STIMULUS OVERLAP CHANGES MODEL COMPARISONS IN CROSS-SUBJECT BRAIN DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-subject brain decoding holds out the listener, not the story: on MEG-MASC every listener hears the same four stories, so the test story reaches the model twice, through the other listeners in pretraining and, under the standard split, through calibration. A subject-only split does not check this overlap, and the tested model comparison is sensitive to it. Under the standard protocol a dynamic encoder (no residual connections; its pooled feature adds the latent's start-to-end displacement) beats a matched static convolutional one at low budgets (criterion pre-specified in $\times$chance at $k{=}100$; in rank percentile $+0.043$ and $+0.014$, $n{=}10$ paired). With story-disjoint calibration (subject split and nominal pretraining volume fixed) the advantage is undetectable; withholding the story from pretraining too, no narrative–budget combination resolves in its favour; at $k{=}2000$ the static encoder leads under the standard protocol and with the story withheld. This shows split-sensitivity at the tested sample size, not equal architectures. Transfer survives: withholding the story from pretraining at matched story count (not segment count) leaves about 54% of above-chance signal ($k{\geq}500$; three narratives, three seeds). Returning 59% of a narrative's training blocks to a reproduced reference pipeline at matched volume prices the exposure: five exposed runs separate completely from three unexposed ($+1.01$ to $+4.59$ top-10 points), and the full dose gives $+3.00$, same-signed on a second narrative ($+2.00$), each from a single full-dose run. Still open: whether split or adjacency guard does more, how much of the fully corrected cell (which also withholds a pretraining story) is the split, and how far unequal segment counts move the halved signal. Repeated stories are legitimate for some targets; for claims about unheard stimuli, hold the story out of pretraining and calibration and report a stimulus-derived baseline. We release two stdlib-only tools that flag this overlap; the main results rest on one MEG corpus.

1 INTRODUCTION

Brain foundation models, pretrained on pooled recordings (Jiang et al., 2024; Xiao et al., 2025), are credited with transfer to novel subjects (Jayalath et al., 2025). Pretrain an encoder on the MEG recordings of nine listeners, fine-tune it on k segments from a tenth, and score how well it retrieves the language-model embedding of what the listener heard. A higher score reads as better transfer, assuming that holding out the listener holds out what they heard. In MEG-MASC (Gwilliams et al., 2023) it does not: all 27 listeners heard the same four stories, and stimulus-locked responses are correlated across people (Hasson et al., 2004; Chen et al., 2015), so a model can locate the new listener's segment by matching it to the same story moment in other brains—time-point identification (Figure 1), which scores as decoding. The story reaches the model twice, in pretraining and in calibration, so a comparison may rank encoders by how well they exploit an already-heard story.

The overlap is structural: hyperalignment and shared response models require a common, time-locked stimulus (Haxby et al., 2011; Chen et al., 2015), and public naturalistic corpora present each story to many listeners. Stimuli are routinely held out of *training* in encoding models (Kay et al., 2008) and in cross-subject retrieval (Zhang et al., 2026) and in fine-tuning (Jayalath and Parker Jones, 2026) on this corpus, but to our knowledge no foundation-model evaluation protocol checks whether the evaluation stimulus was in the *pretraining* corpus: one MEG foundation model

