Skepthical review: *paper*

Summary

This manuscript examines associations among epigenetic age (skin DNA methylation clock), MRI-derived global brain volume, and spatial cognition in Egyptian fruit bats (Secs. 1–3). Using a three-phase foraging task, the authors derive six behavioral metrics intended to separate initial spatial learning, short-term memory, and long-term memory/perseveration (Secs. 2.2.3, 2.3.3), and relate these outcomes to DNAm age and brain volume via regression models adjusted for sex and origin colony (Sec. 2.4.2). The headline results are (i) no detectable association between epigenetic age and total brain volume, (ii) slower initial learning with age (Time_to_First_Reward), and (iii) fewer perseverative errors in later phases with age, alongside little evidence that global brain volume explains age–cognition patterns (Sec. 3, Sec. 4.3–4.4). The study is conceptually interesting and leverages a unique model for healthy aging, but several core elements—outcome modeling (binary/count/censored time), brain-volume quantification validity, specification/reporting of mediation and "resilience" indices, missingness/sample-size inconsistencies, and interpretational scope given a small cross-sectional cohort—require substantial clarification and re-analysis to make the conclusions statistically and methodologically secure (Secs. 2.2–2.4, 3, 4).

Strengths

- Compelling biological motivation: long-lived bats as a model to probe dissociations between aging biomarkers, neuroanatomy, and cognition (Introduction, Sec. 1).
- Multi-modal integration within the same cohort (epigenetic clock, MRI, behavioral task), which is valuable and relatively rare in non-traditional model organisms (Secs. 2.2–2.3).
- Behavioral outcomes are operationalized into multiple phase-specific metrics, enabling a more nuanced profile than a single performance score (Sec. 2.3.3).
- The age-pattern reported—slower initial learning but reduced perseveration later—is intriguing and potentially hypothesis-generating for future mechanistic work (Sec. 3, Sec. 4.3).
- Some transparency about diagnostics and preprocessing steps is present (e.g., mention of skewed brain volume / residual non-normality; bootstrapping mentioned for mediation) (Secs. 2.3, 2.4.2, 3).

Major issues

1. **Outcome distributions are incompatible with the current ordinary least squares (OLS) regression framework, undermining inference for key claims (Sec. 2.4.2, Sec. 3). Several outcomes are binary (STM_Perseverative_Error, LTM_Perseverative_Error), others are counts with likely**

overdispersion/zero inflation (STM/LTM perseveration counts), and Time_to_First_Reward is a time-to-event variable with explicit right-censoring at 3 hours. The manuscript itself notes strong non-normality (e.g., brain volume Shapiro–Wilk $p < 0.001$; non-normal residuals), making p-values/SEs from OLS unreliable for the central results and for downstream residual-based "resilience" indices.

Recommendation: Refit models using outcome-appropriate likelihoods (and report effect sizes with CIs): (i) logistic regression (binomial GLM) for binary outcomes, reporting odds ratios; (ii) Poisson/negative binomial GLMs for counts (assess overdispersion; consider zero-inflated/hurdle variants if many zeros); (iii) survival models for Time_to_First_Reward that treat non-finders as censored at **3 h** (e.g., Cox PH or parametric accelerated failure-time). If linear models are retained for specific reasons, justify explicitly and provide robustness checks (e.g., HC3 robust SEs; transformation sensitivity; influence diagnostics). Update Sec. 2.4.2 to specify each model family and Sec. 3 to report the revised estimates and interpretations.

2. **Global brain volume quantification via "counting non-zero voxels" in skull-stripped mean $b = 0$ images is potentially fragile and may not measure volume as intended (Sec. 2.3.2, Sec. 2.2.2). Non-zero intensity is not equivalent to a brain mask: interpolation/data-type conversion can introduce non-zero background, and true brain voxels can become zero depending on preprocessing. Without confirming that non-brain voxels are exactly zero for every subject, volume estimates could be biased and/or vary with preprocessing quirks rather than anatomy. MRI acquisition/preprocessing details and QC are also too sparse to evaluate measurement validity (Sec. 2.2.2).**

   Recommendation: Clarify the skull-stripping pipeline in Sec. 2.2.2–2.3.2 (software, parameters, whether a binary mask was produced and applied). Prefer computing volume directly from a binarized brain mask in native space rather than intensity-based "non-zero" counting; if thresholding is used, define the threshold and show robustness. Add essential acquisition parameters (scanner, field strength, voxel size, TR/TE, diffusion directions, b-values) and QC steps (e.g., examples of masks, outlier handling, scan/session consistency). Report whether voxel sizes/resolution are identical across scans; if not, show how this is handled. These additions are necessary before concluding there is "no association" between age and global brain volume.

