nature portfolio

Peer Review File

Distinct neuronal populations in the human brain combine content and context

Corresponding Author: Professor Florian Mormann

Any redactions in this file are there to maintain patient confidentiality, the confidentiality of unpublished data, or to remove third-party material.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

Attachments originally included by the reviewers as part of their assessment can be found at the end of this file.

Version 0:

Reviewer comments:

Referee #1

(Remarks to the Author)

In this manuscript, Bausch and colleagues explore a relatively rare dataset of human single unit spiking activity in the medial temporal lobe to examine how representations of content and context may be linked through single unit activity. In their behavioral paradigm, subjects are asked to make decisions about two serially presented images, such as which one is bigger or which is one they like better. The authors consider the different questions as different contexts, and then probe the spiking responses during visual presentations of the images. They find that a small subset of neurons are selectively responsive to individual visual stimuli, recapitulating previous findings. They also find that a largely separate and non-overlapping subset of neurons are selectively responsive to context. They also find a much smaller subset of neurons that respond to the conjunction of both stimulus and context. The focus of their study is mostly on the second subset, what they call mere context neurons since they are selectively modulated by context independent of visual stimuli, and how their activity may be linked to the representations of content (visual stimuli). They show that context information can be decoded from this subset of neurons throughout the different periods of the task, and that activity of these neurons in the hippocampus is preceded by the activity of stimulus responsive neurons in the entorhinal cortex. The use this latter finding to support their primary claim that these different subsets of neurons link their activity in order to form representations of content in context.

Overall, the authors attempt to address an important and key question related to memory formation, which is how information regarding the different aspects of an experience or memory are linked together in the human brain. The authors should be commended for their work, as capturing and analyzing single unit data from the human brain is rare. The finding that some neurons are selectively responsive to different task conditions, or questions, that the authors refer to as context is novel and would be of interest to many different researchers. I think this study has the potential to be a valuable addition to the literature, but there are some issues that would need to be addressed in order to support the primary claims of the paper, and some suggestions that could strengthen the overall manuscript.

Major concerns

One question regarding these findings relates to the behavioral paradigm itself. In this paradigm, subjects are asked to evaluate two serially presented pictures based on some criteria that is presented first as a question. For example, which of the following pictures is bigger. The authors consider this as context for the visual presentations. But this is an interesting, and somewhat unusual, definition of context. When reading about the paradigm, my first reaction was that the paradigm was asking subjects to view images with different task goals. Task goals could be considered as context, indeed, but there are also some potentially confounding factors. For example, if one is asked to think about which item is bigger, this could serve as a targeted salience signal, such that during subsequent visual presentations, the subject is focusing simply on the dimension of size when viewing an object and ignoring the many other possible stimulus features. Does this define context? Or does this instead highlight a possible mechanism whereby neurons are selectively responsive to different stimulus features, in which case the entire interpretation of the results would be different. Moreover, related to this point, the authors

then posit that these data therefore provide insight into memory formation and retrieval and how we link content and context. Yet there is no explicit test here of memory, or specifically how the specific item (picture) is linked to the specific context. It is hard to interpret the neural data as providing evidence regarding the link between context and content for memory if this is not actually what is happening in the task.

A second major issue relates to the statistical analyses performed, and how data are pooled together. The main issue here is that it appears that many of the analyses are performed by pooling data across the entire population of neurons. This is especially true for the correlation analyses in figures 4 and 5, but also for the stimulus responses in figures 1 and 2. These data are collected from multiple different subjects, and from multiple sessions in each subject. At a minimum, the data should be analyzed at the level of sessions, but really at the level of subjects. Even at the level of sessions, are different sessions really considered independent samples. What if all of the main significant effects derive from one subject (with several sessions). Would these results be generalizable. I think the issue of independent sessions could be explained by the possibility that different recording days could capture different neurons, but this should be explicitly shown. In addition, it should also be shown that this does not artificially confound the data. In other words, the same analyses should be performed at the level of subjects, either averaging across sessions within a subject or simply taking one of each subject's session. Analysis at the level of sessions seems to be performed for most of the analyses in Fig 3 for example. The issue of pooling is more problematic though when considering the analyses on correlations. The correlations between neuron pairs in Figure 4 are pooled across all possible pairs from all subjects. But again, how consistent is this across subjects. The correlations of Figure 5b, c are also pooling across all neurons (and possibly all trials separately, although the number of samples here is not clear), which would have the effect of generating extremely low p value just due to the large number of samples. Again, these correlations should be performed within each subject/session and then analyzed across subjects/sessions.

There are several other points of confusion related to the statistics here:

• In some cases, the authors examine 38 sessions, while in others they examine 50. In the methods, they state they analyze 38 sessions, but in results they state they analyze 50 sessions. It looks like 38 sessions are analyzed for context neurons but 50 for stimulus neurons (Fig 3a). Why? Or, for example, in the analysis comparing the ANOVA to stratified label shuffling controls, which confirms the main effect, they report N=50 for both stimulus and context.

• In some cases, the number of neurons that are being analyzed are unclear. For example, in Figure 3f, they report N=400 neurons. Which neurons are these? For the classifier analysis of the pooled data, they analyzed 30 subsamples of 150 neurons. But there are less than 150 mere context neurons. So are they also including contextual stimulus and stimulus-context neurons here.

• Timing. The selection of timing windows for analyses seems to differ between different analyses, but it's not clear how those time windows are chosen. For example, they find that training the decoder during question presentation leads to good decoding during picture presentation. In both cases, they only look at latter half of activity (600-1000ms). Why? Is this derived from the heatmap in Figure 3c? If so, then for these heatmaps, we should be able to identify time regions of significance (across sessions and/or subjects) and use those time regions for subsequent analyses. Or, for example, for the cross-correlation analyses, why look at large window of 1500 ms if the picture is only on screen for 1000 ms (plus 200-300 ISI plus 200 ms orientation). Or, for example, for the analysis looking at the correlation of firing between question and picture, looked at a different window of 500-1300.

Another concern relates to the overall conclusion of the manuscript and the model presented in figure 5. This model largely rests upon the primary result that there are increased correlations between EC stimulus neurons and hippocampal context neurons, with a time delay of tens of millisecond. Thus, this is taken as the primary evidence that there is sequential activity between stimulus response and context response. In addition, the authors observe elevated correlations between EC and H neurons during the experiment and after the experiment, but not prior to the experiment, which could suggest that STDP has increased the link between these sets of neurons. This is a nice result, but there are some potential concerns with drawing these firm conclusions. First, there are the issues related to statistical testing discussed above, especially with regard to the pooled correlations in figure 5. Second, these data demonstrate that there is a single pairwise correlation that is significant (among all pairwise correlations). But in addition, what one would like to see if that these relations between EC stimulus and HC context cells are specific to individual stimuli and individual trials. Are these analyses performed just during the responsive trials, or is this a generic relation regardless of whether the neurons actually respond. With respect to the model, the authors posit that context activity is reinstated once the EC neurons are activated. This could be explicitly shown by examining decoding of the context representations, based on training during the question period, time locked to EC neuronal firing. In addition, related to this point, the reactivation of context signal is much more clearly demonstrated with a decoding classifier approach rather than simply examining the correlations in firing rates between stages of the experiment (e.g., question to picture). Overall, this is a potentially compelling model, but there are some analyses that could help strengthen this claim.

Interestingly, related to this point, the authors also demonstrate that classifiers can be trained on time periods during the question, or during pictures, and used to decode context during other stages of the task. For the hypothesis that the link between context and content is made explicitly between EC stimulus neurons and hippocampal context neurons, one would like to then know why context signals can in fact be decoded in multiple brain regions throughout the task, rather than just in the hippocampus, and whether these context representations are therefore irrelevant (Fig. 3e). In addition, if these are mere context neurons, then we should expect that decoding accuracy should be elevated during the actual question as compared to the picture presentation. The question is, indeed, in this case taken as the context signal, so presumably this should be

the time point of greatest context signal. This goes back to the question of whether what is being shown here really is context, or some other stimulus feature. But in this case, decoding during the actual question appears significantly worse than during other periods of the task like during picture presentation. Why?

There are two other questions that these results raise that could be addressed with these data and that could also have direct bearing on the conclusions and that should be considered.

First, what is happening at the time of choice when subjects have to choose one picture or another. If the link between context neurons and content neurons is related to memory retrieval, as the authors posit, then this should also be apparent when subjects are actually making their choice and therefore retrieving the memory of the two pictures and the associated context to choose correctly. In other words, is there evidence that item in context information is retrieved? Providing these analyses and perhaps showing that these two subsets of neurons are engaged during successful choice (and therefore memory retrieval) would provide stronger evidence that this link is meaningful.

Second, perhaps the strongest evidence here that context and content are linked in this dataset is the presence of stimulus context neurons, those neurons that only respond to the conjunction of a specific stimulus and a specific context. It seems that these are an important population of neurons, yet any further analysis beyond just identifying them is dropped. It would be helpful to know how activity of these neurons is related to the context and content neurons in the different brain regions, and how this may related to the correlated activity between them.

Minor points

In Fig 2c, it looks like most neurons that respond to context are also responsive to stimulus. Yet this appears to be the opposite of what is depicted in the Venn diagram in Figure 2a, where it looks like most context responsive neurons do not respond to stimulus?

For the classifier results in Fig 3a, what is the overall level of decoding classification across sessions/subjects? The results here are presented separately for each category.

For the analysis on correlations between mere stimulus and mere contrast neurons, presumably these neurons should only respond on a small subset of trials. But the result here is strange. If there is a non-preferred stimulus (or context), then there should be no response, and yet we are seeing responses. So are these really selectively responding neurons (e.g., mere stimulus or mere contrast neurons).

Referee #2

(Remarks to the Author)
This manuscript describes a study examining the extent to which single neurons in different regions of the human medial temporal lobe (MTL) demonstrate sensitivity to specific stimulus events, operationalized as one of 4 different pictures, and ‘contexts’, operationalized in terms of one of 5 different comparative questions asked of picture pairs, as well as whether any neurons demonstrate either joint sensitivity to these factors, or respond to them conjunctively. The study also examined temporal relationships, in the form of lagged cross-correlations, between firing patterns in different MTL regions. The findings are interpreted as evidence that, for the most part, individual stimulus items and contexts are separately represented at the single neuron level and that there is a temporal dependency between ‘context’ neurons in the hippocampus and earlier firing item-related neurons in entorhinal cortex.

This study has some notable strengths – the conclusions are based on what, at the level of single neurons (though see below) is a large sample, and the signal analysis methods employed are rigorous and comprehensive. Additionally, the question of how the brain and, notably, brain regions implicated in episodic memory such as the hippocampus, represent stimulus identities and disambiguate stimuli according to their contexts is a timely and important one. These strengths are offset by what in my opinion are a number of weaknesses, the most salient of which are described below.

i) The manipulation of ‘context’ employed here involves variation in the nature of the comparative judgment made on a pair of visual images (i.e. bigger; last seen; older; more expensive; etc.). This manipulation has the unfortunate consequence of changing not only the context, but those features of the stimulus events toward which attention is directed. For example, the features of the image of a truck that receive most attention will be quite different when it is part of an affective comparison rather than a comparison of, say, monetary value. Thus, the nature of the stimulus representations will change according to context. This might explain why some of the neurons identified in this study demonstrated conjunctive context/identity coding. More generally, the authors should provide the rationale for their decision to employ a manipulation of ‘interactive’ rather than ‘extrinsic’ context.

ii) The procedure employed to select the 4 stimulus items to be employed with any given participant requires justification. As I understand it, the items were selected so as to maximize the likelihood of identifying neurons in the experiment that would

respond selectively to only one of the 4 items. Two issues arise – first, doesn't this procedure bias the data set in favor of stimulus over context identity? In principle, a similar procedure could have been employed to maximize the likelihood of identifying discriminative responses to a sub-set of a larger pool of contexts. Since this wasn't done, it's perhaps unsurprising that far more stimulus-specific neurons were identified than neurons that responded selectively to a single context (e.g. fig 2A). Second, the identification of the 4 pictures subsequently employed in the experiment itself took place in a specific context, just not one that was controlled or manipulated. This raises the possibility that at least some of the neurons responding to these pictures in the experiment proper may be ones that were 'pre-selected' to be relatively impervious to contextual change.

iii) Most of the analyses are at the level of single neurons, with little attempt to address whether the findings generalize across participants (Fig. S1 is an exception). It would be useful to include some statistical analyses in which participants, rather than neurons, are treated as the unit of analysis, preferably in the context of a standard random effects approach.

iv) Neurons are sampled from a 4 different MTL structures that are thought to have very distinct functions and to support quite different computations. Moreover, especially with respect to the entorhinal cortex and, even more so, the hippocampus, different sub-regions of these structures are held to support different kinds of computation (e.g., the DG and CA3 hippocampal subregions are thought to support pattern separation to a much greater extent than CA1). Do the authors have any information about the likely localization of the neurons that they recorded from within these regions? If not, this issue should at least be noted as a limitation and a constraint on the model outlined in fig. 5A.

Referee #3

(Remarks to the Author)
In this manuscript, the authors present single unit data recorded during a visually presented object comparison task in 17 neurosurgical patients. They report that the vast majority of neurons in human MTL represent context and content separately. They find many stimulus-modulated neurons ("content" neurons; not surprising since stimuli were pre-selected), some context-modulated neurons, and a significant minority of neurons encoding the interaction. They suggest that the content and context information streams may be later combined through co-activation, synaptic modification, or reinstatement. They are able to decode context significantly from all neurons and context-neurons only. Decoding accuracy was better when stimuli were presented (i.e. when the context needed to be recalled in order to evaluate the stimulus). The authors also find that firing of stimulus neurons in EC reliably precedes firing of context neurons in HC, suggesting that the firing of EC stimulus neurons triggers a context association in HC, potentially via synaptic plasticity.

I congratulate the authors on their efforts to further disentangle the workings of the human MTL and its role in context-dependent memory. I do have a number of concerns and questions regarding methodology and interpretation. The major concerns are those that influence interpretability and impact. The specific questions need answers, but the answers are unlikely to significantly influence impact.

Major Concerns

The first paragraph of the results states 50 experimental sessions across the 17 patients, whereas the Methods say 38 sessions. The total number of units is reported to be 3,127, meaning either 63 or 82 units per recording session, which is extremely high. How many microwire bundles were placed? Typical yield per microwire is 0.5-1 unit per wire in optimal conditions, so either there were many bundles placed or the yield per bundle is higher than most groups report. Either of these possibilities warrants further discussion.

A related question is about uniqueness of units across sessions. What was the distribution of time interval between recording sessions? How confident are the authors that the units in one session were a completely unique set relative to the previous or the next one? Preliminary work from some groups suggests that there is a high degree of carry-over, such that only a fraction of units are different from one to the next session, especially with short delay intervals (hours). This effect would further reduce the total number of independently counted units. In fact, this point of neuronal population stability across sessions is a critical requirement of this experimental paradigm. Since neurons were pre-screened based on response to stimuli, the authors rely on the fact that there is stability of that population from the pre-screening step to the actual experimental step. How often was the pre-screening performed? What was the distribution of intervals between each pre-screening step and its respective experimental session(s)? The only way to incorporate both features (i.e., a pre-screen methodology and assumption of uniqueness of units across all experimental sessions) would be for each experimental session to have its own pre-screen step that immediately preceded it (interval of minutes) and for there to be a relatively much longer interval between each pres-screen/experimental session pair (many hours to days). Given the constraints of the EMU paradigm, I doubt this was the case. I think the most likely situation is that pre-screening does work (as evidenced by the high number of content encoding neurons) because there is significant stability of unit identity across sessions.

Regarding general statistical tests, why was alpha = 0.001 chosen? Why not compute p-values for individual neurons using stratified label-shuffling? I'm not sure why statistics were computed for individual data instead of shuffle averaged data over sessions.

The conclusion of spike timing dependent plasticity is going too far out on a limb. I suggest softening the instances when this possibility is mentioned as an explanation for the EC-HC results. This mechanism can be listed as a hypothesis to test but should not be as closely associated with explanatory power for the current data given the absence of specific evidence. In

fact, the authors should spell out the experiments that would need to be done (and are compatible to be done in humans) to test this hypothesis.

Since stimuli were pre-selected independent of context, I suggest removing the stimulus-selectivity statistic from abstract.

Specific Questions/Requests

Do any analyses include the response variable (whether the picture was presented 1st vs. 2nd)?

Please include fractions / %'s of cells in specific regions in abstract (instead of, or in addition to, absolute numbers. If keeping absolute numbers, please include population size).

Was there always a correct answer for each question and combination of stimuli? Did participants answer the same question consistently across different trial repetitions?

In the paragraph titled "Stimulus and context are mainly encoded separately and rarely conjunctively", did the ANOVA model used include the interaction term?

"Next, we assessed pooled decoding performances from 30 random subsamples of 151 context neurons or 151 stimulus neurons" --> Please clarify the purpose of the subsampling in this analysis?

Please present the pooled decoding accuracy next to the per-session decoding accuracies in figure 3a? (using the same training / testing procedure)

Can the authors clarify this sentence in statistical terms in light of their example: "we also identified contextual stimulus neurons with an additional main effect of context": neuron with significant main effect of stimulus and significant interaction between stimulus and context? --> no, neurons with main effects for both stimulus and context (clarified in methods). Please clarify in main text.

Consider more distinguishing terminology for contextual stimulus neurons (main effect of context and stimulus) vs. stimulus-context neurons (interaction between stimulus and context).

The data availability statement needs links to data and code.

Version 1:

Reviewer comments:

Referee #1

(Remarks to the Author)
The authors have provided a substantial revision to their original manuscript, and have addressed many of the concerns raised in the initial round of reviews. Specifically, they have provided several additional analyses at the level of individual subjects, and have introduced a decoding analysis based on an SVM. All of these new additions strengthen the overall conclusions of the manuscript.

The two main outstanding questions, though, are related to how to interpret these results.

The first question is whether indeed this reflects a context signal. This was raised by the other reviewers as well. I appreciate the point that different contexts can demand different behaviors (e.g., considering the price of food in a store versus the taste of food at home), and the authors make this point to note that the demands of their behavioral task may be interpreted as a generalization of different contexts. This is an interesting point. However, there still remains the alternative explanation that what is being asked of the subjects is to specifically focus on or pay attention to one of the stimulus features. As the authors note, this could indeed happen in different contexts, but the actual manipulation here is not of the contexts but instead of the task demands. In fact, one may posit that if a subject were in a supermarket and asked to either purchase an item or taste an item, both of those different demands would occur within the same context. It is unlikely in that case that the authors would argue that these different demands therefore represent different contexts. I think this has some bearing on how to interpret and view these results, because it seems more likely that what we are observing here are neurons that are tuned to the particular task demands or stimulus features, which then would change the overall interpretation of the interaction between the encoding of content and in this case the encoding of the salient stimulus feature.

This relates to the second question, which is whether this interaction is related to memory. Here the authors do not provide an explicit memory test, and yet their conclusions are based on the interpretation that the integration of content and context is important for memory. In their revision, the authors introduce a new analysis in which they examine the consistency and transitivity of the behavioral preferences. This is important, of course, and as the authors note, it would be difficulty to be consistent in your answers if you do not remember the order of the stimuli that are presented. But this does not feel like an explicit test of memory, let alone a memory that requires the integration of content and context. In fact, if one were to just provide a behavioral response to an image and asked to provide a judgement, one would expect that the same image would

evoke the same judgement. This is consistent (and would also exhibit transitivity since other objects would also have consistent judgements), but this does not seem to involve memory. The memory component here seems to be the fact that the subjects remember which picture is which when making their judgement. This could, as the authors argue, mean that they are remembering the specific item and the question being asked. However, another possible behavioral strategy would be to pay attention to a specific stimulus feature, and then when viewing the images note the feature of the first, and whether the second is bigger or smaller, for example. This does not seem like an integration of context and content in the traditional sense. Without an explicit test of memory, for example remembering the occurrence of a particular event or stimulus within a particular context, it's not clear how these data here would provide evidence that such interactions are related to memory encoding. This is of course confounded by the point that these interactions may not truly reflect context and content, but instead some stimulus feature and the stimulus content.

To be clear, the authors have indeed provided a very thorough revision to the original comments, and there are several valuable aspects of this study that would be of broad interest to a large community of researchers. Certainly the analyses of the single unit data are sophisticated and well done, and the new additional analyses on decoding and examining correlations and activity at the level of individual subjects are important. The question of how to interpret the data and the conclusion regarding context and content processing in the service of memory is less clear.

Referee #2

(Remarks to the Author)
The authors have done a good job in addressing my methodological concerns and have been attentive to some of the conceptual/theoretical issues that were raised as well as the need to discuss limitations to the study. I have only two further comments:
i) The authors still do a poor job in how they handle the notion of a 'context'. They now argue in the discussion, unconvincingly in my view, that 'real-word' contexts invariably impact how stimulus content is processed, as exemplified by their melon/supermarket, melon/kitchen example. But as their own example at the beginning of the paper (same friend, two different restaurants) attests, this is only true of 'interactive' contexts; 'independent' or 'background' contexts (such as the restaurant) do not impact the processing of the the associated stimulus content, at least not to anything like the same degree. There is a rich literature on this distinction in the experimental psychology of memory, see for example, Baddeley, A. D. (1982). Domains of recollection. Psychological Review, 89(6), 708–729. I suggest that the authors acknowledge this distinction, and simply make it clear that their experiment is concerned with the neural representation of interactive (task) context, rather than making a somewhat spurious claim for the ecological validity for their chosen contextual manipulation.
ii) The abstract is far from clear,and appears to be missing a sentence (lines 17-22 cannot be parsed, at least in my case).

Referee #3

(Remarks to the Author)
Many points from the previous round of revision have been answered, including bias towards stimulus selective units, pooling of data, unit yield, and others.

I have a few lingering concerns:

1. Regarding stimulus-context neurons and their relationship to stimulus or context neurons. It is understandable that the number of these neurons is too low to perform cross correlations, but they represent such an interesting subset that perhaps the authors could discuss them a bit further. How is the activity of these neurons related to that of context and content neurons in the different brain regions past an assessment of cross correlations? For example, were these neurons more likely to be found in areas where there were higher proportions of stimulus or context neurons? Was their activity better related to trial accuracy than the other groups?
2. Behavioral performance. The authors enhance the manuscript by including performance metrics that allow for subjectivity. However, for the questions in which an objective response is present (e.g. Bigger?) it would be possible to measure accuracy. Additionally, if the activity is relevant to the task the authors should be able to show a connection between task performance and neural activity, e.g., is neural activity predictive of behavioral errors of the objective questions? There is one sentence in the methods describing a single patient who failed the task and had no context sensitive neurons-that is a very interesting result that doesn't seem to be highlighted. Can this be seen on a trial by trial or session by session basis within patients performing the task?

Version 2:

Reviewer comments:

Referee #1

(Remarks to the Author)
The authors have revised their manuscript and have performed a commendable job in addressing some of the outstanding questions raised during the last review. In particular, I appreciated their efforts in clarifying the distinction between 'task' context and environmental or background context, and their emphasis here on task context which seems appropriate. They have also sufficiently addressed the point about memory. Overall, they revised the manuscript well, and this is a valuable

and well conducted study and will be of interest to the broader community.

Referee #2

(Remarks to the Author)
The authors have satisfactorily addressed my remaining concerns in this iteration of the manuscript.

Referee #3

(Remarks to the Author)
The paper is careful, rigorous, and adds valuable human single-unit data. The dataset of thousands of human single units is especially impressive. All technical concerns on the analytical rigor have been adequately answered. However, the experimental design falls short of supporting the broad claims made in the discussion. First, the authors are limit themselves to stimuli that maximally activate units. Second, as the authors acknowledge, they are limited to "task context," a far narrower concept. As such, rather than an expansive description of how the mesial temporal lobe pairs stimulus and context in a naturalistic setting, it is instead limited in both domains. Therefore, this paper makes a solid contribution that is not paradigm-shifting but is worthy of general attention in the community.

Referees' comments:

Referee #1 (Remarks to the Author):

In this manuscript, Bausch and colleagues explore a relatively rare dataset of human single unit spiking activity in the medial temporal lobe to examine how representations of content and context may be linked through single unit activity. In their behavioral paradigm, subjects are asked to make decisions about two serially presented images, such as which one is bigger or which is one they like better. The authors consider the different questions as different contexts, and then probe the spiking responses during visual presentations of the images. They find that a small subset of neurons are selectively responsive to individual visual stimuli, recapitulating previous findings. They also find that a largely separate and non-overlapping subset of neurons are selectively responsive to context. They also find a much smaller subset of neurons that respond to the conjunction of both stimulus and context. The focus of their study is mostly on the second subset, what they call mere context neurons since they are selectively modulated by context independent of visual stimuli, and how their activity may be linked to the representations of content (visual stimuli). They show that context information can be decoded from this subset of neurons throughout the different periods of the task, and that activity of these neurons in the hippocampus is preceded by the activity of stimulus responsive neurons in the entorhinal cortex. The use this latter finding to support their primary claim that these different subsets of neurons link their activity in order to form representations of content in context.

Overall, the authors attempt to address an important and key question related to memory formation, which is how information regarding the different aspects of an experience or memory are linked together in the human brain. The authors should be commended for their work, as capturing and analyzing single unit data from the human brain is rare. The finding that some neurons are selectively responsive to different task conditions, or questions, that the authors refer to as context is novel and would be of interest to many different researchers. I think this study has the potential to be a valuable addition to the literature, but there are some issues that would need to be addressed in order to support the primary claims of the paper, and some suggestions that could strengthen the overall manuscript.

We appreciate the reviewer’s succinct summary and positive evaluation of our manuscript. Based on highly valuable questions and analysis suggestions, the manuscript has been extended and strengthened significantly, both analytically and conceptually. Among others, we refined our discussion of context representations, demonstrated that subjects remembered contents (i.e. pictures) together with their associated context (i.e. question) and replicated the vast majority of analyses on a subject level (or in a few cases on a session-level with second-level neural analyses). Additionally, we added four novel SVM decoding as well as other analyses that strengthen our previous findings and model, which are now incorporated into figure extensions and multiple Supplementary Figures.

Major concerns

One question regarding these findings relates to the behavioral paradigm itself. In this paradigm, subjects are asked to evaluate two serially presented pictures based on some criteria that is presented first as a question. For example, which of the following pictures is bigger. The authors consider this as context for the visual presentations. But this is an interesting, and somewhat unusual, definition of context. When reading about the paradigm, my first reaction was that the paradigm was asking subjects to view images with different task goals. Task goals could be considered as context, indeed, but there are also some potentially confounding factors. For example, if one is asked to think about which item is bigger, this could serve as a targeted salience signal, such that during subsequent visual presentations, the subject is focusing simply on the dimension of size when viewing an object and ignoring the many other possible stimulus features. Does this define context? Or does this instead highlight a possible mechanism whereby neurons are selectively responsive to different stimulus features, in which case the entire interpretation of the results would be different.

We thank the reviewer for their insightful assessment of the nature of context representations in our data. The reviewer's distinctions and considerations led to much more thorough discussion of the effects.

First, we agree that task goals could be considered as context (and also have been, e.g. see Polyn, Norman, & Kahana, 2009) and more generally, that they constitute an important component of context. To our mind real-world contexts define which stimulus features need to be attended to and what content-specific information should be retrieved for adequate decision-making (e.g. assessing a food's price in a supermarket as opposed to its taste at home). Our task operationalizes these characteristics of real-world contexts. Different questions determine both what features need to be attended to (or are salient as the reviewer rightly points out), but also which content-specific memories need to be retrieved and how this information is to be used to guide decision-making.

However, we also agree that it is valuable to consider the generality of context representations in light of the data and specifically, whether the effects could be explained by targeted salience signals. To our mind, implications differ between each of the described populations.

For example, context neuron activity could in theory reflect a focus-on size signal that represents attention towards size during picture presentations and would comprise a context signal in the narrow sense (even though it could still indicate size judgements in a broader network of context representations). However, since the questions at the beginning of the trial do not have any physical size (or price, etc.) such a saliency signal would not be expected to occur during question presentations. Importantly, context neurons already represent contexts during question presentations as evidenced by both previous as well as novel correlation and decoding analyses inspired by the reviewer (Fig. 3e, Fig. 5b Supplementary Figure 6a). While we agree that at least some context neurons could encode attention or saliency and now address this possibility in the discussion, a vast predominance of saliency representations would not be consistent with the reinstatement of question activity patterns during subsequent picture presentations (Fig. 3e, Fig. 5e), correlated context neuron activity between question and subsequent picture presentations (Fig. 5b, Supplementary Figure 6a,c), a stimulus-driven enhancement of context representations (Fig. 5f) or a persistence of context representations after picture offsets until a decision is made (Fig. 3f).

We also agree that task instructions might accentuate specific stimulus features or modulate attention dynamics, etc. when it comes to stimulus neurons. A subset of contextual stimulus neurons might therefore reflect such effects. However, this only underlines the importance of our finding that

most stimulus neurons were not modulated by context. Attention/saliency effects alone would not explain the primary modulation by either stimulus or context (Fig. 2a,e) or the distinct properties of context versus stimulus (i.e. content) neurons (see regional asymmetry in Fig. 4a as schematized in Fig. 5a).

We added the following paragraph to the discussion section to address these considerations:

> **"Operationalization of real-world contexts**
>
> Real-world contexts are associated with particular contents. Together they not only shape which stimulus features need to be attended to but also which memories are relevant for decision-making. For instance, in a supermarket (context), a discounted melon's (content) usual price (memory) is relevant for a buying decision. However, in a different context such as the kitchen at home, memories of the melon's typical appeal or taste are more relevant for deciding if it should be prepared and eaten. Our task operationalizes these characteristics of real-world contexts (e.g. see Polyn, Norman, & Kahana, 2009,[13] with different contexts for different word lists). Different questions comprise contexts that are to be associated with picture contents, and together not only guide attention, but also shape which content-specific memories are relevant for adequate decision-making in each respective context. While context-neuron activity could in principle at least partly reflect targeted saliency signals (e.g., a focus-on-size signal), our data better aligns with more general context representations. For example, contexts were already represented during question presentations at the beginning of the trial when a targeted salience signal would not be expected (as questions have no physical size, etc.; Fig. 5b; Supplementary Figure. 6a,c). Moreover, contexts were reinstated during subsequent picture presentations (Fig. 3e, Fig. 5e), they were represented more strongly after pre-activation (Fig. 5b, Supplementary Figure 6a,c) or when preferred versus non-preferred pictures of entorhinal stimulus neurons were shown (Fig. 5f), and they even persisted after picture presentations until a decision was made (Fig. 3f). Furthermore, attention or saliency effects alone would not explain the primary modulation by either stimulus or context (see next paragraph; Fig. 2a,e) or the distinct properties of context versus content neurons (e.g. see regional asymmetry in Fig. 4a as schematized in Fig. 5a).
>
> **Context representations across MTL regions**
>
> While the large majority of stimulus neurons (88%) were invariant to context, and most context neurons (63.5%) were invariant to stimulus, a small but significant fraction of stimulus neurons represented either both or specific conjunctions of context and stimulus, particularly in the hippocampus (see Fig. 2b-d). Overall, conjunctive representations of specific pictures and questions were most abundant in hippocampus (2.29%), particularly among stimulus neurons (10.34%). **It should be noted that the context invariance of most stimulus neurons (Fig. 2a,e) is particularly remarkable given that attention fluctuations among different question contexts could potentially lead to artificial context-modulation.**"

Moreover, related to this point, the authors then posit that these data therefore provide insight into memory formation and retrieval and how we link content and context. Yet there is no explicit test here of memory, or specifically how the specific item (picture) is linked to the specific context. It is hard to interpret the neural data as providing evidence regarding the link between context and content for memory if this is not actually what is happening in the task.

