unclear meaning and/or discuss questions raised by unstated background assumptions. Here are some specifics:

- "It also suggests that subjective experience is more tightly associated with accumulating information from multiple timepoints rather than processing any single object" (lines 41-43): In context, this seems to come out of nowhere and the meaning is unclear. What contrast is being drawn by "accumulating information from multiple timepoints rather than processing any single object"? What is a "single object"? And how do the results referred to by "it" support this conclusion? To be clear, I'm not asking the authors to explain the answers to me -- I think I have inferred them from the paper as a whole. I'm saying that, at the point where this statement is made, readers don't have enough background to interpret it.

Thank you for highlighting our admittedly obscure phrasing. We changed the abstract to be clearer about this point (p.2):

> *"This encoding also persisted during mind-wandering with a weak decline compared to on-task periods, pointing towards a decreased ability to integrate information from multiple words."*

- The authors use "consciousness" in contexts when I think they should instead use "attention". For example, in saying that speech processing "can occur when consciousness is reduced" (lines 55-56), it sounds like means reduced consciousness in the sense of e.g., sleepy or anaesthetized, which isn't the intended meaning. I find it strange to talk about shifts in attention as reductions in consciousness -- the person is fully conscious, they're just attending to something else. Similar usages occur on line 69 and elsewhere. This terminological choice also seems to have the substantive effect of inviting conflations with distinct concepts. For example, while using "consciousness" to refer to attention, the authors then try to connect their study to ongoing debates about AI consciousness in the sense of subjective experience, which seem as far as I can tell to be irrelevant in substance to what was studied here. It's just the terminological overlap that invites the comparison in the first place.

We appreciate the opportunity to clarify our terminology.

1. "Conscious awareness" or "attention": We used consciousness in terms of *content* consciousness (what is experienced), rather than *state* consciousness (whether

one has any experience; see, for example, section 2(iii) in Chalmers, 2000 or Bayne et al., 2016, p1). That is, we are referring to conscious awareness, in the sense of having a subjective experience, or awareness, *of something* (thoughts/audiobook). Indeed, mind wandering reports refer to the contents of participants' subjective experience (note that this was the question we asked – "what did you experience at the moment before the pause").

While attention may be one of the prerequisites for conscious awareness/subjective experience, attention and awareness are not typically treated as synonymous (for reviews and empirical evidence see De Brigard & Prinz, 2010; Koch & Tsuchiya, 2007; Naccache et al., 2002). Further, measures of mind-wandering are not always correlated with measures of sustained attention even though both of them correlate with behavior (Chidharom et al., 2025). As we measured subjective reports of experience, rather than manipulating or measuring a proxy of attention, we believe our terminology is the one closest to the data without making assumptions on attention.

Nevertheless, we agree that "consciousness" out of context was confusing, and therefore changed the phrasing in several places throughout the manuscript to clarify we refer to the contents of consciousness\subjective experience. See, for example:

> *"Importantly, speech processing involves not only the decoding of acoustic inputs, but also the* ***conscious, subjective experience of what is being said****" (p.3)*
>
> *"...it remains unclear which components of speech processing are associated with the subjective experience of engaged listening, and which can occur when* ***the conscious awareness of speech contents is reduced****" (p.3)*
>
> *"...This question is especially relevant today, as the capabilities of deep learning models spark discussions regarding whether they can ultimately meet the criteria* ***for conscious awareness (in our case, awareness of contents)****" (p.3)*
>
> *"Here, we use spontaneous fluctuations* ***in the contents of conscious awareness away from the task****, known as mind-wandering, as a naturalistic case* ***of reduced experience of relevant and audible external contents****." (p.3)*
>
> *"These results indicate that semantic and predictive processing at the single-word level do not necessarily entail* ***full conscious awareness of the contents*** *(p.24)"*

2. LLM & subjective experience: given the above, we believe the data support the original interpretation: when people report a reduction in subjective experience of speech contents, the processing of meaning and surprisal remains robust. This also allows us to argue that such properties are not a signature of conscious awareness of content (words in this case), and therefore cannot be used as support that LLMs are aware.