3. **Key confounds for interpreting global brain volume (and its null association with age) are not addressed: body size/allometry and scan/session effects (Secs. 2.1–2.2, Sec. 3, Sec. 4.4). Total brain volume is strongly related to head/body size and can be sensitive to hydration/positioning/protocol variation. Without controlling for body size proxies (e.g., body mass, forearm length, head size) or verifying protocol uniformity (or including ses-**

sion covariates), between-subject variability could mask age effects or create spurious associations (including the reported brain volume–STM perseverative error relationship).

Recommendation: If available, include a body-size covariate (or intracranial volume/head size proxy) in brain-volume and brain-volume→cognition models (Sec. 2.4.2, Sec. 3). Explicitly state whether all animals were scanned on the same scanner and protocol; if multiple sessions/protocol variations exist, include session/date as a covariate or random effect and report sensitivity. If these covariates are unavailable, add a clear limitation and temper claims about (lack of) atrophy/resistance in Sec. 4.4.

4. **The epigenetic clock variable is insufficiently documented, and the manuscript is unclear about chronological age vs DNAm age usage, limiting biological interpretation of "epigenetic age" effects (Sec. 2.1, Sec. 2.2.1, Sec. 3). The clock is referenced but not fully cited/described (training sample size/age range, tissue, performance metrics such as R^2/MAE). It is also unclear whether the age range reported is DNAm age or chronological age, how closely they correspond in this cohort, and whether "age acceleration" (DNAm residual vs chronological age) is considered.**

   Recommendation: In Sec. 2.2.1, provide a full citation and a concise description of the clock (species/tissue, training N, age range, cross-validated performance—R^2 and MAE). In Sec. 2.1 and Sec. 3, state explicitly which age variable is used in each analysis (DNAm age only vs chronological vs both). If chronological age exists, report its correlation with DNAm age and consider age-acceleration analyses (or justify not doing so). If chronological age is unknown/unreliable, state that clearly and moderate language equating DNAm age with "biological aging" (Sec. 4.3–4.4).

5. **Behavioral paradigm and derived metrics are under-specified, and missingness/censoring handling is not transparent, reducing reproducibility and interpretability (Secs. 2.2.3, 2.3.3, 2.4.1, Sec. 3). Critical task details (arena geometry, number/layout of boxes, fixed vs randomized box identities across bats/phases, phase durations, inter-phase intervals) are missing. Time_to_First_Reward appears capped at 3 hours, but the analytic treatment of non-finders (censoring vs fixed maximum) is unclear. Reported degrees of freedom vary across outcomes (e.g., $t(29)$ vs $t(24)$), conflicting with the stated "final analytical sample size of 33 bats with complete data," implying outcome-specific missingness not described.**

   Recommendation: Expand Sec. 2.2.3 and Sec. 2.3.3 with a concise but complete apparatus/procedure description (arena size, box number/positions, randomization, phase lengths, delays—explicitly including the 18-hour LTM delay—and reward rules). Explicitly define how "no reward within 3 h" cases are handled (preferably as censored in a survival model; see Major Issue 1). Add a per-metric missingness table: N used in

each model in Sec. 3, number censored for Time_to_First_Reward, and reasons for missing data (non-participation, tracking failures, incomplete logs). Update Sec. 2.4.1 to describe whether listwise deletion, inner-join merging, or metric-specific inclusion was used.

