Supplementary Information

A new impact for an AI-mediated science

Fabio Favoretto

Author. Fabio Favoretto, School of Biological and Marine Sciences, University of Plymouth, Plymouth, UK. Correspondence: fabio.favoretto@plymouth.ac.uk

This document accompanies the main Analysis article "A new impact for an AI-mediated science." It contains the formal definition of the AI Impact Index (AIII), the question and corpus protocols, the language-model prompting and memorization-probe procedures, the benchmark panel against citation-family metrics, the externally-anchored validation tests, the three use-cases (infrastructure funding, the epistemic shift in AI-mediated literature, and the equity audit), the robustness panel, the reference event-logging architecture, and a ten-point limitations and threats-to-validity discussion. Every number reproduced from the main text is regenerated by `python scripts/analysis_v2/run_all.py` *from open data and open code.*

Contents

S1 Formal definition of the AI Impact Index

S1.1 Channels

The AI Impact Index (AIII) is defined at two levels and along three channels. The three channels correspond to the principal mechanisms by which a large language model engages a scientific literature: retrieval (R), generation (G) and memorization (M). Each channel measures an *engagement footprint*; see Boxes 1 and 2 in the main text for the explicit scoping.

S1.2 Paper-level scoring formula

For a paper P in a domain corpus D queried with a set Q of $|Q| = 30$ research questions, the per-channel AIII is:

$$\mathrm{AIII}_c(P) = w_b \cdot \tilde{n}(\mathrm{QB}) + w_t \cdot \tilde{n}(\mathrm{TB}) + w_s \cdot \tilde{n}(\mathrm{SS}) + w_p \cdot \tilde{n}(\mathrm{RWR}), \tag{1}$$

where $c \in \{R, G, M\}$ indexes the channel and the four components are:

- *Question breadth* (QB): the number of distinct queries for which P appears in the top-k, divided by $|Q|$;
- *Task-type breadth* (TB): the number of distinct task types (literature review, methods question, data lookup, conceptual question) for which P surfaces, divided by 4;
- *Semantic strength* (SS): the mean cosine similarity over surfacings of P;
- *Rank-weighted relevance* (RWR): the mean reciprocal rank of P across surfacings.

$\tilde{n}(\cdot)$ denotes within-domain min-max normalisation. The four components are mathematically independent: QB counts distinct questions, TB counts distinct task types, SS scores match quality, and RWR scores rank prominence.

S1.3 Default weights and sensitivity analysis

Default weights are $(w_b, w_t, w_s, w_p) = (0.35, 0.15, 0.20, 0.30)$. Ranking stability under weight perturbation is tested by drawing 200 four-component weight vectors from a Dirichlet distribution centred on the defaults and recomputing AIII per draw. Mean Kendall τ over draws is 0.96 (min 0.82), confirming that rankings are stable under modest perturbation of the weights. Figure S1 reports the full sensitivity panel.

S1.4 Package-level AIII

At the package level, $\mathrm{AIII}_{\mathrm{pkg}}$ combines four components: (i) retrieval frequency of the package's origin paper through R-AIII, (ii) ecosystem embedding (GitHub stars + CRAN reverse-dependencies),

(iii) functional usage (CRAN/PyPI downloads) and (iv) the magnitude of the citation–usage gap, with default weights $(0.35, 0.20, 0.25, 0.20)$. Crucially, AIII_{pkg} uses *adaptive weighting*: when a component is not measurable for a given package because it has no origin paper, no CRAN listing or no public GitHub mirror, the weight redistributes proportionally over the remaining measurable components. This allows all 86 registry entries to be scored, not only the 29 with full coverage.

S2 Corpus and question protocol

S2.1 Registry construction

The registry contains 86 software and data packages selected by domain-expert curation as operationally central to marine ecology (24 packages), biomedical genomics (29), and climate and Earth science (28). Each package was classified by citation provenance: *full origin paper* (28 entries), *partial citation path* via a tutorial or methods reference (3 entries), *cite-only* (25 entries) and *paperless* (29 entries). The class distribution per domain is in Table S1.

