

Skeptical review: Microstructural Brain Signatures of Adaptive Cognitive Strategies in Long-Lived Bats: An ROI-based DTI and Behavioral Resilience Analysis

Summary

This manuscript investigates cognitive aging and proposed “cognitive resilience” in Egyptian fruit bats (*Rousettus aegyptiacus*) by combining (i) a custom dynamic foraging task with several derived behavioral metrics (perseverative errors, switch cost, lose–shift index, exploration–exploitation; Sec. 2.2.1–2.2.3) and (ii) ROI-based diffusion tensor imaging (DTI) measures (FA/MD/AD/RD) extracted from a 24-region atlas (Sec. 2.3–2.4.2). Linear models using age (named as a DNAm/skin-clock variable in places) and sex report no detectable age associations for any behavioral metric (Sec. 3.1) and no significant age effects for ROI DTI metrics after FDR correction (Sec. 3.2). Critically, the paper then documents a catastrophic atlas-to-subject misregistration (Dice similarity reported as 0.0; Sec. 3.2, Fig. 6–9) alongside physiologically implausible FA values, implying the ROI DTI extraction is not biologically interpretable and blocking the planned brain–behavior and residual-based “resilience score” analyses (Sec. 2.4.4, Sec. 3.3). As written, the manuscript’s highest value is as a transparent methodological post-mortem and cautionary case study for non-model-species diffusion MRI and multimodal aging pipelines; however, several key reporting, definitional, and statistical issues (especially around the DTI pipeline, sample accounting, and behavioral metric validation/sensitivity) currently limit interpretability and reproducibility.

Strengths

- Compelling big-picture motivation: long-lived bats as a comparative model for healthspan/cognitive aging, framed clearly in the Introduction and Discussion/Conclusions (Sec. 1, Sec. 4.1–4.4).
- Ambitious multimodal design linking behavioral strategy measures with microstructural MRI, with a clear intended analytic framework (Sec. 2.4.2–2.4.4).
- Notable transparency: the manuscript explicitly documents the DTI registration failure with visual QC and quantitative overlap metrics and avoids overclaiming mechanistic imaging conclusions (Sec. 3.2–3.4; Fig. 6–9).
- Behavioral task goes beyond simple performance measures and attempts to operationalize flexibility/updating and exploration–exploitation, which could be valuable for future bat cognition work (Sec. 2.2.1–2.2.3).
- Consistent use of explicit regression-formula reporting and multiple-comparisons control where appropriate (Sec. 2.4.2–2.4.4; Sec. 3.1–3.2).

Major issues

1. **DTI ROI results are currently invalid due to catastrophic atlas-to-subject misalignment (Dice reported as 0.0 across subjects) and physiologically implausible diffusion values (e.g., near-zero FA), yet the Methods still read as though ROI extraction yielded meaningful data products (Sec. 2.3.1–2.3.3; Sec. 3.2; Fig. 6–9).** This undermines the central mechanistic aim (microstructural aging signatures) and also makes the reported null ROI age associations uninterpretable as “no effect” (they may simply reflect sampling background/incorrect voxels).

Recommendation: Make the manuscript internally consistent by (i) explicitly labeling the ROI-derived DTI outputs as non-interpretable immediately where ROI extraction is described (end of Sec. 2.3.3, with a forward reference to Sec. 3.2), and (ii) either: (A) fix and re-run the diffusion preprocessing/registration/ROI pipeline, demonstrating non-zero overlap and plausible scalar ranges before any inferential statistics; or (B) if a fix is not feasible with the current data, reframe the paper as a methodological case study/post-mortem, and move/remove ROI age statistics and any downstream brain–behavior plans from the main Results (Sec. 3.2–3.3) to clearly marked “invalid exploratory output” or Supplementary material. In both cases, revise the Discussion (Sec. 4.3–4.4) to avoid language implying preserved microstructure.