6. **Mediation and "cognitive resilience" analyses are under-specified and not reported in a way that supports the claims (Sec. 2.4.2, Sec. 3). Mediation (Age → Brain_Volume → Cognition) is described but indirect-effect estimates/CIs are not clearly presented, and mediation is conceptually unlikely given the reported null Age→Brain_Volume path. The resilience indices are residuals from age–cognition regressions, but (i) residualization models appear inconsistent with later covariate inclusion (sex/colony), (ii) residual definitions are not straightforward for binary/count/time-to-event outcomes, and (iii) metric-selection rules for resilience are inconsistent across Methods/Results (e.g., inclusion of STM_Perseveration_Count despite unclear age association).**

   Recommendation: Either (a) remove mediation from the manuscript if it is not central/supported, or (b) implement and report it fully: specify the mediator and outcome models (including covariates), bootstrap details, and report indirect effects with 95% CIs in Sec. 3. For resilience, define it in a model-consistent way: residualize from the full baseline model including DNAm age + sex + colony (and other key covariates), and for non-Gaussian outcomes use appropriate residuals (e.g., deviance residuals for GLMs; martingale/deviance residuals or time-ratio residuals for survival/AFT). Pre-specify or clearly justify which metrics are used and provide a table aligning metric selection between Sec. 2.4.2 and Sec. 3.

7. **Multiplicity, power, and interpretational scope are not adequately addressed given many models, borderline p-values, $N \approx 33$, and a cross-sectional design (Secs. 3, 4.3–4.4). Multiple behavioral outcomes and several brain-volume/cognition/resilience models are tested; selective emphasis on $p \approx 0.05$ findings risks false positives. In addition, concluding "resistance to global brain atrophy" is too strong from cross-sectional null results over a limited age window relative to lifespan.**

   Recommendation: In Sec. 2.4.2, define primary hypotheses/endpoints (or explicitly label the work exploratory) and apply a multiple-comparisons approach across the behavioral family (e.g., Benjamini–Hochberg FDR). In Sec. 3, report effect sizes with 95% CIs (not only p-values) and explicitly note borderline/uncertain results. In Sec. 4.3–4.4, rephrase causal/mechanistic claims as hypotheses ("consistent with…", "may reflect…") and frame the brain-volume null as "no detectable cross-sectional association in this age range with this measurement." Consider adding a brief sensitivity/power statement (e.g., what decline slope could be detected given observed variance and N).

Minor issues

1. Inconsistent implied sample sizes across analyses conflict with the statement of "33 bats with complete data" (Sec. 2.4.1 vs Sec. 3). Some STM models report $t(24)$/ $F(3, 24)$, implying $N = 28$, not 33.

   Recommendation: For every model in Sec. 3 (and in any tables/figures), report the exact N used. Reconcile the "complete data" claim in Sec. 2.4.1 with outcome-specific df, and explain exclusions/missingness explicitly (ideally with a flow diagram or table).

2. Covariate encoding (sex, origin colony) and consistent inclusion across baseline, mediation, and resilience models is not fully specified (Sec. 2.4.2, Sec. 3).

   Recommendation: In Sec. 2.4.2, provide a single model template per outcome type showing all covariates; specify categorical coding (reference levels, contrasts). Ensure the covariate set is consistent between residualization (resilience definition) and subsequent resilience regressions, or explicitly justify differences.

3. Coefficient scaling/units for Brain_Volume effects are unclear ($\mathbf{mm^3}$ vs rescaled or standardized), making effect magnitudes uninterpretable (Sec. 2.3.2, Sec. 3).

   Recommendation: State unambiguously whether coefficients are standardized or raw. If raw, report Brain_Volume units used in the model matrix (e.g., $\mathbf{mm^3}$, mL, per $\mathbf{10^6\,mm^3}$). If standardized, label as such consistently. Add axis labels/units to any relevant figures.

4. Construct validity of "perseveration" vs alternative explanations (strategy, motivation, locomotor speed, sensory differences) is not sufficiently addressed, yet the Discussion leans toward inhibitory-control interpretations (Sec. 4.3–4.4).

   Recommendation: Add a short Discussion paragraph outlining plausible alternative explanations and what additional measures would distinguish them (e.g., activity level/movement speed covariates; trial-by-trial trajectories; latency vs path efficiency). If such covariates exist, include them as sensitivity analyses.