We thank the reviewer for raising this important point. While some contexts required subjective judgement, and therefore did not allow for objective evaluation (e.g. “Last seen in real life?”, “Like better?”), each context still required the assessment of a unique and distinct order among picture contents. Therefore, it was not possible to give consistent and transitive answers without both remembering the context (i.e. question) as well as the two subsequent pictures when making a decision. Furthermore, transitivity suggests that task instructions were actually adhered to. Following the reviewer’s important observation, we now assess memory of contents in context via the consistency and transitivity of the subject’s answers. Their performance greatly exceeded chance and approached ceiling levels in all but one recording session which we now exclude from all analyses. The following figure depicts performance per recording session:

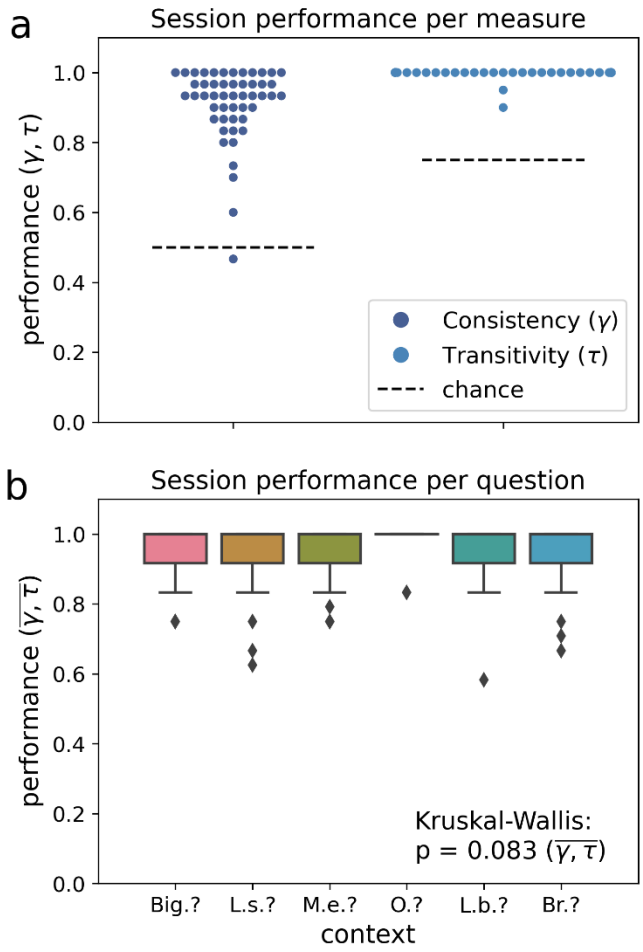


Fig S4. Consistency and transitivity of answers strongly exceeded chance and approached ceiling levels. Pairwise comparisons of contents were both consistent and transitive (a) across different contexts that imposed distinct orders (b), indicating memory of contents in context. Consistency measures how well context-specific preferences for one picture over another are maintained across presentation orders, while transitivity checks if a preference for A over B and B over C also results in a preference for A over C. **a.** Swarmplots depict consistency scores γ (dark blue, left) and transitivity scores τ (light blue, right) for each subject’s answers in each recording session. Dotted lines represent chance performance. Only one recording sessions did not exceed chance levels and was excluded from further analyses. **b.** Boxplots (Q1, median, Q3; whisker: points within ±1.5 IQR) of mean session-wise behavioral performances (averages of consistency and transitivity) across contexts. Performance did not differ significantly between contexts ($p = 0.083$; Kruskal-Wallis-Test).

We added the following sentence to the results section:

> "**Distinct populations encode stimulus and context in a picture comparison task**
> […] Subjects then chose the picture that best answered the question (e.g. which depicted something bigger) and indicated whether it was shown first or second by pressing keys 1 or 2 on the keyboard, respectively. **The majority of answers were highly consistent and transitive, with performance greatly exceeding chance in all but one excluded session and not varying across question contexts (p = 0.083; Kruskal-Wallis-Test; Supplementary Figure 4).** To quantify the representation of stimulus and context during picture presentations, repeated-measures two-way 4x5 ANOVAs with factors picture, question and their interaction were computed for each neuron at an alpha level of 0.001 (see Fig. 2e for distribution of effect sizes). "

The following paragraph was added to the Methods section:

> "**Behavioral performance**
> Memory of contents (i.e. pictures) in context (i.e. after a particular question) was evaluated by determining the consistency and transitivity of the answers that were provided at the end of the trial. Consistency (γ) captures the degree to which preference for one picture over another was maintained when the presentation order was reversed. A picture was considered to be preferred if it was chosen in at least 3 out of the 5 trials that it was paired with a specific other picture in a given temporal order and context. If this picture was preferred both when it was shown first or second (chosen ≥3/5 times in each order), all 10 trials were assigned a consistency value of 1 instead of 0. The overall consistency of a session was computed by averaging the consistency values of zeros and ones across all 300 trials. For random choices a consistency of 0.5 is expected (Supplementary Figure 4a).
> Transitivity (τ) captures the logical consistency of preferences across three pictures (x, y, z) when considered in pairs. For each of the four unordered picture triples we computed preference probabilities P(x>y), P(y>z), and P(x>z) for all three unordered picture pairs based on how many times out of the 10 trials one picture was chosen over the other in a given context. A triple was considered transitive if preference probabilities satisfied the condition of weak stochastic transitivity: P(x>y) ≥ 0.5 and P(y>z) ≥ 0.5, then P(x>z) ≥ 0.5. This condition was tested for all permutations of x, y, and z, and a transitivity value of 1 instead of 0 was assigned to the triple if the condition was satisfied for any permutation. For random choices a transitivity of 0.75 is expected (Supplementary Figure 4a). Finally, context-dependence of performance was assessed by deriving an overall behavioral performance measure from mean of consistency (γ) and transitivity (τ) values per question of each session and comparing them across contexts (Kruskal-Wallis-Test, alpha=0.05; Supplementary Figure 4b)."

Moreover, the following sentence was added to the discussion:

> "**Discussion**
> Contexts constrain which memories are relevant for future decisions[12]. Both semantic content as well as context, i.e. other aspects of a study episode[1], such as tasks[9], episodes[11] or rules[10] are represented in single-neuron activity in the human MTL. It is currently unclear, however, whether and how content and context are combined to form or retrieve integrated item-in-context memories at the single-neuron level in humans. Pairs of pictures were presented on a laptop screen in different contexts, namely five questions that specified how the pictures were to be compared to each other (Fig. 1a). **Sixteen of seventeen subjects remembered these contents (i.e. pictures) in their respective context as they chose both consistently and transitively between pictures despite unique and distinct orders among picture contents in each context (see Supplementary Figure 4).** […]"

A second major issue relates to the statistical analyses performed, and how data are pooled together. The main issue here is that it appears that many of the analyses are performed by pooling data across the entire population of neurons. This is especially true for the correlation analyses in figures 4 and 5, but also for the stimulus responses in figures 1 and 2. These data are collected from multiple different subjects, and from multiple sessions in each subject. At a minimum, the data should be analyzed at the level of sessions, but really at the level of subjects.

Even at the level of sessions, are different sessions really considered independent samples. What if all of the main significant effects derive from one subject (with several sessions). Would these results be generalizable. I think the issue of independent sessions could be explained by the possibility that different recording days could capture different neurons, but this should be explicitly shown.

In addition, it should also be shown that this does not artificially confound the data. In other words, the same analyses should be performed at the level of subjects, either averaging across sessions within a subject or simply taking one of each subject's session.

Analysis at the level of sessions seems to be performed for most of the analyses in Fig 3 for example. The issue of pooling is more problematic though when considering the analyses on correlations.

We thank the reviewer for inquiring about the reproducibility and generalizability of the findings, which is also highly important to us. Unfortunately, epilepsy patients with intracranial micro-wires are extremely rare, and in general it is not feasible to analyze human single-neuron data at the level of sessions, let alone at the level of subjects. This is particularly true for neural interaction analyses that require the concurrent recording of different sparse neural populations across brain regions in the same recording. Nevertheless, with the exception of some of the interaction analyses, we re-computed most figures and most subfigures at the level of subjects (by averaging across sessions within subjects). All of these novel analyses confirmed our previous findings. However, due to extremely small sample sizes, pairwise neural interaction analyses had to be performed at the level of sessions (N=6) and very few secondary analyses are still depicted level of neural pairs (see next response to the reviewer and figures further down).

Regarding the question of whether different sessions can be considered independent samples, it is worth mentioning that, across multiple recording sessions within the same subject, different pictures were presented in the vast majority of cases (96.33 % +/- 5.49 % SD; percent of unique to total pictures). Moreover, time intervals between subsequent sessions were typically long (46.94 hours ± 24.19 SD; range: 16.57 to 118.06 h), which allowed for the drifting of electrodes, as micro-wires are fixed to the skull while the brain is suspended in cerebrospinal fluid (CSF).

We explicitly quantified the resulting degree of neural independence between subsequent sessions in the same subject by computing intra-class correlation coefficients for the numbers of either context or stimulus neurons across channels (i.e. micro-wires). Both context and stimulus neurons showed little dependence across recording channels, as now depicted in Supplementary Figure S8 and further outlined in the methods section:

:

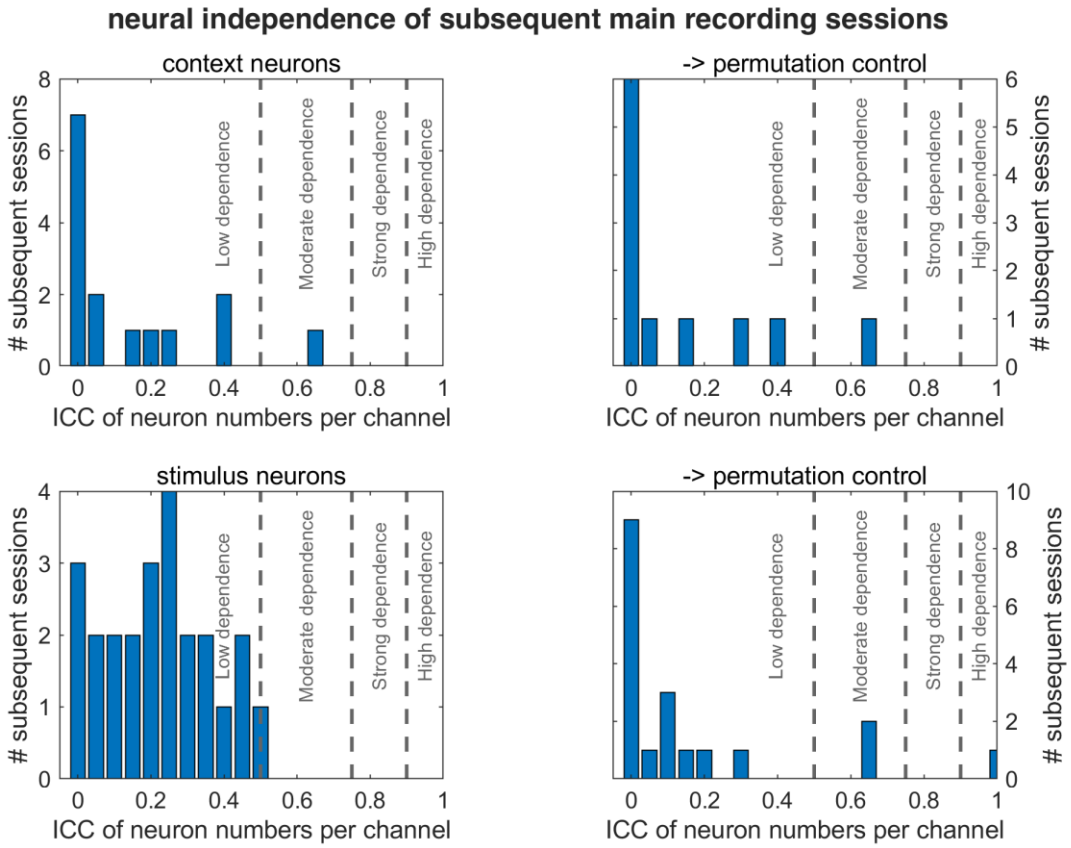


Fig. S8. Numbers of stimulus and context neurons across micro-wires are largely independent for subsequent main recording sessions within subjects. Bar plots of the distribution of intraclass correlation coefficients (ICCs) between either the numbers of either context (top left) or stimulus neurons (bottom left) across channels (i.e. micro-wires) for subsequent main recording sessions. Corresponding channel permutation controls are shown on the right. Vertical dashed lines in gray separate areas of differing effect sizes. Across the long time intervals between subsequent recording sessions (46.94 hours +/- 24.19 SD, range: 16.57 h to 118.06 h), numbers of both numbers context and stimulus neurons exhibited a low degree of dependence comparable to permutation controls.

"**Methods**

Subjects and Recording

[…] One session from one subject was excluded due to low behavioral performance, resulting in a final sample of 49 sessions from 16 subjects. […] **In 13 of the 16 subjects, more than one session was recorded. The mean time interval between multiple sessions was 46.94 hours +/- 24.19 SD (range: 16.57 h to 118.06 h). Neural dependence across 33 consecutive pairs of these sessions was assessed using intraclass correlation coefficients (ICCs(2,1); Shrout & Fleiss, 1979; McGraw & Wong, 1996). ICCs were computed across channels (i.e., micro-wires), separately for context and stimulus neurons. ICCs quantify cross-session stability, with lower values indicating weaker dependence. Observed ICCs were compared to null distributions obtained by permuting channels. ICCs did not strongly exceed these null distributions (Hedges' *g*: stimulus = 0.35; context = 0.16) and only reached significance (α = 0.01) in 6 of 33 stimulus pairs, 3 for context, and none for both, all with low overall effect sizes (see Supplementary Figure 8).**

As figure 1 only introduces the paradigm with examples of neural response types, the following pages depict figures 2 and 3 on a subject level (figures 4 and 5 with novel analyses and additional Supplementary Figures are depicted in later responses):

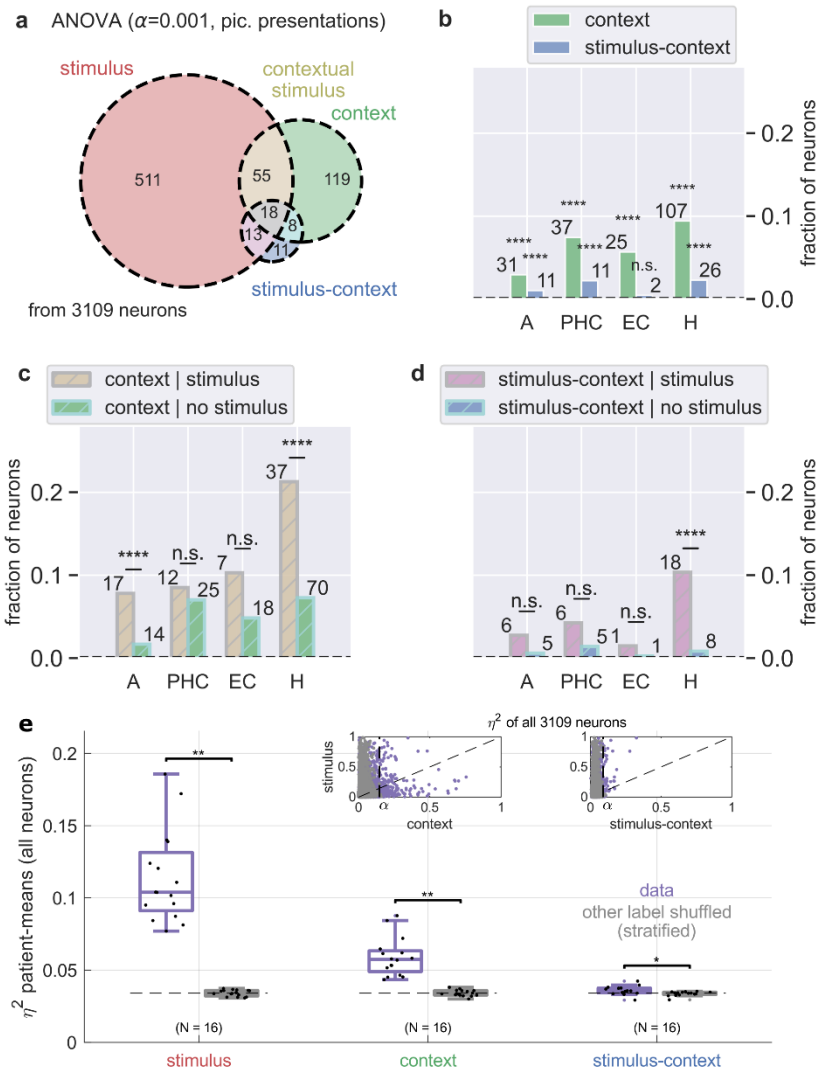


Fig. 2. Stimulus and context are mainly encoded separately and rarely conjunctively. a. Venn diagram of significant number of neuron types defined by repeated-measures two-way ANOVAs with factors stimulus identity (stimulus neurons, red/ocher/gray/pink), question (context neurons, green/ocher/gray/marine) or their interaction (stimulus-context neurons, gray/pink/blue/marine) during picture presentations at an alpha level of 0.001. Intersections denote multiple significances. While most stimulus neurons were not modulated by context, they overlapped with context neurons (contextual stimulus neurons) or stimulus-context neurons. **b.** Probabilities of obtaining context or stimulus-context neurons across brain regions (absolute numbers on top). Significance stars depict binomial tests relative to chance (dotted line). Neurons in all MTL regions were strongly modulated by context and stimulus-context. **c.** Analogous to **b** with probabilities of context conditional on stimulus modulation (stimulus / no stimulus). Significance stars depict results of Fisher exact tests computed from regional contingency tables. Stimulus neurons were more likely to be modulated by context (contextual stimulus neurons) than neurons without a significant main effect of stimulus in all MTL regions (significant in A and H), particularly in hippocampus. **d.** Analogous to **c**, with probabilities of significant interactions (stimulus-context). Stimulus-context interaction neurons with conjunctional representations were most frequent in hippocampus. **e.** Boxplots of subject-averages of repeated-measures two-way ANOVA effect sizes from all 3109 neurons ($\eta^2_{partial}$) for factors stimulus, context and interaction (stimulus-context). Data (lavender) are compared to controls (Wilcoxon signed-rank test) obtained via stratified shuffling of the relevant label (stimulus or context, gray). Effect sizes markedly exceeded chance for stimulus (η^2_p(real) = 0.117, η^2_p(control) = 0.034, $p_{subject}$ = 1.313×10^{-3}, $p_{session}$ = 3.330×10^{-9}) and context (η^2_p(real) = 0.059, η^2_p(control) = 0.034, $p_{subject}$ = 4.827 × 10^{-9}, $p_{session}$ = 1.087 × 10^{-9}). Difference for stimulus-context was smaller but significant (η^2_p(real) = 0.036, η^2_p(control) = 0.034, $p_{subject}$ = 1.840 × 10^{-2}, $p_{session}$ = 3.140 × 10^{-3}). Inlets depict scatters plots of $\eta^2_{partial}$ of context (left) or stimulus-context (right) plotted against those of stimulus. Dashed lines indicate the value of the smallest effect size below the alpha level of 0.001 (α). ***$p < 0.001$; **$p < 0.01$ (Bonferroni corrected).

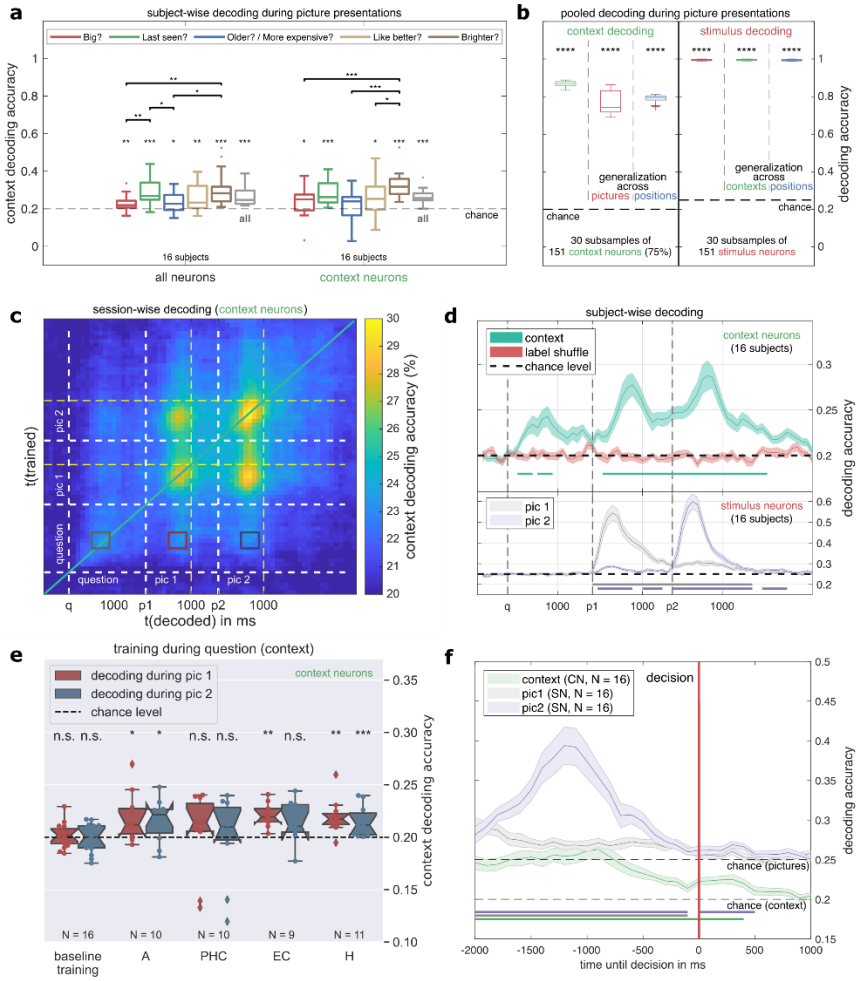


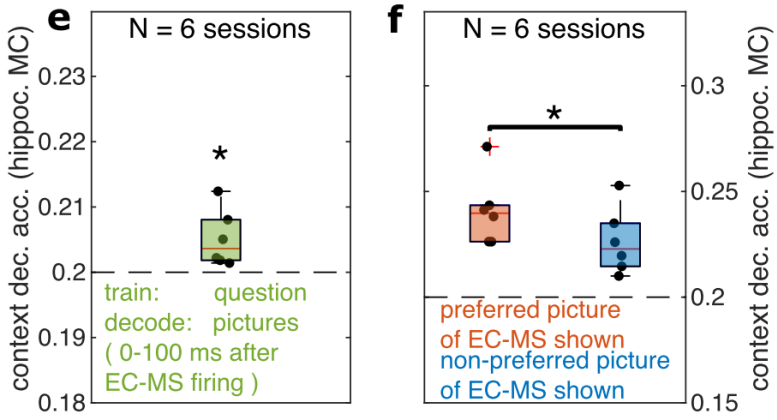
Fig. 3. Generality and temporal dynamics of context representations imply relevance for stimulus processing. Abstract representations of context (a-b) invariant to picture identity or position (b) are combined with those of content (b) across temporal gaps (c-e) both during picture presentations (d) via context reactivation (e) and before context-dependent decisions (f). **a.** Boxplots of subject-wise population-decoding accuracy of context (SVM: support vector machines) during picture presentations (100-1000 ms) for different questions (colors) and for all vs. context neurons. **b.** Boxplots of pooled context (green) or stimulus (red) decoding accuracies (SVM) from subsamples of context (left) and stimulus neurons (right), either obtained by 5-fold cross validation or generalization across pictures (red), contexts (green) and serial picture positions (blue). **c.** Heat map of subject-wise SVM context-decoding accuracies from context neuron activity at different time points. Diagonal line in green corresponds to identical training and decoding times (see **d**). Dashed lines denote onsets (white) and offsets (yellow) of trial events (question, picture 1, picture 2). Boxes indicate particular training (question) and decoding times (late picture presentations) further analyzed in **e**. **d.** Subject-wise decoding accuracies of context (top, context neurons, green) or stimulus (bottom, stimulus neurons, picture 1 and 2 in gray and lavender, respectively). Context was encoded above chance across the entire trial. During picture presentations, decoding accuracies of stimuli and contexts peaked (first stimulus, then context) and remained high in parallel. Solid lines and shaded areas depict means and standard errors, respectively. Data are compared to label-shuffled surrogate data (red) and chance (dashed line in black). Solid lines below indicate significant differences (cluster permutation test, $p < 0.01$). **e.** Boxplots of subject-wise context-decoding accuracies trained with context-neuron activity during question presentations (A: amygdala, PHC: parahippocampal cortex, EC: entorhinal cortex, H: hippocampus) or baseline (BL) and decoded during late first (red) or second (blue) picture presentations (see boxes in **c**). Question activity was recapitulated during first and second picture presentations, particularly in H (and EC), but not in PHC. Neuron numbers were matched across regions (N = 400). **f.** Same as **d**, but preceding context-dependent decisions. Both context and contents were represented above chance (dotted lines) until one picture was chosen over another at the end of the trial (red vertical line) according to the given context. For boxplots in **a,b,e** (Q1, median, Q3; whisker: points within +/- 1.5 IQR) stars denote significant differences from zero (Wilcoxon signed-rank test, uncorrected) or each other (Mann–Whitney *U* test, uncorrected). ****$p < 0.0001$; ***$p < 0.001$; **$p < 0.01$; *$p < 0.05$

The correlations between neuron pairs in Figure 4 are pooled across all possible pairs from all subjects. But again, how consistent is this across subjects.

Only six sessions from four subjects included simultaneous recordings of entorhinal stimulus and hippocampal context neurons, limiting the number of sessions and subjects in which interaction effects could be assessed. Nevertheless, shift-corrected cross correlations were replicated at the level of sessions (see Fig. 4 below) which could be considered statistically independent (see previous response). Moreover, three of the six sessions (50%) from two of the four subjects (50%) exhibited significant cross-correlation peaks around -40 ms (cluster permutation test, $p<0.05$).

Intracranial microwire recordings in humans are rare, and capturing distinct neuron types in different regions within the same session is even less common. Given the random placement of micro-wires, the chance of sampling identical cell types such as entorhinal stimulus neurons and hippocampal context neurons in homologous subregions across individuals (or even sessions) is inherently low. In this context, we consider our findings rare and valuable evidence for a previously uncharacterized mechanism linking contents to contexts in the human medial temporal lobe.

Importantly, interaction results were also supported and replicated by additional second-level analyses prompted by the reviewer (see also later responses for more details). First, SVM decoders trained with session-wise mere context neuron activity during questions can decode context during subsequent picture presentations time-locked to entorhinal stimulus neuron firing (Fig. 5e). Second, session-wise decoding accuracies of context with hippocampal mere context neurons are significantly higher when preferred instead of non-preferred pictures of corresponding entorhinal content neurons were shown (Fig. 5f):



e. Session-wise context SVM decoding accuracy of H- MC trained with question activity and decoded with subsequent picture activity (100-1500 ms) time-locked to E-MS firing (100 ms windows; 6 sessions: $p = 0.031$; 40 neuron pairs: $p = 5.794 \times 10^{-4}$; two-sided Wilcoxon signed-rank test). **f.** Session-wise context SVM decoding accuracies of H-MC distinguishing whether preferred or non-preferred pictures of the corresponding EC-MS were presented (6 sessions: p = 0.0156; 40 neuron pairs: p = 0.0338; one-sided Wilcoxon signed-rank test). *p < 0.05.

The results have been updated and we now explicitly mention the limited number of sessions and subjects for interaction analyses in Figure 4:

> **"Results**
>
> **[…]**
>
> **Sequential activation of entorhinal stimulus and hippocampal context neurons suggests a storage mechanism of stimulus-context associations**
>
> […] Only for pairs of neurons drawn from entorhinal cortex and hippocampus of the same hemisphere (Fig. 4a, Supplementary Figure 2), did the firing of mere stimulus neurons (entorhinal MS) predict firing of mere context neurons (hippocampal MC) after approximately 40 milliseconds as evidenced by asymmetric shift-corrected cross-correlograms (blue graph with peak on the left), but not that of other neurons (hippocampal ON, black graph without prominent peaks). Cross-correlograms differed significantly between both groups of entorhinal-hippocampus pairs (i.e. MS-MC vs. MS-ON) on short timescales (-131 to -17 ms for $p < 0.05$; **two-sided cluster permutation test on a session-level, solid lines in red, Fig. 4a**), but not for pairs drawn from different regions (Supplementary Figure 2) or across hemispheres (Supplementary Figure 3). Findings were therefore consistent with unidirectional synaptic changes reflecting stimulus-context associations. **Overall, only six sessions from four subjects included simultaneous recordings of entorhinal stimulus and hippocampal context neurons. Three of the six sessions (50%) from two of the four subjects (50%) exhibited significant cross-correlation peaks around 40 milliseconds (cluster permutation test, $p<0.05$).**"

Figure 4 is depicted on the following page at a session level with secondary analyses performed at the level of neural pairs:

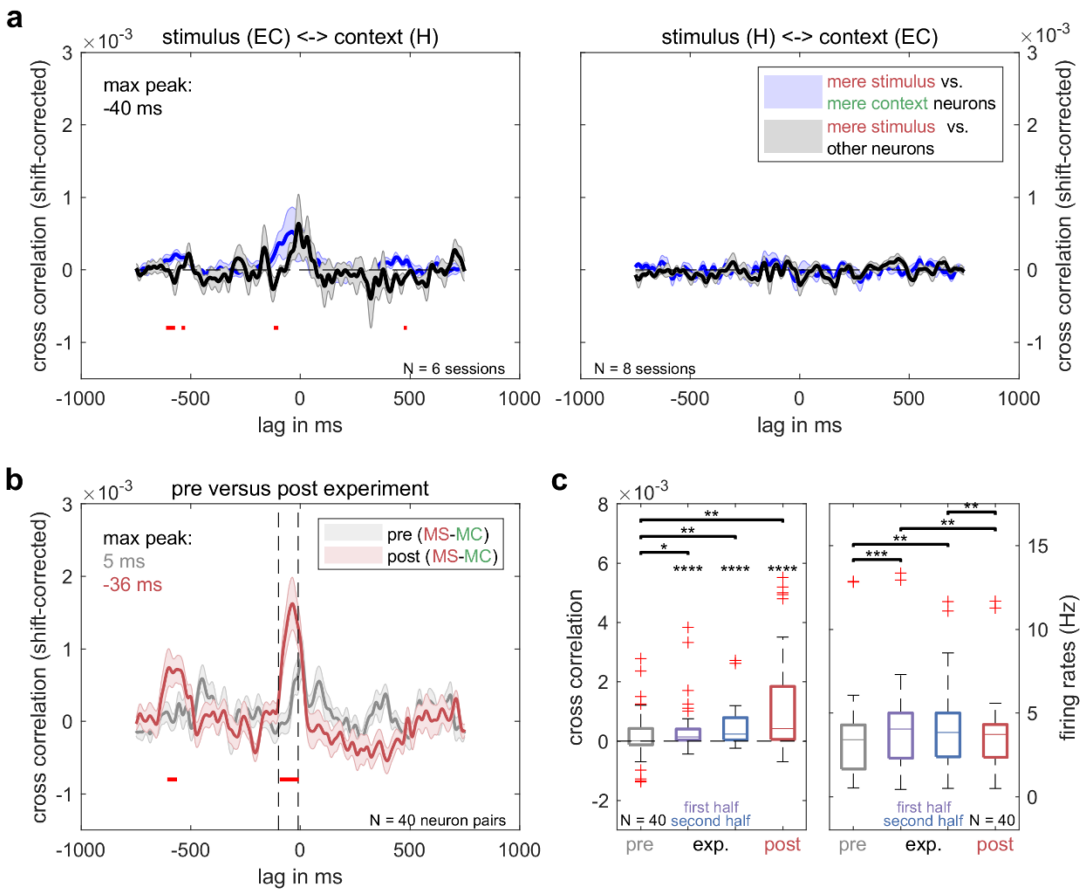


Fig 4. Sequential activation of entorhinal stimulus and hippocampal context neurons suggests a storage mechanism of stimulus-context associations. a. Left: Session means of trial-by-trial cross-correlograms during picture presentations (shift-corrected, see Methods) either between pairs of entorhinal mere stimulus and hippocampal mere context neurons (MS or MC with only one of two factors $p < 0.001$ in ANOVA, blue) or a matched number of entorhinal mere stimulus and other neurons in hippocampus without significant main effects (ON, both factors and interaction in ANOVA $p > 0.1$, black). Data are presented as mean values ± SEM of session means (solid lines and shaded areas). Red horizontal lines indicate significant differences ($p < 0.05$, two-sided cluster permutation test) between pair types (blue vs. black). Maximum peak lag time is depicted in upper left corner. The two groups (i.e. MS-MC vs. MS-ON) differed significantly on small time scales (-30 to -40 ms). Firing of stimulus neurons in EC predicted firing of context neurons in H (but not of other neurons in H) after approximately 30 to 40 milliseconds, potentially reflecting a storage mechanism of stimulus-context associations for the re-instatement of context during stimulus presentations. Right: Same as left for neuron pairs from same brain regions in opposite order. Here, activity of hippocampal stimulus neurons did not predict entorhinal context neuron firing. **b.** Shift-corrected cross correlations of all 40 entorhinal mere stimulus and hippocampal mere context neuron pairs (see **a**, left, blue), but here either before (pre, gray) or after the experiment (post, red). Asymmetric cross correlations at approximately -40 ms were not present before the experiment, but emerged during its course (see **a,c**) and persisted after its termination ($p < 0.01$, two-sided cluster permutation test). **c.** Emergence of interactions during the experiment. Left: Boxplots of mean cross correlations (shift-corrected) between same neuron pairs as in **b** from -10 to 100 ms (see vertical dashed lines). Mean cross correlations did not differ significantly from zero (two-sided Wilcoxon signed-rank test) before the experiment ($p = 0.29$, pre in gray), but exceeded chance levels (all $p < 0.001$) both during the first (exp., violet) and second (exp., blue) half of the experiment as well as after its termination (post, red). Cross correlation sizes of all subsequent time periods were significantly larger than before the experiment (p(pre, exp1) = 3.258×10^{-2}; p(pre, exp2) = 6.625×10^{-3}; p(pre, post) = 3.106×10^{-3}. Right: In contrast, mean firing rates did not differ significantly before and after the experiment, but were significantly lower ($p < 0.01$) compared to the first and second half of the experiment. ****$p < 0.0001$; ***$p < 0.001$; **$p < 0.01$ (uncorrected).