2. **The diffusion preprocessing/registration pipeline is under-specified and partially ambiguous, preventing diagnosis and reproducibility of the failure mode (Sec. 2.3.1–2.3.3; Abstract wording about scans being “stretched to uniform dimensions”; Sec. 3.2).** Key missing details include: exact software/toolchain and versions; eddy/motion and susceptibility correction; b-vector handling; denoising/bias correction; brain extraction/masking; coordinate conventions (RAS/LPS), affines/headers; which images live in which space (native diffusion, b_0 , FA space, template/atlas space); registration type (linear vs nonlinear), cost function and constraints; interpolation for label resampling; and precisely how Dice was computed (which masks; thresholding; binarization).

Recommendation: Expand Sec. 2.3.1–2.3.3 into a stepwise, reproducible pipeline description with explicit file spaces and transforms. At minimum, report for each relevant image (atlas labels, subject diffusion/ b_0 , FA map, brain mask): voxel size, dimensions, orientation convention, and affine; describe all preprocessing steps and parameters (eddy/motion, susceptibility, bvec rotation, denoising, bias-field correction if any); specify the registration path (e.g., subject b_0 → template b_0 /FA; atlas → template; or subject → atlas) and whether it is affine or nonlinear, with cost function and masking strategy. Define Dice computation precisely (whole-brain masks vs union of atlas labels; binarization/thresholds). Add a brief “post-mortem” subsection in Sec. 3.2 identifying the most likely cause(s) (e.g., header mismatch, orientation swap, wrong space, non-binarized masks) and concrete safeguards/QC gates to prevent recurrence.

3. **Sample sizes, inclusion criteria, and age ranges are inconsistent across the manuscript** (e.g., references to an initial cohort of 41, DTI $N = 33$, “complete data” $N = 31$, and conflicting max ages such as 15.07 vs 13.8; Sec. 2.1; Sec. 3.1–3.2; Fig. 1–3 captions/text). These inconsistencies impede interpretation, power assessment, and reproducibility.

Recommendation: Provide a consolidated cohort accounting (ideally a CONSORT-style/funnel table) in Sec. 2.1–2.4.1 and/or a new table/figure: counts at each stage separately for behavior and imaging; explicit inclusion/exclusion criteria and reasons (missingness, QC failures, unusable scans); and the exact N used for each model family in Sec. 3.1–3.3. Reconcile and standardize the min/max age for each analytic subset (behavior-only, DTI-only, multimodal intersection) and ensure figure captions state the N and age range for the plotted subset.

4. **Behavioral metric definitions are promising but not yet rigorous/validated enough to support the central interpretation that “no age effects” implies preserved cognition rather than limited sensitivity, floor/ceiling effects, or confounding by time-on-task/motivation** (Sec. 2.2.1–2.2.3; Sec. 3.1; Sec. 4.3–4.4). Several metrics also have edge cases that can materially affect estimates (e.g., never visiting the correct box in a phase; exploitation ≈ 0 leading to unstable ratios; how “\text{Absolute_Time}” is defined across phases; how phase boundaries are handled).

Recommendation: Strengthen Sec. 2.2.1–2.2.3 by adding precise operational definitions (including edge-case handling) for each metric—ideally with pseudocode or a worked example in an Appendix. Explicitly define timing variables (e.g., whether latency is phase-relative or continuous across the session), rules for missing/undefined events (never finding correct box; no ‘Lose’ events; denominators of 0), and whether/how metrics are normalized by phase duration or number of visits. In Sec. 3.1, add basic psychometric/robustness reporting: metric distributions, missingness per phase, within-bat variability, correlations among metrics, and checks for floor/ceiling effects. If feasible, add sensitivity analyses using alternative formulations (e.g., log-latency or ratio-based switch cost; lose–shift conditional on having previously exploited the correct location; exploration normalized by visit count/time). Update Sec. 4.3–4.4 to more sharply separate “resilience” from “insufficient task sensitivity/limited age span.”