5. The unexpected positive association between brain volume and STM_Perseverative_Error is not tested for robustness and may reflect outliers or model misspecification (Sec. 3, Sec. 4.3).

   Recommendation: Report the effect size with CI and N for this association, show influence diagnostics (e.g., Cook's distance) and re-estimate under appropriate outcome models (logistic/GLM; see Major Issue 1). Discuss as tentative and possibly multiplicity-driven unless robust across specifications.

6. MRI methods are currently too high-level for reproducibility (Sec. 2.2.2): missing acquisition parameters, motion/QC criteria, and whether preprocessing involved interpolation that could affect intensity-based volume estimation.

Recommendation: Add key acquisition and preprocessing details in Sec. 2.2.2 (with full parameters in Supplement if needed) and document QC/exclusion criteria (e.g., motion artifacts, failed skull stripping).

7. Data linkage and ID harmonization description focuses on implementation details but does not state how correctness of multimodal matching was validated (Sec. 2.3.1).

   Recommendation: Summarize validation steps (cross-checking sex/colony metadata, manual verification of mismatches) and state whether any animals were dropped due to unresolved ID discrepancies. Move low-level string-replacement details to Supplement or code comments.

8. Data/code availability and ethics oversight are not fully specified (Sec. 2.1; general).

   Recommendation: Add explicit Data and Code Availability statements (repository link or "upon request," with constraints). Expand the ethics statement in Sec. 2.1 to include approving body and protocol identifiers and confirm adherence to relevant animal research guidelines.

9. Time_to_First_Reward definition may be ambiguous if "Absolute_Time" is a timestamp rather than elapsed duration (Sec. 2.3.3).

   Recommendation: Clarify whether Absolute_Time is elapsed-from-phase-start. If it is wall-clock time, redefine Time_to_First_Reward as (first_reward_time − phase_start_time) and ensure consistency in all analyses/figures.

Very minor issues

1. Typographical/LaTeX and placeholder artifacts reduce clarity: broken words (e.g., "ex\nceeding"), duplicated descriptions (e.g., STM_Perseverative_Error), malformed figure references ("Figure LABEL:fig:fig:…" / "Figure ??"), inconsistent variable-name formatting, and placeholder affiliation text in the title block (Secs. 1, 3; front matter).

   Recommendation: Thoroughly proofread and clean LaTeX/build artifacts; ensure all figures are present, numbered, and referenced consistently; remove duplicated paragraphs; standardize variable-name formatting (code font) across Secs. 2–4; replace placeholder affiliation content with journal-appropriate information.

2. Keywords include irrelevant terms (e.g., "Astronomy data analysis") and omit key domain concepts (Abstract).

   Recommendation: Replace keywords with relevant terms (e.g., epigenetic clock, cognitive aging, MRI, spatial memory, Rousettus aegyptiacus, cognitive resilience) and remove unrelated items.

3. Notation is used (β, model formulas with "∼" and "+")$without defining whether$ \beta$ is standardized/unstandardized or what error assumptions are made (Methods).

Recommendation: Add a brief notation paragraph defining coefficient interpretation, standardization, and (where relevant) link functions/error distributions for each model family.

Mathematical consistency audit

This section audits **symbolic/analytic** mathematical consistency (algebra, derivations, dimensional/unit checks, definition consistency).

Maths relevance: light

The manuscript contains limited explicit mathematics, primarily operational definitions of derived variables (brain volume and behavioral metrics) and linear regression model specifications (including a residual-based 'resilience index' and a conceptual mediation analysis). The main internal-consistency checks therefore center on (i) definition-to-usage consistency, (ii) degrees-of-freedom/sample-size consistency across models, and (iii) unit/scaling clarity for coefficients and derived measures.