Table S1: Registry coverage by citation provenance and domain.

Domain	Full	Partial	Cite-only	Paperless	Total
Marine ecology	14	1	5	9	24
Biomedical genomics	9	2	15	3	29
Climate / Earth science	6	0	5	17	28
Total	**29**	**3**	**25**	**29**	**86**

S2.2 Paper corpus

A paper-level corpus of 1,350 papers was assembled as a stratified sample drawn from OpenAlex, with 450 papers per domain. For each domain the sample was stratified by paper type (empirical, methods, data, review) and by year-of-publication band to approximate the composition of the underlying field literature, then inspected for metadata quality. Abstracts were reconstructed from the OpenAlex inverted index and embedded with `sentence-transformers/all-MiniLM-L6-v2`. The corpus is persisted in a ChromaDB vector store under cosine similarity. Full text was not used.

S2.3 Question set

A set of 90 research questions (30 per domain) was assembled by adapting representative tasks from the published literature and verified by domain experts as the kind of questions a working researcher might bring to an AI assistant. Each question is paired with a task-type label to enable per-task breakdowns. The question set is released with the code.

S3 Generation and memorization channels

S3.1 Generation channel (G-AIII)

For each of the 90 questions, Claude Sonnet 4.5 was prompted to write a literature-style answer of 150–200 words from training memory only, with no retrieval augmentation. The prompt explicitly instructed the model to integrate real methods, authors and datasets into prose without producing fabricated citations or specific numerical claims. Each answer was embedded with the same sentence-transformer used for the corpus, and G-channel similarity ranks corpus papers by cosine similarity to the answer (not to the query). G-AIII applies the four-component formula to the resulting similarity profile across the 90 questions. Prompt templates, raw responses and embeddings are released with the code.

S3.2 Memorization channel (M-AIII)

A stratified sample of 150 papers (50 per domain) was drawn from the corpus and probed by asking Claude Sonnet 4.5 to summarise each paper's main findings from training memory only. Each paper was probed three times with distinct prompt templates to test reproducibility. A conservative *UNKNOWN* rule allowed the model to decline rather than confabulate; UNKNOWN responses score 0. Of 150 probed papers, 50 (33%) were recognised with specific content. Memorization strength is the cosine similarity between the model's from-memory summary and the paper's actual abstract. M-AIII is computed only over the recognised subset.

S3.3 Cross-channel structure

The three channels have only moderate cross-channel correlations on shared-paper subsets (Table S2). R and G correlate at Spearman $\rho = 0.45$–0.54 across domains; R–M and G–M correlations are near zero on samples of $n = 9$–18. The channels are not redundant.

Table S2: Cross-channel agreement (Spearman ρ on shared-paper subsets).

Domain	Pair	ρ	p-value	n
Marine	R vs G	0.447	9.3×10^{-5}	71
Biomed	R vs G	0.466	9.2×10^{-8}	119
Biomed	R vs M	0.080	0.79	14
Biomed	G vs M	−0.072	0.77	19
Climate	R vs G	0.544	7.4×10^{-8}	85
Climate	R vs M	−0.148	0.57	17

S3.4 Per-channel paper-type bias

Each channel surfaces papers with a distinct paper-type composition (Table S3). R-AIII favours method papers over empirical papers in all three domains and disfavours reviews; G-AIII shows a similar attenuated method bias; M-AIII has the most variable pattern with an extreme review deficit in biomedicine and a strong method bias in climate.

Table S3: Per-channel paper-type bias: mean within-domain AIII difference between method and empirical papers, and between review and empirical papers.