While the argument about LLM consciousness is not the core focus of the paper, we do believe that with the correct reservations (in bold below, and after avoiding the confusion that the reviewer pointed to), this claim is important to make in reference to content, rather than state, consciousness. We believe the current version of paragraph shows these reservations, although we are open to change it if the reviewer finds any specific part inaccurate, confusing or unclear (p. 26-27):

> *"LLMs are highly successful at verbal communication, promoting discussions regarding their ability to consciously experience and understand meaning. The capacity of such models stems from autoregressive predictions - that is, predicting the upcoming word given a broad context. Here, however, we show that in humans, some of these capacities persist when individuals report not having a conscious experience of the text.* ***With caution, and assuming the reports of participants reflect a reduction in conscious experience of the audiobook contents****, this suggests that predictive capacities cannot be used as an argument for LLMs having subjective experience,* ***at least not in the same way humans do****."*

We are thankful to the reviewer for raising this important point, and believe that following the edits, our discussion was made more accurate.

References for this section:

Bayne, T., Hohwy, J., & Owen, A. M. (2016). Are There Levels of Consciousness? *Trends in Cognitive Sciences*, *20*(6), 405–413. https://doi.org/10.1016/j.tics.2016.03.009

Chalmers, D. J. (2000). What is a neural correlate of consciousness. *Neural Correlates of Consciousness: Empirical and Conceptual Questions*, *17*. https://home.cs.colorado.edu/~mozer/Teaching/syllabi/3702/readings/Chalmers2000.pdf

Chidharom, M., Bonnefond, A., Vogel, E. K., & Rosenberg, M. D. (2025). Objective markers of sustained attention fluctuate independently of mind-wandering reports. *Psychonomic Bulletin & Review*, *32*(4), 1689–1699. https://doi.org/10.3758/s13423-025-02640-6

De Brigard, F., & Prinz, J. (2010). Attention and consciousness. *WIREs Cognitive Science*, *1*(1), 51–59. https://doi.org/10.1002/wcs.27

Koch, C., & Tsuchiya, N. (2007). Attention and consciousness: Two distinct brain processes. *Trends in Cognitive Sciences*, *11*(1), 16–22. https://www.cell.com/trends/cognitive-sciences/abstract/S1364-6613(06)00303-2

Naccache, L., Blandin, E., & Dehaene, S. (2002). Unconscious Masked Priming Depends on Temporal Attention. *Psychological Science*, *13*(5), 416–424. https://doi.org/10.1111/1467-9280.00474

- The meaning of the authors' use of the term "high-level" is unclear to me. What types of representations would count for them as being "high-level" vs not?

We apologise for this oversight – we indeed omitted the (rather crucial) definition of "high-level", which now appears in the first occurrence of the term:

> *"Abundant evidence shows that when listening to speech or reading text, we continuously make **high-level predictions about upcoming words – their content, meaning, and relationship to overall narrative**" (p.2, abstract, 1st sentence)*

> *"Speech processing happens at multiple levels, including the low-level physical properties of the stimulus and its phonetics, but also high-level features like word meaning, and whole narratives." (p.3)*

- The claim in line 101 that LLMs and humans "share computational resources" as worded sounds like some sort of cyborg, but I think all that may be meant is that LMs can predict brain activity? If so, the latter wording would be clearer.

We agree, and changed the phrasing to:

> *"For context-related activity, we extracted the representations that LLMs use to generate predictions. Such contextual embeddings were previously shown to align with electrophysiological signals recorded in the human brain" (p.4)*

- The authors make a terminological distinction between "conceptual" and "semantic", which seems like a false opposition because semantic interpretation requires context. Because these labels map respectively onto word2vec vs Mistral, I think a clearer description of what is actually being studied would be "contextualized" vs "uncontextualized", but then additional motivation would be helpful as to why this is a distinction that matters for their research question.

We see now that the way we used "semantic" might indeed sound like it refers to the meaning of the narrative, rather than the uncontextualized meaning of the word. We changed several sentences in the MS to clarify we mean "context-independent encoding of semantic features" (e.g., pages 4, 8, 10, 13, 14, 24).