Checked items

1. ✔ **Brain volume from voxel count** (Sec. 2.3.2, p.3)
    - **Claim:** Total brain volume is computed as (number of non-zero voxels in skull-stripped mean $b = 0$ image) × (voxel volume from affine), expressed in $\mathbf{mm}^3$.
    - **Checks:** dimensional/units, definition consistency
    - **Verdict:** PASS; confidence: medium; impact: moderate
    - **Assumptions/inputs:** Voxel dimensions extracted correspond to physical spacing d_x, d_y, d_z., Skull-stripped images have exactly zero outside-brain values (so 'non-zero' equals in-brain)., Averaging $b = 0$ volumes preserves the zero-valued background.
    - **Notes:** $d_x \cdot d_y \cdot d_z$ yields $\mathbf{mm}^3$ and multiplying by voxel count yields $\mathbf{mm}^3$, consistent with the stated output units. Potential edge case: averaging can create small non-zero background if masking is not strict, but the paper asserts skull-stripped images, so symbolically the formula is coherent.
2. ⚠ $b = 0$ **averaging step** (Sec. 2.3.2, p.3)
    - **Claim:** The first three volumes ($b = 0$) are extracted from a $4D$ NIfTI and averaged along the 4th dimension to form a $3D$ structural image.
    - **Checks:** logic/derivation
    - **Verdict:** UNCERTAIN; confidence: low; impact: minor
    - **Assumptions/inputs:** The first three $4D$ volumes correspond to $b = 0$ images., Averaging is performed voxelwise.

- **Notes:** No symbolic issue, but the manuscript does not justify/define that 'first three volumes' are indeed $b = 0$ for all subjects; this is a missing assumption needed to verify the correctness of the described extraction rule.

3. ⚠ **Phase 1 Time_to_First_Reward definition** (Sec. 2.3.3, p.3)
   - **Claim:** Time_to_First_Reward is the 'Absolute_Time' of the first reward retrieval ('F') in the correct box; if none, set to total phase duration (**3** hours).
   - **Checks:** definition consistency, units/dimensional sanity
   - **Verdict:** UNCERTAIN; confidence: medium; impact: moderate
   - **Assumptions/inputs:** Absolute_Time is a duration from phase start (not a wall-clock timestamp).
   - **Notes:** If Absolute_Time is a timestamp, the metric is not a time-to-event duration and mixing it with a fixed **3**-hour duration is dimensionally inconsistent. The paper does not define whether Absolute_Time is elapsed-time or absolute timestamp.

4. ✔ **Errors_before_First_Reward definition** (Sec. 2.3.3, p.3)
   - **Claim:** Errors_before_First_Reward counts incorrect box entries ('E') occurring before Time_to_First_Reward in Phase 1.
   - **Checks:** definition consistency
   - **Verdict:** PASS; confidence: medium; impact: minor
   - **Assumptions/inputs:** Time_to_First_Reward is a comparable time measure to the event times being filtered., Events are totally ordered by the time variable used.
   - **Notes:** Given a well-defined time variable, the counting rule is logically consistent.

5. ✔ **Regression specification: Age → Brain Volume** (Sec. 2.4.2, p.4)
   - **Claim:** Fit Brain_Volume ∼ DNAmAge + Sex + Origin_colony and report coefficient and model statistics.
   - **Checks:** notation/definition consistency
   - **Verdict:** PASS; confidence: high; impact: moderate
   - **Assumptions/inputs:** Linear model with intercept is used., Sex and Origin_colony are encoded as covariates (e.g., dummy variables).
   - **Notes:** Model formula is coherent and matches the stated goal.

6. ✔ **Degrees of freedom consistency for Age → Brain Volume model** (Sec. 3, p.5)
   - **Claim:** For $N = 33$, the model reports $t(29)$ and $F(3, 29)$.
   - **Checks:** algebra/df sanity
   - **Verdict:** PASS; confidence: high; impact: minor

- **Assumptions/inputs:** Model includes **3** predictors (DNAmAge, Sex, Origin_colony) plus intercept., $N = 33$ observations were used.
- **Notes:** Residual df $= N - (p + 1) = 33 - 4 = 29$, matching $t(29)$. Numerator df for F is $p = 3$, matching $F(3, 29)$.