Channel	Domain	Method − Empirical	Review − Empirical
R-AIII	Biomed	+0.070	−0.050
R-AIII	Climate	+0.018	−0.052
R-AIII	Marine	+0.077	−0.013
G-AIII	Biomed	+0.067	−0.052
G-AIII	Climate	+0.053	+0.007
G-AIII	Marine	+0.060	−0.019
M-AIII	Biomed	−0.048	−0.419
M-AIII	Climate	+0.378	—
M-AIII	Marine	+0.017	−0.050

S4 Benchmark panel and validation

S4.1 Four-test benchmark panel

AIII was compared against four candidate metric families currently used in research evaluation: total citations, citations-per-year, log-transformed citations and a transparent Altmetric-style attention proxy. The benchmark is structured as a four-test panel:

- **T1 Infrastructure coverage.** Fraction of the 86-package registry that the metric can score with a discriminative (non-zero) value.
- **T2 Paper-type balance.** 1 minus the Jensen–Shannon divergence between the metric's top-50 paper-type distribution and the corpus base rate (higher = more balanced).
- **T3 Informativeness beyond citations.** Mean absolute partial Spearman ρ between the metric and a within-corpus journal-prestige proxy, controlling for total citations.
- **T4 Channel independence.** Hold-one-out Spearman ρ predicting each AI channel from the others vs from total citations (Table S5).

The full panel of six metrics on T1–T3 is in Table S4.

Table S4: Four-test benchmark panel. Higher values indicate better performance in every column. Multi-AIII wins T1 decisively (it is the only metric that can score paperless infrastructure); citations win T4-style consensus tests by construction; T2 and T3 separate metrics weakly.

Metric	T1 Coverage	T2 Balance	T3 Informativeness
Total citations	0.651	0.994	0.000
Altmetric proxy	0.663	0.994	0.212
R-AIII	0.093	0.977	0.095
G-AIII	0.000	0.980	0.096
M-AIII	0.000	0.982	0.154
Multi-AIII	**0.779**	0.978	0.059

S4.2 Hold-one-channel-out cross-validation

The hold-one-channel-out test treats each AI channel as a pseudo-external target, predicted by the remaining two AI channels and, in parallel, by total citations as a baseline (Table S5). The two retained AI channels predict the held-out channel at Spearman $\rho = 0.43$ (G held out) to $\rho = 0.49$ (R held out), with 95% bootstrap confidence intervals that exclude zero. Total citations predict no channel ($\rho = 0.02$–0.08, CIs include zero). This is the central external-validation result of the framework: AI engagement is internally coherent across mechanisms, and the citation system does not reproduce it.

Table S5: Hold-one-channel-out validation (Spearman ρ with 95% bootstrap CI).

Held-out channel	Predictor	ρ	95% CI	n
R-AIII	G + M channels	$+0.49$	$[+0.39, +0.56]$	409
R-AIII	total citations	$+0.08$	$[-0.02, +0.18]$	409
G-AIII	R + M channels	$+0.43$	$[+0.34, +0.52]$	354
G-AIII	total citations	$+0.02$	$[-0.09, +0.13]$	354
M-AIII	R + G channels	-0.01	$[-0.29, +0.24]$	50
M-AIII	total citations	-0.28	$[-0.52, -0.00]$	50

S4.3 Package-level external validation

For 63 of 86 registry packages, an External Usage Index (EUI) can be constructed from three signals independent of any AI channel: CRAN/PyPI downloads, GitHub stars, and CRAN reverse-dependencies. The package-level external validation correlates each candidate paper-level metric against the EUI (Table S6). The result is a null finding for every paper-level metric, citations and AI channels alike. The honest interpretation is that paper-level AI engagement and package-level operational usage are different constructs, and the framework treats them as such throughout.

Table S6: Package-level external usage validation (Spearman ρ with 95% bootstrap CI; $n = 63$ packages).

Metric	Independent of EUI?	ρ vs EUI	95% CI
Total citations	yes	-0.20	$[-0.46, +0.09]$
Altmetric-style proxy	yes	-0.20	$[-0.46, +0.09]$
R-AIII (retrieval only)	yes	-0.35	$[-0.50, -0.15]$
AIII_{pkg} (full)	**no** (uses downloads)	$+0.33$	$[+0.06, +0.57]$

S5 Use cases

S5.1 Use case 1 — Infrastructure funding priorities

A central policy question for research funders is which scientific software and datasets should be supported as infrastructure rather than as project-grant residuals. The citation record answers this badly. Applied to the full 86-entry registry, AIII_{pkg} surfaces a top-ranked set per domain that is dominated by entries the citation system cannot see (paperless) or sees only as a dataset DOI with no usage signal (Table S7).