The correlations of Figure 5b, c are also pooling across all neurons (and possibly all trials separately, although the number of samples here is not clear), which would have the effect of generating extremely low p value just due to the large number of samples. Again, these correlations should be performed within each subject/session and then analyzed across subjects/sessions.

We thank the reviewer for this inquiry as the results section and the figure caption were not entirely clear and have been updated. Data points were obtained by averaging the normalized activity of each of the 127 mere context neurons during the respective trial phase for each of the 5 questions: 5 questions x 127 neurons resulting in 635 data points. The data is now depicted per subject and the caption has been adapted to increase clarity:

> **"b.** Scatter plot of mean z-values for each of the 5 contextual questions (baseline normalization) during question versus subsequent picture presentations computed for each mere context neuron and averaged per subject (N = 5 × 16)."

We still report the original neuron-wise analysis (see Supplementary Figure S6a,b) and furthermore computed Pearson correlation effect sizes for each mere context neuron from activity between question and subsequent picture presentations or between baseline and subsequent picture presentations (see Supplementary Figure S6c below).

The results section now reads as follows:

> **"Pre-activation of context neurons by preferred questions predicts subsequent stimulus-neuron interaction and context-neuron reactivation**
> Our final analyses addressed whether a pre-activation of context neurons by preferred contexts (i.e. preferred questions) could modulate their activity during subsequent picture presentations as well as their response likelihood following the activation of stimulus neurons. First, we assessed whether normalized activity (baseline normalization, see Methods) of mere context neurons during question presentations (Fig. 5b as opposed to baseline in Fig. 5c) is predictive of activity during subsequent picture presentations. **Subject means of question (r = 0.46, p = 2.07×10^{-5})**, but not baseline **activity (r = 0.03, p = 7.79×10^{-1})** significantly (p < 0.01, Pearson correlation) predicted subsequent responses of mere context neurons to picture presentations (Fig. 5b-c). **The same results were obtained for all mere context neurons (Supplementary Figure S6a,b). Similarly, question activity of mere context neurons predicted subsequent picture activity significantly more strongly than baseline activity on a trial by trial basis (Supplementary Figure S6c). Specifically, subject means of Pearson correlation effect sizes were significantly higher for correlations between question and picture activity ($\bar{r}$ = 0.12, left in green) than between baseline and picture activity ($\bar{r}$ = 0.03, right in gray; p = 6.13× 10^{-3}; two-sided Wilcoxon signed-rank test).**

The method section also has been updated accordingly:

> **"Interaction analyses between stimulus and context neurons**
> **[…]**
> **Next,** Pearson correlations between **(subject means of)** the mean question-wise activity of mere context neurons during question presentations (500 to 1300 ms, Fig. 5b**, Supplementary Figure 6a**) or its baseline (-700 to 100 ms, Fig. 5c**, Supplementary Figure 6b**) versus subsequent picture presentations (500 to 1300 ms) were computed ($p < 0.01$), normalized to baseline and visualized as scatter plots with regression lines (Fig. 5b-c**, Supplementary Figure 6a-b). Finally, as a proxy for the excitability of context neurons Pearson correlation effect sizes were computed for each mere context neuron from activity between question and subsequent picture presentations or between baseline and subsequent picture presentations across all trial conditions (60 combinations of 5 questions and 12 picture pairs) and averaged per subject. Afterwards effect sizes of both conditions were compared and visualized as boxplots (two-sided Wilcoxon signed-rank test; Supplementary Figure 6c).**"

The updated Figure 5 and Supplementary Figure 6 are depicted on the next pages:

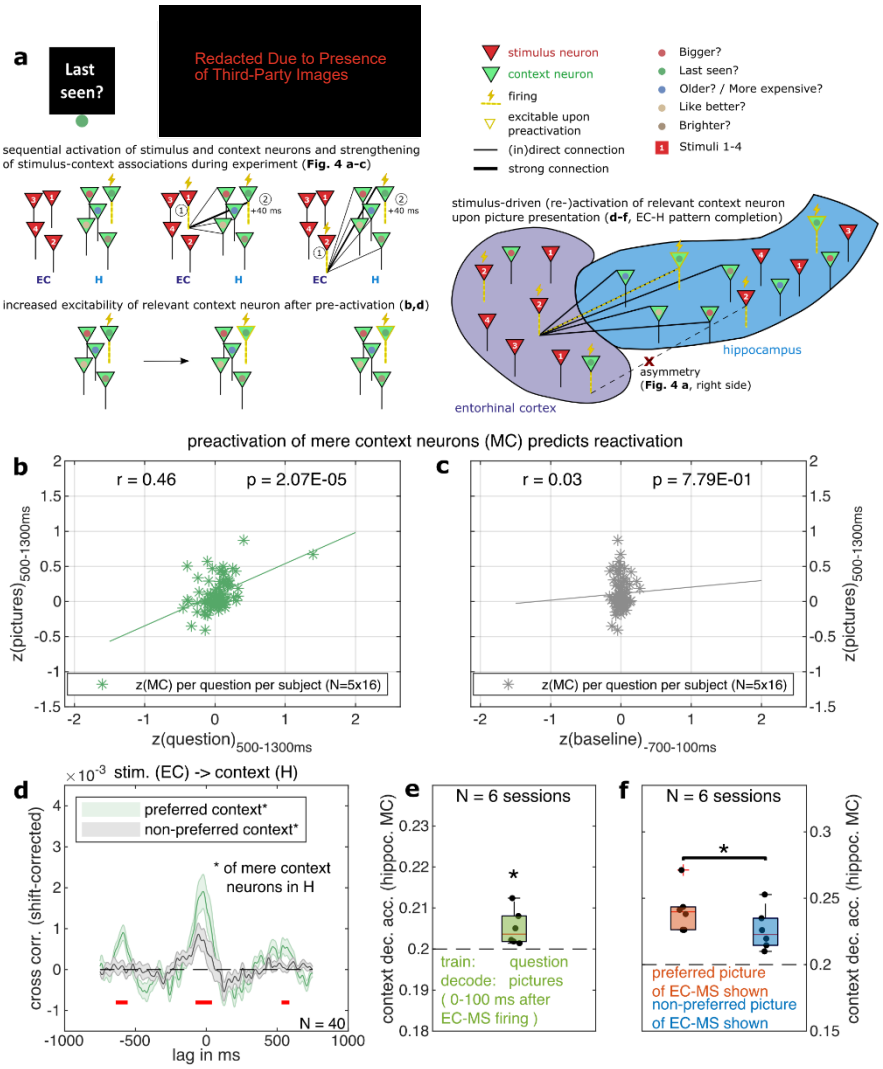


Fig 5. Sequential activation after pre-activation of context neurons by their preferred context could guide stimulus-driven pattern completion. a. Schematic model of stimulus-context storage. Upper left: Activation patterns of entorhinal (EC) mere stimulus (red) and hippocampal (H) mere context neurons (green) during different trial events. Specific context neurons in H respond to particular questions (see inner colored circles) and are reactivated during subsequent picture presentations. Picture presentations elicit responses in selected stimulus neurons (see inner numbers), whose activity predicts firing of context neurons after 40 ms (see Figure 4), resulting in strengthening of connections that could encode stimulus-contexts associations. Lower left: Question response strengths predict activity and excitability (yellow triangles) of context neurons during subsequent picture presentations (see **b,d**, and Supplementary Figure 6a-c). Right: After connections have been strengthened, pre-activation by preferred contexts can selectively guide stimulus-driven pattern completion in hippocampal context neurons (dashed line in yellow). **b.** Scatter plot of mean z-values for each of the 5 contextual questions (baseline normalization) during question versus subsequent picture presentations computed for each mere context neuron and averaged per subject (N = 5 × 16). Pearson correlation strengths and *p*-values (uncorrected) are shown at the top and visualized by regression lines. Question response strengths significantly predict responses to subsequent pictures ($r = 0.46$, $p = 2.07 \times 10^{-5}$). **c.** Same as **b**, but comparing baseline to picture presentations**.** Baseline response strengths do not predict subsequent picture responses ($r = 0.03$, $p = 0.78$). **d.** Shift-corrected cross correlations of entorhinal mere stimulus (EC-MS) and hippocampal mere context neurons (H-MC) during picture presentations after a preferred (max. response, green) versus a non-preferred context (gray) was presented. Red horizontal lines indicate significant differences ($p < 0.01$, two-sided cluster permutation test). Firing of MS predicts MC firing significantly more strongly in preferred contexts. **e.** Session-wise context SVM decoding accuracy of H-MC trained with question activity and decoded with subsequent picture activity (100-1500 ms) time-locked to E-MS firing (100 ms windows; 6 sessions: $p = 0.031$; 40 neuron pairs: $p = 5.794 \times 10^{-4}$; two-sided Wilcoxon signed-rank test). **f.** Session-wise context SVM decoding accuracies of H-MC distinguishing whether preferred or non-preferred pictures of the corresponding EC-MS were presented (6 sessions: $p = 0.0156$; 40 neuron pairs: $p = 0.0338$; one-sided Wilcoxon signed-rank test). *$p < 0.05$.

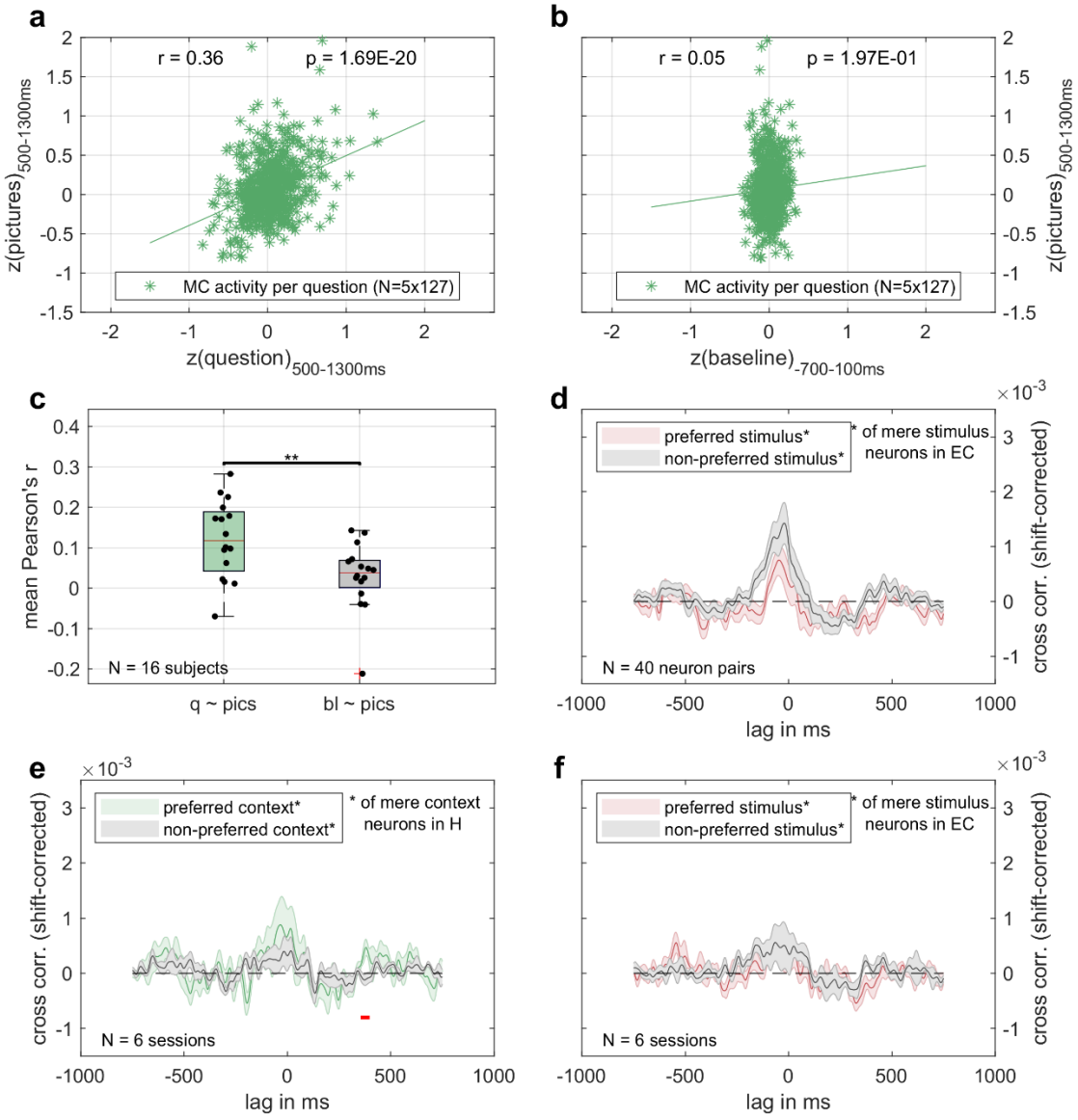


Fig. S6. Sequential activation after pre-activation of context neurons by their preferred context could guide stimulus-driven pattern completion. a. Scatter plot of mean z-values for each of the 5 contextual questions (baseline normalization) during question versus subsequent picture presentations computed for each of the 127 mere context neurons (N = 5 × 127). Pearson correlation strengths and *p* values (uncorrected) are shown at the top and visualized by regression lines. Question response strengths significantly predict responses to subsequent pictures ($r = 0.36$, $p = 1.68 \times 10^{-20}$). **b.** Same as **a**, but comparing baseline to subsequent picture presentations**.** Baseline response strengths do not predict subsequent picture responses ($r = 0.05$, $p = 0.20$). **c.** Subject means of Pearson correlation effect sizes computed from mere context neuron activity between question and subsequent picture presentations across trial conditions. Activity between question and subsequent picture presentations ($\bar{r} = 0.12$, left in green) was correlated significantly more strongly than between baseline and picture activity ($\bar{r} = 0.03$, right in gray; $p = 6.13 \times 10^{-3}$; two-sided Wilcoxon signed-rank test). **d.** Shift-corrected cross correlations of entorhinal mere stimulus and hippocampal mere context neurons during picture presentations depending on whether a preferred stimulus (red, maximum response) versus a non-preferred stimulus (gray, randomly drawn different stimuli) was presented. Red horizontal lines indicate significant differences ($p < 0.01$, green vs gray, two-sided cluster permutation test). Firing of stimulus neurons predicts that of context neurons significantly more strongly in preferred versus non-preferred contexts. **e.** Same as **c,** but depending on whether a preferred context (green, maximum response) versus a non-preferred context (gray, randomly drawn different contexts) was shown. Picture preference of stimulus neurons did not affect cross-correlations. **f.** Same as **d**, but averaged per session.

There are several other points of confusion related to the statistics here:

We apologize for these points of confusion, thank the reviewer for raising them, and have tried our best to improve clarity as detailed below.

• In some cases, the authors examine 38 sessions, while in others they examine 50. In the methods, they state they analyze 38 sessions, but in results they state they analyze 50 sessions. It looks like 38 sessions are analyzed for context neurons but 50 for stimulus neurons (Fig 3a). Why?

As the reviewer rightly observed, the Method section falsely stated that the dataset comprised 38 sessions. This has now been amended. Furthermore, only 38 sessions contained context neurons and therefore analyses pertaining to context neurons only included those sessions. The Methods and Statistics and Reproducibility sections have been updated as follows:

> **"Methods**
> **Subjects and Recording**
> […] Our original data set consisted of 50 experimental sessions from 17 subjects with recordings from the amygdala (A), parahippocampal cortex (PHC), entorhinal cortex (EC), and hippocampus (H). One session from one subject was excluded due to low behavioral performance, resulting in a final sample of 49 sessions from 16 subjects. Remarkably, the excluded session stemmed from the only subject for which no context neurons could be detected (others exhibited an average of 12.5). […]
>
> **Statistics and Reproducibility**
> 49 sessions from 16 subjects were analyzed (one session of one subject was excluded based on a low behavioral performance). A total of 200 context neurons were detected in 38 sessions from the 16 subjects. 25 sessions contained at least three stimulus and context neurons each, 12 sessions more than five.**"**

Or, for example, in the analysis comparing the ANOVA to stratified label shuffling controls, which confirms the main effect, they report N=50 for both stimulus and context.

This is because effect sizes of all neurons were considered. Not just of context neurons, which were only present in 38 sessions. We now state this more clearly in the results section:

> **"Stimulus and context are mainly encoded separately and rarely conjunctively**
> […] In order to account for potential statistical dependencies between factors or neurons, ANOVA effect sizes of all neurons were compared to stratified label-shuffling controls, irrespective of significance (Fig. 2e, see inlets with all 3109 recorded neurons)."

• In some cases, the number of neurons that are being analyzed are unclear. For example, in Figure 3f, they report N=400 neurons. Which neurons are these? For the classifier analysis of the pooled data, they analyzed 30 subsamples of 150 neurons. But there are less than 150 mere context neurons. So are they also including contextual stimulus and stimulus-context neurons here.

We thank the reviewer for pointing out ambiguities with respect to the number and identity of analyzed neurons. N=400 in Fig. 3f (now Fig. 3e) refers to 400 neurons randomly drawn from all recorded neurons of each respective brain region (irrespective of their response behavior). This number was chosen because at least 400 neurons were recorded in each brain region (A=1051, PHC=482, EC=440, H=1135).

Pooled decoding analyses were performed with “context neurons”. By context neurons we meant to refer to the 200 neurons that (at least) exhibited a significant main effect of question. These contain the 127 mere context neurons that only exhibited a significant main effect of question. Only the 200 “context neurons”, were included in pooled decoding analyses (neither contextual stimulus, nor stimulus-context neurons).

In the methods section we now clarified ambiguities regarding the definition of context neurons:

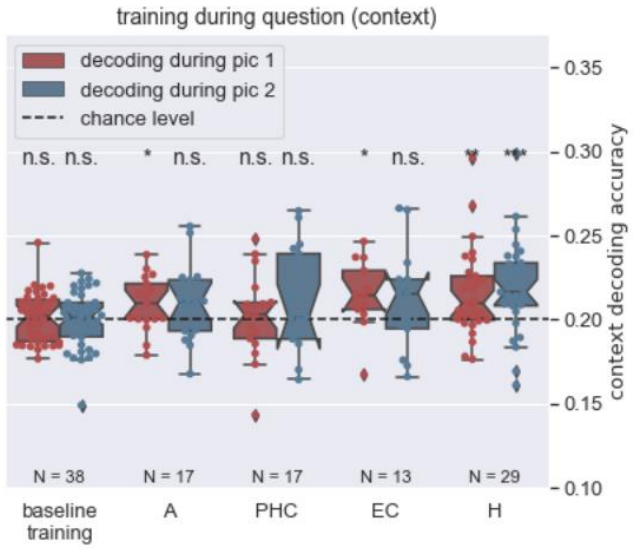
> **“Definition of neural populations and assessment of their significance**
>
> For each unit, repeated two-way repeated-measures ANOVAs with factors stimulus (picture), context (question) and interactions were computed from the activity during picture presentations (100-1000 ms) at an alpha level of 0.001. **Units with (at least) a significant main effect of stimulus or context are referred to as stimulus or context neurons, respectively. If a unit only (at most) showed a significant main effect of either stimulus or context it is termed a mere stimulus or a mere context neuron, and if both main effects were significant it is called a contextual stimulus neuron.** […]**”**

Furthermore, we added a clarification regarding which neurons are analyzed in Fig. 3f (now Fig. 3e) to the results section:

> **“Generality and temporal dynamics of context representations imply relevance for stimulus processing**
>
> […] Lastly, we determined whether activity patterns associated with the presentations of questions were recapitulated during picture presentations (Fig. 3f). For this purpose, session-wise SVM decoders were trained with regional context neuron activity either during question presentations (600-1000 ms) or preceding baseline intervals (-500-100 ms) while contexts were decoded in the late phases of first (red; 600-1000 ms) or second (blue; 600-1000 ms) picture presentations. Training and decoding intervals correspond to colored boxes in Figure 3c. **In order to account for regional differences in the number of recorded neurons (A=1051, PHC=482, EC=440, H=1136), 400 neurons were randomly drawn from each brain region for decoding analyses.** […]”

• Timing. The selection of timing windows for analyses seems to differ between different analyses, but it’s not clear how those time windows are chosen. For example, they find that training the decoder during question presentation leads to good decoding during picture presentation. In both cases, they only look at latter half of activity (600-1000ms). Why? Is this derived from the heatmap in Figure 3c? If so, then for these heatmaps, we should be able to identify time regions of significance (across sessions and/or subjects) and use those time regions for subsequent analyses.

We thank the reviewer for inquiring about the varying time windows. 400 ms bins were chosen in this case to ensure comparability with a short baseline interval between -500 to -100 ms and we chose the particular time windows of 600-1000 ms in order to capture activity at a time when MTL neurons usually respond to written text and become reactivated. As context effects of the heat map are significant almost across the whole trial, we kept the original windows but re-analyzed the data in 100-1000 ms time windows relative to question and picture onsets, which reproduced the findings with strongest and most consistent effects in hippocampus (see below):



Or, for example, for the cross-correlation analyses, why look at large window of 1500 ms if the picture is only on screen for 1000 ms (plus 200-300 ISI plus 200 ms orientation).

This window was chosen in order to compute meaningful correlations for lags between -750 ms to +750 ms.

Or, for example, for the analysis looking at the correlation of firing between question and picture, looked at a different window of 500-1300.

In this case a time windows of 500-1300 ms was chosen because single neuron reactivation responses tend to occur later than original responses to stimuli. However, we re-computed the analysis with different time windows, which consistently led to even stronger effect sizes:

500-1300 ms (original analysis):
z(question) vs z(pictures):
r = 0.26, p = 5.91 x 10^-04

z(baseline) vs z(pictures):
r = 0.04, p = 0.592

Here are the results for alternative windows on a subject level for which the effects are markedly stronger:

100-1000ms: z(question) vs z(pictures): r = 0.52, p = 5.33 x 10^{-13} z(baseline) vs z(pictures): r = 0.03, p = 0.724	100-1500ms: z(question) vs z(pictures): r = 0.83, p = 2.70 x 10^{-43} z(baseline) vs z(pictures): r = 0.02, p = 0.762	500-1500ms: z(question) vs z(pictures): r = 0.53, p = 3.54 x 10^-13 z(baseline) vs z(pictures): r = 0.04, p = 0.630

Another concern relates to the overall conclusion of the manuscript and the model presented in figure 5. This model largely rests upon the primary result that there are increased correlations between EC stimulus neurons and hippocampal context neurons, with a time delay of tens of

millisecond. Thus, this is taken as the primary evidence that there is sequential activity between stimulus response and context response. In addition, the authors observe elevated correlations between EC and H neurons during the experiment and after the experiment, but not prior to the experiment, which could suggest that STDP has increased the link between these sets of neurons. This is a nice result, but there are some potential concerns with drawing these firm conclusions.

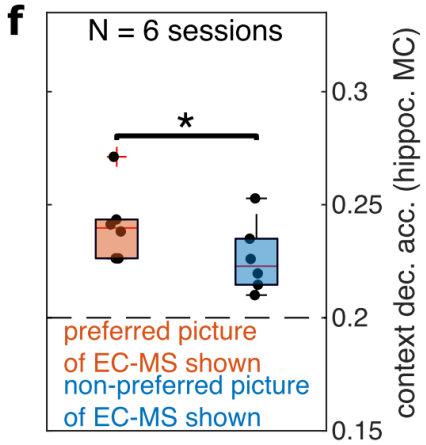
First, there are the issues related to statistical testing discussed above, especially with regard to the pooled correlations in figure 5.

Correlations are now shown per subject (Fig. 5b) and are also supported by additional analyses (see Figure 5 and Supplementary Figure 6 above).

Second, these data demonstrate that there is a single pairwise correlation that is significant (among all pairwise correlations). But in addition, what one would like to see if that these relations between EC stimulus and HC context cells are specific to individual stimuli and individual trials. Are these analyses performed just during the responsive trials, or is this a generic relation regardless of whether the neurons actually respond.

We thank the reviewer for this relevant question. The analyses in the manuscript were performed across all trials. To our mind, the amplitude of shift-corrected cross correlations peaks only reflects the degree to which a neuron A increases the likelihood of another neuron B to fire at specific time lags, regardless of how active neuron A actually is. We interpret this finding as reflecting changes in (indirect) synaptic conductivity which can be detected after multiple pairings of all contents with all contexts, regardless of whether entorhinal stimulus neurons actually responded. Therefore, it should not surprise that shift-corrected cross correlation peaks between entorhinal mere stimulus and hippocampal mere context neurons do not distinguish whether a “preferred” picture of the entorhinal stimulus neurons was depicted.

However, based on the reviewer’s valuable suggestion we added novel decoding analyses that confirm the relevance of responsive trials since context-decoding accuracies of hippocampal mere context neurons were significantly higher when preferred versus non-preferred pictures of the paired entorhinal stimulus neurons were presented. This is now shown in Fig.5f:



f. Session-wise context SVM decoding accuracies of H-MC distinguishing whether preferred or non-preferred pictures of the corresponding EC-MS were presented (6 sessions: $p = 0.0156$; 40 neuron pairs: $p = 0.0338$; one-sided Wilcoxon signed-rank test). *$p < 0.05$ (uncorrected).

The following paragraphs have been added to the results, methods, and discussion sections:

"**Results**

[…]

Pre-activation of context neurons by preferred questions predicts subsequent stimulus-neuron interaction and context neuron reactivation.

[…]

Second, we computed session-wise context SVM decoding accuracies of hippocampal mere context neurons for trials during which the preferred versus non-preferred pictures of the corresponding entorhinal stimulus neuron were presented. Hippocampal mere context neurons represented context more strongly when preferred (0.241) versus non-preferred pictures (0.226) of entorhinal stimulus neurons were presented (Fig. 5f; 6 sessions: $p = 0.016$; 40 neuron pairs: $p = 0.034$; one-sided Wilcoxon signed-rank test). […]

Methods

[…]

Interaction analyses between stimulus and context neurons

[…]

Second, decoders were trained with hippocampal mere context neuron activity per session for trials during which the preferred versus non-preferred pictures of the corresponding entorhinal stimulus neuron were presented. […]

Discussion

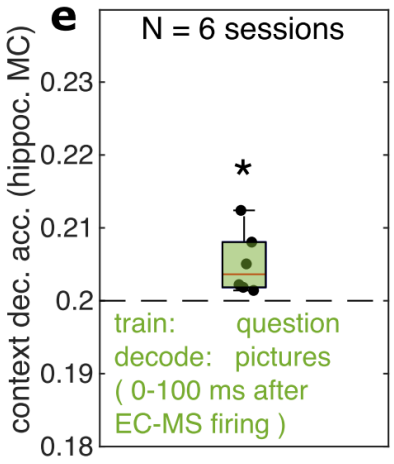
[…] Furthermore, increased excitability of context neurons after pre-activation by their preferred context suggested a gating mechanism of stimulus-driven pattern completion in the hippocampus (Fig. 5a,d) **that resulted in enhanced context representations (Fig. 5f).**

[…]

Moreover, contexts were reinstated during subsequent picture presentations (Fig. 3e, Fig. 5e), they were represented more strongly after pre-activation (Fig. 5b, Supplementary Figure 6a,c) **or when preferred versus non-preferred pictures of entorhinal stimulus neurons were shown (Fig. 5f)** and they even persisted after picture presentations until a decision was made (Fig. 3f).

With respect to the model, the authors posit that context activity is reinstated once the EC neurons are activated. This could be explicitly shown by examining decoding of the context representations, based on training during the question period, time locked to EC neuronal firing.

We thank the reviewer for this great suggestion. In a novel analysis, context decoders were trained with hippocampal mere context neuron activity during question presentations (100-1500 ms) and decoded with their average activity during subsequent picture presentations (100-1500 ms) time-locked in 0-100 ms windows to entorhinal stimulus neurons firing. These analyses are now part of Figure 5e:



e. Session-wise context SVM decoding accuracy of H-MC trained with question activity and decoded with subsequent picture activity (100-1500 ms) time-locked to E-MS firing (100 ms windows; 6 sessions: $p = 0.031$; 40 neuron pairs: $p = 5.794 \times 10^{-4}$; two-sided Wilcoxon signed-rank test).

The following paragraphs have been added to the results and methods:

> "**Results**
> […]
> **Pre-activation of context neurons by preferred questions predicts subsequent stimulus-neuron interaction and context neuron reactivation**.
> […] For further validation of this model, we performed additional SVM decoding analyses regarding the role of entorhinal stimulus and hippocampal context neuron interactions in stimulus-driven reinstatement of context. First, we computed session-wise context SVM decoding accuracies of hippocampal mere context neurons trained with question activity and decoded with subsequent picture activity (100-1500 ms) time-locked to entorhinal stimulus neurons firing (100 ms windows). Decoding accuracies slightly, but significantly, exceeded chance (0.21; Fig. 5e; 6 sessions: $p = 0.03$; 40 neuron pairs: $p = 5.79 \times 10\text{-}4$; two-sided Wilcoxon signed-rank test).
>
> **Methods**
> […]
> **Interaction analyses between stimulus and context neurons**
> […]
> As a further validation of this model, two additional SVM interaction decoding analyses were performed. First, decoders were trained with hippocampal mere context neuron activity per session during question presentations and decoded with their activity during subsequent picture presentations (100-1500 ms) time-locked to entorhinal mere stimulus neurons firing (in 100 ms windows after each EC-MS action potential)."

Moreover, the novel results are now referenced in the discussion, e.g.:

> **"Discussion**
> […] Overall, our results imply that stimulus and context neurons contribute to the context-dependent processing of stimuli via selective reinstatement of context during picture presentations (Fig. 3e**, Fig. 5e,f**)."

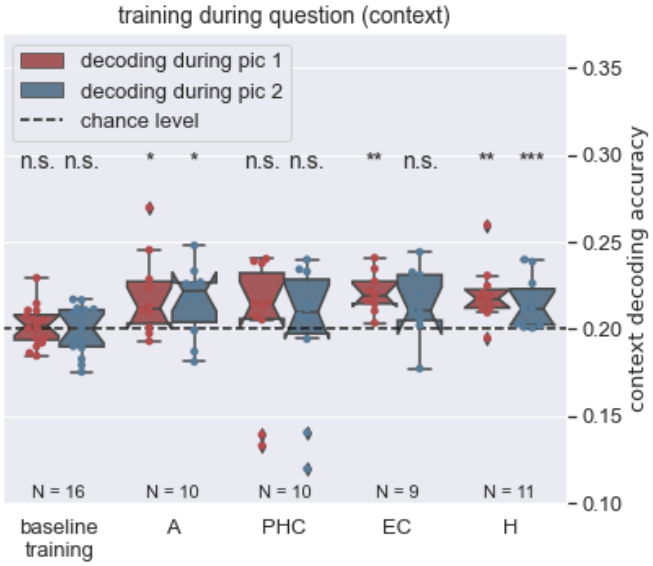
In addition, related to this point, the reactivation of context signal is much more clearly demonstrated with a decoding classifier approach rather than simply examining the correlations in firing rates between stages of the experiment (e.g., question to picture). Overall, this is a potentially compelling model, but there are some analyses that could help strengthen this claim.

In addition to our SVM context reinstatement analysis (now Fig.3e), the valuable decoding classifier suggestions of the reviewer (Fig. 5e,f) were of great assistance for strengthening the model.

Interestingly, related to this point, the authors also demonstrate that classifiers can be trained on time periods during the question, or during pictures, and used to decode context during other stages of the task. For the hypothesis that the link between context and content is made explicitly between EC stimulus neurons and hippocampal context neurons, one would like to then know why context signals can in fact be decoded in multiple brain regions throughout the task, rather than just in the hippocampus, and whether these context representations are therefore irrelevant (Fig. 3e).

Based on the literature, we assume that there are multiple representations of context. Context information is likely maintained within the frontal cortex and sent to multiple MTL regions before converging in the hippocampus. Repeated pairings of pictures and questions and potentially resulting synaptic changes could then aid in the retrieval and reinstatement of previously presented contexts across brain regions within the MTL.