7. ✖ **Degrees of freedom inconsistency for STM models** (Sec. 3, p.6 (STM Perseveration Count; STM Perseverative Error))
    - **Claim:** STM models report $t(24)$ and $F(3, 24)$ despite earlier claims of a complete $N = 33$ dataset.
    - **Checks:** algebra/df sanity, definition-to-results consistency
    - **Verdict:** FAIL; confidence: high; impact: critical
    - **Assumptions/inputs:** Same covariate structure (DNAmAge + Sex + Origin_colony + intercept)., Reported dfs correspond to standard OLS output.
    - **Notes:** If the model has **3** predictors + intercept and residual df is **24**, then $N = 24 + 4 = 28$, not **33**. This contradicts the stated inner-join complete-data master dataset (Sec. 2.4.1). Either per-metric missingness exists (not described) or reported dfs/results are inconsistent.

8. ✔ **Binary outcomes modeled with linear regression** (Secs. 2.4.2 and 3, pp.4–6 (STM/LTM Perseverative Error models))
    - **Claim:** Binary perseverative error variables are analyzed with linear regression and interpreted via the sign of β.
    - **Checks:** assumption clarity, interpretation sanity
    - **Verdict:** PASS; confidence: medium; impact: minor
    - **Assumptions/inputs:** OLS is used even for binary dependent variables.
    - **Notes:** Symbolically consistent: the sign of an OLS coefficient still indicates direction of association in a linear probability model. However, the paper does not label this explicitly and notes assumption violations; this is more a modeling-clarity issue than an algebraic contradiction.

9. ⚠ **Mediation analysis criterion** (Sec. 2.4.2, p.4)
    - **Claim:** Mediation is inferred by reduced DNAmAge coefficient when adding Brain_Volume and a significant Brain_Volume coefficient; indirect effect assessed by bootstrap CI.
    - **Checks:** definition completeness
    - **Verdict:** UNCERTAIN; confidence: medium; impact: moderate
    - **Assumptions/inputs:** A specific indirect-effect estimand (not stated) is used for bootstrap., Age→Brain_Volume and Brain_Volume→Cognition paths are modeled linearly.
    - **Notes:** The description is conceptual but does not define the indirect effect mathematically (e.g., which coefficients are multiplied, whether covariates are included in both path models). Missing explicit estimand blocks verifica-

tion of internal analytic consistency.

10. ✔ **Resilience index as regression residual** (Sec. 2.4.2, p.5; Sec. 3, p.7)
    - **Claim:** Resilience index is residuals from Cognitive_Metric ∼ DNAmAge; then regress residuals on Brain_Volume + Sex + Origin_colony.
    - **Checks:** algebra/definition consistency
    - **Verdict:** PASS; confidence: medium; impact: minor
    - **Assumptions/inputs:** Residuals are computed from an intercept-including OLS fit., Residuals are treated as a continuous response in the second-stage model.
    - **Notes:** Using residuals as a derived variable is algebraically consistent. Potential definition mismatch exists if resilience is intended to be covariate-adjusted, but the procedure as written is internally coherent.
11. ⚠ **Coefficient/unit clarity for Brain_Volume in reported results** (Sec. 3, pp.5–7)
    - **Claim:** Reported β values for Brain_Volume-related models are meaningful without explicit unit/scaling statement.
    - **Checks:** units/dimensional consistency, notation consistency
    - **Verdict:** UNCERTAIN; confidence: medium; impact: moderate
    - **Assumptions/inputs:** Brain_Volume is in $\mathbf{mm^3}$ as stated in Methods.
    - **Notes:** Because β values are reported without units and appear on very different scales across outcomes, it is unclear whether coefficients are standardized or whether Brain_Volume was rescaled. This is a definition/notation gap rather than a provable algebra error.
12. ✔ **Spearman correlation notation and interpretation** (Sec. 3, p.6)
    - **Claim:** Spearman rank correlations are reported as ρ with accompanying p-values.
    - **Checks:** notation sanity
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** ρ denotes Spearman's rho.
    - **Notes:** Notation is standard and internally consistent within the manuscript.

