nature portfolio

Peer Review File

Plasticity and Language in the Anesthetized Human Hippocampus

Corresponding Author: Dr Sameer Sheth

Any redactions in this file are there to maintain patient confidentiality, the confidentiality of unpublished data, or to remove third-party material.

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

Version 0:

Reviewer comments:

Referee #1

(Remarks to the Author)
General Overview:

This is an interesting paper showing that the brain performs pattern recognition in both the sensory and language domains in the absence of consciousness. The authors were able to mitigate current technical problems with inserting neuropixels probes into deep structures by taking advantage of two stages of an anterior temporal lobectomy for pharmacoresistant epilepsy. In a clever design anterior lateral temporal cortex was first removed to gain access to the amygdala and hippocampus. The authors then pause the surgery for 30 minutes and inserted the neuropixels probe into the hippocampus and commence the recording to either an oddball sequence (n=3 patients) or podcasts (n=2 patients). The full anterior lobectomy including the amygdala, hippocampus and adjacent medial temporal structures was then performed. Note the patients consented to this 30-minute surgical pause to enable the auditory recording scenario. The authors report that under Propofol anesthesia the hippocampus detects change in the classic oddball/mismatch negativity task. Notably the authors also report that the oddball response increases with the length of study implying a learning effect. They additionally found that the hippocampus was processing language at the level of detecting both semantic and grammatical information. The manuscript provides the first evidence of the value of Neuropixels in deep structures and given the important claims of this paper several issues need to be addressed. The paper also provides an opportunity to define the roles of not only SUAs but also high gamma in decoding and how they compare and also address the role of lower frequency oscillations in SUA and HG tuning (see below)

The authors need to be clear that their results do not necessarily extend to all forms of unconscious behavior such as sleep and are restricted to Propofol. A more accurate title might be Learning and language in the anesthesized human hippocampus.

Figure 1J shows robust oddball responses for SUAs, ERPs and high gamma. The decoding accuracy seems highest for the high gamma response. Is this true? Given that sEEG research now cannot employ Neuropixels other that in cortical sites some discussion of the relationship of SUAs and high gamma is warranted given increasing evidence that high gamma is a proxy for nearby SUA activity and can be used for decoding numerous cognitive phenomena as well as being used for BCI (see Leonard et al, Nature 2024). This also raises the question if the results in Fig 3 which are focused on SUA activity and an RNN model could also be obtained with their high gamma findings. A similar query could be raised for Fig 4, could high gamma be used for the linguistic decoding effects reported?

The authors frame their podcast results as providing evidence for semantic and grammatical processing in the anesthesized hippocampus. sEEG data has also provided theta band evidence for semantic processing in the hippocampus (Piai at al, PNAS, 2016). This raises the question of whether other bands besides high gamma are engaged during sleep processing. Similarly, abundant evidence has shown theta-SUA and theta-HG phase amplitude coupling in the hippocampus in a range of tasks. The paper would benefit if the authors examined these issues for both SUAs and high gamma. Finally, the discussion is very short given the wealth of empirical findings and proposed hippocampal mechanisms.

Specific Points:

Please define data cleaning procedures to mark epileptic episodes. In that regard what was the epileptic source: amygdala, hippocampus, parahippocampus etc. Did the resected hippocampal tissue around the probe show any evidence of tissue abnormalities such as gliosis etc.?

The authors should mention earlier their use of random SOA (1 to 3 seconds; now specified on line 512) for the presentation of tones in the oddball task as both random and fixed SOA has been employed with the oddball task using SEEG (Tzovara et al., Cerebral Cortex, 2024, cited by the authors when describing the classic oddball task). The authors should discuss how their choice of a temporally unpredictable sequence may have impacted their low decoding accuracy of oddball identity in light of the same cited publication (Tzovara et al., 2024) showing the hippocampus is mainly sensitive to predictable deviants (in wakefulness). The statement that the oddball response grew in magnitude over the course of the experiment needs better substantiation (line 108-109). The authors show an example unit (Fig. 3A) with a higher spike rate towards the end of the experiment and do not show or report results substantiating this statement. The authors focus on decoding accuracy (line 110-111; Fig. 3C and 3D).

BIS monitor values were kept between 45 and 60 for the duration of the surgical case. Did BIS values change similarly or systematically for patients during the course of the experiment? A supplemental figure showing the evolution of BIS values over the course of the experiment would be helpful to support the authors interpretation of the change in decoding of oddball trials as a result of learning rather than depth of anesthesia.
Fig. 3A: Trace for spike rate for oddball trials 1-20 appear to be missing.

Figure 2b - intertrial interval is 1-3 seconds, but there are multiple peaks much after tone presentation but before 1 seconds. How many tones were in each trial? How many oddballs? This is unclear from the methods.

Spike sorting - Was there any attempt to verify that the units resemble a biologically plausible action potential. This is not clear from Fig 1 D. Examples in the supplement would be nice.

Figure 2E - The claim is that the neuron is selective for oddballs, but is this a statistically significant statement? Furthermore, the smoothing doesn't seem to match the raster plot. The rasters show that on oddball trials there are more spikes from 0-100 msec, but the smoothed plot shows a higher firing rate from 100-200 msec.

Was there any re-referencing performed for the ERP or Gamma bands? It's typically to reference according to closest electrodes to extract local activity.

Figure 2J - Figure caption mentions encoding accuracy, while axis legend mentions decoding accuracy. If using SVM, i think decoding is what is meant, but this should be checked.

Figure 3A - one of the curves appears to be missing (I think it's oddball trials 1-20).

Figure 3B-E - It should be clearer that the writers mean 1st and 2nd half OF EACH BLOCK. Without that clarification, these results don't mean much because oddball and frequency would be collinear, and without any regularization the results could be noise.

Figure 3H - There is r no reason to use an EI net for this when they are not actively using that detail in a meaningful way. The methods should be clearer about what the input to this network is. Is it sine waves at different frequencies, and the goal is to produce fixed points in response to this entraining input? Given that most human single neuron results show encoding of variables in spike timing relative to oscillations, I'm hesitant to buy in to the analogy of this system to a rate-coded EI net. More discussion on the utility of this model beyond 'local computation' would be appreciated

Figure 3I - Decoding of an RNN, but the paper only mentions a statistical test. What decoding was being done here? An SVM like with the neural data? No information in the main text or methods as far as i could see.

The average correlation is low for figure 4.
Is there any held out data to corroborate last claim that we can predict firing rates to a word based on responses to other words. If not, the claim should not be made.

Figure 4F was not cited in the text in text.

4j - 38% is a reliable but not a strong correlation

SVM for language decoding needs to specify why chance performance is 50%. They balanced the number of oddball and standard tones, not clear for the language analysis.

4M - Could this same analysis be done with the best model from the oddball experiment to compare. Unclear if this shows a true next word prediction or highlights the caveat that nearby data in a continuous time series will be similar.

Referee #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

Referee #3

(Remarks to the Author)
In their work, Katlowitz and colleagues describe a rare set of experiments in anesthetized neurosurgical patients with drug-resistant epilepsy. Neuropixels microelectrodes were implanted to brain tissue just before its resection, and sounds were presented – either a series of tones in an “oddball” paradigm, or natural language audio streams (podcasts). The authors put forward a provocative statement – that what they term ‘high-order perception’, as reflected in hippocampal activity, occurs in unconscious surgical anesthesia. Undoubtedly, if the paper would convincingly establish that high-level processing of external events during surgical anesthesia is the norm, this would be of immediate interest to wide audiences.

The authors obtained extremely rare intraoperative data of high-density neuronal recordings in humans and expanded on previous studies that recorded spontaneous cortical activity in humans with Neuropixels probes. Here, data were obtained in hippocampus, and while presenting auditory and language stimuli. In the first experiment, the extent of ‘high-order perception’ is addressed by examining the characteristics of tone responses, their selectivity for tone frequency or oddball status (standard vs. deviant), and dynamics throughout the recording session (what the author term learning). Next, the extent of ‘high-order perception’ is addressed by examining activity recorded during podcast presentation and investigation of the relation between neuronal activities and linguistic features including lexical frequency, semantic embedding, grammatical features, as well as representation of recent or upcoming words.

Unfortunately, the data do not support the strong conclusions put forward and their far-reaching implications. Bold statements require strong supporting data, and this is not the case here. Results of the first experiment (tones) have various shortcomings including weak effects, over-interpretation, and novelty. Results of the second experiment (podcast) are more intriguing, but rely on two participants, are very indirect (based on decoding rather than raw data) and - unless it is established that decodability is comparable to awake data - could potentially reflect residual hidden statistical regularities in the language stimuli (e.g. imagine that, on average, every noun in natural language tends to be followed by longer interval of silence before next word compared to verb; just as one example to illustrate the type of regularities that may exist).

Some major issues include the following:
• The lack of awake data in the same individuals makes all comparisons and interpretations difficult.
• Sound intensity at/around subject location was not measured, limiting the ability to infer about frequency selectivity. Even though tones had identical intensity as digital waveforms, the cables, speaker, and room inevitably attenuate some frequencies more than others; therefore, results interpreted as frequency selectivity could still easily reflect effects of different sound intensities. “Hippocampal units encoded tone frequency” can only be stated if one proves that intensity is identical at the ear…
• Novelty: It is widely established that differential responses to stimuli occur during anesthesia even beyond early sensory regions, for example object-selective responses in IT cortex of monkeys; more specifically, responses to deviant tones occur during anesthesia (typically termed stimulus-specific adaptation), this also true in hippocampus e.g. https://www.ibroneuroreports.org/article/S2667-2421(23)00689-9/fulltext. Furthermore, differential responses to deviant tones are observed in multiple unconscious conditions ranging from natural sleep in rodents to coma in humans (e.g. Nourski et al J. Neuroscience 2018 for intracranial human study under anesthesia). This is the basis for many studies employing the “local/global paradigm” to separate local deviance associated with simple adaption-related processes (known to be preserved in unconscious states) from true deviance detection (as the authors imply here).
• Authors report that 70% of hippocampal neurons significantly respond to pure tones during passive listening. This is very perplexing, I am not aware of anything similar in any species or state; for example, in the work of Aronov et al in rats (Tank group, Nature 2017) during passive playback without the task, only 1.7% of CA1 neurons respond; also, if 70% of neurons respond, that seems to contradict the black trace in Fig. 2F (which seems to show that, unless the sound is deviant, the average response is not different than baseline); how can this co-exist with “most units … showed tone-evoked responses”?
• Main effect for oddball as seen in Fig 2J reflects marginal decodability around 0.52-0.56
• “Oddball decodability“ increases throughout sessions of patients P5 and P6. This could reflect various processes. While this is “consistent with learning effects”, taking this as evidence that the brain learns during unconscious surgical anesthesia – as the authors state in the title – seems like overinterpretation.
• Statistics: neurons recorded in the same individual and electrode and session are not statistically independent. A proper statistical approach (increasingly used in recent years by many similar studies) would be to employ a hierarchical mixed model taking into account both # of patients and # of neurons per patient.

Additional minor issues:
• Fig 1E is not clear
• Why is STG data of P3 not shown?
• Smoothing PSTHs with kernel of 150ms is highly unorthodox and reflects weak data
• It may be suboptimal to focus on high-gamma in anesthesia since anesthesia often yields a decrease in frequency of induced power changes and responses may preferentially occur in the low-gamma range

Referee #4

(Remarks to the Author)
This paper reports neuronal recordings from human neurosurgical patients using high-density neuropixels probes in the hippocampus. Under anesthesia, they played auditory oddball sounds and podcasts, and examined neuronal firing. They found that hippocampal neurons were sensitive to rare oddballs, and that this sensitivity changed over the course of the experiment, which they refer to as learning, and which was recapitulated in an artificial RNN model. When listening to natural language stories, neurons were sensitive to lexical, semantic, and syntactic features of words, including both the current and previous/upcoming words. Notably, all of these effects were observed under anesthesia, suggesting a role for hippocampus in sensory and linguistic processing even without consciousness.

This is a very interesting paper, bolstered by a truly unique dataset with human hippocampal neurons while the patient is in an unconscious state. The analyses are generally sophisticated and rigorous, getting a lot out of data that have some inherent limitations (e.g., time constraints, electrode placement, etc). The paper raises several interesting and important questions about the relationship between sensory and linguistic processing, consciousness, and the hippocampus.

I have some questions and suggestions that I hope will improve the manuscript:

MAJOR

1) The framing of changes across trials in the oddball task being a form of “learning” does not seem well-justified to me. The structure of an oddball task like the one used here can be recognized almost immediately once the first rare target is presented, so to the extent that some sort of implicit learning of this task structure happens, it’s not clear why it should be expected to change linearly. So in that sense, I’m not sure how to interpret the change over time. While I agree that it seems like they have ruled out adaptation as an explanation for this change, other possibilities include gradual changes in conscious state affecting firing rates (did they measure BIS continually? See for example Nourski et al 2018, J Neuro), or the possibility that they are not holding the same population of neurons over the entire experiment. If some neurons drop out/in, this could lead to both the change in response magnitude as well as the proportional change in oddball and frequency coding. For the latter point, it would also be useful to know to what extent the relative prevalence of oddball/frequency encoding is true at the single unit level vs the population, since this gives insight into the mechanism of how auditory information is being tracked.

2) The inclusion of the RNN model is a strength of the paper, but one that does not seem fully realized. It seems that the main claim from the RNN is that “the divergence of representations can be due to local computations rather than inherited from other networks”, yet this is not necessarily true depending on how one views the analog of the units in the network and the brain. What is the justification for saying the RNN only models what’s happening in the hippocampus, or that the hippocampus is just one network (see above point about sub-populations of neurons)? Similarly, it would be more interesting if the RNN could be used to explain a specific mechanism for how and why oddball sensitivity emerges from a model trained only on frequency discrimination. This would help allow the paper to make a claim about what the hippocampus is doing in processing statistically-structured sensory input.

3) The finding that hippocampal neurons can discriminate among both semantic and grammatical categories is interesting, but as written, this section of the paper feels too descriptive and it’s hard for me to understand the implications of these results. For example, is there some sort of relationship between the type of semantic and grammatical information encoded within an individual neuron or population that would tell us more about the representation? Is there anything about the structure of the semantic space in hippocampus that could be related to previous work on cortical semantic maps (e.g., Huth et al, 2016)? Even at a more basic level, if a neuron discriminates among 5 or even 10 semantic categories, what does that mean as far as its selectivity for a particular semantic feature? Do these neurons primarily just “structure the space” without actually coding for individual features, or does selectivity to specific features or semantic dimensions somehow give rise to this kind of multi-category discriminability? What is missing is a coherent model of what these neurons are actually doing when listening to language.

4) The comparison between neocortical (MTG) and hippocampal recordings is interesting and potentially important for helping the paper make more of a contribution to understanding the role of the hippocampus in both sensory and linguistic processing networks. However, the cortical recording here is just 2 insertions from a single patient, so it’s hard to know how strongly to interpret these results. As is, the paper limits claims about comparing regions to firing rate and motion, and in my view this is a distraction from the more interesting results, so I would remove it. At the same time, to the extent that they can leverage other data (for example previously published data from the co-authors at MGH), it could be a great opportunity to include additional datasets to make a more complete comparison. Of course, the data from other cortical regions may not be collected using the same task or anesthetic state, so they should consider whether this makes sense.

5) Regarding point 4, what is the model we should take away from these results, in particular as it relates to how the hippocampus should be incorporated into networks involved in language processing (see e.g., work by Melissa Duff)? My understanding is that even bilateral hippocampal lesions don’t lead to specific language deficits, so the question is to what extent the results here suggest something distinct from other cortical areas.

Some more specific points that I think should be addressed:

6) The point about word frequency being a significant predictor of firing rate is interesting, but there are two questions. First, why is the relationship between frequency and FR positive, when the majority of prior work shows a negative correlation? Second, it's not clear to me why they are focusing on frequency over duration – was the duration term in the linear model also significant? Both are interesting to find here, and potentially point to related but distinct features of the words in the stimulus.

7) In figure 2, it's not clear to what extent the early and later components of the response are different subpopulations of neurons. For example, in Fig 2B, there are 2 peaks, and the methods say the ITI was 1-3s (longer than the time window here), so what does that second peak reflect?

8) Also in figure 2, why does the firing rate start to rise before the onset of the tone? Is this just smoothing for the PSTH (though it's also observed in the LFP and gamma plots)?

9) Fig. 1D: what are the x- and y- axes? Relative depth and the x-coordinate on the probe? Also, do they make anything of the lack of positive spiking waveforms? These have been observed pretty consistently in a lot of the recent human single unit recordings, so I wonder if this is a difference between hippocampus and cortex.

10) Similarly, I did not see any details about putative cell types, unit depth (potentially mapped onto specific cellular layers/strata in the hippocampus), or other common characteristics.

11) In Fig 3A, it seems like there should be 4 lines plotted, but I can only make out 3 lines.

12) Fig. 4M: how is this related to word surprisal? It seems like if this truly reflects prediction and/or online language processing, this effect should be modulated by how likely each word is in context.

13) For the RNN, in Fig 3I is chance really 50%? It seems like a model that always just predicts the standard tone would achieve 80% accuracy.

14) The sample sizes are quite small, but given that they recorded from both left and right hippocampus, is there any evidence for laterality effects, particularly for the language task?

15) Lines 237-241: this seems like too strong of a claim without more specific controls for both conscious state and memory consolidation. It's an interesting hypothesis, but I don't see what the evidence is that it's actually related to memory consolidation.

16) Fig 1C: How is the probe included in the micro CT reconstruction if it was retracted from the tissue prior to resection? My understanding is that the site is identified as part of the resection margin, the recording is done, and then the probe is removed so that the resection can take place. Is this not the case here?

17) Lines 670-672: was PCA applied to the word vectors or neural data?

Version 1:

Reviewer comments:

Referee #1

(Remarks to the Author)
Katlowitz et al Plasticity and Language in the Anesthetized Human Hippocampus

This is an outstanding response to our concerns. Given the novelty of this work there are a few additional issues to address:

1. The Neuropixel data is exciting as note in the initial review. But the main finding of the paper is on the role of the anesthesized hippocampus in neural processing. As such the findings that various LFP bands also provide compelling evidence regarding the hippocampus and non-conscious processing should also be clearly noted. For instance, the abstract might say:

Using high density Neuropixel microelectrodes and LFPs…. And hippocampal neurons and oscillations… provided complimentary evidence for …..

2. The brunt of the LFP data is in the supplement even though it supports the main claims of the paper. In this regard we suggest that Fig S3 be included in the main text. It seems the 3 patients had recordings in the oddball task but only P5 and P6 LFP data are in the supplemental data. Please include P4.

3. Decoding accuracy is consistent across bands (Figure S5), suggesting that an aperiodic slope change might be carrying the signal rather than a specific oscillatory band. Have the authors examined this possibility?

4. The authors state in the rebuttal that:
In summary, we found that LFP features were sufficient to decode stimuli in both tone and language tasks, though with a generally lower performance, particularly in language, compared to that of units.

But in the text of the manuscript they only say: Analyses of the LFP corresponding to the ones presented below are found in Supplementary figures S10-S14. Please discuss the LFP language results in the main text.

5. The authors were asked to evaluate SUA and high gamma phase amplitude coupling and responded they preferred not to since it was a thorny topic and subject to another paper in progress. PAC is robust across species and given this paper is likely to be a classic in hippocampal function it would be nice if the authors reconsidered their position.

6. In the manuscript, the authors state:
we expect our units to be drawn from CA1 and dentate gyrus.

Not sure how this expectation arose since in the rebuttal they state:

As far as cellular layers/strata, this information is not given by the Neuropixel electrode. Intraoperative navigation and post operative histology have confirmed the placement of the Neuropixel within the hippocampal head, yet the exact strata and cellular composition vary significantly by depth and superior-inferior position. The exact currently developing novel anatomical reconciliation approaches to address these concerns in the future.

The strong expectation needs to be altered to maybe the hippocampus proper since CA3 might also be a source. So, maybe just say hippocampal head or dorsal/anterior hippocampus.

7. There is a small typo in the Figure S12 caption (LEP instead of LFP)

Bob Knight Eduardo Sandoval

We were Reviewer 1 in the initial review

Referee #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

Referee #3

(Remarks to the Author)
The authors have made meaningful improvements to the manuscript, including adding two additional patients for the surgical operating room paradigm and incorporating microwire EMU data from ten patients. The inclusion of LFP analyses—for example, high-gamma oddball decodability in Fig. S2—also strengthens the work, as does the use of mixed effect model approach. There are still a few remaining issues that warrant attention.

1. My main concern remains that central claims rely on effects that are statistically robust but relatively subtle in magnitude and not readily appreciable in the raw data—particularly at the single-trial, single-unit level. While statistical significance is key, effect size and response magnitudes also carry interpretational weight. Along this line, the reported proportion of responsive hippocampal neurons (~70%) raises the possibility that the threshold may be too permissive. The use of a $p<0.05$ sign-rank test across many hundreds of neurons, without adjusting for multiple comparisons, may contribute to this. Even in the EMU microwire dataset (e.g., Franch et al., Fig. 1C), responses appear modest, and the examples shown—presumably among the clearer ones—still show only moderate increases in firing. It would be very helpful if the authors could report whether the key findings remain stable when applying more conservative thresholds to define responsive neurons (e.g., $p<0.01$ or $p<0.001$).

2. The distinction between “tone frequency” and “tone identity” remains somewhat unclear. Although the authors changed terminology in the main text to “tone identity,” references to “frequency” persist in several figures (e.g., Fig. S4, S5). It would be valuable to know whether responses—both spike rates and high-gamma power—are consistently stronger for one tone than the other across patients and across channels/units. This would help clarify whether the two tones were matched in intensity. While the constraints of the operating room setting are understandable, such information is essential for interpreting decoding performance. For instance, high-gamma distinguishes the two tones at ~0.7–0.8 accuracy; to what extent does this indicate robust hippocampal discrimination under anesthesia, and to what extent may it reflect differences in stimulus intensity?

3. The abstract still contains some formulations that risk overstating the conclusions. Slightly more measured wording could strengthen the manuscript. Examples include:
• Instead of “high-order perception,” a term such as “higher-order pattern recognition” (suggested by reviewer #1), “sensory and linguistic processing without awareness” (suggested by reviewer #3) or the authors’ own phrasing (“higher-order functions associated with parsing semantic and syntactic features of natural speech”) may be more precise.

• I agree with the other two reviewers that some hidden technical factors may contribute to dynamics of oddball discrimination during the session. It could reflect plasticity/learning, but it would be better to tone down the second sentence "Here we demonstrate the persistence of plasticity… under general anesthesia."
• The statement that "hippocampus neurons could reliably detect oddball tones" may overstate the strength of the decoding (0.52–0.56). A more neutral phrasing—such as "hippocampal neurons and field potentials retained some detection of oddball tones"—may better reflect the results.

4. Finally, the qualitative and quantitative similarity of responses in awake and anesthetized conditions raises an interesting broader question: which neural signatures actually differentiate conscious from unconscious processing? It may be worth including a brief discussion. The current findings nicely show that such differences are not evident in hippocampal activity. This contrasts with studies in the visual domain showing clear distinctions in hippocampus unit activities between conscious and unconscious perception (e.g., in flash suppression Kreiman et al PNAS 2002, or backward masking Quian Quiroga et al PNAS 2008). Additionally, intracranial human studies comparing auditory responses across wakefulness, propofol anesthesia, and sleep (e.g., Krom et al., PNAS 2020; Hayat et al., Nat. Neurosci. 2022) are highly relevant and should be cited. These studies report attenuated responses in high-order auditory cortex in unconscious states. Taken together with the current results, this may suggest two partially independent processes: (i) higher-order pattern recognition within the hippocampus that persists irrespective of consciousness, and (ii) a cortical auditory hierarchy that is more sensitive to conscious perception and perturbed by anesthesia. A short discussion of this interpretation would add valuable context.

Referee #4

(Remarks to the Author)
I appreciate the authors' revisions and new analyses/data. I think the changes have improved the manuscript and clarified some key issues. I particularly think that the ablation analyses of the RNN model provide important mechanistic insights into oddball effects.
There are, however, some remaining issues that I think need to be addressed.

1) This information is buried in the initial rebuttal, but as I understand it, the oddball effect is significant only in 1 out of the 2 patients. Even though it is significant when combining data across patients, this fact needs to be made clear in the text, and I think appropriate caveats should be included so readers are aware that this already small effect is not as strong as it may seem.

2) Fig. 4M: they claim in the text that "overall decodability is greater under anesthesia of future words than for awake patients". While I think I can see this at word 0, it's not clear to me that it's the case for past or future words, and no statistical test is used to back up this claim. In fact, if this is true, then doesn't that suggest something very surprising is happening? How could it be that prediction of future words is BETTER in an unconscious state? While not an impossibility, this raises red flags for me that suggest either the data are noisy (maybe n is too low for the anesthetized group), or these neurons are actually encoding something else.

3) The point about a positive correlation between firing rate and word frequency requires more explanation. They argue in the rebuttal that this may be a difference between single neuron and fMRI/MEG/EEG, though this isn't supported. Indeed, single neuron prediction error for auditory stimuli is quite well established (for just one recent example, see https://www.nature.com/articles/s41467-017-02038-6). I'm not arguing with their findings, but I don't think it's sufficient to just brush it off and assume that the single neuron result is "truth".

4) The point about smoothing kernels used to compute PSTHs doesn't seem to address the underlying question that was raised in the first review. Whether the filter is causal or non-causal (and I agree that there is no perfect solution), they show the individual spikes in Fig. 2E for that example neuron. The higher spike density for oddball starts before t=0. Since they used a variable ISI, I find this very surprising, both that there could be such a fast or even predictive evoked response (in hippocampus, no less), and also that it's specific to oddball stimuli. While I don't think the core claims of the paper are dependent on this timing question, I think that this will be noticed by readers and raise questions that (perhaps unfairly) distract from the main points.

5) While I think it makes sense to change the framing from "learning", simply replacing that term with "plasticity" doesn't solve the underlying problem. I still think they haven't ruled out plausible alternatives for the changes they see over the course of the experiment. In fact, something like stimulus specific adaptation (SSA) could explain the decreased tone identity decoding, which itself could be completely independent from the very modest increase in oddball decoding. They now argue in the Discussion that it's not likely to be something like adaptation because those occur over faster timescales, but I'm not sure this result from sensory cortex is known to translate to hippocampus. In fact, looking at the data in Fig 3D (ignoring the linear fits), I wonder how strong of a negative correlation there is between tone and oddball. If anything, the linear trends seem misleading, as both effects appear to be non-monotonic, and ultimately not be related to each other.

Version 2:

Reviewer comments:

Referee #1

(Remarks to the Author)
The authors did a thorough job of addressing our input. We would suggest addressing a few additional points.

The authors report: For each trial and channel, we computed the modulation index (Tort et al. 2010) to quantify phase-amplitude coupling of gamma (30-80 Hz) to theta (4-8 Hz) power.

However, high gamma at 70to150 or 200 Hz best correlates with SUA. The authors should analyze PAC for Theta-High gamma not just low gamma.

The spike-frequency coupling was only analyzed in the oddball experiment, but not the speech experiment. Spike-frequency coupling is typically related to a behaviorally relevant task. If the claim is that the 'task' is pattern recognition, and they're getting better over time, then one might expect that spike-frequency coupling is stronger in the second half of the experiment vs. the first half of the experiment for the oddball task only for units that encoded features above chance (neurons not encoding those features wouldn't make sense to have stronger coupling). One might expect this to not be the case in the language experiment, and instead for coupling to generally be stronger for neurons encoding semantic categories across the recording session. This analysis should also be performed for theta-HFA.

Referee #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

Referee #3

(Remarks to the Author)
The authors have addressed the concerns.
Congratulations for this original study that will surely make a strong impact and spark future work

Referee #4

(Remarks to the Author)
The authors have addressed most of my concerns in this revision.

Before they finalize the paper, I suggest they double check the results related to the point I raised previously about neuronal firing starting before stimulus onset in the tone paradigm. While it helps to have replaced the example neuron in Fig 2E with one that shows a more plausible raster, the plots in F, G, and H have the same issue that was raised earlier. I suppose it's possible that the mean spike rate and gamma amplitude responses start before t=0 due to the filtering issue they brought up, but I don't see how this is possible for panel G, which just shows the raw voltage traces. Unless I'm misunderstanding what they mean by "ERP", this should be a largely unfiltered signal except for hardware lowpass and anti-aliasing. Even though it's an average across multiple channels on the probe, I don't see an explanation for this implausibly early response. Furthermore, since it's there in the raw data, this suggests to me that the reason it appears in the filtered and smoothed versions of the data is not simply due to the signal processing.