Moreover, it should also be noted, that effect sizes were strongest in (EC and) H, particularly on a subject level:



e. Boxplots of subject-wise context-decoding accuracies trained with context-neuron activity during question presentations (A: amygdala, PHC: parahippocampal cortex, EC: entorhinal cortex, H: hippocampus) or baseline (BL) and decoded during late first (red) or second (blue) picture presentations (see boxes in **c**). **Question activity was recapitulated during first and second picture presentations, particularly in H (and EC), but not in PHC.** Neuron numbers were matched across regions (N = 400).

In addition, if these are mere context neurons, then we should expect that decoding accuracy should be elevated during the actual question as compared to the picture presentation. The question is, indeed, in this case taken as the context signal, so presumably this should be the time point of greatest context signal. This goes back to the question of whether what is being shown here really is context, or some other stimulus feature. But in this case, decoding during the actual question appears significantly worse than during other periods of the task like during picture presentation. Why?

We thank the reviewer for their careful consideration of the encoding dynamics. Studies that analyze reactivation usually screen for a representation and then analyze whether this representation is re-instantiated at some later time point. Here we screened for "context neurons" during picture presentations, at a time when we assume the reinstatement of context to occur. The fact that context decoding peaks are highest during picture presentations at least partially results from our definition of context neurons as neurons with a significant main effect of question during **picture presentations**. If we instead define context neurons as neurons with a significant main effect of question during **question presentations**, context representations are strongest during question presentations with subsequent lower but statistically significant reactivation peaks during picture representations as the reviewer rightly expected.

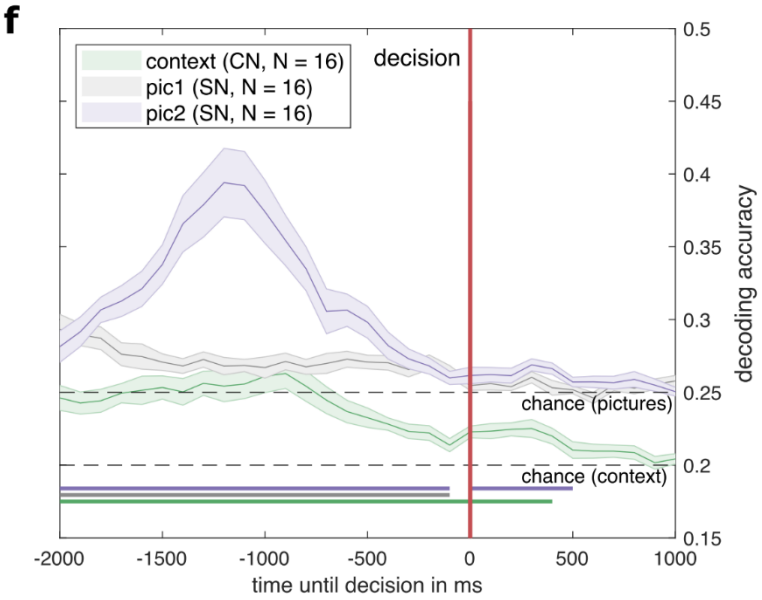
However, it should be noted that defining context neurons during the 300 question presentations instead of the 600 picture presentations considerably lowers the neuron yield and therefore requires lowering the alpha level, which in turn markedly reduces specificity and increases false positives. This is the main reason why we screened for context neurons during picture presentations.

There are two other questions that these results raise that could be addressed with these data and that could also have direct bearing on the conclusions and that should be considered.

First, what is happening at the time of choice when subjects have to choose one picture or another. If the link between context neurons and content neurons is related to memory retrieval, as the authors posit, then this should also be apparent when subjects are actually making their choice and therefore retrieving the memory of the two pictures and the associated context to choose correctly. In other words, is there evidence that item in context information is retrieved? Providing these analyses and perhaps showing that these two subsets of neurons are engaged during successful choice (and therefore memory retrieval) would provide stronger evidence that this link is meaningful.

We thank the reviewer for this important question. Based on this valuable suggestion we could confirm that both subsets of neurons are indeed engaged when one picture is chosen over another according to context at the end of the trial.

We moved Figure 3e (regional context decoding analyses) to the supplement and added a Figure 3f with SVM decoding analyses analogous to Figure 3d that depict these novel findings at the subject-level:



f. Same as **d**, but preceding context-dependent decisions. Both context and contents were represented above chance (dotted lines) until one picture was chosen over another at the end of the trial (red vertical line) according to the given context.

The results section has been updated accordingly:

> **"Generality and temporal dynamics of context representations imply relevance for stimulus processing**
> […]
>
> Finally, we tested whether content and context representations of stimulus and context neurons persist until the end of the trial when subjects chose one picture over another according to context (Fig. 3f). Both context as well as picture contents could be decoded above chance until a decision was indicated (two-sided cluster permutation test $p < 0.01$)."

The method section now reads:

> **"Support vector machine population decoding**
> […]
>
> Lastly, we assessed decoding accuracies of stimulus and context neurons preceding the end of the trial when one picture was chosen over the other according to context. Again, subject-wise SVM decoding accuracies of contents and context were computed from the activity of stimulus and context neurons, respectively and compared to chance (Fig. 3f; two-sided cluster permutation tests $p < 0.01$).

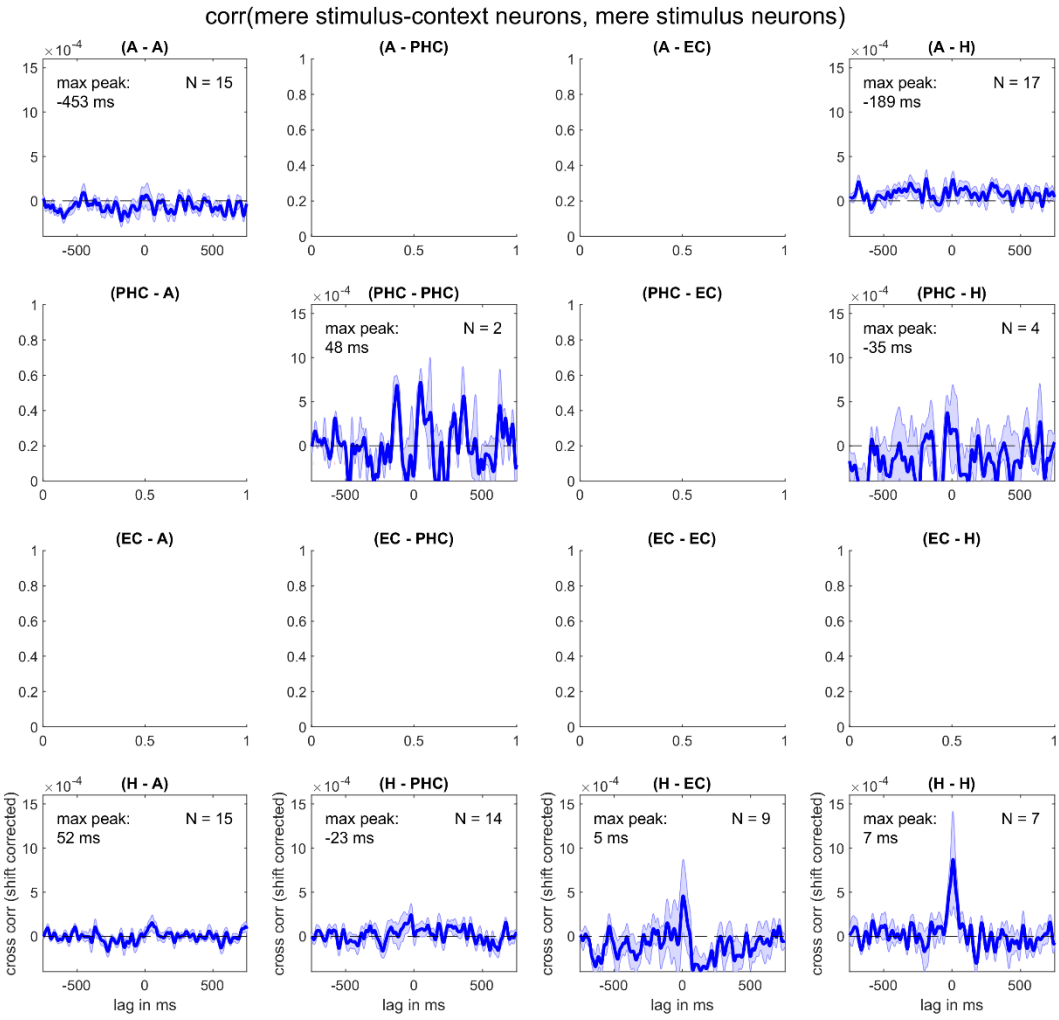
Furthermore, the discussion has been updated:

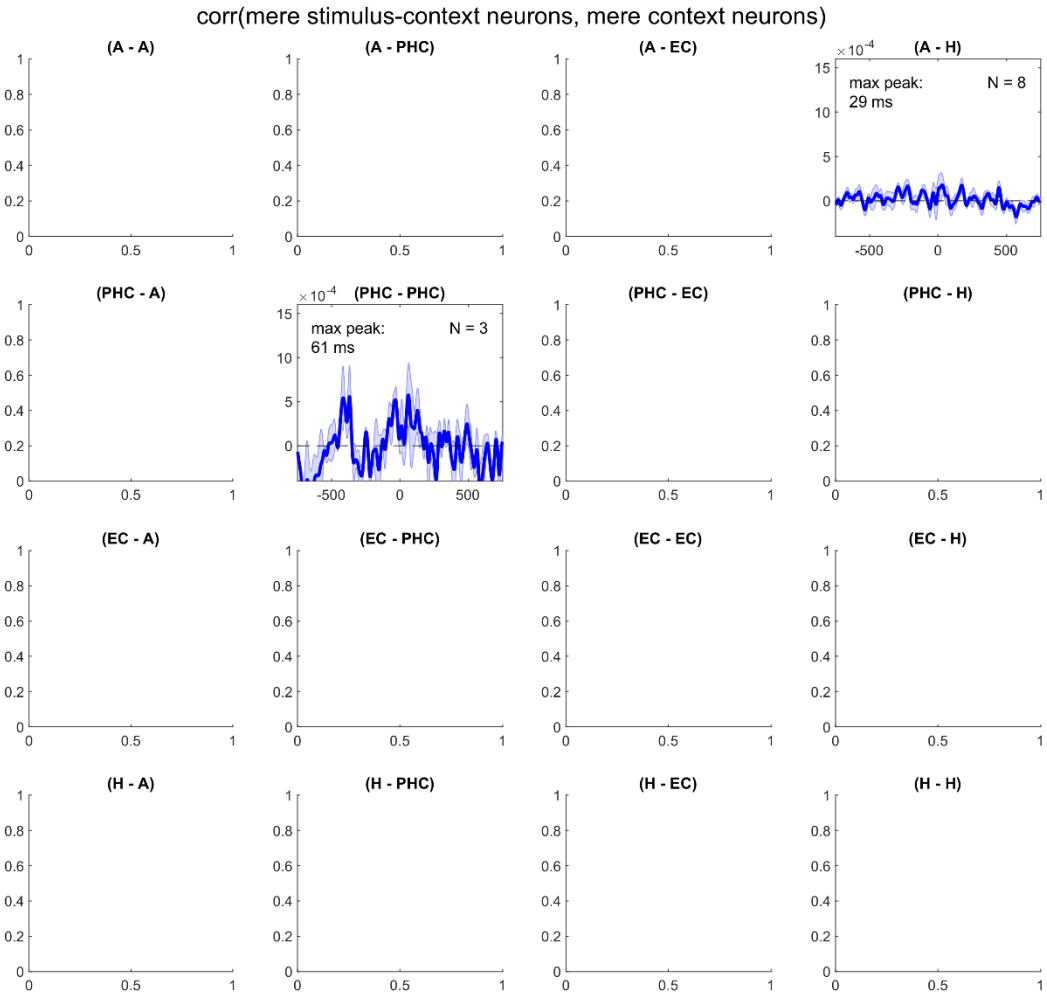
> "**Discussion**
> […]
> During picture presentations when questions and picture contents became task-relevant, context neurons encoded question context, and stimulus neurons encoded picture contents. (Fig. 3). **Importantly, both context and content representations persisted until the end of the trial when one picture was chosen over the other according to the given context (Fig. 3f).**"

Second, perhaps the strongest evidence here that context and content are linked in this dataset is the presence of stimulus context neurons, those neurons that only respond to the conjunction of a specific stimulus and a specific context. It seems that these are an important population of neurons, yet any further analysis beyond just identifying them is dropped. It would be helpful to know how activity of these neurons is related to the context and content neurons in the different brain regions, and how this may related to the correlated activity between them.

We agree with the reviewer that further interaction analyses involving stimulus-context neurons would indeed be valuable. Unfortunately, however, there are too few neuron pairs of mere stimulus-context neurons and mere stimulus or context neurons across brain regions within the same recording session in our data.

Nevertheless, we computed shift-corrected cross-correlations for these neural pairs:





Minor points

In Fig 2c, it looks like most neurons that respond to context are also responsive to stimulus. Yet this appears to be the opposite of what is depicted in the Venn diagram in Figure 2a, where it looks like most context responsive neurons do not respond to stimulus?

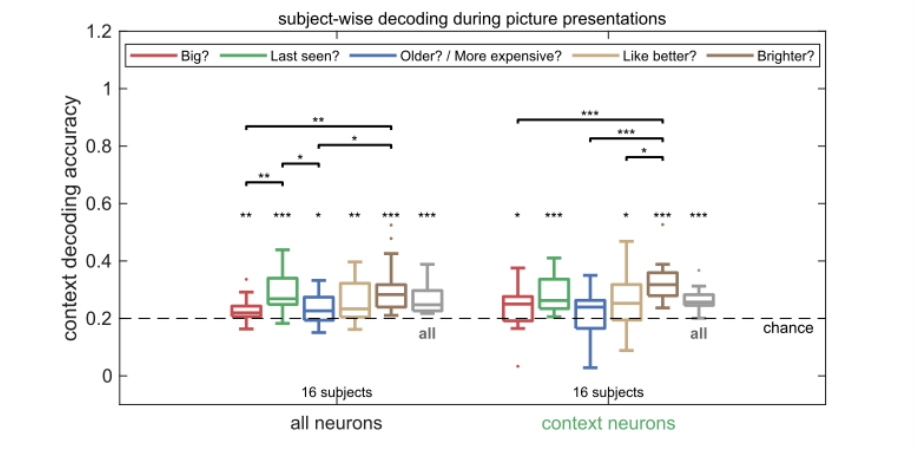
It is actually not the case that most neurons that respond to context are also responsive to stimulus. Rather, Figure 2c depicts the fraction that exhibit a main effect of context in the ANOVA among neurons that either do exhibit a main effect of stimulus or that do not:

(#context & #stimulus) / (#stimulus) -> approximately 12.15 %
versus
(#context & #no_stimulus) / (#no_stimulus) -> approximately 5%

However, the reviewer raises an interesting question regarding the fraction of context neurons that also exhibit a main effect of stimulus, which was 36.5% in alignment with Figure 2a: (55+18)/(55+18+8+119).

For the classifier results in Fig 3a, what is the overall level of decoding classification across sessions/subjects? The results here are presented separately for each category.

The overall decoding performance across all contexts amounted to 0.263 for all neurons and to 0.264 for context neurons, and both strongly exceeded chance ($p < 0.001$). Novel boxes (see "all" in gray) were added to Fig. 3a:



Furthermore, the results have been updated:

> "**Generality and temporal dynamics of context representations imply relevance for stimulus processing**
>
> […] Confirming previous ANOVA results, decoding accuracies significantly exceeded chance levels of 0.2 (one out of five questions, two-sided Wilcoxon signed-rank test, uncorrected) for all contexts when all neurons were included (Bigger?: $p_{session} = 1.981\times 10^{-3}$, $p_{subject} = 4.094 \times 10^{-3}$; Last seen?: $p_{session} = 9.142 \times 10^{-7}$, $p_{subject} = 6.430 \times 10^{-4}$; Older / More expensive?: $p_{session} = 5.187 \times 10^{-3}$, $p_{subject} = 3.861 \times 10^{-2}$; Like Better?: $p_{session} = 1.456 \times 10^{-3}$, $p_{subject} = 6.624 \times 10^{-3}$; Brighter? : $p_{session} = 8.016 \times 10^{-8}$, $p_{subject} = 4.378 \times 10^{-4}$; **all contexts: $p_{session} = 4.378 \times 10^{-4}$, $p_{subject} = 4.378 \times 10^{-4}$**). Similar results were obtained when restricting these analyses to context neurons (Bigger?: $p_{session} = 1.713 \times 10^{-2}$, $p_{subject} = 4.938 \times 10^{-2}$; Last seen?: $p_{session} = 5.888 \times 10^{-5}$, $p_{subject} = 4.378 \times 10^{-4}$; Older / More expensive?: $p_{session} = 3.875 \times 10^{-2}$, $p_{subject} = 4.08 \times 10^{-1}$; Like Better?: $p_{session} = 8.126 \times 10^{-3}$, $p_{subject} = 4.373 \times 10^{-2}$; Brighter? : $p_{session} = 9.826 \times 10^{-7}$, $p_{subject} = 4.368 \times 10^{-4}$; **all contexts: $p_{session} = 4.368 \times 10^{-4}$, $p_{subject} = 4.368 \times 10^{-4}$**).

For the analysis on correlations between mere stimulus and mere contrast neurons, presumably these neurons should only respond on a small subset of trials. But the result here is strange. If there is a non-preferred stimulus (or context), then there should be no response, and yet we are seeing responses. So are these really selectively responding neurons (e.g., mere stimulus or mere contrast neurons).

See above. To our mind, the pairing of all pictures with all questions increases synaptic gating among all pairs of stimulus and context neurons which results in a shifted cross-correlation peak regardless of which picture is actually presented. The specificity arises from the fact that the activation of an entorhinal stimulus neuron particularly drives the hippocampal context neurons which was previously activated by the question (see novel Fig. 5e,f). Findings from the novel analyses suggested by the reviewer (Fig. 5f) nicely support this view.

Referee #2 (Remarks to the Author):

This manuscript describes a study examining the extent to which single neurons in different regions of the human medial temporal lobe (MTL) demonstrate sensitivity to specific stimulus events, operationalized as one of 4 different pictures, and 'contexts', operationalized in terms of one of 5 different comparative questions asked of picture pairs, as well as whether any neurons demonstrate either joint sensitivity to these factors, or respond to them conjunctively. The study also examined temporal relationships, in the form of lagged cross-correlations, between firing patterns in different MTL regions. The findings are interpreted as evidence that, for the most part, individual stimulus items and contexts are separately represented at the single neuron level and that there is a temporal dependency between 'context' neurons in the hippocampus and earlier firing item-related neurons in entorhinal cortex.
This study has some notable strengths – the conclusions are based on what, at the level of single neurons (though see below) is a large sample, and the signal analysis methods employed are rigorous and comprehensive. Additionally, the question of how the brain and, notably, brain regions implicated in episodic memory such as the hippocampus, represent stimulus identities and disambiguate stimuli according to their contexts is a timely and important one.

We thank the reviewer for this concise summary and positive evaluation of our manuscript. Based on the reviewer's highly relevant questions and suggestions, the manuscript has been extended and strengthened significantly. For example, design choices are now explained more thoroughly, statistical findings have been reproduced at the subject level and now include behavioral performance. Furthermore, multiple additional supporting findings are now presented in new Supplementary figures.

These strengths are offset by what in my opinion are a number of weaknesses, the most salient of which are described below.

i) The manipulation of 'context' employed here involves variation in the nature of the comparative judgment made on a pair of visual images (i.e. bigger; last seen; older; more expensive; etc.). This manipulation has the unfortunate consequence of changing not only the context, but those features of the stimulus events toward which attention is directed. For example, the features of the image of a truck that receive most attention will be quite different when it is part of an affective comparison rather than a comparison of, say, monetary value. Thus, the nature of the stimulus representations will change according to context. This might explain why some of the neurons identified in this study demonstrated conjunctive context/identity coding. More generally, the authors should provide the rationale for their decision to employ a manipulation of 'interactive' rather than 'extrinsic' context.

We appreciate the reviewer's careful consideration of our task design and the way we operationalized context. In a new paragraph of the discussion we now provide a rationale for our interactive task design and evaluate potential effects of attention in light of the data.

We agree that attention can co-vary as a function of context in our task. To our mind, however, real-world contexts share this property as they also modulate which stimulus features need to be attended to, and what content-specific information needs to be retrieved for adequate decision-making (e.g. assessing a food's price in a supermarket as opposed to its taste at home). Our task captures these characteristics of real-world contexts. Different questions determine what features need to be attended to (as the reviewer rightly points out), which content-specific memories need to be retrieved according to context, and how this information is to be used to guide decision-making.

In order to assess the effects of attention specifically, different populations of this study need to be analyzed separately.

Context neurons:

To our mind, context neurons do not appear to primarily reflect attention effects in our data. For example, the questions at the beginning of the trial do not have features such as physical size or a price, etc., and are thus not expected to modulate attention during question presentations themselves. Nevertheless, context neurons already encode context during question presentations (Fig. 3e, Fig. 5b Supplementary Figure 6a), reinstate question activity patterns during subsequent picture presentations (Fig. 3e), and fire more strongly during subsequent picture presentations following strong responses to questions (Fig. 5b, Supplementary Figure 6a,c). They even exhibit a stimulus-driven enhancement of context representations (novel analysis in Fig. 5f), and context representations persist after picture offsets until a decision is made (novel analysis in Fig. 3f).

Stimulus(-context) neurons:

We agree that stimulus(-context) neuron activity could reflect attention and now deliberate on this possibility in the discussion. However, to our mind this only underlines the importance of our main finding that most stimulus neurons were not modulated by context. Additionally, attention effects do not explain the primary modulation by either stimulus or context (Fig. 2a,e) or the distinct properties of context versus content neurons (see for instance the regional asymmetry in Fig. 4a as schematized in Fig. 5a). We have therefore added the following paragraphs to the discussion:

> **"Operationalization of real-world contexts**
>
> Real-world contexts are associated with particular contents. Together they not only shape which stimulus features need to be attended to but also which memories are relevant for decision-making. For instance, in a supermarket (context), a discounted melon's (content) usual price (memory) is relevant for a buying decision. However, in a different context such as the kitchen at home, memories of the melon's typical appeal or taste are more relevant for deciding if it should be prepared and eaten. Our task operationalizes these characteristics of real-world contexts (e.g. see Polyn, Norman, & Kahana, 2009,[13] with different contexts for different word lists). Different questions comprise contexts that are to be associated with picture contents, and together not only guide attention, but also shape which content-specific memories are relevant for adequate decision-making in each respective context. […] Furthermore, attention or saliency effects alone would not explain the primary modulation by either stimulus or context (see next paragraph; Fig. 2a,e) or the distinct properties of context versus content neurons (e.g. see regional asymmetry in Fig. 4a as schematized in Fig. 5a).
>
> **Context representations across MTL regions**
>
> While the large majority of stimulus neurons (88%) were invariant to context, and most context neurons (63.5%) were invariant to stimulus, a small but significant fraction of stimulus neurons represented either both or specific conjunctions of context and stimulus, particularly in the hippocampus (see Fig. 2b-d). Overall, conjunctive representations of specific pictures and questions were most abundant in hippocampus (2.29%), particularly among stimulus neurons (10.34%). **It should be noted that the context invariance of most stimulus neurons (Fig. 2a,e) is particularly remarkable given that attention fluctuations among different question contexts could potentially lead to artificial context-modulation.**"

ii) The procedure employed to select the 4 stimulus items to be employed with any given participant requires justification. As I understand it, the items were selected so as to maximize the likelihood of identifying neurons in the experiment that would respond selectively to only one of the 4 items. Two issues arise – first, doesn't this procedure bias the data set in favor of stimulus over context identity? In principle, a similar procedure could have been employed to maximize the likelihood of identifying discriminative responses to a sub-set of a larger pool of contexts. Since this wasn't done, it's perhaps unsurprising that far more stimulus-specific neurons were identified than neurons that responded selectively to a single context (e.g. fig 2A).

The reviewer is absolutely right in pointing out that stimulus neurons were screened for in a preceding recording sessions, while context neurons were not, and that it therefore not surprising that there are more stimulus neurons than context neurons. This is by design and to our mind it is therefore all the more remarkable that we find a comparable number of context neurons. This design-intrinsic bias concerning the number of stimulus neurons is now stated more clearly both in the results section and in the discussion (see response to the reviewer's next point).

> "**Results**
>
> **Distinct populations encode stimulus and context in a picture comparison task**
>
> […] While we expected to find many neurons with a significant main effect of stimulus due to our pre-selection of response-eliciting pictures **which artificially increased the number of stimulus neurons**, it was unclear to what degree and how context would be encoded."

Second, the identification of the 4 pictures subsequently employed in the experiment itself took place in a specific context, just not one that was controlled or manipulated. This raises the possibility that at least some of the neurons responding to these pictures in the experiment proper may be ones that were 'pre-selected' to be relatively impervious to contextual change.

The reviewer raises the intriguing question of whether the screening procedure could have resulted in a bias towards context invariance, i.e. whether a substantial subset of stimulus neurons encoded pictures differentially in the screening versus the main experiment due to context modulation, leaving only context-invariant neurons to be measured in the subsequent main experiment.

To our mind, even the existence of some subset of stimulus neurons that is context invariant and could potentially contribute to the memories of contents across contexts would constitute an important finding by itself. Moreover, if a neuron only encodes pictures during the screening, but not in the subsequent main experiment (or vice versa), it is difficult to disentangle technical reasons, such as electrode drift or clustering effects, from genuine context effects.

Nevertheless, we highly appreciate the reviewer's point and assessed whether most stimulus neurons encoded pictures in an analogous manner across experiments.

Specifically, for each recording session, we trained a SVM decoder with neural activity from channels, i.e. micro-wires, that contained stimulus neurons (versus those that did not) during the presentation of the four pictures in the main experiment (see new Supplementary Figure 7 below). The session-wise decoding accuracy of these four pictures during the pre-screening vastly exceeded chance and was almost identical to the decoding performance within the main experiment itself.

This strongly argues against a strong bias due to pre-selection and further strengthens the claim that the vast majority of stimulus neurons is truly context-invariant.

We added the following passages to the Discussion:

"Discussion

Context representations across MTL regions

[…] Moreover, while stimulus neurons were overrepresented due to a pre-selection of response eliciting pictures in a screening before the main experiment, **this did not appear to introduce a bias towards context invariance as stimulus representations were maintained from pre-screenings to main experiments (further supporting their context-invariance, see Supplementary Figure 7).**"

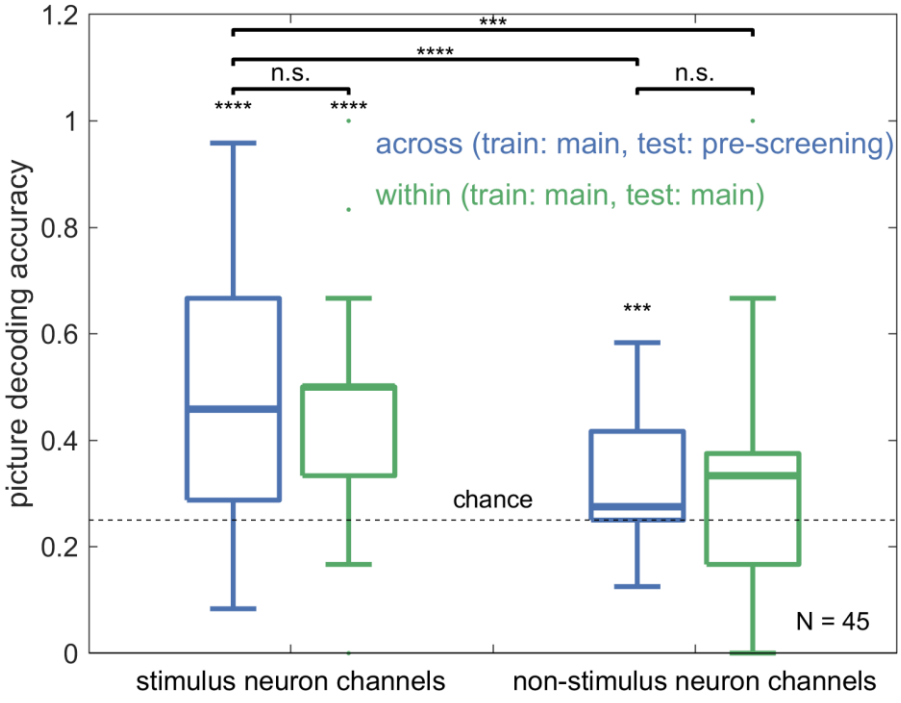


Fig. S7. Stimulus representations were maintained from pre-screenings to subsequent main experiments, particularly in micro-wires with stimulus neurons. Boxplots (Q1, median, Q3; whisker: points within ±1.5 IQR) of session-wise SVM picture decoding accuracies of the four pictures shown in the main experiment. SVM decoders were trained on main experiment activity and tested either on pre-screening (across in blue) or main experiment activity (within in green, leave-out cross-validation with matched trial counts). Dotted line in black denotes chance decoding of 0.25. For 45 recording sessions, all four pictures of the main experiment were part of the pre-screening before the experiment (N = 45). Picture decoding accuracies highly exceeded chance, particularly for micro-wire channels that contained stimulus neurons (left side; stimulus-across: 0.4854 +/- 0.2243, $p = 8.229 \times 10^{-7}$; stimulus-within: 0.4444 +/- 0.2072, $p = 1.017 \times 10^{-6}$) versus those that did not (right side; non-stimulus-across: 0.3241 +/- 0.1155, $p = 3.017 \times 10^{-4}$; non-stimulus-within: 0.2889 +/- 0.2114, $p = 3.728 \times 10^{-1}$). Importantly, there were no significant differences in decoding accuracy when testing on pre-screening data (across) or on main experiment activity (within), both for stimulus neuron channels ($p_{diff} = 0.301$) and for non-stimulus neuron channels ($p_{diff} = 0.231$; two-sided Wilcoxon signed-rank test). Overall, stimulus representations were maintained across the relatively short time intervals (4.71 hours +/- 2.02 SD, min: 1.40 h, max: 9.15 h) between pre-screenings and subsequent main experiments, specifically in channels that contained stimulus neurons. ****$p < 0.0001$; ***$p < 0.001$; **$p < 0.01$ (uncorrected).

Furthermore, the novel results are now described in the Methods section:

> "**Methods […]**
>
> **Definition of neural populations and assessment of their significance**
>
> […] In order to assess the stability of stimulus representations across pre-screenings and main experiments, Support vector machine (SVM) decoders were trained on main experiment activity and tested either on pre-screening (across) or main experiment activity (within, leave-out cross-validation with matched trial counts). For 45 recording sessions, all four pictures of the main experiment were part of the pre-screening before the experiment. Picture decoding accuracies highly exceeded chance, particularly for micro-wire channels that did versus did not contain stimulus neurons (Supplementary Figure 7). Importantly, there were no significant differences in decoding accuracy when testing on pre-screening data (across) or on main experiment activity (within), both for stimulus neuron channels and for non-stimulus neuron channels (two-sided Wilcoxon signed-rank test)."

iii) Most of the analyses are at the level of single neurons, with little attempt to address whether the findings generalize across participants (Fig. S1 is an exception). It would be useful to include some statistical analyses in which participants, rather than neurons, are treated as the unit of analysis, preferably in the context of a standard random effects approach.

We thank the reviewer for inquiring about the reproducibility and generalizability of the findings, which is clearly important. Unfortunately, epilepsy patients with intracranial micro-wires are extremely rare, and in general it is not feasible to analyze human single-neuron data at the level of sessions, let alone at the level of subjects. This is indeed an absolute exception and rarely found in any of the human single-neuron studies published to date.

It is even more true for neural interaction analyses that require the concurrent recording of different sparse neural populations across brain regions in the same recording. Nevertheless, with the exception of some of the interaction analyses, we re-computed most figures and subfigures at the level of subjects (after averaging across sessions within subjects). All of these novel analyses confirmed our previous findings. However, due to extremely small sample sizes, pairwise neural interaction analyses still can only be performed at the level of sessions (N=6) and few select analyses are still depicted at the level of neural pairs (see below). The overall findings consistently remain the same.

Moreover, we now also added random-effects analyses of all recorded neurons across participants to assess whether ANOVA effect sizes (partial eta squared) exceeded chance levels determined by our label-shuffling controls. These analyses revealed significant differences between real and permutation-based effect sizes for all three effects tested in the ANOVA, while accounting for subject-level variability. Specifically, we fitted separate linear mixed-effects models for the main effect of stimulus, the main effect of context, and the stimulus-context interaction, each modeling the respective difference between real and permuted effect sizes as a function of a fixed intercept (capturing grand population effect size differences) and subject-specific random intercepts (e.g., stimulus_diff ~ 1 + (1|subject)). We found significant effects for the main effect of stimulus (β = 0.080, 95 % CI [0.065, 0.095], $p = 2.177\times10^{-24}$), the main effect context (β = 0.025, 95 % CI [0.019, 0.032], $p = 4.186\times10^{-14}$), and the stimulus-context interaction (β = 0.002, 95 % CI [0.001, 0.003], $p = 2.269\times10^{-6}$).

