

Retention and transformation of internal experiences in autobiographical memory narratives

Corresponding Author: Dr Hongmi Lee

This file contains all editorial decision letters in order by version, followed by all author rebuttals in order by version.

Version 0:

Decision Letter:

**** Please ensure you delete the link to your author homepage in this e-mail if you wish to forward it to your coauthors ****

Dear Dr Lee,

Thank you for your patience during the peer-review process. Your manuscript titled "Retention and transformation of internal experiences in autobiographical memory narratives" has now been seen by 4 reviewers, and I include their comments at the end of this message. They find your work of interest but raised some important points. We are interested in the possibility of publishing your study in Communications Psychology, but would like to consider your responses to these concerns and assess a revised manuscript before we make a final decision on publication.

We therefore invite you to revise and resubmit your manuscript, along with a point-by-point response to the reviewers. Please highlight all changes in the manuscript text file.

Editorially, we consider several points raised by the reviewers to be essential for revision. The most critical issues concern the methodological clarity and validity of your approach: you will need to provide a stronger justification for the use of your internal/external coding scheme and terminology relative to existing frameworks (e.g., the Autobiographical Interview), clarify the participant inclusion criteria and event selection procedures, and address the potential artifact introduced by the variable-length cue summaries used as recall prompts. In addition, your analysis of added versus retained details should be restructured to account for the unequal base rates of internal and external details, and inter-rater reliability statistics should be reported. Finally, one of your results—a comparison of two correlation coefficients—is reported as being “marginally significant.” Such terminology is not permitted in this journal; please revise to indicate that the comparison did not achieve significance.

We are committed to providing a fair and constructive peer-review process. Please don't hesitate to contact us if you wish to discuss the revision in more detail.

I am attaching an Editorial Requests Table that details critical reporting requirements for the revised manuscript. Please attend to each item and ensure your manuscript is fully compliant. If your revised manuscript is not aligned with these requests on major issues, such as those concerning statistics, it may be returned to you for further revisions without re-review.

Please submit the following items:

- Revised manuscript
- Point-by-point response to the referees' comments
- Cover letter (as a separate document)
- [Nature Research Reporting Summary](https://www.nature.com/documents/nr-reporting-summary.pdf)
- Completed Editorial Request Table (attached).

via this link: Link Redacted .

** This url links to your confidential home page and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage first **

Additional guidance is available in our style and formatting guide Communications Psychology formatting guide.

We hope to receive your revised paper within 8 weeks; please let us know if you aren't able to submit it within this time so that we can discuss how best to proceed. If we don't hear from you, and the revision process takes significantly longer, we may close your file. In this event, we will still be happy to reconsider your paper at a later date, provided it still presents a significant contribution to the literature at that stage.

We would appreciate it if you could keep us informed about an estimated timescale for resubmission, to facilitate our planning.

We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

Best regards,

Troby Lui, on behalf of

Jesse Rissman

Troby Lui, PhD
Associate Editor
Communications Psychology

Jesse Rissman, PhD
Editorial Board Member
Communications Psychology
orcid.org/0000-0001-8889-5539

REVIEWER EXPERTISE:

Reviewer #1: autobiographical memory; narrative analysis / NLP

Reviewer #2: autobiographical memory; narrative analysis / NLP

Reviewer #3: memory; narrative analysis / NLP

REVIEWER REPORTS:

Reviewer #1 (Remarks to the Author):

Using a large, open-access dataset of autobiographical memories sampled twice, several weeks apart (test-retest approach), the authors examined the consistency of internal versus external details, defined as unobservable or observable, respectively. To highlight some main findings: Internal details were less consistent. However, emotional intensity for such details protected memories from change.

This is an interesting, well-written, creative paper, and I enjoyed reading it. The topic is important and timely. I do have several concerns and suggestions for the authors to consider and I hope you will find my comments helpful and constructive.

Nomenclature and Scoring Concerns/Suggestions

1. The scoring protocol uses the nomenclature internal and external, which is a bit confusing, given the well-known scoring protocol by Levine et al. where these terms are used quite differently (to refer to episodic and non-episodic content, respectively). I worry this will cause confusion in the literature. Do not get me wrong, I like the distinction the authors are making, but the naming is a concern for me. I urge the authors to consider a name change (observable versus unobservable or something else). For simplicity, I will use your naming in my review.

2. Putting nomenclature aside, I wondered why the authors did not just use the Levine protocol (time, place, perceptual, spatial would map onto what the authors call "external" and could be summed into one category, and thought/emotions would map onto what the authors call "internal"). I know the Levine protocol very well and it is not clear to me why the authors opted to develop a new scoring system—I am sorry if I am missing something. The authors address this by saying, that for internal details, "Prior studies have typically reported them as merely one among many categories of memory details,

without systematically characterizing them at the level of individual details.” But is that really true? Many studies do report the individual detail types.

3. Are there any ambiguous statements that fall between internal and external? E.g., “I thought I might faint” —at first glance, this seems squarely internal. But this also involves how one perceives the environment relative to their own body (room is spinning, sounds are muffled). What about “we picked a chill spot in the park” —this is scored as category 1 by the authors but seems to contain two details: one for it being “chill” and one for “the park” (internal and external)

4. It seems the authors score at the sentence level but some statements contain many clauses: “I was hiking near Lake Ontario with my partner and our two grandkids” —this would be 4 details remembered but it seems it is given just 1? Sorry if I misunderstood.

5. I had trouble understanding the difference between internal thoughts and feelings versus speculations about others —how are those different? Don’t both refer to internal thoughts one is having at the time of the event. (I do understand why general comments would be separated—this is what Levine would call “metacognition” or “other” in his scoring protocol and makes sense as a distinct category).

6. Can the authors report ICCs for each detail type? Sorry if I missed it.

7. How do the brief summaries get factored into the scoring? Assuming those would more likely reference “external” content, could that account for the differences between internal and external (i.e., it gives external a leg up)?

8. I could not make up my mind about any potential circularity in using the narratives themselves to compute emotional intensity/valence. In the end I decided it was ok. But more broadly, how does VADER deal with clauses with ambiguity. My sister died (easy to score), she left me all her books (not so easy to score unless you use the context but even then, is this a good thing or a bad thing?).

9. A bit more information on the consistency scoring in the main text would be helpful—what are the different possibilities—some details can be consistent, some new etc etc. What is the math that is being done? What happens when a detail is something like, “I was a bit scared” versus “I was a bit nervous” versus “I was a bit apprehensive” —are these all “the same”?

Relationships to other literature

10. There is a literature on the consistency of emotions themselves. This literature is relevant but potentially overlooked in the paper and I think including it would had additional richness to your work: <https://pubmed.ncbi.nlm.nih.gov/38712802/>

11. Two of the main results are quite consistent with two recent papers: Wardell et al., 2025, BRM and Wardell & Palombo, 2025, now in *Emotion*), the former showing that emotions/thoughts are lower in consistency relative to other detail types, the latter showing that emotional narratives are more consistent overall than neutral ones. The authors cite this work but I wonder if it should be highlighted a bit more. Not because it is my own work, but in light of how similar the approach and findings are, especially with Wardell & Palombo, 2025. In my view, it is a strength that this work aligns well as it bolsters confidence in the findings. There are some divergencies important to note as well (In Wardell and Palombo, emotion seemed to affect both “external” and “internal” information, Figure 2C.)—this can be an interesting point of discussion.

12. The authors write: “This raises the question of whether remembering internal thoughts and feelings influences how people evaluate the significance of a recalled event. To test this, we used participants’ self-reported ratings of the perceived importance of their memories (i.e., how impactful, significant, or personal the event felt) on a 5-point scale, as provided in the Hippocampus dataset” —this statement appears causal. Can’t the opposite be true (more important leads to more consistency?). The authors address this later in the discussion but I would caveat this sentence as well as it could be misleading.

13. I liked the authors discussion of limitations and the distinction between forgetting and choosing not to share a detail. I wonder if the authors could speculate on whether their results would look different if they were sampling narratives from people who were sharing their memory to the same audience versus a different one. After all, in real life, we tend to share our memories across distinct contexts and conditions. Who we narrate to and who we narrate with (shared events) will change the memories. Perhaps inconsistency is underestimated when we create identical recall scenarios (as in a laboratory study).

-Daniela Palombo

Reviewer #2 (Remarks to the Author):

This study aimed to examine whether there exist differences in how internal and external experiences are remembered and preserved across time. The authors found that internal experiences were more susceptible to forgetting than external experiences, however, this susceptibility to forgetting may be lessened by factors such as high emotional intensity and increased semantic connections. Moreover, internal, but not external details promote a greater sense of perceived importance. Overall, this manuscript was well written. However, there are several major conceptual/theoretical and methodological concerns that need to be addressed and the novelty of this work needs to be better established.

Major Conceptual/Theoretical Points:

1. I do not find the study as novel as the authors claim it to be. While their analyses are novel, other research has already demonstrated the expected effects (as the authors themselves cite). For example, there are many studies of narrative and narrative coherence (including number of internal details, for example) that are for single events as opposed to the life story. Tilman Habermas, David Pillemer, Robyn Fivush, Adrial Boals, and David Rubin all have published studies looking at the narrative content of individual events. The authors might consider specifically the literature on meaning-making, which includes the use of internal details as individuals try to make sense of their memories. As another example, the authors claim in the discussion that previous research has had small sample sizes; there is a robust literature on autobiographical memory narratives and phenomenology that includes robust sample sizes.

2. The authors note throughout the paper that they are studying memories of thoughts and feelings (e.g., noted on page 16);

it is not entirely clear given their methodology that is the case. Rather, the authors are studying internal details that people recall when remembering external events. The paper should be revised to capture the nuance that they are studying internal thought or feelings that come up as part of the reconstructive process.

3. At several points in the Introduction and the Discussion, the authors argue that internal experiences are more susceptible to forgetting and distortion than observable external experiences. While I agree that the data suggests internal experiences are more prone to forgetting, it is not entirely clear what finding points to higher rates of "distortion" or exactly what "distortion" means in this context. How are the authors defining distortion? Does this refer to participants adding additional details in the second recall (which, based on the data, would point in the opposite direction)? In other words, aren't the changes that are specific to internal experiences all to do with forgetting? If so, the authors should clarify what distortion means in this context and what data specifically reflect distortion. See below for a concern regarding how the memories were sampled and how the addition/subtraction of details was considered by the authors.

4. The authors suggest that internal experiences are particularly biased by one's current emotions, beliefs, goals, and evaluations of past events. When mentioned in the Introduction and subsequently as an explanation of the findings in the Discussion, I expected there to be data that could speak to this point. For example, a measurement of participants' general emotional state at the time of recollection. Was there data to speak to this explanation specifically? And if not, the authors might consider rephrasing their discussion of this explanation as to not imply this could be measured in the present study.

Major Methodological Points:

1. A methodological artifact may be driving these results. Participants were asked to provide a brief summary (2-3 sentences) of their event to be used as a cue for the second recall. At this point, were the participants specifically told that they were providing a summary of the event to be used for a later recall? Additionally, was the length of these statements controlled (either at the time of recollection or in data preprocessing)? It is possible that some participants gave more details in the brief summary statement that they could use to promote remembering of more details during the second recollection. The authors should provide more details on participants' brief summaries, including length and number of details. If there was a lot of variability in length or number of details, it might be wise to control for this in the analyses. I understand that these are existing data so the information may not be available, but the authors should mention these as potential limitations if unable to control for it.

2. Additionally, according to the methods, participants were asked to remember an event from the past "three to six months". This implies that participants should NOT recall an event that occurred within 3 months of the first recollection. However, the authors then report that they "included participants who recalled an event that occurred within three months (90 days) prior to the first recall session." If my understanding is correct and participants followed instructions, there should be no memories that occurred within 3 months of the first recall session. Moreover, the average age of events at the first recall was 41.8 days old (just over 1 month). The authors should provide clarification of these points.

3. Also related to how participants were prompted to recall a memory, were participants told to recall a "specific" event that is not repeated (e.g., they recalled a hike, but they hike every Sunday)? If not, it is possible that participants recalled an event that was repeated several times between the first and second recall sessions. This may explain forgetting and/or added details and should be considered more thoroughly.

4. The analysis of the retention and addition of details is flawed. When analyzing the retention and addition of memory details, you show that internal details made up a significantly lower proportion of added details compared to external details. While interesting, this is not all that surprising given the overall finding that internal details make up a smaller proportion of all recalled details, as you suggest. A more insightful analysis would assess whether participants are more likely to generate new external details by accounting for the fact that internal details are generally recalled at lower rates. For example, you could calculate the proportion of new internal details relative to the overall number of internal details recalled and compare to the proportion of new external details relative to the overall number of external details.

5. The preserved recall of external details may have less to do with their external nature and more to do with schema and event sequencing. An external detail (e.g., location) is likely to cue other external details and aid in the preservation or addition of these details on subsequent recalls. For example, work by Marsh on retellings suggests that recall structure (schema) guides and biases later event memories. As a note, her work also includes data on internal details.

6. On the perceived importance of memories (page 14) - memories of internal experiences are more "emotionally charged". This represents a design flaw; by definition, emotion is internal. Additionally, the correlation between internal details and perceived importance is also flawed. Importance, which might be a corollary for meaning-making, would also be an internal detail. We tend to ruminate on more important events.

Additional design and analysis concerns:

I was unable to access the pre-registration. For any analyses not pre-registered, the authors should specify that they are exploratory.

Why is inter-rater reliability not reported? IRR should be reported at least for the training of coders.

Add alpha criteria for relevant tests (e.g., t-tests, ANOVAs, correlations). Why was there no alpha correction for multiple comparisons?

Add effect sizes and CIs for t-tests on top of pg. 7.

A note should be added to Table 1 describing what (unc.) means.

Add range for age of memories on pg. 5.

Reviewer #3 (Remarks to the Author):

Note: This review was co-written with a graduate trainee.

In this study, Su et al. utilized an existing large-scale dataset of participant-written autobiographical narratives (recalled at two time points) to investigate the memorability of internal details – i.e., memory for thoughts and feelings that occurred during an experience. Using a mixture of manual annotation and natural language processing, the authors identified internal and external (i.e., externally observable details from a remembered event) details within individual autobiographical memories and examined how they changed across two retrieval attempts, separated by several weeks. The authors found that internal details occur less frequently than external details and that they were more prone to distortion or forgetting. However, when an internal detail was more emotionally intense and semantically related to external details, it was more likely to be remembered. Despite their malleability, memories composed of a greater proportion of internal details were more likely rated as “important” by the rememberer.

Overall, the present manuscript is well-written and presents novel findings on the memorability of internal details in autobiographical memory. We both enjoyed reading this paper. The combination of qualitative and quantitative techniques is commendable. Also, we agree with the authors that the internal aspects of an experience is an understudied aspect of autobiographical memory research and thus these results will be highly valuable to the field. Below are some comments and suggestions that aim to help strengthen this paper.

(1) Definition of internal vs. external:

In this manuscript, the authors define internal and external details based on whether they refer to aspects of the remembered experience with a locus within the rememberer (e.g., their thoughts, feelings) or outside the rememberer (i.e., details that that could be corroborated by an external observer). This is intuitive, I think. But – the autobiographical memory literature, specifically work using the Autobiographical Interview (AI; Levine et al., 2002), uses the same terms to refer to a different rubric. According to the AI, internal details refer to “...those that pertained directly to the main event described by the participant, were specific to time and place, and were considered to reflect episodic reexperiencing.” (p. 679; Levine et al., 2002) External details were defined as “semantic (factual information or extended events that did not require recollection of a specific time and place).” (p. 679; Levine et al., 2002) In other words, internal details are episodic – irrespective of whether they are about the rememberer’s thoughts/feelings or observable details from an experience. For example, recalling that an event happened in a certain floor of the library at their university would count as an internal detail according to the AI, and (if we are understanding the authors’ work correctly) an external detail using this manuscript’s rubric. It also seems like the authors occasionally conflate these two distinct definitions of internal/external (e.g., “Memories of internal thoughts and feelings have rarely been the central focus of research using autobiographical narratives, even though they are commonly recalled and explicitly included as a distinct category in standardized coding protocols such as the Autobiographical Interview37.” (p. 16)

The authors should explicitly address the distinction between internal and external details as defined here and as defined in the AI in the manuscript. Perhaps calling them “internally-oriented” and “externally-oriented” details would help better delineate them?

(2) Is this actually forgetting?

Forgetting, in this study, was defined as a detail being mentioned in the first retrieval session that was not mentioned in the second retrieval session. But if we consider remembering as a communicative act (e.g., Marsh, 2007; Mahr & Csibra, 2018), not mentioning a detail might also be caused by not wanting to repeat oneself, whether implicitly or explicitly. Also, given that we might change the nature of what we recall given the communicative goals of recalling an experience (e.g., Dudukovic, Marsh & Tversky, 2004) yet another possibility could be that the internal experiences are simply less central/necessary to the demands of the specific recall instruction associated with this task (e.g., the participant might implicitly/explicitly conceive that whoever is asking them to recall this experience mainly cares about what actually happened – thus external details, by the authors’ definition). The authors mention this limitation at the end of the discussion, but the rest of the paper is heavily framed using forgetting [e.g., “...internal details ($M = 38.3\%$, $SD = 32.7$) were significantly less likely to be retained than external details ($M = 48.9\%$, $SD = 19.2$), indicating that internal experiences are more susceptible to forgetting over time”; “Although second recalls were generally shorter than the first—indicating some degree of forgetting—...” (p. 11)].

Some tempering of the use of forgetting language seems necessary here, given the above points. Also, it would be helpful if the authors more substantively/explicitly discussed some of the reasons, outside of forgetting, why a rememberer might not include a detail from the first recall session in the second recall session.

(3) Effects of delay?

By looking at the mean and SD, there seems to be a considerable amount of variability in the delay between the two retrieval sessions. Reporting a bit more detail about the distribution of delays would be helpful, given the importance of delay with respect to forgetting. I recommend including the range in addition to mean and SD, alongside a histogram of delays, perhaps as part of Figure 1 – or, at least, in the supplemental materials.

Furthermore, the authors refer to memory transformation over time throughout the manuscript. However, as far as I am aware, the effect of time was not explicitly modelled. Are internal details less likely to be repeated amongst individuals who happened to have a longer delay between recall sessions? It seems to me that delay could be added to the mixed-effects logistic regression models on memory retention. This seems worthwhile given all the work demonstrating the relationship

between delay and forgetting.

(4) Manual coding of details.

The raters manually coded each memory into one of the five following categories: (i) internal details, (ii) external details, (iii) inferences about others, (iv) general comments, and (v) other details. The authors mentioned that these categories were predefined. How did the authors end up deciding on these definitions? What was a systematic inductive coding procedure that the authors used? In any case, some more details about how the authors settled on using these categories is warranted.

(5) Interrater reliability.

The authors stated the following: annotators' "performance was monitored by one of the authors (C.K.), and they were encouraged to ask questions before and during the task to clarify any uncertainties. C.K. also conducted a quality check of each completed recall narrative and requested revisions when any issues were identified." Were any interrater reliability measures taken? For example, did C.K. or another annotator independently segment, categorize, and/or match subsets of the data coded by the annotators. Were metrics such as percent agreement or Cohen's kappa computed? Given that details were deductively coded as either internal or external (i.e., "Annotators were also provided with detailed definitions and examples for each category."), rigorous interrater reliability checks may not be as applicable to the present study relative to other, more exploratory qualitative work. That said, other studies also using manual coding, for example, to analyze distinct aspects of autobiographical memory narratives (e.g., Sotgiu et al., 2025), computed Cohen's kappa values as an index of interrater reliability among coders. The authors of the present manuscript should specify whether any such measures were taken, and if not, should provide justification for their decisions.