Table S7: Use case 1: top AIII_{pkg} entries per domain. Paperless entries (marked) are invisible to any citation-based metric.

Domain	Name	Coverage	Components	AIII_{pkg}	Signal
Biomed	**pandas**	paperless	2/4	0.79	7.8B dl/yr; 48.6k stars; no paper
Biomed	numpy	complete	3/4	0.76	9.2B dl/yr; 31.6k stars
Biomed	1000 Genomes	cite-only	1/4	0.71	dataset, no usage index
Biomed	**scikit-learn**	paperless	2/4	0.59	2.4B dl/yr; 65.9k stars; no paper
Climate	**xarray**	paperless	2/4	1.00	239M dl/yr; 4.1k stars; no paper
Climate	CMIP6	cite-only	1/4	1.00	dataset, no usage index
Climate	wrf-python	complete	3/4	0.34	158k dl/yr; 0.5k stars
Climate	**cartopy**	paperless	2/4	0.20	16.5M dl/yr; 1.6k stars; no paper
Marine	sf	complete	3/4	0.96	73M dl/yr; 1.4k stars
Marine	**terra**	paperless	2/4	0.39	1.9M dl/yr; 0.6k stars; no paper
Marine	lme4	complete	3/4	0.36	33M dl/yr; 0.7k stars

S5.2 Use case 2 — The epistemic shift in AI-mediated literature

A researcher whose view of a literature is shaped by AI retrieval is exposed to a systematically different evidence base than one shaped by citation ranking. Comparing the top-10 R-channel retrievals (union across the 90 questions) with a matched citation-ranked set produces three effects consistent across all three domains (Table S8). The method-bias effect is significant under a Mann–Whitney U test of method versus empirical papers across domains ($p = 6.1 \times 10^{-4}$).

Table S8: Use case 2: paper-type composition and recency of the AI-retrieved top-50 versus the citation top-50, by domain.

	Biomedical		Climate		Marine	
	AI top-50	Cite top-50	AI top-50	Cite top-50	AI top-50	Cite top-50
% methods	38.4	37.7	30.9	17.9	24.8	15.5
% reviews	5.0	0.6	4.1	1.6	12.4	14.0
Median year	2015	2006	2011	1999	2014	2009

S5.3 Use case 3 — Equity audit

AIII is a different aggregation of the same underlying literature. Comparing its top-50 to the citation top-50 tests whether AI retrieval reproduces, attenuates or reshapes the venue and open-access concentrations the citation-based system is known to carry. The signal is heterogeneous and domain-specific (Table S9).

Table S9: Use case 3: equity audit of top-50 AI-retrieved vs top-50 cited papers by domain.

Domain	% Open access (AI / cite)	Unique journals (AI / cite)	Journal HHI (AI / cite)	Median year (A
Biomedical	90 / 86	25 / 19	0.076 / 0.101	201
Climate	74 / 38	31 / 38	0.044 / 0.034	201
Marine	52 / 56	37 / 41	0.036 / 0.026	201

S6 Robustness panel

Ranking sensitivity was tested against six perturbations:

- **Dirichlet weight perturbation.** 200 draws from a Dirichlet centred on the default weights; mean Kendall $\tau = 0.96$ (min 0.82).
- **Retrieval depth k.** Comparing $k = 5$ vs $k = 10$: per-domain Kendall $\tau = 0.87$–0.93.
- **Query paraphrasing.** Each of the 90 queries was deterministically paraphrased; ranking stability mean Kendall $\tau = 0.84$.
- **Cross-encoder replication.** R-channel was recomputed with `all-mpnet-base-v2`; Jaccard set-overlap between the encoders ranges 0.40–0.70 across domains, with the largest divergence in the marine corpus.
- **Scoring formula variant.** Replacing rank-weighted relevance with raw rank; ranking $\tau \geq 0.91$.