To clarify the importance of using these two metrics, and their difference, we added in p.4:

> *"...we focused on two complementary processing metrics which represents the ongoing construction of narrative: the overall context that is used for predicting upcoming words, and the representation of single-word meaning out-of-context. "*

Our initial introduction of the term was also changed to clarify what we mean by "context-independent":

> *"For context-independent processing, we examined the neural encoding of word2vec embeddings, which reflect context-independent semantic dimensions at the word level. These embeddings were pre-trained on an external corpus, and represent semantic features of the specific word, regardless of what preceded it." (p.4)*

- I don't understand the sentence on lines 505-507: "Our results carry immediate implications for decoupling accounts of mind-wandering21,22,24, which implies that externally-oriented processes are inactive during such periods"

We have revised the sentence for clarity and flow, and we hope it is clearer now:

> *"Our results carry immediate implications for traditional perceptual decoupling accounts of mind-wandering, which often postulate that high-level, externally-oriented processing is suspended during such periods, effectively decoupling the wanderer from the external environment."* (p.26)

- Lines 511-512 say "Considering that we do not find a major effect of MW on predictive signals, this postulated mechanism appears to come without any processing cost." I don't see how this follows. Is the assumption that no processing occurs during MW? Isn't that what this paper is arguing against? What is the signature of processing cost that the authors believe is lacking in their results?

What we meant to say is that if the split or multiplexed awareness holds, this splitting comes without cost, at least as far as predictive processing is concerned. Our intention was to raise doubt regarding split-awareness mechanisms. In hindsight, we believe this notion may be more confusing than helpful and decided to remove this sentence altogether (p.26).

4. It would be a helpful sanity check to see whether comprehension question accuracy is correlated with the percentage of time on task, as an additional piece of validation for the self report data.

We appreciate this suggestion for validating the self-report data. We conducted the analysis and found no significant correlation between comprehension question accuracy and the percentage of time on task ($r(23) = 0.06$). This is not surprising as we originally only intended for these simple questions to provide motivation for the participants to listen and used them only as crude validation that they did, rather than as a sensitive measure of ongoing comprehension. Furthermore, an even-odds split-half analysis revealed that the comprehension questions lack the necessary reliability

(that is, people are not significantly correlated with themselves, r(23) = 0.05), making them a noisy metric for this comparison (for comparison, the on-task rate provided split half r(23) = .57, p = .03).

Given these limitations, and our interest in subjective experience, we believe the subjective interest questions serve as a more robust metric for validation. When analyzing those, we found that when an individual rated a specific audiobook chapter as more interesting, they also report being on-task in higher rates (mixed-effects model, slope of subject-specific interest level, t(50.9) = 3.43, p = .001. See also Fig. S1). This, alongside the ability to classify mind-wandering from several frequency bands, support the validity of the self reports for capturing different states.

Minor:

5. Please clarify how subword tokenization was handled when computing probabilities.

We now added clarifications to the Methods sections regarding the computations of both probabilities (sum of log probabilities) and embeddings (mean of sub-word embeddings).

For embeddings:

> *"Punctuation tokens were treated as word boundaries and excluded from the embeddings. In cases where the Mistral tokenizer segmented individual words into multiple tokens, the hidden states of all constituent subword tokens were averaged, producing a single contextual embedding per word. This means that for the words in the phrase "the charismatic person" the embedding we used to predict the brain response for "person" was the mean of the embeddings based on the tokens for "char", "is" and "matic", an a-priori choice made to provide a stable estimate for an embedding that includes the last word. In practice, a comparison to another popular method, taking the last token, showed the two methods provide nearly identical embeddings (Mean of word-based cosine similarity across methods = 0.964, SD = 0.016)." (p.9)*

For probability:

> *"Mistral 7B was also used to extract model-assigned probabilities of word occurrence, which were then transformed to negative log probability (-log(p)), a theoretical and empirical gold standard for calculating contextual surprise score[38,51,52]. These negative log-probabilities were summed across sub-tokens of the same word, which is equivalent to the negative log of joint probability of all tokens in the word." (p.9)*

6. It would be helpful to match the y/z-axis bounds in Fig 3, to make the effect sizes more comparable.

Axes limits were matched when reproducing Figure 3 with the new analyses (with log probability instead of log-odds). The same changes were made for Figure 4.