5. **Null findings are not supported with sufficient quantitative reporting and sensitivity/power considerations. The manuscript does not systematically provide effect sizes, uncertainty (CI), or exact p/q values for age effects (behavior; DTI), and does not estimate the smallest detectable effects given N and multiple-testing burden** (Sec. 2.4.2–2.4.4; Sec. 3.1–3.2; Sec. 4.3–4.4). This makes it difficult to judge what magnitudes of aging effects the study can rule out.

Recommendation: Add summary tables (main text or Supplementary) for Sec. 3.1 and the intended Sec. 3.2 models listing: $\beta(\text{age})$, SE, 95% CI, p -value, and (where applicable) FDR-adjusted q -value, plus $\beta(\text{sex})$ and coding details. Include a brief sensitivity/power analysis (post hoc is acceptable) indicating the minimum detectable standardized age effect sizes for behavioral metrics given $N \approx 31$ and for imaging given the number of tests (24 ROIs \times 4 metrics). Use these results to temper claims in Sec. 4.3–4.4 (e.g., “small-to-moderate declines cannot be excluded”).

6. **The manuscript’s “cognitive resilience score” (residual-based framework) is presented as a key conceptual contribution, but its role is confused because (i) no robust behavioral age decline is observed, and (ii) imaging is unusable; additionally, some wording suggests residuals are mathematically impossible to compute without significant decline (Sec. 2.4.4; Sec. 3.3). There is also a specification mismatch: age-effect models include sex, but residualization is described as $\text{Behavioral_Metric} \sim \text{Age}$ (omitting sex), which can leave systematic sex differences in the residuals.**

Recommendation: Rewrite Sec. 2.4.4 and Sec. 3.3 to distinguish planned vs executed analyses and to clarify that residuals are always computable but may not be interpretable as “resilience to decline” when the estimated age slope is ~ 0 . Either (a) define the score as “age-adjusted performance” and keep interpretation cautious, or (b) propose alternative resilience operationalizations that do not require observable decline in this cohort (e.g., learning parameters from a reinforcement-learning model; performance conditional on task difficulty/phase transitions). Ensure the residualization model matches the covariates used elsewhere ($\text{Behavioral_Metric} \sim \text{Age} + \text{Sex}$, and consider colony/origin if relevant; see below), or justify exclusions explicitly.

7. **The age predictor is inconsistently described and named (chronological age vs DNA methylation age; variable name indicates a skin clock), which is conceptually central for an “aging” study and affects interpretation of regression results (Sec. 2.1; Sec. 2.4.2–2.4.4; Fig. 2–3; Sec. 3).**

Recommendation: Standardize terminology and explicitly define the primary age variable in Sec. 2.1 and Sec. 2.4.2: whether it is chronological age, DNAm-predicted age, or both. If both exist, report their relationship (correlation, bias) and state which is used in each analysis. Replace opaque phrases/labels (e.g., “Skin+Sexaging analyses”) with clear wording throughout the Results and figure captions.

Minor issues

1. Figure set (especially Fig. 1–4 and Fig. 8–10) has several presentation problems that reduce standalone interpretability: missing in-figure legends/keys, unclear axes/labels, inconsistent N s across panels, limited reporting of missingness/attrition reasons, and

accessibility concerns (color palettes, font sizes). Some captions also describe analyses not shown or rely on color alone for meaning (Sec. 3.1–3.3).

Recommendation: Revise Fig. 1–4 and Fig. 8–10 captions and layouts to be self-contained: state N and age range for each panel; add embedded legends and color keys; use colorblind-safe palettes; increase font sizes; and explicitly annotate missingness/attrition reasons (especially in the cohort funnel). For inferential displays (e.g., heatmaps), prefer effect sizes with direction and significance overlays rather than p -values alone, and clearly state correction methods in captions.