One additional minor point they should double check is the p6 oddball delta result in Fig. 2K. It's hard to tell with the notches on the box plot as it's rendered, but it's surprising to me that the effect is significant, especially compared to others in the same plot that are not (e.g., p5 alpha oddball).

Version 3:

Reviewer comments:

Referee #1

(Remarks to the Author)
The authors have done an excellent job addressing all our concerns.

We would like to congratulate the authors for this timely and outstanding scientific contribution.

Robert Knight and Eduardo Sandoval

(Remarks on code availability)

Referee #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

(Remarks on code availability)

Referee #4

(Remarks to the Author)
The authors have addressed my concerns.

(Remarks on code availability)

Reply to Reviewers

We thank the reviewers for their careful reading of our manuscript and their constructive suggestions. In response to these suggestions, we have made significant additions and changes, which we feel have greatly increased the impact of the findings. Our responses to each comment are indented and in bold face.

Thank you for your patience while we waited for our referees to return their reports on your manuscript, entitled "Learning and language in the unconscious human hippocampus". It has now been seen by 3 referees, whose comments are attached below. While they find your work of potential interest, as do we, they have raised important concerns that in our view need to be addressed before we can consider publication in Nature.

We appreciate your interest in our manuscript. We have carefully addressed all reviewer concerns, and we believe the manuscript is now much improved.

Should further experimental data and analyses allow you to address especially the concerns raised by R2 and provide strong and compelling support for your conclusions, we would be happy to consider a revised manuscript (unless something similar has been accepted at Nature or appeared elsewhere in the meantime).

Our revised manuscript includes two new kinds of data to address the concerns raised by Reviewer 2.

- **<u>First</u>, we include data from two additional participants who underwent hippocampal Neuropixel recordings for the language task (doubling the number of participants from whom we collected these rare data). Our new data robustly replicate and strengthen our original claims.**
- **<u>Second</u>, we include a new comparison group of 10 participants in whom we performed single unit hippocampal recordings using microwire electrodes while they were listening to the same podcasts but while awake and undergoing seizure monitoring in the epilepsy monitoring unit. These recordings in awake participants provide direct quantitative comparison to our recordings from anesthetized patients, as specifically requested by the reviewer. In short, the data from awake patients resemble those from anesthetized patients, both**

quantitatively and qualitatively, lending further credibility to our original claims.

We also extend our previous analyses by including 14 new supplemental figures, providing intriguing new findings from the analyses suggested by the reviewers.

Referee #1:

General Overview:

This is an interesting paper showing that the brain performs pattern recognition in both the sensory and language domains in the absence of consciousness. The authors were able to mitigate current technical problems with inserting neuropixels probes into deep structures by taking advantage of two stages of an anterior temporal lobectomy for pharmacoresistant epilepsy. In a clever design, anterior lateral temporal cortex was first removed to gain access to the amygdala and hippocampus. The authors then pause the surgery for 30 minutes and inserted the neuropixels probe into the hippocampus and commence the recording to either an oddball sequence (n=3 patients) or podcasts (n=2 patients). The full anterior lobectomy including the amygdala, hippocampus and adjacent medial temporal structures was then performed. Note the patients consented to this 30-minute surgical pause to enable the auditory recording scenario. The authors report that under Propofol anesthesia the hippocampus detects change in the classic oddball/mismatch negativity task. Notably the authors also report that the oddball response increases with the length of study implying a learning effect. They additionally found that the hippocampus was processing language at the level of detecting both semantic and grammatical information. The manuscript provides the first evidence of the value of Neuropixels in deep structures and given the important claims of this paper several issues need to be addressed. The paper also provides an opportunity to define the roles of not only SUAs but also high gamma in decoding and how they compare and also address the role of lower frequency oscillations in SUA and HG tuning (see below)

This summary precisely describes our manuscript. We appreciate the kind words and have done our best to refine the manuscript to address the Reviewer's concerns.

The authors need to be clear that their results do not necessarily extend to all forms of unconscious behavior such as sleep and are restricted to Propofol. A more accurate title might be Learning and language in the anesthesized human hippocampus.

The reviewer is entirely correct. There are multiple forms of unconsciousness and our paper does not address all of them. We have revised the title of our manuscript accordingly. Due to other reviewer concerns (found below) we have also removed the word "learning" from the title. The new title is:

Plasticity and Language in the Anesthetized Human Hippocampus

Figure 1J shows robust oddball responses for SUAs, ERPs and high gamma. The decoding accuracy seems highest for the high gamma response. Is this true?

Yes, that is true, although not statistically significant. The SVM decoding accuracy as analyzed in Fig. 2J was highest for high gamma in both tasks. However, the difference between high gamma and other groups was significant for tone, but not oddball detection. Specifically:

Tone: median across-fold SVM decoding accuracy of 0.7229 for high gamma vs. 0.6519 for units (p=0.0038) vs. 0.6523 for ERP (p = 0.0149).

Oddball: median across-fold SVM decoding accuracy of 0.5308 for high gamma vs. 0.5138 for units (p = 0.578) vs. 0.5183 for ERP (p = 0.601).

This lack of significant difference means we cannot draw firm conclusions from the difference between measures. However, we would emphasize that the point of the paper is not comparing LFP to units. Indeed, that comparison would better be done in more well understood regions with more well understood signals (like visual responses in visual cortex). Instead, the goal is to demonstrate that this information is available in the brain even during anesthesia.

Summary of the relevant results:

	Units	ERP	High Gamma
Median accuracy (tone)	0.6519	0.6523	0.7229
CI of difference vs. high-gamma (tone)	[0.0338 0.1097] p = .0038	[0.0291 0.1173] p = .0149	N/A

Median accuracy (oddball)	0.5138	0.5183	0.5308
CI of difference vs. high-gamma (oddball)	[-0.1011 0.1260] p = 0.5776	[-0.1163 0.1417] p = 0.601	N/A

Given that sEEG research now cannot employ Neuropixels other that in cortical sites some discussion of the relationship of SUAs and high gamma is warranted given increasing evidence that high gamma is a proxy for nearby SUA activity and can be used for decoding numerous cognitive phenomena as well as being used for BCI (see Leonard et al, Nature 2024).

This also raises the question if the results in Fig 3 which are focused on SUA activity and an RNN model could also be obtained with their high gamma findings.

This is an interesting point and clearly one of current interest. We now directly compare the information content and confirm that the high gamma activity can differentiate oddballs at rates higher than unit activity within this current paradigm. We include this information in the Supplement. In addition, we replicate all the paper's core analyses using all bands and include those results in the Supplement.

In short, the major results are mostly (but not entirely) reproduced using both gamma and high gamma, and results are less consistent using other bands.

For convenience, the major new additions are reproduced here:

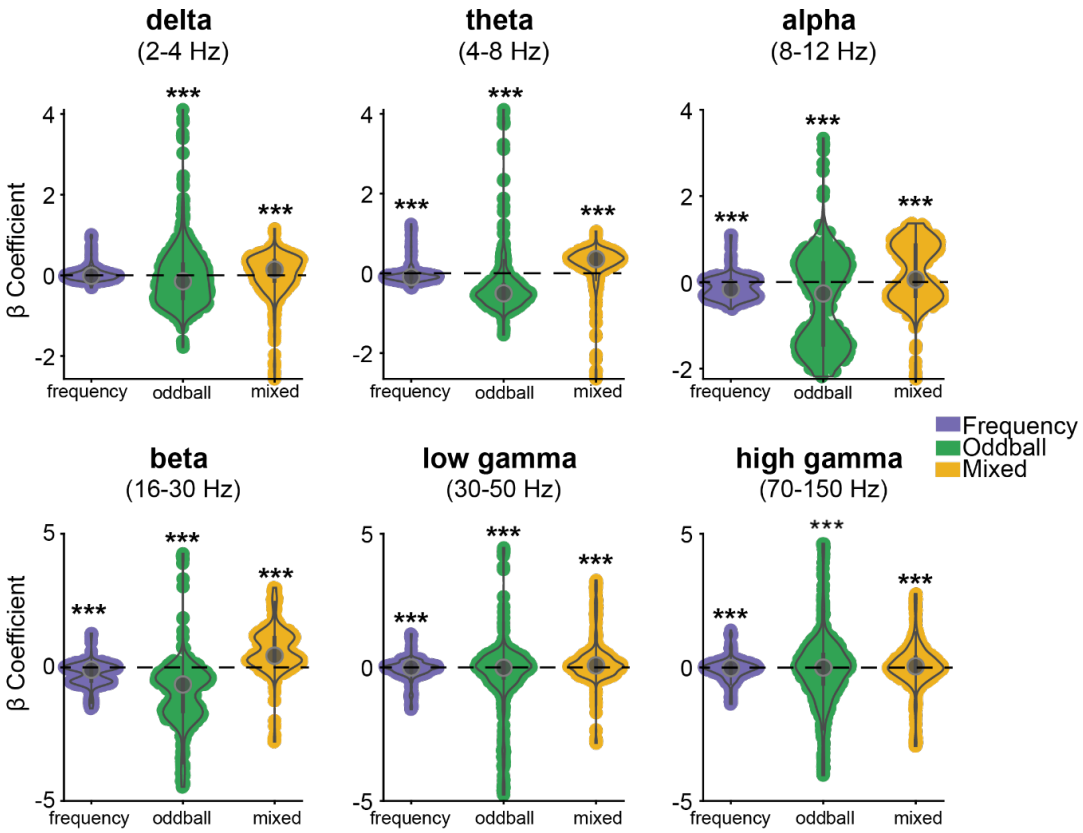


Figure S1. Violin plots showing the distribution of β coefficients obtained from a linear regression model run per channel for each frequency band, to determine response modulation as a function of tone identity (tone frequency β, purple, left), oddball identity (oddball β, green, middle), and an interaction/mixed term (mixed β, yellow, right). Asterisks denote statistical significance determining if the distribution of individual β coefficients is distinct from a distribution with a zero median value (nonparametric Wilcoxon's signrank test). *** denotes p-value < 0.0001.

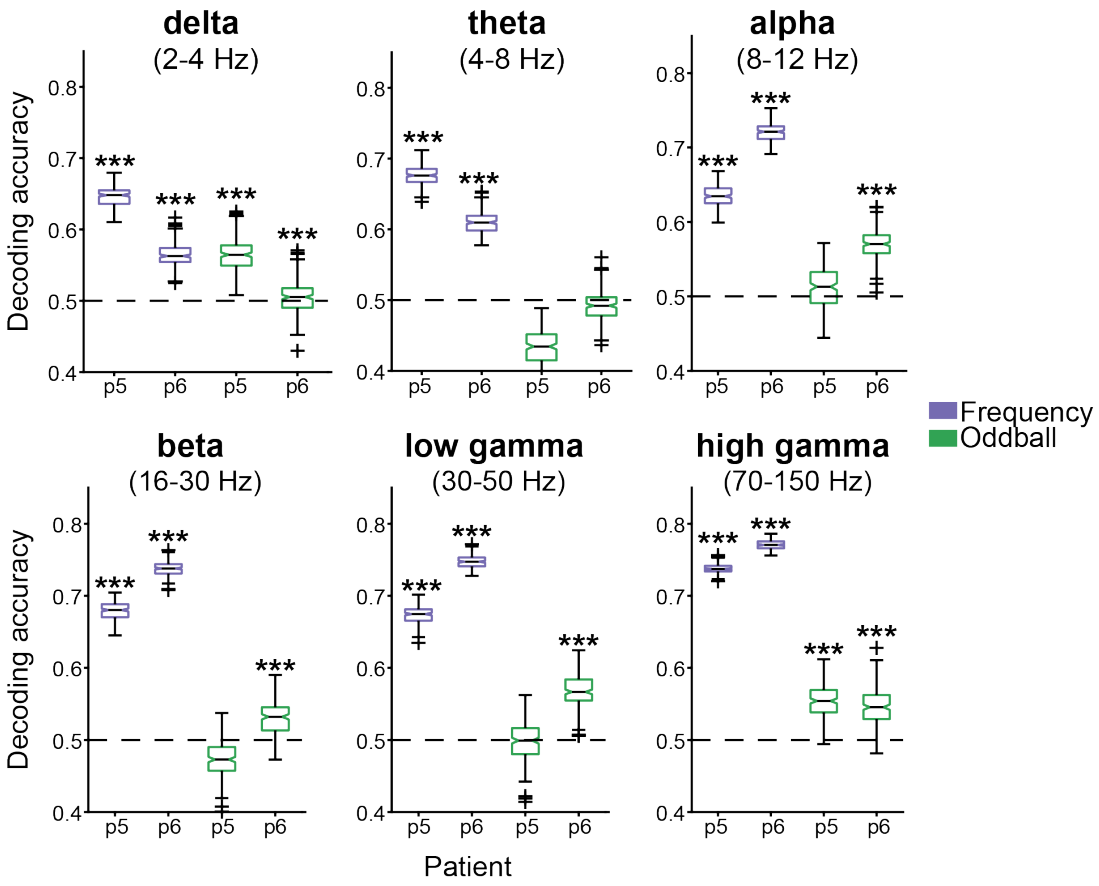


Figure S2. Box and whisker plots of decoding accuracy for tone identity (purple) and oddball identity (green), obtained using a Support Vector Machine (SVM) decoder for P5 and P6, across the population of recorded channels within each pre-defined frequency band. Pluses are outliers and asterisks denote statistical significance relative to change. *** denotes p-value <0.0001.

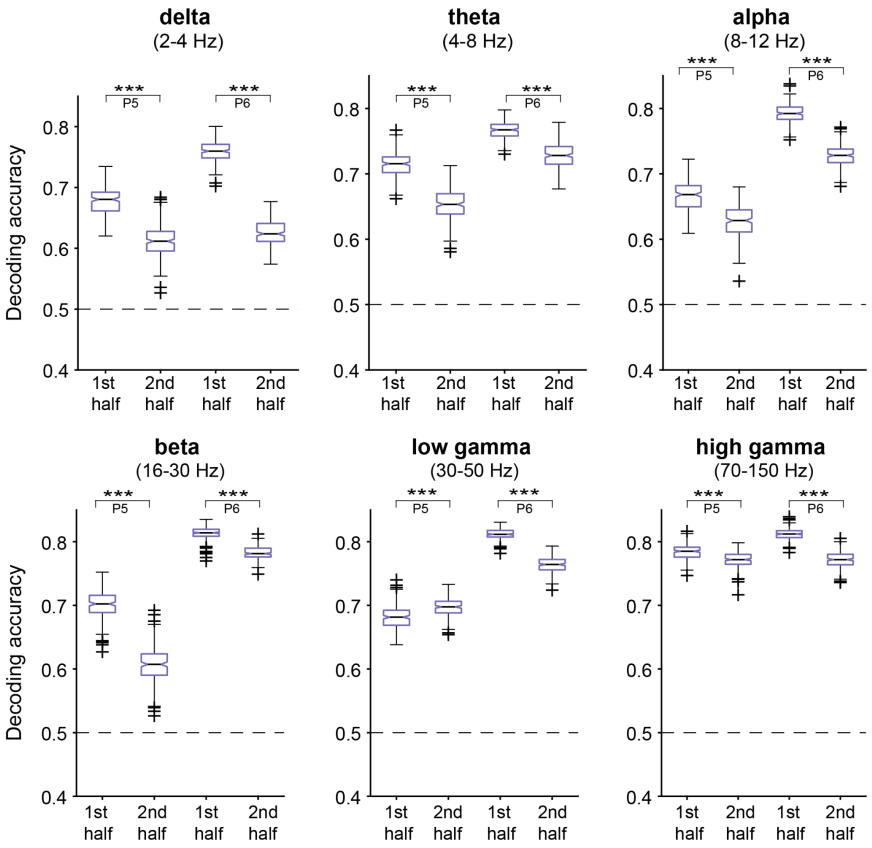


Figure S3. Accuracy of tone identity decoding across the population of recorded channels within each pre-defined frequency band for patients P5 and P6, for the first half trials (left) or second half trials (right), combined across both blocks. Statistically significant differences indicated with an asterisk. *** denotes p-value <0.0001

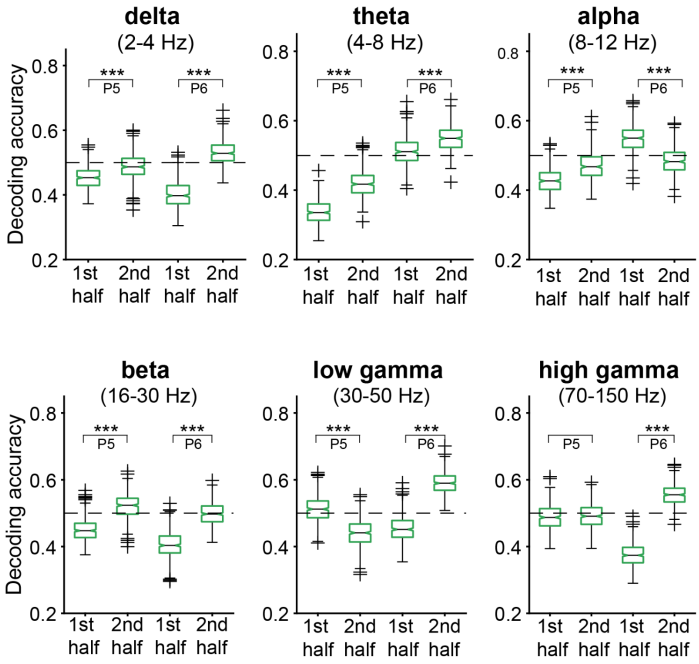


Figure S4. Accuracy of oddball identity decoding across the population of recorded channels within each pre-defined frequency band for patients P5 and P6, for the first half trials (left) or second half trials (right), combined across both blocks. Statistically significant differences indicated with an asterisk. *** denotes p-value <0.0001.

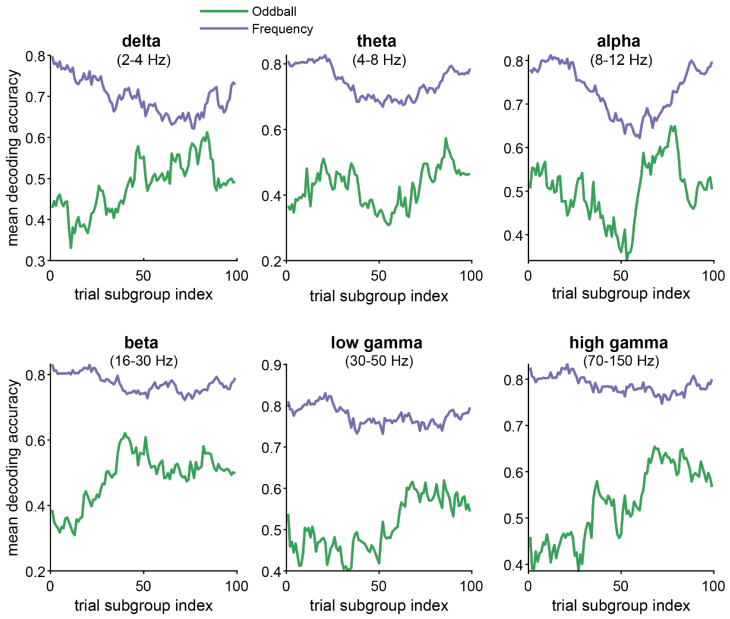


Figure. S5. Decoding accuracy as a function of trial position across both patients. Each point represents SVM accuracy within a set of 50 trials starting at the index location. Decoding accuracy for tone frequency is shown in purple, and for oddball identity in green.

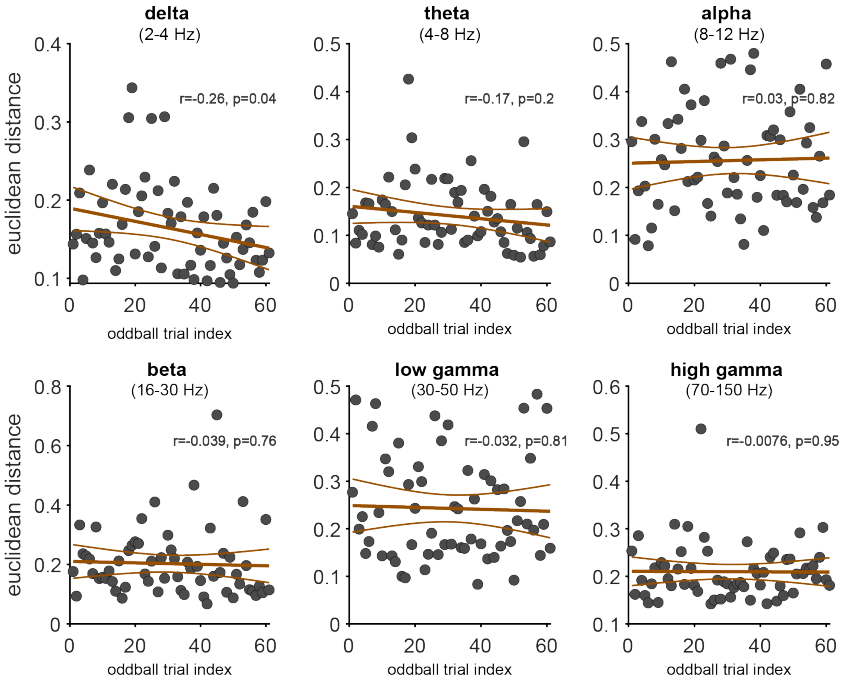


Figure. S6. Euclidean distance between standard and oddball population response vectors comprising all channels across both patients (n=756 channels), computed for each oddball trial, separately for responses within distinct frequency bands. Each datapoint (in black) indicates Euclidean distance per trial, with the lines showing a linear fit with 95% confidence intervals. Analysis of cosine angle (not shown) demonstrated a highly similar trend: for all bands, R was within 0.01 of the equivalent result for Euclidean distance.

The following new text in the Supplemental Results explains these figures:

Z-scored sensory responses for all channels (computed separately for each frequency band) were modeled as a function of tone identity, context (standard vs. oddball), and their interaction using separate linear regression models. We observed comparable encoding for nearly all terms and for each frequency band ($p<0.0001$; n=756 channels): i) delta band (2-4 Hz). 2% of channels showed tone encoding; 4% channels showed oddball encoding; 4% channels showed an interaction; ii) theta band (4-8 Hz). 3% channels showed tone encoding; 3% channels showed oddball encoding; 3% channels showed an interaction; iii) alpha band (8-12 Hz). 26% of channels showed tone encoding; 45%

showed oddball encoding; 35% showed an interaction; iv) beta band (16-30 Hz). 46% of channels showed tone encoding; 45% showed oddball encoding; 45% showed an interaction; v) low gamma band (30-50 Hz). 12% of channels showed tone encoding; 9% showed oddball encoding; 10% showed an interaction; vi) high gamma band (70-150 Hz). 22% of channels showed tone encoding; 20% showed oddball encoding; 18% showed an interaction. Across the population, the distribution of beta weights was significantly different from a distribution centered at zero median for the following (signrank test on beta values, *** denotes $p<0.0001$, Figure S1.): i) delta band (2-4 Hz). βtone, $p=0.1$; βoddball, $p<0.0001$; βmixed, $p<0.0001$; ii) theta band (4-8 Hz). βtone, $p=0.1$; βoddball, $p<0.0001$; βmixed, $p<0.0001$; iii) alpha band (8-12 Hz). βtone, $p=0.1$; βoddball, $p<0.0001$; βmixed, $p<0.0001$; iv) beta band (16-30 Hz). βtone, $p=0.1$; βoddball, $p<0.0001$; βmixed, $p<0.0001$; v) low gamma band (30-50 Hz). βtone, $p<0.0001$; βoddball, $p=0.03$; βmixed, $p<0.0001$; vi) high gamma band (70-150 Hz). βtone, $p<0.0001$; βoddball, $p=0.02$; βmixed, $p=0.002$.

Leveraging the power of large-scale recordings, we used a 10-fold cross-validated support vector machine (SVM) to decode stimulus features on a trial-by-trial level across the neuronal population (n=383 channels for p5, n=373 channels for p6). Tone identity was robustly represented in both patients across frequency bands, with mean accuracy ranging between 0.52 to 0.79 ($p<0.0001$ for all, t-test. Figure S2.). Oddball identity could also be decoded above chance in some frequency bands for both patients ($p<0.0001$ for delta and high gamma bands for p5; and for all frequency bands except theta band for p6), albeit at lower levels, ranging from 0.35 to 0.62 (accuracy rates on shuffled data ranged from 0.47 to 0.55).

Splitting our task into halves, we found a significant increase in oddball encoding for both patients (paired t-test, $p<0.0001$ Figure S4.). Replicating our findings from the SUA, we also observed a concomitant decrease in tone encoding, again raising the possibility of compensatory mechanisms ($p<0.0001$, paired t-test for both) (Figure S3).

Using a sliding window of subsets of 50 trials, across most frequency bands, we found a continuous increase in oddball decoding accuracy across the approximately 10-minute duration of the experiment (Figure S5, purple), accompanied by an initial decrease in tone encoding ($p<0.0001$, Figure S5, green). Linear fits for evaluating the correlation of the decoding of tone identity and oddball identity as a function of time showed: i) delta band (2-4 Hz). βtone = -0.001, $p<0.0001$; βoddball=0.00l, $p<0.0001$; ii) theta band (4-8 Hz). βtone = -6x10^-4, $p<0.0001$; βoddball=6x10^-4, $p<0.0001$; iii) alpha band (8-12 Hz). βtone=-5x10^-4, $p=0.005$; βoddball=4x10^-4, $p=0.01$; iv) beta band (16-30 Hz). βtone=-4x10^-4, $p<0.0001$; βoddball=0.001, $p<0.0001$; v) low gamma band (30-50 Hz). βtone= -4x10^-4, $p<0.0001$; βoddball=0.001, $p<0.0001$; vi) high gamma band (70-150 Hz). βtone= -3x10^-4, $p<0.0001$; βoddball=0.002, $p<0.0001$. These effects are comparable to those observed for single units, further supporting our hypothesis that the neural population was sacrificing its tone responses for the sake of oddball

representations over the course of the experiment, suggesting that the hippocampal responses were shifting to represent the salient features of the stimulus.

We created neural vectors of the average standard tone response as well as each individual oddball trial (756-dimensional vectors composed of the mean response of the oddball units). In contrast to our single unit data results, we did not find a statistically significant divergence between standard and oddball vectors over the course of the session.

We also include new text in the revised Discussion elaborating on this point:

Some aspects of the local field potential correlate with single unit activity, although this relationship may be task-dependent or may fluctuate[38]. Recently, a good deal of evidence has supported the idea that the gamma band, in particular, may have close alignment with single units[39,41]. Our results have some resonance with these ideas; in particular, we find that the gamma band is generally aligned with neural activity, although not perfectly so. Our results also show that some effects have an alignment with other bands or even have a broadband effect. One limitation of our results is that failure to achieve significant effects may reflect an insufficiency of data rather than a true disjunction between unit activity and LFP activity. Thus, it is difficult to draw firm negative conclusions from observed differences between unit and LFP responses. In any case, our results do indicate that LFP activity, especially in the gamma band, is likely to be a useful proxy for single unit responding, although it may have other functional roles as well. Another limitation is related to our modeling. Our modeling does not indicate that the processes we describe are unique to the hippocampus. Indeed, the RNN relies critically on the formatting of its input into binary categories, which must be inherited from other regions. Instead, our goal is to understand how oddball-dependent responses can arise from such input.

This table reports decoding accuracy values computed using all trials (as in the LFP version of Fig. 2J)

	Delta	Theta	Alpha	Beta	Low gamma	High Gamma
median	0.53	0.46	0.53	0.5	0.53	0.55
Ranksum (freq band vs. h.gamma)	p<0.0001	p<0.0001	p<0.0001	p<0.0001	p<0.0001	p<0.0001

A similar query could be raised for Fig 4, could high gamma be used for the linguistic decoding effects reported?

As above, we now present similar analyses for the language analyses below. In summary, LFP features can decode some aspects of language, but with higher variability across task categories and reduced accuracy compared to single units. Our results suggest that low gamma, beta and alpha frequency bands have the greatest value for language. Interestingly, the word frequency effects are reversed when looking at gamma.