- **Bootstrap resampling.** 1,000 bootstrap iterations on shared-paper subsets; 95% CIs on all reported Spearman ρ values.

Figure S1 shows the weight-sensitivity and k-sensitivity panels. Figure S2 shows the cross-encoder replication results.

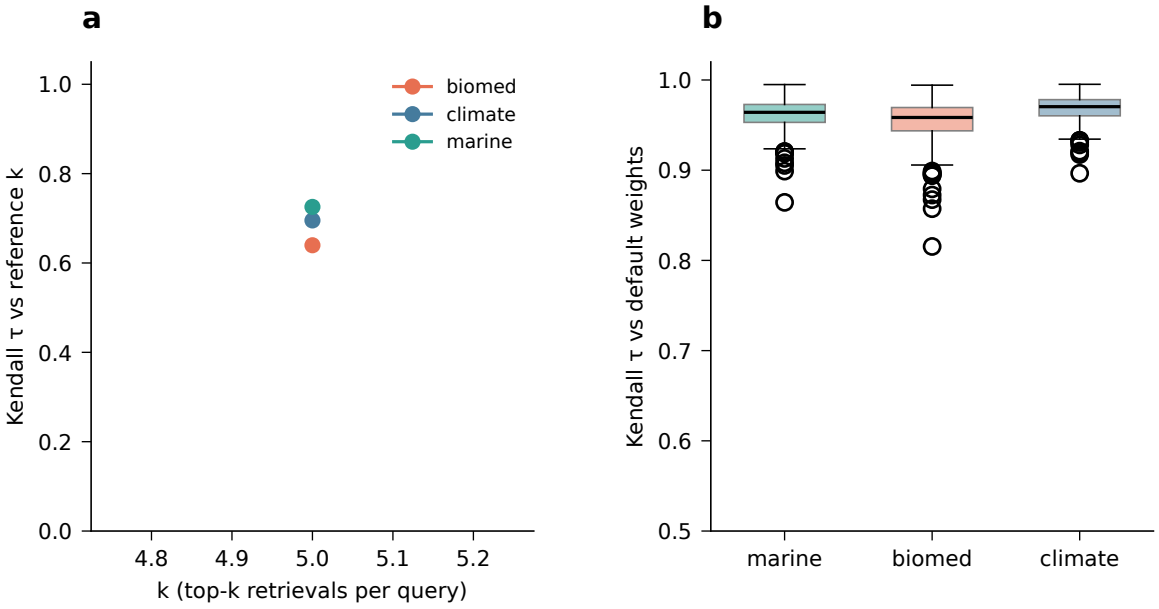

Figure S1: Robustness Figure S1: ranking sensitivity. Left: Dirichlet weight perturbation, showing the distribution of Kendall τ over 200 draws relative to the default weight ranking. Right: k-sensitivity ($k = 5$ vs $k = 10$).

S7 Reference event-logging architecture

The components of an AI-engagement measurement system already exist: DataCite identifies software and data, Crossref Event Data records non-citation events, and ORCID binds contributions to researchers. A reference architecture for adding AI-engagement event types to that stack is sketched below.

1. **Retrieval events.** Logged at the vector-store layer of an AI-assisted scientific search system. The event payload contains the query, the retrieved paper identifier (DOI or DataCite), the retrieval similarity score, the rank in the top-k, and a system-side timestamp. Privacy-preserving aggregation through k-anonymous bucketing prevents per-user tracking.

2. **Generation events.** Logged at the LLM gateway after answer generation. The event payload includes a hash of the generated answer, the corpus papers whose abstracts pass a similarity threshold against the answer, and the LLM identifier and version.