(6) Mixed-effects logistic regression predictors.

To predict the retention of individual memory details, the authors conducted a mixed-effects logistic regression analysis, using emotional valence, emotional intensity, semantic connection to external elements, and word count as predictors. How were these predictors selected? Were they theoretically informed? Were they selected based on natural language processing constraints? The authors should clarify their decisions regarding predictor selection, specifically explaining why they expect such variables (as opposed to others) to predict the likelihood of memory retention. The rationale is briefly touched on in the "Features predicting memory detail retention" section of the results (pp. 12-13), but it would be beneficial to discuss it earlier on (perhaps in the methods section) to set readers up for understanding why specific predictors, and not others, were selected for focus.

(7) Word count as an index of informational richness.

The authors indicated that word count per detail acted as an indicator of information richness. Did the authors control for stop words in word count calculations? Stop words may inflate the degree of "information richness" given their contribution to word count, despite their relative lack of "richness." In general, more details on any automated text preprocessing steps used would also be helpful.

(8) Semantic similarity of matched details.

The authors report that internal details show less semantic similarity to their matched counterparts across recall sessions than matching external details (see Figure 3C). Why is this the case? One possibility is that internal details are truly less stable across recall attempts – which is the interpretation that the authors use. However, might another possibility be that this is a property of the SBERT embeddings themselves? For example, if SBERT vectors better capture the semantics of concrete sentences better than abstract ones – and if internal details tend to describe abstract mental states/emotions – might that be another reason to observe this effect? Personally, I am not aware of much work examining the effect of word/sentence concreteness on embedding quality, but if the authors have any thoughts on this possibility, I'd be curious to hear them.

Minor comments:

(1) The sentence "...researchers often assess the consistency of memories across repeated sessions rather than their accuracy relative to what actually occurred" (p. 1) reads a bit unclear. Consider cutting "relative to what actually occurred".

(2) For memory detail segmentation, categorization, and matching, how did annotators decide which narratives they were responsible for? Was there any random selection/assignment involved with respect to which memories were assigned to which raters?

(3) In the methods subsection, "Memory detail categorization", the authors provide examples of all detail categories except unrelated details. Consider elaborating on and providing an example of this category, too, within the main manuscript.

(4) At a couple points in the manuscript, I was confused whether the semantic similarity analyses were referring to semantic similarity across matched details (e.g., internal detail *i* from recall session 1 to internal detail *i* from recall session 2) or local/global similarity, which is yoked to external details (e.g., Figure 5C, which I think was measuring the former not the latter). Perhaps adding a sentence clarifying which is used when might be useful – but it could have just been me.

(5) Also, with respect to local similarity, the calculation was described as targeting "typically two external details preceding and following" — why typically and not always? Is it something to do when a detail doesn't have 2 preceding it or following it? If so, it would be useful to add a sentence clarifying when 2 before and after was not used. And, to clarify, this approach was used to calculate local/global similarity for both internal and external details? In other words, local and global similarity always refer to a target detail's semantic similarity to either a small set (i.e., local) or a large set (i.e., global) of external details?

(6) On page 13, the authors state: "Overall, compared to external details, internal details were more emotionally intense, but showed weaker semantic links to external details and were shorter in length (Table 1)". The fact that internal details are more emotionally intense isn't surprising, right? Given they are defined as an individual's thoughts/feelings from an

experience? In reading this, it sounded like a finding, but it seems more like a sanity check. Perhaps a sentence clarifying this would be helpful. Perhaps not.

References:

- Dudukovic, N. M., Marsh, E. J., & Tversky, B. (2004). Telling a story or telling it straight: The effects of entertaining versus accurate retellings on memory. *Applied Cognitive Psychology*, 18(2), 125-143.
- Levine, B., Svoboda, E., Hay, J. F., Winocur, G., & Moscovitch, M. (2002). Aging and autobiographical memory: dissociating episodic from semantic retrieval. *Psychology and aging*, 17(4), 677.
- Mahr, J. B., & Csibra, G. (2018). Why do we remember? The communicative function of episodic memory. *Behavioral and brain sciences*, 41, e1.
- Marsh, E. J. (2007). Retelling is not the same as recalling: Implications for memory. *Current Directions in Psychological Science*, 16(1), 16-20.
- Sotgiu, I., Consonni, S., & Marengo, D. (2025). Testing order effects in autobiographical memory research. *Memory*, 1–12. <https://doi.org/10.1080/09658211.2025.2561999>

Reviewer #4 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Communications Psychology initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

* TRANSPARENT PEER REVIEW: Communications Psychology uses a transparent peer review system. This means that we publish the editorial decision letters including Reviewers' comments to the authors and the author rebuttal letters online as a supplementary peer review file. However, on author request, confidential information and data can be removed from the published reviewer reports and rebuttal letters prior to publication. If your manuscript has been previously reviewed at another journal, those Reviewers' comments would not form part of the published peer review file.

** Visit Nature Research's author and referees' website at www.nature.com/authors for information about policies, services and author benefits**

Communications Psychology is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the Manuscript Tracking System by clicking on 'Modify my Springer Nature account' and following the instructions in the link below. Please also inform all co-authors that they can add their ORCIDs to their accounts and that they must do so prior to acceptance.

<https://www.springernature.com/gp/researchers/orcid/orcid-for-nature-research>

For more information please visit <http://www.springernature.com/orcid>

If you experience problems in linking your ORCID, please contact the Platform Support Helpdesk.

Version 1:

Decision Letter:

** Please ensure you delete the link to your author homepage in this e-mail if you wish to forward it to your coauthors **

Dear Dr Lee,

Your manuscript titled "Retention and transformation of internal experiences in autobiographical memory narratives" has now been seen by our reviewers, whose comments appear below. In light of their advice I am delighted to say that we are happy, in principle, to publish a suitably revised version in Communications Psychology.

We therefore invite you to revise your paper one last time to address the remaining concerns of our reviewers and a list of editorial requests. At the same time we ask that you edit your manuscript to comply with our format requirements and to maximise the accessibility and therefore the impact of your work.

EDITORIAL REQUESTS:

Please review our specific editorial comments and requests regarding your manuscript in the attached "Editorial Requests Table". Please outline your response to each request in the right hand column. Please upload the completed table with your manuscript files as a Related Manuscript file.

If you have any questions or concerns about any of our requests, please do not hesitate to contact me.

SUBMISSION INFORMATION:

In order to accept your paper, we require the files listed here <https://www.nature.com/documents/commsj-file-checklist.pdf>.

OPEN ACCESS:

Communications Psychology is a fully open access journal. Articles are made freely accessible on publication. For further information about article processing charges, open access funding, and advice and support from Nature Research, please visit <https://www.nature.com/commspsychol/open-access>

At acceptance, you will be provided with instructions for completing the open access licence agreement on behalf of all authors. This grants us the necessary permissions to publish your paper. Additionally, you will be asked to declare that all required third party permissions have been obtained, and to provide billing information in order to pay the article-processing charge (APC).

*** TRANSPARENT PEER REVIEW:** Communications Psychology uses a transparent peer review system. On author request, confidential information and data can be removed from the published reviewer reports and rebuttal letters prior to publication. If you are concerned about the release of confidential data, please let us know specifically what information you would like to have removed. Please note that we cannot incorporate redactions for any other reasons.

*** CODE AVAILABILITY:** All Communications Psychology manuscripts must include a section titled "Code Availability" at the end of the methods section. We require that the custom analysis code supporting your conclusions is made available in a publicly accessible repository at this stage; please choose a repository that generates a digital object identifier (DOI) for the code; the link to the repository and the DOI must be included in the Code Availability statement. Publication as Supplementary Information will not suffice.

*** DATA AVAILABILITY:**

All Communications Psychology manuscripts must include a section titled "Data Availability" at the end of the Methods section. More information on this policy, is available in the Editorial Requests Table and at <http://www.nature.com/authors/policies/data/data-availability-statements-data-citations.pdf>.

Please use the following link to submit the above items:

Link Redacted

****** This url links to your confidential home page and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage first ******

We hope to hear from you within two weeks; please let us know if you need more time.

Best regards,

Troby Lui

Troby Lui, PhD
Associate Editor
Communications Psychology

Jesse Rissman, PhD
Editorial Board Member
Communications Psychology
orcid.org/0000-0001-8889-5539

REVIEWERS' COMMENTS:

Reviewer #1 (Remarks to the Author):

Review

The authors did a thorough job with the reviewer comments, including my own. It was very helpful to see how consistency is calculated as this will allow replication and comparison across protocols, where the denominator differs. I have only one minor comment for the authors.

Consider a small edit:

Change: A recent study examining repeated autobiographical recalls similarly found that thoughts and emotions were recalled less consistently than other types of memory details

To: A recent study examining repeated autobiographical recalls similarly found that thoughts and emotions were recalled less consistently than most other types of memory details (add "most")

Reviewer #2 (Remarks to the Author):

I am satisfied with the author's responses to my comments (Reviewer #2)

Reviewer #3 (Remarks to the Author):

The authors have addressed all of my concerns. The revised version is more clear and precise. Also, I appreciate their effort in conducting a retrospective calculation of inter-rater reliability for their manual coding scheme. While this certainly took some time, I think these scores provide useful qualifiers for the reader about the nature of these categories.

Reviewer #4 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Communications Psychology initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

**** Visit Nature Research's author and referees' website at www.nature.com/authors for information about policies, services and author benefits****

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were

made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

EDITORS' COMMENTS:

Editorially, we consider several points raised by the reviewers to be essential for revision. The most critical issues concern the methodological clarity and validity of your approach: you will need to provide a stronger justification for the use of your internal/external coding scheme and terminology relative to existing frameworks (e.g., the Autobiographical Interview), clarify the participant inclusion criteria and event selection procedures, and address the potential artifact introduced by the variable-length cue summaries used as recall prompts. In addition, your analysis of added versus retained details should be restructured to account for the unequal base rates of internal and external details, and inter-rater reliability statistics should be reported. Finally, one of your results—a comparison of two correlation coefficients—is reported as being “marginally significant.” Such terminology is not permitted in this journal; please revise to indicate that the comparison did not achieve significance.

: We thank the editors for their consideration of our manuscript. We have thoroughly revised the manuscript in response to the editors' and reviewers' requests, focusing on providing clearer explanations and stronger justification for our methodological approach. Detailed responses to all issues raised by both the editors and reviewers are included in the point-by-point responses below. Here, we additionally address the issues specifically raised by the editors. First, regarding the participant inclusion criteria and event selection procedures, we have revised the corresponding subsections of the Methods to provide a clearer description of the initial composition of the Hippocorpus dataset and the criteria and procedures used for data selection.

p.4. The Hippocorpus dataset⁴³ comprises 6,854 first-person narratives describing **recalled or imagined** memorable life events, collected from crowdworkers on Amazon Mechanical Turk.

p.6. Since our study focused on the retention and transformation of memory across repeated recalls, we initially selected data from 1,308 participants who **described the same event in two recall sessions, with the summary generated in the first session provided as the cue for the second recall**. Participants who provided imagined narratives or who completed only the first, but not the second, recall session were excluded from analysis.

We then applied additional **step-by-step** screening procedures **based on criteria pre-defined in our preregistration protocol** (<https://aspredicted.org/9zmr-djp5.pdf>), resulting in a final sample of 210 participants included in our analyses. **First, to ensure recall quality, we selected 822 participants who were not distracted during memory recall. These participants responded “Not at all” to the question, “How distracted were you while writing your story?” during both the first and second recall sessions. Next, to focus on memories of clearly identifiable, specific events, we included only participants whose first and second recalls described the same specific event that occurred within a single day. Participants were excluded if they recalled different events across sessions, described situations that unfolded over several days or weeks (e.g., changes in life after returning to college), provided only general statements about recurring events, or**

produced narratives that were disorganized and lacked a clear episodic focus. This selection was performed manually by one of the authors (C.K.), who reviewed the content and quality of both recalls. This step resulted in a remaining pool of 330 participants. Finally, we selected 210 participants who recalled an event that occurred within three months (90 days) prior to the first recall session. This criterion was intended to focus on relatively recent memories from the Hippocampus dataset⁴³, as older, more stabilized memories typically allow less room for transformation between recall sessions^{30,56}.

Regarding the use of the term ‘marginally significant,’ we have revised the relevant sentence in the Results section to indicate that the difference between the two correlation coefficients was not statistically significant.

p.19. However, greater semantic similarity of internally oriented details across sessions was associated with an increase in perceived importance ($r(125) = .19, p = .037, 95\% \text{ CI} = [.01, .35]$), whereas no such relationship emerged for externally oriented details ($r(207) = -.004, p = .951, 95\% \text{ CI} = [-.14, .13]$; Fig 5c). Although the difference between the correlations for internally and externally oriented details did not reach statistical significance ($t(123) = 1.89, p = .061, 95\% \text{ CI} = [-.01, .41]$, Cohen’s $q = .21$), the positive correlation for internally oriented details suggests that consistent and precise retention of internal experiences may help buffer against the general decline in perceived importance over time.

Please note that we made an additional revision to our report of participants’ ages, although this was not requested by the editors or reviewers. During our closer inspection of the original questionnaires used in the Hippocampus data collection, we discovered that participants reported their age by selecting a predefined age group (e.g., 18–24, 25–29), rather than entering their exact age in years. Accordingly, the values in the dataset’s “age” column represented the lower bounds of these age groups, not actual ages. Because the exact ages were not available, we now report the percentage of participants in each age group instead of the mean and standard deviation of age.

p.6. Based on data collected during the first recall session, the final sample of 210 participants included 6.7% aged 18–24, 20.0% aged 25–29, 23.3% aged 30–34, 18.1% aged 35–39, 7.1% aged 40–44, 9.1% aged 45–49, 8.1% aged 50–54, and 7.6% aged 55 or older.

Finally, during the revision process, we found a minor reporting error in the 95% confidence interval for the importance rating comparison on page 18. The original submission reported the CI as [.02, .41], but the correct 95% CI is [.20, .41]. We have updated this value in the revised manuscript. This correction does not change any interpretation or conclusions.

p.18. Indeed, importance ratings were significantly higher in the first recall ($M = 4.33, SD = .90$) than in the second ($M = 4.02, SD = 1.03; t(209) = 5.82, p < .001, 95\% \text{ CI} =$

[.20, .41], Cohen's $d = .31$).

REVIEWER REPORTS:

Reviewer #1 (Remarks to the Author):

Using a large, open-access dataset of autobiographical memories sampled twice, several weeks apart (test-retest approach), the authors examined the consistency of internal versus external details, defined as unobservable or observable, respectively. To highlight some main findings: Internal details were less consistent. However, emotional intensity for such details protected memories from change.

This is an interesting, well-written, creative paper, and I enjoyed reading it. The topic is important and timely. I do have several concerns and suggestions for the authors to consider and I hope you will find my comments helpful and constructive.

: We thank the reviewer for their thorough consideration and positive assessment of our work.

Nomenclature and Scoring Concerns/Suggestions

1. The scoring protocol uses the nomenclature internal and external, which is a bit confusing, given the well-known scoring protocol by Levine et al. where these terms are used quite differently (to refer to episodic and non-episodic content, respectively). I worry this will cause confusion in the literature. Do not get me wrong, I like the distinction the authors are making, but the naming is a concern for me. I urge the authors to consider a name change (observable versus unobservable or something else). For simplicity, I will use your naming in my review.

: We thank the reviewer for their suggestion. Reviewer 3 raised a similar concern in their Comment 1. We fully agree that the terms “internal details” and “external details” may be confusing to readers familiar with their usage in the Autobiographical Interview scoring method. Following the reviewers’ request, we have replaced these terms with “internally oriented details” and “externally oriented details” (suggested by Reviewer 3) throughout the manuscript where appropriate.

2. Putting nomenclature aside, I wondered why the authors did not just use the Levine protocol (time, place, perceptual, spatial would map onto what the authors call “external” and could be summed into one category, and thought/emotions would map onto what the authors call “internal”). I know the Levine protocol very well and it is not clear to me why the authors opted to develop a new scoring system—I am sorry if I am missing something. The authors address this by saying, that for internal details, “Prior studies have typically reported them as merely one among many categories of memory details, without systematically characterizing them at the level of individual details.” But is that really true? Many studies do report the individual detail types.

: We appreciate the reviewer for raising this important point. Reviewer 3 also raised a related issue in their Major Comment 4. We agree that the ‘internally oriented’ category in our study corresponds to the internal ‘thoughts/emotions’ details defined in the Levine protocol (provided that the thoughts/emotions were the participants’ own), and that the ‘externally oriented’ category corresponds to the combination of internal ‘event,’ ‘time,’ ‘place,’ and ‘perceptual’ details in the Levine protocol. The ‘other’ category in our study corresponds to the ‘external’ details in the Levine protocol.

We did not use the Levine protocol in its original form and instead adopted a simplified coding scheme, as our interest was in the broader distinction between internally and externally oriented experiences. Because we did not plan to differentiate among specific types of externally oriented details or details unrelated to the recalled event, it was unnecessary to include all categories from the Levine protocol. This approach also allowed us to conserve limited time and resources required for manual coding. Similarly, we used a coarser-grained narrative segmentation than the Levine protocol (a point also relevant to our response to Comment 4). Specifically, we treated clauses as the smallest units of analysis rather than dividing them into more fine-grained elements, consistent with approaches used in traditional narrative comprehension and memory research (e.g., Trabasso & van den Broek, 1995, *J. Mem. Lang.*) and in recent work on automatic narrative segmentation and text analysis of narrated recall (e.g., Georgiou et al., 2023, *bioRxiv*).

Thus, our initial simplified set of memory detail categories included internally oriented, externally oriented, and other details. However, during pilot coding used to test and refine the coding instructions for the annotators, we identified several potential misclassification issues. First, as the reviewer noted in Comment 5, descriptions of participants’ perceptions of other people’s thoughts and feelings could be mistakenly coded as internally oriented details, which were meant to capture only the participants’ own thoughts and feelings accessible solely to themselves. Second, thoughts and feelings that arose during the act of recall, rather than during the original event, could also be misclassified as internally oriented details, even though the category was intended to include only internal states experienced at the time of the event (this issue is also relevant to Reviewer 2’s Conceptual Point 2). Because these types of details often used language similar to that of internally oriented details to describe internal states, they were easily confusable. To address these issues, we created two additional categories, ‘inferences about others’ and ‘general comments or evaluations’, to help annotators avoid misclassification. Making annotators explicitly aware of these types of details and instructing them to categorize such instances separately allowed us to identify and exclude them from analysis, as they were not relevant to the main goal of our study.

We have revised the ‘Memory detail categorization’ subsection of the Methods to more clearly justify our categorization scheme and compare it with the Levine protocol, also in relation to the reviewer’s Comment 1.

p.8. Since the primary goal of the current study was to examine the retention and transformation of internal thoughts and feelings over repeated recalls, and to compare these patterns with those of objective external elements, these two categories **were defined and analyzed as the primary categories of interest. Given our focus on the broad**

internal-external distinction, we did not further subdivide externally oriented details into finer-grained categories (e.g., time, location, perception), as is done in previously established protocols (e.g., ref.³⁹). The categories of ‘speculations or inferences about others’ and ‘general comments or evaluations’ were additionally introduced to help annotators avoid misclassification. The former distinguished other people’s inferred internal experiences from participants’ own, while the latter separated thoughts or feelings generated during the act of recall from those that occurred during the original event. Because internally oriented details and these additional types of details all described internal states using similar language, they were easily confused. By explicitly flagging these categories for annotators and instructing them to classify such details separately, we aimed to more effectively identify and exclude them from the main analysis—together with the ‘other’ category—as they were not relevant to the study’s primary goal.