The stimulus effect, which was inflated due to stimulus pre-screenings, was approximately 3.2 times larger than the context effect and 36.4 times larger than the interaction effect. Nonetheless, all effects were highly significant across participants.

The novel results have been added to the results and methods, respectively:

> "**Results […]**
> **Stimulus and context are mainly encoded separately and rarely conjunctively**
> "[…] Confirming previous results, subject-averaged effect sizes of stimulus (η^2_p(real) = 0.117, η^2_p(control) = 0.034, $p_{subject}$ = 1.313×10^{-3}, $p_{session}$ = 3.330×10^{-9}), context (η^2_p(real) = 0.059, η^2_p(control) = 0.034, $p_{subject}$ = 1.313×10^{-3}, $p_{session}$ = 4.827×10^{-9}) and their interaction η^2_p(real) = 0.036, η^2_p(control) = 0.034, $p_{subject}$ = 1.840×10^{-2}, $p_{session}$ = 3.140×10^{-3}) all significantly differed from controls (two-sided Wilcoxon signed-rank tests, Bonferroni-corrected). Linear mixed-effects models with a fixed intercept (capturing grand population effect size differences) and random intercepts for subjects, fitted across all recorded neurons, confirmed that real effect sizes significantly exceeded permutation-based baselines for all three ANOVA factors: stimulus (β = 0.080, 95 % CI [0.065, 0.095], p = 2.177×10^{-24}), context (β = 0.025, 95 % CI [0.019, 0.032], p = 4.186×10^{-14}), and stimulus-context interaction (β = 0.002, 95 % CI [0.001, 0.003], p = 2.269×10^{-6}).
>
> **Methods […]**
> **Definition of neural populations and assessment of their significance**
> […] To additionally account for subject-level variability, we fit separate linear mixed-effects models for each ANOVA factor across all recorded neurons. Each model included a fixed intercept (capturing grand population effect size differences) and subject-specific random intercepts."

Revised Figures with novel subject (or session) -level analyses are presented on the following pages:

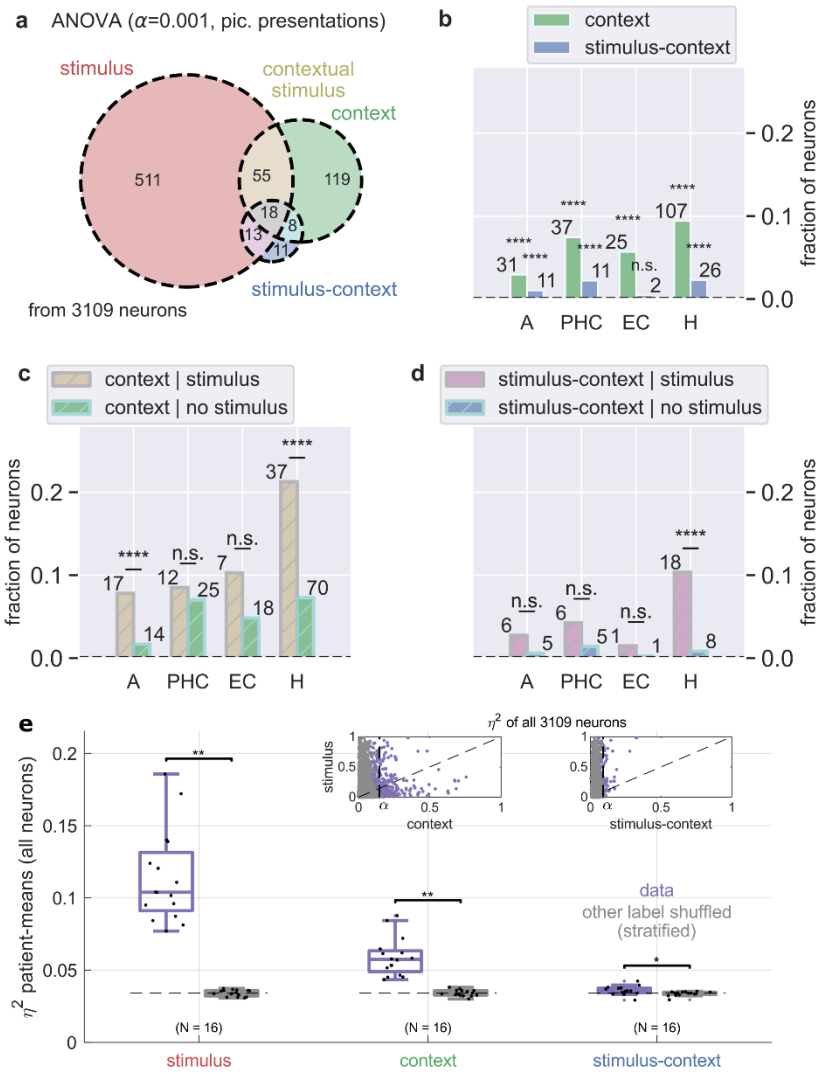


Fig. 2. Stimulus and context are mainly encoded separately and rarely conjunctively. a. Venn diagram of significant number of neuron types defined by repeated-measures two-way ANOVAs with factors stimulus identity (stimulus neurons, red/ocher/gray/pink), question (context neurons, green/ocher/gray/marine) or their interaction (stimulus-context neurons, gray/pink/blue/marine) during picture presentations at an alpha level of 0.001. Intersections denote multiple significances. While most stimulus neurons were not modulated by context, they overlapped with context neurons (contextual stimulus neurons) or stimulus-context neurons. **b.** Probabilities of obtaining context or stimulus-context neurons across brain regions (absolute numbers on top). Significance stars depict binomial tests relative to chance (dotted line). Neurons in all MTL regions were strongly modulated by context and stimulus-context. **c.** Analogous to **b** with probabilities of context conditional on stimulus modulation (stimulus / no stimulus). Significance stars depict results of Fisher exact tests computed from regional contingency tables. Stimulus neurons were more likely to be modulated by context (contextual stimulus neurons) than neurons without a significant main effect of stimulus in all MTL regions (significant in A and H), particularly in hippocampus. **d.** Analogous to **c**, with probabilities of significant interactions (stimulus-context). Stimulus-context interaction neurons with conjunctional representations were most frequent in hippocampus. **e.** Boxplots of subject-averages of repeated-measures two-way ANOVA effect sizes from all 3109 neurons ($\eta^2_{partial}$) for factors stimulus, context and interaction (stimulus-context). Data (lavender) are compared to controls (Wilcoxon signed-rank test) obtained via stratified shuffling of the relevant label (stimulus or context, gray). Effect sizes markedly exceeded chance for stimulus (η^2_p(real) = 0.117, η^2_p(control) = 0.034, $p_{subject}$ = 1.313×10^{-3}, $p_{session}$ = 3.330×10^{-9}) and context (η^2_p(real) = 0.059, η^2_p(control) = 0.034, $p_{subject}$ = 4.827 × 10^{-9}, $p_{session}$ = 1.087 × 10^{-9}). Difference for stimulus-context was smaller but significant (η^2_p(real) = 0.036, η^2_p(control) = 0.034, $p_{subject}$ = 1.840 × 10^{-2}, $p_{session}$ = 3.140 × 10^{-3}). Inlets depict scatters plots of $\eta^2_{partial}$ of context (left) or stimulus-context (right) plotted against those of stimulus. Dashed lines indicate the value of the smallest effect size below the alpha level of 0.001 (α). ***$p < 0.001$; **$p < 0.01$ (Bonferroni corrected).

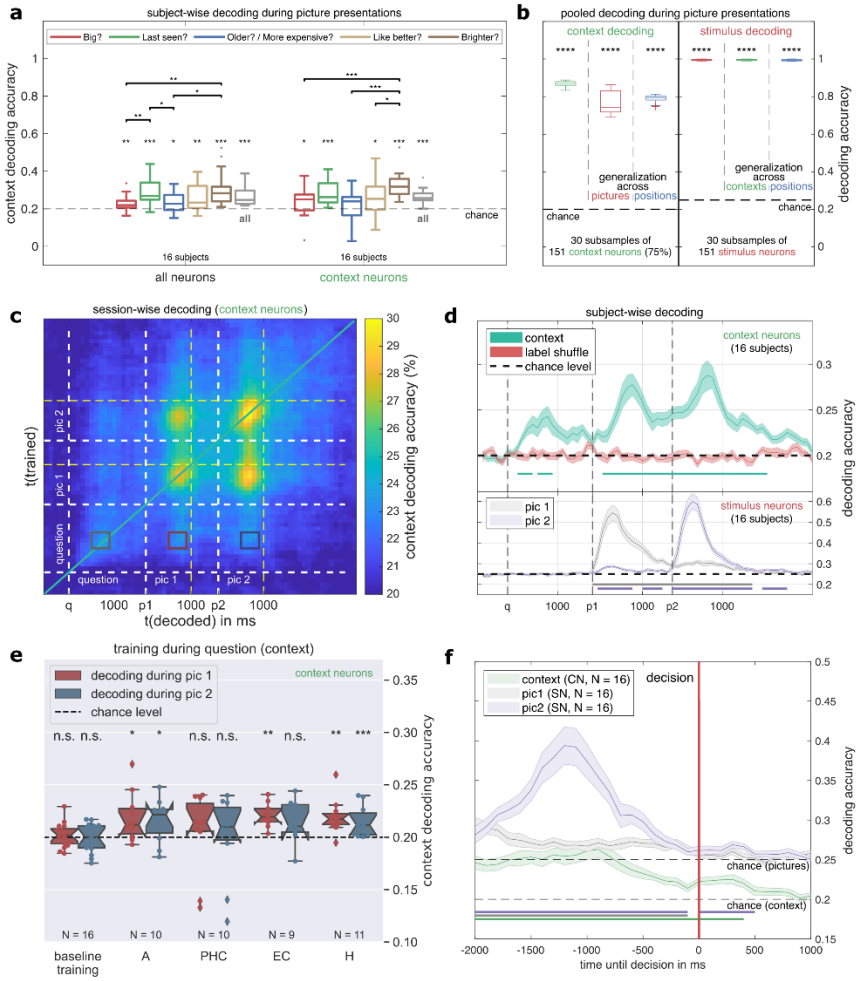


Fig. 3. Generality and temporal dynamics of context representations imply relevance for stimulus processing. Abstract representations of context (a-b) invariant to picture identity or position (b) are combined with those of content (b) across temporal gaps (c-e) both during picture presentations (d) via context reactivation (e) and before context-dependent decisions (f). **a.** Boxplots of subject-wise population-decoding accuracy of context (SVM: support vector machines) during picture presentations (100-1000 ms) for different questions (colors) and for all vs. context neurons. **b.** Boxplots of pooled context (green) or stimulus (red) decoding accuracies (SVM) from subsamples of context (left) and stimulus neurons (right), either obtained by 5-fold cross validation or generalization across pictures (red), contexts (green) and serial picture positions (blue). **c.** Heat map of subject-wise SVM context-decoding accuracies from context neuron activity at different time points. Diagonal line in green corresponds to identical training and decoding times (see **d**). Dashed lines denote onsets (white) and offsets (yellow) of trial events (question, picture 1, picture 2). Boxes indicate particular training (question) and decoding times (late picture presentations) further analyzed in **e**. **d.** Subject-wise decoding accuracies of context (top, context neurons, green) or stimulus (bottom, stimulus neurons, picture 1 and 2 in gray and lavender, respectively). Context was encoded above chance across the entire trial. During picture presentations, decoding accuracies of stimuli and contexts peaked (first stimulus, then context) and remained high in parallel. Solid lines and shaded areas depict means and standard errors, respectively. Data are compared to label-shuffled surrogate data (red) and chance (dashed line in black). Solid lines below indicate significant differences (cluster permutation test, $p < 0.01$). **e.** Boxplots of subject-wise context-decoding accuracies trained with context-neuron activity during question presentations (A: amygdala, PHC: parahippocampal cortex, EC: entorhinal cortex, H: hippocampus) or baseline (BL) and decoded during late first (red) or second (blue) picture presentations (see boxes in **c**). Question activity was recapitulated during first and second picture presentations, particularly in H (and EC), but not in PHC. Neuron numbers were matched across regions (N = 400). **f.** Same as **d**, but preceding context-dependent decisions. Both context and contents were represented above chance (dotted lines) until one picture was chosen over another at the end of the trial (red vertical line) according to the given context. For boxplots in **a,b,e** (Q1, median, Q3; whisker: points within +/- 1.5 IQR) stars denote significant differences from zero (Wilcoxon signed-rank test, uncorrected) or each other (Mann–Whitney *U* test, uncorrected). ****$p < 0.0001$; ***$p < 0.001$; **$p < 0.01$; *$p < 0.05$

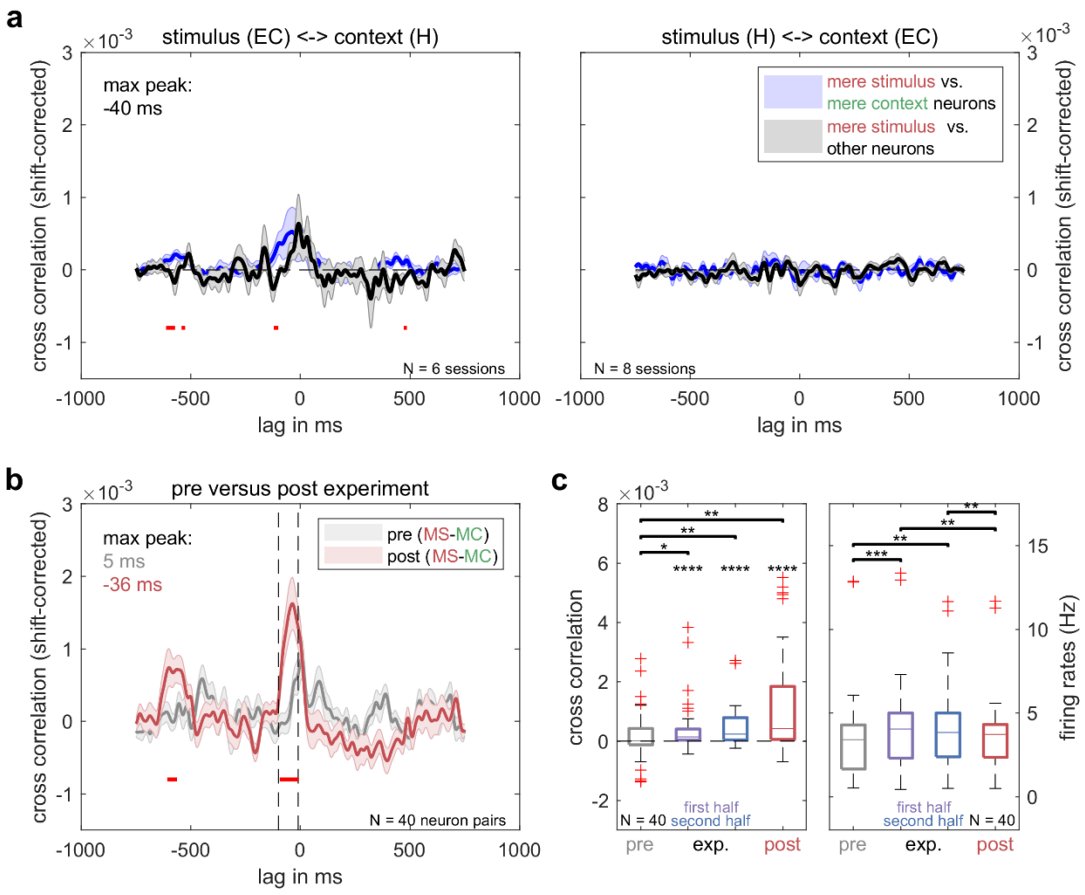


Fig 4. Sequential activation of entorhinal stimulus and hippocampal context neurons suggests a storage mechanism of stimulus-context associations. a. Left: Session means of trial-by-trial cross-correlograms during picture presentations (shift-corrected, see Methods) either between pairs of entorhinal mere stimulus and hippocampal mere context neurons (MS or MC with only one of two factors $p < 0.001$ in ANOVA, blue) or a matched number of entorhinal mere stimulus and other neurons in hippocampus without significant main effects (ON, both factors and interaction in ANOVA $p > 0.1$, black). Data are presented as mean values ± SEM of session means (solid lines and shaded areas). Red horizontal lines indicate significant differences ($p < 0.05$, two-sided cluster permutation test) between pair types (blue vs. black). Maximum peak lag time is depicted in upper left corner. The two groups (i.e. MS-MC vs. MS-ON) differed significantly on small time scales (-30 to -40 ms). Firing of stimulus neurons in EC predicted firing of context neurons in H (but not of other neurons in H) after approximately 30 to 40 milliseconds, potentially reflecting a storage mechanism of stimulus-context associations for the re-instatement of context during stimulus presentations. Right: Same as left for neuron pairs from same brain regions in opposite order. Here, activity of hippocampal stimulus neurons did not predict entorhinal context neuron firing. **b.** Shift-corrected cross correlations of all 40 entorhinal mere stimulus and hippocampal mere context neuron pairs (see **a**, left, blue), but here either before (pre, gray) or after the experiment (post, red). Asymmetric cross correlations at approximately -40 ms were not present before the experiment, but emerged during its course (see **a,c**) and persisted after its termination ($p < 0.01$, two-sided cluster permutation test). **c.** Emergence of interactions during the experiment. Left: Boxplots of mean cross correlations (shift-corrected) between same neuron pairs as in **b** from -10 to 100 ms (see vertical dashed lines). Mean cross correlations did not differ significantly from zero (two-sided Wilcoxon signed-rank test) before the experiment ($p = 0.29$, pre in gray), but exceeded chance levels (all $p < 0.001$) both during the first (exp., violet) and second (exp., blue) half of the experiment as well as after its termination (post, red). Cross correlation sizes of all subsequent time periods were significantly larger than before the experiment (p(pre, exp1) = 3.258×10^{-2}; p(pre, exp2) = 6.625×10^{-3}; p(pre, post) = 3.106×10^{-3}. Right: In contrast, mean firing rates did not differ significantly before and after the experiment, but were significantly lower ($p < 0.01$) compared to the first and second half of the experiment. ****$p < 0.0001$; ***$p < 0.001$; **$p < 0.01$ (uncorrected).

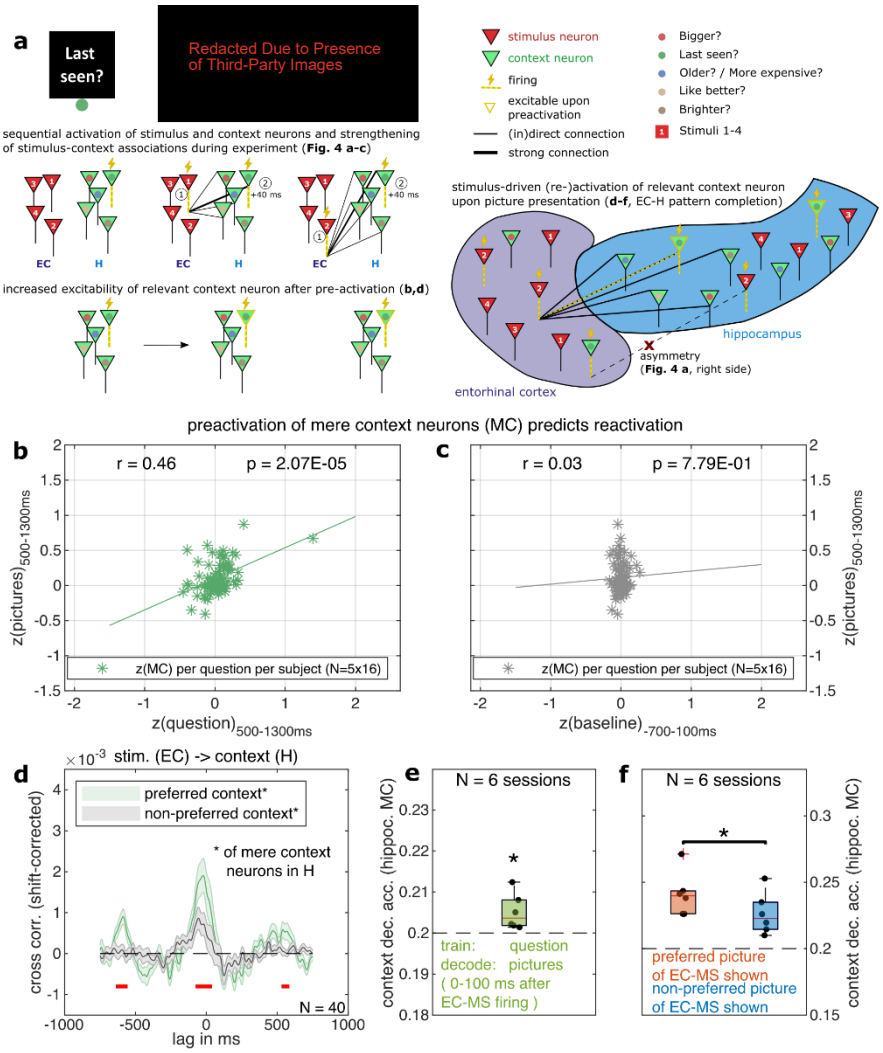


Fig 5. Sequential activation after pre-activation of context neurons by their preferred context could guide stimulus-driven pattern completion. a. Schematic model of stimulus-context storage. Upper left: Activation patterns of entorhinal (EC) mere stimulus (red) and hippocampal (H) mere context neurons (green) during different trial events. Specific context neurons in H respond to particular questions (see inner colored circles) and are reactivated during subsequent picture presentations. Picture presentations elicit responses in selected stimulus neurons (see inner numbers), whose activity predicts firing of context neurons after 40 ms (see Figure 4), resulting in strengthening of connections that could encode stimulus-contexts associations. Lower left: Question response strengths predict activity and excitability (yellow triangles) of context neurons during subsequent picture presentations (see **b,d**, and Supplementary Figure 6a-c). Right: After connections have been strengthened, pre-activation by preferred contexts can selectively guide stimulus-driven pattern completion in hippocampal context neurons (dashed line in yellow). **b.** Scatter plot of mean z-values for each of the 5 contextual questions (baseline normalization) during question versus subsequent picture presentations computed for each mere context neuron and averaged per subject (N = 5 × 16). Pearson correlation strengths and *p*-values (uncorrected) are shown at the top and visualized by regression lines. Question response strengths significantly predict responses to subsequent pictures ($r = 0.46$, $p = 2.07 \times 10^{-5}$). **c.** Same as **b**, but comparing baseline to picture presentations**.** Baseline response strengths do not predict subsequent picture responses ($r = 0.03$, $p = 0.78$). **d.** Shift-corrected cross correlations of entorhinal mere stimulus (EC-MS) and hippocampal mere context neurons (H-MC) during picture presentations after a preferred (max. response, green) versus a non-preferred context (gray) was presented. Red horizontal lines indicate significant differences ($p < 0.01$, two-sided cluster permutation test). Firing of MS predicts MC firing significantly more strongly in preferred contexts. **e.** Session-wise context SVM decoding accuracy of H-MC trained with question activity and decoded with subsequent picture activity (100-1500 ms) time-locked to E-MS firing (100 ms windows; 6 sessions: $p = 0.031$; 40 neuron pairs: $p = 5.794 \times 10^{-4}$; two-sided Wilcoxon signed-rank test). **f.** Session-wise context SVM decoding accuracies of H-MC distinguishing whether preferred or non-preferred pictures of the corresponding EC-MS were presented (6 sessions: $p = 0.0156$; 40 neuron pairs: $p = 0.0338$; one-sided Wilcoxon signed-rank test). *$p < 0.05$.

iv) Neurons are sampled from a 4 different MTL structures that are thought to have very distinct functions and to support quite different computations. Moreover, especially with respect to the entorhinal cortex and, even more so, the hippocampus, different sub-regions of these structures are held to support different kinds of computation (e.g., the DG and CA3 hippocampal subregions are thought to support pattern separation to a much greater extent than CA1). Do the authors have any information about the likely localization of the neurons that they recorded from within these regions? If not, this issue should at least be noted as a limitation and a constraint on the model outlined in fig. 5A.

We agree that analyses of sub-regions would both be expected to differ in terms of pattern separation, pattern completion, etc., and could also strengthen the model, but unfortunately recordings in humans do not allow for a reliable localization of micro-wires beyond individual medial temporal lobe subregions, e.g., distinguishing between subiculum and the CA/DG region. We now explicitly address this limitation in the discussion:

> "**Discussion**
> **Context representations across MTL regions**
> […]
> Furthermore, the exact placement of micro-wires within a target region (e.g. within hippocampus or entorhinal cortex) cannot be determined reliably with the noninvasive imaging methods allowed in humans. Therefore, it is unclear whether hippocampal micro-wires were recorded from CA1/CA3 dental gyrus et cetera."

Referee #3 (Remarks to the Author):

In this manuscript, the authors present single unit data recorded during a visually presented object comparison task in 17 neurosurgical patients. They report that the vast majority of neurons in human MTL represent context and content separately. They find many stimulus-modulated neurons ("content" neurons; not surprising since stimuli were pre-selected), some context-modulated neurons, and a significant minority of neurons encoding the interaction. They suggest that the content and context information streams may be later combined through co-activation, synaptic modification, or reinstatement. They are able to decode context significantly from all neurons and context-neurons only. Decoding accuracy was better when stimuli were presented (i.e. when the context needed to be recalled in order to evaluate the stimulus). The authors also find that firing of stimulus neurons in EC reliably precedes firing of context neurons in HC, suggesting that the firing of EC stimulus neurons triggers a context association in HC, potentially via synaptic plasticity.

I congratulate the authors on their efforts to further disentangle the workings of the human MTL and its role in context-dependent memory. I do have a number of concerns and questions regarding methodology and interpretation. The major concerns are those that influence interpretability and impact. The specific questions need answers, but the answers are unlikely to significantly influence impact.

We highly appreciate the reviewer's positive assessment as well as the valuable feedback and questions. As a result, the manuscript has been improved and strengthened, and multiple novel analyses now provide additional evidence supporting our conclusions.

Major Concerns

The first paragraph of the results states 50 experimental sessions across the 17 patients, whereas the Methods say 38 sessions.

We thank the reviewer for catching this, and we apologize for the confusion. Indeed, the phrasing in the methods section was wrong in the previous manuscript. In general, all 50 sessions were analyzed (e.g. for the definition of neuron types, Fig 2e, Fig. 3a, etc. – note that now 1 session, and consequently one subject, was excluded in the new analyses due to low behavioral performance, see below). Our analyses of context neurons, however, could only be performed in the 38 sessions they could be detected in. Accordingly, we changed the method section. which now reads:

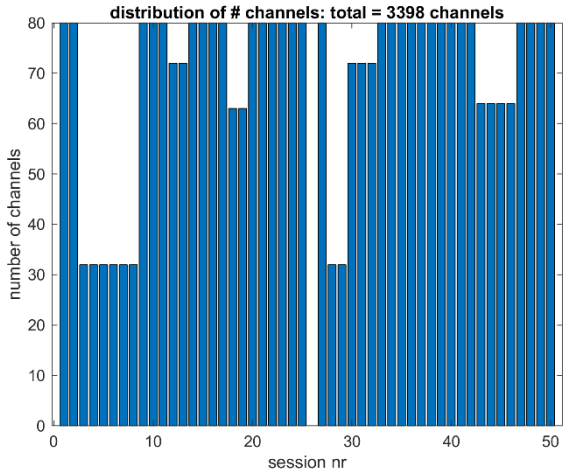
> "**Methods**
>
> **Subjects and Recording**
>
> […] Our original data set consisted of **50** experimental sessions from 17 subjects with recordings from the amygdala (A), parahippocampal cortex (PHC), entorhinal cortex (EC), and hippocampus (H). One session from one subject was excluded due to low behavioral performance, resulting in a final sample of 49 sessions from 16 subjects. Remarkably, the excluded session stemmed from the only subject for which no context neurons could be detected (others exhibited an average of 12.5). […]"

Moreover, we modified a sentence in the "Statistics and Reproducibility" section:

> "49 sessions from 16 subjects were analyzed (one session of one subject was excluded based on a low behavioral performance). A total of 200 context neurons were detected in 38 sessions from the 16 subjects. 25 sessions contained at least three stimulus and context neurons each, 12 sessions more than five."

The total number of units is reported to be 3,127, meaning either 63 or 82 units per recording session, which is extremely high. How many microwire bundles were placed? Typical yield per microwire is 0.5-1 unit per wire in optimal conditions, so either there were many bundles placed or the yield per bundle is higher than most groups report. Either of these possibilities warrants further discussion.

We thank the reviewer for this correct assessment. In our case, many bundles were placed, usually ten with 80 micro-wires, resulting in 69.35 channels per session on average (range: 32 to 80) and 3398 channels in total (see figure below). Therefore, the yield of units was 0.92 per channel, consistent both with our overall high recording quality and the reviewer's suggestion:



We added the following sentence to the methods section:

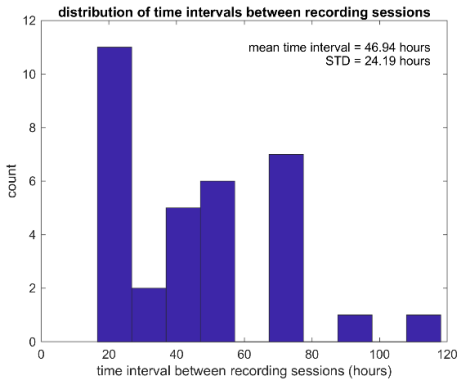
"**Methods**
Subjects and Recording

[…] Each depth electrode contained a micro-wire bundle consisting of nine micro-wires, consisting of a reference electrode with low impedance and eight high-impedance recording electrodes (AdTech, Racine, WI), which protruded from the tip of the electrode by approximately 4 mm. All bundles were localized using a post-implantation CT scan co-registered with a pre-implantation MRI scan normalized to Montreal Neurological Institute space. The differential signal from the micro-wires was amplified using a Neuralynx ATLAS system (Bozeman, MT), filtered between 0.1 Hz and 9,000 Hz, and sampled at 32 kHz. These recordings were stored digitally for further analysis. Our original data set consisted of 50 experimental sessions from 17 subjects with recordings from the amygdala (A), parahippocampal cortex (PHC), entorhinal cortex (EC), and hippocampus (H). One session from one subject was excluded due to low behavioral performance, resulting in a final sample of 49 sessions from 16 subjects. Remarkably, the excluded session stemmed from the only subject for which no context neurons could be detected (others exhibited an average of 12.5). **For most subjects, ten micro-wire bundles (median of 80 channels) were placed with an average of 69.34 recorded channels (range: 32 to 80) and a neuron yield of 0.92 per channel.** […]"

A related question is about uniqueness of units across sessions. What was the distribution of time interval between recording sessions?

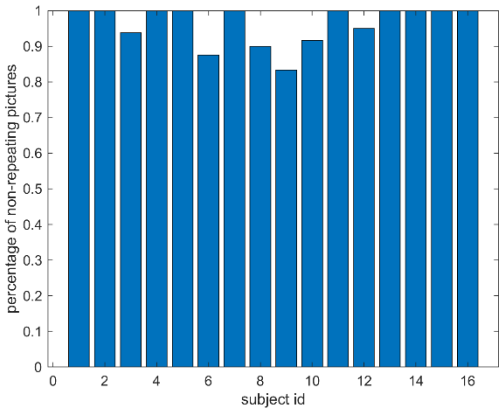
We thank the reviewer for raising this question. The mean time interval between recording sessions was 46.94 hours +/- 24.19 STD (range: 16.57 to 118.06 h) for the 13 of 16 subjects with more than one recording session (see reviewer figure below). We added the following passage to the methods section:

> "**Methods**
>
> **Subjects and Recording**
>
> […] . Our original data set consisted of 50 experimental sessions from 17 subjects with recordings from the amygdala (A), parahippocampal cortex (PHC), entorhinal cortex (EC), and hippocampus (H). One session from one subject was excluded due to low behavioral performance, resulting in a final sample of 49 sessions from 16 subjects. For most subjects, ten micro-wire bundles (80 channels) were placed with an average of 69.34 recorded channels (min: 32, max: 80) and a neuron yield of 0.92 per channel. **In 13 of the 16 subjects, more than one recording session was recorded. The mean time interval between multiple sessions was 46.94 hours +/- 24.19 SD (range: 16.57 to 118.06 h).**"



How confident are the authors that the units in one session were a completely unique set relative to the previous or the next one? Preliminary work from some groups suggests that there is a high degree of carry-over, such that only a fraction of units are different from one to the next session, especially with short delay intervals (hours). This effect would further reduce the total number of independently counted units.