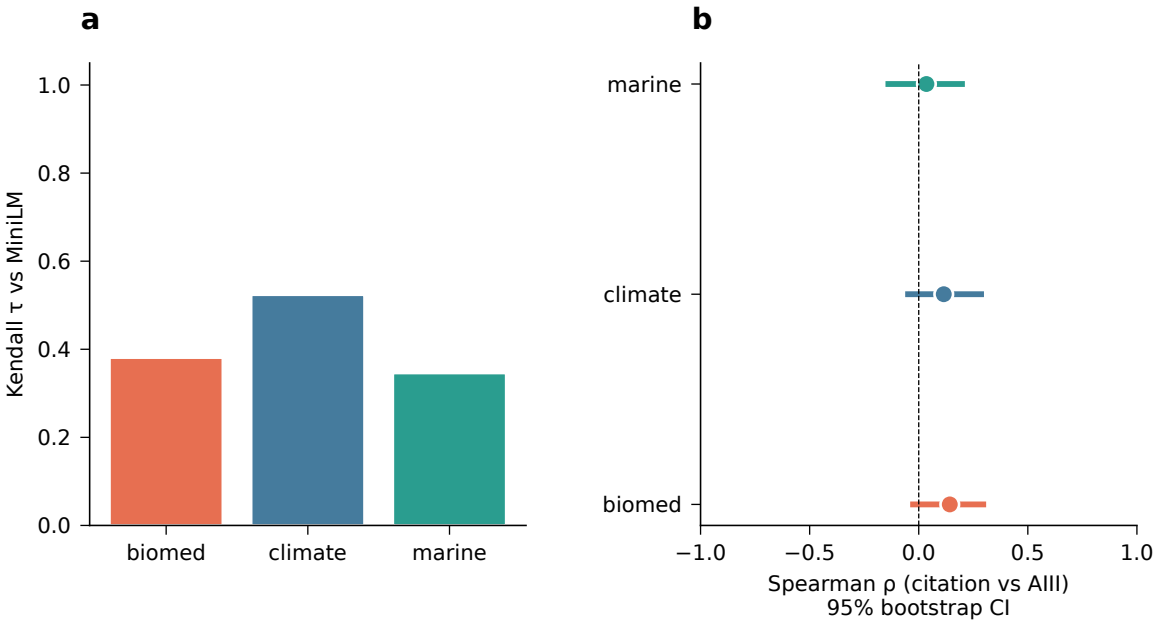

Figure S2: Robustness Figure S2: cross-encoder replication of R-AIII with `all-mpnet-base-v2`. Per-domain Jaccard set-overlap and Kendall τ on the top retrieval sets.

3. **Memorization probes.** Run periodically as an audit, not at user-request time, with a fixed probe set per LLM release. Probe results are stored as memorization-strength scores per paper per model version.

4. **Event ingestion.** A common event schema feeds into Crossref Event Data. Each event is tagged by channel (R, G or M), source (AI system identifier), and timestamp.

5. **Aggregation and reporting.** Per-paper and per-package AIII scores are computed from the event log on a fixed cadence, with bootstrap CIs.

The architecture is a design proposal supported by a reference implementation in the project repository. It is not a deployed system; production-scale telemetry, operational integrations and user evaluations are needed before any deployment claim can be made.

S8 Limitations and threats to validity

L1 – The internal coherence check is not external validation. An earlier draft framed papers surfaced by ≥ 2 of {R, G, M} channels as a "ground truth" set and reported AUC ≈ 0.97 for Multi-AIII against it. Peer review correctly identified this as a self-consistency test. The analysis has been removed from the main argument; the externally-anchored hold-one-channel-out (Table S5) and package-level (Table S6) validations carry the load instead.

L2 – G-AIII measures alignment, not causal use. Cosine similarity between an LLM answer and a paper abstract does not establish that the model used the paper. G-AIII captures field-level semantic overlap, common terminology and distributional knowledge, not traceable influence. The manuscript uses "AI engagement" rather than "AI use" where the distinction matters.

L3 – M-AIII is underpowered. Only 50 of 150 probed papers were recognised by the model with specific content (33%). Within-domain shared-paper subsets for cross-channel correlations involving M-AIII have $n = 9$–18. M-AIII findings are exploratory; the main-text and supplementary results that lean least on M-AIII are the most robust.