It is also important to note that our internally and externally oriented memory detail categories differ from the “internal” and “external” details defined in the widely used Autobiographical Interview (AI) protocol³⁹. In the AI protocol, internal details refer to episodic elements directly related to the main recalled event, whereas external details encompass information not directly tied to that event (e.g., semantic knowledge). Under this framework, both the internally oriented and externally oriented details in our study correspond to the AI protocol’s internal details, while our ‘other’ category corresponds to the AI protocol’s external details.

Regarding the statement, “Prior studies have typically reported them as merely one among many categories of memory details, without systematically characterizing them at the level of individual details,” we agree that earlier work has also reported individual detail types, including thoughts and feelings as one of the subcategories, which we acknowledged in the same paragraph of the original manuscript (p.19: “(memories of internal thoughts and feelings) are commonly recalled and explicitly included as a distinct category in standardized coding protocols such as the Autobiographical Interview”). Our intended point was that, because these studies did not specifically target internally oriented details, they did not prioritize analyzing the characteristics or features of individual internally oriented details (e.g., identifying which internally oriented details within a narrative are most likely to be recalled). We have revised the paragraph to clarify this point.

p.19. Internal thoughts and feelings have rarely been the central focus of **cognitive psychology research on memory** using autobiographical narratives, even though they are commonly recalled and explicitly included as a distinct category in standardized coding protocols such as the Autobiographical Interview³⁹. Prior studies **using these protocols** have typically reported the **overall proportion and reliability of thought- and feeling-related details** as one category among many^{30,39,48}. However, because internal experiences were not the primary target of this work, their features and characteristics at the level of individual details have received relatively little attention. The present study extends this prior research by conducting **more fine-grained** analyses of internal

experiences within narratives, focusing on individual details and identifying the factors that contribute to their relative durability.

3. Are there any ambiguous statements that fall between internal and external? E.g., “I thought I might faint” —at first glance, this seems squarely internal. But this also involves how one perceives the environment relative to their own body (room is spinning, sounds are muffled). What about “we picked a chill spot in the park” —this is scored as category 1 by the authors but seems to contain two details: one for it being “chill” and one for “the park” (internal and external)

: We appreciate the reviewer’s careful examination of our approach. The reviewer is correct that the distinction between internal and external experiences was not always completely clear, and some ambiguous cases were possible. Whether a memory detail (i.e., clause; please see our response to Comment 4 below) was categorized as internal or external depended on the annotators’ best subjective judgment and common-sense understanding of what constitutes a thought or feeling. Ambiguous or mixed cases were categorized according to the relatively dominant meaning or focus of the detail as a whole. Therefore, “we picked a chill spot in the park” was categorized as an external experience because it primarily describes an observable action—moving to and sitting at a particular location in the park. In contrast, “I thought I might faint” was categorized as an internal experience because it describes a subjective internal state accessible only to the participant, consistent with our definition of the internal category (p.7. “participant’s personal emotions, thoughts, intentions, or interpretations during the recalled event. These were internal states accessible only to the participant”). Although the feeling of almost fainting may involve perceptions of the external environment (e.g., the room seeming to spin), this experience reflects the participant’s internal physiological or emotional state: others in the same situation would not experience it (as the room is not actually spinning), and the internal feeling itself is not observable unless expressed through overt behavior (e.g., stumbling).

We have slightly revised the ‘Memory detail categorization’ subsection of the Methods to clarify the categorization procedure.

p.7. Annotators received spreadsheets containing the segmented recall narratives, with each memory detail listed in a separate row. In an adjacent column, they assigned a category to each detail by entering a number corresponding to one of the predefined categories (see Supplementary Table 1 for an example). **Annotators used their best subjective judgment, classifying each detail according to its dominant meaning or focus.**

In addition, although some details were ambiguous, we do not believe they were prevalent enough to blur the overall distinction between internal and external categories. We observed clear differences between the two categories in both memory performance measures (e.g., Figure 3) and other associated features (e.g., Table 1). Thus, while some misclassification is possible, we do not expect it to substantially alter the main patterns reported in our study.

4. It seems the authors score at the sentence level but some statements contain many clauses: “I was hiking near Lake Ontario with my partner and our two grandkids” —this would be 4 details

remembered but it seems it is given just 1? Sorry if I misunderstood.

: We thank the reviewer for this question. As the reviewer noticed, we segmented the recall narratives primarily into sentences (using automated text processing). Our annotators then additionally segmented these sentences into clauses (defined as a group of words that contains both a subject and a verb, which can be a complete sentence or part of a sentence) in cases where they contained different ideas—especially if they could be categorized as different types of the five details we used. The example given in the reviewer’s comment, “I was hiking near Lake Ontario with my partner and our two grandkids,” is actually a single clause: “I” (subject) + “was hiking near Lake Ontario with my partner and our two grandkids” (predicate).

We used clauses as the smallest segments rather than dividing them further into even more fine-grained pieces of information (e.g., I was hiking / near Lake Ontario / with my partner / and our two grandkids) for a few different reasons. A clause that expresses an integrated, cohesive idea has commonly served as the standard unit of analysis both in the traditional literature on discourse and narrative comprehension and memory (e.g., Trabasso & van den Broek, 1995, *J. Mem. Lang.*) and in more recent work on automatic narrative segmentation and text analysis of narrated recall (e.g., Georgiou et al., 2023, *bioRxiv*). There was also a practical consideration: segmenting and identifying individual pieces of information within each clause would substantially increase the time and effort required during segmentation. Moreover, it would greatly increase the number of details requiring categorization and cross-session matching, making this level of granularity impractical for our manual coding procedures given the available time and resources. We have revised the ‘Memory detail segmentation’ subsection of the Methods to more clearly explain our rationale for not subdividing clauses into finer-grained units. Please note that we mistakenly used the term “phrase” instead of “clause” in the original manuscript, and this has now been corrected.

p.7. Specifically, sentences containing multiple distinct ideas were split into separate **clauses**. For example, the sentence “It was really nice outside, which made me happy.” would be divided into two lines: “It was really nice outside” and “which made me happy.” **Clauses have commonly served as the standard unit of analysis in the literature on narrative comprehension and memory^{57,58}. Thus, to preserve the integrated, cohesive idea conveyed within each clause, we did not further subdivide them into finer-grained pieces of information.**

5. I had trouble understanding the difference between internal thoughts and feelings versus speculations about others —how are those different? Don’t both refer to internal thoughts one is having at the time of the event. (I do understand why general comments would be separated—this is what Levine would call “metacognition” or “other” in his scoring protocol and makes sense as a distinct category).

: We thank the reviewer for raising this point. The “speculations or inferences about others” category included descriptions of participants’ perception of other people’s internal experiences (e.g., thoughts and feelings) during the event. This category was distinct from internal thoughts and feelings in two important ways: (1) it referred to thoughts and feelings attributed to other

people (e.g., “He was sad”), rather than the participant’s own (e.g., “I was sad”); and (2) such attributions were likely based on externally observable behaviors or expressions, since participants cannot directly access others’ internal states and can only infer them from what is observable. It is also worth noting that direct descriptions of other people’s observable expressions or their explicitly stated emotions (e.g., “He said he was sad”) were classified as external details. We included this category in our coding scheme to prevent annotators from misclassifying these descriptions as the participant’s own internal thoughts and feelings, given that both categories involve descriptions of internal states. We have revised the ‘Memory detail categorization subsection’ of the Methods to clarify the distinction between these two categories.

pp.7-8. The third category, speculations and inferences **about others**, included the participant’s interpretations or assumptions about other people’s thoughts, feelings, or intentions during the recalled event (e.g., “My son wasn’t aware of what was about to happen”), **in contrast to their own internal states. Descriptions in this category were typically based on inferences drawn from others’ observable behaviors or expressions, as participants cannot directly access another person’s internal experiences. However, direct reports of others’ statements or actions (e.g., “She said she was happy”) were not included in this category; instead, they were** classified as specific objective external experiences.

6. Can the authors report ICCs for each detail type? Sorry if I missed it.

: We appreciate the reviewer’s suggestion. Reviewer 2 and Reviewer 3 also requested that we report inter-rater reliability. In our original coding procedure, each narrative was coded by only one annotator due to limited time and resources, making it impossible to compute inter-rater reliability for the full dataset. To address the reviewers’ comment, we estimated inter-rater reliability by recruiting two additional independent annotators and having them code a subset of the data consisting of 40 narratives from 20 participants (including both first and second recalls). These narratives were randomly selected from those originally coded by the annotator who had coded the largest number of narratives. The two additional annotators received the same instructions as the original annotators and categorized every memory detail in each narrative into one of the five nominal categories used in the study (external, internal, inferences about others, general comments, or other).

In our case, Intraclass Correlation Coefficients (ICC) were not appropriate because ICC assumes continuous, interval-scaled measurements. Therefore, we used Fleiss’s Kappa (κ), which is the appropriate statistic for assessing chance-corrected agreement among multiple raters when the data are nominal (Fleiss, 1971). We computed κ by pooling all coded details across the 40 narratives and computing agreement among the three raters. The overall κ was .54 (95% CI = [.52, .56]), indicating moderate agreement according to conventional benchmarks (Landis & Koch, 1977). Category-specific κ values for external, internal, inferences about others, general comments, or other categories were .64, .57, .54, .39, and .42, respectively. When each recall session was considered separately, κ for the first and second recalls was .53 and .57, respectively. Additionally, we calculated percent agreement across

annotators, which had a mean of 59.4% (SD = 18.5) across the 40 narratives. The percent agreement for the first and second recalls was 59.2% (SD = 18.6) and 60.7% (SD = 18.9), respectively. We have revised the Methods section to report the κ values.

p.9. In addition, we retrospectively assessed the inter-rater reliability of memory detail categorization by having two additional independent annotators code a randomly selected subset of data from 20 participants. The overall Fleiss's Kappa across the three annotators (including the original annotator) was .54, indicating moderate agreement. Category-specific Kappa values were .64 for externally oriented, .57 for internally oriented, .54 for inferences about others, .39 for general comments, and .42 for other categories.

7. How do the brief summaries get factored into the scoring? Assuming those would more likely reference “external” content, could that account for the differences between internal and external (i.e., it gives external a leg up)?

: We thank the reviewer for this insightful comment. Reviewer 2 raised a similar point in their Major Methodological Point 1. In our original analysis, we focused only on the full recall narratives and did not examine the brief summaries or their potential effects. We agree with the reviewer that the summaries likely contain more external details, which could inflate recall performance for external details. To address this concern, we conducted an additional analysis of the brief summaries. One additional independent annotator segmented and categorized the summaries following the same procedures used in the original analysis. As expected, the summaries contained more external (65.8%) than internal details (13.6%). To control for the possible influence of summary content, we excluded any memory details from both the first and second recalls that overlapped with those mentioned in the summary. We then compared the proportion of retained and added memory details for external versus internal details using only the remaining (non-overlapping) memory details. The results showed that external details were still significantly more likely to be retained ($t(133) = 5.03, p < .001, 95\% \text{ CI} = [7.54, 17.30]$, Cohen's $d = 0.43$) and added ($t(207) = 16.61, p < .001, 95\% \text{ CI} = [35.95, 45.63]$, Cohen's $d = 1.15$), consistent with our original findings. Thus, the influence of summary details does not appear to be critical to the main effects reported in the manuscript. These additional analyses, along with the descriptive statistics for the summary details, have been included in the Supplementary Results.

Effects of memory details included in the brief summary

In the Hippocampus dataset¹, each participant generated a brief summary (two to three sentences) of the recalled event during the first recall session, which was later used as the memory cue in the second session. Because participants were instructed to include key details in this summary, it is possible that details mentioned in the summary could have influenced recall performance during the second recall. To account for this possibility, we conducted additional analyses of the brief summaries to assess and control for their potential effects.

We recruited one additional independent annotator to segment and code the summary responses for all 210 participants included in our study. Following the same instructions used for the main narrative coding, the annotator divided each summary into single-idea segments and categorized each segment into one of the five detail types (externally oriented, internally oriented, inferences about others, general comments, or other details). Participants' brief summaries contained an average of 32.8 words ($SD = 8.93$; range = 16–61) and 3.74 details ($SD = 1.11$). Importantly, the number of summary details was not significantly correlated with the number of details produced in the second recall ($r(208) = .05$, $p = .434$, 95% CI = $[-.08, .19]$), suggesting that summary length was not a critical factor influencing overall recall performance.

We next examined the distribution of detail categories included in the summaries. On average, participants' brief summaries consisted of 65.8% externally oriented details ($SD = 27.6$), 13.6% internally oriented details ($SD = 19.8$), 1.3% inferences about others ($SD = 5.8$), 18.1% general comments ($SD = 23.0$), and 1.2% other content ($SD = 6.1$). Consistent with the pattern observed in the full recall narratives, participants included significantly more externally oriented than internally oriented details in their summaries ($t(209) = 18.21$, $p < .001$, 95% CI = $[.47, .58]$, Cohen's $d = 1.26$).

To control for the potential influence of differing proportions of externally and internally oriented details in the brief summaries, we re-ran the retention and addition analyses after excluding any memory details that overlapped with those included in the summaries. Because the summary sentences, by definition, often condensed the full narratives, manually identifying specific matching details between summaries and full recalls could be more effortful and more prone to error than matching details across the two full-narrative sessions. Therefore, to apply a more objective criterion, we used semantic similarity scores derived from the SBERT model² to identify overlapping content. For each participant, we computed the cosine similarity between the sentence-embedding vector of every summary detail and every detail from both the first and second recalls. Any detail with a similarity score of .40 or higher was excluded from subsequent analyses. In addition, when a first-recall detail exceeded the .40 threshold, the corresponding matching detail(s) in the second recall were also removed, and vice versa. The .40 threshold was chosen based on validation analyses comparing manually matched versus non-matched detail pairs across sessions (Supplementary Figure 1). Across participants, the average semantic similarity of non-matched pairs consistently fell below .40, whereas the majority of manually matched pairs showed similarity values above this level.

After excluding the details that overlapped with the summary content, the remaining counts were 12.60 details ($SD = 6.85$) for the first recall and 8.81 details ($SD = 5.20$) for the second recall. Among these remaining details, 17.8% ($SD = 19.8$) of all detail types reported in the first recall were retained in the second. Importantly, when externally oriented and internally oriented details were analyzed separately, internally oriented details ($M = 11.7\%$, $SD = 22.1$) were again significantly less likely to be retained than externally oriented details ($M = 24.1\%$, $SD = 28.7$), indicating that internally oriented experiences were more likely to be omitted over time ($t(133) = 5.03$, $p < .001$, 95% CI = $[7.54, 17.30]$, Cohen's $d = 0.43$). Similarly, among the details newly introduced in the

second recall, internally oriented details comprised a much smaller proportion ($M = 11.3\%$, $SD = 15.7$) than externally oriented details ($M = 52.1\%$, $SD = 27.8$; $t(207) = 16.61$, $p < .001$, 95% CI = [35.95, 45.63], Cohen's $d = 1.15$), again reflecting their lower overall prevalence. These results are qualitatively the same as those obtained in the original analysis conducted without excluding summary details, suggesting that presentation of the summary was not a critical driver of the higher retention and addition rates for externally oriented details.

We have also revised the 'The Hippocampus dataset' subsection of the Methods and the 'Retention and addition of memory details' subsection of the Results to reference the Supplementary Results in the main text.

p.5. They **were** also **instructed to** provide a brief summary (2–3 sentences) of the event, including key details to help them remember what they had written when recalling the event in the future (see Supplementary Results for analyses of the memory details contained in these summaries).

pp.13-14. This pattern remained qualitatively unchanged after excluding memory details mentioned in the brief summary sentences used as recall cues (see Supplementary Results). The retention rate was also not significantly influenced by the length of the delay between the two recall sessions for either internally oriented or externally oriented details (see Supplementary Results).

8. I could not make up my mind about any potential circularity in using the narratives themselves to compute emotional intensity/valence. In the end I decided it was ok. But more broadly, how does VADER deal with clauses with ambiguity. My sister died (easy to score), she left me all her books (not so easy to score unless you use the context but even then, is this a good thing or a bad thing?).

: We appreciate the reviewer's question. According to the original published proceeding (Hutto & Gilbert, 2014) and the documentation for the Valence Aware Dictionary and sEntiment Reasoner (VADER; <https://vadersentiment.readthedocs.io/en/latest/index.html>), VADER is a rule-based sentiment analysis tool designed to estimate the emotional tone of short text segments. It relies on a large lexicon of words that were empirically rated by independent human judges, with each word assigned a valence score ranging from -4 ("extremely negative") to $+4$ ("extremely positive"). For example, "okay" has a score of 0.9, "good" 1.9, "great" 3.1, and "horrible" -2.5 . VADER applies a series of linguistic rules, such as handling negations (e.g., "not good"), intensifiers ("very good"), punctuation, and capitalization, to adjust these scores before combining them into a single normalized score between -1 and $+1$. Because VADER relies primarily on the lexical ratings of individual words, it is not sensitive to the broader narrative context. In other words, VADER provides a rough, surface-level estimate of expressed affect rather than a deep contextual interpretation. Thus, ambiguous sentences that contain mostly neutral words and convey emotional significance only through broader context (or information not explicitly stated in the text) will be rated as neutral. We have revised the relevant paragraph

in the ‘Predicting the retention of individual memory details’ subsection of Methods to include these clarifications.

pp.10-11. Specifically, we applied the rule-based VADER (Valence Aware Dictionary and sEntiment Reasoner) sentiment analysis tool⁶¹, implemented in the Natural Language Toolkit (NLTK) Python library (version 3.8.1; <https://www.nltk.org>). **VADER relies on a large lexicon of words that were empirically rated by independent human judges, with each word assigned a valence score. These scores are then adjusted using a set of linguistic rules, such as handling negations (e.g., “not good”), intensifiers (“very good”), punctuation, and capitalization. VADER combines and normalizes this information to generate a single polarity score for each text (i.e., memory detail), ranging from -1 to 1, where negative values indicate more negative sentiment and positive values indicate more positive sentiment. Emotional valence was represented directly by these polarity scores. Emotional intensity was calculated as the absolute value of the polarity score, with higher values indicating stronger emotional expression, regardless of whether the sentiment was positive or negative. Notably, these scores provide surface-level estimates of expressed affect for individual memory details and do not incorporate broader contextual interpretation.**

In addition, we have revised the limitations paragraph in the Discussion to explicitly acknowledge the constraints of relying on purely text-based methods to estimate sentiment or emotion.

p.21. **Fourth**, because the Hippocampus dataset⁴³ did not include participants’ subjective ratings of individual memory details, we relied on text analysis to characterize them. **However, emotion ratings based solely on the text of each individual detail cannot account for the broader context and may not accurately reflect the emotions actually experienced by participants.**

9. A bit more information on the consistency scoring in the main text would be helpful—what are the different possibilities—some details can be consistent, some new etc etc. What is the math that is being done? What happens when a detail is something like, “I was a bit scared” versus “I was a bit nervous” versus “I was a bit apprehensive” —are these all “the same”?

: We thank the reviewer for raising this point. As described in the Methods and Supplementary Table 2, human annotators matched details from the first recall to those from the second recall based on their meaning rather than on exact wording. The matching outcome was binary (i.e., matching vs. non-matching); there was no intermediate category such as “similar but not identical.” Thus, details with similar meanings such as those in the reviewer’s examples would be labeled as matching if the annotators judged them to be sufficiently similar in meaning. This procedure follows the approach used in our prior work on narrative memory, in which we compared written descriptions of movie events with their later narrated recall to identify remembered versus forgotten content (e.g., Lee & Chen, 2022, Nature Communications).