We thank the reviewer for inquiring about the statistical independence between recording sessions. As described in response to the reviewer's previous question, the interval between multiple sessions was quite long (46.94 hours +/- 24.19 SD (range: 16.57 h to 118.06 h)). These long intervals across days (far exceeding intervals between screenings and subsequent main experiments, 4.71 hours +/- 2.02 SD, see below) are expected to result in a high degree of statistical independence, for example due to electrode drift. It is worth noting that the micro-wires are fixed to the skull, while the brain floats in cerebrospinal fluid (CSF). It is therefore plausible that head movements during sleep accelerate this drift.
Additionally, for multiple recording sessions in the same subject, in the vast majority of cases, different pictures were presented (96.33 % +/- 5.49 % SD). The following figure depicts the ratio of non-repeating pictures calculated by dividing the set of unique pictures across multiple sessions of the same subject by the total number of presented pictures across these sessions:



Finally, we explicitly quantified the degree of neural independence between sessions by computing intra-class correlation coefficients of the numbers of either context or stimulus neurons across channels (i.e. micro-wires) between consecutive recording sessions within subjects. Across the long time intervals between subsequent recording sessions (46.94 hours +/- 24.19 SD, range: 16.57 to 118.06 h), numbers of both context and stimulus neurons exhibited a low degree of dependence, comparable to permutation controls. We added the following Supplementary Figure S8 which is now described in the methods:

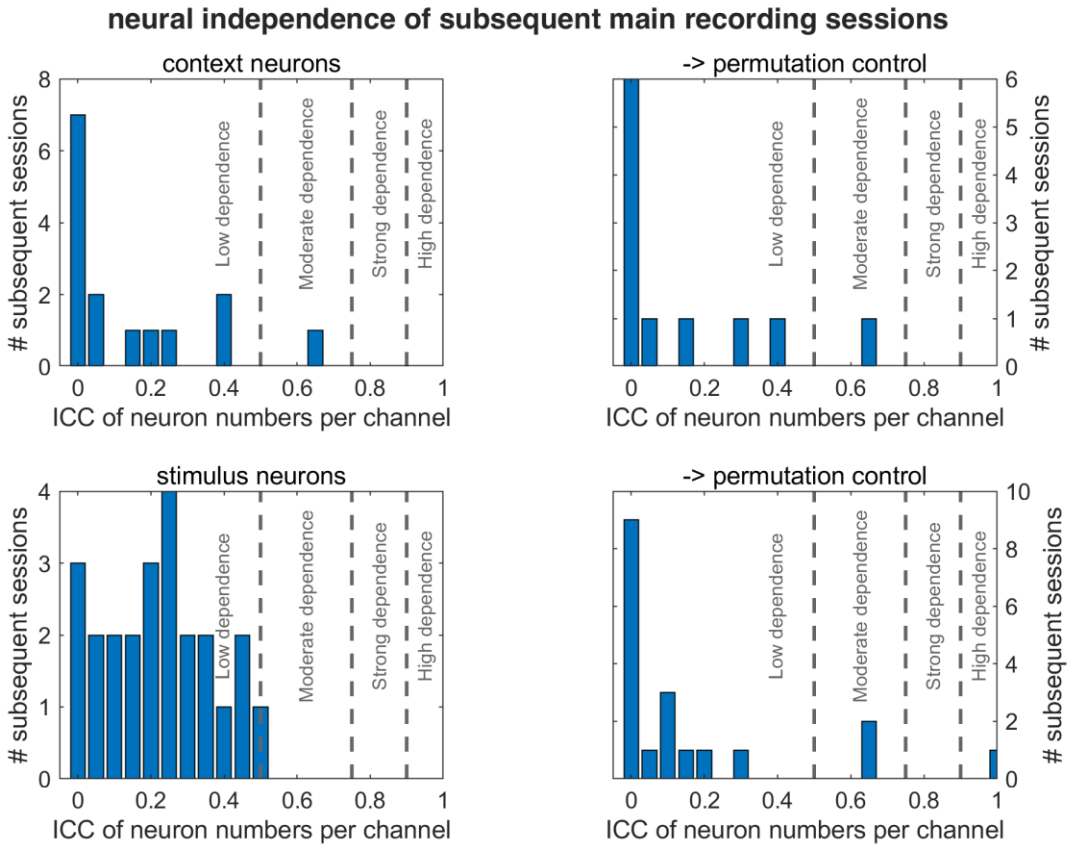


Fig. S8. N**umbers of stimulus and context neurons across micro-wires are largely independent for subsequent main recording sessions within subjects.** Bar plots of the distribution of intraclass correlation coefficients (ICCs) between either the numbers of either context (top left) or stimulus neurons (bottom left) across channels (i.e. micro-wires) for subsequent main recording sessions. Corresponding channel permutation controls are shown on the right. Vertical dashed lines in gray separate areas of differing effect sizes. Across the long time intervals between subsequent recording sessions (46.94 hours +/- 24.19 SD, min: 16.57 h, 118.06 h), numbers of both numbers context and stimulus neurons exhibited a low degree of dependence comparable to permutation controls.

"**Methods**

Subjects and Recording

[…] One session from one subject was excluded due to low behavioral performance, resulting in a final sample of 49 sessions from 16 subjects. […] **In 13 of the 16 subjects, more than one session was recorded. The mean time interval between multiple sessions was 46.94 hours +/- 24.19 SD (range: 16.57 h to 118.06 h). Neural dependence across 33 consecutive pairs of these sessions was assessed using intraclass correlation coefficients (ICCs(2,1); Shrout & Fleiss, 1979; McGraw & Wong, 1996). ICCs were computed across channels (i.e., micro-wires), separately for context and stimulus neurons. ICCs quantify cross-session stability, with lower values indicating weaker dependence. Observed ICCs were compared to null distributions obtained by permuting channels. ICCs did not strongly exceed these null distributions (Hedges' g: stimulus = 0.35; context = 0.16) and only reached significance (α = 0.01) in 6 of 33 stimulus pairs, 3 for context, and none for both, all with low overall effect sizes (see Supplementary Figure 8).**

In fact, this point of neuronal population stability across sessions is a critical requirement of this experimental paradigm. Since neurons were pre-screened based on response to stimuli, the authors rely on the fact that there is stability of that population from the pre-screening step to the actual experimental step.

Unlike the long time intervals between multiple recording sessions (46.94 hours +/- 24.19 SD, range: 16.57 to 118.06 h), time intervals between each screening and subsequent main experiment - that the reviewer inquires about below - were much shorter (4.65 hours +/- 2.05 SD, range: 1.40 to 9.15 h). These shorter time intervals are highly conducive to neuronal stability, which we now demonstrate in novel decoding analyses depicted in an additional Supplementary Figure 7 below.

We added the following passage to the Discussion:

"**Discussion**

Context representations across MTL regions

[…] Moreover, while stimulus neurons were overrepresented due to a pre-selection of response-eliciting pictures in a screening before the main experiment, this did not appear to introduce a bias towards context invariance **as stimulus representations were maintained from pre-screenings to main experiments (further supporting their context-invariance, see Supplementary Figure 7).**"

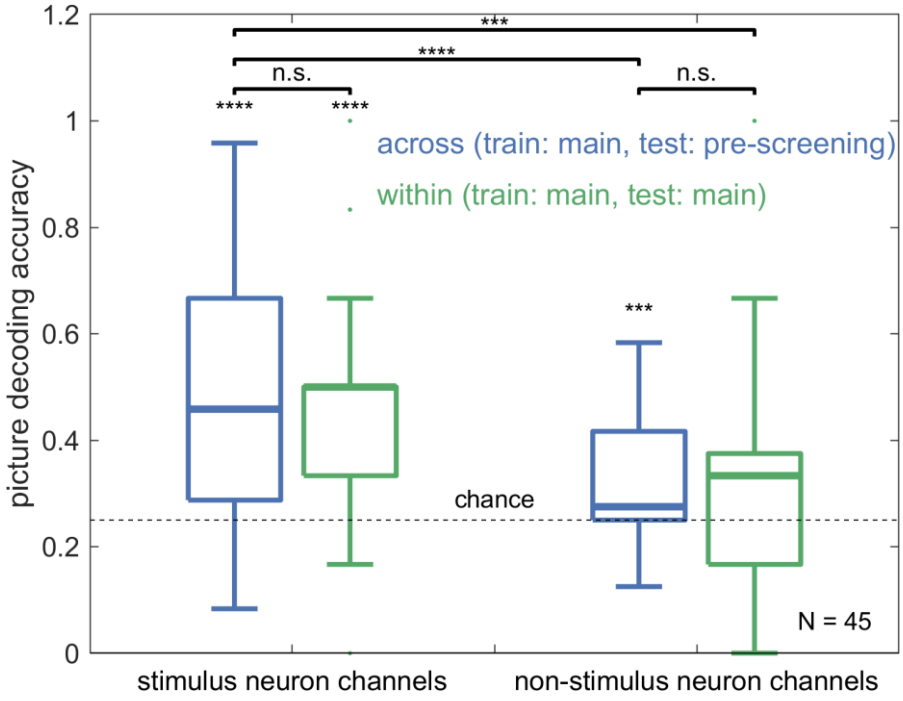


Fig. S7. Stimulus representations were maintained from pre-screenings to subsequent main experiments, particularly in micro-wires with stimulus neurons. Boxplots (Q1, median, Q3; whisker: points within ±1.5 IQR) of session-wise SVM picture decoding accuracies of the four pictures shown in the main experiment. SVM decoders were trained on main experiment activity and tested either on pre-screening (across in blue) or main experiment activity (within in green, leave-out cross-validation with matched trial counts). Dotted line in black denotes chance decoding of 0.25. For 45 recording sessions, all four pictures of the main experiment were part of the pre-screening before the experiment (N = 45). Picture decoding accuracies highly exceeded chance, particularly for micro-wire channels that contained stimulus neurons (left side; stimulus-across: 0.4854 +/- 0.2243, $p = 8.229 \times 10^{-7}$; stimulus-within: 0.4444 +/- 0.2072, $p = 1.017 \times 10^{-6}$) versus those that did not (right side; non-stimulus-across: 0.3241 +/- 0.1155, $p = 3.017 \times 10^{-4}$; non-stimulus-within: 0.2889 +/- 0.2114, $p = 3.728 \times 10^{-1}$). Importantly, there were no significant differences in decoding accuracy when testing on pre-screening data (across) or on main experiment activity (within), both for stimulus neuron channels ($p_{diff} = 0.301$) and for non-stimulus neuron channels ($p_{diff} = 0.231$; two-sided Wilcoxon signed-rank test). Overall, stimulus representations were maintained across the relatively short time intervals (4.71 hours +/- 2.02 SD, min: 1.40 h, max: 9.15 h) between pre-screenings and subsequent main experiments, specifically in channels that contained stimulus neurons. ****$p < 0.0001$; ***$p < 0.001$; **$p < 0.01$ (uncorrected).

Furthermore, the novel results are now described in the Methods section:

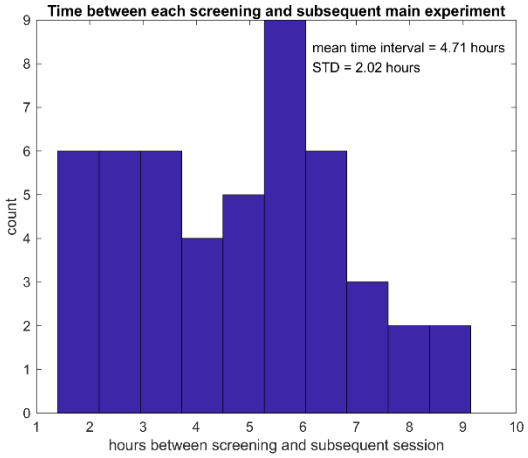
> "**Methods […]**
>
> **Definition of neural populations and assessment of their significance**
>
> […] In order to assess the stability of stimulus representations across pre-screenings and main experiments, Support vector machine (SVM) decoders were trained on main experiment activity and tested either on pre-screening (across) or main experiment activity (within, leave-out cross-validation with matched trial counts). For 45 recording sessions, all four pictures of the main experiment were part of the pre-screening before the experiment. Picture decoding accuracies highly exceeded chance, particularly for micro-wire channels that did versus did not contain stimulus neurons (Supplementary Figure 7). Importantly, there were no significant differences in decoding accuracy when testing on pre-screening data (across) or on main experiment activity (within), both for stimulus neuron channels and for non-stimulus neuron channels (two-sided Wilcoxon signed-rank test)."

How often was the pre-screening performed?

Exactly once before each of the recorded experimental sessions. We regard a pre-screening session as valid only for the same day on which it was performed. In other words: new recording day = new pre-screening.

What was the distribution of intervals between each pre-screening step and its respective experimental session(s)?

We thank the reviewer for this relevant question. The following figure depicts the distribution of time intervals between each pre-screening and its respective (unique and singular) experimental session. The mean time interval between screenings and subsequent experiments was 4.71 hours +/- 2.02 SD (range: 1.40 to 9.15 h).



We added the following sentences to the methods section:

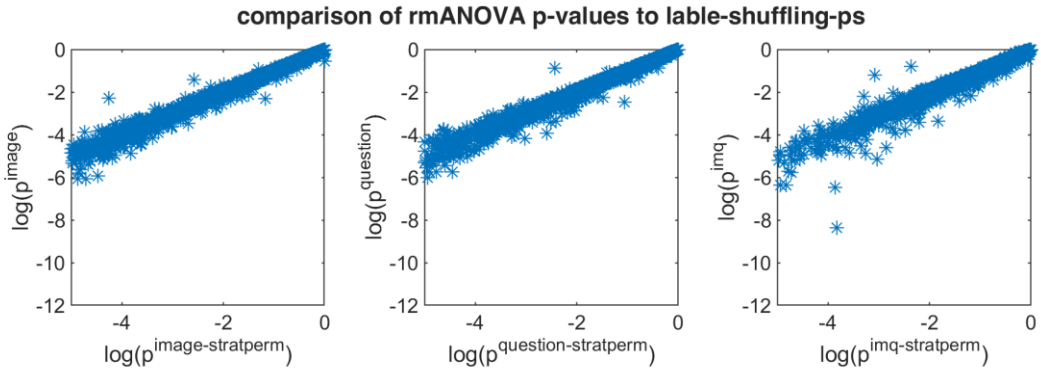
> "**Methods**
>
> **Subjects and Recording**
>
> […] Our original data set consisted of 50 experimental sessions from 17 subjects with recordings from the amygdala (A), parahippocampal cortex (PHC), entorhinal cortex (EC), and hippocampus (H). One session from one subject was excluded due to low behavioral performance, resulting in a final sample of 49 sessions from 16 subjects. Remarkably, the excluded session stemmed from the only subject for which no context neurons could be detected (others exhibited an average of 12.5). For most subjects, ten micro-wire bundles (median of 80 channels) were placed with an average of 69.34 recorded channels (range: 32 to 80) and a neuron yield of 0.92 per channel. **Each experimental session was recorded after a screening for visually selective responses in the morning of the same day. The mean time interval between screenings and subsequent experiments was 4.71 hours +/- 2.02 SD (range: 1.40 h to 9.15 h).**"

The only way to incorporate both features (i.e., a pre-screen methodology and assumption of uniqueness of units across all experimental sessions) would be for each experimental session to have its own pre-screen step that immediately preceded it (interval of minutes) and for there to be a relatively much longer interval between each pres-screen/experimental session pair (many hours to days). Given the constraints of the EMU paradigm, I doubt this was the case. I think the most likely situation is that pre-screening does work (as evidenced by the high number of content encoding neurons) because there is significant stability of unit identity across sessions.

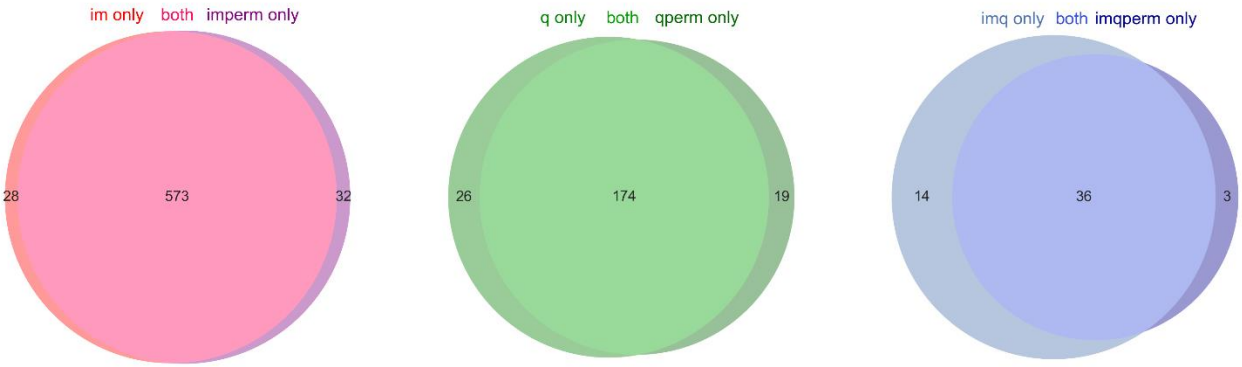
A novel pre-screening (typically with different stimuli) was performed on average 4.71 hours +/- 2.02 SD before each new recording session. Only one session was recorded per day and the next session took place on average 46.94 hours +/- 24.19 SD (range: 16.57 to 118.06 h) later (again with a new screening session a few hours before the experiment). See above for the demonstration of neural stability between pre-screenings and main experiments on the one hand (Supplementary Figure 7), and a high degree of neural independence between subsequent main experiments (Supplementary Figure 8) on the other hand.

Regarding general statistical tests, why was alpha = 0.001 chosen? Why not compute p-values for individual neurons using stratified label-shuffling? I'm not sure why statistics were computed for individual data instead of shuffle averaged data over sessions.

We thank the reviewer for this inquiry. Label shuffling of *p*-values for each individual neuron does not necessarily take care of statistical dependencies between (multiple recorded or) statistically dependent units within one recording session (at least if different permutations are employed for each neuron). However, we explicitly checked whether populations defined by permutation tests would differ. Namely, we re-computed *p*-values by determining the fraction of effect sizes (partial eta squared) from stratified label-shuffled controls (1,000 permutations, resampled with replacement to 10,000 values) that exceeded the observed effect size for each neuron. Thus obtained *p*-values strongly correlated with original *p*-values:



Furthermore, populations defined by these stratified label-shuffling controls largely overlapped with the populations from our manuscript:



Therefore, we did not re-compute our analyses but added the following sentence to the methods section:

> "**Methods**
>
> **Definition of neural populations and assessment of their significance**
>
> […] Importantly, neural populations remained largely consistent, regardless of whether they were identified via label-shuffled ANOVA effect sizes or analytical ANOVA p-values."

The conclusion of spike timing dependent plasticity is going too far out on a limb. I suggest softening the instances when this possibility is mentioned as an explanation for the EC-HC results. This mechanism can be listed as a hypothesis to test but should not be as closely associated with explanatory power for the current data given the absence of specific evidence. In fact, the authors should spell out the experiments that would need to be done (and are compatible to be done in humans) to test this hypothesis.

We thank the reviewer for carefully assessing spike timing dependent plasticity in the manuscript. While second-level and additional analyses were also consistent with potential synaptic modifications (e.g. Fig. 4a or Fig 5e-f), we softened the instances when this possibility is mentioned, and now present it as one possible hypothesis that could be strengthened by future experiments in the discussion section:

> "**Discussion**
> […]
> During and after, but not before the experiment, firing of entorhinal stimulus neurons predicted the activity of hippocampal context neurons after tens of milliseconds (Fig. 4a-c), consistent **for example** with the storage of stimulus-context associations via spike-timing-dependent plasticity (STDP).
> […]
> **Context representations across MTL regions**
> […]
> It is presumed that through a process called pattern completion, inputs from entorhinal cortex (representing content) are complemented in the hippocampal formation (thereby retrieving associated context)[25]. Supporting this view, during and after, but not before the experimental pairing of stimuli and context, firing of entorhinal stimulus neurons predicted the activity of hippocampal context neurons after tens of milliseconds~~**, consistent with synaptic modifications via STDP**~~. Picture presentations served both as cues for the recall of

content (from the current or the preceding picture) and context (relevant question). Correspondingly, representations of current visual stimuli were followed by re-instatement of previous stimuli or contexts by stimulus and context neurons, respectively. **Synaptic modifications of either direct or indirect connections between both populations via STDP offer one potential explanation for these findings that could be further strengthened by future experiments. For example, if only a subset of contents and contexts were to be paired, shift-corrected cross correlations should only change for those stimulus and context neurons that represent these pairings. Moreover, electrostimulation of hippocampal context neurons approximately 800 ms after picture onsets during context reactivation peaks following stimulus representations (see Fig. 3d) should interfere with cross correlation patterns, reinstatement patterns and behavior.**
Increased excitability of context neurons offers a potential gating mechanism of stimulus-driven pattern completion in hippocampus. Specifically, after the strengthening of multiple connections between stimulus and context neurons, increased excitability of hippocampal context neurons after pre-activation by their preferred context could account for their subsequent increased preferential firing following entorhinal stimulus-neuron activation. **Again, future experiments could validate this model by altering the excitability of context neurons via electrostimulation or modulation of attention during question presentations.**
In light of these considerations, representations and experiment-induced neural interactions between both neuronal populations hold the potential to contribute to memory of stimuli in context. Resulting co-activations in turn could guide contextual memory processing of stimuli, e.g. via indexing of brain areas that contain context-specific stimulus information."

Since stimuli were pre-selected independent of context, I suggest removing the stimulus-selectivity statistic from abstract.

We added a caveat to the abstract indicating that a pre-screening of stimulus neurons was performed:

"[…] In a context-dependent picture comparison task, co-firing of 597 (pre-screened) stimulus-modulated and 200 context-modulated neurons […]"

Moreover, we emphasized the resulting bias more in the results and discussion section:

"**Results**

Distinct populations encode stimulus and context in a picture comparison task

[…] While we expected to find many neurons with a significant main effect of stimulus due to our pre-selection of response-eliciting pictures **which artificially increased the number of stimulus neurons**, it was unclear how context would be encoded."

"**Discussion**

Context representations across MTL regions

[…] **Moreover, while stimulus neurons were overrepresented due to a pre-selection of response-eliciting pictures in a screening before the main experiment,** […]**"**

Specific Questions/Requests

Do any analyses include the response variable (whether the picture was presented 1st vs. 2nd)?

We thank the reviewer for inquiring about the influence of the serial picture positions as our previous terminology could have been more precise. Some analyses examined whether pictures were presented first or second. For example, in Figure 3b, we analyzed whether context or stimulus representations generalized across serial picture positions, i.e. whether decoders trained to decode context or stimulus during the first serial picture position could still decode context during the second serial picture position. We now speak of "serial picture positions" in the caption of Figure 3b and throughout the manuscript:

> "**b.** Boxplots of pooled context (green) or stimulus (red) decoding accuracies (SVM) from subsamples of context (left) and stimulus neurons (right), either obtained by 5-fold cross validation or generalization across pictures (red), contexts (green) and **serial** picture positions (blue)."

Similarly, Figure 3e shows that context-neuron activity during question presentations is reactivated both during first and second picture presentations.

Additionally, we computed analogous repeated-measures ANOVAs with a main effect of serial picture position (alpha=0.001). Only 14.74 % of stimulus neurons, 18% of context neurons and 12% of stimulus-context neurons exhibited a significant main effect of position, confirming earlier findings.

Please include fractions / %'s of cells in specific regions in abstract (instead of, or in addition to, absolute numbers. If keeping absolute numbers, please include population size).

We added both the population size and fractions in the abstract:

> "[…] In a context-dependent picture comparison task, co-firing of 597 (pre-screened) stimulus-modulated and 200 context-modulated neurons (total: 3109, amygdala: 2.95%, parahippocampal cortex: 7.68%, entorhinal cortex: 5.68%, hippocampus: 9.42%) from 16 neurosurgical patients combined different comparison questions (contexts) with two subsequent to-be-compared items (stimuli) through neural reinstatement of question contexts."

Was there always a correct answer for each question and combination of stimuli? Did participants answer the same question consistently across different trial repetitions?

We thank the reviewer for raising this important point. While some contexts required subjective judgement, and therefore did not allow for objective evaluation (e.g. "Last seen in real life?", "Like better?"), each context still required the assessment of a unique and distinct order among picture contents. Therefore, it was not possible to give consistent and transitive answers without both remembering the context (i.e. question) as well as the two subsequent pictures when making a decision. Furthermore, transitivity suggests that task instructions were actually adhered to. We now assess memory of contents in context via the consistency and transitivity of the subjects' answers. Their performance greatly exceeded chance and approached ceiling levels in all but one recording session which we now exclude from all analyses. The following figure depicts performance per recording session:

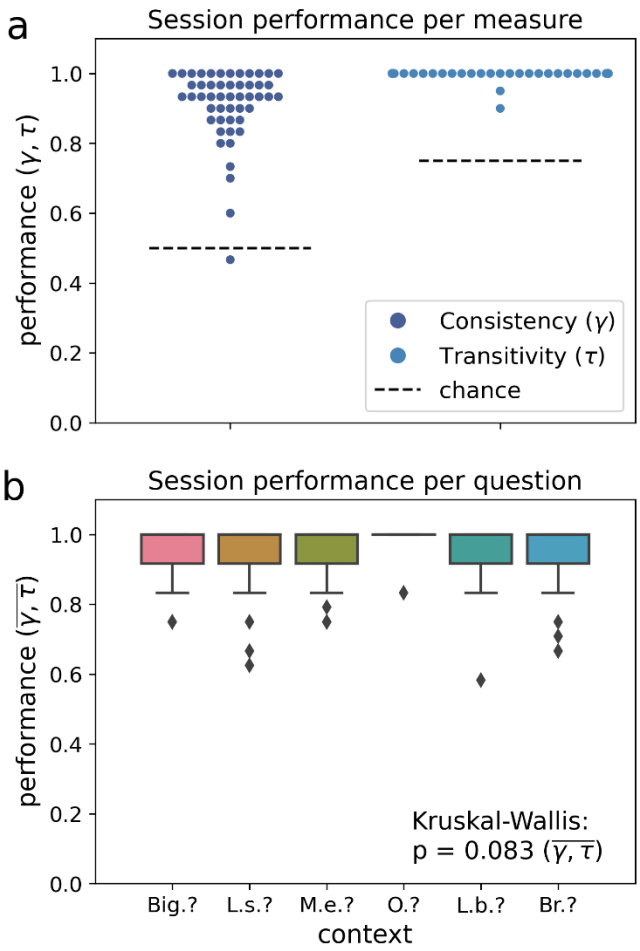


Fig S4. Consistency and transitivity of answers strongly exceeded chance and approached ceiling levels. Pairwise comparisons of contents were both consistent and transitive (a) across different contexts that imposed distinct orders (b), indicating memory of contents in context. Consistency measures how well context-specific preferences for one picture over another are maintained across presentation orders, while transitivity checks if a preference for A over B and B over C also results in a preference for A over C. **a.** Swarmplots depict consistency scores γ (dark blue, left) and transitivity scores τ (light blue, right) for each subject's answers in each recording session. Dotted lines represent chance performance. Only one recording sessions did not exceed chance levels and was excluded from further analyses. **b.** Boxplots (Q1, median, Q3; whisker: points within ±1.5 IQR) of mean session-wise behavioral performances (averages of consistency and transitivity) across contexts. Performance did not differ significantly between

We added the following sentence to the results section:

> "**Distinct populations encode stimulus and context in a picture comparison task**
>
> […] Subjects then chose the picture that best answered the question (e.g. which depicted something bigger) and indicated whether it was shown first or second by pressing keys 1 or 2 on the keyboard. **The majority of answers were highly consistent and transitive, with performance greatly exceeding chance in all but one excluded session and not varying across question contexts (*p* = 0.083; Kruskal-Wallis-Test; Supplementary Figure 4).** To quantify the representation of stimulus and context during picture presentations, repeated-measures two-way 4x5 ANOVAs with factors picture, question and their interaction were computed for each neuron at an alpha level of 0.001 (see Fig. 2e for distribution of effect sizes). "

The following paragraph was added to the Methods section:

"**Behavioral performance**

Memory of contents (i.e. pictures) in context (i.e. after a particular question) was evaluated by determining the consistency and transitivity of the answers that were provided at the end of each trial. Consistency (γ) captures the degree to which preference for one picture over another was maintained when the presentation order was reversed. A picture was considered to be preferred if it was chosen in at least 3 out of the 5 trials that it was paired with a specific different picture in a given temporal order and context. If this picture was preferred both when it was shown first or second (chosen ≥3/5 times in each order), all 10 trials were assigned a consistency value of 1 instead of 0. The overall consistency of a session was computed by averaging the consistency values of zeros and ones across all 300 trials. For random choices a consistency of 0.5 is expected (Supplementary Figure 4a).
Transitivity (τ) captures the logical consistency of preferences across three pictures (x, y, z) when considered in pairs. For each of the four unordered picture triples we computed preference probabilities P(x>y), P(y>z), and P(x>z) for all three unordered picture pairs based on how many times out of the 10 trials one picture was chosen over the other in a given context. A triple was considered transitive if preference probabilities satisfied the condition of weak stochastic transitivity: P(x>y) ≥ 0.5 and P(y>z) ≥ 0.5, then P(x>z) ≥ 0.5. This condition was tested for all permutations of x, y, and z, and a transitivity value of 1 instead of 0 was assigned to the triple if the condition was satisfied for any permutation. For random choices a transitivity of 0.75 is expected (Supplementary Figure 4a). Finally, context-dependence of performance was assessed by deriving an overall behavioral performance measure from mean of consistency (γ) and transitivity (τ) values per question of each session and comparing them across contexts (Kruskal-Wallis-Test; alpha=0.05; Supplementary Figure 4b)."

Moreover, the following sentence was added to the discussion:

"**Discussion**

Contexts constrain which memories are relevant for future decisions[12]. Both semantic content as well as context, i.e. other aspects of a study episode[1], such as tasks[9], episodes[11] or rules[10] are represented in single-neuron activity in the human MTL. It is currently unclear, however, whether and how content and context are combined to form or retrieve integrated item-in-context memories at the single-neuron level in humans. Pairs of pictures were presented on a laptop screen in different contexts, namely five questions that specified how the pictures were to be compared to each other (Fig. 1a). **Sixteen of seventeen subjects remembered these contents (i.e. pictures) in their respective context as they chose both consistently and transitively between pictures despite unique and distinct orders among picture contents in each context (see Supplementary Figure 4).** […]"

In the paragraph titled "Stimulus and context are mainly encoded separately and rarely conjunctively", did the ANOVA model used include the interaction term?

Yes. The ANOVA included the interaction term and we now state this in the paragraph mentioned by the reviewer in order to avoid confusion:

"**Results** […]

Stimulus and context are mainly encoded separately and rarely conjunctively

Context was represented during picture presentations in substantially more neurons than expected by chance. Out of 3109 recorded units, 200 exhibited a main effect of question $p < 0.001$ in a repeated measures two-way ANOVA with factors stimulus and question **(including their interaction**, Fig. 2a), largely exceeding the expected number of roughly 3 units (two-sided binomial test, $p < 2.225 \times 10^{-308}$)."

"Next, we assessed pooled decoding performances from 30 random subsamples of 151 context neurons or 151 stimulus neurons" --> Please clarify the purpose of the subsampling in this analysis?