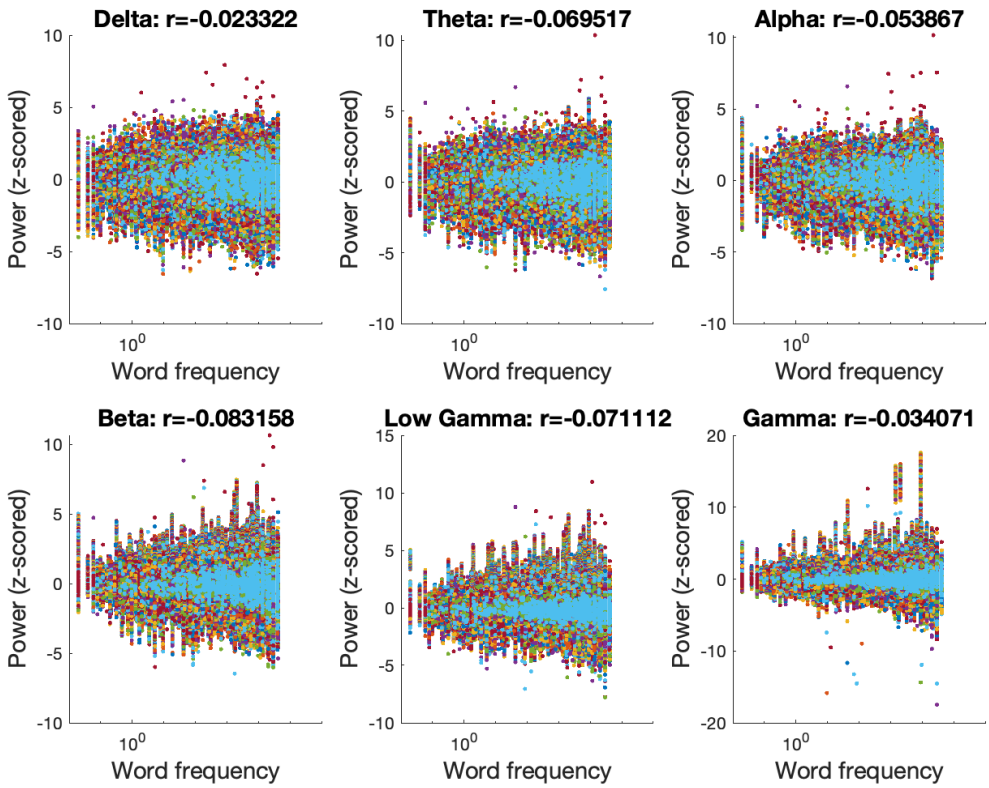


Figure S10. Correlation between word frequency and evoked power. Following analysis of the tone experiment (Fig. S1-S6), Z-scored LFP power features were similarly computed for each trial, frequency band, and channel in the 4 patients that participated in the language task. In all 6 bands, power demonstrated a modest but significantly negative correlation with word frequency ($p<0.0001$; n = 384 channels in each of 4 patients), contrasting the results of single unit analysis.

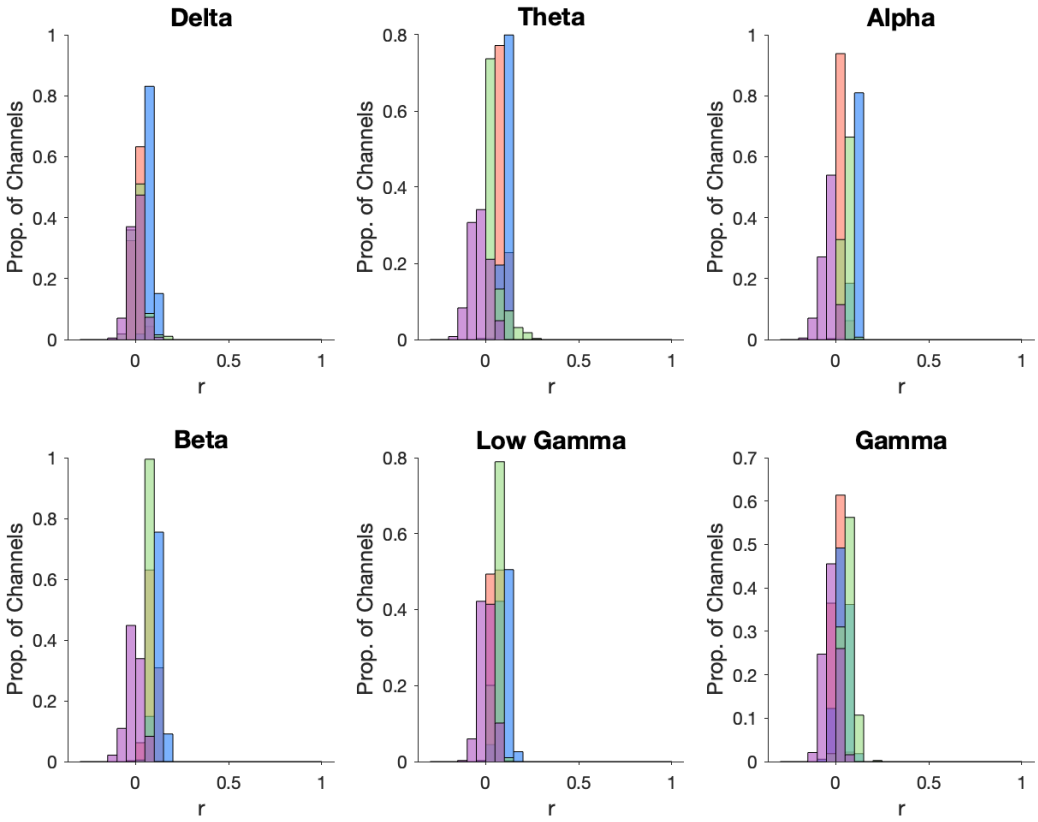


Figure S11. Semantic embedding prediction performance by LFP power band. Each panel recreates the analysis of Fig. 4C, showing the average correlation between true power and predicted power of a linear model regressing LFP power in the given band vs. the semantic embedding of each word, computed separately for each channel and grouped by patient. Colors indicate distributions for 4 different patients (pink, blue, orange, green correspond to 6, 8, 9, and 11, respectively). For all bands, the RMSE of a linear model significantly outperformed models trained on shuffled data, though with reduced prediction performance relative to single units ($p < 0.05$; delta: mean R = 0.029, 46% of channels were significant; theta: mean R = 0.054, 75.8% significant channels; alpha: mean R = 0.039, 73.7% significant channels; beta: mean R = 0.067, 76.4% significant channels; low gamma: mean R = 0.052, 76.3% significant channels; gamma: mean R = 0.021, 39.3% significant channels).

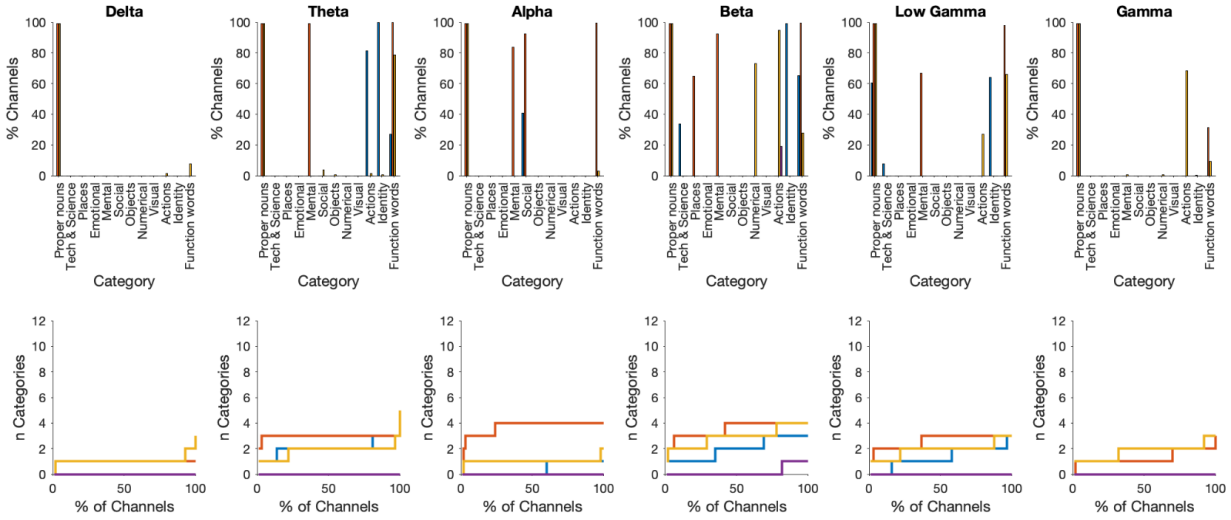


Figure S12. Semantic category prediction performance by LFP power band. Each column recreates the analysis of Fig. 4F and Fig. 4G, quantifying the percentage of channels significantly encoding a specific semantic category vs. other categories. Colors indicate distributions for 4 patients. LFP prediction was overall less significant than single units, with higher variability across patients. Beta and low gamma bands were the strongest predictors of semantic category. Overall, semantic category representation was significant but weaker than that of single units. Category discrimination was greatest in the beta band, followed by alpha, theta, and then low gamma. In the delta band, 50% of channels predicted one category and 2% of channels discriminated between 2 categories. In the theta band, 67% discriminated between 2 categories and 31% discriminated between 3 categories. In the alpha band, 60% of channels predicted 1 category, 26% discriminated between 3, and 19% discriminated between 4 categories. In the beta band, 80% of channels predicted 1 category, 50% discriminated between 3 categories, and 21% discriminated between 4. In the low gamma band, 71% of channels predicted 1 category and 20% of channels discriminated between 3 categories. In the gamma band, 50% of channels predicted 1 category and 25% of channels discriminated between 2. For both types of word feature, category discrimination using LFP was more variable across patients and categories than was observed in the single unit analysis. For example, proper nouns were discriminated in 100% of channels in the alpha band for one patient but 0% of channels in another. This observation may reflect the lower-dimensional signal content of LEP recordings, as well as their high signal correlation across densely packed channels of the Neuropixel array.

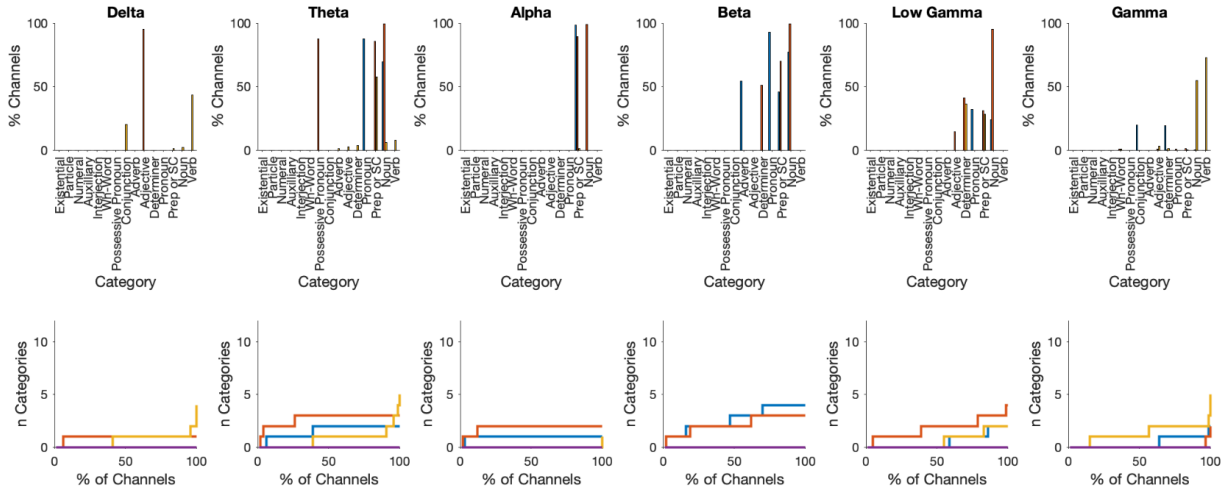


Figure S13. Part of speech prediction performance by LFP power band. Each column recreates the analysis of Fig. 4I and Fig. 4J, quantifying the percentage of channels significantly encoding a specific part of speech vs. other categories. Colors indicate distributions for 4 patients. Part of speech discrimination by LFP power was overall less pronounced than semantic category discrimination but demonstrated comparable trends across bands. Part of speech discrimination was greatest in the beta band, followed by theta, then low gamma, and alpha. In the delta band, 39% of channels predicted one category and 1% of channels discriminated between 2 categories. In the theta band, 42% discriminated between 2 categories and 20% discriminated between 3 categories. In the alpha band, 50% of channels predicted 1 category, and 22% discriminated between 2. In the beta band, 42% of channels discriminated between 2 categories, 23% discriminated between 3 categories, and 8% discriminated between 4. In the low gamma band, 46% of cells predicted 1 category, 24% of cells discriminated between 2 categories, and 6% discriminated between 3. In the gamma band, 32% of cells predicted 1 category and 12% of cells discriminated between 2.

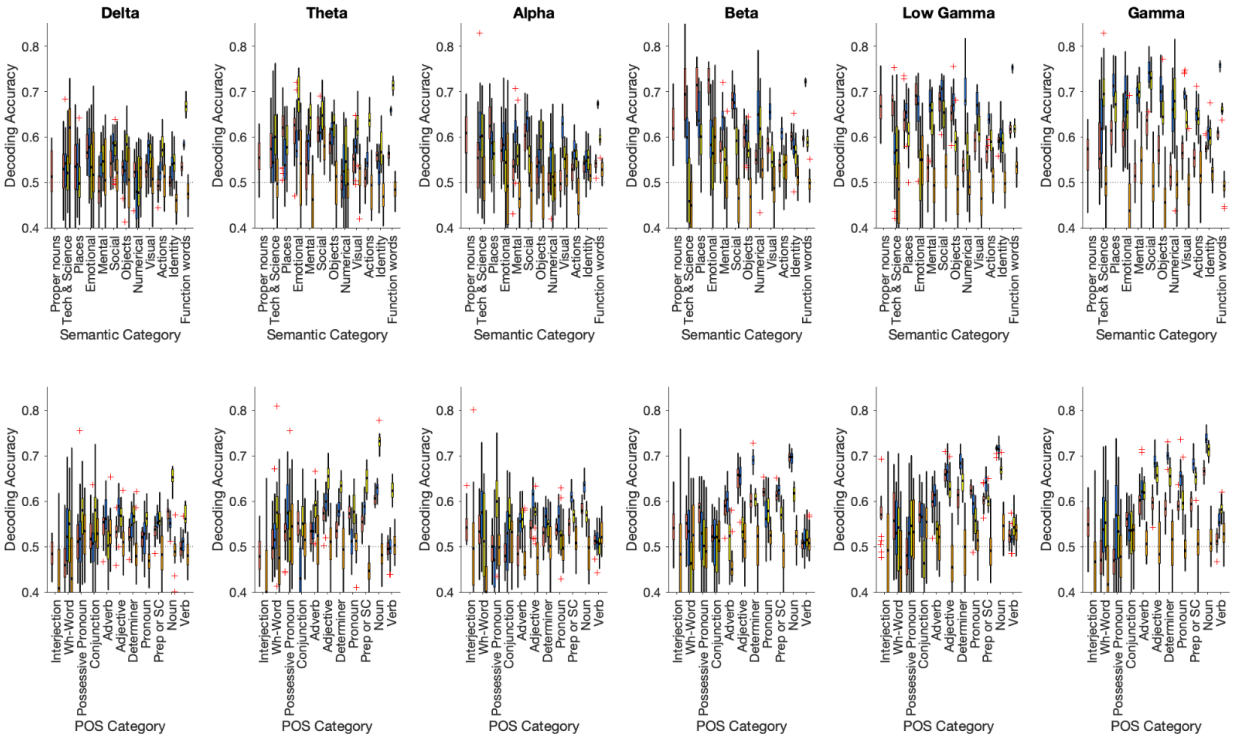


Figure S14. SVM classifier prediction performance by LFP power band. Each column recreates the analysis of Fig. 4K (top) and Fig. 4L (bottom), quantifying classifier decoding performance for each word category from bandpower values across channels. Colors indicate distributions for 4 patients. LFP prediction was comparable to that of single units. Frequency bands in the alpha range and above were the strongest predictors for semantic category and part of speech. Decoding accuracy was significantly greater than 0.5 for all bands, but overall lower than that of single units. Semantic category prediction performance was highest for gamma, followed by low gamma and alpha bands (delta: mean accuracy = 0.520, 72.9% significant categories; theta: mean accuracy = 0.536, 81.3% significant categories; alpha: mean accuracy = 0.555, 77.1% significant categories; beta: mean accuracy = 0.517, 85.4% significant categories; low gamma: mean accuracy = 0.564, 85.4% significant categories; gamma, mean accuracy = 0.572, 85.4% significant categories). Part of speech prediction performance was also significant, though modestly lower than that of semantic category prediction, mirroring results of the single unit analysis. Decoding accuracy was highest for alpha power, followed by gamma and low gamma bands (delta: mean accuracy = 0.515, 58.3% significant categories; theta: mean accuracy = 0.523, 60.0% significant categories; alpha: mean accuracy = 0.564, 60.0% significant categories; beta: mean accuracy = 0.473, 68.3% significant categories; low gamma: mean accuracy = 0.535, 70.0% significant categories; gamma, mean accuracy = 0.536, 71.7% significant categories).

The following new text in the Supplemental Results explains these figures:

For both types of word feature, category discrimination using LFP was more variable across patients and categories than was observed in the single unit analysis. For example, proper nouns were discriminated in 100% of channels in the alpha band for one patient but 0% of channels in another. This observation may reflect the lower-dimensional signal content of LEP recordings, as well as their high signal correlation across densely packed channels of the Neuropixel array.

In summary, we found that LFP features were sufficient to decode stimuli in both tone and language tasks, though with a generally lower performance, particularly in language, compared to that of units. We believe that this discrepancy may arise from the high-dimensional structure of language and lower-dimensional, highly correlated nature of the LFP signal. We observed high heterogeneity of neuronal responses to language, finding that different units can respond to single or multiple semantic categories with specificity - a property that could aid robust prediction of a specific category from the population-level data. In comparison, the reduced performance of LFP may be due to its higher homogeneity: because they reflect the aggregated activity of many neurons and synaptic inputs, different LFP channels are less likely to provide sufficient variability in their responses to stimuli that could aid the prediction of specific language subcategories. In comparison, LFP performance was higher for tones, where the task structure is low-dimensional and uses stimuli of high sensory salience.

The authors frame their podcast results as providing evidence for semantic and grammatical processing in the anesthesized hippocampus. sEEG data has also provided theta band evidence for semantic processing in the hippocampus (Piai at al, PNAS, 2016). This raises the question of whether other bands besides high gamma are engaged during sleep processing.

We now include data for the theta band (as well as all major bands) in the revised Supplement (see above) and cite Piai as well as others within the discussion.

Similarly, abundant evidence has shown theta-SUA and theta-HG phase amplitude coupling in the hippocampus in a range of tasks. The paper would benefit if the authors examined these issues for both SUAs and high gamma.

We agree with the reviewer that the relationship between spike timing and the oscillatory phase is a very interesting question. However, these are thorny issues that require a good deal of additional analysis. Moreover, doing spike/phase or HG/phase alignment is not really something that can

be readily added on - it requires a number of assumptions (such as identifying the optimal boundaries for the gamma and theta bands, which taper methods are best for dealing with spectral leakage, ways to test for significance when there are many ways to get effects). These are all addressable questions, but they would require several figures and would basically make the paper into something else entirely. We are planning to study this topic in a separate manuscript and therefore would prefer to not address it in this study.

Finally, the discussion is very short given the wealth of empirical findings and proposed hippocampal mechanisms.

Our original Discussion was indeed cursory. We now include a 6-paragraph Discussion, which is reproduced here:

We identified neural signatures of plasticity of sensory response and of semantic processing in the anesthetized human hippocampus. These are core cognitive functions often assumed to be absent in the anesthetized state. Our findings do not have obvious explanations based solely on low-level sensory responses. For example, the long and slow increase in oddball detection over the course of 10 minutes is unlikely to reflect adaptation or repetition suppression, which both take place on much shorter timescales. Likewise, the representation of semantic features in speech listening requires specific processing of acoustic information. We therefore show that within anesthetic-induced unconsciousness some high-level process of sensory integration is preserved, suggesting that it is consolidation that is compromised. These results provide a potential explanation for previous reports of post-anesthesia implicit recall, which would depend on sensory processing and plasticity processes despite the absence of consciousness. Broadly speaking, these results confirm ideas previously developed in animal and human studies showing preserved neural responses to sensory stimuli, including oddball stimuli, during anesthesia, and extend them in a few important ways. First, we find a gradual change in representation over the timescale of several minutes, showing that the anesthetized brain is capable of the kinds of integration over those timescales usually associated with wakeful learning. Second, we link those changes to a specific plausible biocomputational model that accounts for the effects without the types of executive control that are presumably diminished during anesthesia. Third, we show that high-level language information is available, beyond the lowest level of auditory processing.

The present results complement our past studies, which were performed using awake patients in the epilepsy monitoring unit (EMU), on hippocampus neuron-level representation of lexical semantics[35] and semantic contexualization[36]. In those studies, we found support for the idea that the hippocampus has a largely vectorial representation of meaning, and that that representation changes in systematic ways with adjacent words. The present results suggest that these vectorial representations of meaning, and at least

some contextualization, can occur in the absence of conscious awareness. That in turn raises the question of how far linguistic processing can go in the absence of awareness[37].

Some aspects of the local field potential correlate with single unit activity, although this relationship may be task-dependent or may fluctuate[38]. Recently, a good deal of evidence has supported the idea that the gamma band, in particular, may have close alignment with single units[39,41]. Our results have some resonance with these ideas; in particular, we find that the gamma band is generally aligned with neural activity, although not perfectly so. Our results also show that some effects have an alignment with other bands or even have a broadband effect. One limitation of our results is that failure to achieve significant effects may reflect an insufficiency of data rather than a true disjunction between unit activity and LFP activity. Thus, it is difficult to draw firm negative conclusions from observed differences between unit and LFP responses. In any case, our results do indicate that LFP activity, especially in the gamma band, is likely to be a useful proxy for single unit responding, although it may have other functional roles as well. Another limitation is related to our modeling. Our modeling does not indicate that the processes we describe are unique to the hippocampus. Indeed, the RNN relies critically on the formatting of its input into binary categories, which must be inherited from other regions. Instead, our goal is to understand how oddball-dependent responses can arise from such input.

While the hippocampus is not a well-known part of the core language network, it has a clear and well established role in language[35,42-44] , including in recent single unit recording studies[35,36,45]. The hippocampus is well-situated to perform the core elements of semantic contextualization: (1) it is closely associated with mapping functions that are related to temporal position encoding[46,47], (2) it integrates information across multiple modalities and uses that information to contextualize representations in a variety of domains[48], (3) and it is closely associated with the processes of prediction that drive contextualization[49,50]. Finally, (4) the hippocampus has a long-established role in memory, one that has many links with semantic knowledge and with semantic representations in particular[51,52]. We suspect some of the reason for the absence of the hippocampus in classical language models is negative evidence - the hippocampus can be difficult to image using fMRI and lesions isolated to the dominant hippocampus are rare. However, another factor is that the hippocampus is clearly not a site uniquely aligned to language coding processes, so it does not fit modern uniqueness-focused definitions of language areas[53]. In any case, we surmise that future work on the functional architecture of language will need to incorporate the hippocampus into its theorizing, specifically for applying rapidly changing contexts.

There are several important limitations for the current study. First, anesthesia, while medically important, has an uncertain relationship with waking life[54]. Moreover, it remains unclear whether what we have learned will apply to other non-conscious states, such as sleep and coma.[55] Indeed, because all of our patients were anesthetized with propofol, it is not even clear our results will generalize to other anesthetics. Another limitation is that we did not have enough patients to test for lateralization / dominance. This means that the effects we observe do not tell us about language processing in the

dominant, and presumably more important, hemisphere. It is well understood that, even in highly lateralized individuals, language is a bilateral process, so it is not surprising that we do observe some language in the non-dominant hemisphere. Another limitation is related to our modeling. Our modeling does not indicate that the processes we describe are unique to the hippocampus. Indeed, the RNN relies critically on the formatting of its input into binary categories, which must be inherited from other regions. Instead, the goal is to understand how oddball-dependent responses can arise from such input. Finally, our choice to use temporally unpredictable tones may have weakened our sensitivity to tone-related effects. There is some evidence that temporally predictable deviant events may maximally drive neural responding[11].

These results highlight the robust coding of cognitive variables in the hippocampus in the absence of consciousness. Such task-correlated patterns of neural activity are typically thought of as neural correlates of their corresponding cognitive processes, so our observations raise the possibility that those processes may occur in the absence of consciousness. That idea, in turn, suggests that consciousness may have some association with those processes but is not a *sine qua non*. Instead, those processes may occur in a subconscious manner, as for example, we may monitor subconsciously others' conversations at a cocktail party[56,57]. A good deal of past research has emphasized the central positioning of the hippocampus within the anatomical hierarchy of brain areas; indeed, it may be among a small number of regions most distant from both input and output ends of brain processing[58]. It is therefore generally assumed that the hippocampus will have the most attenuated inputs in the absence of consciousness; our results also invite a reconsideration of that idea.

Specific Points:

Please define data cleaning procedures to mark epileptic episodes.

There was no data cleaning for epileptic episodes. The patients were under propofol, a powerful anti-epileptic drug, throughout the experiment. There were no seizure episodes during the recording periods. We now indicate this in the revised results:

Because patients were under propofol anesthesia, they did not experience seizures during the surgery, so we did not do any seizure-related data-cleaning. The lack of seizures was confirmed by analysis of the waveform activity by a trained neurologist.

In that regard what was the epileptic source: amygdala, hippocampus, parahippocampus etc. Did the resected hippocampal tissue around the probe show any evidence of tissue abnormalities such as gliosis etc.?

The tissue surrounding the probe was not examined for gliosis, given that it was a ~15 minute recording there was no expectation of gliosis within the recording. However, all tissue was inspected with a microscope by Dr. Heilbronner and her lab, and no abnormalities were reported. We now indicate this in the revised Methods:

All tissue was inspected with a microscope by Dr. Heilbronner and her lab, and no abnormalities were reported.

The authors should mention earlier their use of random SOA (*Stimulus Onset Asynchrony*) (1 to 3 seconds; now specified on line 512) for the presentation of tones in the oddball task as both random and fixed SOA has been employed with the oddball task using SEEG (Tzovara et al., Cerebral Cortex, 2024, cited by the authors when describing the classic oddball task).

We now add this information earlier at the start of the Results:

Tones were presented with random, rather than fixed stimulus onset asynchrony (1-3 seconds, randomized by tone).

The authors should discuss how their choice of a temporally unpredictable sequence may have impacted their low decoding accuracy of oddball identity in light of the same cited publication (Tzova ra et al., 2024) showing the hippocampus is mainly sensitive to predictable deviants (in wakefulness).

We agree it is a good idea to provide more context about this stimulus decision. We now include the following new text in the Discussion:

Finally, our choice to use temporally unpredictable tones may have weakened our sensitivity to tone-related effects. There is some evidence that temporally predictable deviant events may maximally drive neural responding (Tzovara et al., 2024).

The statement that the oddball response grew in magnitude over the course of the experiment needs better substantiation (line 108-109). The authors show an example unit (Fig. 3A) with a higher spike rate towards the end of the experiment and do not show or report results substantiating this statement. The authors focus on decoding accuracy (line 110-111; Fig. 3C and 3D).

We apologize for the misleading writing. Overall firing rates across neurons are relatively unimportant; what is important is the information available in the distribution of firing rates across the population, which is best assessed with decodability. The revised text now makes our point clear:

In oddball-responsive units (n=43), we found that the oddball response grew more distinct (as inferred from decodability) over the course of the experiment (~10 minutes, example unit, **Figure 3A)**.

BIS monitor values were kept between 45 and 60 for the duration of the surgical case. Did BIS values change similarly or systematically for patients during the course of the experiment? A supplemental figure showing the evolution of BIS values over the course of the experiment would be helpful to support the authors interpretation of the change in decoding of oddball trials as a result of learning rather than depth of anesthesia.

Unfortunately, we do not have access to these records. However, it is very unlikely that BIS value changes, even modestly, over the course of the experiment, given that the patient is closely monitored, the experiment duration (~15 minutes for the active recording part) is short relative to duration of the surgery, and the anesthesiologist is attempting to keep the patient stably in a single plane of anesthesia. We acknowledge that we did not make this clear at first. We now do so in the revised Methods:

Specific anesthesia depth was flat across the brief time of the experiment. First, recordings took place several hours after anesthesia induction and several hours before the end of the procedure, so patients were well into the stable portion of the surgery. Second, the anesthesiologist was maintaining active monitoring and stably controlled anesthesia levels.

Fig. 3A: Trace for spike rate for oddball trials 1-20 appear to be missing.

We apologize for this error and we have now fixed the issue. The revised figure panel is reproduced here.