2. Quality control (QC) is currently presented late, after ROI extraction, but given the centrality of registration it should be a gated step with explicit pass/fail criteria (Sec. 2.3; Sec. 3.2).

Recommendation: Add an explicit QC subsection in Sec. 2.3 (or at the start of Sec. 3.2) describing automated and visual QC performed for every subject (not “several”), including thresholds/criteria for acceptable alignment (Dice/Jaccard ranges, mutual information, visual checklist). State how QC outcomes affect inclusion/exclusion in subsequent analyses and reflect this in the sample accounting table.

3. Potential grouping/confounding variables (e.g., colony/origin referenced in figures) are not addressed in the modeling, despite being plausible sources of systematic behavioral or imaging differences (Sec. 2.1; figures referencing origin; Sec. 2.4.2–2.4.4).

Recommendation: Either incorporate colony/origin (and any other major husbandry/site factors) as covariates or random effects where appropriate, or justify their exclusion (e.g., perfectly confounded with other variables, too few levels). At minimum, report whether age distributions differ by origin and whether origin correlates with key behavioral metrics.

4. Model specification details are incomplete: centering/scaling of predictors, sex coding, transformations for skewed outcomes (especially latency-based metrics), and diagnostic checks are not reported (Sec. 2.4.2–2.4.4).

Recommendation: In Sec. 2.4.2–2.4.4, state whether predictors were centered/standardized, how sex was coded (reference level), whether outcomes were transformed (e.g., log-latency), and which diagnostics were used (residual plots, heteroscedasticity, influential points). If transformations are adopted, reflect them consistently in figures and interpretation (Sec. 3.1).

5. Atlas/ROI provenance and anatomical interpretability are under-described: ROIs are largely numeric IDs without mapping to anatomical structures, and the atlas’ origin/validation and resolution are unclear (Sec. 2.3.2; Sec. 3.2).

Recommendation: Add a table (main text or Supplementary) mapping ROI IDs to anatomical labels and providing atlas provenance: how it was created, species-specificity, resolution, and any validation. Even if current ROI results are unusable, this

improves the manuscript’s value as a reusable methodological resource.

6. Behavioral task description lacks certain operational details needed for reproducibility and for interpreting strategy metrics (layout/number of boxes, phase duration/transition rules, training, handling aborted/invalid trials; Sec. 2.2.1–2.2.3).

Recommendation: Expand Sec. 2.2.1–2.2.3 with concrete task parameters: number/arrangement of boxes, reward schedule per phase, how/when phase transitions occur, session termination criteria, training procedures, and how aborted/invalid trials are handled. Ensure metric definitions reference these parameters unambiguously.

7. Discussion sometimes blends data-driven statements with speculative interpretations about resilience/preserved cognition and (attempted) microstructure, without clearly flagging speculative status (Sec. 4.3–4.4; also parts of Sec. 3.4).

Recommendation: Revise Sec. 4.3–4.4 (and any relevant Results phrasing) to clearly separate: (i) what is supported (no detectable behavioral age effects within this age range; DTI ROI pipeline failure) from (ii) hypotheses (true resilience; later-life decline not sampled; task insensitivity; low power). Add a concise limitations paragraph in Conclusions summarizing these interpretive constraints.

8. Ethics and welfare reporting is brief for animal MRI work (approvals, anesthesia/monitoring, post-procedure care; Sec. 2.1; Sec. 2.3).

Recommendation: Add explicit ethical approvals (committee, protocol ID if permissible) and a short description of anesthesia/monitoring and recovery procedures in Sec. 2.1 or a dedicated ethics subsection, following standard animal-research reporting expectations.

Very minor issues

1. Copyediting/formatting inconsistencies (heading styles, mixed quotation/LaTeX formatting, OCR-like artifacts such as “Skin+Sexaging analyses,” and promotional adjectives like “pioneering/robust” without qualification) reduce clarity (Sec. 1; Sec. 2.2–2.4; Sec. 3; Sec. 4.1–4.4).