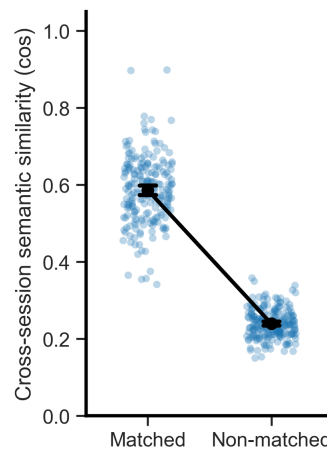
A match indicated that the detail was retained; unmatched details from the first recall were considered omitted. If no match was found for a detail in the second recall, the detail was classified as newly added. The retention rate was calculated as the number of first-recall details with a corresponding match in the second recall divided by the total number of details produced in the first recall. We did not compute a separate omission rate because it is simply $1 - \text{retention rate}$. For memory addition, we calculated the proportion of detail categories within the added details (i.e., the number of newly added details of a given category divided by the total number of added details). We have revised the 'Matching details across recall sessions' subsection of the Methods to clarify the matching procedure and our definitions of memory detail retention, omission, and addition.

p.9. Annotators were given two spreadsheets per participant, one for each recall session, with each memory detail listed in a separate row. Details from the first recall were numbered sequentially. For each detail in the second recall, annotators recorded the number(s) of the corresponding detail(s) from the first recall, if a match was identified (see Supplementary Table 2 for an example). **If no match was found, the cell was left blank. Thus, the matching outcome was binary: a detail from the first recall either matched or did not match a detail from the second recall; no intermediate category such as "similar but not identical" was allowed. Annotators were instructed to use their subjective judgment to pair only those details that described the same content (e.g., actions, people). Exact wording was not required, but the meaning had to be equivalent or highly similar. For example, "I started crying and said, 'yes, definitely yes!'" and "Of course I said yes!" were considered a match. Cross-session semantic similarity was higher for matched compared to non-matched details (Supplementary Fig. 1), confirming that annotators were indeed pairing details that were similar in meaning.**

Once the matching procedure was complete, details from the first recall that had corresponding matches in the second recall were classified as retained, whereas those without a match were classified as omitted. Likewise, details in the second recall that matched details from the first recall were categorized as retained, and those without a match were classified as newly added. The retention rate was calculated as the number of first-recall details with a corresponding match in the second recall divided by the total number of details produced in the first recall. Retention rates were also computed separately within each detail category. For memory addition, we calculated the proportion of each detail category within the newly added details. Specifically, for each category, this proportion was defined as the number of newly added details of that category in the second recall divided by the total number of added details.

Additionally, although not directly requested by the reviewer, we conducted an analysis to validate the manual cross-session matching of memory details. As described above, the meaning-based matching relied on annotators' subjective judgments. If annotators were indeed matching details that were similar in meaning, then semantic similarity, measured using the SBERT model, should be higher for matched details than for non-matched details. For each participant, every memory detail was converted into an SBERT embedding vector. We then computed cosine similarity between each first-recall detail and its corresponding matched detail

from the second recall, and averaged these similarity values across all matched pairs. When a single detail from the first recall corresponded to multiple sentences in the second recall, those sentences were concatenated into one string prior to embedding and similarity computation. For non-matched pairs, we exhaustively calculated cosine similarities for all possible first–second recall detail combinations that had not been annotated as matches, and averaged these values as well. Matched pairs showed a mean similarity of .59 ($SD = .09$), whereas non-matched pairs averaged .24 ($SD = .04$). A paired-samples t -test confirmed that matched details exhibited significantly higher semantic similarity than non-matched details ($t(209) = 59.59$, $p < .001$, 95% $CI = [.34, .36]$, Cohen's $d = 4.11$), validating the meaning-based matching procedure. We have reported this analysis in Supplementary Figure 1.



Supplementary Figure 1. Mean cross-session semantic similarity between matched and non-matched memory details. For each participant, every memory detail was represented as a 768-dimensional embedding vector using the SBERT model². We then computed cosine similarity between each detail from the first recall and its corresponding detail from the second recall. For matched details (as identified by human annotators), we calculated the mean cosine similarity across all matched pairs. If a single detail from the first recall corresponded to multiple sentences in the second recall, those sentences were concatenated into one string prior to embedding and similarity computation. For non-matched pairs, we exhaustively calculated cosine similarities between all possible first–second recall detail combinations not annotated as matches and averaged these values. The mean similarity was significantly higher for matched pairs ($M = .59$, $SD = .09$) than for non-matched pairs ($M = .24$, $SD = .04$; two-tailed paired-samples $t(209) = 59.59$, $p < .001$, 95% $CI = [.34, .36]$, Cohen's $d = 4.11$), indicating that the manual matching procedure effectively identified semantically similar memory detail pairs across sessions. Black circles indicate group means, colored dots represent individual participants, and error bars show 95% confidence intervals.

Relationships to other literature

10. There is a literature on the consistency of emotions themselves. This literature is relevant but potentially overlooked in the paper and I think including it would had additional richness to your work: <https://pubmed.ncbi.nlm.nih.gov/38712802/>

: We thank the reviewer for mentioning this relevant work. We were aware of this study and referenced it, alongside other related works on the consistency of memories of emotions, across

multiple sections of our manuscript (see below).

p.3. Memories of feelings, in particular, are susceptible to bias from one's current emotions and evaluations of the past event (Safer et al., 2001; Castillo et al., 2024; Levine & Safer, 2002)

p.19. Additionally, previous research **assessing the accuracy of** memories of real-world internal experiences—**and comparing them with external experiences**—has often relied on small samples, sometimes including the experimenters themselves (Brewer, 1988; Conway et al., 1996; Larsen & Conway, 1997), or **has** focused on highly selective emotional events such as the 9/11 attacks (Castillo et al., 2024; Hirst et al., 2009; Hirst et al., 2015).

p.20. As a result, recalling internal experiences often involves inference: people reconstruct past thoughts or feelings based on semantic self-knowledge (e.g., “I always get nervous at parties”) or on current appraisals (e.g., “I dislike X, so I must have felt angry”), rather than retrieving the original experience (Robinson & Clore, 2002). This reliance on reconstruction may make internal memories particularly susceptible to bias, as they are often reshaped by current beliefs and goals, as demonstrated in prior work (Levine et al., 2009; Castillo et al., 2024; Levine & Safer, 2002).

11. Two of the main results are quite consistent with two recent papers: Wardell et al., 2025, BRM and Wardell & Palombo, 2025, now in *Emotion*), the former showing that emotions/thoughts are lower in consistency relative to other detail types, the latter showing that emotional narratives are more consistent overall than neutral ones. The authors cite this work but I wonder if it should be highlighted a bit more. Not because it is my own work, but in light of how similar the approach and findings are, especially with Wardell & Palombo, 2025. In my view, it is a strength that this work aligns well as it bolsters confidence in the findings. There are some divergencies important to note as well (In Wardell and Palombo, emotion seemed to affect both “external” and “internal” information, Figure 2C.)—this can be an interesting point of discussion.

: We appreciate the reviewer's suggestion. We agree that the two papers mentioned report results similar to our findings, and it would be helpful to highlight them more explicitly in our manuscript. Regarding the results from Wardell & Palombo (2025), which show that emotion influences both external (event/perceptual) details and internal (emotion) details, we believe these findings are actually consistent—not divergent—with our own. In our study, emotional intensity positively predicted retention of both internal and external details when each feature was modeled separately (Figure 4d), although the effect for external details did not reach statistical significance when other features were included in the model. Our interpretation was that emotional intensity supported the retention of both types of details, but exerted relatively stronger effects on internal details (p.20: “Notably, the effects of emotion and semantic connection extended beyond internal experiences to external ones, suggesting that they reflect general principles supporting the recall of memory details”). As requested by the reviewer, we

have revised the relevant paragraphs in the Discussion to better incorporate the findings of these two papers and have updated the reference for Wardell & Palombo (2025) to the published version.

p.20. Our findings indicate that internal experiences are frequently remembered in everyday life, but they are also more vulnerable to **omission** and **semantic** distortion than external experiences, in line with prior empirical work^{21,22,24}. **A recent study examining repeated autobiographical recalls similarly found that thoughts and emotions were recalled less consistently than other types of memory details³⁰.**

p.20. Prior studies have also shown that emotional autobiographical events are remembered with greater detail and vividness than neutral ones^{40,48,60}. **Similarly, emotion has been found to enhance the consistency of autobiographical recall across repetitions, improving both externally and internally oriented information⁴⁰.**

12. The authors write: “This raises the question of whether remembering internal thoughts and feelings influences how people evaluate the significance of a recalled event. To test this, we used participants’ self-reported ratings of the perceived importance of their memories (i.e., how impactful, significant, or personal the event felt) on a 5-point scale, as provided in the Hippocampus dataset” —this statement appears causal. Can’t the opposite be true (more important leads to more consistency?). The authors address this later in the discussion but I would caveat this sentence as well as it could be misleading.

: We agree with the reviewer that the original sentence could be misleading, as it implies a specific causal direction (i.e., remembering internal thoughts affects perceived importance). We have revised the sentence to remove any assumption of causality:

p.17. Prior work shows that emotion enhances phenomenological qualities of memory, such as vividness, reliving, and perceived importance^{54,55,63}. This raises the question of whether remembering internal thoughts and feelings **is associated with** how people evaluate the significance of a recalled event.

13. I liked the authors discussion of limitations and the distinction between forgetting and choosing not to share a detail. I wonder if the authors could speculate on whether their results would look different if they were sampling narratives from people who were sharing their memory to the same audience versus a different one. After all, in real life, we tend to share our memories across distinct contexts and conditions. Who we narrate to and who we narrate with (shared events) will change the memories. Perhaps inconsistency is underestimated when we create identical recall scenarios (as in a laboratory study).

: We thank the reviewer for this thoughtful comment. Reviewer 3 also raised a related point (Comment 2) regarding the communicative function of memory recall. We agree that the content and quality of autobiographical memory reports can vary across social contexts and communicative goals, as shown in the prior work cited in Reviewer 3’s comment (e.g.,

Dudukovic, Marsh & Tversky, 2004). Thus, we would expect recall narratives to be less consistent across sessions if participants were addressing different audiences. In our dataset, all repeated recalls were generated under the same task instructions and context (i.e., writing a diary-like narrative to oneself, with the possibility of being read by experimenters), so the inconsistency observed here may represent a lower bound on the variability found in real-world settings. These contextual effects may also be relevant for internally oriented experiences, which are the primary focus of our study. For example, people may be more likely to share internal thoughts and feelings with close others than with unfamiliar individuals such as experimenters.

We have revised the limitations paragraph in the Discussion to incorporate this point, along with the potential reasons for memory detail omission raised by Reviewer 3. We kept the discussion relatively brief because our manuscript was already far above the journal's recommended word limit (~5,000 words), but we would be happy to elaborate further should the reviewer or editors consider additional detail necessary.

p.21. Second, in narrated free recall, it is difficult to determine whether the absence of certain details reflects genuine forgetting or a choice not to report them. **Because memory recall is often a communicative act⁸⁴, participants may omit details they perceive as redundant, peripheral to the task goal, or irrelevant to a potential audience^{85,86}. Details omitted due to retrieval neglect can also re-emerge when appropriate cues are provided²³. Indeed, details missing from an initial recall frequently appeared in later recalls, as shown in prior studies^{30,62} and in the present work. Future research could address this issue by incorporating detailed probes or prompts following narrated free recall. Third, **also related to the communicative function of memory recall, the quality and characteristics of autobiographical narratives may vary depending on the social context in real life⁸⁵⁻⁸⁷. For example, people may share more thoughts and feelings when recalling memories for close others than they do in experimental contexts such as the one used in the present study.****

-Daniela Palombo

Reviewer #2 (Remarks to the Author):

This study aimed to examine whether there exist differences in how internal and external experiences are remembered and preserved across time. The authors found that internal experiences were more susceptible to forgetting than external experiences, however, this susceptibility to forgetting may be lessened by factors such as high emotional intensity and increased semantic connections. Moreover, internal, but not external details promote a greater sense of perceived importance. Overall, this manuscript was well written. However, there are several major conceptual/theoretical and methodological concerns that need to be addressed and the novelty of this work needs to be better established.

: We thank the reviewer for their valuable and important comments on our work.

Major Conceptual/Theoretical Points:

1. I do not find the study as novel as the authors claim it to be. While their analyses are novel, other research has already demonstrated the expected effects (as the authors themselves cite). For example, there are many studies of narrative and narrative coherence (including number of internal details, for example) that are for single events as opposed to the life story. Tilman Habermas, David Pillemer, Robyn Fivush, Adrial Boals, and David Rubin all have published studies looking at the narrative content of individual events. The authors might consider specifically the literature on meaning-making, which includes the use of internal details as individuals try to make sense of their memories. As another example, the authors claim in the discussion that previous research has had small sample sizes; there is a robust literature on autobiographical memory narratives and phenomenology that includes robust sample sizes.

: We are grateful to the reviewer for acknowledging the novelty of our methodological approach and for directing us to relevant prior work that we had not been aware of. As the reviewer noted, we cited many previous studies in cognitive psychology examining memories of thoughts and feelings in the Introduction and Discussion to appropriately situate our work within the existing literature in the field. However, we agree that additional relevant studies may not have been recognized in the original manuscript, particularly those grounded in perspectives beyond cognitive psychology, such as clinical psychology, developmental psychology, or psychotherapy. Following the reviewer's suggestion, we conducted a more extensive literature search, focusing in particular on the authors highlighted by the reviewer. We now cite these studies in the appropriate places in the Introduction and Discussion sections and discuss their relevance, especially regarding meaning-making. We kept the discussion relatively brief because our manuscript was already far above the journal's recommended word limit (~5,000 words), but we would be happy to elaborate further if the reviewer and editors feel that additional detail is necessary. We would also be happy to incorporate further citations if the reviewer has additional recommendations, and we would appreciate full reference information for any specific papers they would like us to include.

p.3. However, **relatively** little is known about how memories of naturally occurring internal experiences are expressed and preserved over time within autobiographical narratives. This gap is notable because narrative—accounts of temporally unfolding, interrelated events involving agents, goals, and actions—serves as a primary format through which real-life episodes are stored, retrieved, and studied (Lee et al., 2020; Palombo, 2024; Rubin & Greenberg, 2003; **Nelson & Fivush, 2020; Habermas & Bluck, 2000**). In autobiographical memory research, participants typically generate narratives by describing, in their own words, **what** they remember about a specific life event (Levine et al., 2002; Wardell & Palombo, 2025; **Rubin, 2011; Habermas & Berger, 2011**).

p.21. More broadly, our findings suggest that internal experiences may serve as a key bridge between event and self, capturing not only what happened but also its personal meaning—consistent with the view that autobiographical memory supports self-

coherence (Conway, 2005; Conway & Pleydell-Pearce, 2000; Bluck et al., 2005). These findings also align with perspectives showing that people make meaning from life events by constructing narratives around their memories (Habermas & Bluck, 2000; Fivush et al., 2017; Pillemer, 2001), with thoughts and emotions playing a significant role in this process (Fivush et al., 2008; Habermas, 2019).

p.22. These findings may also have implications for clinical contexts, given that the ways people recall their thoughts and emotions are often associated with mental health and well-being (Booker et al., 2020; Boals & Klein, 2005; Vanaken et al., 2021).

Regarding our statement in the Discussion about prior studies with relatively small sample sizes, we were specifically referring to work in which the primary research design and questions focused on assessing the accuracy of memories of thoughts and feelings in real-world settings and/or comparing them with external details. The reviewer is correct that many studies of autobiographical narratives included several hundred participants or more; however, these larger studies did not always center on contrasting the accuracy (as estimated from repeated recalls) of memories of internal experiences with those of external events. We have revised the relevant paragraph in the Discussion to clarify this point.

p.19. Additionally, previous research assessing the accuracy of memories of real-world internal experiences—and comparing them with external experiences—has often relied on small samples, sometimes including the experimenters themselves^{21,22,64}, or has focused on highly selective emotional events such as the 9/11 attacks^{24,32,56}.

We have also revised a sentence in the same paragraph to clarify the scope of the research discussed in the paragraph.

p.19. Internal thoughts and feelings have rarely been the central focus of cognitive psychology research on memory using autobiographical narratives, even though they are commonly recalled and explicitly included as a distinct category in standardized coding protocols such as the Autobiographical Interview³⁹.

2. The authors note throughout the paper that they are studying memories of thoughts and feelings (e.g., noted on page 16); it is not entirely clear given their methodology that is the case. Rather, the authors are studying internal details that people recall when remembering external events. The paper should be revised to capture the nuance that they are studying internal thought or feelings that come up as part of the reconstructive process.

: We thank the reviewer for their comment. If we have understood the concern correctly, the main issue is whether the internal thoughts and feelings targeted in our study may instead reflect experiences that occurred during recall, rather than during the original event. The reviewer is correct that recall narratives often included thoughts and feelings likely experienced at the moment of recall rather than during the original event, reflecting the reconstructive nature of remembering (e.g., “It was one of the happiest moments of my life”). Importantly, however,

we defined a separate memory detail category, “general comments or evaluations”, to identify and distinguish these from the internal details of interest. As described in the ‘Memory detail categorization’ subsection of Methods (copied below), this category differs from internal details in terms of timing: it captures post-event thoughts and feelings generated during recall, rather than emotions experienced during the original event. Because these details were coded separately and excluded from analysis, the internal details included in our analyses are likely to primarily reflect actual thoughts and feelings that occurred during the recalled event.

p.8. The fourth category, general comments or evaluations, encompassed reflections, interpretations, or assessments made about the event during the memory recall sessions (e.g., “It was one of the happiest moments of my life”). A key distinction between this category and subjective internal thoughts or feelings lies in the timing of the reflection. For example, the statement “I had a great time” would be classified as a general comment or evaluation, as it reflects a post-event assessment made after the experience had ended. In contrast, “I was having a great time” would be categorized as a subjective internal thought or feeling, as it describes an emotional experience occurring during the event itself.

To clarify the distinction between internal experiences that occurred during the original event and thoughts or feelings that arose during the act of recall as a reconstructive process, we have also revised a paragraph in the ‘Memory detail categorization’ subsection of the Methods.

p.8. The **categories of ‘speculations or inferences about others’ and ‘general comments or evaluations’ were additionally introduced to help annotators avoid misclassification. The former distinguished other people’s inferred internal experiences from participants’ own, while the latter separated thoughts or feelings generated during the act of recall from those that occurred during the original event. Because internally oriented details and these additional types of details all described internal states using similar language, they were easily confused. By explicitly flagging these categories for annotators and instructing them to classify such details separately, we aimed to more effectively identify and exclude them from the main analysis—together with the ‘other’ category—as they were not relevant to the study’s primary goal.**

In addition, if participants mentioned any thoughts or feelings that were not specifically tied to the event being recalled (e.g., thoughts or feelings experienced while imagining the future during the recall), they were categorized as “other” and also excluded from analysis. We have added a new paragraph explaining the “other” category in more detail to address this point and to respond to Reviewer 3’s minor comment 3.

p.8. **Finally, details that did not align with any of the categories described above were classified as ‘other.’ These were not directly related to the main recalled event and included, but were not limited to, general semantic knowledge about oneself or others (e.g., “He is normally a happy guy”), meta-memory comments (e.g., “I can’t remember”), thoughts or feelings about the future related to the recalled event (e.g., “I will have to get**

another job”), and recollections of past events other than the main event being recalled (e.g., “Last year I didn’t have a very good birthday”).