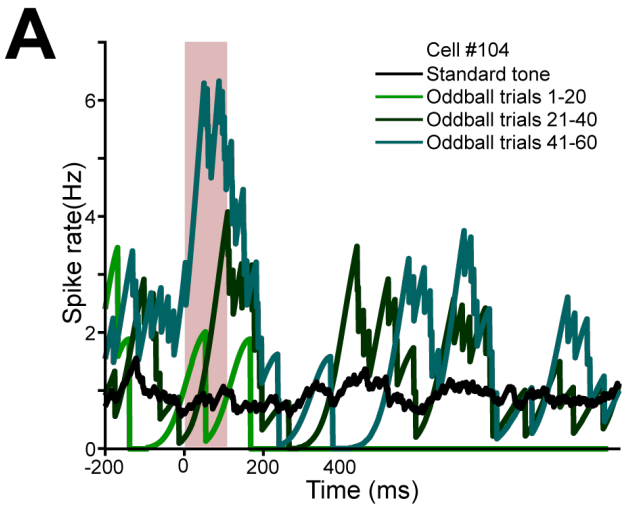


Figure 2b - intertrial interval is 1-3 seconds, but there are multiple peaks much after tone presentation but before 1 seconds. How many tones were in each trial? How many oddballs? This is unclear from the methods.

> **The reviewer is correct that the results show a clear biphasic response. That response is not due to a second tone; instead, it is due to the temporal structure of the neural response, which has two clear phases. We have now added a sentence indicating this to the results.**
>
> In general, these responses had a biphasic temporal envelope, meaning that they had distinct early and late components.
>
> **In addition, we have added the following information to the revised Results:**
>
> Each trial consisted of a single pure tone, played for a duration of 100 ms. The minimum possible inter-tone interval was 1 second, so additional peaks following each tone are not the result of other stimuli occurring shortly after the tone. The oddball tones were presented randomly with a pre-defined probability of 20%. We played 300 tones total; 80% (n=240) were standard; the remainder were oddballs.

Spike sorting - Was there any attempt to verify that the units resemble a biologically plausible action potential. This is not clear from Fig 1 D. Examples in the supplement would be nice.

> **This is a good question, and we appreciate the opportunity to make the quality of isolation observed clearer. In short the quality of isolation is very**

high (consistent with what others have seen with Neuropixels in animal models). We now provide example spikes in the revised Figure 1. They are reproduced here:

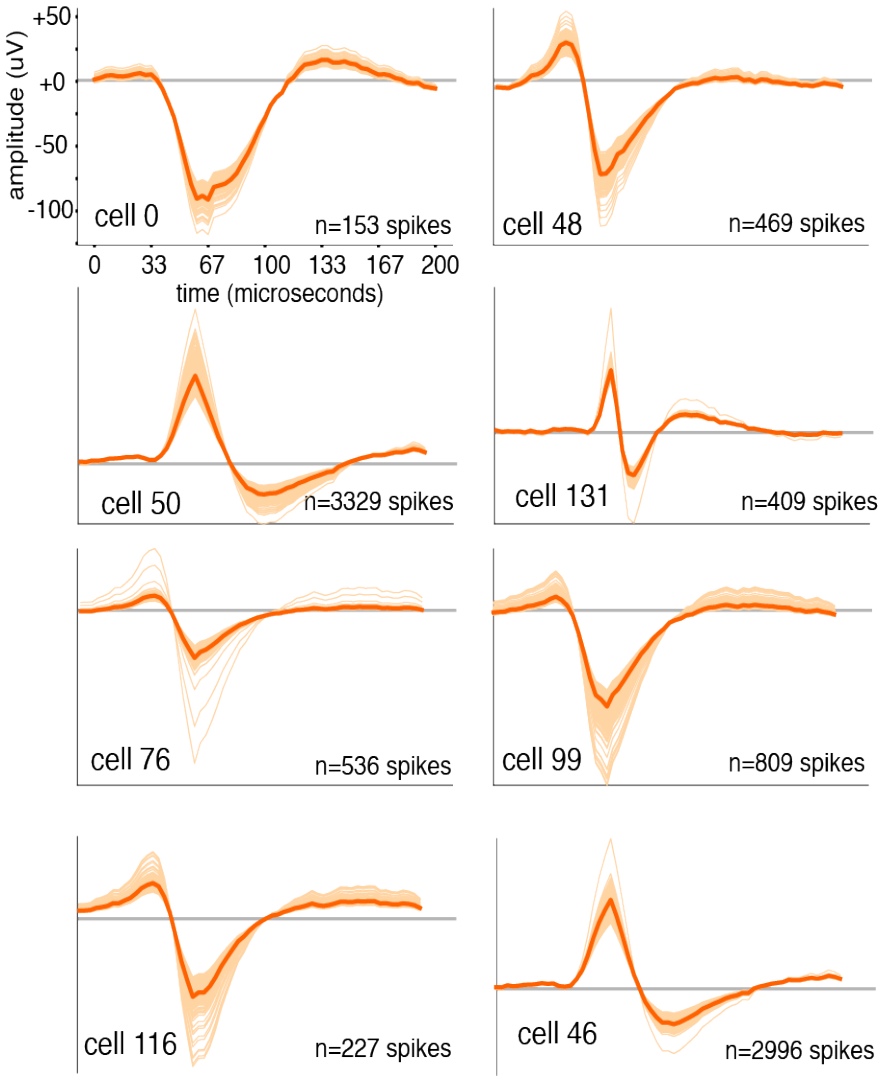
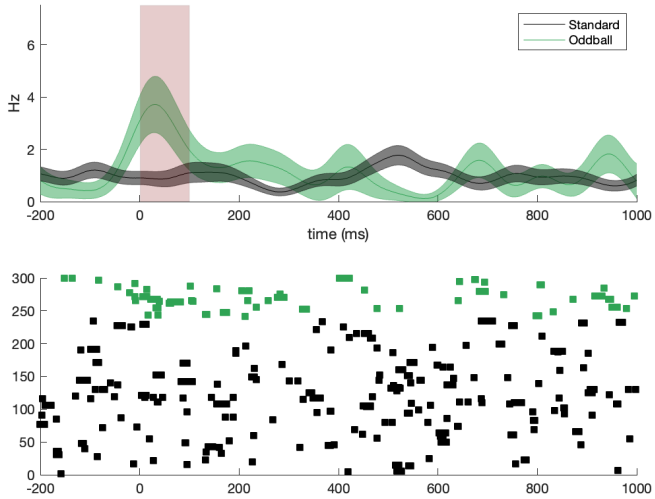


Figure 2E - The claim is that the neuron is selective for oddballs, but is this a statistically significant statement?

The neuron referred to in this reviewer question is selective for oddballs ($p=0.0078$). However, given other reviewers' questions about this panel we have replaced this example with a different one. That one is also significant (in this case, at $p<0.0001$).



Furthermore, the smoothing doesn't seem to match the raster plot. The rasters show that on oddball trials there are more spikes from 0-100 msec, but the smoothed plot shows a higher firing rate from 100-200 msec.

> **This effect resulted from our use of a causal smoothing filter, which only included past time series values to estimate instantaneous firing rate. The decision about smoothing filters is a complex one, and there is no single correct answer. Given the reviewer's concerns, we have altered our smoothing method to instead use a symmetric non-causal Gaussian filter which equally weighs past and future time points. However, we know that while this approach is more temporally accurate, it has its own limitations. In particular, it implies there is some information before the tone. We could easily use a different filter, which would not have that information, but it would be less accurate on the timing of the peak. For clarity, we have included the following text:**
>
> Note that we use a symmetric non-causal Gaussian filter, which equally weights past and future time points. The small amount of responding before the tone is an artifact of this filter.

Was there any re-referencing performed for the ERP or Gamma bands? It's typically to reference according to closest electrodes to extract local activity.

> **The reviewer raises the prospect of re-referencing ERP signals, which serves the goal of removing common noise across channels. This is fairly straightforward with conventional electrodes, but not as clear how to do**

with Neuropixel electrodes. This practice is common for human studies with orders of magnitude larger distances between them, allowing for the local field potentials to be relatively uncorrelated, leaving primarily the noise to be common across channels. The significantly higher channel density in our data means that ERP and other events are visibly well-correlated across neighboring channels, and therefore re-referencing would discard much of the physiological signal rather than only noise. In any case, the risk of common noise is low here, due to the nature of Neuropixel electrodes. We note that at least some prior studies using the same electrode arrays (e.g. Paulk et al., 2022) did not include a re-referencing step in their LFP analyses either.

Figure 2J - Figure caption mentions encoding accuracy, while axis legend mentions decoding accuracy. If using SVM, i think decoding is what is meant, but this should be checked.

The reviewer is correct, and we apologize for the error. We have now revised the text to say "decoding".

Figure 3A - one of the curves appears to be missing (I think it's oddball trials 1-20).

We apologize for this error and we have now fixed the issue. The revised figure panel is reproduced here.

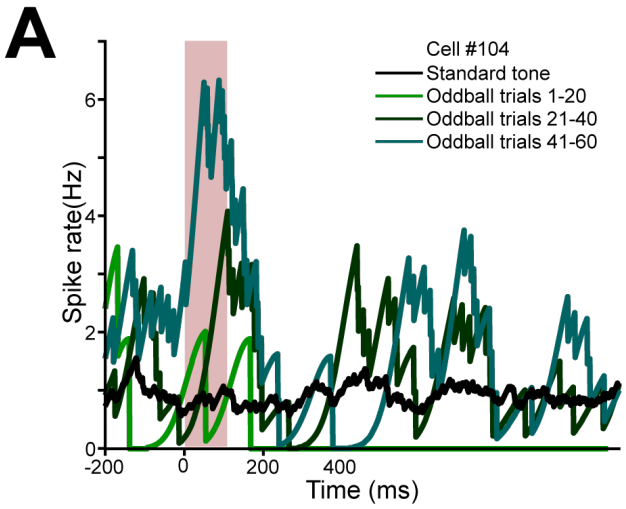


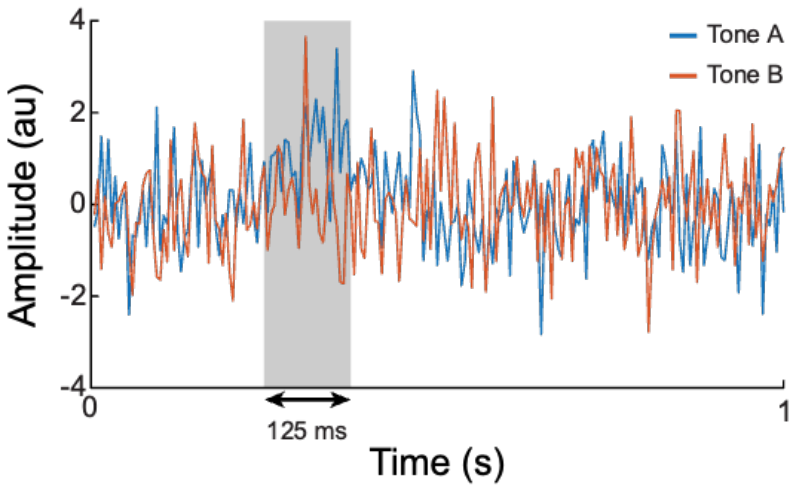
Figure 3B-E - It should be clearer that the writers mean 1st and 2nd half OF EACH BLOCK. Without that clarification, these results don't mean much because oddball and frequency would be collinear, and without any regularization the results could be noise.

This is a good point. We now make this clearer in the caption and in the revised Results where these are mentioned. Revised Results text in blue, in context, is shown here:

Splitting our task into first and second halves (for each block), we found a significant increase in oddball encoding for both patients (P5: p=0.01, P6: p<0.001, t-test, Figure 3C). Surprisingly, we also observed a concomitant decrease in frequency encoding from the first to the second half of each block, raising the possibility of compensatory mechanisms (p<0.001, t-test for both)

Figure 3H - There is no reason to use an EI net for this when they are not actively using that detail in a meaningful way. The methods should be clearer about what the input to this network is. Is it sine waves at different frequencies, and the goal is to produce fixed points in response to this entraining input? Given that most human single neuron results show encoding of variables in spike timing relative to oscillations, I'm hesitant to buy in to the analogy of this system to a rate-coded EI net. More discussion on the utility of this model beyond 'local computation' would be appreciated

The input to our RNN model consisted of two Gaussian noise channels, one of which was transiently elevated by a constant value (+1.0) during the stimulus window to indicate the presence of a particular tone (either Tone A or Tone B). An example of the input signals is shown below.



Example input signals to the RNN for Tone A and Tone B channels during a Tone A trial. A transient increase in the Tone A channel during the stimulus window (shaded area) indicates the presented tone, while both channels contain background Gaussian noise.

To clarify these details, we have now revised the manuscript to include the following lines:

To simulate the two auditory tones, the model received Gaussian white noise across two input channels, each corresponding to one of the two tones (Tone A and Tone B). On each trial, a transient +1.0 bias was added to one of the two channels during a brief stimulus window, indicating the presence of the corresponding tone.

We agree with the reviewer that spike timing plays a critical role in single-neuron coding, and our modeling approach does not aim to discount these mechanisms. Rather, our goal was to identify possible circuit-level principles underlying flexible tone discrimination and oddball detection. The following new text highlights this idea:

Rate-based RNNs provide a tractable framework for investigating how E-I interactions give raise to these mechanisms. To this end, we used an E-I RNN model to reflect the known composition of cortical circuits, which are predominantly excitatory (~80%) with a smaller proportion of inhibitory neurons (~20%). As the reviewer pointed out, to more fully leverage the EI structure of the model, we conducted systematic lesioning analyses in which we selectively lesioned each of the four recurrent connection subtypes (E->I, E->E, I->E, I->I) by setting the corresponding weights to zero. We then re-computed SVM decoding performance for both tone frequency and oddball context.

As shown in the figure below, lesioning inhibitory-to-excitatory (I->E) and inhibitory-to-inhibitory (I->I) connections led to the most pronounced drop in decoding accuracy for both tone frequency and oddball context. These findings indicate that inhibitory feedback, both directly onto excitatory neurons and within inhibitory populations, plays a critical role in shaping population-level representations. These results, then, show a potential parallel between the modeling and neural data.

We now include these additional results in the revised manuscript:

Systematic lesioning of each of the four recurrent synaptic connection types (E→E, E→I, I→E, I→I) revealed that inhibitory connections are essential for encoding both tone frequency and oddball context (see Supplementary Figure 9).

Figure S9 is reproduced here:

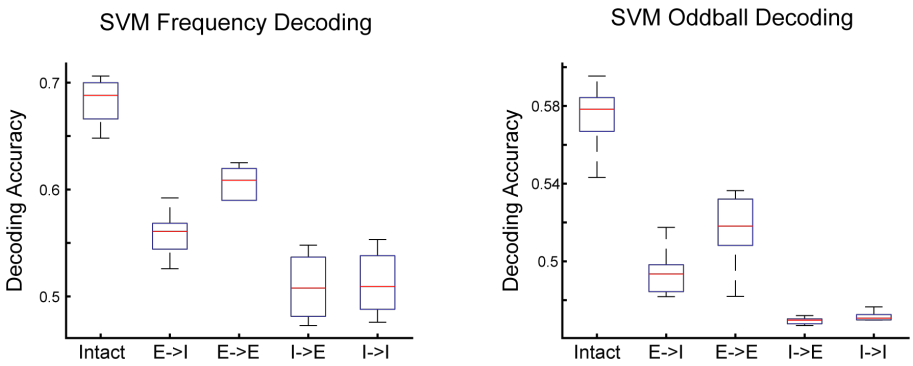


Figure S9. Inhibitory connections are necessary for encoding both tone frequency and oddball context. Each subtype of recurrent connection in the trained EI-RNN (E->I, E->E, I->E, and I->I) was lesioned by setting the corresponding weights to zero.

Figure 3I - Decoding of an RNN, but the paper only mentions a statistical test. What decoding was being done here? An SVM like with the neural data? No information in the main text or methods as far as i could see.

We thank the reviewer for pointing this out. We now clarify in the main text that SVM decoding was used to assess the model's ability to classify both tone frequency and oddball context:

Notably, despite being explicitly trained only on tone identity discrimination, the model was able to perform not only identity discrimination (tone identity, $p<0.005$, signed Wilcoxon test vs. shuffled data) but also context classification (oddball identity, $p<0.005$, signed Wilcoxon test vs. shuffled data, **Figure 3I**) via linear SVM decoding (see **Methods**).

We also edited the **Figure 3** caption:

I. SVM decoding accuracy of RNN for tone frequency (purple, left) and oddball identity (green, right).

Corresponding details have also been added to the Methods section:

To evaluate the model's ability to decode both tone identity and oddball context, we performed linear SVM decoding using population activity from the recurrent units. For each decoding analysis, we generated 100 trials for each condition. A linear SVM classifier was trained using 70% of the trials and tested on the remaining 30%, and this

procedure was repeated 100 times to estimate decoding accuracy. Separate SVM classifiers were trained for tone identity and oddball context classification.

The average correlation is low for figure 4. Is there any held out data to corroborate last claim that we can predict firing rates to a word based on responses to other words. If not, the claim should not be made.

We apologize for the lack of clarity here. First, the reviewer is correct that the average correlation is low; this is precisely what we expect with neuron-level (and neural population-level) firing rate data, which has an inherently high noise level. This is why statistical significance is important. As the reviewer points out, the probative value is increased if we use a held-out data approach. That approach reduces the risk of over-fitting. In our original manuscript, we used an approach with held-out data, for precisely the reasons the reviewer raises. However, we did not make this clear in the text. In the revised text we do so.

(Note that our encoding model approach makes use of held-out data to prevent overfitting; **see Methods**)

Figure 4F was not cited in the text in text.

We apologize. This is now cited (in the sentence following the citation for 4E).

4j - 38% is a reliable but not a strong correlation

This is a reasonable point. The revised text now reads:

Interestingly, we found a modest, but significant correlation between the number of semantic categories and the number of part of speech categories represented across neurons ($r=0.38$, $p<0.001$ Spearman's correlation)

SVM for language decoding needs to specify why chance performance is 50%. They balanced the number of oddball and standard tones, not clear for the language analysis.

As with the oddball data, we subsampled the majority class such that positive and negative data were represented with equal frequency. We have added this clarification to the Results. The new sentence is reproduced here:

(Note that chance level is 0.5 because we subsampled the majority class such that positive and negative data were represented with equal frequency).

4M - Could this same analysis be done with the best model from the oddball experiment to compare.

This is a clever idea, however it is not workable in practice. It would require training a separate RNN model specifically on a language-based task. Training the same type of biologically constrained RNN on language data would be quite challenging, and to our knowledge, hasn't been done before. Moreover, doing it would involve a series of decisions about how exactly to go about it, and the results would almost certainly be reflective of those decisions, not of the underlying tone-selective model. However, the underlying idea is sound.

Unclear if this shows a true next word prediction or highlights the caveat that nearby data in a continuous time series will be similar.

The reviewer is likely referring to an important ongoing discussion in the field relating to the analysis of neural correlates of semantic embedding and whether they really implement prediction. The key paper in this debate is this one (which we now cite at the appropriate location in the Discussion):

Schönmann, I., Szewczyk, J., de Lange, F. P., & Heilbron, M. (2025). Stimulus dependencies—rather than next-word prediction—can explain pre-onset brain encoding during natural listening. bioRxiv, 2025-03.

We agree with the reviewer (and with Schönmann et al.) that these are difficult, but important, issues, and that correlations can cause spurious effects that look like predictions. However, that debate is somewhat distinct from our interests here, and we don't think that our data have particular relevance either way to that debate.

Indeed, we suspect that a full resolution to this debate may be impossible because sufficiently contextualized representations may be isomorphic to prediction. In other words, prediction may be based entirely on processes that are encompassed into what we call contextualization, and therefore it may be impossible to ever demonstrate prediction apart from

contextualization, and, even if it were possible, it may not be the kind of contextualization-based prediction that is interesting.

But, again, these debates are somewhat distinct form the goals of our paper. For these reasons, we think it is best to make two changes to our manuscript for clarity.

First, we have altered the framing of the results in Panel 4M, so that we do not make strong claims about prediction. Instead, we say the following:

These findings demonstrate that not only is recent speech being actively tracked, but encoding of words is contextualized such that we can decode future words from these encodings; this type of contextualization is a high level cognitive function crucial to speech comprehension that depends on engagement of the language network.[33] (Note that these results do not necessarily imply active prediction above and beyond contextualization, Schönmann et al., 2024).

Second, we now include a complementary result showing a surprisal effect in our responses. Again, this surprisal result does not prove prediction (see, for example, Pickering and Gambi, 2018). However, it does support our contention that neural correlates of linguistic information incorporate information above and beyond acoustic and primary semantic processes:

Supporting the idea that these effects were driven by prediction, and not just semantic association, firing rates were modulated by surprisal, a metric that evaluates the relative probability of each word as a function of the prior words,[24] in many single units (r=0.06 +/- 0.0023, 246/375 units significant at α<0.05, Figure 4N and 4O). Similar results were observed in awake patients (r=0.0386 +/- 0.0013, 245/356 neurons significant.

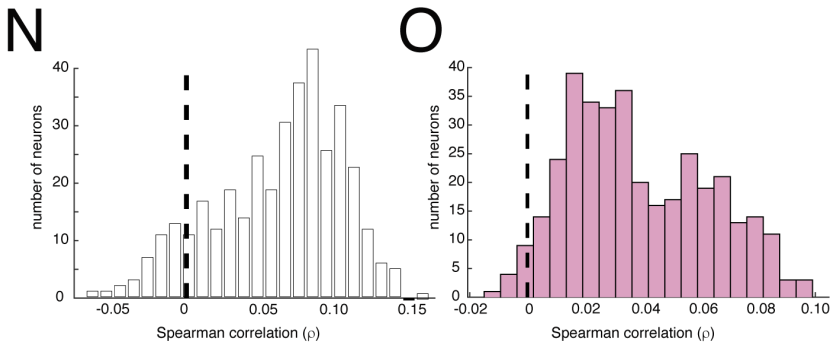


Figure 4N. Histogram of correlation coefficient (Spearman) for each neuron between surprise and firing rate. Overall neural firing rates were higher for surprising words. O. Corresponding data for awake patients showing that effects are similar for awake and anesthetized patients.

Referee #2

In their work, Katlowitz and colleagues describe a rare set of experiments in anesthetized neurosurgical patients with drug-resistant epilepsy. Neuropixels microelectrodes were implanted to brain tissue just before its resection, and sounds were presented – either a series of tones in an “oddball” paradigm, or natural language audio streams (podcasts). The authors put forward a provocative statement – that what they term ‘high-order perception’, as reflected in hippocampal activity, occurs in unconscious surgical anesthesia. Undoubtedly, if the paper would convincingly establish that high-level processing of external events during surgical anesthesia is the norm, this would be of immediate interest to wide audiences.

The authors obtained extremely rare intraoperative data of high-density neuronal recordings in humans and expanded on previous studies that recorded spontaneous cortical activity in humans with Neuropixels probes. Here, data were obtained in hippocampus, and while presenting auditory and language stimuli. In the first experiment, the extent of ‘high-order perception’ is addressed by examining the characteristics of tone responses, their selectivity for tone frequency or oddball status (standard vs. deviant), and dynamics throughout the recording session (what the author term learning). Next, the extent of ‘high-order perception’ is addressed by examining activity recorded during podcast presentation and investigation of the relation between neuronal activities and linguistic features including lexical frequency, semantic embedding, grammatical features, as well as representation of recent or upcoming words.

We thank you for your kind words and appreciate the summary.

Unfortunately, the data do not support the strong conclusions put forward and their far-reaching implications. Bold statements require strong supporting data, and this is not the case here. Results of the first experiment (tones) have various shortcomings including weak effects, over-interpretation, and novelty. Results of the second

experiment (podcast) are more intriguing, but rely on two participants, are very indirect (based on decoding rather than raw data) and - unless it is established that decodability is comparable to awake data - could potentially reflect residual hidden statistical regularities in the language stimuli (e.g. imagine that, on average, every noun in natural language tends to be followed by longer interval of silence before next word compared to verb; just as one example to illustrate the type of regularities that may exist).

All concerns are addressed in detail below. In brief, the reviewer here raises two major concerns that we address with new data that were not included in the original manuscript.

1. **First, the reviewer argues that the sample size of the two language patients (n=195 units) is too low to draw reliable conclusions. In the revised manuscript, we present data from two more patients, bringing the total number of patients to four and the total number of units recorded to 375. In short, we replicate and confirm all the claims made with the first two patients in the new sample of two more patients.**

2. **The reviewer asks us to confirm that the claims we made about language use during anesthesia are also observed, at similar levels, in awake patients. We now include a comparison group of 10 awake patients with a total of 356 units to address this concern. In short, all patterns are similar in the two groups. Of note, 3 patients were the same patients that participated in the intraoperative experiment, recorded in both awake and anesthetized conditions, so we have the advantages of within-participant comparison. Again, we find the same results, even in the same individuals. Indeed, we can confirm that all major findings were replicated in this subset of three patients who participated in both tasks.**

Finally, we want to clarify one point related to the reviewer's concern that our analyses are based on decoding. We note that three of these analyses (those found in Figure 4C, 4D, and 4M) use encoding models, rather than decoding models. (and Figure 4C is arguably the key figure in the language part of the paper). In other words, these do not simply show that neurons can discriminate between categories, but instead build a specific prediction for how each neuron will fire, and show that the prediction is better than chance. For this reason, many people feel that encoding models go beyond simply showing information is there, by proposing and validating a specific

theory for how that information is transformed into spikes. We clarify this point in the revised manuscript.

Some major issues include the following:

• The lack of awake data in the same individuals makes all comparisons and interpretations difficult.

We have now included data from a cohort of 10 awake patients listening to podcast stimuli. These results are summarized in the revised Figure 4. In short, we find qualitatively identical and quantitatively similar results in the awake and anesthetized patients.

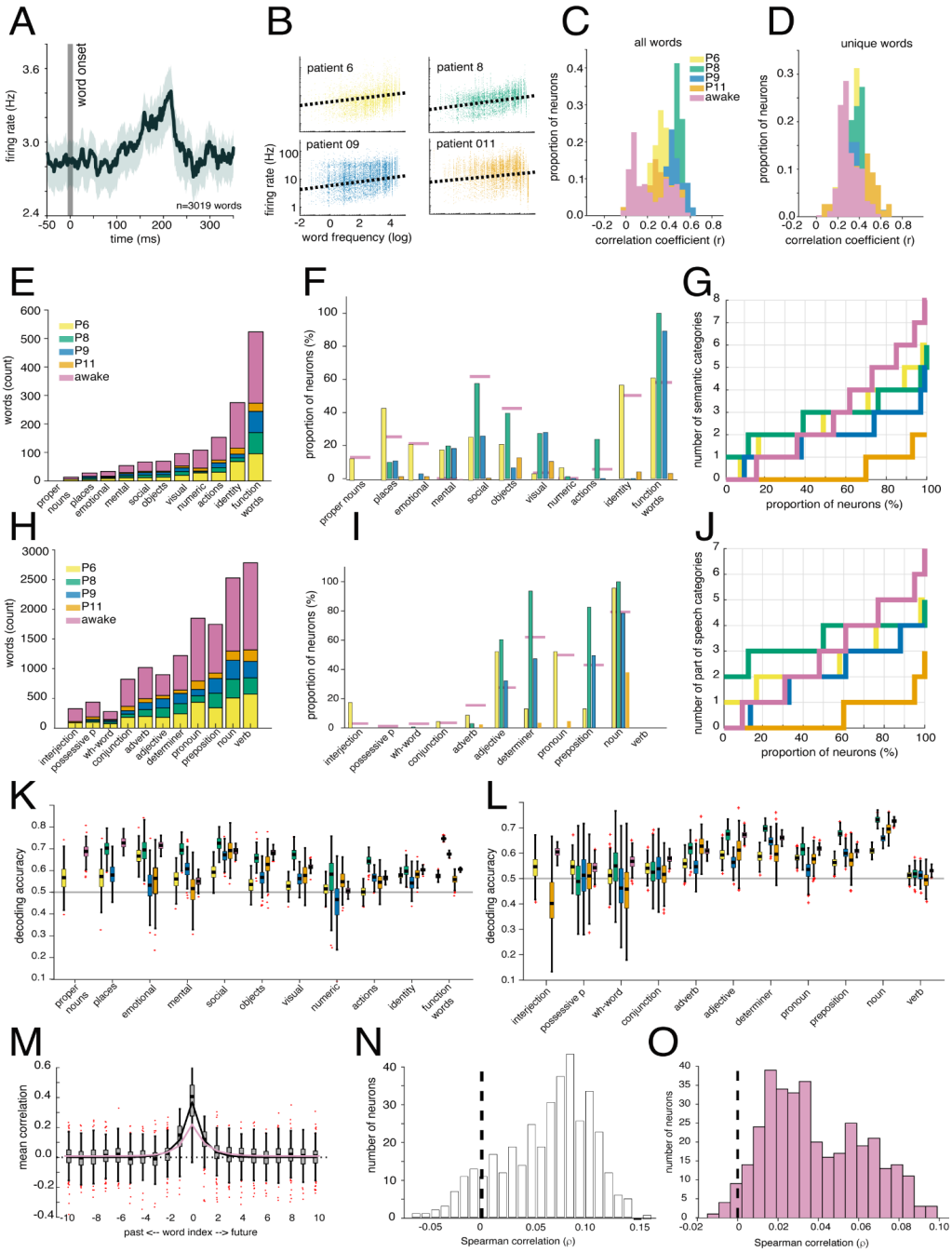


Figure 4. Neuronal responses to natural language in the anesthetized human hippocampus. **A.** Average neuronal response (spike rate, Hz) of an example unit across all words presented (n=3019 words), shown aligned to word onset (indicated with a vertical dashed line). Shaded area represents +/- SEM. In all analyses, activity more than 50 ms past word offset was removed to reduce potential contamination from adjacent words. **B.** Neuronal responses (spike rate, Hz) as a function of word frequency shown for each neuron. Each data point corresponds to an individual single unit and word, color-coded according to patient. Dashed line corresponds to a linear fit for that patient. **C-D.** Distribution of Pearson's correlation coefficient for predicted vs. actual firing rates for all words (**C**) and unique words (**D**), data shown per patient (colors correspond

to panel 4B). In all cases, correlations are greater than zero, indicating that the models are better than chance at fitting neural responses. **E.** Distribution of words within semantic categories sorted by frequency found within the podcast episodes, shown per patient. **F.** Percentage of units selective for each semantic category, compared to all other semantic categories, shown per patient. **G.** Number of categories decoded by individual units, shown per patient. Neurons generally participate in coding of multiple categories. **H-J.** Similar to E-G but for part of speech categories. **K-L.** ~~Box plots for~~ Decoding accuracy of a support vector machine (SVM) across units for semantic (**K**) and part of speech (**L**) categories, shown per patient. Dashed lines represent chance and pluses indicate outliers. **M.** Correlation coefficients for predicted vs. actual firing rates as a function of word index. Index=0: current word. Positive indices correspond to future words (right) and negative indices to past words (left), shown per patient. **N.** Histogram of correlation coefficient (Spearman) for each neuron between surprisal and firing rate. Overall neural firing rates were higher for surprising words. **O.** Corresponding data for awake patients showing that effects are similar between awake and anesthetized patients.