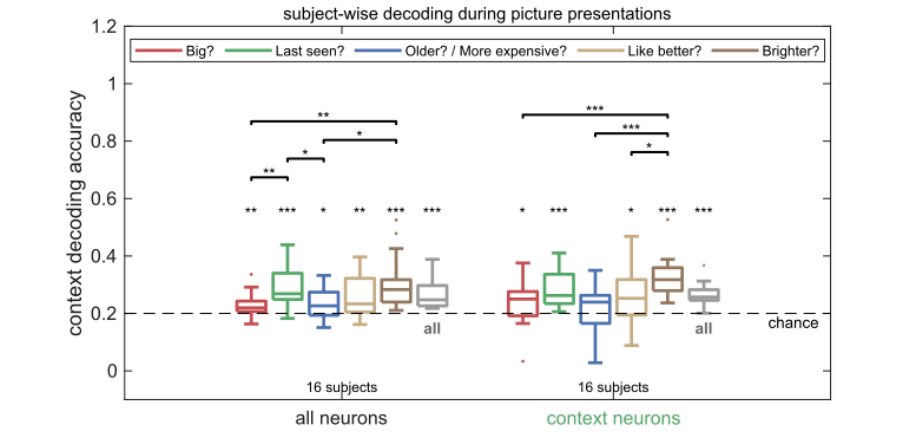
We thank the reviewer for catching this imprecision. Pooled decoding analyses were performed with neurons across all recording sessions. In order to be able to estimate the variance of these pooled decoders, 30 random subsamples of 75% of context neurons, i.e. 150 context neurons, and an equal number of stimulus neurons were randomly drawn 30 times, as reported elsewhere (e.g. see Minxha et al). This is now stated more clearly:

"**Results** […]

Next, we assessed decoding performances obtained by pooling neurons across all sessions. In order to estimate their variance, random subsamples of 75% of context neurons, i.e. 151 context neurons, and an equal number of stimulus neurons were drawn 30 times. Pooled analyses resulted in highly significant (two-sided Wilcoxon signed-rank test against 0.2 or 0.25, uncorrected) context (0.866, p = 1.708 × 10-6, green) as well as stimulus (0.995, p = 1.627 × 10-6, red) decoding accuracies (Fig. 3b left sub-columns on left or right). […]"

Please present the pooled decoding accuracy next to the per-session decoding accuracies in figure 3a? (using the same training / testing procedure)

Novel boxes (see "all", i.e. pooled across all questions, in gray) have been added to Fig. 3a:



Furthermore, the results have been updated:

> **"Generality and temporal dynamics of context representations imply relevance for stimulus processing**
>
> […] Confirming previous ANOVA results, decoding accuracies significantly exceeded chance levels of 0.2 (one out of five questions, two-sided Wilcoxon signed-rank test, uncorrected) for all contexts when all neurons were included (Bigger?: $p_{session} = 1.981\times 10^{-3}$, $p_{subject} = 4.094 \times 10^{-3}$; Last seen?: $p_{session} = 9.142 \times 10^{-7}$, $p_{subject} = 6.430 \times 10^{-4}$; Older / More expensive?: $p_{session} = 5.187 \times 10^{-3}$, $p_{subject} = 3.861 \times 10^{-2}$; Like Better?: $p_{session} = 1.456 \times 10^{-3}$, $p_{subject} = 6.624 \times 10^{-3}$; Brighter? : $p_{session} = 8.016 \times 10^{-8}$, $p_{subject} = 4.378 \times 10^{-4}$; **all contexts: $p_{session} = 4.378 \times 10^{-4}$, $p_{subject} = 4.378 \times 10^{-4}$**). Similar results were obtained when restricting these analyses to context neurons (Bigger?: $p_{session} = 1.713 \times 10^{-2}$, $p_{subject} = 4.938 \times 10^{-2}$; Last seen?: $p_{session} = 5.888 \times 10^{-5}$, $p_{subject} = 4.378 \times 10^{-4}$; Older / More expensive?: $p_{session} = 3.875 \times 10^{-2}$, $p_{subject} = 4.08 \times 10^{-1}$; Like Better?: $p_{session} = 8.126 \times 10^{-3}$, $p_{subject} = 4.373 \times 10^{-2}$; Brighter? : $p_{session} = 9.826 \times 10^{-7}$, $p_{subject} = 4.368 \times 10^{-4}$; **all contexts: $p_{session} = 4.368 \times 10^{-4}$, $p_{subject} = 4.368 \times 10^{-4}$**).

Can the authors clarify this sentence in statistical terms in light of their example: "we also identified contextual stimulus neurons with an additional main effect of context": neuron with significant main effect of stimulus and significant interaction between stimulus and context? --> no, neurons with main effects for both stimulus and context (clarified in methods). Please clarify in main text.

We now state this in the main text:

> **"Results**
> **Distinct populations encode stimulus and context in a picture comparison task**
> […]
> Lastly, while mere stimulus neurons (only significant main effect of stimulus) were much more common (Fig. 1b), we also identified contextual stimulus neurons defined by significant main effects of both stimulus and context (irrespective of the interaction term), e.g. selectively responding to the picture of a hamburger, yet particularly strongly when the context was "Last seen in real life?" (Fig 1e)."

Consider more distinguishing terminology for contextual stimulus neurons (main effect of context and stimulus) vs. stimulus-context neurons (interaction between stimulus and context).

We thank the reviewer for this suggestion. In the main text, stimulus-context neurons are now always referred to as stimulus-context interaction neurons for greater clarity.

The data availability statement needs links to data and code.

A GitHub repository with initial data and code has been added and will be completed in the course of the further review process.

Referee #1 (Remarks to the Author):

The authors have provided a substantial revision to their original manuscript, and have addressed many of the concerns raised in the initial round of reviews. Specifically, they have provided several additional analyses at the level of individual subjects, and have introduced a decoding analysis based on an SVM. All of these new additions strengthen the overall conclusions of the manuscript. The two main outstanding questions, though, are related to how to interpret these results.

We thank the reviewer for recognizing our substantial revision, for noting that our new analyses strengthen the overall conclusions, and for the valuable feedback provided throughout both rounds of review. We appreciate the reviewer's identification of these two main questions about interpretation, which we address below.

The **first question** is whether indeed this reflects a **context signal.** This was raised by the other reviewers as well. I appreciate the point that different contexts can demand different behaviors (e.g., considering the price of food in a store versus the taste of food at home), and the authors make this point to note that the demands of their behavioral task may be interpreted as a generalization of different contexts. This is an interesting point. However, there still remains the alternative explanation that what is being asked of the subjects is to specifically focus on or pay attention to one of the stimulus features.

As the authors note, this could indeed happen in different contexts, but the actual manipulation here is not of the contexts but instead of the task demands. In fact, one may posit that if a subject were in a supermarket and asked to either purchase an item or taste an item, both of those different demands would occur within the same context. It is unlikely in that case that the authors would argue that these different demands therefore represent different contexts.

I think this has some bearing on how to interpret and view these results, because it seems more likely that what we are observing here are neurons that are tuned to the particular task demands or stimulus features, which then would change the overall interpretation of the interaction between the encoding of content and in this case the encoding of the salient stimulus feature.

The reviewer makes an important distinction that maps onto established types of context. We now disentangle these in accordance with Baddeley (1982), distinguishing interactive/task contexts (which influence how contents are processed and encoded) from independent/background contexts (which may be encoded alongside contents without altering their processing). We agree that task demands can change within the same physical location and should be conceptually separated from place. Importantly, in the memory literature, both types constitute valid forms of context. Within Baddeley's framework, "purchase" versus "taste" would indeed constitute different interactive contexts even within the same supermarket, as they fundamentally change how the items are processed and encoded.

The Introduction now states that our experiment specifically investigates interactive (task) context and the "Operationalization of real-world contexts" section of the Discussion has been revised accordingly. It now also addresses the question of whether merely stimulus features are represented. Specifically, stimulus neurons represented picture contents rather than isolated features, as shown by consistent encoding across contexts (Fig. 3b) as well as reactivation of first-picture representations during second-picture presentations in all five questions (Supplementary Figure 11). Moreover, context neurons encoded the question (context) rather than transient stimulus features, distinguishing contexts before stimulus onset, reinstating them during picture viewing, and persisting until the decision (Fig. 3d–f; Fig. 5b,e; Supplementary Figure 6a,c). Together, these patterns are inconsistent with a mere feature-specific representation.

The introduction now reads:

> "Introduction
> [...]
> Recording from sixteen patients who underwent surgery to treat epilepsy, we investigated how single neurons in the human MTL combine neural representations of items with those of context. For this purpose, we devised a picture-comparison task in which pairs of pictures were presented on a laptop screen in different contexts, namely questions (e.g., "Bigger?") that specified how the pictures should

be compared to each other. The task required subjects to remember items (i.e., two pictures) in a specific context. **Here, "context" refers to interactive or task context rather than independent or background context (Baddeley, 1982)[12].**"

Additionally, the Discussion has been updated to:

"**Discussion**
[...]
Operationalization of real-world contexts
Real-world contexts include both independent elements (encoded alongside contents without altering processing) and interactive elements that influence how contents are processed (Baddeley, 1982)[12]. For example, evaluating a melon's price in a supermarket versus evaluating its taste in the kitchen at home are examples of interactive context effects. Our task operationalizes such interactive contexts, as questions in our task define comparison rules that shape stimulus processing (e.g. see Polyn, Norman, & Kahana, 2009)[14]. Behaviorally, the task captured these interactive context effects as pictures were ranked differently across questions (median Kendall's W = 0.232) yet consistently and transitively within questions, tracking objective ground truths (mean ρ = 0.760; Supplementary Fig. 10). This confirms the questions functioned as interactive contexts rather than eliciting fixed stimulus values. Our neural data suggests that we successfully captured both interactive contexts and the contents that they acted on. First, stimulus neurons represented picture contents rather than isolated features, as evidenced by consistent encoding across contexts (Fig. 3b) and reactivation of first-picture representations during second-picture presentations for all questions (Supplementary Figure 11), not only for a feature-relevant question as expected for a feature-specific encoding. Second, context neurons represented interactive context rather than transient attention effects by distinguishing questions before stimulus onset (Fig. 3d; Fig. 5b; Supplementary Figure 6a,c), reinstating representations during picture viewing (Fig. 3e, Fig. 5e), and persisting until the decision (Fig. 3f). Supporting their functional relevance, both context and stimulus decoding were higher on behaviorally correct trials (Supplementary Figure 12). Overall, the task design captures how interactive context shapes memory processing in the human MTL. Nonetheless, future studies should also examine these mechanisms under independent context manipulations and explore them in naturalistic settings."

This relates to the **second question**, which is whether this interaction is related to memory. Here the authors do not provide an explicit memory test, and yet their conclusions are based on the interpretation that the integration of content and context is important for memory. In their revision, the authors introduce a new analysis in which they examine the consistency and transitivity of the behavioral preferences. This is important, of course, and as the authors note, it would be difficulty to be consistent in your answers if you do not remember the order of the stimuli that are presented. But this does not feel like an explicit test of memory, let alone a memory that requires the integration of content and context. In fact, if one were to just provide a behavioral response to an image and asked to provide a judgement, one would expect that the same image would evoke the same judgement. This is consistent (and would also exhibit transitivity since other objects would also have consistent judgements), but this does not seem to involve memory.

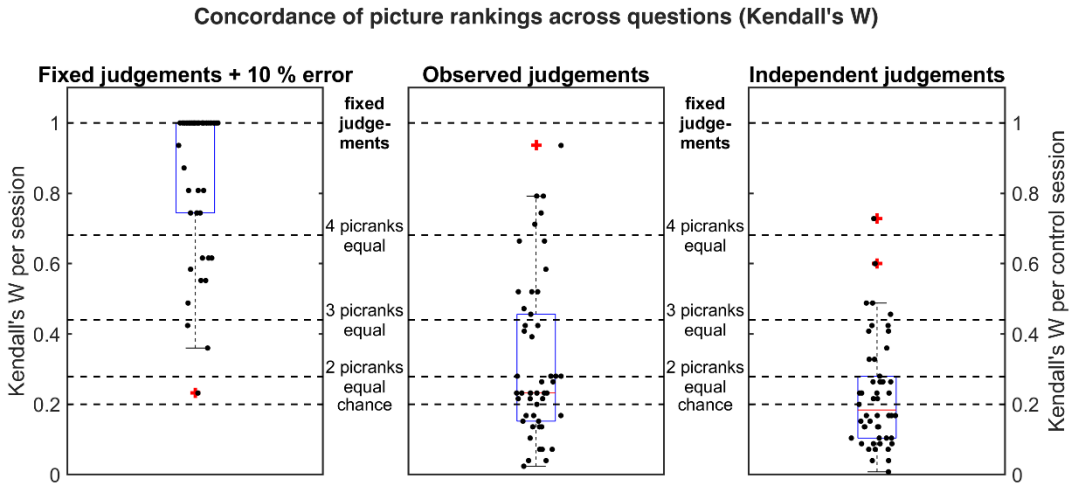
We thank the reviewer for this and the following important points, which have led us to perform multiple new behavioral and neural analyses (including an explicit memory test and evidence for memory of contents rather than features; see also next points) that strongly support memory of contents in context rather than simple value judgments.

While fixed value judgements of individual pictures could theoretically result in consistent and transitive answers, they would also result in highly similar rankings of pictures (inferred from pairwise choices) across all questions. Our data strongly argue against this fixed-judgement account. Subjects ranked pictures very differently across questions, yet their responses remained consistent and transitive within each question.

This differential ranking of pictures across questions can be demonstrated by computing Kendall's coefficient of concordance, W, for the picture rankings across different questions for each session. Fixed value judgments would predict a median Kendall's W of 1.0. In contrast, our behavioral data yielded a median W of 0.232, a value

close to the theoretical chance level of 0.2 and not significantly different from chance baselines (p=0.0575, Wilcoxon rank-sum test).

These low session-wise Kendall's W values (see below) demonstrate that picture ranks varied substantially across questions, which is incompatible with a fixed judgement of each picture irrespective of the question:



Reviewer Figure R1.1. Concordances of picture rankings (Kendall's W) across questions for each session. Observed data (middle) are compared against two models (left and right): **Left (Fixed-judgment model):** Expected concordance if picture judgments were fixed across all questions, resulting in near-perfect agreement (median W=1.0 despite 10% permuted picture ranks to approximate observed error rates). **Middle (Observed judgement data):** Actual concordance from subject behavior, which is low (median W=0.232). **Right (Chance model of independent judgements):** Concordance expected if picture ranks were completely independent for each question (median W=0.184, obtained via random permutation of picture ranks). Horizontal dashed lines mark reference values for perfect agreement (W=1.0) and scenarios with partial agreement across questions.

The observed data in the middle panel of Reviewer Figure R1.1. clearly align better with the chance model rather than the fixed-judgment model. Observed picture rank concordances exceeded chance levels only slightly (as expected as picture rankings might sometimes be correlated among different questions), yet remained far below the values near 1.0 predicted by fixed-judgment model. The observed concordance was not statistically different from completely independent judgments (W=0.232 vs 0.184; Wilcoxon rank-sum, two-sided, p = 0.0575). By contrast, observed concordances differed dramatically from the fixed-judgment model (W = 0.232 vs 1; Wilcoxon rank-sum, $p = 3.7 \times 10^{-15}$).

This indicates that subjects adapted their ratings to the current question, which in turn suggests that they remembered the specific question asked during that trial, i.e., context, at the time of evaluation. Consequently, at picture presentations, subjects needed to retrieve the question to respond differentially while remaining consistent and transitive.

Overall, the behavioral performance therefore implies memory of the specific question context and picture information (see also evidence below that the latter constitutes content-based rather than feature-based memory).

The memory component here seems to be the fact that the subjects remember which picture is which when making their judgement.

We agree that remembering which picture is which is a memory component, but this alone is insufficient to explain subjects' performance. The same picture pairs frequently elicited opposite responses depending on the question asked (e.g., when comparing pictures of an elephant and a butterfly, choosing elephant for 'Bigger?' but butterfly for 'Like better?'). This context-dependent reversal cannot be explained by memory for picture identities alone, as shown by near-chance picture rank concordance across questions (median Kendall's W =

0.232, p = 0.0575 vs. chance, versus W≈1.0 expected if subjects relied solely on picture identity). Successful performance thus requires memory for both stimulus identity and the current question context.

This could, as the authors argue, mean that they are remembering the specific item and the question being asked. However, another possible behavioral strategy would be to pay attention to a specific stimulus feature, and then when viewing the images note the feature of the first, and whether the second is bigger or smaller, for example. This does not seem like an integration of context and content in the traditional sense.

The reviewer raises an important distinction between remembering specific picture features versus more abstract picture contents in their entirety. If performance relied only on specific stimulus features (e.g., size), one would expect a reactivation of first-stimulus-related activity during second-picture presentations only for the corresponding question (e.g., "Bigger?"), but not for other questions.

However, we find that neural representations of first pictures were reactivated during the late phase of second-picture presentations, i.e., precisely when they were to be compared to the first pictures, and this occurred across all five question contexts, highlighting that reactivations reflected content-based rather than feature-based picture representations. The following new Supplementary Figure depicts the novel findings:

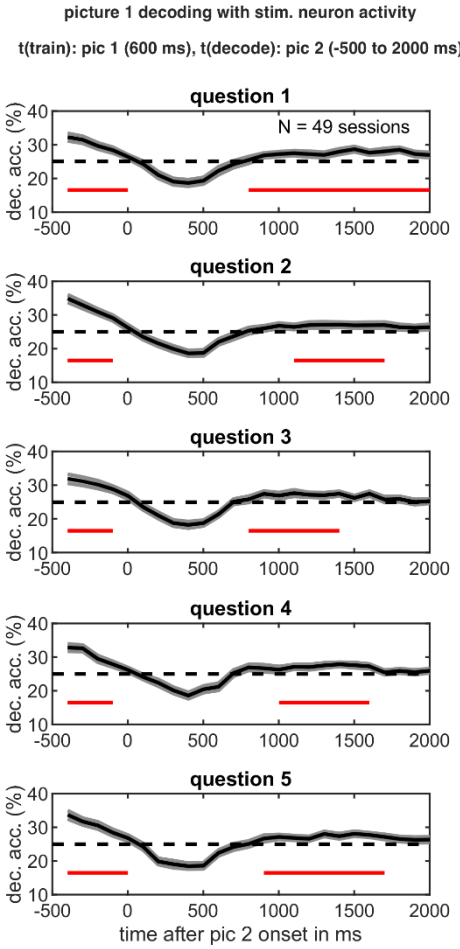


Fig. S11. Stimulus neurons recapitulate first-picture activity during second-picture presentations in all question contexts, indicating content-based rather than feature-specific encoding. Session-wise decoding accuracies of first stimuli during second-picture presentations in all five question contexts using SVM decoders trained with stimulus neuron activity ~ 600 ms after first-picture onset (400-800 ms bin). Solid lines and shaded areas depict means and standard errors, respectively. Data are compared to chance (25%, dashed horizontal black line). Solid lines in red below indicate significant differences (cluster permutation test, two-sided, $p < 0.05$). During second-picture presentations, first-picture decoding accuracies exceeded chance beginning ~ 600-1000 ms after second-picture onsets in all question contexts, arguing against a mere recollection of stimulus features.

Combined with the parallel reactivation of question contexts (Fig. 3d-e), these reactivation patterns support memory of integrated content-context information rather than selective, feature-specific comparisons.

We summarized the novel findings in the Results, Discussion and Methods sections, respectively:

> "Results […]
>
> Generality and temporal dynamics of context representations imply relevance for stimulus processing
>
> […] Figure 3d depicts subject-wise context decoding accuracies of context neurons (top, green) for identical training and decoding times (corresponding to the green diagonal line in Fig. 3c) together with stimulus decoding accuracies of pictures 1 and 2 by stimulus neurons (bottom, gray and lavender). Context decoding accuracies (Fig. 3d; top) exceeded chance for the entire duration of the trial (solid line in green, two-sided cluster permutation test; $p < 0.01$), even when compared to controls with shuffled question labels (red line; two-sided cluster permutation test $p < 0.01$). In parallel, stimulus neurons significantly encoded currently depicted pictures with high accuracies (Fig. 3d; bottom), as well as previous or even upcoming pictures (see solid lines in gray and lavender). The latter was possible since second pictures could be inferred stochastically by design (probability of 1/3 instead of 1/4 due to no repetition). **Notably, neural representations of first pictures were reactivated during second-picture presentations in all five question contexts (Supplementary Fig. 11), confirming that contents rather than question-related features were represented by stimulus neurons.** […]"

> "Discussion
>
> […] In our study, MTL neurons exhibited a relatively low, but detectable degree of pattern separation (stimulus-context interaction neurons in hippocampus) whose structure could account for both generalization and differentiation of contents and contexts. In agreement with Minxha et al.[9], stimulus neurons reliably distinguished picture identities irrespective of context. **Notably, stimulus-neuron representations reflected picture contents rather than question-related features, as evidenced by consistent first-picture reactivations during second-picture presentations across all question contexts.** In addition, we found neurons that represented context irrespective of picture identity or position. The invariance of both populations appears suited to generalize memories across contents and contexts, respectively […]."

> "Methods […]
>
> Support vector machine population decoding
>
> […] Decoding accuracies of either first (Fig. 3d bottom in gray) or second pictures (Fig. 3d bottom in lavender) obtained by subject-wise training and decoding with the activity of stimulus neurons in the course of the experiment were plotted and compared to chance (1/4) in an analogous manner (Fig. 3d, bottom). **Additionally, we examined whether neural activity patterns of stimulus neurons reflected content rather than feature representations (Supplementary Figure 11). Session-wise linear SVMs were trained on stimulus-neuron activity (400-800 ms after first-picture onset) to decode first picture identities from neural activity throughout second picture presentations (400 ms bins, stride = 100 ms) for each of the five question contexts. Statistical significance across sessions was assessed using cluster-based, two-sided permutation tests against chance ($\alpha = 0.05$, chance = 1/4).** […]"

Without an explicit test of memory, for example remembering the occurrence of a particular event or stimulus within a particular context, it's not clear how these data here would provide evidence that such interactions are

related to memory encoding. This is of course confounded by the point that these interactions may not truly reflect context and content, but instead some stimulus feature and the stimulus content.

In addition to the aforementioned evidence for memory of contents (rather than features) and for encoding of contents in context, we also added a more explicit behavioral analysis of memory as suggested by the reviewer.

Specifically, we assessed subjects' memory performance more explicitly by correlating, for each session and question, the behaviorally determined picture ranks with the corresponding objective ground truth

We used Spearman's rank correlation for questions with determinable objective answers ("Bigger?" and "More expensive?"/"Older?") to quantify picture rank agreement, excluding trials where objective ranks could not be determined. The question "Brighter image?" was also excluded from these analyses as perceived brightness can deviate from physical luminance due to contextual factors and contrast effects. For each session, correlation values were averaged across the available objective questions.

As shown in Supplementary Figure 10 (see below), rank agreement with objective ground truth was strong (mean Spearman's $\rho = 0.760$) and highly significant. All N = 50 sessions showed positive correlations (two-sided Wilcoxon signed-rank test against zero, $p = 6.47 \times 10^{-10}$), and the observed mean correlation significantly exceeded chance levels determined by permuting picture ranks (1000 permutations; two-sided, $p < 0.001$).

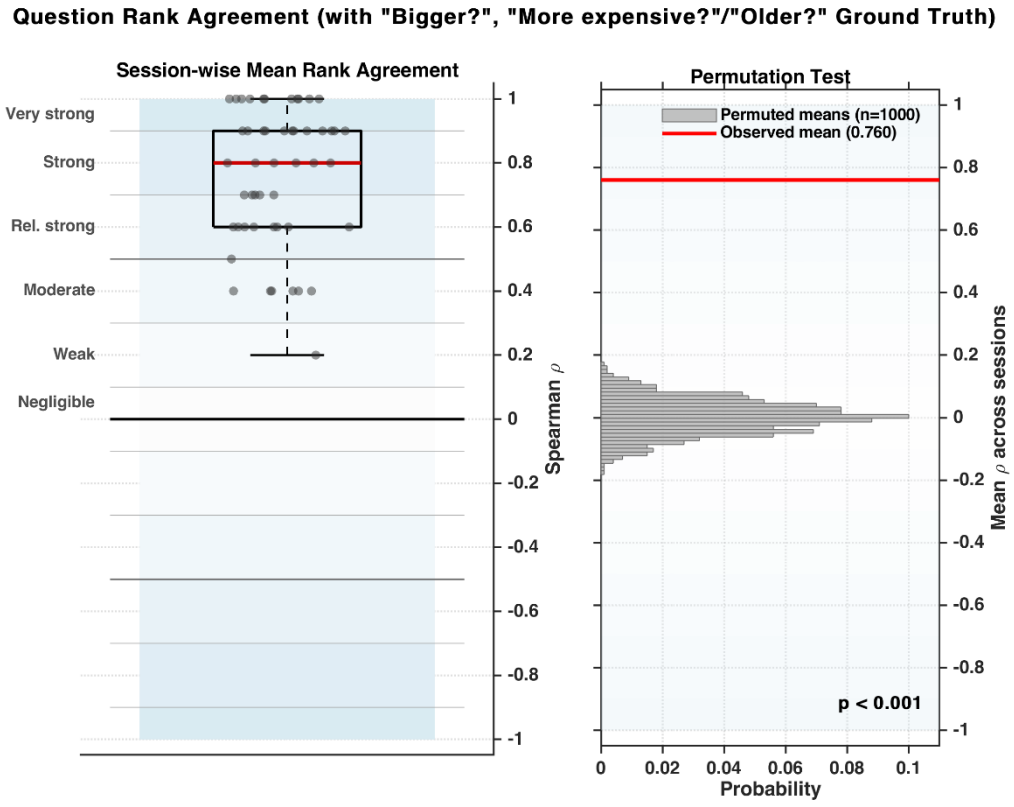


Fig. S10. Session-wise rank agreement with objective ground truth validates successful task performance. **Left:** Boxplot (Q1, median, Q3; whiskers: points within ±1.5 IQR) of session-wise mean Spearman rank correlations between behaviorally determined picture ranks and objectively determined picture ranks for the questions "Bigger?" and "More expensive?"/"Older?". Grey dots represent individual sessions (N=50). All session means were significantly greater than zero ($p = 6.47 \times 10^{-10}$, Wilcoxon signed-rank test). Shaded background areas indicate effect size categories. **Right:** Null distribution of mean correlations across sessions from 1,000 permutations (grey histogram), where picture ranks were randomly shuffled within each session-question pair before averaging. Observed mean shown in red ($\rho = 0.760$). The observed mean significantly exceeded the permutation null distribution ($p < 0.001$).

We summarized the novel findings via the following new text passages in the Results, Discussion and Methods sections, respectively:

> "Results […]
>
> Distinct populations encode stimulus and context in a picture-comparison task
>
> […] The majority of answers were highly consistent and transitive, with performance greatly exceeding chance in all but one excluded session, with these performance metrics not differing across question contexts ($p = 0.083$; Kruskal-Wallis-Test; Supplementary Figure 4). **Furthermore, answers strongly correlated with ground truths for "Bigger?" and "More expensive?"/"Older?" (mean $\rho = 0.760$, $p = 6.47 \times 10^{-10}$; Supplementary Figure 10).**"
>
> "Discussion
>
> Contexts constrain which memories are relevant for future decisions[13]. Both semantic content as well as context, i.e., other aspects of a study episode[1], such as tasks[9], episodes[11] or rules[10] are represented in single-neuron activity in the human MTL. It is currently unclear, however, whether and how content and context are combined to form or retrieve integrated item-in-context memories at the single-neuron level in humans. Pairs of pictures were presented on a laptop screen in different contexts, namely five questions that specified how the pictures were to be compared to each other (Fig. 1a). Sixteen of seventeen subjects remembered these contents (i.e., pictures) in their respective context as they chose both consistently and transitively between pictures despite unique and distinct orders among picture contents in each context (see Supplementary Figure 4). **Moreover, their rankings strongly matched ground truths (Supplementary Figure 10).**"
>
> "Methods […]
>
> Behavioral performance
>
> […] Finally, context-dependence of performance was assessed by deriving an overall behavioral performance measure from mean of consistency (γ) and transitivity (τ) values per question of each session and comparing them across contexts (Kruskal-Wallis-Test, $\alpha = 0.05$; Supplementary Figure 4b). **Moreover, for questions with objective ground truths ("Bigger?" and "More expensive?"/"Older?"), we validated behavioral performance by computing Spearman correlations between behaviorally determined picture ranks and objectively determined ranks, averaged across available questions per session. Session-wise correlations were tested against zero using a Wilcoxon signed-rank test, and the observed mean across sessions was compared to a null distribution generated from 1,000 permutations of picture ranks within each session-question pair (Supplementary Figure 10).**"

Finally, we related this objective behavioral measure to neural representations of content and context, by stimulus and context neurons, respectively (Supplementary Fig. S12). During picture presentations, picture decoding by stimulus neurons and question decoding by context neurons were significantly higher for trials in which choices matched objective ground truth than for trials that did not. This links the neural signatures to verifiable, context-appropriate memory. Together with content-based representations, reactivation of both content and context, and low cross-question concordance of picture ranks, these converging lines of evidence demonstrate memory encoding and retrieval of integrated content-context information, not merely fixed-value judgments or feature-specific attention.

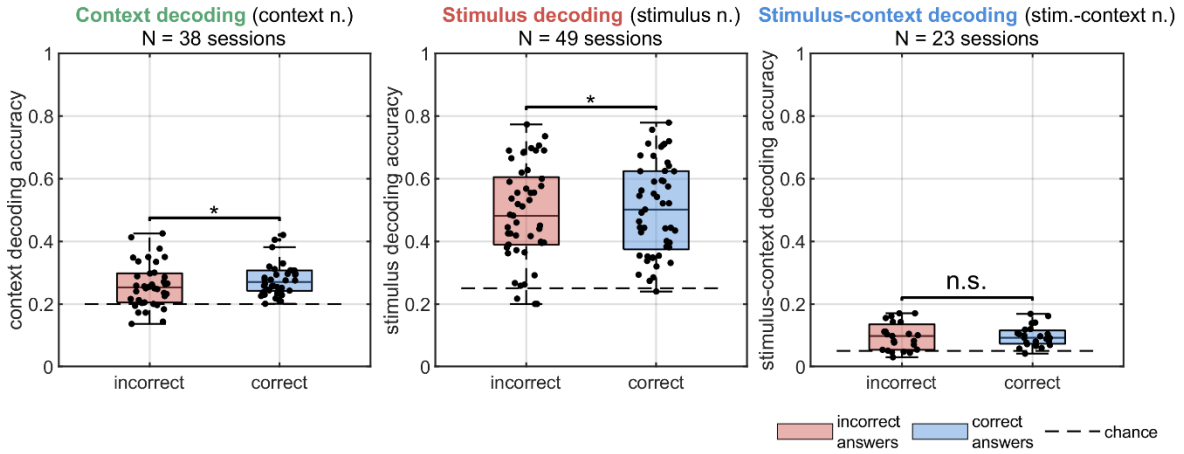


Fig. S12. Decoding of stimulus and context is enhanced when behavioral answers correspond to ground truth. Boxplot (Q1, median, Q3; whiskers: points within ±1.5 IQR) of session-wise mean decoding accuracies during picture presentations for trials with incorrect (red) versus correct (blue) answers relative to objective ground truth. **Left:** context decoding from context neurons (N = 38; chance = 1/5). **Middle:** stimulus (picture) decoding from stimulus neurons (N = 49; chance = 1/4). **Right:** stimulus-context (picture and question) decoding from stimulus-context neurons (N = 23; chance = 1/20). Dashed lines indicate chance levels. Decoding accuracies were significantly higher for correct versus incorrect trials for context ($p = 0.0128$) and stimulus neurons ($p = 0.0351$), while stimulus-context neurons showed no reliable difference ($p = 0.4098$, n.s.). *$p < 0.05$ (uncorrected, Wilcoxon signed-rank test, one-sided, correct > incorrect).

The novel findings are now described in the Results, Discussion and Methods sections, respectively:

> "**Results [...]**
>
> **Generality and temporal dynamics of context representations imply relevance for stimulus processing**
>
> [...] Finally, we tested whether content and context representations of stimulus and context neurons persist until the end of the trial when subjects chose one picture over another according to context (Fig. 3f). Both context as well as picture contents could be decoded above chance until a decision was indicated (two-sided cluster permutation test $p < 0.01$), **with significantly enhanced decoding when choices matched objective ground truth (context: $p = 0.0128$; stimulus: $p = 0.0351$, one-sided Wilcoxon signed-rank test, Supplementary Figure 12).**"
>
> "**Methods [...]**
>
> **Behavioral performance**
>
> **[...]**
>
> Finally, context-dependence of performance was assessed by deriving an overall behavioral performance measure from mean of consistency (γ) and transitivity (τ) values per question of each session and comparing them across contexts (Kruskal-Wallis test, $\alpha = 0.05$; Supplementary Figure 4b). Moreover, for questions with objective ground truth ("Bigger?" and "More expensive?"/"Older?"), we validated behavioral performance by computing Spearman correlations between behaviorally determined picture ranks and objectively determined picture ranks, averaged across available questions per session. Session-wise correlations were tested against zero using a Wilcoxon signed-rank test, and the observed mean across sessions was compared to a null distribution generated from 1,000 permutations of picture ranks within each session-question pair (Supplementary Figure 10). **Additionally, decoding accuracies were related to behavioral performance by comparing correct trials (where behavioral choices matched ground-truth picture rankings) versus incorrect trials (where choices deviated from ground truth). Session-wise linear SVMs (5-fold cross-validation, 5 repeats) were trained on context, stimulus, or stimulus-context neuron activity during picture presentations (100-1000 ms) to decode questions, pictures, and picture-question combinations, respectively. Decoding accuracies for**

correct versus incorrect trials were compared using one-sided Wilcoxon signed-rank tests (correct > incorrect; $\alpha = 0.05$). Chance levels were 1/4 for stimulus, 1/5 for context, and 1/20 for stimulus-context decoding (Supplementary Figure 12).