Reviewer #2 (Remarks to the Author):

Summary

This manuscript investigates the dissociation between high-level predictive speech processing and conscious awareness during mind-wandering. Using EEG recordings from 25 participants listening to continuous audiobook narratives, the authors employed a neural encoding framework to compare neural responses during self-reported mind-wandering (MW) versus on-task (OT) states. Specifically, they examined neural measures of word onset, word-level surprisal, semantic content, and predictive contextual embeddings. The results indicate that while MW is characterized by distinct spectral changes and attenuated early responses to word onsets, neural signatures of word-level surprisal and semantic encoding remain robust. Notably, the encoding of predictive context persists during MW, albeit with reduced strength compared to OT, suggesting an impaired capacity for multi-word integration rather than a complete cessation of prediction. These findings support the authors' conclusion that single-word predictive mechanisms can operate independently of subjective experience, which may instead rely on the accumulation of information across extended temporal windows.

Overall, this is an interesting study that advances our understanding of the neural encoding of continuous speech during mind-wandering. The analysis is methodologically sound, and the manuscript is clearly written and accessible. However, several major concerns and a few minor issues require further consideration.

Major:

1. While the manuscript provides details on experimental procedures and data preprocessing, it critically lacks a comprehensive description of the EEG data acquisition process. To ensure the study's reproducibility and allow for a proper assessment of data quality, the authors must include a dedicated section or paragraph detailing the recording hardware and setup. Specifically, please clarify the make and model of the EEG system, the electrode configuration (e.g., channel count, montage layout, and reference placement), and the sampling parameters. Furthermore, it is essential to state whether electrooculography (EOG) was recorded to monitor eye blinks and movements, and if so, how these signals were integrated into the artifact correction pipeline. Without these fundamental acquisition details, it is difficult to fully evaluate the robustness of the subsequent preprocessing and analysis results.

> We apologize for this glaring oversight. This necessary section was accidentally deleted between versions and we failed to notice it. We now added back the following paragraph to p.7-8:

"EEG was recorded using an ActiveTwo system (Biosemi, The Netherlands) with 64 active electrodes embedded in an elastic cap according to the extended 10-20 system (https://www.biosemi.com/pics/cap_64_layout_medium.jpg). Eight additional flat electrodes were placed: two on the mastoid processes, two horizontal electrooculogram (EOG) electrodes on the outer canthus of each eye, two vertical EOG electrodes below and above the right eye, a single EOG channel under the left eye, and a channel on the tip of the nose. All channels were referenced during recording to a common-mode signal (CMS) electrode between POz and PO3. The sampling rate was 2048 Hz using an online low-pass filter with a cutoff of 1/5 the sampling rate. "

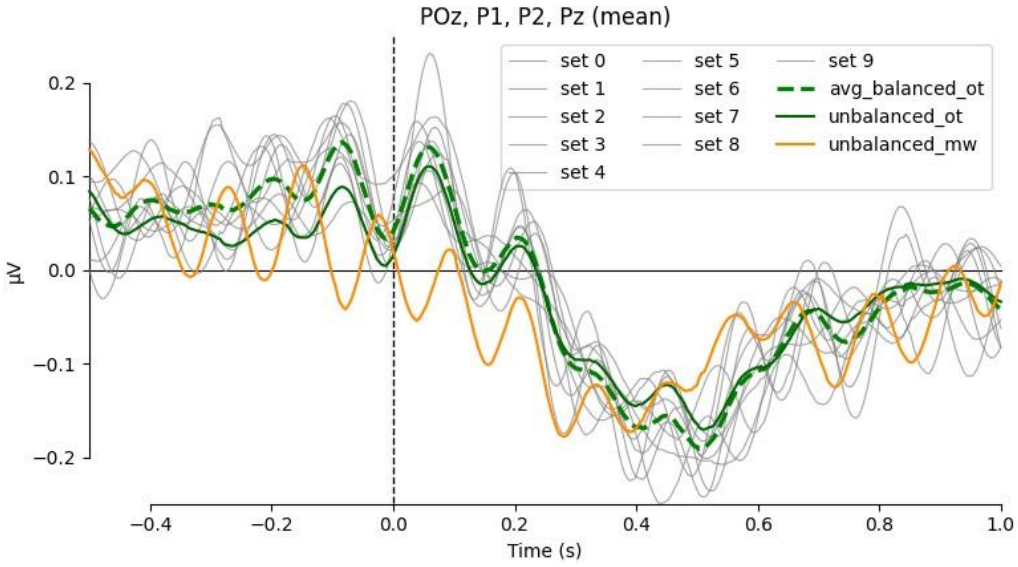
The following paragraph addressed data cleaning using ICA (p. 8):

"To clean activity associated with eye movements and blinks, an Independent Component Analysis was trained on the raw data. Following inspection within-participant, we zeroed the weights of components with topography and time course that reflect blinks or eye movement activity (Mean components removed = 2.72 components, SD = 0.98, range 2-5)."