• Sound intensity at/around subject location was not measured, limiting the ability to infer about frequency selectivity. Even though tones had identical intensity as digital waveforms, the cables, speaker, and room inevitably attenuate some frequencies more than others; therefore, results interpreted as frequency selectivity could still easily reflect effects of different sound intensities. “Hippocampal units encoded tone frequency” can only be stated if one proves that intensity is identical at the ear…

Below, we report more information on the tone stimuli, sound delivery equipment and their frequency response characteristics (now included in the methods). Nevertheless, we acknowledge that the stimuli are not perfectly controlled as they would be in an auditory experiment (and we make this clearer in our revised text). However, this point, while important, is irrelevant to our central claims, which are concerned with oddball stimuli, not with pure tone discrimination. We have therefore reworded claims such as “Hippocampal units encoded tone frequency” to the more accurate “Hippocampal units encoded tone identity,” emphasizing the categorical difference between the stimuli rather than claiming that this encoding is due to a specific auditory characteristic.

Revised text:

Sound stimuli for the auditory oddball task consisted of high- and low-pitched tones. The low-pitched tone was 100-ms duration and 200 Hz, approximating a square wave. The high-pitched tone was 100-ms duration and 5kHz frequency, also approximating a square wave. These stimuli were constructed to have distinct perceived pitch and salient onset structure.

Sounds were delivered in stereo, using a sound delivery system that was calibrated in the testing suite (B&K type 4939-A-011 calibration mic and NEXUS 4939-A-011 conditioning amplifier). Both speakers had a relatively flat frequency response (+/-5dB) across the utilized frequency range (200-6000Hz) and no high or low frequency roll-off.

• Novelty: It is widely established that differential responses to stimuli occur during anesthesia even beyond early sensory regions, for example object-selective responses in IT cortex of monkeys; more specifically, responses to deviant tones occur during anesthesia (typically termed stimulus-specific adaptation), this also true in hippocampus e.g. https://www.ibroneuroreports.org/article/S2667-2421(23)00689-9/fulltext.

We acknowledge that tone encoding in the hippocampus under anesthesia is previously known. However, we believe that our study goes beyond established research, in showing these effects (1) at the unit level in human hippocampus, (2) linking it to a specific computational model, and (3) incorporating it into a larger story about cognitive capacities under anesthesia, including language results. In the revised text, we now try to make these findings clearer with a revised text that includes greater emphasis on novelty. Most notably, that includes a new discussion, especially the following two paragraphs:

We identified neural signatures of plasticity of sensory response and of semantic processing in the anesthetized human hippocampus. These are core cognitive functions often assumed to be absent in the anesthetized state. Our findings do not have obvious explanations based solely on low-level sensory responses. For example, the long and slow increase in oddball detection over the course of 10 minutes is unlikely to reflect adaptation or repetition suppression, which both take place on much shorter timescales. Likewise, the representation of semantic features in speech listening requires specific processing of acoustic information. We therefore show that within anesthetic-induced unconsciousness some high-level process of sensory integration is preserved, suggesting that it is consolidation that is compromised. These results provide a potential explanation for previous reports of post-anesthesia implicit recall, which would depend on sensory processing and plasticity processes despite the absence of consciousness. Broadly speaking, these results confirm ideas previously developed in animal and human studies showing preserved neural responses to sensory stimuli, including oddball stimuli, during anesthesia, and extend them in a few important ways. First, we find a gradual change in representation over the timescale of several minutes, showing that the anesthetized brain is capable of the kinds of integration over those timescales usually associated with wakeful learning. Second, we link those changes to a specific plausible biocomputational model that accounts for the effects without the types of executive control that are presumably diminished during anesthesia. Third, we show that high-level language information is available, beyond the lowest level of auditory processing.

> The present results complement our past studies, which were performed using awake patients in the epilepsy monitoring unit (EMU), on hippocampus neuron-level representation of lexical semantics[35] and semantic contexualization[36]. In those studies, we found support for the idea that the hippocampus has a largely vectorial representation of meaning, and that that representation changes in systematic ways with adjacent words. The present results suggest that these vectorial representations of meaning, and at least some contextualization, can occur in the absence of conscious awareness. That in turn raises the question of how far linguistic processing can go in the absence of awareness[37].

Furthermore, differential responses to deviant tones are observed in multiple unconscious conditions ranging from natural sleep in rodents to coma in humans (e.g. Nourski et al J. Neuroscience 2018 for intracranial human study under anesthesia). This is the basis for many studies employing the “local/global paradigm” to separate local deviance associated with simple adaption-related processes (known to be preserved in unconscious states) from true deviance detection (as the authors imply here).

> **We appreciate these suggestions and have now included citations to the new references in our revised text. In any case we now emphasize, in our revised Discussion, the novel elements of our paper, which is (1) the fact that these results are single units in humans, which are more localized and amenable to studies of mechanism; (2) that there is a gradual change over time; and (3) the links between this pattern of change and the modeling.**

• Authors report that 70% of hippocampal neurons significantly respond to pure tones during passive listening. This is very perplexing, I am not aware of anything similar in any species or state; for example, in the work of Aronov et al in rats (Tank group, Nature 2017) during passive playback without the task, only 1.7% of CA1 neurons respond;

> **We agree that this result is surprising, however, it may not be *that* surprising.**
>
> **First, there are many idiosyncratic factors that would change the measured proportion of neurons responding, including the amount of data, and the definitions used for responding.**
>
> **Second, it is well established that humans are highly specialized for auditory processing, including for language, and have many years of experience processing subtle tonal differences that have communicative meaning for humans and that are not particularly important for monkeys and rodents. Moreover, this observation of a large proportion of responsive**

units is consistent with results we have observed in language tasks in other studies (Franch et al., 2025; Katlowitz et al., 2025B).

Although the reviewer does not cite it specifically, the more direct comparison within the human hippocampus is with the concept cell literature, wherein low percentages of selective cells are typically observed in hippocampus. For that debate (and for animal models as well), there is an important confounding issue, which is response screening, which can greatly reduce estimates of relevant neurons. In our study, we did not pre-screen our neurons, which is the statistically most neutral approach, and tends to produce higher numbers of relevant units. For more on this debate, see:

Suthana, N., Ekstrom, A. D., Yassa, M. A., & Stark, C. (2021). Pattern separation in the human hippocampus: Response to Quiroga. *Trends in cognitive sciences*, *25*(6), 423.

also, if 70% of neurons respond, that seems to contradict the black trace in Fig. 2F (which seems to show that, unless the sound is deviant, the average response is not different than baseline); how can this co-exist with “most units … showed tone-evoked responses”?

The reviewer raises a subtle and interesting point. The confusion here stems from the fact that responses are equally likely to involve an increase and a decrease in firing rate; these cancel out. In contrast, the oddball responses have a strong bias towards the positive direction, resulting in the effect shown in Figure 2F. We now explain this in the revised manuscript:

Note that positive-deflecting and negative-deflecting tone-evoked responses tended to cancel each other out, so there is no visible peak in the averaged response for the standard tone.

• Main effect for oddball as seen in Fig 2J reflects marginal decodability around 0.52-0.56

The reviewer is entirely correct that these results are marginal in their effect size. In this case, the key question is whether these effects are greater than would be expected by chance. And the appropriate question is

not effect size, but statistical significance. We have added the following new text to the manuscript to address this point:

Oddball identity was decodable at above chance levels for both patients combined ($p<0.05$ for the two patients combined), albeit at lower levels, ranging from 0.52 to 0.56. These results therefore indicate that oddball signals are present, although with substantially weaker strength than pure tone.

Note, moreover, that the core claim that the *brain* (not limited to units) distinguishes oddballs is further validated by our analysis of local field potentials, found in Figure S2:

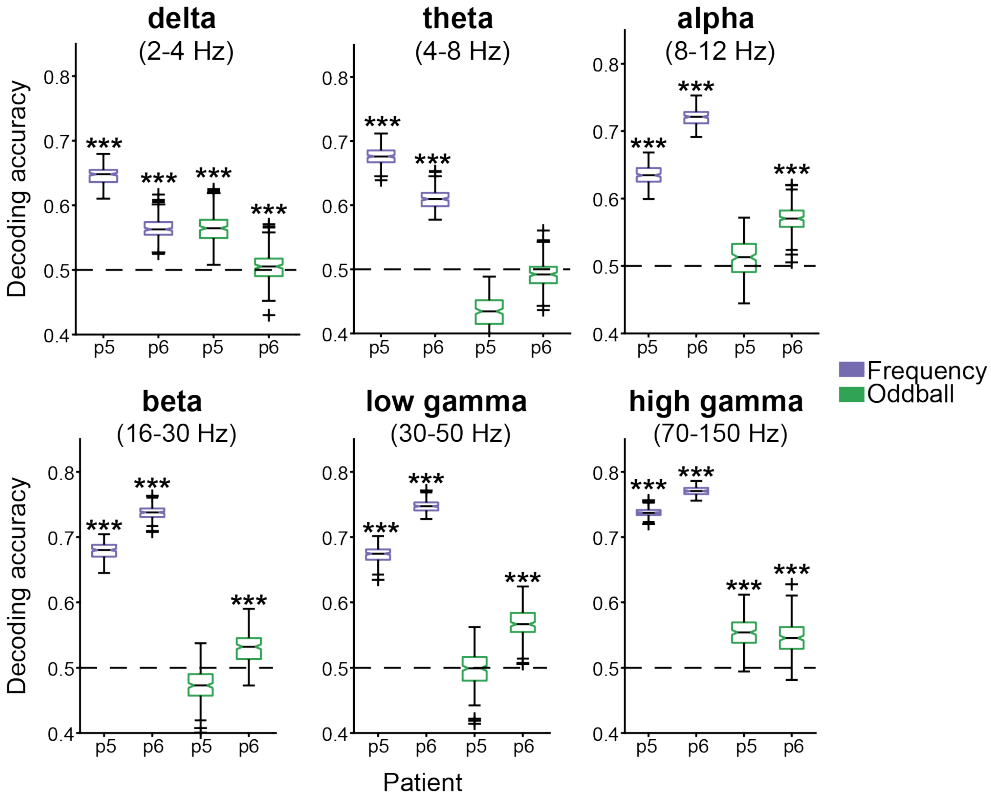


Figure S2. Box and whisker plot of decoding accuracy for tone identity (purple) and oddball identity (green), obtained using a Support Vector Machine (SVM) decoder for P5 and P6, across the population of recorded channels within each pre-defined frequency band. Pluses are outliers and asterisks denote statistical significance relative to change. *** denotes p-value <0.0001.

• “Oddball decodability“ increases throughout sessions of patients P5 and P6. This could reflect various processes. While this is “consistent with learning effects”, taking this as evidence that the brain learns during unconscious surgical anesthesia – as the authors state in the title – seems like overinterpretation.

Following this comment and the comment of another reviewer, we have decided that the most prudent path is to make clearer that the evidence for learning is suggestive, not probative.

Thus, we retitle the paper with the more neutral: Plasticity and Language in the Anesthetized Human Hippocampus

And we have edited the manuscript throughout to emphasize that the idea of learning is speculative. Most relevant, this new paragraph from the Discussion deals with this issue:

> We identified neural signatures of plasticity of sensory response and of semantic processing in the anesthetized human hippocampus. These are core cognitive functions often assumed to be absent in the anesthetized state. Our findings do not have obvious explanations based solely on low-level sensory responses. For example, the long and slow increase in oddball detection over the course of 10 minutes is unlikely to reflect adaptation or repetition suppression, which both take place on much shorter timescales. Likewise, the representation of semantic features in speech listening requires specific processing of acoustic information. We therefore show that within anesthetic-induced unconsciousness some high-level process of sensory integration is preserved, suggesting that it is consolidation that is compromised. These results provide the foundation for previous reports of post-anesthesia implicit recall, which would depend on sensory processing and plasticity processes despite the absence of consciousness. Broadly speaking, these results confirm ideas previously developed in animal and human studies showing preserved neural responses to sensory stimuli, including oddball stimuli, during anesthesia, and extend them in a few important ways. First, we find a gradual change in representation over the timescale of several minutes, showing that the anesthetized brain is capable of the kinds of integration over those timescales usually associated with wakeful learning. Second, we link those changes to a specific plausible biocomputational model that accounts for the effects without the types of executive control that are presumably diminished during anesthesia. Third, we show that high-level language information is available, beyond the lowest level of auditory processing.

• Statistics: neurons recorded in the same individual and electrode and session are not statistically independent. A proper statistical approach (increasingly used in recent years

by many similar studies) would be to employ a hierarchical mixed model taking into account both # of patients and # of neurons per patient.

We have added statistical analysis in applicable results sections to use a mixed effects model approach, following the reviewer's suggestion. This change did not affect significance or our interpretation of the results. For some analyses this modeling approach was not applicable, for example classifier-based decoding results that are aggregated across neurons by design.

We have included the following description in the methods:

Where applicable, we used mixed effects models to quantify how task conditions affect spike count and other neurophysiological variables while accounting for the hierarchical data structure of multiple subjects, neurons, and channels. For analyses of spike count, we summed spikes over equivalent durations across task conditions and fit a generalized linear mixed effects model using a log link function and modeling spike counts as Poisson-distributed. For LFP variables, a linear mixed effects model was used. All analyses used a random effects structure for neurons or channels nested within-subject.

Below is a summary of updated statistical results for applicable analyses.

Figure	Analysis	p-value
2D	Frequency response	8.691e-10
2F	Oddball response	4.0174e-37
2F	Frequency/oddball interaction	0.00048775
2I	ERP frequency response	0.224
2I	ERP oddball response	0.038
2I	ERP interaction	0.0186
2I	High gamma frequency response	0.0006184
2I	High gamma oddball response	0.024374
2I	High gamma interaction	0.00028012
3E	Euclidean dist. evolution	0.0036755
3E	Cosine dist. evolution	9.9095e-05

4B	Word frequency	0

Additional minor issues:
• Fig 1E is not clear

We apologize for the lack of clarity. We have revised the text for clarity. The revised text reads:

E. Drift maps illustrating the low drift rates, and high stability, of Neuropixel recordings in human hippocampus. X-axis: time across session. Y-axis: depth along electrode shank. Points (white marks) indicate individual spikes. Banding pattern reflects degree of motion before (left panels) and after (right panels) motion correction.

• Why is STG data of P3 not shown?

Following the suggestion of a different reviewer, we have removed the STG data from the manuscript.

• Smoothing PSTHs with kernel of 150ms is highly unorthodox and reflects weak data

In the revised manuscript, we use a more conservative and more standard smoothing window of 100 ms.

In any case, we acknowledge that this kernel is relatively long (although not at all rare in the field, including, for example, past published work of both senior authors). Indeed, long kernels like this are common in studies with relatively low trial numbers and firing rates, just like this one. In any case, the key question is one of statistical significance, which is provided in our analyses. Most importantly, note that the smoothing here is used solely for presentation; none of the statistical claims in our manuscript rely on smoothing; all statistical analyses are performed on binned unsmoothed data.

• It may be suboptimal to focus on high-gamma in anesthesia since anesthesia often yields a decrease in frequency of induced power changes and responses may preferentially occur in the low-gamma range.

We now include low gamma results alongside the high gamma results. In our case, we do not find a marked consistent difference between these two bands. In any case, we think the reviewer has a good point and it is worth including both for completeness.

Referee #3:

This paper reports neuronal recordings from human neurosurgical patients using high-density neuropixels probes in the hippocampus. Under anesthesia, they played auditory oddball sounds and podcasts, and examined neuronal firing. They found that hippocampal neurons were sensitive to rare oddballs, and that this sensitivity changed over the course of the experiment, which they refer to as learning, and which was recapitulated in an artificial RNN model. When listening to natural language stories, neurons were sensitive to lexical, semantic, and syntactic features of words, including both the current and previous/upcoming words. Notably, all of these effects were observed under anesthesia, suggesting a role for hippocampus in sensory and linguistic processing even without consciousness.

This is a very interesting paper, bolstered by a truly unique dataset with human hippocampal neurons while the patient is in an unconscious state. The analyses are generally sophisticated and rigorous, getting a lot out of data that have some inherent limitations (e.g., time constraints, electrode placement, etc). The paper raises several interesting and important questions about the relationship between sensory and linguistic processing, consciousness, and the hippocampus.

We appreciate the accurate portrayal of our findings and appreciate the kind words.

I have some questions and suggestions that I hope will improve the manuscript:

MAJOR

1) The framing of changes across trials in the oddball task being a form of “learning” does not seem well-justified to me. The structure of an oddball task like the one used here can be recognized almost immediately once the first rare target is presented, so to the extent that some sort of implicit learning of this task structure happens, it’s not clear why it should be expected to change linearly. So in that sense, I’m not sure how to interpret the change over time. While I agree that it seems like they have ruled out adaptation as an explanation for this change, other possibilities include gradual changes

in conscious state affecting firing rates (did they measure BIS continually? See for example Nourski et al 2018, J Neuro), or the possibility that they are not holding the same population of neurons over the entire experiment. If some neurons drop out/in, this could lead to both the change in response magnitude as well as the proportional change in oddball and frequency coding.

The idea that these results may not be attributable to learning is one raised by another reviewer. Following these comments, we have decided that the idea of learning may be worth mentioning as speculation, but that the findings are suggestive, not probative. However, it is certainly not the case that humans or other animals learn instantly whenever any new stimulus appears, and immediately and fully update all priors. Indeed, the gradual formation of expectations is a hallmark of learning and learning-like processes, which is why we use the term "learning curve" instead of "learning step function". Thus, while it is worth being cautious and avoiding definitive statements associated with the word "learning", it is also not unreasonable as far as speculation goes.

Following this comment and the comment of another reviewer, we have decided that the most prudent path is to make clearer that the evidence for learning is suggestive, not probative.

We retitle the paper to remove the word "learning": Plasticity and Language in the Anesthetized Human Hippocampus

And we have edited the manuscript throughout to emphasize that the idea of learning is speculative. Most relevant, this new paragraph from the Discussion deals with this issue:

We identified neural signatures of plasticity of sensory response and of semantic processing in the anesthetized human hippocampus. These are core cognitive functions often assumed to be absent in the anesthetized state. Our findings do not have obvious explanations based solely on low-level sensory responses. For example, the long and slow increase in oddball detection over the course of 10 minutes is unlikely to reflect adaptation or repetition suppression, which both take place on much shorter timescales. Likewise, the representation of semantic features in speech listening requires specific processing of acoustic information. We therefore show that within anesthetic-induced unconsciousness some high-level process of sensory integration is preserved, suggesting that it is consolidation that is compromised. These results provide a potential explanation for previous reports of post-anesthesia implicit recall, which would depend on sensory processing and plasticity processes despite the absence of consciousness.

Broadly speaking, these results confirm ideas previously developed in animal and human studies showing preserved neural responses to sensory stimuli, including oddball stimuli, during anesthesia, and extend them in a few important ways. First, we find a gradual change in representation over the timescale of several minutes, showing that the anesthetized brain is capable of the kinds of integration over those timescales usually associated with wakeful learning. Second, we link those changes to a specific plausible biocomputational model that accounts for the effects without the types of executive control that are presumably diminished during anesthesia. Third, we show that high-level language information is available, beyond the lowest level of auditory processing.

Second, while we did not measure anesthesia, these are measured and carefully controlled by the anesthesia team.

While we do not have access to these records, it is very unlikely that BIS value changes, even modestly, over the short course of the experiment, given that the patient is closely monitored, and the anesthesiologist is attempting to keep the patient at a stable anesthetic depth. We acknowledge that we did not make this clear at first. We now do so in the revised Methods:

Specific anesthesia depth was likely flat across the brief time of the experiment. First, recordings took place several hours after anesthesia induction several hours before the end of the procedure, so patients were well into the stable portion of the surgery. Second, the anesthesiologist was maintaining active monitoring and stably controlled anesthesia levels.

Finally, because we were performing single-unit recording, we were able to use standardized (and manually checked) cell-sorting techniques to ensure that our isolation was stably maintained across the brief 15-minute session. Indeed, our sorting software (Spike Interface) contains an explicitly computed metric (*presence ratio*) that purports to identify likely changes in neuron identity. In our patients, it did not detect changes in units.

For the latter point, it would also be useful to know to what extent the relative prevalence of oddball/frequency encoding is true at the single unit level vs the population, since this gives insight into the mechanism of how auditory information is being tracked.

This request is addressed in the revised manuscript. Specifically, we now include both single unit (the GLME models) and population (decoding and rotation analyses). In short, we find strong effects at both levels.

2) The inclusion of the RNN model is a strength of the paper, but one that does not seem fully realized. It seems that the main claim from the RNN is that “the divergence of representations can be due to local computations rather than inherited from other networks”, yet this is not necessarily true depending on how one views the analog of the units in the network and the brain. What is the justification for saying the RNN only models what’s happening in the hippocampus, or that the hippocampus is just one network (see above point about sub-populations of neurons)?

> **The reviewer brings up an excellent point. It was not our intention to use the RNN to localize this activity entirely to the hippocampus. In fact, the RNN relies on the formatting of the input into binary categories, a job assumed to be inherited from other areas. The goal of the RNN was to understand how this might arise. We now clarify this in the revised Discussion.**
>
> Another limitation is related to our modeling. Our modeling does not indicate that the processes we describe are unique to the hippocampus. Indeed, the RNN relies critically on the formatting of its input into binary categories, which must be inherited from other regions. Instead, the goal is to understand how oddball-dependent responses can arise from such input.

Similarly, it would be more interesting if the RNN could be used to explain a specific mechanism for how and why oddball sensitivity emerges from a model trained only on frequency discrimination. This would help allow the paper to make a claim about what the hippocampus is doing in processing statistically-structured sensory input.

> **Motivated by the reviewer’s helpful suggestion, we have added a new analysis to the paper.**
>
> **To more fully leverage the model, we conducted systematic lesioning analyses in which we selectively lesioned each of the four recurrent connection subtypes (E->I, E->E, I->E, I->I) by setting the corresponding weights to zero. We then re-computed SVM decoding performance for both tone frequency and oddball context. As shown in the figure below, lesioning inhibitory-to-excitatory (I->E) and inhibitory-to-inhibitory (I->I) connections led to the most pronounced drop in decoding accuracy for both tone frequency and oddball context. These findings indicate that inhibitory feedback, both directly onto excitatory neurons and within inhibitory populations, plays a critical role in shaping population-level representations.**

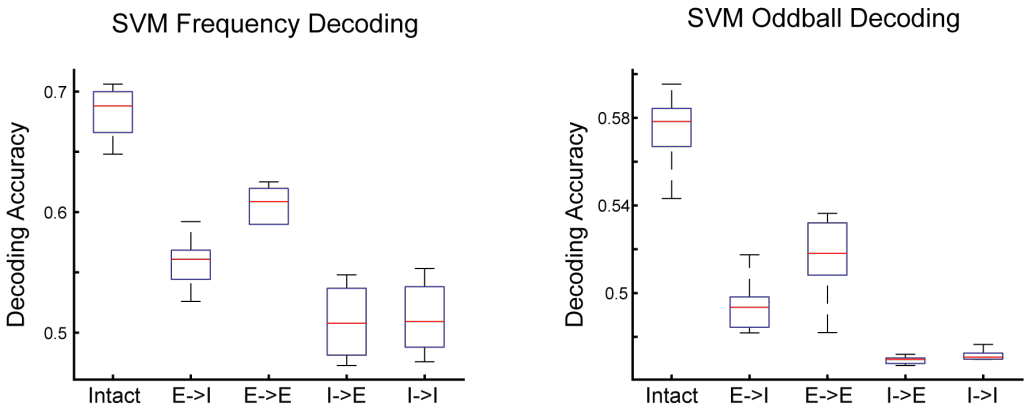


Figure S9. Inhibitory connections were important for encoding both tone frequency and oddball context. Each subtype of recurrent connection in the trained EI-RNN (E->I, E->E, I->E, and I->I) was lesioned by setting the corresponding weights to zero.

We now include these additional results in the revised manuscript:

We used an excitatory-inhibitory RNN model to reflect the known composition of cortical circuits, which are predominantly excitatory (~80%) with a smaller proportion of inhibitory neurons (~20%). Systematic lesioning of each of the four recurrent synaptic connection types (E→E, E→I, I→E, I→I) revealed that inhibitory connections are essential for encoding both tone and oddball categories (see Supplementary Figure 7). Lesioning inhibitory-to-excitatory (I->E) and inhibitory-to-inhibitory (I->I) connections led to the most pronounced drop in decoding accuracy for both tone frequency and oddball context. These findings indicate that inhibitory feedback, both directly onto excitatory neurons and within inhibitory populations, plays a critical role in shaping population-level representations. Inhibitory connections were important for encoding both tone identity and oddball context. Each subtype of recurrent connection in the trained EI-RNN (E->I, E->E, I->E, and I->I) was lesioned by setting the corresponding weights to zero.

3) The finding that hippocampal neurons can discriminate among both semantic and grammatical categories is interesting, but as written, this section of the paper feels too descriptive and it's hard for me to understand the implications of these results.

For example, is there some sort of relationship between the type of semantic and grammatical information encoded within an individual neuron or population that would tell us more about the representation? Is there anything about the structure of the semantic space in hippocampus that could be related to previous work on cortical semantic maps (e.g., Huth et al, 2016)? Even at a more basic level, if a neuron discriminates among 5 or even 10 semantic categories, what does that mean as far as its selectivity for a particular semantic feature? Do these neurons primarily just "structure the space" without actually coding for individual features, or does selectivity to specific features or semantic dimensions somehow give rise to this kind of multi-

category discriminability? What is missing is a coherent model of what these neurons are actually doing when listening to language.

We agree with the reviewer that these are fascinating questions. The focus of this paper is on the response properties of hippocampal neurons in anesthetized patients, and for that goal we are necessarily concentrating on demonstrating a response to relevant linguistic features. The reviewer's questions, while interesting, are best asked in a more conventional paradigm. Indeed, several of these questions are answered in a recent manuscript from our lab (Franch et al., 2025). That study has the advantage of focusing on awake patients, which are, for obvious reasons, more representative of natural language use.