Recommendation: Perform a careful copyedit: standardize headings and numbering, fix OCR artifacts and variable names, use consistent math formatting (e.g., “ $p > 0.17$ ”, “95%”), and moderate promotional descriptors unless tied to specific evidence.

2. Figure captions sometimes omit essential metadata (N , correction method, definition of color scales) and occasionally rely on color alone to convey meaning (Fig. 1–10; Sec. 3.1–3.3).

Recommendation: Ensure every caption states N , defines all visual encodings (color/lines/symbols), specifies statistical corrections (e.g., FDR), and avoids color-only meaning (add labels/annotations where feasible).

3. Use of code-like variable/file names directly in prose can impede readability (e.g., DNAmAgeBat.Rousettus.aegyptiacus_Skin; Atlas.nii; CorrectBox_P1) (Sec. 2.2–2.4).

Recommendation: Format variable/file names consistently (inline code or a glossary/table) and keep prose focused on conceptual meaning, with technical identifiers relegated to Methods tables or parentheses.

Mathematical consistency audit

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

Maths relevance: light

The paper contains no extended derivations; the mathematical content is limited to (i) algorithmic definitions of behavioral metrics (counts, latencies, ratios) and (ii) linear regression model specifications with FDR correction plus a residual-based 'cognitive resilience' score. Internal consistency is generally acceptable, but several metric definitions are mathematically incomplete in edge cases (undefined denominators/undefined latencies), and one Results claim incorrectly treats residual computation as impossible absent a significant age effect.

Checked items

1. **Perseverative Error Count definition** (Sec. 2.2.3, p.3)

- **Claim:** Perseverative Error Count equals the number of Phase 2 visits to the Phase 1 correct box occurring before the first Phase 2 visit to the Phase 2 correct box.
- **Checks:** definition consistency, edge-case well-posedness
- **Verdict:** UNCERTAIN; confidence: high; impact: moderate
- **Assumptions/inputs:** Phase 1 and Phase 2 correct boxes are well-defined for each bat, A 'visit' event stream exists in Phase 2 with ordered timestamps
- **Notes:** The definition is coherent if the bat eventually visits the Phase 2 correct box. If the bat never visits `\text{CorrectBox_P2}` in Phase 2, 'prior to the first visit' is undefined. The paper does not specify whether to set the count to all visits to `\text{CorrectBox_P1}`, mark as NA, or use a censoring rule.

2. **LatencyToFirstCorrect definition** (Sec. 2.2.3, p.3)

- **Claim:** LatencyToFirstCorrect for a phase is the `\text{Absolute_Time}` of the first successful visit to the correct box.
- **Checks:** units/dimensional sanity, edge-case well-posedness
- **Verdict:** UNCERTAIN; confidence: medium; impact: moderate

- **Assumptions/inputs:** $\text{\text{Absolute_Time}}$ is comparable within a phase and has a meaningful zero/reference point, There exists at least one correct-box visit in the phase
- **Notes:** Using a timestamp as a latency is fine only if time is measured from phase start; if $\text{\text{Absolute_Time}}$ is global-clock time, then differences can still work but the term 'latency' is imprecise. Also undefined if a bat never visits the correct box in that phase; no rule is given.

3. ✓ **Switch Cost (Phase 2) as latency difference** (Sec. 2.2.3, p.3)

- **Claim:** $\text{Switch Cost (Phase 2)} = \text{LatencyToFirstCorrect}\{P2\} - \text{LatencyToFirstCorrect}$.
- **Checks:** algebra, units/dimensional sanity, definition consistency
- **Verdict:** PASS; confidence: medium; impact: minor
- **Assumptions/inputs:** Latencies are measured in the same units and are comparable across phases, Both latencies are defined (bat finds the correct box in both phases)
- **Notes:** Subtracting two times yields a time difference (still in time units), consistent with 'switch cost' as additional latency. Comparability depends on $\text{LatencyToFirstCorrect}$ being a phase-relative measure; otherwise interpretation is weaker but algebra is consistent.