2. The current results indicate a significant class imbalance in the labeled data, with On-Task (OT) states comprising approximately 72% of probes compared to only 28% for Mind-Wandering (MW) states. It is crucial to clarify whether specific strategies were employed to mitigate this imbalance. For instance, the authors should provide a control analysis demonstrating that the main effects persist when the sample sizes for both conditions are strictly equated (e.g., via downsampling). Without such evidence, it remains uncertain whether the reported neural findings reflect genuine state-dependent mechanisms or are merely artifacts of unequal data quantities.

We appreciate the concern, and therefore followed the reviewer's suggestion and re-analyzed the data while undersampling the on-task condition to match the MW prevalence in each subject. This resulted in qualitatively similar waveforms for the OT condition, undermining the concerns from the imbalanced sets. The waveforms for surprise response during unlabeled periods and MW were near-identical (they were based the same dataset), indicating the overall regression model is robust to the balance between MW and OT samples. To ensure our results are not due to a specific undersampling set, we repeated the procedure 10 times.

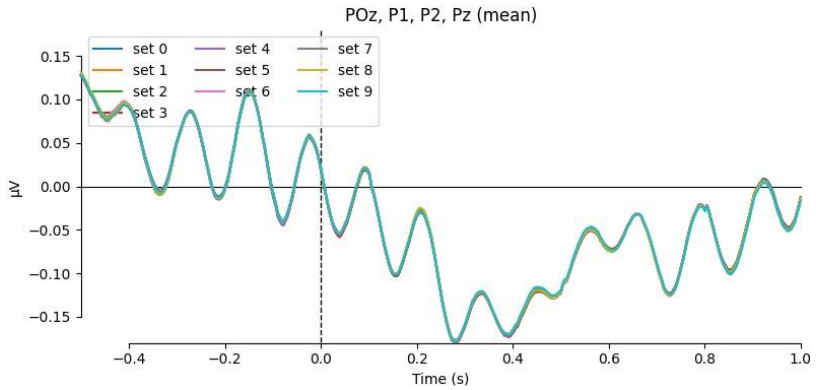
The figure below illustrates the On-task waveforms for the surprise response with under-sampling iterations in gray. For reference, the unbalanced MW and the OT waveform from the manuscript are shown in color. All ten downsampling sets showed the OT-MW effect on the surprise response, and no significant cluster emerged when comparing MW and OT (9/10 sets showed BF01 > 4 between 300-500 ms). This indicates that our results vis-à-vis the surprise response are robust to the imbalance.



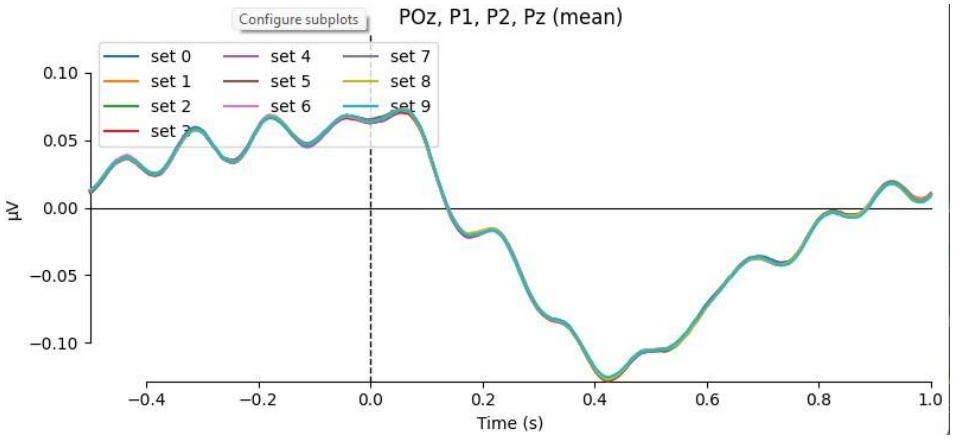
This appears as Figure S9 in the manuscript:

"This analysis is similar to the one presented in Figure 3 for the surprise response, but it is based on different sub-samples of the words in the OT category, so that the number of words is equivalent within each participant and response category (MW/OT). Dashed green line is the average of all sub-samples. The solid lines illustrate the MW and OT effects presented in Figure 3 for reference ."