However, we also acknowledge that even the specific recall of internal thoughts and feelings experienced during the original event may be susceptible to the reconstructive nature of memory recall. Reconstruction can occur in any format of recall and for any category of memory detail. However, internal experiences may be especially vulnerable to this process compared to objective facts, as demonstrated by prior studies showing that remembered emotions can be biased to align with one’s current states, such as emotions, goals, and attitudes (e.g., Safer et al., 2001; Levine & Safer, 2002; Castillo et al., 2024). We incorporated this point into our manuscript by mentioning this literature in the Introduction as relevant prior work and as part of the rationale for hypothesizing that internal details may be less well remembered in autobiographical narratives than external details.

p.3. In these studies, internal experiences are generally remembered less consistently than external aspects of events, such as actions^{21,22,24}. Memories of feelings, in particular, are susceptible to bias from one’s current emotions and evaluations of the past event^{31–33}.

p.4. We hypothesized that memories of internal experiences would be more susceptible to forgetting and **semantic** distortion than memories of external experiences, since internal thoughts and feelings generally lack sensory or perceptual anchors¹¹ and are readily reshaped by current goals and appraisals, **as shown in prior work**^{1,33}.

We also addressed this susceptibility to reconstruction in the Discussion when interpreting our results.

p.20. Our findings indicate that internal experiences are frequently remembered in everyday life, but they are also more vulnerable to **omission** and **semantic** distortion than external experiences, in line with prior empirical work^{21,22,24}. **A recent study examining repeated autobiographical recalls similarly found that thoughts and emotions were recalled less consistently than other types of memory details**³⁰. Unlike externally oriented events, which benefit from multisensory cues (e.g., visual, auditory, spatial) that provide multiple retrieval pathways, internal states generally lack such direct perceptual anchors^{11,32}. They are also less likely to form rich, image-based traces that support episodic recollection²². As a result, recalling internal experiences often involves inference: people reconstruct past thoughts or feelings based on semantic self-knowledge (e.g., “I always get nervous at parties”) or on current appraisals (e.g., “I dislike X, so I must have felt angry”), rather than retrieving the original experience⁶⁵. This reliance on reconstruction may make internal memories particularly susceptible to bias, as they are often reshaped by current beliefs and goals, **as demonstrated in prior work**^{1,32,33}.

3. At several points in the Introduction and the Discussion, the authors argue that internal

experiences are more susceptible to forgetting and distortion than observable external experiences. While I agree that the data suggests internal experiences are more prone to forgetting, it is not entirely clear what finding points to higher rates of “distortion” or exactly what “distortion” means in this context. How are the authors defining distortion? Does this refer to participants adding additional details in the second recall (which, based on the data, would point in the opposite direction)? In other words, aren’t the changes that are specific to internal experiences all to do with forgetting? If so, the authors should clarify what distortion means in this context and what data specifically reflect distortion. See below for a concern regarding how the memories were sampled and how the addition/subtraction of details was considered by the authors.

: We appreciate the reviewer’s suggestion for clarification. In the current manuscript, we used the term “distortion” to refer to semantic changes in the descriptions of recalled memory details, as opposed to “forgetting” (now replaced with “omission” in response to Reviewer 3’s comment), which refers to the failure to mention a memory detail during recall. As defined in this way, “distortion” is specifically quantified as the semantic similarity between the first and second recalls of the same memory detail across sessions, as reported in the “Semantic similarity across recall sessions” subsection of the Results. Following the reviewer’s suggestion, we have clarified this definition by elaborating the relevant sentence in that subsection and by revising the term “distortion” to “semantic distortion” throughout the manuscript.

p.2. We found that internal experiences were more prone to **omission and semantic distortion** over time compared to external ones.

p.4. We hypothesized that memories of internal experiences would be more susceptible to forgetting and **semantic distortion** than memories of external experiences, since internal thoughts and feelings generally lack sensory or perceptual anchors¹¹ and are readily reshaped by current goals and appraisals, **as shown in prior work^{1,33}**.

p.14. Thus far, we have analyzed the binary presence or absence of specific memory details. However, even when a detail was present in both recalls, its precise meaning may have shifted over time, **reflecting semantic distortion**.

p.15. If memories of internal experiences are more susceptible to **omission and semantic distortion** than memories of external experiences, what factors might make them more durable?

p.19. Using natural language processing to analyze a large-scale dataset of autobiographical narratives (the Hippocampus dataset^{43,44}), we found that internal experiences were more susceptible to **omission and semantic distortion** than observable external experiences.

p.20. Our findings indicate that internal experiences are frequently remembered in everyday life, but they are also more vulnerable to **omission and semantic distortion** than

external experiences, in line with prior empirical work^{21,22,24}.

4. The authors suggest that internal experiences are particularly biased by one's current emotions, beliefs, goals, and evaluations of past events. When mentioned in the Introduction and subsequently as an explanation of the findings in the Discussion, I expected there to be data that could speak to this point. For example, a measurement of participants' general emotional state at the time of recollection. Was there data to speak to this explanation specifically? And if not, the authors might consider rephrasing their discussion of this explanation as to not imply this could be measured in the present study.

: We thank the reviewer for their comment. Our mention of internal experiences being vulnerable to bias from one's current emotions, goals, and appraisals was intended to reference prior work (e.g., Safer et al., 2001; Levine & Safer, 2002; Castillo et al., 2024), not to suggest that we would test this idea using the Hippocampus dataset in the present study. These prior studies typically asked participants to rate the emotions they experienced during (or immediately after) highly salient events (e.g., the 9/11 attacks, the death of a spouse) and later re-rate those emotions during recall, demonstrating that remembered emotions often shift to align with current attitudes, emotions, and appraisals. We cited this literature to highlight a previously established susceptibility to bias that might help explain why internal experiences could be less likely to be retained (or more likely to undergo semantic change) over time compared to external details. We have revised the relevant sentences in the Introduction and Discussion to clarify that these points refer to findings from prior research.

p.4. It also helps reduce potential demand characteristics associated with direct questioning, which may be especially relevant for memories of internal experiences given their **previously reported** vulnerability to distortion³³.

p.4. We hypothesized that memories of internal experiences would be more susceptible to forgetting and **semantic** distortion than memories of external experiences, since internal thoughts and feelings generally lack sensory or perceptual anchors¹¹ and are readily reshaped by current goals and appraisals, **as shown in prior work**^{1,33}.

p.20. This reliance on reconstruction may make internal memories particularly susceptible to bias, as they are often reshaped by current beliefs and goals, **as demonstrated in prior work**^{1,32,33}.

Major Methodological Points:

1. A methodological artifact may be driving these results. Participants were asked to provide a brief summary (2-3 sentences) of their event to be used as a cue for the second recall. At this point, were the participants specifically told that they were providing a summary of the event to be used for a later recall? Additionally, was the length of these statements controlled (either at the time of recollection or in data preprocessing)? It is possible that some participants gave more details in the brief summary statement that they could use to promote remembering of

more details during the second recollection. The authors should provide more details on participants' brief summaries, including length and number of details. If there was a lot of variability in length or number of details, it might be wise to control for this in the analyses. I understand that these are existing data so the information may not be available, but the authors should mention these as potential limitations if unable to control for it.

: We appreciate the reviewer for raising this important point. Reviewer 1 raised a similar point in their Comment 7. First, participants were informed during the initial recall session that the summary they provided would serve as a cue for a later recall. They were instructed to include key events from the story so that they could easily identify the specific event they had described when recalling it again in the future. The required length of the summary was 2–3 sentences (100–300 characters, including spaces). We have revised 'The Hippocampus dataset' subsection of the Methods to clarify this point.

p.5. They **were also instructed to provide a brief summary (2–3 sentences) of the event, including key details to help them remember what they had written when recalling the event in the future (see Supplementary Results for analyses of the memory details contained in these summaries).**

Since our original analysis focused only on the full recall narratives, we did not analyze the brief summaries and therefore did not examine or control for their length. To address the reviewer's concern, we conducted an additional analysis of the brief summaries. An independent annotator segmented and categorized the summaries following the same procedures used in the original analysis. On average, participants produced 32.8 words ($SD = 8.93$; range = 16–61) and 3.74 details ($SD = 1.11$) in their brief summaries. Importantly, the number of summary details was not associated with the number of details produced in the second recall ($r(208) = .05$, $p = .434$, 95% CI = $[-.08, .19]$), suggesting that summary length did not inflate overall recall. This is likely because the summary length was strictly limited to 2–3 sentences regardless of the length of participants' full narratives.

However, the summaries generally contained more external details (65.8%) than internal details (13.6%), which could have inflated recall performance for external details. To control for the possible influence of summary content, we excluded any memory details from both the first and second recalls that overlapped with those mentioned in the summary. We then compared the proportion of retained and added memory details for external versus internal details using only the remaining (non-overlapping) memory details. The results showed that external details were still significantly more likely to be retained ($t(133) = 5.03$, $p < .001$, 95% CI = $[7.54, 17.30]$, Cohen's $d = 0.43$) and added ($t(207) = 16.61$, $p < .001$, 95% CI = $[35.95, 45.63]$, Cohen's $d = 1.15$), consistent with our original findings. Thus, the influence of summary details does not appear to be critical to the main effects reported in the manuscript. These additional analyses, along with the descriptive statistics for the summary details, have been included in the Supplementary Results.

Effects of memory details included in the brief summary

In the Hippocampus dataset¹, each participant generated a brief summary (two to three sentences) of the recalled event during the first recall session, which was later used as the memory cue in the second session. Because participants were instructed to include key details in this summary, it is possible that details mentioned in the summary could have influenced recall performance during the second recall. To account for this possibility, we conducted additional analyses of the brief summaries to assess and control for their potential effects.

We recruited one additional independent annotator to segment and code the summary responses for all 210 participants included in our study. Following the same instructions used for the main narrative coding, the annotator divided each summary into single-idea segments and categorized each segment into one of the five detail types (externally oriented, internally oriented, inferences about others, general comments, or other details). Participants' brief summaries contained an average of 32.8 words ($SD = 8.93$; range = 16–61) and 3.74 details ($SD = 1.11$). Importantly, the number of summary details was not significantly correlated with the number of details produced in the second recall ($r(208) = .05$, $p = .434$, 95% CI = $[-.08, .19]$), suggesting that summary length was not a critical factor influencing overall recall performance.

We next examined the distribution of detail categories included in the summaries. On average, participants' brief summaries consisted of 65.8% externally oriented details ($SD = 27.6$), 13.6% internally oriented details ($SD = 19.8$), 1.3% inferences about others ($SD = 5.8$), 18.1% general comments ($SD = 23.0$), and 1.2% other content ($SD = 6.1$). Consistent with the pattern observed in the full recall narratives, participants included significantly more externally oriented than internally oriented details in their summaries ($t(209) = 18.21$, $p < .001$, 95% CI = $[.47, .58]$, Cohen's $d = 1.26$).

To control for the potential influence of differing proportions of externally and internally oriented details in the brief summaries, we re-ran the retention and addition analyses after excluding any memory details that overlapped with those included in the summaries. Because the summary sentences, by definition, often condensed the full narratives, manually identifying specific matching details between summaries and full recalls could be more effortful and more prone to error than matching details across the two full-narrative sessions. Therefore, to apply a more objective criterion, we used semantic similarity scores derived from the SBERT model² to identify overlapping content. For each participant, we computed the cosine similarity between the sentence-embedding vector of every summary detail and every detail from both the first and second recalls. Any detail with a similarity score of .40 or higher was excluded from subsequent analyses. In addition, when a first-recall detail exceeded the .40 threshold, the corresponding matching detail(s) in the second recall were also removed, and vice versa. The .40 threshold was chosen based on validation analyses comparing manually matched versus non-matched detail pairs across sessions (Supplementary Figure 1). Across participants, the average semantic similarity of non-matched pairs consistently fell below .40, whereas the majority of manually matched pairs showed similarity values above this level.

After excluding the details that overlapped with the summary content, the remaining counts were 12.60 details ($SD = 6.85$) for the first recall and 8.81 details (SD

= 5.20) for the second recall. Among these remaining details, 17.8% ($SD = 19.8$) of all detail types reported in the first recall were retained in the second. Importantly, when externally oriented and internally oriented details were analyzed separately, internally oriented details ($M = 11.7\%$, $SD = 22.1$) were again significantly less likely to be retained than externally oriented details ($M = 24.1\%$, $SD = 28.7$), indicating that internally oriented experiences were more likely to be omitted over time ($t(133) = 5.03$, $p < .001$, 95% CI = [7.54, 17.30], Cohen's $d = 0.43$). Similarly, among the details newly introduced in the second recall, internally oriented details comprised a much smaller proportion ($M = 11.3\%$, $SD = 15.7$) than externally oriented details ($M = 52.1\%$, $SD = 27.8$; $t(207) = 16.61$, $p < .001$, 95% CI = [35.95, 45.63], Cohen's $d = 1.15$), again reflecting their lower overall prevalence. These results are qualitatively the same as those obtained in the original analysis conducted without excluding summary details, suggesting that presentation of the summary was not a critical driver of the higher retention and addition rates for externally oriented details.

We have also revised the 'Retention and addition of memory details' subsection of the Results to reference the Supplementary Results in the main text.

pp.13-14. Importantly, when externally and internally oriented details were analyzed separately (Fig. 3a), internally oriented details ($M = 38.3\%$, $SD = 32.7$) were significantly less likely to be retained than externally oriented details ($M = 48.9\%$, $SD = 19.2$), indicating that internal experiences are more susceptible to omission over time ($t(178) = 3.73$, $p < .001$, 95% CI = [4.98, 16.19], Cohen's $d = 0.39$). This pattern remained qualitatively unchanged after excluding memory details mentioned in the brief summary sentences used as recall cues (see Supplementary Results).

2. Additionally, according to the methods, participants were asked to remember an event from the past “three to six months”. This implies that participants should NOT recall an event that occurred within 3 months of the first recollection. However, the authors then report that they “included participants who recalled an event that occurred within three months (90 days) prior to the first recall session.” If my understanding is correct and participants followed instructions, there should be no memories that occurred within 3 months of the first recall session. Moreover, the average age of events at the first recall was 41.8 days old (just over 1 month). The authors should provide clarification of these points.

: We thank the reviewer for their careful consideration. We used the phrase “within the past three to six months” to remain consistent with the original study instructions (Sap et al., 2020), which stated: “a memorable event that happened in the last three to six months” and “(an event that) was recent, i.e., within the last 3–6 months.” The reviewer is correct that, in a literal sense, “in the last three to six months” could exclude events younger than three months. However, in conversational contexts, this phrase could also be interpreted more broadly as “within the past six months (approximately three to six months ago),” so participants may have reported more recent memories unless explicitly told otherwise. Indeed, several hundred participants reported memories younger than three months, suggesting that participants commonly understood the

instruction in this more flexible, conversational sense, focusing on “recent” rather than the exact age of the memory (or disregarded the precise timeframe). Following the reviewer’s request to avoid confusion, we have slightly revised the sentence to de-emphasize the exact timeframe.

p.5. In the first session, participants were instructed to write a 15–25 sentence diary-like narrative about a memorable event that had occurred **recently (approximately** within the past three to six months).

However, because the phrase “three to six months” comes directly from the original study’s instructions, we chose to retain it. That said, we would be happy to change it to “within the past six months” to further minimize confusion, if the reviewer and editors consider it necessary.

3. Also related to how participants were prompted to recall a memory, were participants told to recall a “specific” event that is not repeated (e.g., they recalled a hike, but they hike every Sunday)? If not, it is possible that participants recalled an event that was repeated several times between the first and second recall sessions. This may explain forgetting and/or added details and should be considered more thoroughly.

: We thank the reviewer for this important question. In the original data collection (Sap et al., 2020), participants were instructed to recall a recent memorable event, which implied that it should be a specific past occurrence. However, there were no instructions explicitly preventing participants from recalling an event that might have occurred more than once, making it possible that some participants repeatedly experienced events similar to the one recalled. We agree with the reviewer that it would be interesting to examine the potential effects of event repetition, should it have occurred. Unfortunately, however, because the Hippocampus dataset did not include information about whether similar events occurred between the two recall sessions, it does not allow us to determine if such repetitions took place or to systematically assess their effects.

That said, we believe that participants’ recalling of nonspecific events was unlikely to be a major factor driving the results reported in our study. Because participants were instructed to recall the same event they had described in the first session, prompted by a summary cue, it is likely that they recalled that particular event even if similar ones occurred between sessions, assuming they were healthy individuals with intact episodic memory. Furthermore, to ensure that our analyses focused only on specific events, we performed a data selection step in which we included only participants whose first and second recalls described the same single event that occurred within one day. We have revised the Methods section to clarify this step.

p.6. **Next, to focus on memories of clearly identifiable, specific events,** we included only participants whose first and second recalls described the same specific event that occurred within a single day. Participants **were excluded if they** recalled different events across sessions, **described situations that unfolded over several days or weeks (e.g., changes in life after returning to college), provided only general statements about recurring events,** or **produced** narratives **that** were disorganized and lacked a clear episodic **focus**. This selection was performed manually by one of the authors (C.K.), who

reviewed the content and quality of both recalls. This step resulted in a remaining pool of 330 participants.

Finally, in any cases where participants mentioned repeated (similar) events that occurred between the first and second recalls, those details were coded as “other details unrelated to the main event recalled” and excluded from analysis. Therefore, it is unlikely that they affected the reported comparisons between internal and external details.

4. The analysis of the retention and addition of details is flawed. When analyzing the retention and addition of memory details, you show that internal details made up a significantly lower proportion of added details compared to external details. While interesting, this is not all that surprising given the overall finding that internal details make up a smaller proportion of all recalled details, as you suggest. A more insightful analysis would assess whether participants are more likely to generate new external details by accounting for the fact that internal details are generally recalled at lower rates. For example, you could calculate the proportion of new internal details relative to the overall number of internal details recalled and compare to the proportion of new external details relative to the overall number of external details.

: We appreciate the reviewer’s thoughtful attention to this part of the data. The reviewer is correct that the higher proportion of external vs. internal details among newly generated details largely reflects the overall predominance of external details in recall. As the reviewer noted, this pattern is not unexpected, but it still remains informative because it complements the overall proportion analyses in Fig. 2: it shows that the relative scarcity of internal details is evident not only in the details recalled initially but also in the new details introduced during the second recall. Importantly, if the pattern had been different, for example, if newly added details were predominantly internal, this would have warranted a very different interpretation, such as the possibility that internal details play a larger role in memory reconstruction during repeated recall. We have clarified this point in the revised text in the ‘Retention and addition of memory details’ subsection of the Results.

p.14. On average, 55.7% ($SD = 17.6$) of the details reported in the second recall were newly introduced. Among these new details, internally oriented details comprised a significantly smaller proportion ($M = 10.5\%$, $SD = 11.6$) than externally oriented details ($M = 54.0\%$, $SD = 24.8$; $t(208) = 21.05$, $p < .001$, 95% CI = [39.45, 47.60], Cohen’s $d = 2.25$). This pattern mirrors the overall proportions reported in Fig. 2, indicating that the relative scarcity of internally oriented details is evident not only in the details recalled initially but also in the new details introduced during the second recall.

We also agree with the reviewer that it is informative to examine the proportion of newly added internal and external details relative to the total number of details of the same type in the second recall. In response to the reviewer’s request, we performed this analysis and found that, on average, 43.9% ($SD = 39.8$) of internal details and 48.6% ($SD = 20.6$) of external details were newly generated; this difference was not statistically significant ($p = .094$). We now report this result in the ‘Retention and addition of memory details’ subsection of the Results.

p.14. However, within each category, the proportion of newly added details relative to all details of the same type did not differ significantly between internally oriented ($M = 43.9\%$, $SD = 39.8$) and externally oriented details ($M = 48.6\%$, $SD = 20.6$; $t(209) = 1.68$, $p = .094$, 95% CI = $[-.01, .10]$, Cohen's $d = .12$).