4. ✓ **Switch Cost (Phase 3) as latency difference** (Sec. 2.2.3, p.3)

- **Claim:** $\text{Switch Cost (Phase 3)} = \text{LatencyToFirstCorrect}\{P3\} - \text{LatencyToFirstCorrect}$.
- **Checks:** algebra, units/dimensional sanity
- **Verdict:** PASS; confidence: medium; impact: minor
- **Assumptions/inputs:** Latencies are comparable across phases, Both latencies are defined
- **Notes:** Same comments as Switch Cost Phase 2; algebraically consistent.

5. △ **Lose-Shift Index definition** (Sec. 2.2.3, p.3)

- **Claim:** $\text{Lose-Shift Index} = (\text{number of shifts immediately following a lose event}) / (\text{number of lose events})$.
- **Checks:** normalization/constraints, edge-case well-posedness
- **Verdict:** UNCERTAIN; confidence: high; impact: moderate
- **Assumptions/inputs:** A lose event is any visit to an incorrect box, A shift event is defined as the next visit being a different box
- **Notes:** As a conditional proportion, the index should lie in $[0,1]$ when defined. It is undefined if the total number of lose events is 0 (e.g., bat goes directly to the correct box and never visits an incorrect one). The paper does not specify how such cases are handled.

6. ✓ **Exploration Score definition** (Sec. 2.2.3, p.3)

- **Claim:** Exploration Score equals the number of unique incorrect boxes visited within a phase.
- **Checks:** definition consistency, sanity constraints
- **Verdict:** PASS; confidence: high; impact: minor
- **Assumptions/inputs:** A set of incorrect boxes can be determined given the correct box label
- **Notes:** Well-posed and nonnegative; upper-bounded by (number of boxes -1) if the environment has a fixed set of boxes.

7. △ **Exploitation Score definition** (Sec. 2.2.3, p.3)

- **Claim:** Exploitation Score equals the number of visits to the correct box after the first successful visit to that box.
- **Checks:** definition consistency, edge-case well-posedness
- **Verdict:** UNCERTAIN; confidence: high; impact: moderate
- **Assumptions/inputs:** A first successful correct visit exists in the phase
- **Notes:** If the bat never visits the correct box, the 'after the first successful visit' set is undefined; if the bat visits it exactly once, the exploitation score is 0, which then affects the ratio metric. Handling rules are not specified.

8. △ **Exploration–Exploitation Ratio definition** (Sec. 2.2.3, p.3)

- **Claim:** Exploration–Exploitation Ratio = Exploration Score / Exploitation Score.
- **Checks:** algebra, edge-case well-posedness
- **Verdict:** UNCERTAIN; confidence: high; impact: moderate
- **Assumptions/inputs:** Exploitation Score > 0
- **Notes:** Algebra is trivial, but the ratio is undefined when Exploitation Score = 0 (e.g., only one or zero correct visits). The paper does not specify safeguards (e.g., add-one smoothing, defining ratio as $+\infty$, or NA).

9. ✓ **DTI age-effect regression model specification** (Sec. 2.4.2, p.4)

- **Claim:** For each ROI and each DTI metric, fit a linear regression $\text{DTI_Metric} \sim \text{Age} + \text{Sex}$.
- **Checks:** notation consistency, model-form consistency
- **Verdict:** PASS; confidence: high; impact: minor
- **Assumptions/inputs:** DTI_Metric refers to the extracted mean within a given ROI for a given scalar (FA/MD/AD/RD), Age is in years and Sex is a categorical covariate
- **Notes:** The formula is internally consistent with the stated goal and with later mentions of controlling for sex.