In the revised version of the manuscript, we provide a direct comparison of the awake and anesthetized patient groups, including several new analyses that did not appear in the Franch et al. manuscript. That direct comparison draws more explicit links to the Franch paper and addresses many of the reviewer's concerns about lack of overall interpretation. First, because data from those patients were included in the Franch et al. paper, the specific big picture questions the reviewer asks for are answered. Second, it tells us the answers to the reviewer's questions about linguistic representation, including the relationship between wakefulness and anesthesia. Third, we have added two new analyses to the present manuscript, replicating key results from the Franch et al. paper. One is showing a correlation between semantic embedding (using GPT2) and neural embeddings. This new result confirms that the most important features of the semantic embedding space are represented in the anesthetized hippocampus as well. The other demonstrates a surprise effect in the data, demonstrating that neurons in the anesthetized HPC are sensitive to lexical surprisal (Figure 4N and 4O).

Finally, we want to clarify one point related to the reviewer's concern that we are just showing discrimination. We note that three of these analyses (those found in Figure 4C, 4D, and 4M) use encoding models, rather than decoding models. In other words, these do not simply show that neurons can discriminate between categories but instead build a specific prediction for how each neuron will fire, and show that the prediction is better than chance. For this reason, many people feel that encoding models go beyond simply showing information is there, by proposing and validating a specific theory for how that information is transformed into spikes.

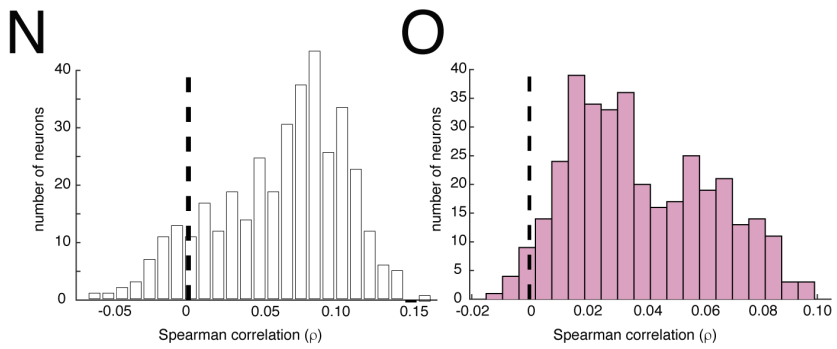


Figure 4N. Histogram of correlation coefficient (Spearman) for each neuron between surprise and firing rate. Overall neural firing rates were higher for surprising words. O. Corresponding data for awake patients showing that effects are similar for awake and anesthetized patients.

4) The comparison between neocortical (MTG) and hippocampal recordings is interesting and potentially important for helping the paper make more of a contribution to understanding the role of the hippocampus in both sensory and linguistic processing networks. However, the cortical recording here is just 2 insertions from a single patient, so it's hard to know how strongly to interpret these results. As is, the paper limits claims about comparing regions to firing rate and motion, and in my view this is a distraction from the more interesting results, so I would remove it.

> **Given the reviewer's concerns, we have adopted their suggestion and removed these results.**

At the same time, to the extent that they can leverage other data (for example previously published data from the co-authors at MGH), it could be a great opportunity to include additional datasets to make a more complete comparison. Of course, the data from other cortical regions may not be collected using the same task or anesthetic state, so they should consider whether this makes sense.

> **We agree with the reviewer, both that a direct comparison would be useful, and that it is difficult to do so with existing data. Given the focus of this study on anesthetized patients, and the uniqueness of these data, we aren't sure adding the MGH data (which are in a different region with a different stimulus) could shed much light. However, what would make sense (and is suggested by another reviewer) is a comparison with awake data. These are now included in the revised Figure 4 and accompanying text and are reproduced here.**

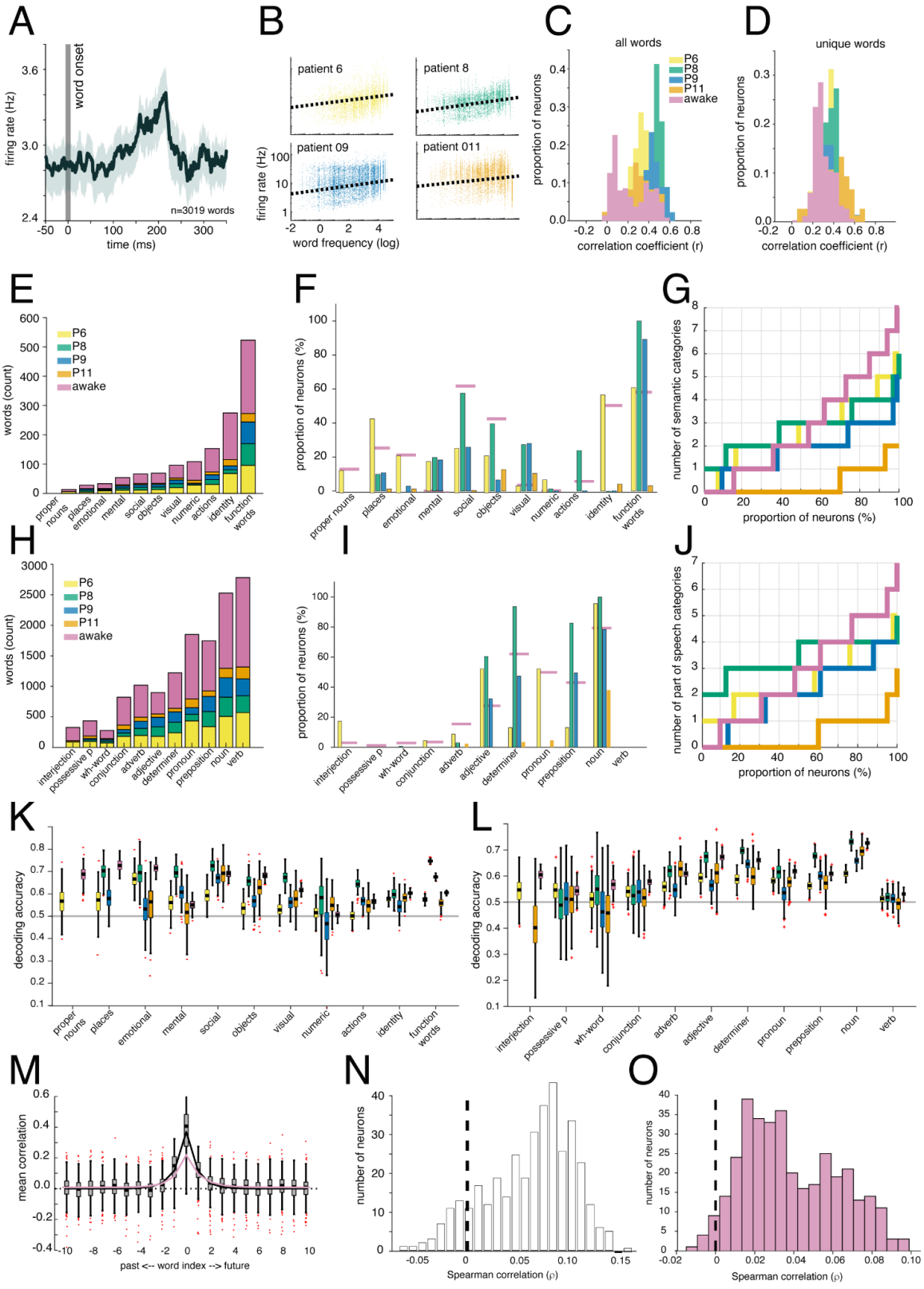

A
word onset
firing rate (Hz)
time (ms)
n=3019 words
B
patient 6
patient 8
patient 09
patient 011
firing rate (Hz)
word frequency (log)
C
all words
P6
P8
P9
P11
awake
proportion of neurons
correlation coefficient (r)
D
unique words
proportion of neurons
correlation coefficient (r)
E
words (count)
proper nouns
places
emotional
mental
social
objects
visual
numeric
actions
identity
function words
F
proportion of neurons (%)
G
number of semantic categories
proportion of neurons (%)
H
words (count)
interjection
possessive p
wh-word
conjunction
adverb
adjective
determiner
pronoun
preposition
noun
verb
I
proportion of neurons (%)
J
number of part of speech categories
proportion of neurons (%)
K
decoding accuracy
L
decoding accuracy
M
mean correlation
past <-- word index --> future
N
number of neurons
Spearman correlation (ρ)
O
number of neurons
Spearman correlation (ρ)

Figure 4. Neuronal responses to natural language in the anesthetized human hippocampus. **A.** Average neuronal response (spike rate, Hz) of an example unit across all words presented (n=3019 words), shown aligned to word onset (indicated with a vertical dashed line). Shaded area represents +/- SEM. In all analyses, activity more than 50 ms past word offset was removed to reduce potential contamination from adjacent words. **B.** Neuronal responses (spike rate, Hz) as a function of word frequency shown for each neuron. Each data point corresponds to an individual single unit and word, color-coded according to patient. Dashed line corresponds to a linear fit for that patient. **C-D.** Distribution of Pearson's correlation coefficient for predicted vs. actual firing rates for all words (**C**) and unique words (**D**), data shown per patient (colors correspond to panel 4B). In all cases, correlations are greater than zero, indicating that the models are better than chance at fitting neural responses. **E.** Distribution of words within semantic categories sorted by frequency found within the podcast episodes, shown per patient. **F.** Percentage of units selective for each semantic category, compared to all other semantic categories, shown per patient. **G.** Number of categories decoded by individual units, shown per patient. Neurons generally participate in coding of multiple categories. **H-J.** Similar to E-G but for part of speech categories. **K-L.** Decoding accuracy of a support vector machine (SVM) across units for semantic (**K**) and part of speech (**L**) categories, shown per patient. Dashed lines represent chance and pluses indicate outliers. **M.** Correlation coefficients for predicted vs. actual firing rates as a function of word index. Index=0: current word. Positive indices correspond to future words (right) and negative indices to past words (left), shown per patient. **N.** Histogram of correlation coefficient (Spearman) for each neuron between surprisal and firing rate. Overall neural firing rates were higher for surprising words. **O.** Corresponding data for awake patients showing that effects are similar between awake and anesthetized patients.

5) Regarding point 4, what is the model we should take away from these results, in particular as it relates to how the hippocampus should be incorporated into networks involved in language processing (see e.g., work by Melissa Duff)? My understanding is that even bilateral hippocampal lesions don’t lead to specific language deficits, so the question is to what extent the results here suggest something distinct from other cortical areas.

This is a fascinating question. We believe that the language community has historically undervalued the hippocampus, largely because of its inaccessibility for fMRI, EEG, and MEG studies. However, we believe it has a clear role, one that will become more well understood in future years as human single neuron recordings grow in quantity.

As the reviewer says, hippocampal lesions don't lead to classic language deficits. Of course, many studies say the same thing about Broca's Area despite its (nearly) universally accepted role in speech production. So, lack of effects in lesion studies may not be that suggestive. A small sampling of these studies:

- **Andrews, J. P., Cahn, N., Speidel, B. A., Chung, J. E., Levy, D. F., Wilson, S. M., ... & Chang, E. F. (2022). Dissociation of Broca's area from Broca's aphasia in patients undergoing neurosurgical resections. Journal of neurosurgery, 138(3), 847-857.**
- **Dronkers, N. F., Wilkins, D. P., Van Valin Jr, R. D., Redfern, B. B., & Jaeger, J. J. (2004). Lesion analysis of the brain areas involved in language comprehension. Cognition, 92(1-2), 145-177.**
- **Gajardo-Vidal, A., Lorca-Puls, D. L., Team, P., Warner, H., Pshdary, B., Crinion, J. T., ... & Price, C. J. (2021). Damage to Broca's area does not contribute to long-term speech production outcome after stroke. Brain, 144(3), 817-832.**
- **Herron, T. J., Schendel, K., Curran, B. C., Lwi, S. J., Spinelli, M. G., Ludy, C., ... & Baldo, J. V. (2024). Is Broca's area critical for speech and language? Evidence from lesion-symptom mapping in chronic aphasia. Frontiers in Language Sciences, 3, 1398616.**
- **Tsuruya, N., Kobayakawa, M., Futamura, A., Sugimoto, A., & Kawamura, M. (2013). Does a lesion in Broca's area cause apraxia?. Neurology and Clinical Neuroscience, 1(2), 55-62.**

Additionally, more nuanced studies have demonstrated key deficits deriving from hippocampal lesions in "semantic-binding" processes that are crucial for language comprehension (e.g., McKay 1998).

Having said that, in light of the wealth of literature on the hippocampus, it seems likely that whatever its contributions, the hippocampus is not uniquely involved in language processing. It participates in a large number of other non-linguistic processes. Indeed, we anticipate that the growth of hippocampal recordings in language tasks will help identify ways in which language processes intersect with other aspects of cognition. In any case, a full survey of them is beyond what we think belongs in this paper. We now include the following brief summary in the revised Discussion:

While the hippocampus is not a well-known part of the core language network, it has a clear and well established role in language[35,42-44] , including in recent single unit

recording studies[35,36,45]. The hippocampus is well-situated to perform the core elements of semantic contextualization: (1) it is closely associated with mapping functions that are related to temporal position encoding[46,47], (2) it integrates information across multiple modalities and uses that information to contextualize representations in a variety of domains[48], (3) and it is closely associated with the processes of prediction that drive contextualization[49,50]. Finally, (4) the hippocampus has a long-established role in memory, one that has many links with semantic knowledge and with semantic representations in particular[51,52]. We suspect some of the reason for the absence of the hippocampus in classical language models is negative evidence - the hippocampus can be difficult to image using fMRI and lesions isolated to the dominant hippocampus are rare. However, another factor is that the hippocampus is clearly not a site uniquely aligned to language coding processes, so it does not fit modern uniqueness-focused definitions of language areas[53]. In any case, we surmise that future work on the functional architecture of language will need to incorporate the hippocampus into its theorizing, specifically for applying rapidly changing contexts.

Some more specific points that I think should be addressed:

6) The point about word frequency being a significant predictor of firing rate is interesting, but there are two questions. First, why is the relationship between frequency and FR positive, when the majority of prior work shows a negative correlation?

This is a fascinating question, although we are not sure that the premise is correct. The number of studies looking at lexical frequency and firing rate of single neurons in hippocampus is quite low (indeed, we do not know of any). The reviewer may be talking about fMRI, MEG, or EEG studies; however, there are many discontinuities between single units and these mass action measures. Indeed, this is one reason why unit-level results are important.

In any case, in the original manuscript, we made this claim with two Neuropixel participants. We have now added two additional participants. In the new cases, we replicated the original finding of a positive correlation between firing rate and word frequency.

Moreover, the LFP analysis suggested by Reviewer 1 shows a clear (and statistically significant) <u>negative</u> correlation in all six bands individually. That demonstrates (1) a disjunction between the single unit and LFP, and (2) suggests that the earlier results the reviewer is pointing to are

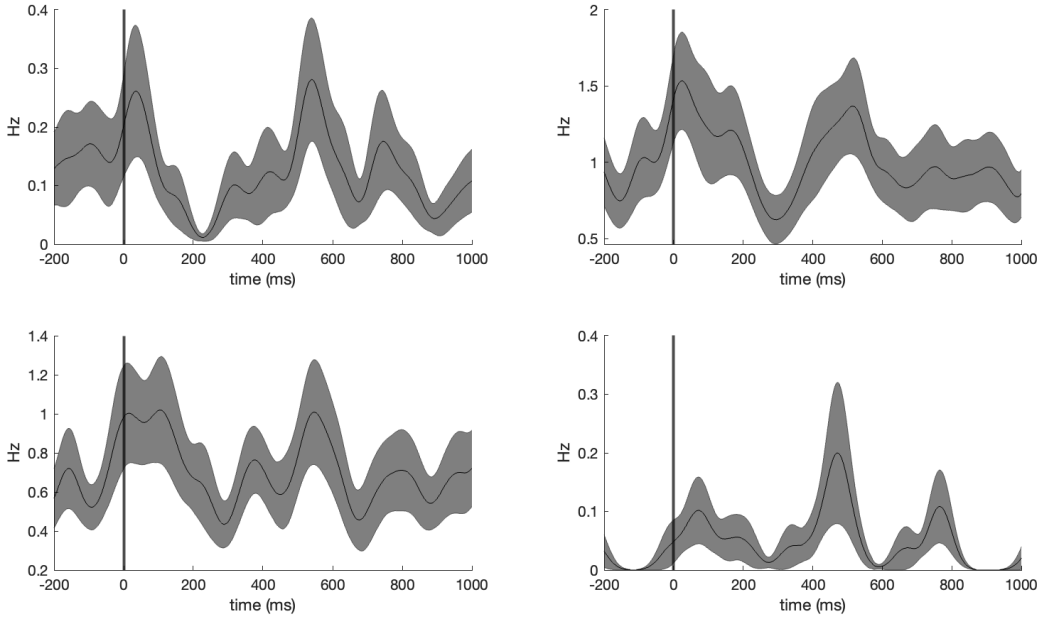
replicated in our data, and the problem is with the expectation that SUA and LFP should match, not with our experiment.

Second, it's not clear to me why they are focusing on frequency over duration – was the duration term in the linear model also significant? Both are interesting to find here, and potentially point to related but distinct features of the words in the stimulus.

The reason we focused on the frequency term rather than duration is that lexical frequency is a high-level abstract feature that is learned through years of exposure to language, thus one very much in line with the core questions we are asking, namely, whether these types of abstract features are encoded in the absence of anesthesia. Meanwhile duration is not as directly related - it is a potentially low-level feature (regardless of whether it is encoded or not), thus that result does not strongly support our hypothesis. In any case, lexical duration is very strongly encoded, in all four patients individually as well as for the group. We now report this fact in our revised Results.

7) In figure 2, it's not clear to what extent the early and later components of the response are different subpopulations of neurons. For example, in Fig 2B, there are 2 peaks, and the methods say the ITI was 1-3s (longer than the time window here), so what does that second peak reflect?

This effect was largely driven by bimodal responses at the level of individual neurons, rather than two different populations. Here are representative examples of neurons that responded bimodally:



Four example neurons demonstrating a temporally bimodal peak response to tones. t=0: onset of the tone stimulus. Y-axis: firing rate in Hz. Traces show mean +/- SEM of the smoothed firing rate curve across repeated trials.

We quantified this effect by applying a peak detection method to trial-averaged smoothed firing rate curves of each neuron. 48 cells (32%) displayed a significant response peak near both tone onset and the delayed response present near 600 ms in Fig. 2b. Only 20 cells (13.3%) responded at the delayed 600 ms peak without a response at tone onset.

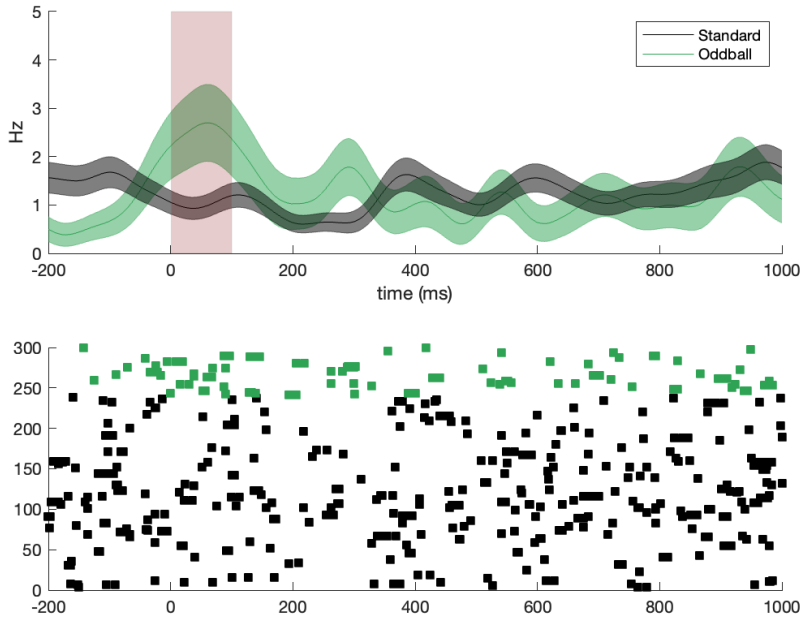
8) Also in figure 2, why does the firing rate start to rise before the onset of the tone? Is this just smoothing for the PSTH (though it's also observed in the LFP and gamma plots)?

This effect resulted from our use of a causal smoothing filter, which included past time series values to estimate instantaneous firing rate. The decision about smoothing filters is a complex one, and there is no single correct answer. We have adapted our smoothing method to instead use a symmetric non-causal Gaussian filter which equally weighs past and future time points. However, while this approach does not produce the problematic peak, it still implies there is some information before the tone. We could easily use a different filter, which would not have that

information, but it would be less accurate on the timing of the peak. For clarity, we have included the following text:

Note that we use a symmetric non-causal Gaussian filter, which equally weights past and future time points. The small amount of responding before the tone is an artifact of this filter.

Below we reproduce the example in the figure using the adjusted smoothing approach:



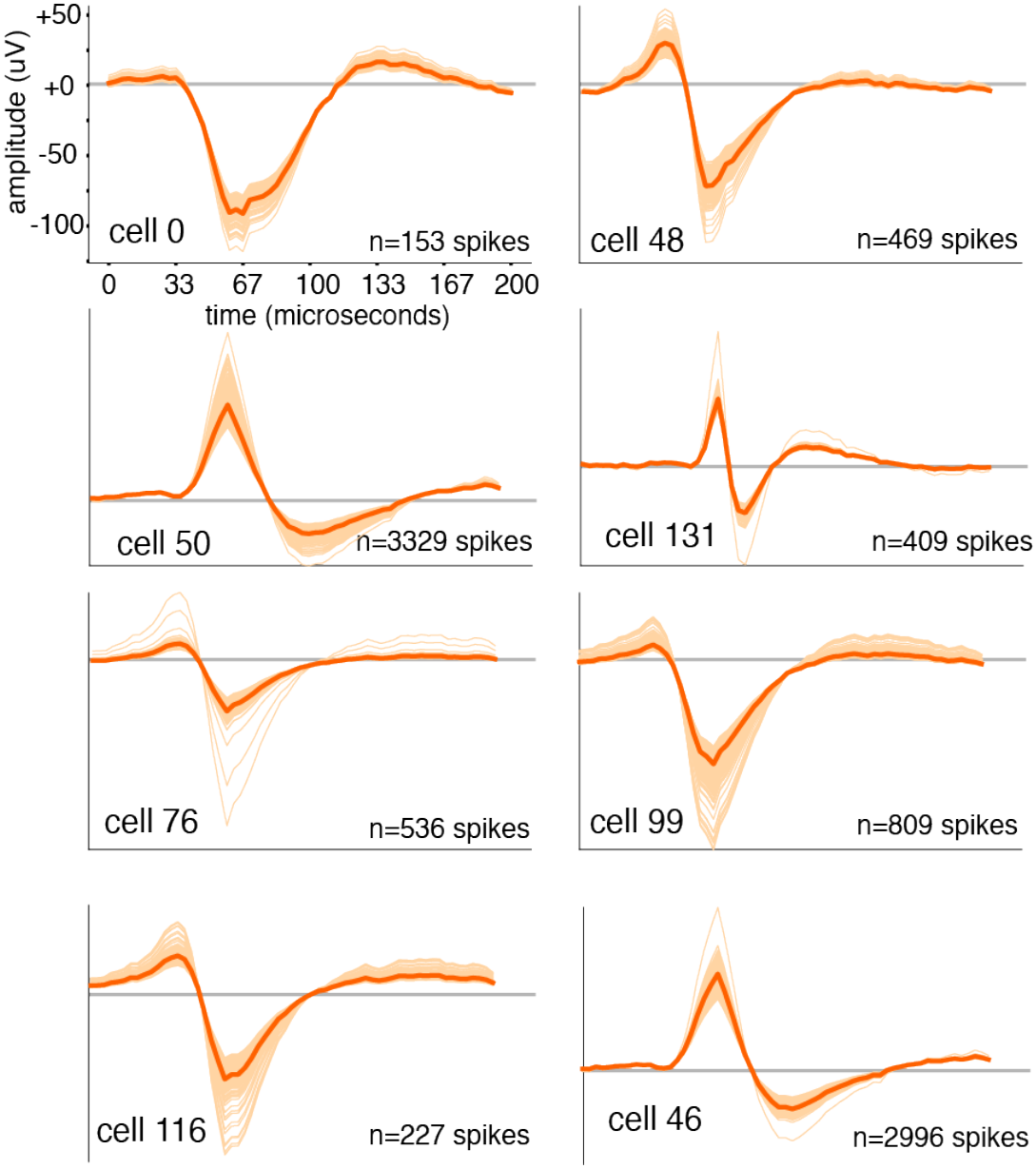
9) Fig. 1D: what are the x- and y- axes? Relative depth and the x-coordinate on the probe?

We apologize for not including this information. It is included in the revised text. Specifically, the revised caption includes the following new text:

X-axis refers to each of the four columns in the columnar arrangement of sensors. Y-axis: position along electrode shank.

Also, do they make anything of the lack of positive spiking waveforms? These have been observed pretty consistently in a lot of the recent human single unit recordings, so I wonder if this is a difference between hippocampus and cortex.

This is an interesting idea, but we do have plenty of other examples of both kinds of waveforms. In our revision, we now include a few randomly selected neurons, including ones with both positive and negative spiking:



10) Similarly, I did not see any details about putative cell types, unit depth (potentially mapped onto specific cellular layers/strata in the hippocampus), or other common characteristics.

We now include the requested information.

In the tone/oddball experiment (Figs. 2 and 3), we classified 45.3% of cells as pyramidal cells, 51.2% of cells as narrow interneurons, and 3.5% cells as wide interneurons using conventional analysis tools (CellExplorer) based on their spike waveform and autocorrelogram features. In the language analysis (**Figure 4**), we classified 65.3% of cells as pyramidal cells, 24.6% of cells as narrow interneurons, and 10% of cells as wide interneurons.

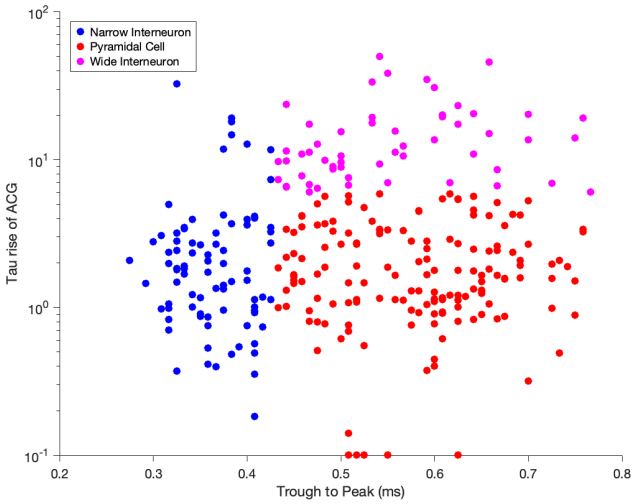


Figure S7: Categorization of cell types from unit features.

The tone response was significant in a larger proportion of interneurons than pyramidal cells (pyramidal cells: 46.7% responders; interneurons: 73.6% responders; p = 0.0142, chi-squared test).