It is important to note, however, that these relative proportions do not directly address our primary research question, namely, whether internal or external details were better remembered across recalls. A high proportion of newly added details could reflect lower retention of originally recalled details, increased likelihood of generating additional details during the second recall, or a combination of both. Thus, these proportions alone do not provide a clear interpretation regarding the relative memorability of internal versus external details.

5. The preserved recall of external details may have less to do with their external nature and more to do with schema and event sequencing. An external detail (e.g., location) is likely to cue other external details and aid in the preservation or addition of these details on subsequent recalls. For example, work by Marsh on retellings suggests that recall structure (schema) guides and biases later event memories. As a note, her work also includes data on internal details.

: We thank the reviewer for their thoughtful comments and for directing us to relevant prior work. Reviewer 1 and Reviewer 3 also raised related points (with an emphasis on the communicative goals of recall), and Reviewer 3 similarly cited work by Marsh and colleagues (e.g., Dudukovic, Marsh & Tversky, 2004; Marsh, 2007). We agree that recall structure (schemas, goals, or contexts) can influence the quality and characteristics of narrated recall. In the Hippocampus dataset used in this study, all repeated recalls were produced under relatively neutral task instructions and context (i.e., writing a diary-like narrative to oneself, with the possibility of being read by experimenters), without explicitly encouraging participants to emphasize external over internal details. Nonetheless, if participants implicitly perceived the task as requiring an accurate recounting of what happened, it is possible that external details were prioritized in their narratives. In response to the reviewer's comment (along with comments from Reviewer 1 and Reviewer 3), we have revised the limitations paragraph in the Discussion to address the potential effects of recall contexts and goals, and we now cite the relevant literature. We kept the discussion concise in line with the journal's recommended word limit, but we would be happy to elaborate further if the reviewer or editors find additional detail necessary.

p.21. Second, in narrated free recall, it is difficult to determine whether the absence of certain details reflects genuine forgetting or a choice not to report them. **Because memory recall is often a communicative act⁸⁴, participants may omit details they perceive as redundant, peripheral to the task goal, or irrelevant to a potential audience^{85,86}. Details omitted due to retrieval neglect can also re-emerge when appropriate cues are provided²³. Indeed, details missing from an initial recall frequently appeared in later recalls, as shown in prior studies^{30,62} and in the present work. Future research could address this issue by incorporating detailed probes or prompts following narrated free recall. Third, also related to the communicative function of memory recall,**

the quality and characteristics of autobiographical narratives may vary depending on the social context in real life^{85–87}. For example, people may share more thoughts and feelings when recalling memories for close others than they do in experimental contexts such as the one used in the present study.

We also agree with the reviewer that an external detail can cue other external details and support their preservation or reappearance in subsequent recalls. Indeed, our mixed-model logistic regression analysis reported in the ‘Features predicting memory detail retention’ section illustrates this pattern clearly: external details that had stronger semantic connections with other external details were more likely to be retained in the second recall. Importantly, however, this effect was not unique to external details. Internal details were also more likely to be retained when they were semantically connected to external details, as shown in Figure 4. The parameter estimates for semantic connections with external details were not lower for internal details (global = .24, local = .27) than for external details (global = .22, local = .16). In other words, external details can cue both external and internal details, thereby promoting their retention across recall sessions.

6. On the perceived importance of memories (page 14) - memories of internal experiences are more “emotionally charged”. This represents a design flaw; by definition, emotion is internal. Additionally, the correlation between internal details and perceived importance is also flawed. Importance, which might be a corollary for meaning-making, would also be an internal detail. We tend to ruminate on more important events.

: We thank the reviewer for raising this point. Reviewer 3 also commented on a related issue in their Minor Comment 6 regarding the emotional nature of internal details. The statement that “memories of internal experiences...tend to be more emotionally charged” refers back to the descriptive results shown in Table 1, which were included solely to summarize the characteristics of the memory details. As both reviewers noted, the higher emotional intensity associated with internal details was an expected outcome based on their definition in the current study, and was presented primarily as a sanity check rather than as a novel finding of interest (as also acknowledged by Reviewer 3). We have slightly revised the relevant sentence in the “Features predicting memory detail” retention subsection to clarify that this was an expected result reflecting the inherent nature of internal details.

p.15. Overall, compared to externally oriented details, internally oriented details were more emotionally intense, consistent with their definition, but showed weaker semantic links to externally oriented details and were shorter in length (Table 1).

Regarding the correlation between internal details and perceived importance, we agree with the reviewer that perceived importance is an internally oriented construct and is closely tied to the meaning-making functions of autobiographical memory. We have acknowledged this point in the Discussion section of the manuscript.

pp.20-21. Our results showed that memories containing more internal thoughts and

feelings were perceived as more important. Because this relationship is correlational, at least two explanations are possible. One possibility is that events of greater personal significance naturally elicit stronger internal experiences during encoding¹, which in turn are more likely to be recalled due to the memory-enhancing effects of emotion⁷⁴. Alternatively, recalling more internal experiences may enhance the phenomenological qualities of the memory at retrieval, which are associated with perceived importance⁵⁵..... Precise retelling of internal experiences may reinstate the original appraisals and emotions of the event⁷⁶, reengaging the sense of self-relevance that initially made the event important. More broadly, our findings suggest that internal experiences may serve as a key bridge between event and self, capturing not only what happened but also its personal meaning—consistent with the view that autobiographical memory supports self-coherence^{3,77,78}. **These findings also align with perspectives showing that people make meaning from life events by constructing narratives around their memories^{38,79,80}, with thoughts and emotions playing a significant role in this process^{81,82}.**

However, the fact that perceived importance is also an internally oriented, subjective phenomenon does not preclude examining its association with metrics derived from internal details extracted from the narrative text. These variables were designed and measured separately to capture different aspects of autobiographical memory recall: perceived importance ratings capture participants' subjective judgments of how impactful, important, or personal the entire event was, whereas the metrics concerning internal details quantify how much the recalled narrative included specific thoughts or feelings experienced during the event, as well as how precisely these details were reproduced in the second recall. Our correlation analyses therefore assessed the relationship between potentially conceptually related but distinctly measured variables, which is a widely used approach in research on psychological constructs.

Additional design and analysis concerns:

I was unable to access the pre-registration. For any analyses not pre-registered, the authors should specify that they are exploratory.

: The preregistration information and corresponding link were provided in the first paragraph of the Methods section of the original manuscript (<https://aspredicted.org/9zmr-djp5.pdf>). We have double-checked the URL and confirmed that it remains publicly accessible. In addition, we have revised the paragraph to clarify that the mixed-effects model analysis and the analysis involving perceived importance were exploratory.

p.4. The primary analyses comparing memory details that described internal versus external experiences were preregistered on AsPredicted (<https://aspredicted.org/9zmr-djp5.pdf>; Preregistration date: March 2nd, 2024). **Analyses examining (1) predictors of memory detail retention and (2) perceived memory importance were exploratory and not included in the original preregistration.**

Why is inter-rater reliability not reported? IRR should be reported at least for the training of coders.

: We appreciate the reviewer's comment. Reviewer 1 and Reviewer 3 also requested that we report inter-rater reliability. In our original coding procedure, each narrative was coded by only one annotator due to limited time and resources, making it impossible to compute inter-rater reliability for the full dataset. To address the reviewers' comment, we estimated inter-rater reliability by recruiting two additional independent annotators and having them code a subset of the data consisting of 40 narratives from 20 participants (including both first and second recalls). These narratives were randomly selected from those originally coded by the annotator who had coded the largest number of narratives. The two additional annotators received the same instructions as the original annotators and categorized every memory detail in each narrative into one of the five nominal categories used in the study (external, internal, inferences about others, general comments, or other).

We used Fleiss's Kappa (κ), which is the appropriate statistic for assessing chance-corrected agreement among multiple raters when the data are nominal (Fleiss, 1971). We computed κ by pooling all coded details across the 40 narratives and computing agreement among the three raters. The overall κ was .54 (95% CI = [.52, .56]), indicating moderate agreement according to conventional benchmarks (Landis & Koch, 1977). Category-specific κ values for external, internal, inferences about others, general comments, or other categories were .64, .57, .54, .39, and .42, respectively. When each recall session was considered separately, κ for the first and second recalls was .53 and .57, respectively. Additionally, we calculated percent agreement across annotators, which had a mean of 59.4% (SD = 18.5) across the 40 narratives. The percent agreement for the first and second recalls was 59.2% (SD = 18.6) and 60.7% (SD = 18.9), respectively. We have revised the Methods section to report the κ values.

p.9. In addition, we retrospectively assessed the inter-rater reliability of memory detail categorization by having two additional independent annotators code a randomly selected subset of data from 20 participants. The overall Fleiss's Kappa across the three annotators (including the original annotator) was .54, indicating moderate agreement. Category-specific Kappa values were .64 for externally oriented, .57 for internally oriented, .54 for inferences about others, .39 for general comments, and .42 for other categories.

Add alpha criteria for relevant tests (e.g., t-tests, ANOVAs, correlations). Why was there no alpha correction for multiple comparisons?

: We thank the reviewer for raising this point. Our statistical tests used the conventional alpha level of $p = .05$, as reflected in the 95% confidence intervals reported throughout the manuscript. For transparency, we also reported exact p values, except for very small values, which were reported as $p < .001$ in accordance with the journal's guidelines. Regarding multiple comparisons correction, most of our analyses did not strictly require such adjustments under standard conventions, as they involved targeted comparisons between internal and external

details for a small number of dependent variables that were selected a priori. The exception, however, was the set of mixed-effects logistic regression analyses in which five separate models were used to examine multiple memory-detail features for exploratory purposes. In response to the reviewer's request, we applied the Benjamini–Hochberg false discovery rate (FDR) correction to these results. We have added a sentence to the 'Predicting the retention of individual memory details' subsection of the Methods to describe the correction method.

p.11. Statistical significance of each predictor was determined using a likelihood ratio test, comparing the full model (with all fixed and random effects) to a null model that excluded the predictor of interest. **When individual predictors were tested in separate models, we applied a Benjamini–Hochberg false discovery rate (FDR) correction across the five likelihood ratio tests to adjust the resulting *p* values for multiple comparisons.**

The pattern of results remained unchanged after correction. However, we have revised the text in the 'Features predicting memory detail retention' subsection of the Results and provided detailed statistics, including both uncorrected and FDR-corrected *p* values, in a new Supplementary Table 5. This table was generated to comply with the journal's guidelines by reporting the full statistics for each test or predictor, rather than summarizing them in the text (e.g., as *ps* < .001).

p.16. Likelihood ratio tests showed that local similarity, global similarity, and emotional intensity each positively predicted the retention of internally **oriented** details, **whereas** emotional valence and word count were not significant predictors (**see Supplementary Table 5 for full statistics**).

p.16. When tested individually in separate models, all features except emotional valence positively predicted the retention of externally **oriented** details (Fig. 4d; **Supplementary Table 5**).

Supplementary Table 5. Mixed-effects logistic regression and likelihood ratio test results for predictors of memory detail retention.

Detail category	Feature	β	95% CI	$\chi^2(1)$	<i>p</i> (uncorrected)	<i>p</i> (FDR corrected)
Internally oriented	Emotional valence	.14	[-.03, .31]	2.77	.096	.122
	Emotional intensity	.20	[.03, .37]	5.67	.017	.030
	Global semantic similarity	.24	[.07, .40]	7.78	.005	.014
	Local semantic similarity	.27	[.10, .44]	10.21	.001	.008
	Word count	.09	[-.07, .25]	1.19	.275	.273

Externally oriented	Emotional valence	.02	[-.05, .10]	.40	.528	.527
	Emotional intensity	.09	[.01, .16]	5.23	.022	.028
	Global semantic similarity	.22	[.14, .31]	27.99	< .001	< .001
	Local semantic similarity	.16	[.08, .24]	16.37	< .001	< .001
	Word count	.28	[.20, .36]	49.72	< .001	< .001

We also revised Figures 4c-d and the figure legend so that the statistical significance markers now reflect FDR-corrected p values.

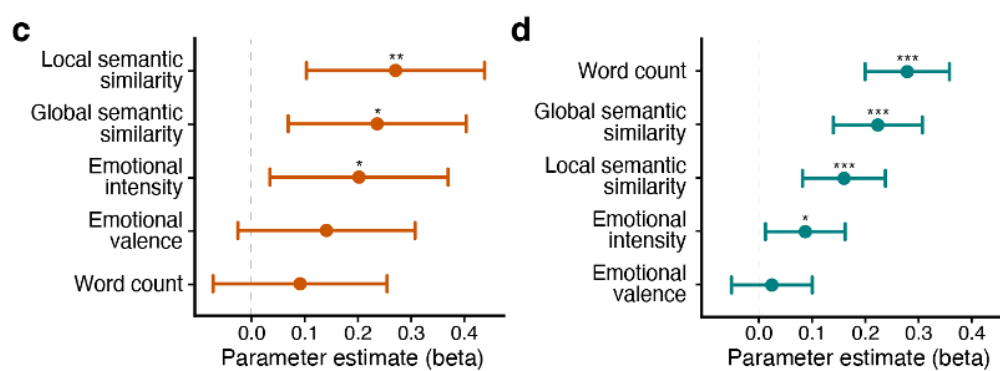


Fig. 4. Predicting memory detail retention from first-recall features. (c) and externally oriented (d) details using different memory detail features. Circles represent parameter estimates, and error bars indicate 95% confidence intervals. Statistical significance for each feature was assessed with a likelihood ratio test, comparing the full model to the null model excluding the feature of interest. **Supplementary Table 5 reports the full statistics for each feature.** * $p < .05$, ** $p < .01$, *** $p < .001$ (FDR corrected).

Add effect sizes and CIs for t-tests on top of pg. 7.

: We appreciate the reviewer's suggestion. However, no t-tests were reported on page 7 of the original manuscript or the supplementary information. If the reviewer was referring to the Williams tests comparing two correlation coefficients (reported on pages 15-16), we did not report Cohen's d because, unlike standard t-tests, the Williams test is not based on mean differences, even though it yields a t statistic. Instead, the statistic is derived from the covariance structure of the correlations being compared. Consequently, the R function we used (`r.test` in the `psych` package) does not provide effect sizes or confidence intervals in the way that typical t -tests do.

However, to provide more detailed statistics in response to the reviewer's request, we computed Cohen's q (Cohen, 1988) as the effect size for the difference between two

correlations. Cohen's q is defined as the difference between the Fisher z -transformed correlations, with conventional benchmarks of 0.10 (small), 0.30 (medium), and 0.50 (large). To obtain confidence intervals, we followed Zou's (2007) method for constructing 95% CIs for the difference between two dependent correlations. These statistics are now reported in the 'Perceived importance of memory' subsection of the Results.

p.18. Williams tests confirmed that the correlations for internally versus externally oriented details differed significantly in both the first ($t(207) = 2.51, p = .013, 95\% \text{ CI} = [.06, .49], \text{Cohen's } q = .28$) and second recalls ($t(207) = 3.62, p < .001, 95\% \text{ CI} = [.18, .59], \text{Cohen's } q = .39$).

p.19. Although the difference between the correlations for internally and externally oriented details did not reach statistical significance ($t(123) = 1.89, p = .061, 95\% \text{ CI} = [-.01, .41], \text{Cohen's } q = .21$), the positive correlation for internally oriented details suggests that consistent and precise retention of internal experiences may help buffer against the general decline in perceived importance over time.

A note should be added to Table 1 describing what (unc.) means.

: We appreciate the reviewer's suggestion for clarification. We added a note to the table indicating that we were reporting uncorrected (unc.) raw p -values.

Table 1. Comparison of selected features between externally and internally oriented details recalled during the first session.

Feature	Externally oriented mean (SD)	Internally oriented mean (SD)	Externally oriented vs. Internally oriented comparison				
			t	df	p^*	95% CI	Cohen's d
Emotional valence	.10 (.15)	.15 (.35)	-1.79	178	.075	[-.09, .00]	.17
Emotional intensity	.21 (.10)	.41 (.19)	-11.87	178	< .001	[-.22, -.16]	1.24
Global semantic similarity	.28 (.06)	.24 (.06)	8.60	177	< .001	[.03, .05]	.69
Local semantic similarity	.34 (.06)	.28 (.08)	8.62	177	< .001	[.04, .07]	.83
Word count	10.4 (2.3)	9.6 (3.9)	2.45	178	.015	[.14, 1.32]	.23

*Uncorrected two-tailed p value.

Add range for age of memories on pg. 5.

: We appreciate the reviewer's suggestion. We have reported the range of memory ages for both the first and second recalls in the "Data selection" subsection of the Methods section:

p.6. On average, recalled events were 41.8 days old ($SD = 29.3$, range = 7–90) at the first session and 126.6 days old ($SD = 71.1$, range = 21–540) at the second session.

Reviewer #3 (Remarks to the Author):

Note: This review was co-written with a graduate trainee.

In this study, Su et al. utilized an existing large-scale dataset of participant-written autobiographical narratives (recalled at two time points) to investigate the memorability of internal details – i.e., memory for thoughts and feelings that occurred during an experience. Using a mixture of manual annotation and natural language processing, the authors identified internal and external (i.e., externally observable details from a remembered event) details within individual autobiographical memories and examined how they changed across two retrieval attempts, separated by several weeks. The authors found that internal details occur less frequently than external details and that they were more prone to distortion or forgetting. However, when an internal detail was more emotionally intense and semantically related to external details, it was more likely to be remembered. Despite their malleability, memories composed of a greater proportion of internal details were more likely rated as “important” by the rememberer.

Overall, the present manuscript is well-written and presents novel findings on the memorability of internal details in autobiographical memory. We both enjoyed reading this paper. The combination of qualitative and quantitative techniques is commendable. Also, we agree with the authors that the internal aspects of an experience is an understudied aspect of autobiographical memory research and thus these results will be highly valuable to the field. Below are some comments and suggestions that aim to help strengthen this paper.

: We thank the reviewer for their enthusiasm and positive assessment of our study.

(1) Definition of internal vs. external:

In this manuscript, the authors define internal and external details based on whether they refer to aspects of the remembered experience with a locus within the rememberer (e.g., their thoughts, feelings) or outside the rememberer (i.e., details that that could be corroborated by an external observer). This is intuitive, I think. But – the autobiographical memory literature, specifically work using the Autobiographical Interview (AI; Levine et al., 2002), uses the same terms to refer to a different rubric. According to the AI, internal details refer to “...those that pertained directly to the main event described by the participant, were specific to time and place, and were considered to reflect episodic reexperiencing.” (p. 679; Levine et al., 2002) External details were defined as “semantic (factual information or extended events that did not require recollection of a specific time and place).” (p. 679; Levine et al., 2002) In other words, internal details are episodic – irrespective of whether they are about the rememberer’s thoughts/feelings or observable details from an experience. For example, recalling that an event happened in a certain floor of the library at their university would count as an internal detail

according to the AI, and (if we are understanding the authors' work correctly) an external detail using this manuscript's rubric. It also seems like the authors occasionally conflate these two distinct definitions of internal/external (e.g., "Memories of internal thoughts and feelings have rarely been the central focus of research using autobiographical narratives, even though they are commonly recalled and explicitly included as a distinct category in standardized coding protocols such as the Autobiographical Interview³⁷." (p. 16)

The authors should explicitly address the distinction between internal and external details as defined here and as defined in the AI in the manuscript. Perhaps calling them "internally-oriented" and "externally-oriented" details would help better delineate them?