These are the MW surprise responses for a regression model with under-sampling of OT periods. The regression procedure with different sub-samples of OT produced MW estimates which are nearly identical in all runs, making them hard to distinguish:



Unlabeled surprise response for a regression model with undersampling of OT:



We also added this result in the text (p.19):

> *"To address concerns regarding the 72% on-task to 28% mind-wandering class imbalance, we performed a control analysis by undersampling the on-task data to match mind-wandering prevalence. Across 10 iterations of this procedure, all sets maintained the mind-wandering effect, with 9 out of 10 iterations yielding a Bayes Factor (BF10) < 0.25 for comparing between states during the 300-500 ms period, supporting the conclusions of the original analysis (see Fig. S9). "*

3. Regarding the Encoding Analyses (lines 315-318, and 433-436), I have a concern about the data distribution used for the training procedure. The authors utilized unlabeled data as the training set, while using WM and OT data as test sets. Since the OT state was reported disproportionately more often than the WM state during the experiment, the unlabeled training data is presumably dominated by OT-related neural patterns. Consequently, the trained encoding models might be biased towards capturing OT-specific features, potentially compromising their validity when predicting WM states. Could the authors clarify how they addressed this potential bias?

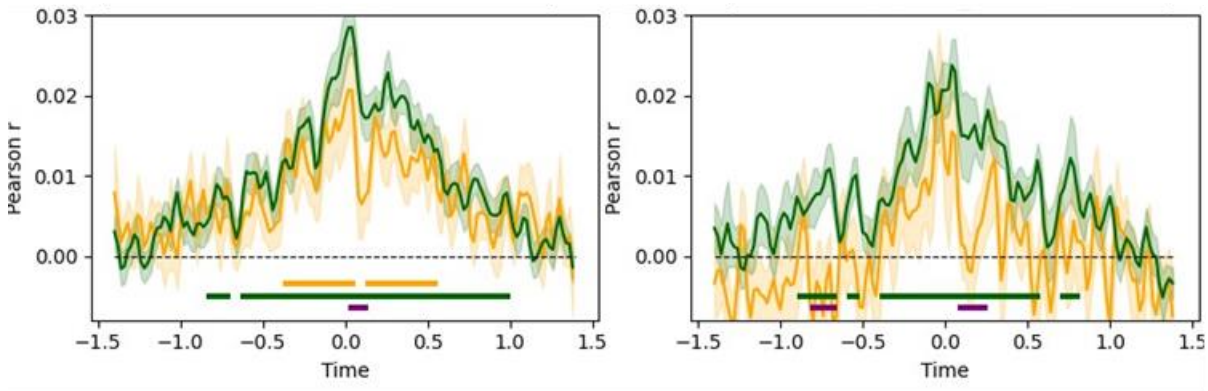
If we correctly understand, the reviewer suggests that the reduced encoding accuracy for MW compared to OT shown in Figure 5 might be due to a bias towards OT in the training test. This is a valid concern, if the contextualized processing during MW differs in its spatio-temporal patterns (used for encoding) from OT states.

Under the above assumption, our procedure introduces a potential confound to cases where OT shows stronger effect than MW (because the model is more tuned towards OT-associated representations). This contains the contextual processing effect, but not the persistent semantic encoding during MW.

To address this concern, we now added an additional control. We augmented the training set by oversampling the MW labelled trials and adding them to the unlabelled training set, in order to conserve the original training set size (see full detail below

and Fig. S10). If indeed the reduced encoding accuracy in the MW compared to OT was due to scarcity of MW examples in the training set, this augmentation should improve the encoding of MW words.

As seen in the figure below, this was not the case. Rather, the augmentation impaired the encoding of both OT and MW words, even more so for the latter. This does not support the hypothesis that the difference between OT and MW encoding accuracy was due to the unbalanced training, but leaves open the possibility that the difference is due to noisier MW signals (that is, noisy/weaker encoding of contextual information), which we now acknowledge. We are grateful to the reviewer for pushing us on this point.