Limitations

- The provided content contains no explicit equation numbering and references to figures are unresolved ("Figure ??"), limiting cross-checking of whether plotted quantities and axes match the variable definitions and units.
- Several analytic procedures (bootstrapped mediation indirect effect, exact encoding of categorical covariates, possible variable rescaling/standardization) are described verbally without explicit mathematical definitions, preventing full verification of those

components.

- This audit is restricted to internal symbolic/analytic consistency and does not assess whether chosen statistical models are appropriate for the data beyond basic definitional/assumption clarity.

Numerical results audit

This section audits **numerical/empirical** consistency: reported metrics, experimental design, baseline comparisons, statistical evidence, leakage risks, and reproducibility.

All **19** automated numeric checks passed. Cohort counts and subgroup totals reconcile exactly, mean age lies within the stated range, and reported p-values for multiple t- and F-statistics match computed two-sided/upper-tail values within the stated tolerances (typically with sub-**0.001** absolute differences).

Checked items

1. ✔ **C1** (Page 2 (Methods → 2.1 Animal Cohort))
    - **Claim:** Initial cohort size is **41**, and final analytical sample is **33** after exclusions.
    - **Checks:** parts_vs_total
    - **Verdict:** PASS
    - **Notes:** Inequality check only.
2. ✔ **C2** (Page 5 (Results, first paragraph))
    - **Claim:** Initial cohort comprised **41** subjects with **8** exclusions, yielding analytical cohort $N = \mathbf{33}$.
    - **Checks:** parts_vs_total
    - **Verdict:** PASS
    - **Notes:** Exact integer arithmetic.
3. ✔ **C3** (Page 2 (Methods → 2.1 Animal Cohort))
    - **Claim:** Sex distribution was **22** males and **19** females in the study cohort (implies $N = 41$).
    - **Checks:** parts_vs_total
    - **Verdict:** PASS
    - **Notes:** Exact integer sum.
4. ✔ **C4** (Page 2 (Methods → 2.1 Animal Cohort))
    - **Claim:** Origin colonies: Aseret ($N = \mathbf{23}$) and Herzliya ($N = 18$) (implies $N = 41$).
    - **Checks:** parts_vs_total
    - **Verdict:** PASS

- **Notes:** Exact integer sum.

5. ✔ **C5** (Page 5 (Results, first paragraph))
   - **Claim:** Analytical cohort ($N = 33$) consisted of 19 males and 14 females.
   - **Checks:** parts_vs_total
   - **Verdict:** PASS
   - **Notes:** Exact integer sum.
6. ✔ **C6** (Page 2 (Methods → 2.1 Animal Cohort) and Page 5 (Results, first paragraph))
   - **Claim:** Age range 6.62 to 15.07 years; mean ± SD 9.87 ± 1.96 years. Mean should lie within the stated range.
   - **Checks:** range_check
   - **Verdict:** PASS
   - **Notes:** Mean-within-range sanity check.
7. ✔ **C7** (Page 5 (Results, brain volume regression))
   - **Claim:** For epigenetic age predicting total brain volume: $t(29) = 0.040$ and $p = 0.968$ should be consistent for a two-sided t-test.
   - **Checks:** p_value_from_test_statistic
   - **Verdict:** PASS
   - **Notes:** Allow rounding; verify two-sided p.
8. ✔ **C8** (Page 5 (Results, brain volume regression))
   - **Claim:** Overall model $F(3, 29) = 1.185$ and $p = 0.333$ should be consistent.
   - **Checks:** p_value_from_F_statistic
   - **Verdict:** PASS
   - **Notes:** Allow rounding.
9. ✔ **C9** (Page 5 (Results, brain volume regression))
   - **Claim:** Reported $R^2 = 0.109$ should be consistent with $F(3, 29) = 1.185$ given n and k.
   - **Checks:** F_R2_consistency
   - **Verdict:** PASS
   - **Notes:** R^2 and F rounded; compare reconstructed F.
10. ✔ **C10** (Page 5 (Results))
    - **Claim:** Shapiro-Wilk normality test: $W = 0.735$ with $p < 0.001$. Check that p-threshold statement is numerically plausible (not exact p).
    - **Checks:** inequality_check
    - **Verdict:** PASS

- **Notes:** Only a cheap sanity check of bound value.