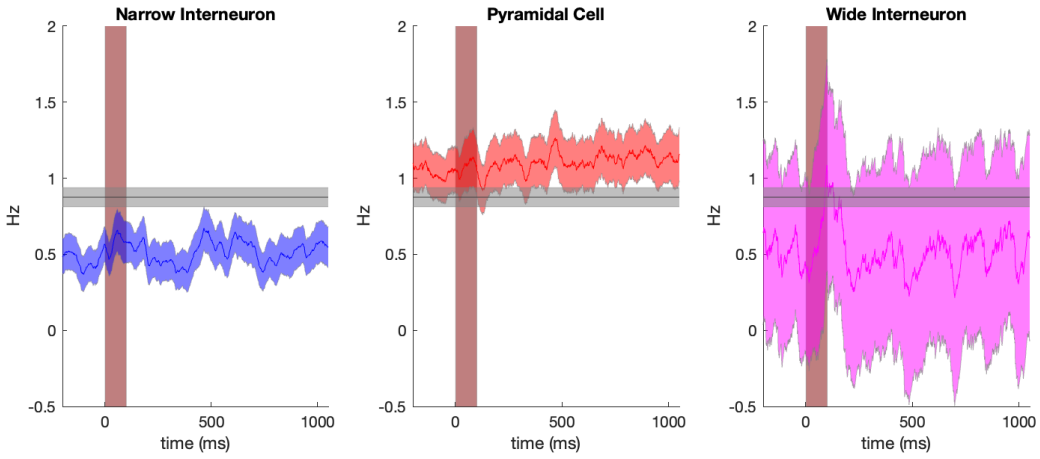
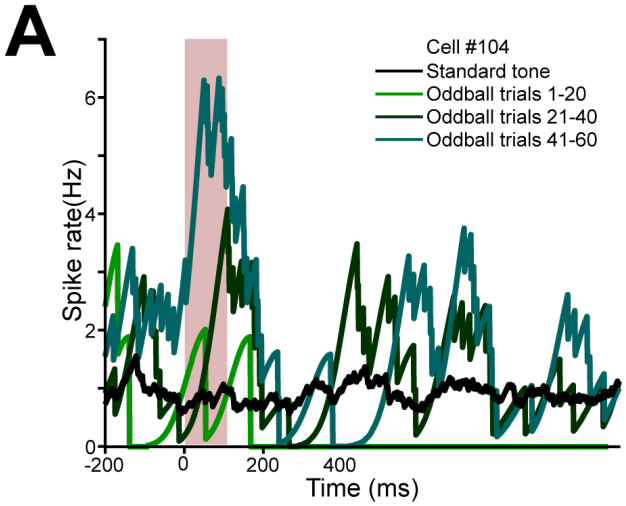


Figure S8: Tone response by cell type. Each column plots the firing rate over time (mean ± SEM across cells) in response to tones. Vertical line: duration of the tone response. Horizontal line: baseline firing rate (mean ± SEM across all cells).

As far as cellular layers/strata, this information is not given by the Neuropixel electrode. Intraoperative navigation and post operative histology have confirmed the placement of the Neuropixel within the hippocampal head, yet the exact strata and cellular composition vary significantly by depth and superior-inferior position. The exact localization, unfortunately, is still an open question in the field. We (and others) are currently developing novel anatomical reconciliation approaches to address these concerns in the future.

11) In Fig 3A, it seems like there should be 4 lines plotted, but I can only make out 3 lines.

We apologize for this error, and we have now fixed the issue. The revised figure panel is reproduced here.



12) Fig. 4M: how is this related to word surprisal? It seems like if this truly reflects prediction and/or online language processing, this effect should be modulated by how likely each word is in context.

The reviewer raises a good point. We did not include any surprisal results in the original manuscript, but it's a good thing to look at. We now include a surprisal analysis. As anticipated by the reviewer, firing of neurons is modulated by lexical surprisal. From the revised Methods:

Word surprisal values were calculated using the GPT-2 large model[12] from the Hugging Face Transformers library,[13] computing the negative log probability of each word conditioned on the preceding context

And the revised Results:

There was also a consistent relationship of the neural responses with the relative surprise of each word ($r=0.11 \pm 0.03$, 187/195 units significant at $\alpha<0.05$),

And the revised Figure 4:

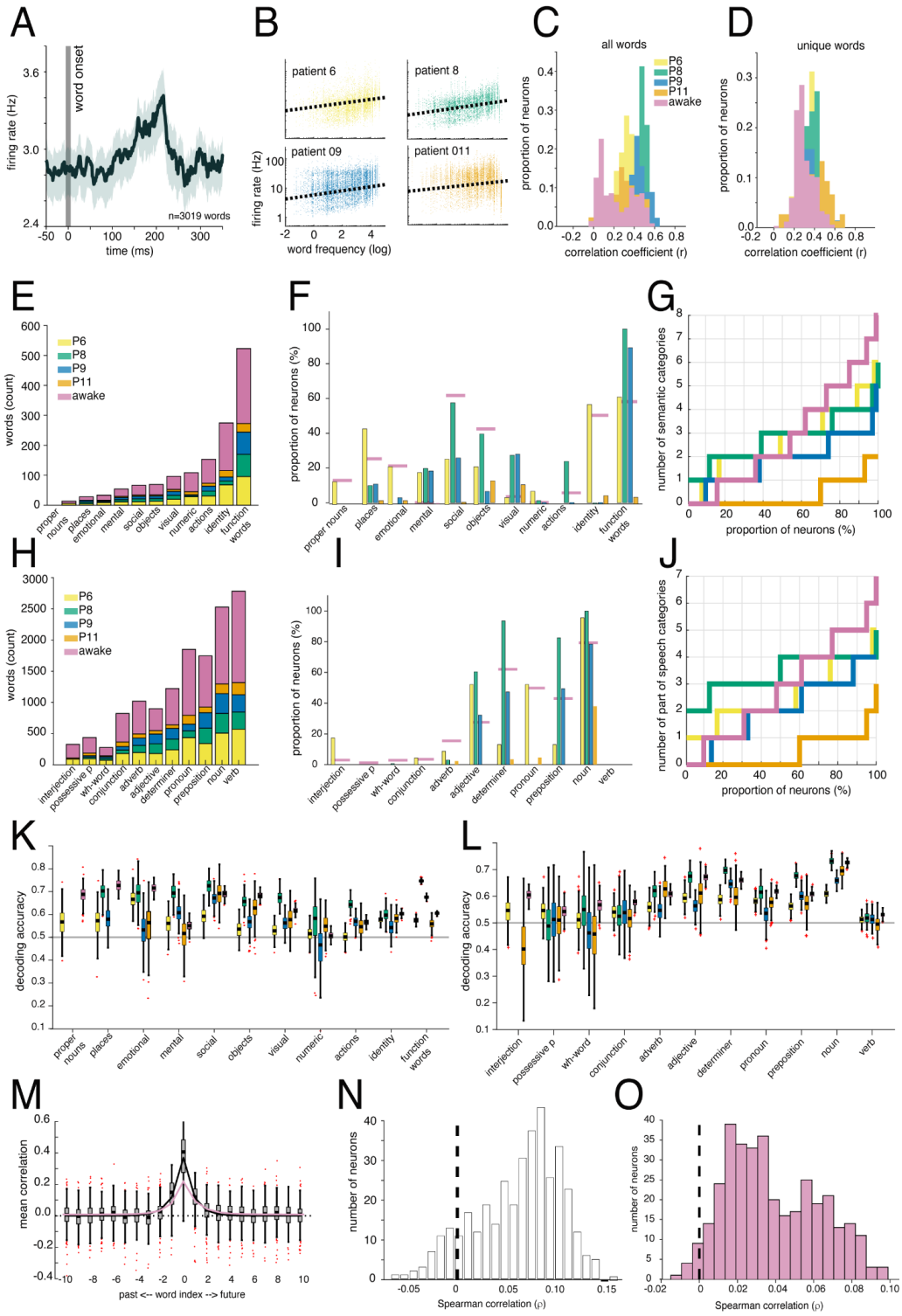


Figure 4. Neuronal responses to natural language in the anesthetized human hippocampus. **A.** Average neuronal response (spike rate, Hz) of an example unit across all words presented (n=3019 words), shown aligned to word onset (indicated with a vertical dashed line). Shaded area represents +/- SEM. In all analyses, activity more than 50 ms past word offset was removed to reduce potential contamination from

adjacent words. **B.** Neuronal responses (spike rate, Hz) as a function of word frequency shown for each neuron. Each data point corresponds to an individual single unit and word, color-coded according to patient. Dashed line corresponds to a linear fit for that patient. **C-D.** Distribution of Pearson's correlation coefficient for predicted vs. actual firing rates for all words (**C**) and unique words (**D**), data shown per patient (colors correspond to panel 4B). In all cases, correlations are greater than zero, indicating that the models are better than chance at fitting neural responses. **E.** Distribution of words within semantic categories sorted by frequency found within the podcast episodes, shown per patient. **F.** Percentage of units selective for each semantic category, compared to all other semantic categories, shown per patient. **G.** Number of categories decoded by individual units, shown per patient. Neurons generally participate in coding of multiple categories. **H-J.** Similar to E-G but for part of speech categories. **K-L.** Decoding accuracy of a support vector machine (SVM) across units for semantic (**K**) and part of speech (**L**) categories, shown per patient. Dashed lines represent chance and pluses indicate outliers. **M.** Correlation coefficients for predicted vs. actual firing rates as a function of word index. Index=0: current word. Positive indices correspond to future words (right) and negative indices to past words (left), shown per patient. **N.** Histogram of correlation coefficient (Spearman) for each neuron between surprisal and firing rate. Overall neural firing rates were higher for surprising words. **O.** Corresponding data for awake patients showing that effects are similar between awake and anesthetized patients.

13) For the RNN, in Fig 3I is chance really 50%? It seems like a model that always just predicts the standard tone would achieve 80% accuracy.

> **This is a good question, and we apologize for the lack of clarity. Although the standard tone occurs in 80% of trials, each iteration of the model was given balanced data, and the network was trained to make a binary choice between two alternatives. Therefore, the theoretical chance performance remains 50%. We now make this clearer in the revised text.**

14) The sample sizes are quite small, but given that they recorded from both left and right hippocampus, is there any evidence for laterality effects, particularly for the language task?

> **This is a very interesting question. However, with 4 patients, and one recording site each, there is no mathematical way we can make any claims about laterality effects in our data, even with our balanced classes (by chance, we had 2 dominant, 2 non-dominant). We are pursuing these analyses, however, in our larger sample of awake patients, although even there, we are waiting on additional data so that we can do valid statistics.**

15) Lines 237-241: this seems like too strong of a claim without more specific controls for both conscious state and memory consolidation. It's an interesting hypothesis, but I don't see what the evidence is that it's actually related to memory consolidation.

We appreciate the caution and now dial back the claim. Note that this sentence doesn't really have a direct analogue given the large revisions to the Discussion. The closest analogue to that passage is found here:

We identified neural signatures of plasticity of sensory response and of semantic processing in the anesthetized human hippocampus. These are core cognitive functions often assumed to be absent in the anesthetized state. Our findings do not have obvious explanations based solely on low-level sensory responses. For example, the long and slow increase in oddball detection over the course of 10 minutes is unlikely to reflect adaptation or repetition suppression, which both take place on much shorter timescales. Likewise, the representation of semantic features in speech listening requires specific processing of acoustic information. We therefore show that within anesthetic-induced unconsciousness some high-level process of sensory integration is preserved, suggesting that it is consolidation that is compromised. These results provide the foundation for previous reports of post-anesthesia implicit recall, which would depend on sensory processing and plasticity processes despite the absence of consciousness. Broadly speaking, these results confirm ideas previously developed in animal and human studies showing preserved neural responses to sensory stimuli, including oddball stimuli, during anesthesia, and extend them in a few important ways. First, we find a gradual change in representation over the timescale of several minutes, showing that the anesthetized brain is capable of the kinds of integration over those timescales usually associated with wakeful learning. Second, we link those changes to a specific plausible biocomputational model that accounts for the effects without the types of executive control that are presumably diminished during anesthesia. Third, we show that high-level language information is available, beyond the lowest level of auditory processing.

16) Fig 1C: How is the probe included in the micro CT reconstruction if it was retracted from the tissue prior to resection? My understanding is that the site is identified as part of the resection margin, the recording is done, and then the probe is removed so that the resection can take place. Is this not the case here?

We apologize for the lack of clarity. When technically achievable, we kept the probe in the tissue during the resection process. We now make this clearer in the Revised Results.

Figure 1C shows the probe in the tissue, imaged in P8, wherein the probe was resected in situ with the surrounding tissue prior to the full hippocampal resection.

17) Lines 670-672: was PCA applied to the word vectors or neural data?

We apologize for the lack of clarity. We now make clear that PCA was applied to the semantic embeddings (the word vectors) not the neural data. Specifically, the corresponding text in the revision is:

To prevent overfitting, Principal Component Analysis (PCA) was used to reduce the dimensionality of the semantic embedding vectors to account for 30% of the variance prior to modeling.

Referee #1

This is an outstanding response to our concerns. Given the novelty of this work there are a few additional issues to address:

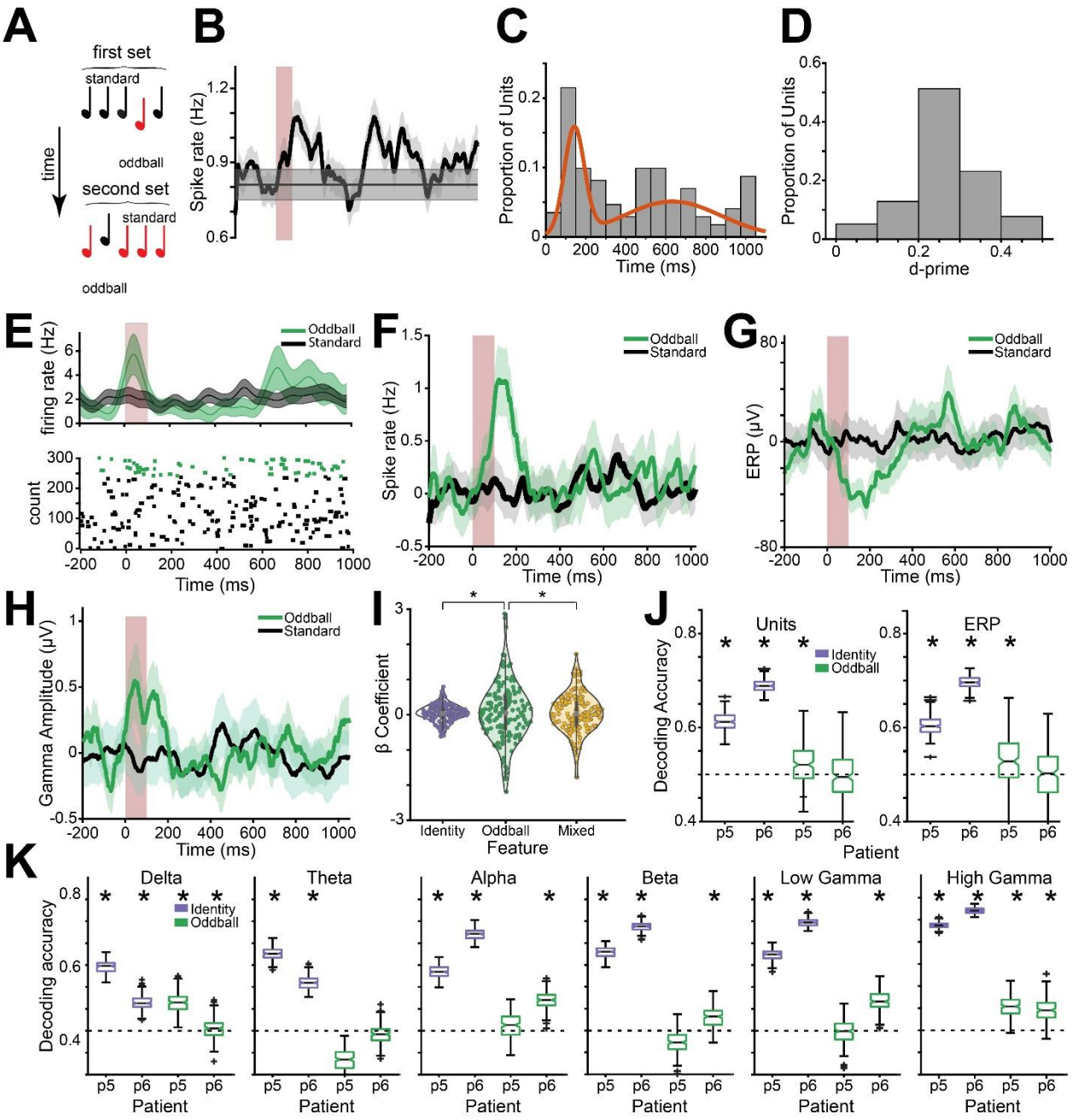
1. The Neuropixel data is exciting as noted in the initial review. But the main finding of the paper is on the role of the anesthetized hippocampus in neural processing. As such the findings that various LFP bands also provide compelling evidence regarding the hippocampus and non-conscious processing should also be clearly noted. For instance, the abstract might say:

Using high density Neuropixel microelectrodes and LFPs…. And hippocampal neurons and oscillations… provided complimentary evidence for …..

> *The reviewer's suggestion is a good one. We have revised the abstract per the reviewer's suggestions. The revised version is reproduced here, with modified text in blue ink.*
>
> **Consciousness is a fundamental component of cognition,[1] but the degree to which higher-order pattern recognition relies on it remains disputed.[2,3] Here we demonstrate the persistence of oddball discrimination, semantic processing, and online prediction in individuals under general anesthesia-induced loss of consciousness.[4,5] Using high-density Neuropixels microelectrodes[6] to record both single unit and local field potential neural activity in the human hippocampus while playing a series of tones to anesthetized patients, we found that hippocampal neurons and oscillations retained detection of oddball tones. This effect size grew over the course of the experiment (~10 minutes), demonstrating representational plasticity. A biologically plausible recurrent neural network model showed that learning and oddball representation are an emergent property of flexible tone discrimination. Moreover, when we played language stimuli, single units and local field potentials carried information about the semantic and grammatical features of natural speech, even predicting semantic information about upcoming words. Together these results indicate that in the hippocampus, which is anatomically and functionally distant from primary sensory cortices,[7] complex processing of sensory stimuli occurs even in the unconscious state.**

2. The brunt of the LFP data is in the supplement even though it supports the main claims of the paper. In this regard we suggest that Fig S3 be included in the main text.

> *In line with the reviewer's suggestion, we now have rather added Figure S2 into the main text at the bottom of Figure 2 (reproduced below), as it more fundamentally represents the presence of a tone and oddball within the hippocampus and was highlighted by Reviewer #3.*



K. SVM decoding accuracy for tone identity (purple) and oddball identity (green) for P5 and P6, across the population of recorded channels within each pre-defined frequency band. Pluses are outliers and asterisks denote statistical significance relative to chance. * denotes p-value <0.01.

It seems the 3 patients had recordings in the oddball task but only P5 and P6 LFP data are in the supplemental data. Please include P4.

The final version of the oddball task was only completed in two patients. Patient 4 performed a similar version without proper counterbalancing of tone and oddball trials, so we could not differentiate between tone selectivity and oddball selectivity. Their results are only used for analyses of general tone response characteristics (Fig. 2A-D). We apologize that we did not make this clearer in the text. The specific differences in task design for p4 are further clarified in the methods and the main text.

Main Text, Line 111
For two patients (P5 and P6), we balanced tone identity and oddball status (Figure 2A, n=150 units).

Methods, Line 789
For P5 and P6 we counterbalanced the tones. A sequence of auditory stimuli (pure tones; f1=200Hz, f2=5kHz) were presented with 80-20 probability distribution, while switching the tone frequency designated as the auditory oddball stimulus halfway through (first half, n=150 trials, f2 is auditory oddball; second half, n=150 trials, f1 is auditory oddball). We interleaved a washout period (30 trials) using the same auditory stimuli but presented at 50-50 probability distribution in between the two counterbalanced tasks.

3. Decoding accuracy is consistent across bands (Figure S5), suggesting that an aperiodic slope change might be carrying the signal rather than a specific oscillatory band. Have the authors examined this possibility?

We have performed an additional investigation of aperiodic slope for key oddball and language analyses and added as a separate text in the supplement and main results (text reproduced below). In short, aperiodic slope largely mirrors the LFP findings, for both tone and language, but does not reliably detect oddballs. (As the reviewer's comment indicates, the positive result is not particularly surprising; the null result is difficult to interpret, possibly because we are noise-limited). The following new text appears in the revised supplement (and, given other reviewer comments, is summarized alongside LFP analyses in the results):

Tones
Main Text, Line 219
Further analyses of the information encoded in LFP, including band-limited analysis and aperiodic slope can be found in Supplementary Figures S1-S5.

Supplement, Page 6

Given that multiple LFP bands were sufficient to decode tone and oddball stimuli, we further investigated whether similar results could be obtained using a single aperiodic slope feature. For each trial and channel, we computed the aperiodic slope over the same time interval as other LFP features. We computed the power spectral density between 2 Hz and 40 Hz, performed a log-log transformation, applied a median filter to mitigate the influence of narrow spectral peaks, and computed the linear slope fit (Voytek et al., 2015). Using generalized linear models, aperiodic slope values were modeled as a function of tone identity, oddball status, and their interaction. We found that 12.3% of channels demonstrated significant tone identity encoding, 6.6% of channels demonstrated significant oddball encoding, and 6.1% of channels demonstrated an interaction ($p < 0.05$ for each channel). Leveraging a 10-fold cross-validated SVM, aperiodic slope reliably predicted tone identity (p5: accuracy = 0.628, $p<0.0001$; p6: accuracy = 0.677, $p<0.0001$). Aperiodic slope was insufficient for decoding oddball identity ($p>0.05$ in both patients).

Language
Main Text, Line 406
Aperiodic slope greatly outperformed individual spectral features in predicting semantic embeddings (mean r = 0.32 +/- 0.17, rank sum test, $p<0.001$ for all bands) and robustly encoded multiple semantic and POS categories (62.5% and 25.0% encoded at least 8 categories, respectively).

4. The authors state in the rebuttal that:
In summary, we found that LFP features were sufficient to decode stimuli in both tone and language tasks, though with a generally lower performance, particularly in language, compared to that of units.

But in the text of the manuscript they only say: Analyses of the LFP corresponding to the ones presented below are found in Supplementary figures S10-S14. Please discuss the LFP language results in the main text.

The reviewer's suggestion is a good one. We have now added the requested information to the main text of the manuscript. That is reproduced here:

Main Text, Line 398
Further analyses of the LFP are found in Supplementary Figures S10-S13. In short, we found that semantics were encoded in all six bands (Figure S10), although less faithfully than in the single unit data (Figure 4C). We also found encoding of semantic category in all bands (Figure S11), and, again, less strongly than with the single units (Figure 4F-G). Category discrimination was greatest in the beta band,

followed by alpha, theta, and then low gamma. Likewise, we found significant encoding of part of speech in all frequency bands (Figure S12), although less strongly than semantic category (Figure 4I-J). Notably, however, we found that an SVM approach was able to classify both part of speech and semantic category (Figure S13) about as well as the single units in all six bands (Figure 4K-L). Aperiodic slope greatly outperformed individual spectral features in predicting semantic embeddings (mean r = 0.32 +/- 0.17, rank sum test, p<0.001 for all bands) and robustly encoded multiple semantic and POS categories (62.5% and 25.0% encoded at least 8 categories, respectively). Theta-gamma phase-amplitude coupling also outperformed spectral features in predicting semantic embeddings (mean r = 0.17 +/- 0.13, rank sum test, p<0.001 for all bands) and similarly encoded multiple semantic and POS categories (74.5% and 73.0% encoded at least 8 categories, respectively). However, spike-frequency coupling in the theta band did not significantly encode semantic embeddings (mean r=0.002 +/- 0.004) or categories (7% and 2% of units encoded one semantic and POS category, respectively; none encoded 2 or more). Together, these results attest to the robust recoverability of linguistic information in the anesthetized brain.

5. The authors were asked to evaluate SUA and high gamma phase amplitude coupling and responded they preferred not to since it was a thorny topic and subject to another paper in progress. PAC is robust across species and given this paper is likely to be a classic in hippocampal function it would be nice if the authors reconsidered their position.

We have added analyses of these two topics in the supplement and main text, reproduced below. In short, phase-amplitude coupling demonstrates predictive power in both oddball and language, and spike-frequency coupling shows the weakest effects.

Supplement, Page 7

Phase-amplitude coupling in the oddball task:

After observing successful prediction using both spike and LFP features, we further examined whether stimuli are represented in the temporal relationship between these features in the form of phase-amplitude coupling and spike-frequency coupling. For each trial and channel, we computed the modulation index (Tort *et al.* 2010) to quantify phase-amplitude coupling of gamma (30-80 Hz) to theta (4-8 Hz) power. Using generalized linear models, modulation index values were modeled as a function of tone identity, oddball status, and their interaction. Phase-amplitude coupling did not distinguish either tone or oddball stimuli: 2.25% of channels demonstrated significant tone encoding, 2.51% of channels demonstrated significant oddball encoding, and 2.78% of channels demonstrated an interaction ($p < 0.05$). Leveraging SVM decoding, phase-amplitude coupling could weakly predict both

tone identity and oddball status in p6 (p6: tone accuracy = 0.505, p < 0.001; oddball accuracy = 0.527, p < 0.005), but not p5 (p > 0.05 for both tone and oddball).

Spike-frequency coupling in the oddball task:
For each cell and trial, we similarly computed the spike-triggered theta power index (STPI) to quantify spike-frequency coupling between each cell's spike timing and instantaneous theta power. For each trial we computed instantaneous theta power from the LFP channel where spike amplitude for the given cell was measured at highest amplitude, sampled theta power values at the time of each spike, and computed the z-score of averaged theta power sampled at spike times relative to the total distribution of theta throughout the trial. Using generalized linear models, STPI values were modeled as a function of tone identity, oddball status, and their interaction. 20% of cells demonstrated significant tone encoding, 14.7% of cells demonstrated significant oddball encoding, and 12.7% of cells demonstrated an interaction (p < 0.05). Using SVM decoding, spike-frequency coupling predicted tone identity in both patients (p5: tone accuracy = 0.589, p < 0.0001; p6: tone accuracy = 0.555, p < 0.0001), but not oddball status (p > 0.05 for both patients).

Main text:
Theta-gamma phase-amplitude coupling also outperformed spectral features in predicting semantic embeddings (mean r = 0.17 +/- 0.13, rank sum test, p<0.001 for all bands) and similarly encoded multiple semantic and POS categories (74.5% and 73.0% encoded at least 8 categories, respectively). However, spike-frequency coupling in the theta band did not significantly encode semantic embeddings (mean r=0.002 +/8- 0.004) or categories (7% and 2% of units encoded one semantic and POS category, respectively, none encoded 2 or more).

6. In the manuscript, the authors state: "we expect our units to be drawn from CA1 and dentate gyrus." Not sure how this expectation arose since in the rebuttal they state:
"As far as cellular layers/strata, this information is not given by the Neuropixel electrode. Intraoperative navigation and post operative histology have confirmed the placement of the Neuropixel within the hippocampal head, yet the exact strata and cellular composition vary significantly by depth and superior-inferior position. The exact currently developing novel anatomical reconciliation approaches to address these concerns in the future."
The strong expectation needs to be altered to maybe the hippocampus proper since CA3 might also be a source. So, maybe just say hippocampal head or dorsal/anterior hippocampus.

In the interests of caution, we will go with the reviewer's suggestion and use the general term. This statement has been removed, and the text specifies that recordings were performed in the hippocampal 'anterior body'.

Main Text, Line 56

Hippocampal recordings were conducted in the anterior body after resection of the lateral temporal cortex and prior to resection of the mesial temporal structures such as parahippocampal gyrus and amygdala (Figure 1A-G).

7. There is a small typo in the Figure S12 caption (LEP instead of LFP)

Fixed, thank you

Referee #3

The authors have made meaningful improvements to the manuscript, including adding two additional patients for the surgical operating room paradigm and incorporating microwire EMU data from ten patients. The inclusion of LFP analyses—for example, high-gamma oddball decodability in Fig. S2—also strengthens the work, as does the use of mixed effect model approach. There are still a few remaining issues that warrant attention.

1. My main concern remains that central claims rely on effects that are statistically robust but relatively subtle in magnitude and not readily appreciable in the raw data—particularly at the single-trial, single-unit level. While statistical significance is key, effect size and response magnitudes also carry interpretational weight. Along this line, the reported proportion of responsive hippocampal neurons (~70%) raises the possibility that the threshold may be too permissive. The use of a $p<0.05$ sign-rank test across many hundreds of neurons, without adjusting for multiple comparisons, may contribute to this. Even in the EMU microwire dataset (e.g., Franch et al., Fig. 1C), responses appear modest, and the examples shown—presumably among the clearer ones—still show only moderate increases in firing. It would be very helpful if the authors could report whether the key findings remain stable when applying more conservative thresholds to define responsive neurons (e.g., $p<0.01$ or $p<0.001$).

The reviewer raises concern about the interpretation of the data given statistical significance and the (apparently) modest effect sizes.

First, we want to address the question of correction for multiple comparisons. Given the statistical claims we are making, multiple comparison correction would not be appropriate. Multiple comparison correction is used when there are multiple ways to get the same result, but if there is only a single way, then any multiple comparison correction is overly conservative, and will cause Type II errors. This logic is explained in, for example, this highly cited paper: Rubin, M. (2021). When to adjust alpha during multiple testing: A consideration of disjunction, conjunction, and individual testing. Synthese, 199(3), 10969-11000.