: We thank the reviewer for raising this important point. Reviewer 1 raised a similar concern in their Comment 1. We fully agree that the terms "internal details" and "external details" may be confusing to readers familiar with their usage in the Autobiographical Interview scoring method. Following the reviewers' request, we have replaced these terms with "internally oriented details" and "externally oriented details" throughout the manuscript where appropriate.

Additionally, we clarified the distinction between our coding scheme and the Autobiographical Interview in the 'Memory detail categorization' subsection of the Methods. As the reviewer noted, the internally oriented category in our study corresponds to the thoughts/emotions details within the internal (episodic) category of the AI protocol (when the thoughts or emotions were the participant's own). The externally oriented category corresponds to the combination of event, time, place, and perceptual details within the AI's internal (episodic) category. Finally, the other category in our study corresponds to the AI protocol's external (non-episodic) category.

p.8. It is also important to note that our internally and externally oriented memory detail categories differ from the "internal" and "external" details defined in the widely used Autobiographical Interview (AI) protocol³⁹. In the AI protocol, internal details refer to episodic elements directly related to the main recalled event, whereas external details encompass information not directly tied to that event (e.g., semantic knowledge). Under this framework, both the internally oriented and externally oriented details in our study correspond to the AI protocol's internal details, while our 'other' category corresponds to the AI protocol's external details.

(2) Is this actually forgetting?

Forgetting, in this study, was defined as a detail being mentioned in the first retrieval session that was not mentioned in the second retrieval session. But if we consider remembering as a communicative act (e.g., Marsh, 2007; Mahr & Csibra, 2018), not mentioning a detail might also be caused by not wanting to repeat oneself, whether implicitly or explicitly. Also, given that we might change the nature of what we recall given the communicative goals of recalling an experience (e.g., Dudukovic, Marsh & Tversky, 2004) yet another possibility could be that the internal experiences are simply less central/necessary to the demands of the specific recall instruction associated with this task (e.g., the participant might implicitly/explicitly conceive that whoever is asking them to recall this experience mainly cares about what actually happened –

thus external details, by the authors' definition). The authors mention this limitation at the end of the discussion, but the rest of the paper is heavily framed using forgetting [e.g., "...internal details ($M = 38.3\%$, $SD = 32.7$) were significantly less likely to be retained than external details ($M = 48.9\%$, $SD = 19.2$), indicating that internal experiences are more susceptible to forgetting over time"; "Although second recalls were generally shorter than the first—indicating some degree of forgetting—..." (p. 11)].

Some tempering of the use of forgetting language seems necessary here, given the above points. Also, it would be helpful if the authors more substantively/explicitly discussed some of the reasons, outside of forgetting, why a rememberer might not include a detail from the first recall session in the second recall session.

: We thank the reviewer for this thoughtful comment and for directing us to relevant prior work. Reviewer 1 raised a similar point in her Comment 13. We fully agree that the absence of a mentioned detail does not necessarily indicate that the detail was forgotten. As suggested, we have replaced the term "forgetting" with "omission" throughout the main text where appropriate.

p.2. We found that internal experiences were more prone to **omission and semantic** distortion over time compared to external ones.

p.13. First, as a measure of memory retention or **omission**, we calculated the proportion of memory details reported during the first recall that were also recalled during the second session.

p.13. Importantly, when externally **oriented** and internally **oriented** details were analyzed separately (Fig. 3a), internally **oriented** details ($M = 38.3\%$, $SD = 32.7$) were significantly less likely to be retained than externally **oriented** details ($M = 48.9\%$, $SD = 19.2$), indicating that internal experiences are more susceptible to **omission** over time ($t(178) = 3.73$, $p < .001$, 95% CI = [4.98, 16.19], Cohen's $d = 0.39$).

p.14. Although second recalls were generally shorter than the first—indicating **greater omission**—they also included new memory details not reported in the initial recall, consistent with findings from prior studies^{30,62}.

p.15. If memories of internal experiences are more susceptible to **omission and semantic** distortion than memories of external experiences, what factors might make them more durable?

p.19. Using natural language processing to analyze a large-scale dataset of autobiographical narratives (the Hippocorpus dataset^{43,44}), we found that internal experiences were more susceptible to **omission and semantic** distortion than observable external experiences.

p.20. Our findings indicate that internal experiences are frequently remembered in

everyday life, but they are also more vulnerable to **omission** and **semantic** distortion than external experiences, in line with prior empirical work^{21,22,24}.

p.22. While these internal experiences were prone to **omission**, their retention was supported by emotional intensity and stronger connections to other elements of the memory.

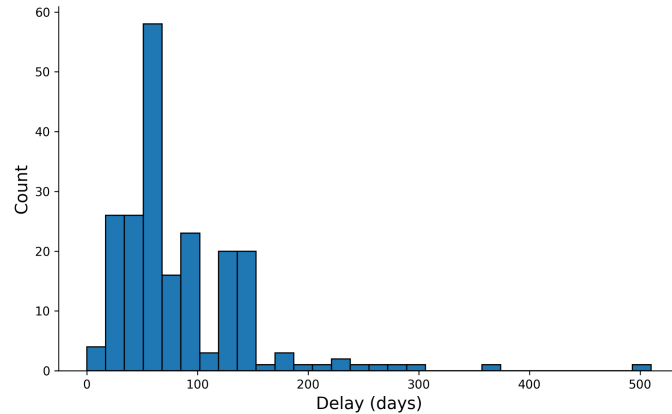
In addition, following the reviewer's suggestion, we have revised the limitations paragraph in the Discussion to elaborate on potential reasons why certain memory details were not included in recall. In response to Reviewer 1's comment, we also added a new point addressing how differences in audience or social context may influence what participants choose to report. We kept the discussion relatively brief because our manuscript was already far above the journal's recommended word limit (~5,000 words), but we would be happy to elaborate further should the reviewer or editors consider additional detail necessary.

p.21. Second, in narrated free recall, it is difficult to determine whether the absence of certain details reflects genuine forgetting or a choice not to report them. **Because memory recall is often a communicative act⁸⁴, participants may omit details they perceive as redundant, peripheral to the task goal, or irrelevant to a potential audience^{85,86}. Details omitted due to retrieval neglect can also re-emerge when appropriate cues are provided²³. Indeed, details missing from an initial recall frequently appeared in later recalls, as shown in prior studies^{30,62} and in the present work. Future research could address this issue by incorporating detailed probes or prompts following narrated free recall. Third, **also related to the communicative function of memory recall, the quality and characteristics of autobiographical narratives may vary depending on the social context in real life⁸⁵⁻⁸⁷. For example, people may share more thoughts and feelings when recalling memories for close others than they do in experimental contexts such as the one used in the present study.****

(3) Effects of delay?

By looking at the mean and SD, there seems to be a considerable amount of variability in the delay between the two retrieval sessions. Reporting a bit more detail about the distribution of delays would be helpful, given the importance of delay with respect to forgetting. I recommend including the range in addition to mean and SD, alongside a histogram of delays, perhaps as part of Figure 1 – or, at least, in the supplemental materials.

: We thank the reviewer for raising this important point. Following their request, we now report the distribution and range of the delay between the first and second recalls in Supplementary Figure 2.



Supplementary Figure 2. Histogram showing the distribution of delays (in days) between the first and the second recall sessions (range = 0–510 days). In each recall session, participants estimated how long ago the reported event had occurred by entering a numerical value and selecting a unit (weeks or months). These responses were converted to days to approximate the event's age at each session. The delay interval was then computed as the difference between the event's estimated age at the first and second recalls. Unlikely values such as 0 days may have resulted from the coarse resolution of the response units or from errors in participants' self-reports.

We also revised the relevant sentence in the Results section to reference the new supplementary figure.

p.12. Participants recalled a memorable life event twice, separated by an average of 84.8 days ($SD = 61.9$; see Supplementary Fig. 2 for the distribution of delay durations).

Please note that the delay distribution includes some unlikely values (e.g., 0 days, 510 days), likely due to the way memory age was collected and/or errors in participants' self-reports. In the original Hippocampus survey, participants were asked in each recall session to estimate how long ago the reported event had occurred ("How long ago [in weeks/months] was the event?"). They entered a numerical value and selected a unit (weeks or months), and these responses were converted to days to approximate the age of the memory (i.e., the interval between the event and the recall session). The delay interval was defined as the difference between the event's estimated age at the first and second recalls. Because delay estimates were based on self-reported duration and unit selection, some measurement error is possible if participants misinterpreted the question or selected an incorrect unit. In addition, the coarse resolution of the available units may have produced delay values that were much longer or shorter than the actual interval. For example, a delay of 0 days could arise if a participant reported the memory as '3 months old' in both recall sessions even though the true delay was 2–3 weeks.

Furthermore, the authors refer to memory transformation over time throughout the manuscript. However, as far as I am aware, the effect of time was not explicitly modelled. Are internal details less likely to be repeated amongst individuals who happened to have a longer delay between recall sessions? It seems to me that delay could be added to the mixed-effects logistic

regression models on memory retention. This seems worthwhile given all the work demonstrating the relationship between delay and forgetting.

: We appreciate the reviewer's suggestion for this analysis. Following the reviewer's request, we examined the effect of delay on memory detail retention. We included the normalized (z-transformed) delay (in days) as a fixed effect in separate mixed-effects logistic regression models for internal and external details. In both models, adding delay did not significantly improve model fit relative to the null model (internal: $\chi^2(1) = .14$, $p = .71$; external: $\chi^2(1) = .01$, $p = .92$), indicating that the time between sessions did not reliably predict whether a detail would be recalled again. Similarly, the correlations between delay and memory-detail retention rate were not significant for either internal ($r(177) = .05$, $p = .50$, 95% CI = [-.10, .20]) or external details ($r(208) = -.01$, $p = .86$, 95% CI = [-.15, .12]). These results remained qualitatively the same after excluding potentially erroneous values (delay = 0 days or > 200 days). Since the delay predictor had essentially no effect when included alone, we did not proceed with adding delay as an additional predictor in the mixed-effects models reported in the manuscript. We speculate that delay may have had limited influence because the memories were already, on average, more than one month old at the time of the first recall. The new analysis results are reported in the Supplementary Results.

Effects of delay between the first and second recalls on memory detail retention

Among the 210 participants included in the analysis, the delay between the first and second recall sessions varied substantially (Supplementary Figure 2). Therefore, we examined whether the length of this interval influenced the likelihood that memory details would be retained across sessions.

We first computed Pearson correlations between the delay interval (in days) and the proportion of memory details retained in the second recall, separately for internally oriented and externally oriented details. There was no significant correlation for either internally oriented details ($r(177) = .05$, $p = .50$, 95% CI = [-.10, .20]) or externally oriented details ($r(208) = -.01$, $p = .86$, 95% CI = [-.15, .12]). We also fit mixed-effects logistic regression models in which the delay (z-transformed) was included as a fixed-effect predictor and participant was included as a random intercept, predicting whether each individual memory detail was retained in the second recall. Separate models were constructed for internally and externally oriented details and compared with corresponding null models (containing only the participant random effect) using likelihood ratio tests. Again, adding delay did not significantly improve model fit for either internally oriented details ($\chi^2(1) = .14$, $p = .71$) or externally oriented details ($\chi^2(1) = .01$, $p = .92$), indicating that the time interval between sessions had little influence on whether a detail was recalled again.

Moreover, the results remained qualitatively unchanged even after excluding outlier and implausible delay values that were likely due to errors (i.e., delay = 0 days or > 200 days). There was no significant correlation between delay and the proportion of retained internally oriented details ($r(165) = .04$, $p = .60$, 95% CI = [-.11, .19]) or externally oriented details ($r(196) = .002$, $p = .98$, 95% CI = [-.14, .14]). Similarly, mixed-effects logistic regression showed that delay did not significantly predict retention of

individual memory details for either internally oriented ($\chi^2(1) = .09, p = .77$) or externally oriented details ($\chi^2(1) = .16, p = .69$).

We also added a sentence to the 'Retention and addition of memory details' subsection in Results to reference the new findings reported in the Supplementary Results.

p.14. The retention rate was also not significantly influenced by the length of the delay between the two recall sessions for either internally oriented or externally oriented details (see Supplementary Results).

(4) Manual coding of details.

The raters manually coded each memory into one of the five following categories: (i) internal details, (ii) external details, (iii) inferences about others, (iv) general comments, and (v) other details. The authors mentioned that these categories were predefined. How did the authors end up deciding on these definitions? What there a systematic inductive coding procedure that the authors used? In any case, some more details about how the authors settled on using these categories is warranted.

: We appreciate the reviewer's suggestion for clarification. Reviewer 1 also raised a related issue in Comment 2 (regarding our rationale for not using the Autobiographical Interview protocol). In developing our coding categories and their definitions, we drew on both deductive considerations based on the primary goals of the study and inductive observations from the narrative data (though the latter were not conducted through a fully formalized procedure). First, because the aim of the study was to contrast memories of internal experiences accessible only to the self with externally observable experiences, these two classes naturally formed our primary categories of interest. Because our study and dataset focused on episodic recall of specific life events, all details needed to be relevant to the event being described. Accordingly, any information not related to the main recalled event was labeled as "other" and excluded from analysis. Moreover, because our focus was on the broad internal-external distinction, we did not further subdivide external details (e.g., time, location, perception) or event-unrelated details (e.g., semantic, repetition) into finer-grained categories, as done in the Autobiographical Interview protocol.

Thus, our initial coding scheme consisted of three categories: internal, external, and other. However, during pilot coding conducted to test and refine the coding instructions for the annotators, we identified several potential misclassification issues within the subset of reviewed data. First, descriptions of participants' perceptions of other people's thoughts and feelings could be mistakenly coded as internal details, which were meant to capture only the participants' own thoughts and feelings accessible solely to themselves (this issue is also relevant to Reviewer 1's Comment 5). Second, thoughts and feelings that arose during the act of recall, rather than during the original event, could also be misclassified as internal details, even though the category was intended to include only internal states experienced at the time of the event (this issue is relevant to Reviewer 2's Conceptual Point 2). Because these types of details often used language similar to that of internal details to describe internal states, they were easily confusable. To address these issues, we created two additional categories,

'inferences about others' and 'general comments or evaluations', to help annotators avoid misclassification. Making annotators explicitly aware of these types of details and instructing them to categorize such instances separately allowed us to identify and exclude them from analysis, as they were not relevant to the main goal of our study.

Following the reviewer's request, we have revised the 'Memory detail categorization' subsection of the Methods to more clearly justify our categorization scheme.

p.8. Since the primary goal of the current study was to examine the retention and transformation of internal thoughts and feelings over repeated recalls, and to compare these patterns with those of objective external elements, these two categories **were defined and analyzed as the primary categories of interest**. Given our focus on the broad internal-external distinction, we did not further subdivide externally oriented details into finer-grained categories (e.g., time, location, perception), as is done in previously established protocols (e.g., ref.³⁹). The categories of 'speculations or inferences about others' and 'general comments or evaluations' **were additionally introduced to help annotators avoid misclassification**. The former distinguished other people's inferred internal experiences from participants' own, while the latter separated thoughts or feelings generated during the act of recall from those that occurred during the original event. Because internally oriented details and these additional types of details all described internal states using similar language, they were easily confused. By explicitly flagging these categories for annotators and instructing them to classify such details separately, we aimed to more effectively identify and exclude them from the main analysis—together with the 'other' category—as they were not relevant to the study's primary goal.

(5) Interrater reliability.

The authors stated the following: annotators' "performance was monitored by one of the authors (C.K.), and they were encouraged to ask questions before and during the task to clarify any uncertainties. C.K. also conducted a quality check of each completed recall narrative and requested revisions when any issues were identified." Were any interrater reliability measures taken? For example, did C.K. or another annotator independently segment, categorize, and/or match subsets of the data coded by the annotators. Were metrics such as percent agreement or Cohen's kappa computed? Given that details were deductively coded as either internal or external (i.e., "Annotators were also provided with detailed definitions and examples for each category."), rigorous interrater reliability checks may not be as applicable to the present study relative to other, more exploratory qualitative work. That said, other studies also using manual coding, for example, to analyze distinct aspects of autobiographical memory narratives (e.g., Sotgiu et al., 2025), computed Cohen's kappa values as an index of interrater reliability among coders. The authors of the present manuscript should specify whether any such measures were taken, and if not, should provide justification for their decisions.

: We thank the reviewer for their thoughtful comment. Reviewer 1 and Reviewer 2 also requested that we report inter-rater reliability. In our original coding procedure, each narrative was coded by only one annotator due to limited time and resources, making it impossible to

compute inter-rater reliability for the full dataset. To address the reviewers' comment, we estimated inter-rater reliability by recruiting two additional independent annotators and having them code a subset of the data consisting of 40 narratives from 20 participants (including both first and second recalls). These narratives were randomly selected from those originally coded by the annotator who had coded the largest number of narratives. The two additional annotators received the same instructions as the original annotators and categorized every memory detail in each narrative into one of the five nominal categories used in the study (external, internal, inferences about others, general comments, or other).

We used Fleiss's Kappa (κ), which is the appropriate statistic for assessing chance-corrected agreement among multiple raters when the data are nominal (Fleiss, 1971). We computed κ by pooling all coded details across the 40 narratives and computing agreement among the three raters. The overall κ was .54 (95% CI = [.52, .56]), indicating moderate agreement according to conventional benchmarks (Landis & Koch, 1977). Category-specific κ values for external, internal, inferences about others, general comments, or other categories were .64, .57, .54, .39, and .42, respectively. When each recall session was considered separately, κ for the first and second recalls was .53 and .57, respectively. Additionally, we calculated percent agreement across annotators, which had a mean of 59.4% (SD = 18.5) across the 40 narratives. The percent agreement for the first and second recalls was 59.2% (SD = 18.6) and 60.7% (SD = 18.9), respectively. We have revised the Methods section to report the κ values.

p.9. In addition, we retrospectively assessed the inter-rater reliability of memory detail categorization by having two additional independent annotators code a randomly selected subset of data from 20 participants. The overall Fleiss's Kappa across the three annotators (including the original annotator) was .54, indicating moderate agreement. Category-specific Kappa values were .64 for externally oriented, .57 for internally oriented, .54 for inferences about others, .39 for general comments, and .42 for other categories.

(6) Mixed-effects logistic regression predictors.

To predict the retention of individual memory details, the authors conducted a mixed-effects logistic regression analysis, using emotional valence, emotional intensity, semantic connection to external elements, and word count as predictors. How were these predictors selected? Were they theoretically informed? Were they selected based on natural language processing constraints? The authors should clarify their decisions regarding predictor selection, specifically explaining why they expect such variables (as opposed to others) to predict the likelihood of memory retention. The rationale is briefly touched on in the "Features predicting memory detail retention" section of the results (pp. 12-13), but it would be beneficial to discuss it earlier on (perhaps in the methods section) to set readers up for understanding why specific predictors, and not others, were selected for focus.

: We appreciate the reviewer's request for clarification, and we agree that it would be helpful to discuss earlier in the Methods section why we selected the specific predictors used in the mixed-effects logistic regression analysis. Our choices were guided by both practical constraints

and prior theoretical and empirical work. First, a major constraint was that all predictors needed to be derivable directly from the recall text itself, because no additional information was available at the level of individual memory details. Although many other factors could plausibly influence memory detail retention, including phenomenological characteristics based on participants' subjective experience (e.g., Stawarczyk & D'Argembeau, 2019; Finley & Brewer, 2024), such measures were not available in this dataset. We noted this constraint as a limitation of the study in the Discussion section.

p.21. **Fourth**, because the Hippocampus dataset⁴³ did not include participants' subjective ratings of individual memory details, we relied on text analysis to characterize them. **However, emotion ratings based solely on the text of each individual detail cannot account for the broader context and may not accurately reflect the emotions actually experienced by participants. Moreover**, certain phenomenological properties known to predict recall, such as vividness and sense of reliving^{23,28}, can only be assessed by participants themselves. Collecting such ratings for individual details could therefore provide a more precise basis for predicting retention.