10. ✓ **FDR correction scope for ROI-wise tests** (Sec. 2.4.2, p.4)

- **Claim:** Apply Benjamini–Hochberg FDR correction for each DTI metric across all ROIs.
- **Checks:** definition consistency
- **Verdict:** PASS; confidence: medium; impact: minor
- **Assumptions/inputs:** A family of p -values exists per metric across 24 ROIs
- **Notes:** The stated correction scope is coherent (per-metric across ROIs). The paper does not specify whether it also accounts for the four metrics jointly; that is a design choice, not an internal inconsistency.

11. ✓ **Behavioral age-effect regression model specification** (Sec. 2.4.3, p.4)

- **Claim:** For each behavioral metric, fit $\text{Behavioral_Metric} \sim \text{Age} + \text{Sex}$.
- **Checks:** notation consistency
- **Verdict:** PASS; confidence: high; impact: minor
- **Assumptions/inputs:** Behavioral metrics are scalar outputs per bat (per phase where relevant)
- **Notes:** Consistent with the DTI modeling approach and with the later Results narrative.

12. ✓ **Cognitive Resilience Score as regression residual** (Sec. 2.4.4, p.4)

- **Claim:** Compute $\text{Cognitive_Resilience_Score}$ as residuals from $\text{Behavioral_Metric} \sim \text{Age}$ (age-only).
- **Checks:** definition consistency, model-spec alignment
- **Verdict:** PASS; confidence: high; impact: moderate
- **Assumptions/inputs:** A linear model is fitted with an intercept unless otherwise stated
- **Notes:** Residuals from an age-only regression are well-defined and do remove the fitted linear age component. However, because Sex is omitted here while used elsewhere, the residuals are not adjusted for sex; the claim of isolating variance 'independent of age' holds, but comparability across sexes may be affected.

13. ✓ **Brain-resilience regression model specification** (Sec. 2.4.4, p.4)

- **Claim:** For significant ROI–DTI pairs, regress $\text{Cognitive_Resilience_Score} \sim \text{ROI_DTI_Metric} + \text{Sex}$.
- **Checks:** notation consistency, model-form consistency
- **Verdict:** PASS; confidence: high; impact: minor

- **Assumptions/inputs:** ROI_DTI_Metric denotes a scalar DTI value in a specific ROI
 - **Notes:** Consistent formula syntax; including Sex here partially addresses sex differences, although residuals were not sex-adjusted.
14. ✘ **Claim that resilience score cannot be calculated without significant age decline** (Sec. 3.3, p.8)
- **Claim:** Because no behavioral metric significantly declined with age, the Cognitive Resilience Score 'could not be meaningfully calculated or applied'.
 - **Checks:** logical implication check
 - **Verdict:** FAIL; confidence: high; impact: moderate
 - **Assumptions/inputs:** Resilience is operationalized strictly as deviation from an age-decline trend
 - **Notes:** Residuals from $\text{Behavioral_Metric} \sim \text{Age}$ are mathematically computable regardless of whether the slope differs significantly from 0. What fails is the interpretation of the residuals as 'resilience to age-related decline' if no decline is evidenced. The text conflates 'not meaningful for the intended interpretation' with 'cannot be calculated/applied'.

Limitations

- The provided PDF text contains almost no explicit mathematical derivations or numbered equations; most checks are limited to the well-posedness and internal consistency of metric definitions and regression formulas.
- No explicit formulas are given for DTI scalars (FA/MD/AD/RD) or for the Dice Similarity Coefficient; therefore their mathematical correctness cannot be audited from the document.
- Figures referenced (e.g., heatmaps/barplots) are descriptive and do not add verifiable analytic derivation steps beyond the stated model forms.

Numerical results audit

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

8 candidate numeric checks were assessed: 6 PASS (exact-sum or basic plausibility/consistency checks), 1 FAIL (Methods vs Results maximum age mismatch), and 1 UNCERTAIN (p -value bound claim not verifiable from available text inputs).

Checked items

1. ✓ **C1_subjects_sex_sum_methods** (Page 2, Section 2.1 Subjects)
 - **Claim:** “The cohort comprised 31 Egyptian fruit bats... included 18 males and 13 females.”