"Discussion […]

[…] During picture presentations when questions and picture contents became task-relevant, context neurons encoded associated question context, and stimulus neurons encoded picture contents (Fig. 3). Importantly, both context and content representations persisted until the end of the trial when one picture was chosen over the other according to the given context (Fig. 3f) and were enhanced during behaviorally correct trials (Supplementary Figure 12)."

To be clear, the authors have indeed provided a very thorough revision to the original comments, and there are several valuable aspects of this study that would be of broad interest to a large community of researchers. Certainly the analyses of the single unit data are sophisticated and well done, and the new additional analyses on decoding and examining correlations and activity at the level of individual subjects are important.

We sincerely thank the reviewer for acknowledging our thorough revision, the valuable aspects of this study that would be of broad interest to a large community of researchers, and our sophisticated single unit analyses. The reviewer's recognition of our new decoding analyses and subject-level correlations is much appreciated.

The question of how to interpret the data and the conclusion regarding context and content processing in the service of memory is less clear.

We believe our revisions address this concern. We have embedded our findings within the theoretical framework of interactive context (Baddeley, 1982), ruled out alternative interpretations of feature-specific attention and fixed judgments, and demonstrated behavioral-neural correspondence with objective ground truth. Together, these additions establish that our findings reveal fundamental mechanisms of flexible, context-dependent memory processing in the human MTL.

Referee #2 (Remarks to the Author):

The authors have done a good job in addressing my methodological concerns and have been attentive to some of the conceptual/theoretical issues that were raised as well as the need to discuss limitations to the study.

We thank the reviewer for their positive assessment and for their valuable feedback throughout the review process.

I have only two further comments:

i) The authors still do a poor job in how they handle the notion of a 'context'. They now argue in the discussion, unconvincingly in my view, that 'real-word' contexts invariably impact how stimulus content is processed, as exemplified by their melon/supermarket, melon/kitchen example. But as their own example at the beginning of the paper (same friend, two different restaurants) attests, this is only true of 'interactive' contexts; 'independent' or 'background' contexts (such as the restaurant) do not impact the processing of the the associated stimulus content, at least not to anything like the same degree. There is a rich literature on this distinction in the experimental psychology of memory, see for example, Baddeley, A. D. (1982). Domains of recollection. Psychological Review, 89(6), 708–729. I suggest that the authors acknowledge this distinction, and simply make it clear that their experiment is concerned with the neural representation of interactive (task) context, rather than making a somewhat spurious claim for the ecological validity for their chosen contextual manipulation.

We greatly appreciate the reviewer directing us towards Baddeley's (1982) important distinction between interactive and independent/background contexts, which has improved our manuscript's theoretical framing.

We fully agree that not all real-world contexts impact stimulus processing equally and have removed passages that could imply otherwise. Following the reviewer's suggestion, we now explicitly acknowledge the interactive vs. independent distinction and now clarify already in the Introduction that our experiment specifically investigates interactive (task) context.

Furthermore, we revised the Discussion accordingly, with a thoroughly rewritten "Operationalization of real-world contexts" section that places our work within this framework.

Additionally, we acknowledge that future studies should investigate these neural populations under independent context manipulations to provide a complete picture of contextual processing in the MTL.

The introduction now reads:

> "Introduction
> […]
> Recording from sixteen patients who underwent surgery to treat epilepsy, we investigated how single neurons in the human MTL combine neural representations of items with those of context. For this purpose, we devised a picture-comparison task in which pairs of pictures were presented on a laptop screen in different contexts, namely questions (e.g., "Bigger?") that specified how the pictures should be compared to each other. The task required subjects to remember items (i.e., two pictures) in a specific context. **Here, "context" refers to interactive or task context rather than independent or background context (Baddeley, 1982)[12].**"

Additionally, the Discussion has been updated to:

> "Discussion
> [...]
> Operationalization of real-world contexts
> Real-world contexts include both independent elements (encoded alongside contents without altering processing) and interactive elements that influence how contents are processed (Baddeley, 1982)[12]. For example, evaluating a melon's price in a supermarket versus evaluating its taste in the kitchen at home are examples of interactive context effects. Our task operationalizes such interactive contexts, as questions in our task define comparison rules that shape stimulus processing (e.g. see Polyn, Norman, & Kahana, 2009)[14]. Behaviorally, the task captured these interactive context effects as pictures were ranked differently across questions (median Kendall's W = 0.232) yet consistently and transitively within questions, tracking objective ground truths (mean ρ = 0.760; Supplementary Fig. 10). This confirms the questions functioned as interactive contexts rather than eliciting fixed stimulus values. Our neural data suggests that we successfully captured both interactive contexts and the contents that they acted on. First, stimulus neurons represented picture contents rather than isolated features, as evidenced by consistent encoding across contexts (Fig. 3b) and reactivation of first-picture representations during second-picture presentations for all questions (Supplementary Figure 11), not only for a feature-relevant question as expected for a feature-specific encoding. Second, context neurons represented interactive context rather than transient attention effects by distinguishing questions before stimulus onset (Fig. 3d; Fig. 5b; Supplementary Figure 6a,c), reinstating representations during picture viewing (Fig. 3e, Fig. 5e), and persisting until the decision (Fig. 3f). Supporting their functional relevance, both context and stimulus decoding were higher on behaviorally correct trials (Supplementary Figure 12). Overall, the task design captures how interactive context shapes memory processing in the human MTL. Nonetheless, future studies should also examine these mechanisms under independent context manipulations and explore them in naturalistic settings."

ii) The abstract is far from clear,and appears to be missing a sentence (lines 17-22 cannot be parsed, at least in my case).

Thanks for catching this. We restructured the problematic lines 17-22 into two clearer sentences and made minor edits throughout the abstract to improve overall clarity.

> "**Abstract:** The medial temporal lobe (MTL), and particularly hippocampus, has been proposed to encode items in context. While hippocampal memory representations are largely context-dependent in rodents, concept cells in humans appear to be context-invariant. It is currently unclear whether and by which mechanism neural representations of item memory and context memory are combined to form or retrieve integrated item-in-context memories at the single-neuron level in humans. In a context-dependent picture-comparison task, **we recorded 3109 neurons from 16 neurosurgical patients, identifying 597 stimulus-modulated (pre-screened) and 200 context-modulated neurons (amygdala: 2.95%, parahippocampal cortex: 7.68%, entorhinal cortex: 5.68%, hippocampus: 9.42%). Their co-firing** combined different comparison questions (contexts) with **two subsequent pictures** (stimuli) through neural reinstatement of question contexts. Both populations were largely separate, generalized across each other's preferred dimension**, and covaried with behavioral peformance**. **Following experimental pairings** of stimuli and context, firing of entorhinal stimulus neurons predicted that of hippocampal context neurons after tens of milliseconds. Overall, synaptic modifications and co-firing of stimulus and context neurons could contribute to memory of items in their respective context, specify which stimulus memories need to be retrieved, and even generalize memories through mutual reinstatement of largely separate, orthogonal representations. In contrast, only 50 stimulus-context interaction neurons were **modulated by specific picture-question pairs, consistent with limited pattern separation in the human MTL and** potentially indexing stimuli-in-context or their attributes **more** explicitly."

Referee #3 (Remarks to the Author):

Many points from the previous round of revision have been answered, including bias towards stimulus selective units, pooling of data, unit yield, and others.

We greatly appreciate the reviewer's positive assessment of our revisions as well as the guidance on stimulus unit bias, data pooling, and unit yield that have strengthened our manuscript.

I have a few lingering concerns:

1. Regarding stimulus-context neurons and their relationship to stimulus or context neurons. It is understandable that the number of these neurons is too low to perform cross correlations, but they represent such an interesting subset that perhaps the authors could discuss them a bit further. How is the activity of these neurons related to that of context and content neurons in the different brain regions past an assessment of cross correlations? For example, were these neurons more likely to be found in areas where there were higher proportions of stimulus or context neurons?

We thank the reviewer for this insightful suggestion of how to explore the relationship between stimulus-context neurons and stimulus or context neurons. To address whether stimulus-context neurons are more prevalent in areas with higher proportions of stimulus or context neurons, we performed correlation analyses across sessions and brain regions.

We computed Spearman correlations between the proportion of stimulus-context neurons and the proportions of stimulus or context neurons across all session-site combinations (n=168). To assess regional specificity, we then performed site-specific analyses using permutation tests (1000 permutations, one-sided test for positive associations). Overall, stimulus-context neuron proportions showed significant positive correlations with both stimulus neurons ($\rho = 0.177$, $p = 0.022$) and context neurons ($\rho = 0.487$, $p < 0.001$), with a notably stronger relationship for context neurons.

Site-specific analyses revealed interesting regional patterns. Hippocampus and amygdala showed significant correlations with both neuron types (hippocampus: stimulus $\rho = 0.274$, $p = 0.032$; context $\rho = 0.471$, $p < 0.001$; amygdala: stimulus $\rho = 0.329$, $p = 0.012$; context $\rho = 0.487$, $p = 0.003$). In contrast, parahippocampal and entorhinal cortices showed significant correlations exclusively with context neurons (PHC: $\rho = 0.569$, $p < 0.001$; EC: $\rho = 0.377$, $p = 0.023$) but not stimulus neurons (PHC: $\rho = 0.033$, $p = 0.387$; EC: $\rho = -0.075$, $p = 0.324$).

These novel analyses reveal that stimulus-context neurons are more frequent in brain regions with more abundant context representations across all regions examined, and additionally with stimulus neuron proportions in hippocampus and amygdala. The distinct regional patterns appear to be compatible with a hierarchical organization of memory processing and a particular role of hippocampus (and amygdala) for integrating both stimulus and context information.

The following Supplementary Figure 9 has been added to the manuscript:

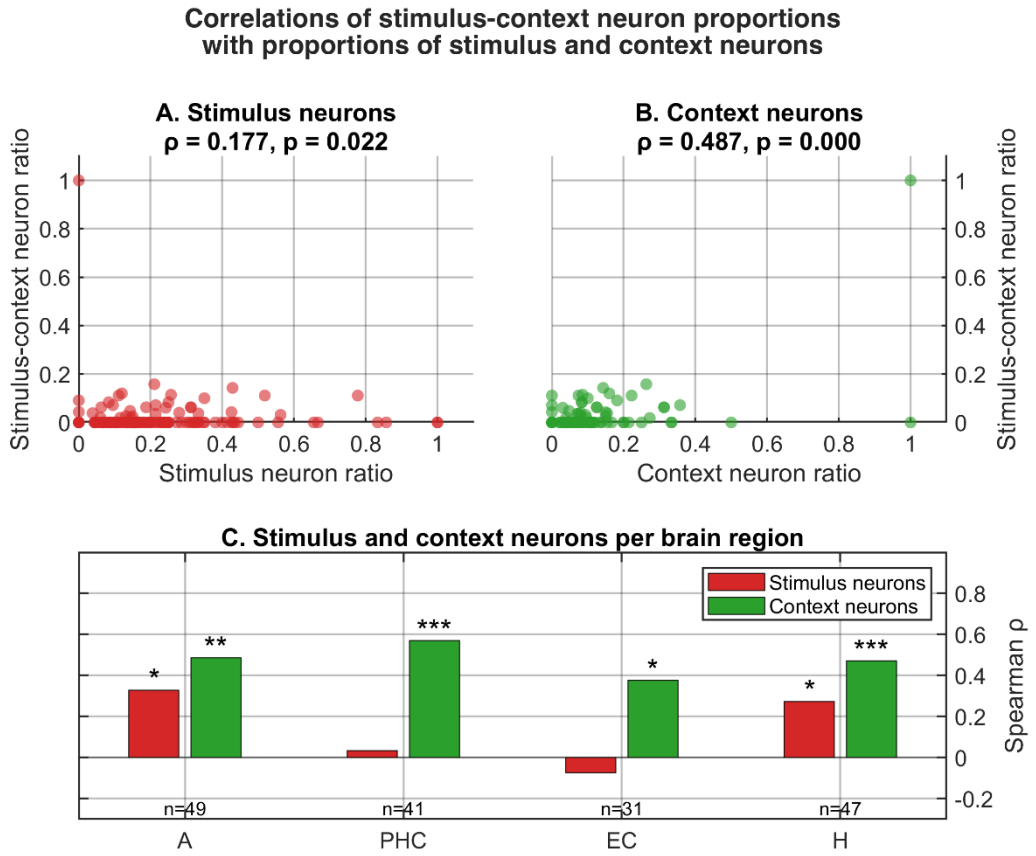


Fig. S9. Stimulus-context neurons are more prevalent in brain regions with higher proportions of context neurons and, to a lesser extent, stimulus neurons. a-b. Scatter plots of stimulus-context neuron ratios versus stimulus neuron ratios (**a**, red) or context neuron ratios (**b**, green) across all session-site combinations (N=168). Spearman correlations revealed weak positive associations with stimulus neurons (ρ = 0.177, p = 0.022) and strong positive associations with context neurons (ρ = 0.487, p < 0.001). **c.** Spearman correlation coefficients between stimulus-context neuron proportions and stimulus (red) or context (green) neuron proportions for each brain region. Statistical significance was assessed using permutation tests (1000 permutations, one-sided). Amygdala (A, N=49): stimulus ρ = 0.329, p = 0.012; context ρ = 0.487, p = 0.003. PHC (N=41): stimulus ρ = 0.033, p = 0.387; context ρ = 0.569, p < 0.001. EC (N=31): stimulus ρ = -0.075, p = 0.324; context ρ = 0.377, p = 0.023. Hippocampus (H, N=47): stimulus ρ = 0.274, p = 0.032; context ρ = 0.471, p < 0.001. ***p < 0.001; **p < 0.01; *p < 0.05 (uncorrected).

Furthermore, the following text passages were added to the Results, Methods, and Discussion sections, respectively:

"Results [...]

Stimulus and context are mainly encoded separately and rarely conjunctively

[...] Finally, we examined whether stimulus-context neurons were more prevalent in areas with higher proportions of stimulus or context neurons (Supplementary Fig. 9). Across all session-site combinations (N=168), stimulus-context neuron proportions correlated significantly with both stimulus (Spearman's ρ = 0.177, p = 0.022) and context neuron proportions (ρ = 0.487, p < 0.001), particularly in hippocampus and amygdala, where stimulus-context proportions correlated with both stimulus (hippocampus: ρ = 0.274, p = 0.032; amygdala: ρ = 0.329, p = 0.012) and context neuron proportions (hippocampus: ρ = 0.471, p < 0.001; amygdala: ρ = 0.487, p = 0.003). In contrast, parahippocampal and entorhinal cortices showed significant correlations exclusively with context (PHC: ρ = 0.569, p < 0.001; EC: ρ = 0.377, p = 0.023) but not with stimulus neuron proportions.

"Methods […]

Definition of neural populations and assessment of their significance

[…] Finally, Spearman correlations were computed between stimulus-context neuron proportions and stimulus or context neuron proportions across all session-site combinations (N=168, Supplementary Figure 9). Site-specific correlations were assessed using permutation tests (1000 permutations of stimulus-context proportions, one-sided test for positive associations)."

"Discussion […]

Context representations across MTL regions

It has been hypothesized that the MTL is organized according to separate parallel streams of content and context that converge in the hippocampus to ensure that contents can cue the retrieval of context and vice versa.[14,15] According to this view, context represented in neocortical areas (such as medial prefrontal, retrosplenial or posterior parietal cortex) is processed via parahippocampal or medial entorhinal cortex, and content representations (e.g., from lateral prefrontal, anterior temporal or inferior angular cortex) via lateral entorhinal or perirhinal cortex before convergence in the hippocampus.[14] […] While the large majority of stimulus neurons (88%) were invariant to context, and most context neurons (63.5%) were invariant to stimulus, a small but significant fraction of stimulus neurons represented either both or specific conjunctions of context and stimulus, particularly in the hippocampus (see Fig. 2b-d). Overall, conjunctive representations of specific pictures and questions were most abundant in hippocampus (2.29%), particularly among stimulus neurons (10.34%). Consistent with this, stimulus-context neurons were more prevalent in sessions with higher proportions of both stimulus and context neurons in hippocampus and amygdala, but only with higher context neuron proportions in parahippocampal and entorhinal cortices (Supplementary Figure 9). This regional pattern aligns with the proposed convergence of parallel streams in the hippocampus, where both information types may interact to form conjunctive representations.

Was their activity better related to **trial accuracy** than the other groups?

We appreciate the reviewer's question regarding whether stimulus-context neurons relate more strongly to trial accuracy than other groups, and address it below in response to the reviewer's insightful point about the link between neural activity and behavioral errors on objective questions.

2. Behavioral performance. The authors enhance the manuscript by including performance metrics that allow for subjectivity. However, for the questions in which an objective response is present (e.g. Bigger?) it would be possible to measure accuracy.

We thank the reviewer for this excellent suggestion. After determining objective answers to the questions "Bigger?" and "More expensive?"/"Older?", we assessed subjects' memory performance more explicitly by correlating, for each session and question, the behaviorally determined picture ranks with the corresponding objectively determined ranks.

We used Spearman's rank correlation to quantify rank agreement, excluding trials where no objective ranks could be assigned. The question "Brighter image?" was also excluded from these analyses as perceived brightness can deviate from physical luminance due to contextual factors and simultaneous contrast illusions. For each session, correlation values were averaged across the available objective questions.

As shown in Supplementary Figure 10, rank agreement with objective ground truth was strong (mean $\rho = 0.760$) and highly significant. All 50 sessions showed positive correlations ($p = 6.47 \times 10^{-10}$, Wilcoxon signed-rank test against zero), and the observed mean correlation significantly exceeded chance levels determined by permuting picture ranks ($p < 0.001$, permutation test). This validates that subjects successfully performed the memory task

and that their choices closely matched the objectively correct answers whenever such answers could be determined.

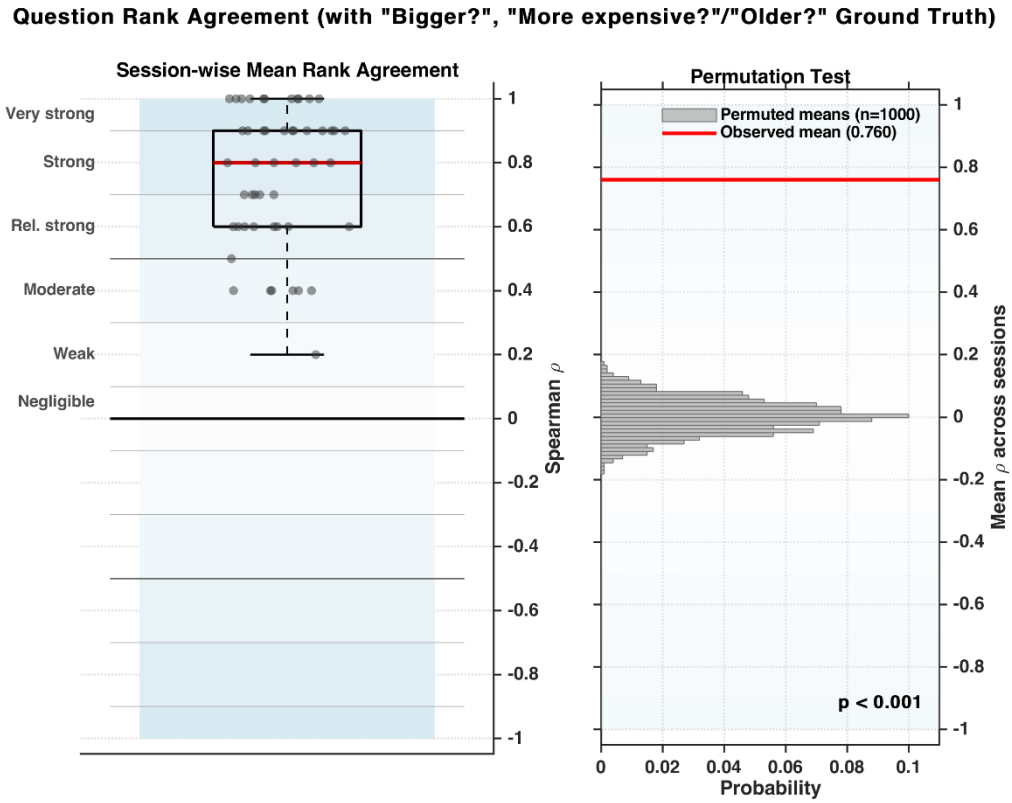


Fig. S10. Session-wise rank agreement with objective ground truth validates successful task performance. **Left:** Boxplot (Q1, median, Q3; whiskers: points within ±1.5 IQR) of session-wise mean Spearman rank correlations between behaviorally determined picture ranks and objectively determined picture ranks for the questions "Bigger?" and "More expensive?"/"Older?". Grey dots represent individual sessions (N=50). All session means were significantly greater than zero ($p = 6.47 \times 10^{-10}$, Wilcoxon signed-rank test). Shaded background areas indicate effect size categories. **Right:** Null distribution of mean correlations across sessions from 1,000 permutations (grey histogram), where picture ranks were randomly shuffled within each session-question pair before averaging. Observed mean shown in red ($\rho = 0.760$). The observed mean significantly exceeded the permutation null distribution ($p < 0.001$).

The following passages were added to the Results, Discussion and Methods sections, respectively:

> "**Results [...]**
>
> **Distinct populations encode stimulus and context in a picture-comparison task**
>
> [...] The majority of answers were highly consistent and transitive, with performance greatly exceeding chance in all but one excluded session, with these performance metrics not differing across question contexts ($p = 0.083$; Kruskal-Wallis-Test; Supplementary Figure 4). **Furthermore, answers strongly correlated with ground truths for "Bigger?" and "More expensive?"/"Older?" (mean $\rho = 0.760$, $p = 6.47 \times 10^{-10}$; Supplementary Figure 10).**"

"Discussion

Contexts constrain which memories are relevant for future decisions[13]. Both semantic content as well as context, i.e., other aspects of a study episode[1], such as tasks[9], episodes[11] or rules[10] are represented in single-neuron activity in the human MTL. It is currently unclear, however, whether and how content and context are combined to form or retrieve integrated item-in-context memories at the single-neuron level in humans. Pairs of pictures were presented on a laptop screen in different contexts, namely five questions that specified how the pictures were to be compared to each other (Fig. 1a). Sixteen of seventeen subjects remembered these contents (i.e., pictures) in their respective context as they chose both consistently and transitively between pictures despite unique and distinct orders among picture contents in each context (see Supplementary Figure 4). **Moreover, their rankings strongly matched ground truths (Supplementary Figure 10).**"

"Methods […]

Behavioral performance

[…] Finally, context-dependence of performance was assessed by deriving an overall behavioral performance measure from mean of consistency (γ) and transitivity (τ) values per question of each session and comparing them across contexts (Kruskal-Wallis-Test, $\alpha = 0.05$; Supplementary Figure 4b). **Moreover, for questions with objective ground truths ("Bigger?" and "More expensive?"/"Older?"), we validated behavioral performance by computing Spearman correlations between behaviorally determined picture ranks and objectively determined picture ranks, averaged across available questions per session. Session-wise correlations were tested against zero using a Wilcoxon signed-rank test, and the observed mean across sessions was compared to a null distribution generated from 1,000 permutations of picture ranks within each session-question pair (Supplementary Figure 10).**"

Additionally, if the activity is relevant to the task the authors should be able to show a connection between task performance and neural activity, e.g., is neural activity predictive of behavioral errors of the objective questions?

There is one sentence in the methods describing a single patient who failed the task and had no context sensitive neurons-that is a very interesting result that doesn't seem to be highlighted. Can this be seen on a trial by trial or session by session basis within patients performing the task?

We thank the reviewer for the highly relevant question of whether neural activity is related to task performance according to ground truth. Session-wise mean decoding accuracies of stimulus, context, and stimulus-context during picture presentations were compared between incorrect answers (not corresponding to ground truth) and correct answers (corresponding to ground truth). Decoding of context from context neurons and of stimulus identity from stimulus neurons was significantly higher on correct versus incorrect trials (context: $p = 0.0128$; stimulus: $p = 0.0351$), whereas stimulus-context decoding showed no reliable difference ($p = 0.4098$). The following Supplementary Figure 12 was added to the manuscript:

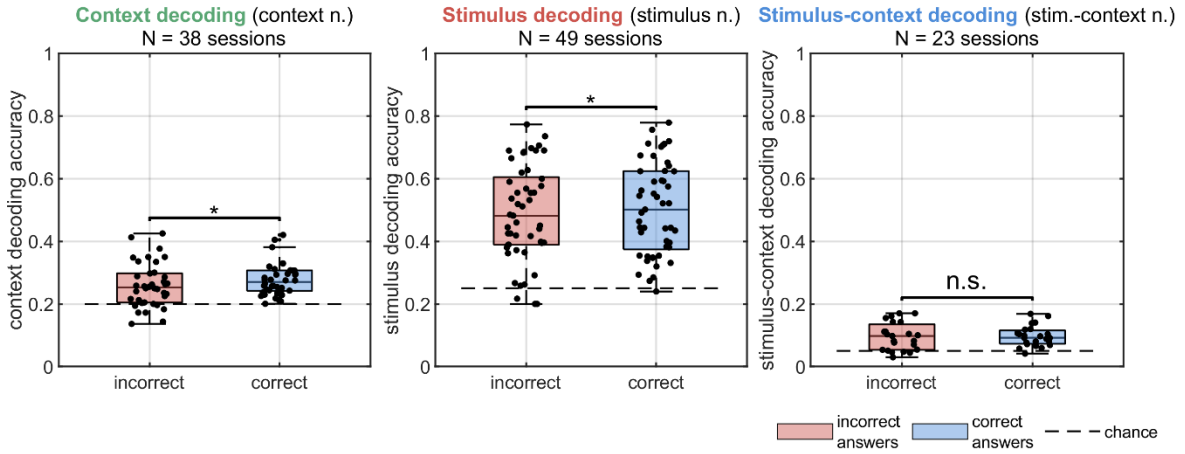


Fig. S12. Decoding of stimulus and context is enhanced when behavioral answers correspond to ground truth.

Boxplot (Q1, median, Q3; whiskers: points within ±1.5 IQR) of session-wise mean decoding accuracies during picture presentations for trials with incorrect (red) versus correct (blue) answers relative to objective ground truth. **Left:** context decoding from context neurons (N = 38; chance = 1/5). **Middle:** stimulus (picture) decoding from stimulus neurons (N = 49; chance = 1/4). **Right:** stimulus-context (picture and question) decoding from stimulus-context neurons (N = 23; chance = 1/20). Dashed lines indicate chance levels. Decoding accuracies were significantly higher for correct versus incorrect trials for context ($p = 0.0128$) and stimulus neurons ($p = 0.0351$), while stimulus-context neurons showed no reliable difference ($p = 0.4098$, n.s.). *$p < 0.05$ (uncorrected, Wilcoxon signed-rank test, one-sided, correct > incorrect).

The novel findings are now described in Results, Discussion and Methods sections, respectively:

> "**Results [...]**
>
> **Generality and temporal dynamics of context representations imply relevance for stimulus processing**
>
> [...] Finally, we tested whether content and context representations of stimulus and context neurons persist until the end of the trial when subjects chose one picture over another according to context (Fig. 3f). Both context as well as picture contents could be decoded above chance until a decision was indicated (two-sided cluster permutation test $p < 0.01$)**, with significantly enhanced decoding when choices matched objective ground truth (context: $p = 0.0128$; stimulus: $p = 0.0351$, one-sided Wilcoxon signed-rank test, Supplementary Figure 12).**"
>
> "**Methods [...]**
>
> **Behavioral performance**
>
> **[...]**
>
> Finally, context-dependence of performance was assessed by deriving an overall behavioral performance measure from mean of consistency (γ) and transitivity (τ) values per question of each session and comparing them across contexts (Kruskal-Wallis-Test, $\alpha = 0.05$; Supplementary Figure 4b). Moreover, for questions with objective ground truths ("Bigger?" and "More expensive?"/"Older?"), we validated behavioral performance by computing Spearman correlations between behaviorally determined picture ranks and objectively determined picture ranks, averaged across available questions per session. Session-wise correlations were tested against zero using a Wilcoxon signed-rank test, and the observed mean across sessions was compared to a null distribution generated from 1,000 permutations of picture ranks within each session-question pair (Supplementary Figure 10).
> **Additionally, decoding accuracies were related to behavioral performance by comparing correct trials (where behavioral choices matched ground-truth picture rankings) versus incorrect trials (where choices deviated from ground truth). Session-wise linear SVMs (5-fold cross-validation, 5 repeats) were trained on**

context, stimulus, or stimulus-context neuron activity during picture presentations (100-1000 ms) to decode questions, pictures, and picture-question combinations, respectively. Decoding accuracies for correct versus incorrect trials were compared using one-sided Wilcoxon signed-rank tests (correct > incorrect; $\alpha = 0.05$). Chance levels were 1/4 for stimulus, 1/5 for context, and 1/20 for stimulus-context decoding (Supplementary Figure 12)."

"Discussion […]

[…] During picture presentations when questions and picture contents became task-relevant, context neurons encoded associated question context, and stimulus neurons encoded picture contents (Fig. 3). Importantly, both context and content representations persisted until the end of the trial when one picture was chosen over the other according to the given context (Fig. 3f) and were enhanced during behaviorally correct trials (Supplementary Figure 12)."

Referees' comments:

Referee #1 (Remarks to the Author):

The authors have revised their manuscript and have performed a commendable job in addressing some of the outstanding questions raised during the last review. In particular, I appreciated their efforts in clarifying the distinction between 'task' context and environmental or background context, and their emphasis here on task context which seems appropriate. They have also sufficiently addressed the point about memory. Overall, they revised the manuscript well, and this is a valuable and well conducted study and will be of interest to the broader community.

We thank the reviewer for recognizing that our revision addressed all remaining concerns, for the valuable feedback provided throughout both rounds of review, and the overall positive assessment of the study's broader significance.

Referee #2 (Remarks to the Author):

The authors have satisfactorily addressed my remaining concerns in this iteration of the manuscript.

We thank the reviewer for recognizing that we satisfactorily addressed all remaining concerns, and for the valuable feedback provided throughout both rounds of review.

Referee #3 (Remarks to the Author):

The paper is careful, rigorous, and adds valuable human single-unit data. The dataset of thousands of human single units is especially impressive. All technical concerns on the analytical rigor have been adequately answered.

We thank the reviewer for recognizing the rigor of our work and the value of our extensive human single-unit dataset.

However, the experimental design falls short of supporting the broad claims made in the discussion.

We appreciate this constructive feedback and have carefully refined our claims to align with our experimental scope. Expansive descriptions have been pruned throughout the discussion.

First, the authors are limit themselves to stimuli that maximally activate units.

We selected pictures based on their likelihood of evoking visually selective responses, which is the standard procedure for investigating concept neurons. These neurons respond selectively to specific concepts across modalities and have been proposed as building blocks of human memory, with recent work from our group supporting this view (https://www.nature.com/articles/s41467-024-52295-5, now cited in the Discussion). Importantly, the stimulus neurons captured in our study maintained stable representations from pre-screening through experiments (Supplementary Figure 7), generalized across contexts and positions (Fig. 3b), and showed memory-consistent reactivation patterns (Supplementary Figure 11). These properties establish them as prime candidates for content representations in everyday memory.

Second, as the authors acknowledge, they are limited to "task context," a far narrower concept.

Our paradigm operationalizes interactive context as defined by Baddeley (1982), which represents a rather broad class of contexts that "influence the memory trace directly by affecting the way in which the stimulus is encoded" and "affect both recall and recognition." Such interactive contexts pervade everyday cognition. Following the reviewer's guidance, we have renamed the section from "Operationalization of real-world contexts" to "Operationalization of context" and removed references to "real-world". The section retains the now emphasized caveat: "While the task captures how interactive context shapes memory processing, future studies should examine these mechanisms under independent context manipulations and in naturalistic settings."

As such, rather than an expansive description of how the mesial temporal lobe pairs stimulus and context in a naturalistic setting, it is instead limited in both domains.

We acknowledge this point and have revised the discussion accordingly. While our paradigm employs controlled conditions, we demonstrate that both stimulus neurons (proxies for content representations) and context neurons (reflecting interactive task contexts) exhibit neural dynamics consistent with mechanisms that could generalize beyond the lab. These findings advance understanding of MTL function within a well-defined experimental framework.

Therefore, this paper makes a solid contribution that is not paradigm-shifting but is worthy of general attention in the community.

We thank the reviewer for their thoughtful engagement with our work. We agree that the paper makes a solid contribution and are confident that, within the now clearly defined scope, it provides important insights into how the human MTL integrates content and context information to support memory and flexible cognition.