L4 – Benchmark T3 was redesigned after peer review. An earlier "novelty over citations" test rewarded any divergence from citations regardless of meaning. T3 has been replaced by an informativeness-beyond-citations test (partial Spearman ρ against a journal-prestige proxy controlling for total citations). The withdrawn AUC = 0.97 claim from v3 is not reintroduced.

L5 – Construct conflation. AI retrievability, semantic alignment, memorisation likelihood, package-level operational usage, and scientific importance are five distinct constructs. AIII measures the first three; the package-level EUI measures the fourth; none speaks directly to the fifth. Boxes 1 and 2 in the main text make this explicit.

L6 – Single embedding family, single LLM, single corpus, single retrieval paradigm. All channels use `all-MiniLM-L6-v2` for embedding; the R-channel robustness check uses `all-mpnet-base-v2`. G-AIII and M-AIII use Claude Sonnet 4.5. Multi-model and multi-corpus replication is the natural next study and is not in this version of the work.

L7 – Researcher-dependent normalisation and weighting. AIII uses min-max normalisation within domain, fixed defaults for component and channel weights, and adaptive redistribution for the package-level index. Dirichlet sensitivity analysis confirms ranking stability under modest perturbation (mean Kendall $\tau = 0.96$), but the framework is closer to a bespoke composite index than to a theoretically derived metric. A learned weighting against an external AI-usage criterion (when such data becomes available) is a natural future-work direction.

L8 – No deployment evidence. The pilot architecture in §S7 is a design proposal supported by a reference implementation, not a deployed system. Production telemetry, operational integrations and user evaluations are needed before any deployment claim can be made.

L9 – Marine corpus composition. OpenAlex concept-ID sampling drew the marine corpus heavily from agricultural and environmental science rather than pure marine ecology, visible as 80% UNKNOWN in the marine high-cited memorisation stratum versus 0% in biomedical and 30% in climate. The structural multi-channel findings hold across all three domains regardless of this corpus issue.

L10 – Memorization probing is conservative. The UNKNOWN rule instructed the model to decline rather than confabulate. The resulting 33% recognition rate is a lower bound, not an unbiased

estimate of memorization in the underlying model.

S9 Reproducibility

The pipeline is implemented in Python 3.11 with `pandas`, `scikit-learn`, `sentence-transformers`, `ChromaDB` and NumPy. Language-model caches ship with the code, so re-runs do not require an API key. A single command,

```
python scripts/analysis_v2/run_all.py
```

regenerates every figure, table and number in the main text and in this Supplementary Information from raw inputs. The repository is at https://github.com/Fabbiologia/ai-impact-index and is licensed MIT. The repository is archived on Zenodo at acceptance.

Supplementary tables

The Supplementary Information references the following Supplementary Tables, available as CSV files in the project repository:

- **Table S1–S3.** Per-paper R-, G-, M-channel AIII with all four components and benchmark metrics.
- **Table S4.** Multi-channel paper AIII with channel-coverage flags.
- **Table S5.** Per-channel paper-type bias (Table S3 expanded with p-values and sample sizes).
- **Table S6.** Per-channel corpus coverage and top-50 composition.
- **Table S7.** Cross-channel Spearman ρ on shared-paper subsets (Table S2 expanded).
- **Tables S8–S9.** Package-level AIII on the full 86-entry registry and on the *complete*-coverage subset.
- **Table S10.** Weight sensitivity (200 Dirichlet draws).
- **Tables S11–S13.** Benchmark correlations and partial-correlation analyses.
- **Tables S14–S17.** Robustness panel (k-sensitivity, paraphrase, cross-encoder, bootstrap).
- **Tables S18–S20.** Use cases 1, 2 and 3 in full.
- **Tables S21–S26.** Auxiliary tables on retrieval selection, coverage summary, metric comparison and historical ground-truth validation curves (retained for completeness; see L1 for the framing change).