- **Checks:** parts_vs_total
 - **Verdict:** PASS
 - **Notes:** $18 + 13 = 31$.
2. ✓ **C2_subjects_age_mean_sd_range_consistency** (Page 2, Section 2.1 Subjects)
- **Claim:** “ages ranging from 6.62 to 15.07 years (mean age = 9.81 ± 1.83 years)”
 - **Checks:** range_mean_plausibility_and_sd_nonnegative
 - **Verdict:** PASS
 - **Notes:** Mean is within [min, max] and SD is nonnegative.
3. ✓ **C3_dti_volumes_total** (Page 3, Section 2.3.1 DTI Metric Map Calculation)
- **Claim:** “DTI scans... provided as 33-volume NIfTI files, comprising 3 non-diffusion-weighted ($b = 0$) volumes and 30 diffusion-weighted volumes.”
 - **Checks:** parts_vs_total
 - **Verdict:** PASS
 - **Notes:** $3 + 30 = 33$.
4. ✓ **C4_rois_count_consistency** (Page 1 Abstract; Page 3 Section 2.3.2; Page 6 Section 3.2)
- **Claim:** The paper repeatedly states there are “24 predefined regions/ROIs” used for ROI-based DTI extraction/analysis.
 - **Checks:** repeated_constant_consistency
 - **Verdict:** PASS
 - **Notes:** All extracted ROI counts equal 24.
5. ✓ **C5_subject_funnel_reduction** (Page 4 Results; Figure 2 caption text in parsed content)
- **Claim:** “The initial cohort of 41 subjects was reduced to 33 for the final DTI analysis.”
 - **Checks:** difference_check
 - **Verdict:** PASS
 - **Notes:** Excluded subjects computed as $41 - 33 = 8$.
6. ✓ **C6_final_dti_sex_sum** (Page 4 Results paragraph; Figure 3 caption text in parsed content)
- **Claim:** For the final DTI cohort: “the cohort comprised 21 males and 12 females” and “final study cohort of 33”.
 - **Checks:** parts_vs_total
 - **Verdict:** PASS
 - **Notes:** $21 + 12 = 33$.

7. ✖ **C7_age_range_inconsistency_methods_vs_results** (Page 2 Section 2.1 vs Page 4 Results and Figure 3 caption)
- **Claim:** Methods state ages 6.62–15.07 years (mean 9.81 ± 1.83) for 31 bats; Results describe a 33-bat DTI cohort with ages 6.6–13.8 years and elsewhere mention “maximum age of approximately 15 years” in sample.
 - **Checks:** cross_section_numeric_consistency
 - **Verdict:** FAIL
 - **Notes:** Minimum ages agree within tolerance (6.62 vs 6.6), but maximum ages do not (15.07 vs 13.8). Narrative “approximately 15 years” is closer to 15.07 than to 13.8.
8. ⚠ **C8_behavioral_pvalues_threshold_statement** (Page 5, Section 3.1)
- **Claim:** “all p -values for age effect > 0.17 ”
 - **Checks:** inequality_internal_logic
 - **Verdict:** UNCERTAIN
 - **Notes:** Not verifiable from the provided inputs because no explicit list/table of the relevant p -values is available to check for contradictions.

Limitations

- Only numeric statements explicitly present in the provided PDF text were used; no external datasets or knowledge were consulted.
- Values embedded only in plots/figures without printed numeric labels cannot be verified without pixel-based extraction, which is out of scope.
- Several statistical claims (e.g., FDR-corrected significance, Dice scores for all subjects) are not accompanied by explicit numeric tables; these cannot be recomputed or verified as FAST checks from the PDF alone.
- One inequality-based claim about p -values (Page 5, Section 3.1: “all p -values for age effect > 0.17 ”) could not be validated against any enumerated p -values from the available inputs.