Second, among the text-based features that could be extracted from each individual memory detail, we focused on those most likely to influence memory performance based on prior theoretical and empirical work that we were aware of (i.e., emotional properties, semantic similarity to external details, and word count). As the reviewer noted, the rationale for selecting these predictors, along with citations to related studies, is provided in the relevant Results section and in the final paragraph of the Introduction.

p.15. First, we examined the emotional properties of each detail (Fig. 4a), given their known influence on autobiographical recall^{47,48}. ... Second, we assessed local and global semantic similarity with external experiences (Fig. 4b), based on the hypothesis that such connections could serve as concrete retrieval cues facilitating recall^{51,52}. ... Finally, we used the word count of each detail as a proxy for informational richness or complexity, assuming that more elaborated details would be more memorable⁴⁹.

p.4. However, we predicted that emotional salience^{47,48}, informational richness^{49,50}, and semantic connections to external elements that may serve as concrete retrieval cues^{51,52} would mitigate this vulnerability.

Following the reviewer's request, we revised the 'Predicting the retention of individual memory details' subsection in the Methods to further explain the rationale behind our predictor selection. We also added a new paragraph clarifying the 'word count' predictor, in response to the reviewer's Comment 7 below.

p.10. We conducted mixed-effects logistic regression analyses to examine which features of individual memory details from the first recall predicted their likelihood of being retained in the second recall. **Since the Hippocampus dataset contained no information about individual memory details beyond the recalled text itself, all predictors**

had to be features that could be extracted through text analysis. Among these text-derived features, we focused on those likely to influence memory retention based on prior findings^{48,49,51}, including emotional valence, emotional intensity, semantic connection to external elements, and word count per detail.

p.10. We selected emotional valence and intensity as predictors because prior work has shown that emotional autobiographical memories are generally remembered better than neutral ones^{40,48,60}. The emotional valence and emotional intensity of each memory detail were automatically rated using a natural language processing approach.....

p.11. Additionally, we examined the semantic connections between each memory detail and other externally oriented details. Earlier research has demonstrated that rich semantic connections among external events can enhance recall for those events⁵¹. For each internally oriented or externally oriented memory detail recalled during the first session, we computed its semantic similarity to external elements from the same session using the SBERT model⁵⁹ described above. To examine the effects of different levels of external context, we calculated both local and global semantic similarity....

p.11. Finally, we included word count per detail as a proxy for how elaborate or informationally rich each memory detail was. More elaborated memories with richer features are generally remembered better^{49,50}. For each detail, we counted the total number of words and did not exclude function words such as articles and prepositions. This approach preserved the natural form of participants' written responses and maintained consistency with other text-based analyses in the current study. Notably, the mixed-effects logistic regression results were qualitatively unchanged when we instead excluded function words and used the number of content words as the predictor (Supplementary Results).

(7) Word count as an index of informational richness.

The authors indicated that word count per detail acted as an indicator of information richness. Did the authors control for stop words in word count calculations? Stop words may inflate the degree of "information richness" given their contribution to word count, despite their relative lack of "richness." In general, more details on any automated text preprocessing steps used would also be helpful.

: We thank the reviewer for this insightful question. In our original analysis, the word count measure included all words in each memory detail, including stop words. This choice preserved participants' original written responses as much as possible and maintained consistency with other text analyses (for example, semantic similarity and sentiment analysis), which were also conducted on the unmodified text. However, to address the reviewer's comment, we recomputed word count after removing stop words using the Natural Language Toolkit corpus of English stop words. After removing stop words from each detail, external details still contained significantly more content words than internal details ($t(178) = 5.34, p < .001$). We then re-ran the mixed-effects logistic regression using the z-transformed word count excluding stop words

as a predictor, separately for internal and external details. The results remained unchanged: word count (excluding stop words) was not a significant predictor for the retention of internal details ($\chi^2(1) = .81, p = .37$), but continued to be a robust predictor for the retention of external details ($\chi^2(1) = 62.02, p < .001$). We now report these results in the Supplementary Results:

Effects of content word count on detail retention

We used the word count of each detail as a proxy for informational richness. However, function words (e.g., articles, prepositions) may inflate this measure without contributing substantive content. Therefore, we recomputed word count after removing stop words using the Natural Language Toolkit (NLTK) corpus of English stop words. After stop words were removed, externally oriented details still contained significantly more content words than internally oriented details (externally oriented $M = 4.71, SD = 1.17$; internally oriented $M = 3.98, SD = 1.75$; $t(178) = 5.34, p < .001, 95\% CI = [.46, .99]$, Cohen's $d = .40$). We then re-ran the mixed-effects logistic regression using the z-transformed content word count as a fixed effect, with participant as a random intercept. Separate models were fit for internally oriented and externally oriented details. The results remained unchanged: word count (excluding stop words) was not a significant predictor of internally oriented detail retention ($\chi^2(1) = .81, p = .37$), but continued to be a robust predictor of externally oriented detail retention ($\beta = .31, 95\% CI = [.23, .39], \chi^2(1) = 62.02, p < .001$).

We also added a paragraph to the 'Predicting the retention of individual memory details' subsection of the Methods to clarify our definition of the word count feature and to reference the above supplementary analysis results.

p.11. Finally, we included word count per detail as a proxy for how elaborate or informationally rich each memory detail was. More elaborated memories with richer features are generally remembered better^{49,50}. For each detail, we counted the total number of words and did not exclude function words such as articles and prepositions. This approach preserved the natural form of participants' written responses and maintained consistency with other text-based analyses in the current study. Notably, the mixed-effects logistic regression results were qualitatively unchanged when we instead excluded function words and used the number of content words as the predictor (Supplementary Results).

(8) Semantic similarity of matched details.

The authors report that internal details show less semantic similarity to their matched counterparts across recall sessions than matching external details (see Figure 3C). Why is this the case? One possibility is that internal details are truly less stable across recall attempts – which is the interpretation that the authors use. However, might another possibility be that this is a property of the SBERT embeddings themselves? For example, if SBERT vectors better capture the semantics of concrete sentences better than abstract ones – and if internal details tend to describe abstract mental states/emotions – might that be another reason to observe this effect? Personally, I am not aware of much work examining the effect of word/sentence

concreteness on embedding quality, but if the authors have any thoughts on this possibility, I'd be curious to hear them.

: We appreciate the reviewer's suggestion of this alternative explanation. SBERT converts each sentence into an embedding such that sentences with similar meanings have embeddings that point in similar directions. Because SBERT is trained on millions of human-judged sentence pairs reflecting similar meaning or logical entailment, independent of whether the language is concrete or abstract, its similarity estimates primarily reflect shared semantic content rather than surface-level word overlap or concreteness. For this reason, it is unlikely that the higher cross-session semantic similarity for external compared to internal details was driven mainly by the concreteness of external details.

This can be seen in simple examples. Two concrete external sentences describing the same type of event (using different wording) yield high similarity. For example, "I bought a drink at the small shop" and "I picked up something to drink from a nearby store" have a similarity of 0.78, and "I crossed the street when the light changed" and "I walked across the road once it turned green" also have a similarity of 0.78. Importantly, however, two abstract internal sentences with similar emotional meaning also produce highly similar embeddings even with limited word overlap. For instance, "I felt relieved when it was finally over" and "I felt calm after everything had ended" have a cosine similarity of 0.79, and "I felt excited about what was coming next" and "I looked forward to what would happen" have a similarity of 0.73. These examples illustrate that SBERT captures semantic similarity for both abstract internal experiences and concrete external events.

We have revised the 'Measuring semantic similarity across recall sessions' subsection of the Methods to clarify this point.

p.10. Each memory detail was converted into a 768-dimensional embedding vector using a pretrained SBERT model⁵⁹ (all-mpnet-base-v2) implemented in the Sentence Transformers Python package (version 2.2.0, <https://www.sbert.net>). **The model is trained so that texts with similar semantic content produce embedding vectors that align closely in the multidimensional space, regardless of whether the information conveyed is abstract or concrete.**

Minor comments:

(1) The sentence "...researchers often assess the consistency of memories across repeated sessions rather than their accuracy relative to what actually occurred" (p. 1) reads a bit unclear. Consider cutting "relative to what actually occurred".

: We appreciate the reviewer's suggestion. We have revised the sentence as requested:

p.3. Because these uncontrolled real-life experiences—especially internal ones—are difficult to objectively verify, researchers often assess the consistency of memories across repeated sessions rather than their accuracy^{29,30}.

(2) For memory detail segmentation, categorization, and matching, how did annotators decide which narratives they were responsible for? Was there any random selection/assignment involved with respect to which memories were assigned to which raters?

: We thank the reviewer for raising this point. For the segmentation and categorization steps, participants were randomly assigned to annotators. To promote efficiency during the matching step, annotators were assigned to match the data of the same participants they had categorized whenever possible, resulting in 179 of 210 participants' recalls being categorized and matched by the same annotators. The remaining 31 participants' data were categorized and matched by different annotators because the annotators completed the task at different rates, resulting in varying numbers of files processed at each step. We have clarified this in the relevant Methods subsections, as shown below:

p.6. Each narrative was segmented by a single annotator, with each annotator processing an average of 52.5 narratives ($SD = 45.1$, range = 4–149). **Participants were randomly assigned to annotators.** For all but one participant, the same annotator segmented both the first and second recall sessions.

p.7. Each narrative was categorized by a single annotator, with each annotator processing an average of 84 narratives ($SD = 88.5$, range = 2–200). For consistency, the same annotator categorized both the first and second recall sessions for each participant. **Participants were randomly assigned to annotators.**

p.9. Each pair of narratives was matched by a single annotator, with each annotator processing an average of 52.5 narrative pairs ($SD = 32.7$, range = 18–90). **To promote efficiency, annotators were assigned to match narratives from the same participants they had categorized whenever possible, resulting in 179 of 210 participants' narratives being categorized and matched by the same annotators.**

(3) In the methods subsection, “Memory detail categorization”, the authors provide examples of all detail categories except unrelated details. Consider elaborating on and providing an example of this category, too, within the main manuscript.

: We appreciate the reviewers' request for clarification. In the original manuscript, we included definitions and examples for each category that were provided to the annotators. Because annotators were instructed to label any detail that did not fit these predefined categories as “other,” no specific definitions or examples were given for this category and, consequently, it was not described in the manuscript. Following the reviewer's suggestion, we have now added a paragraph in the “Memory detail categorization” subsection elaborating on the instructions related to the “other” category and providing representative examples from participants' responses. We believe this addition will help readers better understand the nature of the “other” category, even though it was not a focus of analysis in the current study.

p.8. **Finally, details that did not align with any of the categories described above were**

classified as 'other.' These were not directly related to the main recalled event and included, but were not limited to, general semantic knowledge about oneself or others (e.g., "He is normally a happy guy"), meta-memory comments (e.g., "I can't remember"), thoughts or feelings about the future related to the recalled event (e.g., "I will have to get another job"), and recollections of past events other than the main event being recalled (e.g., "Last year I didn't have a very good birthday").

(4) At a couple points in the manuscript, I was confused whether the semantic similarity analyses were referring to semantic similarity across matched details (e.g., internal detail i from recall session 1 to internal detail i from recall session 2) or local/global similarity, which is yoked to external details (e.g., Figure 5C, which I think was measuring the former not the latter). Perhaps adding a sentence clarifying which is used when might be useful – but it could have just been me.

: We thank the reviewer for raising this issue. To avoid confusion, we have revised the y-axis labels in Figures 3C and 5C to read "Cross-session semantic similarity" and have replaced "semantic similarity" with this term in the relevant sections of the main text and supplementary information where applicable.

p.10. **Cross-session** semantic similarity was then calculated as the cosine similarity between the embedding vectors of each matched pair.

p.10. Additionally, we examined **cross-session** semantic similarity values after accounting for the number of words in each memory detail pair.

p.10. The resulting residuals—**cross-session** semantic similarity values with word count effects removed—were then averaged within each memory detail category to produce summary scores.

p.14. For **cross-session** semantic similarity, retention, and addition metrics for categories beyond **externally** and **internally oriented details**, see Supplementary Table 3.

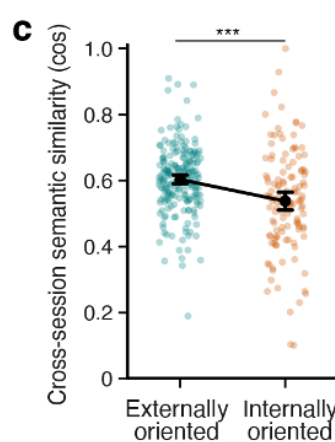


Fig. 3. Transformation of memories for external and internal experiences. c. Mean semantic similarity between matched externally and internally oriented details across recall sessions ($N = 209$ and 127 , respectively), computed as cosine similarity between SBERT-generated text embeddings.

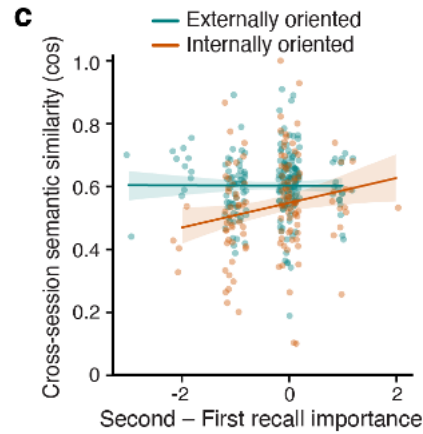


Fig. 5. Relationship between perceived memory importance and the recall of external and internal experiences. c. Correlations between the change in importance ratings (second minus first recall) and the mean semantic similarity of matched details across sessions, analyzed separately for externally oriented ($N = 209$) and internally oriented ($N = 127$) details.

Supplementary Table 3. Memory transformation metrics for Inferences about others, General comments, and Other details categories.

Metric	Inferences about others	General comments	Other
Proportion retained from first recall (%)	37.8 (43.0)	33.0 (37.7)	26.5 (25.4)
Proportion within added details in second recall (%)	2.4 (5.8)	8.2 (10.1)	24.9 (23.7)
Cross-session semantic similarity (cos)	.52 (.17)	.59 (.16)	.57 (.15)

Mean values across participants (SD in parentheses).

Supplementary Table 6. Relationship between subjective memory importance and Inferences about others, General comments, and Other details categories.

Variable pair	Inferences about others		General comments		Other	
	r	p	r	p	r	p

Importance during first recall & Proportion during first recall	-.03	.640	.02	.789	.12	.074
Importance during second recall & Proportion during second recall	-.14	.042	-.02	.770	.20	.004
Second – first recall importance & Proportion retained from first recall	.11	.392	.01	.865	.05	.481
Second – first recall importance & Proportion within added details in second recall	-.14	.050	-.11	.116	.11	.127
Second – first recall importance & Cross-session semantic similarity	.02	.913	.02	.864	.05	.586

r = Pearson correlation coefficient between the pair of variables.

p = uncorrected two-tailed p -value.

(5) Also, with respect to local similarity, the calculation was described as targeting “typically two external details preceding and following” — why typically and not always? Is it something to do when a detail doesn’t have 2 preceding it or following it? If so, it would be useful to add a sentence clarifying when 2 before and after was not used. And, to clarify, this approach was used to calculate local/global similarity for both internal and external details? In other words, local and global similarity always refer to a target detail’s semantic similarity to either a small set (i.e., local) or a large set (i.e., global) of external details?

: We appreciate the reviewer’s request for clarification. We wrote “*typically* the two immediately preceding and following it” because local similarity was calculated between a target memory detail and nearby external details that were *not* immediately adjacent to the target, when one or both of the adjacent details belonged to non-external categories (e.g., if the target detail was the third in the sequence “external–other–internal–external”). The reviewer is also correct that both local and global similarity were computed in the same way for internal and external details: similarity values were calculated between a target detail, internal or external, and nearby (local) or all (global) external details. We have revised the relevant paragraph in the Methods section to clarify these points.

p.11. For each **internally oriented or externally oriented** memory detail recalled during the first session, we computed its semantic similarity to external elements from the same session using the SBERT model⁵⁹ described above. To examine the effects of different levels of external context, we calculated both local and global semantic similarity. Local similarity was defined as the average cosine similarity between the embedding vector of the target detail and those of its **immediately preceding and following externally oriented details**. **If the adjacent details were not externally oriented, the nearest preceding and following externally oriented details were used instead**. Global similarity was defined as the average cosine similarity between the target detail and all other **externally oriented** details from the session.

(6) On page 13, the authors state: "Overall, compared to external details, internal details were more emotionally intense, but showed weaker semantic links to external details and were shorter in length (Table 1)". The fact that internal details are more emotionally intense isn't surprising, right? Given they are defined as an individual's thoughts/feelings from an experience? In reading this, it sounded like a finding, but it seems more like a sanity check. Perhaps a sentence clarifying this would be helpful. Perhaps not.

: The reviewers are correct that higher emotional intensity for internal compared to external details was an expected result, included solely to describe the examined memory detail features. We have slightly revised the relevant paragraph in the Results section to clarify this point.

p.15. Overall, compared to externally **oriented** details, internally **oriented** details were more emotionally intense, **consistent with their definition**, but showed weaker semantic links to externally **oriented** details and were shorter in length (Table 1).

References:

- Dudukovic, N. M., Marsh, E. J., & Tversky, B. (2004). Telling a story or telling it straight: The effects of entertaining versus accurate retellings on memory. *Applied Cognitive Psychology*, 18(2), 125-143.
- Levine, B., Svoboda, E., Hay, J. F., Winocur, G., & Moscovitch, M. (2002). Aging and autobiographical memory: dissociating episodic from semantic retrieval. *Psychology and aging*, 17(4), 677.
- Mahr, J. B., & Csibra, G. (2018). Why do we remember? The communicative function of episodic memory. *Behavioral and brain sciences*, 41, e1.
- Marsh, E. J. (2007). Retelling is not the same as recalling: Implications for memory. *Current Directions in Psychological Science*, 16(1), 16-20.
- Sotgiu, I., Consonni, S., & Marengo, D. (2025). Testing order effects in autobiographical memory research. *Memory*, 1–12. <https://doi.org/10.1080/09658211.2025.2561999>

Reviewer #4 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Communications Psychology initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

: We appreciate the reviewer's efforts in co-reviewing our manuscript.

REVIEWERS' COMMENTS:

Reviewer #1 (Remarks to the Author):

Review

The authors did a thorough job with the reviewer comments, including my own. It was very helpful to see how consistency is calculated as this will allow replication and comparison across protocols, where the denominator differs. I have only one minor comment for the authors.

Consider a small edit:

Change: A recent study examining repeated autobiographical recalls similarly found that thoughts and emotions were recalled less consistently than other types of memory details

To: A recent study examining repeated autobiographical recalls similarly found that thoughts and emotions were recalled less consistently than most other types of memory details (add "most")

: We thank the reviewer for their positive evaluation of the revised manuscript and have revised the sentence accordingly.

Reviewer #2 (Remarks to the Author):

I am satisfied with the author's responses to my comments (Reviewer #2)

: We appreciate the reviewer's evaluation of our manuscript.

Reviewer #3 (Remarks to the Author):

The authors have addressed all of my concerns. The revised version is more clear and precise. Also, I appreciate their effort in conducting a retrospective calculation of inter-rater reliability for their manual coding scheme. While this certainly took some time, I think these scores provide useful qualifiers for the reader about the nature of these categories.

: We appreciate the reviewer's evaluation of our manuscript.

Reviewer #4 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Communications Psychology initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

: We appreciate the reviewer's evaluation of our manuscript.