11. ✔ **C11** (Page 5 (Results, Phase 1 Time to First Reward regression))
    - **Claim:** Time_to_First_Reward model: $t(29) = 2.051$ and $p = 0.049$ should be consistent for a two-sided t-test.
    - **Checks:** p_value_from_test_statistic
    - **Verdict:** PASS
    - **Notes:** Near 0.05 threshold; allow rounding.
12. ✔ **C12** (Page 6 (Results, Phase 1 Errors before First Reward regression))
    - **Claim:** Errors_before_First_Reward model: $t(29) = 0.805$ and $p = 0.427$ should be consistent.
    - **Checks:** p_value_from_test_statistic
    - **Verdict:** PASS
    - **Notes:** Allow rounding.
13. ✔ **C13** (Page 6 (Results, STM Perseveration Count regression))
    - **Claim:** STM_Perseveration_Count age effect: $t(24) = -2.860$ and $p = 0.009$ should be consistent.
    - **Checks:** p_value_from_test_statistic
    - **Verdict:** PASS
    - **Notes:** Small p, rounding likely.
14. ✔ **C14** (Page 6 (Results, STM Perseveration Count regression))
    - **Claim:** STM_Perseveration_Count overall model: $F(3, 24) = 4.816$ and $p = 0.009$ should be consistent.
    - **Checks:** p_value_from_F_statistic
    - **Verdict:** PASS
    - **Notes:** Small p, rounding likely.
15. ✔ **C15** (Page 6 (Results, STM Perseveration Count regression))
    - **Claim:** STM_Perseveration_Count model reports $R^2 = 0.376$; check consistency with $F(3, 24) = 4.816$.
    - **Checks:** F_R2_consistency
    - **Verdict:** PASS
    - **Notes:** R^2 and F are rounded; compare reconstructed F.
16. ✔ **C16** (Page 6 (Results, STM Perseverative Error regression))
    - **Claim:** STM_Perseverative_Error: $t(24) = -1.568$ and $p = 0.130$ should be consistent.
    - **Checks:** p_value_from_test_statistic

- **Verdict:** PASS
- **Notes:** Allow rounding.

17. ✔ **C17** (Page 6 (Results, LTM Perseverative Error regression))
    - **Claim:** LTM_Perseverative_Error: $t(29) = -2.878$ and $p = 0.007$ should be consistent.
    - **Checks:** p_value_from_test_statistic
    - **Verdict:** PASS
    - **Notes:** Small p, rounding likely.
18. ✔ **C18** (Page 6 (Results, LTM Perseveration Count regression))
    - **Claim:** LTM_Perseveration_Count: $t(29) = 0.487$ and $p = 0.630$ should be consistent.
    - **Checks:** p_value_from_test_statistic
    - **Verdict:** PASS
    - **Notes:** Allow rounding.
19. ✔ **C19** (Page 6 (Results, Spearman correlations summary))
    - **Claim:** Spearman correlations: $\rho = 0.29$ (age vs Time_to_First_Reward), $\rho = -0.43$ (age vs STM_Perseveration_Count), $\rho = -0.42$ (age vs LTM_Perseverative_Error), and strong correlations $\rho = -0.70$ and $\rho = -0.68$. Sanity check all $|\rho| \leq 1$.
    - **Checks:** range_check
    - **Verdict:** PASS
    - **Notes:** Exact bound check.

Limitations

- Only parsed text (no tables/figures/equations rendered) was available; figure-referenced numeric content could not be extracted or checked.
- Many statistical claims (coefficients, exact p-values for nonparametric tests, bootstrap CI results) require underlying data or model outputs not present in the PDF text.
- No raw brain volume values, voxel counts, or behavioral logs were provided, preventing direct recomputation of derived metrics.
- Correlation significance statements (p-thresholds) could not be verified because exact p-values depend on effective sample size and tie handling, which were not provided.
- Regression coefficient and resilience-model numeric claims could not be independently verified without underlying model outputs (e.g., standard errors/df) or raw data.