The left plot above is the original result from the manuscript (Fig. 5), and the right one includes the balanced training procedure. As can be seen, the MW effect was preserved during this period. However, this analysis resulted in slightly weaker encoding compared to the original dataset, perhaps because we add training data which was more noisy (due to mind-wandering) compared to the original training set. This suggests that the representations are not state specific, as this would have increased encoding during MW in the balanced case.

Details on the balanced analysis were added to Figure S10):

> *The balanced analysis includes oversampling of MW periods from out-of-test data, conducted to examine whether the training data was biased in favor of the more prevalent OT states. For a fully balanced training set based on strictly OT and MW periods, our effective training set will be diminished, on average, to ~20% its original size. To conduct a stronger analysis, we decided to supplement the original training set with out-of-test MW periods – oversampled several times to match the hypothesized rate of MW in the specific subject.*
>
> *To illustrate the approach, if a participant had 9,000 unlabeled words, and reported MW in 20% of the time – this means approximately 1,800 of the unlabeled words include MW periods, and we would need to oversample the (out of test) MW-labeled words up to 5,400 repetitions, for reaching a total of 7,200 MW-assumed words. This achieves the goal of providing a stable*

training set while reducing the presumed bias that would exist if representations were state-specific.

4. The current analyses rely heavily on several critical empirical parameters, s, e.g., PCA dimensionality (line 159), context window size (lines 161-162), and EEG data segments (lines 180, 206, 243-244), that may influence the feature extraction and modeling outcomes. I recommend that the authors should explicitly justify these parameter selections based on prior literature or theoretical grounds.

As for the time window for EEG segments (1400 before and after word onset), we now added in p.10:

"Our choice of time window is based on our motivation to observe processes close to the onset of the word, while allowing enough time for observing the evolution of temporal dynamics along the timepoints that showed significant encoding in Goldstein et al. (2022)"

Our reasoning for the other parameters mentioned was capturing a wide enough context without overparameterizing models, but we accept that the choice of parameters is necessarily a bit arbitrary. We believe that the best way to address this is by testing the robustness of the results with other choices of parameters.

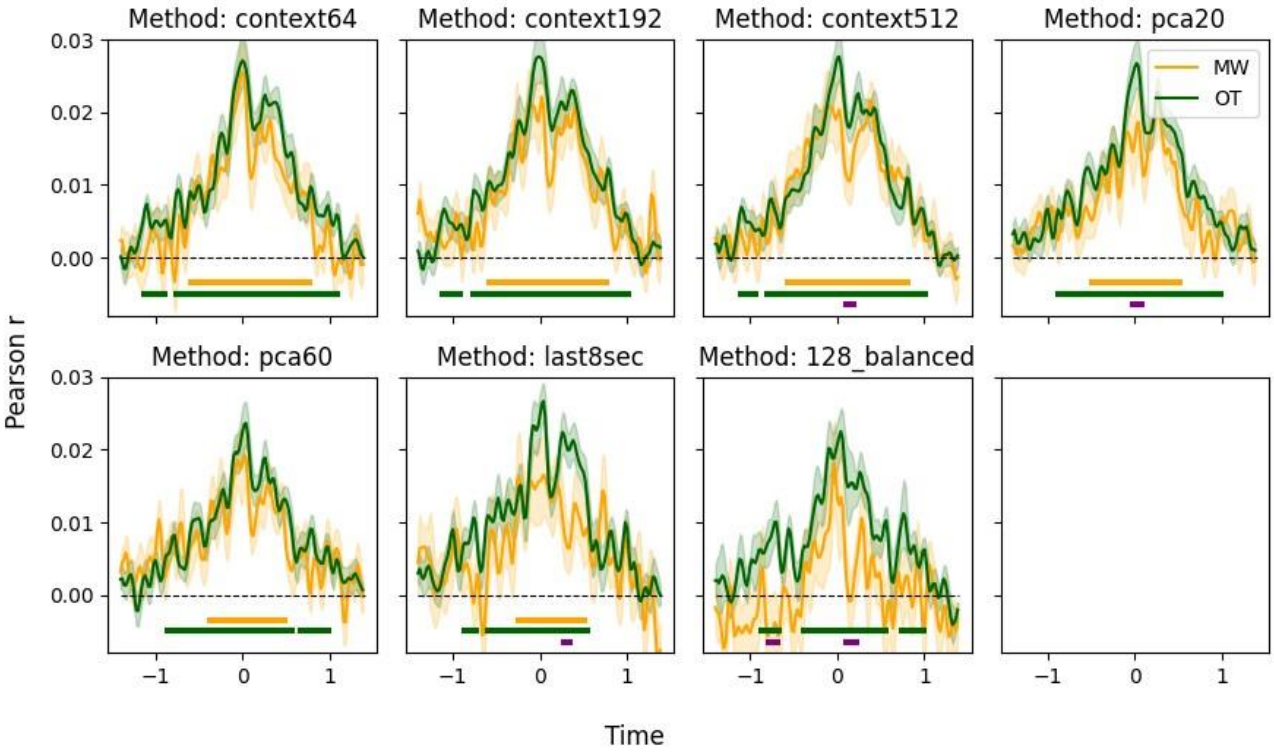
We therefore repeated the analyses with:

1.Using 20-component and 60-component PCA instead of 40

2.Using embeddings based on different context windows of 64, 192 and 512 tokens instead of 128.

3. Using an 8-second window to define MW (effectively creating a smaller, but more accurate, test set).

All these analyses showed robust encoding during MW and OT periods, supporting our main conclusions. However, the difference between MW and OT, while nominally present in all analyses, was not significant for the 64 and 192 context windows, and for the 60-component PCA. Thus, we toned down our discussion of this specific effect while explicitly noting these findings. (p.26-27)



This figure (Figure S10) is a reconstruction of Figure 5 with alternative preprocessing parameters. We used a different number of components for reducing embeddings dimensionality (20/60 instead of 40), a different size of context window (64/192/512 instead of 128), and a different time window defined as MW (8 seconds instead of 12). Solid horizontal lines reflect clusters with $p < .05$ for MW (orange) and OT (green) states.

We added to the text (p.23-24):

> *"To address the robustness of the results to the a-priori chosen parameters, we re-analyzed the data with different parameters for preprocessing context windows (64, 192 and 512 tokens instead of 128), embedding dimensionality (20 and 60 components, in addition to the original 40 components), and time labeled before the probe. We also repeated the analysis with the original parameters with an augmented training set that includes additional out-of-test mind-wandering data, to mitigate possible effects of imbalance in the training set. Crucially, the encoding accuracy during mind-wandering and on-task periods remained significant in all comparisons. The difference between mind-wandering and on-task found using the original parameters around the peak, remained similar in most conditions, but was less robust (significant clusters in 4/7 comparisons, see Fig. S10 for detailed report). "*

5. The authors' finding that high-level word processing remains robust during mind-wandering (MW) states aligns well with prior work, notably Har-shai Yahav & Zion Golumbic (eLife, 2021), which demonstrated neural processing of words and phrases in task-irrelevant continuous speech. The authors should expand their discussion in light

of these findings to clarify whether the linguistic processing observed during MW reflects a similar mechanism to that of actively ignored speech, or if distinct neural signatures differentiate internal distraction from external inattention.

We now expanded our discussion to address this study specifically, as it illustrates the distinction of our approach from selective attention (p.25-26):

> *When examined in comparison to the attention literature, our results seem at odds with studies showing a robust decline in high-level speech processing, and even contextual predictions when a stream of spoken words is not attended.* ***Notably, even a study which used highly sensitive measures (hierarchical frequency-tagging), found that during inattention (Har Shai & Zion-Golumbic, 2021), MEG responses only track word pairs but not full (4-word) sentences.*** *In contrast to such studies, we do find evidence for high-level processing of several words, as implied by the contextual and surprise-related responses. We relied on spontaneous fluctuations in conscious experience, rather than multi-talker presentations in which participants are required to suppress irrelevant speech. This implies that speech processing may remain active during mind-wandering as no active suppression is involved (in fact, the participants aim to listen to the audiobook). "*

Minor:

Some typos need to be edited and corrected.
e.g., "context-based predictions of *a single word" in line 64;
"between *word response" in line 247;
"reflect *a reduction" in line 528.

All typos were corrected, and we conducted an additional review for locating and fixing other grammatical errors. We are thankful for the reviewer for pointing these errors.