Having said that, the reviewer's suggested solution can provide us information about the overall distribution of effect sizes. We have now done so. All core analyses hold up with the two criteria the reviewer suggests. From the revised Results:

Main Text, Line 134
Tone detection effects remain robust at lower significance levels, with 55.2% of cells (p<0.01) and 28.5% of cells (p<0.001) having a significant response to any tone. Tone-selective responses remain robust at p<0.01 (12.2%) and p<0.001 (4.1%). Modeled oddball detection effects remain robust at p<0.01 (14.7%) and p<0.001 (9.3%), and tone/oddball interaction effects remained robust at p<0.01 (16.7%) and p<0.001 (10.0%).

Main Text, Line 357
Word frequency encoding effects remain robust at p<0.01 (98.1%) and at p<0.001 (97.1%). For word frequency encoding in awake patients, the effects are robust at p<0.01 (86.8%) and at p<0.001 (79.8%). Semantic encoding proportions are also robust using p<0.01 (77.9% and 68.5%) and p<0.001 in both groups (awake: 62.1% and 52.0%, respectively). Finally, for part of speech encoding, we observe a similar trend: p<0.01 (71.5% and 62.1%) and p<0.001 (74.4% and 62.6%).

Line 396
Surprisal effects are also robust at p<0.01 and p<0.001 in anesthetized (56.8% and 40.3%, respectively) and in awake patients (56.2% and 43.0%, respectively).

Finally, we want to emphasize that, especially for the language processing, we don't really know what a modest versus large effect size really is - especially given that these results are likely to reflect distributed processing. Additionally, our language stimuli are embedded in continuous, context-rich narrative speech and occur only once. As such, a comparison to more classic literature by averaging over repetitions (to ideally improve our SNR) would inappropriately collapse distinct contextual states. Consequently, modest explained variance is expected. Indeed, our effect sizes fall within the range reported in other naturalistic language studies linking neural activity to semantic embeddings (Huth & Gallant 2016; Goldstein 2024; Leonard & Gwilliams 2023; Karkowski et al. 2025).

2. The distinction between "tone frequency" and "tone identity" remains somewhat unclear. Although the authors changed terminology in the main text to "tone identity," references to "frequency" persist in several figures (e.g., Fig. S4, S5).

We have fixed the error in these figures (and double-checked other figures), replacing the label 'frequency' with 'tone identity'.

It would be valuable to know whether responses—both spike rates and high-gamma power—are consistently stronger for one tone than the other across patients and across channels/units. This would help clarify whether the two tones were matched in intensity. While the constraints of the operating room setting are understandable, such information is essential for interpreting decoding performance. For instance, high-gamma distinguishes the two tones at ~0.7–0.8 accuracy; to what extent does this indicate robust hippocampal discrimination under anesthesia, and to what extent may it reflect differences in stimulus intensity?

It is critical to understand that oddball frequency was fully counterbalanced between the two tone types, so systematic differences in driving effects between the two tone types would have cancelled out. The oddball analysis, which essentially quantifies the interaction between the two main effects, is, mathematically, orthogonal to the main effects of tone type.

Spurred by the reviewer's suggestion, however, we further investigated this difference and found that similar proportions of neurons responded to the two tone types. This new analysis allows us to go further - we can now say that tone identity discrimination effects reflect a balanced shift in the population response rather than a population response that is disproportionately greater for, in fact, any feature of one tone over the other (including stimulus intensity). We have added the following new text to the Results:

Main Text, Line 107
There was no difference in the proportion of neurons that significantly responded to each tone (p=0.184; N = 172 units; chi-squared goodness of fit test), thus tone identity discrimination effects reflect a balanced shift in the population response rather than a preferred acoustic feature of one of the tones.

3. The abstract still contains some formulations that risk overstating the conclusions. Slightly more measured wording could strengthen the manuscript. Examples include:
• Instead of "high-order perception," a term such as "higher-order pattern recognition" (suggested by reviewer #1), "sensory and linguistic processing without awareness" (suggested by reviewer #3) or the authors' own phrasing ("higher-order functions associated with parsing semantic and syntactic features of natural speech") may be more precise.

We have fixed this and gone with the reviewer's first suggestion. The revised abstract is reproduced here.

Consciousness is a fundamental component of cognition,[1] but the degree to which higher-order pattern recognition relies on it remains disputed.[2,3] Here we demonstrate the persistence of oddball discrimination, semantic processing, and online prediction in individuals under general anesthesia-induced loss of consciousness.[4,5] Using high-density Neuropixels microelectrodes[6] to record both single unit and local field potential neural activity in the human hippocampus while playing a series of tones to anesthetized patients, we found that hippocampal neurons and oscillations retained some detection of oddball tones. This effect size grew over the course of the experiment (~10 minutes), demonstrating representational plasticity. A biologically plausible recurrent neural network model showed that learning and oddball representation are an emergent property of flexible tone discrimination. Moreover, when we played language stimuli, single units and local field potentials carried information about the semantic and grammatical features of natural speech, even predicting semantic information about upcoming words. Together these results indicate that in the hippocampus, which is anatomically and functionally distant from primary sensory cortices,[7] complex processing of sensory stimuli occurs even in the unconscious state.

• I agree with the other two reviewers that some hidden technical factors may contribute to dynamics of oddball discrimination during the session. It could reflect plasticity/learning, but it would be better to tone down the second sentence "Here we demonstrate the persistence of plasticity… under general anesthesia."

We have now replaced this text with the more conservative "oddball discrimination" which summarized the result with less interpretation. The revised abstract is reproduced above.

• The statement that "hippocampus neurons could reliably detect oddball tones" may overstate the strength of the decoding (0.52–0.56). A more neutral phrasing—such as "hippocampal neurons and field potentials retained some detection of oddball tones"—may better reflect the results.

We have now implemented this suggestion. The revised abstract is reproduced above.

4. Finally, the qualitative and quantitative similarity of responses in awake and anesthetized conditions raises an interesting broader question: which neural signatures actually differentiate conscious from unconscious processing? It may be worth including a brief discussion. The current findings nicely show that such differences are not evident in hippocampal activity. This contrasts with studies in the visual domain showing clear distinctions in hippocampus unit activities between conscious and unconscious perception (e.g., in flash suppression Kreiman et al PNAS 2002, or backward masking Quian Quiroga et al PNAS 2008). Additionally, intracranial human studies comparing auditory responses across wakefulness, propofol anesthesia, and sleep (e.g., Krom et al., PNAS 2020; Hayat et al., Nat. Neurosci. 2022) are highly relevant and should

be cited. These studies report attenuated responses in high-order auditory cortex in unconscious states. Taken together with the current results, this may suggest two partially independent processes: (i) higher-order pattern recognition within the hippocampus that persists irrespective of consciousness, and (ii) a cortical auditory hierarchy that is more sensitive to conscious perception and perturbed by anesthesia. A short discussion of this interpretation would add valuable context.

We appreciate and agree with the reviewer's hypotheses. We have now added a discussion of these ideas to help the reader. We have now added a significant amount of text to the Discussion section of the manuscript to conclude our manuscript.

Line 529

These results raise the question about what features differentiate conscious from non-conscious processing. Prior studies have demonstrated that recognition enhances the medial temporal lobe response to a visual stimulus (Kreiman et al PNAS 2002, Quian Quiroga et al PNAS 2008) and, conversely, that propofol anesthesia (Krom et al., PNAS 2020) and sleep (Hayat et al., Nat. Neurosci. 2022) can suppress features of responses to auditory stimuli in the temporal cortex. Our results demonstrate that hippocampal auditory responses remain robust and retain the ability to encode information about the environment. One possibility is that there is an anatomical dissociation between higher order pattern recognition processes that occur in the hippocampus (and may occur other places as well) and with cortical processes that are correlated with consciousness, including in the cortical auditory hierarchy. Existing prominent theories consistent with our results include the idea that consciousness involves coordination between different structures (Dehaene & Changeux, 2011; Dehaene et al., 2006; Mashour et al., 2020), global propagation of local signals (Baars, 1988; Baars, 2005; Dehaene et al., 1998), or recurrent processing (Lamme, 2006; Lamme & Roelfsema, 2000; Pascual-Leone & Walsh, 2001). Another possibility is that consciousness is not simply a result of neural activity at a given time, but is instead constructed through repeated revisions of current experiences, and that it is this revision process that is impaired (Dennett, 1991), perhaps detectable in hippocampal replay (which we did not measure).

Referee #4

I appreciate the authors' revisions and new analyses/data. I think the changes have improved the manuscript and clarified some key issues. I particularly think that the ablation analyses of the RNN model provide important mechanistic insights into oddball effects.
There are, however, some remaining issues that I think need to be addressed.

1) This information is buried in the initial rebuttal, but as I understand it, the oddball effect is significant only in 1 out of the 2 patients. Even though it is significant when combining data across patients, this fact needs to be made clear in the text, and I think appropriate caveats should be included so readers are aware that this already small effect is not as strong as it may seem.

> *The reviewer is partially correct. While the oddball effect is not significant in p6 for the single unit analysis of the entire session, it is significant for the second half when the effect size grows (Figure 3C). It is also significant for the entire session for multiple LFP bands (Figure 2K) and phase-amplitude coupling (Supplement, page 7). We have now added the following text to make this caveat clearer in the relevant section.*
>
> *Line 144*
> **Oddball identity was decodable at above chance levels for the two patients combined (p<0.05), albeit not in p6 alone, and at lower levels that range from chance to 0.53. Dividing the LFP signals into individual frequency bands, we found that tone identity could be decoded across all bands (accuracy between 0.52 and 0.79, p<0.001, t-test). Oddball identity could also be decoded above chance in some frequency bands for both patients (p<0.001 for delta and high gamma bands for p5; and for all frequency bands except theta for p6), again at lower levels, ranging from chance to 0.62. These results therefore indicate that tone and oddball information is present in both single unit and LFP signals, although with oddball signals having substantially weaker strength than tone identity.**

2) Fig. 4M: they claim in the text that "overall decodability is greater under anesthesia of future words than for awake patients". While I think I can see this at word 0, it's not clear to me that it's the case for past or future words, and no statistical test is used to back up this claim. In fact, if this is true, then doesn't that suggest something very surprising is happening? How could it be that prediction of future words is BETTER in an unconscious state? While not an impossibility, this raises red flags for me that suggest either the data are noisy (maybe n is too low for the anesthetized group), or these neurons are actually encoding something else.

> *The reviewer raises a valid point that motivated us to devise a more thorough analysis to test this point in detail. Our previous analysis had been based on aggregated data across multiple time bins, which is statistically valid but can miss subtleties in the dynamics. We*

have now performed a fuller analysis comparing each point to its corresponding point in the complementary dataset, and separating past (memory) from future ("prediction" with caveats) conditions. While this approach is fundamentally more conservative (and therefore less statistically sensitive) it provides more detailed information about each specific word position in the past and future. We find that there is a significant prediction effect in both the awake and anesthetized data, but that effect is not significantly different for any particular time bin in the range of +1 to +5. (This is the core range in which effects are observed). The revised text makes this point clear:

Line 382

These numbers are comparable to those for awake patients (β_0=0.227, τ_future = 1.081; τ_past = 0.895); indeed, individual data points do not differ significantly (p>0.05 in all cases in the range +1 to +5, Figure 4M).

3) The point about a positive correlation between firing rate and word frequency requires more explanation. They argue in the rebuttal that this may be a difference between single neuron and fMRI/MEG/EEG, though this isn't supported. Indeed, single neuron prediction error for auditory stimuli is quite well established (for just one recent example, see https://www.nature.com/articles/s41467-017-02038-6). I'm not arguing with their findings, but I don't think it's sufficient to just brush it off and assume that the single neuron result is "truth".

The reviewer is entirely correct that we should not assume single neurons are truer than LFP - and if we accidentally implied this we apologize. We have revised the manuscript to address this concern in two ways. First, generally, we have now worked to incorporate the LFP data widely into the text (this also addresses concerns of other reviewers). Second, specifically, we have revised the place in the text where this is mentioned, reframing it so as not to favor either measure as more correct. That revised text is reproduced here:

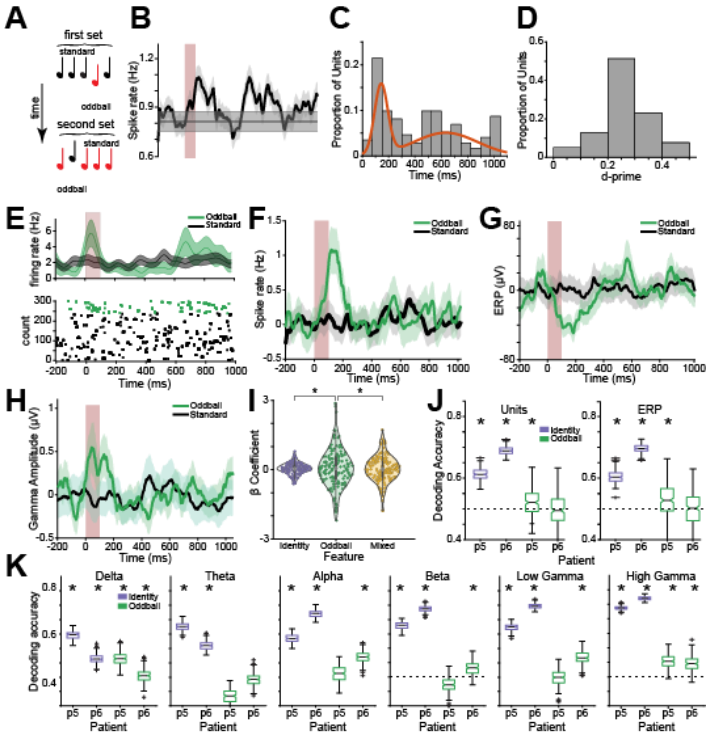
Line 295

Specifically, we find a significant positive correlation in the single unit activity (mean r=0.48+/- 0.06, Spearman's correlation, α<0.05) and across patients (p<0.0001, GLME) and a modest but significantly negative correlation in all six bands of the LFP (range -0.08 in Beta band to -0.02 in Delta band, p<0.001 for all).

We should note, however, that neural responses to infrequent words may be unrelated to their response to errors. Word frequency is not the same as lexical errors. If someone says "throw out the baby with the _____" then "bathwater", a very rare word, is not surprising, whereas "time" (the most common noun in English) would be an error.

4) The point about smoothing kernels used to compute PSTHs doesn't seem to address the underlying question that was raised in the first review. Whether the filter is causal or non-causal (and I agree that there is no perfect solution), they show the individual spikes in Fig. 2E for that example neuron. The higher spike density for oddball starts before t=0. Since they used a variable ISI, I find this very surprising, both that there could be such a fast or even predictive evoked response (in hippocampus, no less), and also that it's specific to oddball stimuli. While I don't think the core claims of the paper are dependent on this timing question, I think that this will be noticed by readers and raise questions that (perhaps unfairly) distract from the main points.

> *We agree with the reviewer that such a short latency response is unlikely, and that this plot may be misleading. Given that this is an example cell, designed to cleanly illustrate a more general pattern, it is worth finding a cleaner example. We have replaced the example figure panel with another (reproduced below in 2E):*



5) While I think it makes sense to change the framing from "learning", simply replacing that term with "plasticity" doesn't solve the underlying problem. I still think they haven't ruled out plausible alternatives for the changes they see over the course of the experiment. In fact, something like stimulus specific adaptation (SSA) could explain the decreased tone identity

decoding, which itself could be completely independent from the very modest increase in oddball decoding. They now argue in the Discussion that it's not likely to be something like adaptation because those occur over faster timescales, but I'm not sure this result from sensory cortex is known to translate to hippocampus. In fact, looking at the data in Fig 3D (ignoring the linear fits), I wonder how strong of a negative correlation there is between tone and oddball. If anything, the linear trends seem misleading, as both effects appear to be non-monotonic, and ultimately not be related to each other.

> *The reviewer proposes that "plasticity" is an unjustified framing, because of the possibility of, for example, stimulus specific adaptation (SSA). While the reviewer is correct that SSA is a possible explanation for our data, we believe that SSA is a form of plasticity. In fact, "plasticity" is the broadest available term that refers to changes such as SSA.*
>
> *The reviewer also raises a concern about the relationship between the oddball effect and the tone identity effect. While stimulus-specific adaptation (SSA) is well characterized in auditory subcortical and cortical areas, the reviewer is correct that timescales and mechanisms by which passive sensory exposure shapes hippocampal population codes remain largely uncharacterized in the anesthetized state. However, we emphasize that there is a strong correlation between the evolution trajectories of oddball and tone decoding accuracy (units r=-0.23, p=0.03). We can also point out that 4/6 LFP trends also showed a significant negative correlation. Note that our data do not allow us to directly identify a mechanism underlying these changes. In particular, stimulus-specific adaptation, inherited sensory dynamics, or slow network-level reweighting (or even more than one) could contribute. However, the coordinated, gradual reorganization of hippocampal population decoding along dimensions aligned with stimulus predictability is more naturally interpreted as a unified representational computation than as two unrelated processes evolving in parallel.*
>
> *We have therefore added quantification of this trend to the results, supporting the possibility that the tradeoff between these two processes reflects a task-related computation, while expanding the discussion to acknowledge gaps in the literature and alternative interpretations including SSA:*
>
> *Line 202*
> **There was a significant negative correlation between the evolution trajectories of oddball and tone decoding accuracy (r = -0.23, p<0.03, Pearson's test).**
>
> *Line 500*

There are several important limitations for the current study. First, anesthesia, while medically important, has an uncertain relationship with waking life[54]. Moreover, it remains unclear whether what we have learned will apply to other non-conscious states, such as sleep and coma.[55] Indeed, because all of our patients were anesthetized with propofol, it is not even clear our results will generalize to other anesthetics. Another limitation is that we did not have enough patients to test for lateralization / dominance. It is well understood that language is a bilateral process even in highly lateralized individuals, so it is not surprising that we do observe some language in the non-dominant hemisphere. Another limitation is related to our modeling. Our modeling does not indicate that the processes we describe are unique to the hippocampus. Indeed, the RNN relies critically on the formatting of its input into binary categories, which must be inherited from other regions. Instead, the goal is to understand how oddball-dependent responses can arise from such input. Furthermore, we cannot definitively conclude that plastic changes in tone identity decoding during the oddball task are a compensatory property of the oddball response, rather than an independent plastic effect such as stimulus-specific adaptation. Finally, our choice to use temporally unpredictable tones may have weakened our sensitivity to tone-related effects as there is some evidence that temporally predictable deviant events may maximally drive neural responding.

Dear Editor,
We appreciate the opportunity to continue to improve our manuscript. We are attaching our replies to the small number of remaining comments here.

Core data used for analysis have been uploaded to DABI:
https://dabi.loni.usc.edu/dsi/U01NS108923

Code used to implement the computational modeling and analyses in this study is publicly available on github:
https://github.com/NuttidaLab/rnn_oddball

Referees' comments:

Referee #1:

The authors did a thorough job of addressing our input. We would suggest addressing a few additional points.

The authors report: For each trial and channel, we computed the modulation index (Tort et al. 2010) to quantify phase-amplitude coupling of gamma (30-80 Hz) to theta (4-8 Hz) power.

However, high gamma at 70 to 150 or 200 Hz best correlates with SUA. The authors should analyze PAC for Theta-High gamma not just low gamma.

We now include results quantifying the role of theta-gamma PAC (phase-amplitude coupling) using both low gamma (30-80 Hz) and high gamma (70-150 Hz) frequency ranges. In summary, high gamma phase-amplitude coupling improved channel-level decoding of all three categories (tone, oddball, and interaction) compared to low gamma. The new analysis using high-gamma PAC modestly improved decoding performance for language as well. This new finding does not change our overall narrative, and, if anything, strengthens it.

We have now amended the Results as follows:

High gamma phase-amplitude coupling reliably detected tone and oddball stimuli: 26.6% of channels demonstrated significant tone encoding, 21.96% of channels demonstrated significant oddball encoding, and 24.07% of channels demonstrated an interaction ($p < 0.05$). Leveraging SVM decoding, high gamma phase-amplitude

coupling predicted tone identity (p5: tone accuracy = 0.553; p6: tone accuracy = 0.542), but could not decode oddballs. Low gamma phase-amplitude coupling did not significantly encode tones, oddballs, or their interaction. Low gamma phase-amplitude coupling could weakly predict both tone identity and oddball status in p6 (p6: tone accuracy = 0.505, p < 0.001; oddball accuracy = 0.527, p < 0.005), but not p5 (p > 0.05 for both tone and oddball).

Phase-amplitude coupling in language:

Theta-low gamma phase-amplitude coupling also outperformed spectral features in predicting semantic embeddings (mean r = 0.170 +/- 0.13, rank sum test, p<0.001 for all bands) and similarly encoded multiple semantic and POS categories (74.5% and 73.0% encoded at least 8 categories, respectively). Phase-amplitude coupling using the high gamma band (70-150 Hz) showed comparable semantic embedding prediction (mean r = 0.166 +/- 0.13) and improved encoding of semantic and POS categories (99.3% and 98.7% encoded at least 8 categories, respectively).

The spike-frequency coupling was only analyzed in the oddball experiment, but not the speech experiment.

In our revision, we actually did include an analysis of spike-frequency coupling for language near figure 4 in the main text, where we indeed found spike-frequency coupling to have the least predictive value for semantics (a result that contradicts the reviewer's expectation). This result was, however, buried at the end of a long paragraph. We apologize that the organization of this information was confusing. We have now revised the manuscript to emphasize this new result, giving it its own paragraph with topic sentence highlighting it: "We also found robust phase-amplitude coupling results for the language data."

Spike-frequency coupling is typically related to a behaviorally relevant task. If the claim is that the 'task' is pattern recognition, and they're getting better over time, then one might expect that spike-frequency coupling is stronger in the second half of the experiment vs. the first half of the experiment for the oddball task only for units that encoded features above chance (neurons not encoding those features wouldn't make sense to have stronger coupling). One might expect this to not be the case in the language experiment, and instead for coupling to generally be stronger for neurons encoding semantic categories across the recording session. This analysis should also be performed for theta-HFA.

We have now performed the new analyses that the reviewer suggested, comparing how spike-frequency coupling and phase-amplitude coupling differ between first and second half of the oddball task. In summary, both features significantly vary vs. time during this task. We have added the following text to the supplement:

Phase-amplitude coupling demonstrated a significant difference in tone decoding accuracy between first half and second half of the oddball experiment (t-test; $p<0.0001$), in both patients. Oddball decoding only significantly increased in the second half of the oddball task for p6 ($p<0.0001$), not p5. Spike-frequency coupling showed a significant increase in both tone and oddball decoding between task halves, and in both patients ($p<0.0001$).

Furthermore, we agree with the reviewer's suggestion; it would not make sense to perform this particular analysis in the language task given the non-significant performance of spike-frequency coupling for semantics, and we do not hypothesize any variation in encoding performance over time during this task.

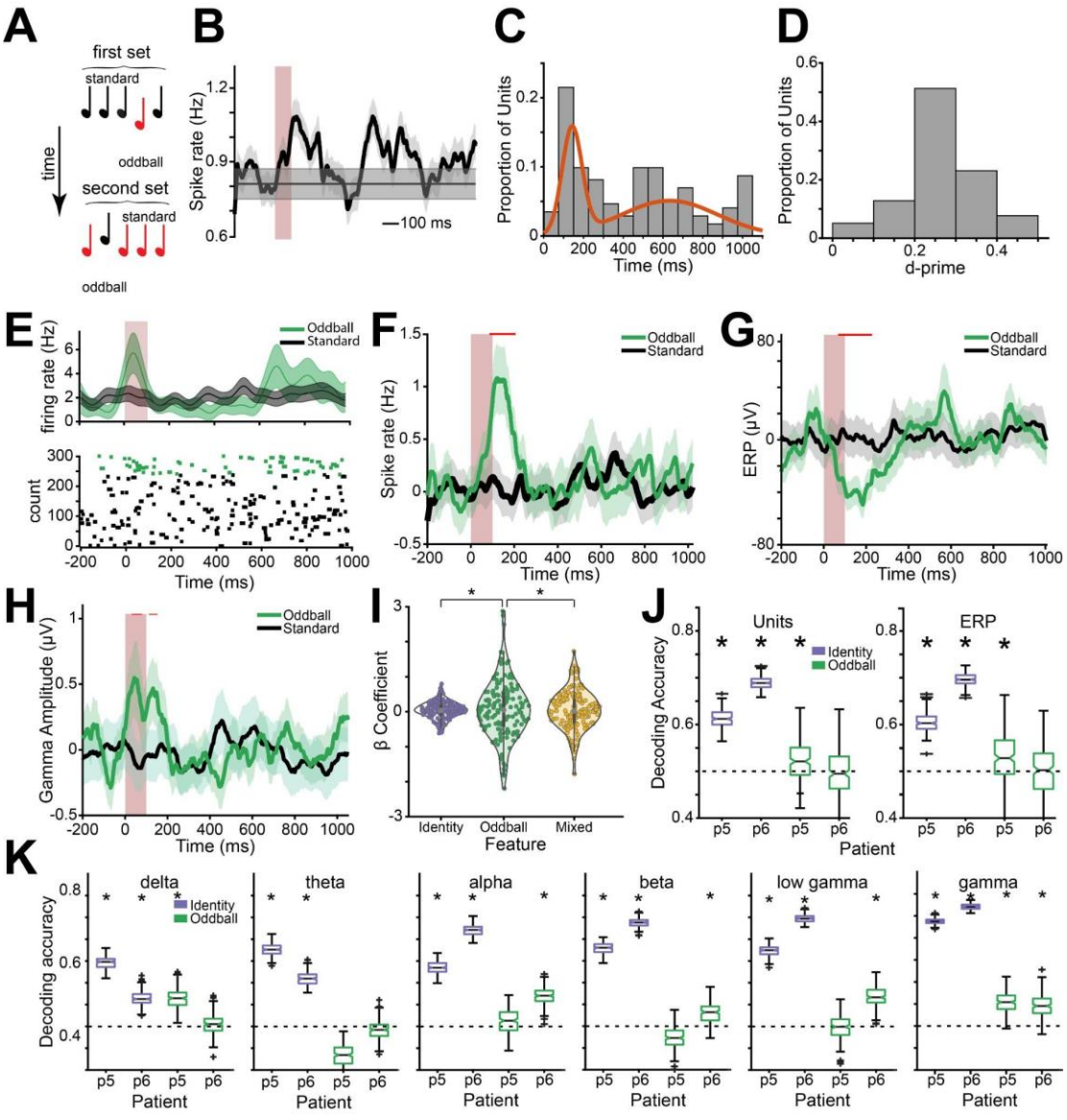
Referee #4:

The authors have addressed most of my concerns in this revision.

Before they finalize the paper, I suggest they double check the results related to the point I raised previously about neuronal firing starting before stimulus onset in the tone paradigm. While it helps to have replaced the example neuron in Fig 2E with one that shows a more plausible raster, the plots in F, G, and H have the same issue that was raised earlier. I suppose it's possible that the mean spike rate and gamma amplitude responses start before t=0 due to the filtering issue they brought up, but I don't see how this is possible for panel G, which just shows the raw voltage traces. Unless I'm misunderstanding what they mean by "ERP", this should be a largely unfiltered signal except for hardware lowpass and anti-aliasing. Even though it's an average across multiple channels on the probe, I don't see an explanation for this implausibly early response. Furthermore, since it's there in the raw data, this suggests to me that the reason it appears in the filtered and smoothed versions of the data is not simply due to the signal processing.

As the review states, panel G is not smoothed. The peak the reviewer identifies is just noise. Indeed, we can quantify the probability that it is some kind of signal (regardless of whether it is real or artifactual) using a

conventional statistical approach. This can be seen from the fact that the error bars overlap between stimulus conditions throughout this pre-stimulus period. To further demonstrate this point, we now include a small red bar above the plot indicating periods of statistical significance. We do this in panel G, as well as in the other panels the reviewer is concerned about - F and H. These results, reproduced below, show that there is no statistically significant signal before time zero. In other words, in all cases, the effect the reviewer notices is demonstrably due to noise. There are various things we could do to eliminate this noise, but we feel it is more honest to present the data as they truly are.



One additional minor point they should double check is the p6 oddball delta result in Fig. 2K. It's hard to tell with the notches on the box plot as it's rendered, but it's surprising to me that the effect is significant, especially compared to others in the same plot that are not (e.g., p5 alpha oddball).

> **This result (oddball decoding using delta power in p6) was not significant (p=0.1159) and we apologize for mistakenly labeling it as such. Fortunately, this does not change the narrative, as many other bands still demonstrate strong effects. We have corrected the text and figure corresponding to this section and double-checked that the remaining p-values are correct.**

> **Oddball identity could also be decoded above chance in some frequency bands for both patients ($p<0.001$ for delta and high gamma bands for p5; and for all frequency bands except delta and theta for p6).**