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

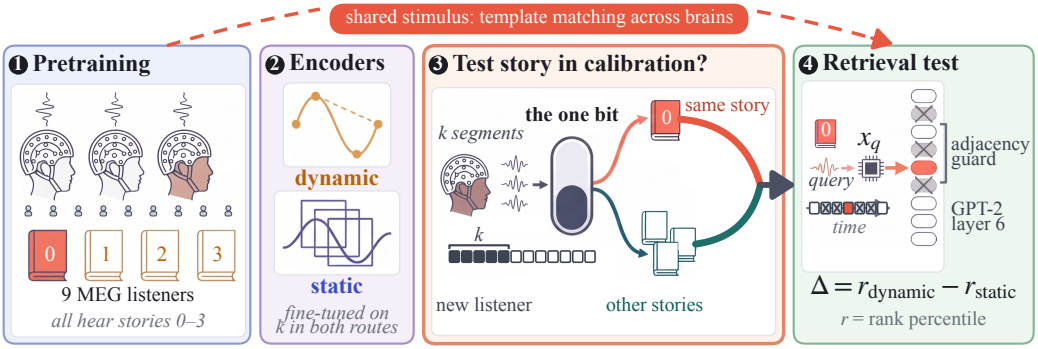

Figure 1: **A subject-level split holds out the listener, but the story can reach the model twice, in pretraining and in calibration; taking it out of calibration alone leaves the dynamic encoder's low-budget advantage undetectable (Table 2).** (1) All nine pretraining MEG listeners hear all four stories, including an evaluation story (coral, drawn as story 0); the dashed arc marks the exposure. (2) The endpoint-augmented *dynamic* encoder (amber; residual-free trunk, not full trajectory dynamics) and the static convolutional encoder (indigo) are pretrained with the same recipe and data, and each is fine-tuned on the k calibration segments in both routes. (3) The one bit: a held-out listener is calibrated on k segments, either from recordings that share stories with the test recordings (coral, the standard recording split) or only from other stories (teal; the story split calibrates on stories 1 and 3 and tests on stories 0 and 2). (4) Test: retrieving the query's head-projected GPT-2 layer-6 target; candidates adjacent to the query are excluded (adjacency guard, a separate factor), r is the rank percentile and the comparison is $\Delta = r_{\text{dynamic}} - r_{\text{static}}$. The schematic idealises the recording-guard cells of Table 2 (four-story pretraining); story sets: §3. No values are plotted.

lists MEG-MASC among its pretraining data (Xiao et al., 2025) and has since been compared on MEG-MASC (Jayalath and Parker Jones, 2026). Leakage is known to inflate scores (Kapoor and Narayanan, 2023; Rosenblatt et al., 2024), and a stimulus-disjoint split collapses brain-to-text generation (Yin et al., 2025); we ask whether the overlap also changes which model wins. If the story helps two encoders unequally, a subject-level split confounds that with architecture (Proposition 1).

We keep the subject split and move only the story. Transfer survives: on a story no brain in pretraining heard, retrieval stays at percentile 0.66–0.71 (chance 0.5), though withholding the story at matched story count removes about half the above-chance signal at $k \geq 500$ (Table 1). The architecture ranking is split-sensitive: a dynamic encoder (residual-free temporal stack, pooled features plus start-to-end displacement) beats a matched static convolutional one at low budgets under the standard protocol but not detectably once the story leaves calibration or pretraining (Tables 2 and 3). The pair is our own; the one published ingredient we audit keeps its advantage under the corrected split, deflated by about a third (Appendix F), so the split can change a comparison's verdict without always doing so. Still open: whether the split or the guard does more, how much of the fully corrected cell (which also withholds a pretraining story) is the split, and how far unmatched volume (UNSEEN arms pretrain on more segments; §4) shifts the halved signal.

Contributions. (1) *An identification result:* a subject-level split measures the architecture difference plus an exposure term it cannot separate (Proposition 1), and a 2×2 guard$\times$split design decomposes the protocol change exactly into two main effects (§3). (2) *A comparison whose verdict depends on the split:* under the standard protocol the dynamic encoder meets its pre-specified criterion, and only there is its advantage resolved at both low budgets (Tables 2 and 3). What changes is the conclusion the comparison reaches, not a confirmed change in the difference between encoders: the paired architecture$\times$split interaction excludes zero only at k=500 and only before correction for multiple comparisons. (3) *A price, and checks for it:* withholding the evaluation story from pretraining retains about half the above-chance signal (Table 1), monotone in exposure at pinned volume; in the reproduced reference pipeline a brain-free word-onset clock beats the decoder, and exposure is worth about half the fix's volume cost (Table 4). We release the corrected MEG-MASC split, the pool-pinning patch and two stdlib-only tools that flag the overlap in a split manifest (failing CI) or in corpus metadata (Appendix J).

2 Related Work

Cross-subject and foundation models for neural recordings. Self-supervised encoders pretrained on pooled recordings are now the default starting point for EEG decoding (Kostas et al., 2021; Yang et al., 2023; Jiang et al., 2024) and are being scaled to intracranial data (Wang et al., 2023; Han et al., 2025); benchmarks compare them across tasks, subjects and recording sites (Wu et al., 2025; Zahorodnii et al., 2025). Independent audits find that tuned supervised baselines often match them (Lee et al., 2025; Wang et al., 2026) and that subject identity can act as a shortcut (Lin et al., 2026). These works hold out a subject or a dataset; none asks whether the evaluation stimulus was in the pretraining corpus, which is the variable we manipulate. Pretraining on other corpora avoids the overlap by construction (Jayalath et al., 2025; Jayalath and Parker Jones, 2026); we measure it when pretraining pools the evaluation corpus.

Decoding speech and language from non-invasive recordings. Contrastive MEG/EEG decoders retrieve the heard speech segment among more than a thousand candidates (Défossez et al., 2023) and have been extended to individual words (d'Ascoli et al., 2025); fMRI decoders reconstruct continuous language (Tang et al., 2023), and within-subject corpora push MEG decoding to tens of hours per participant (Özdogan et al., 2025). The public naturalistic datasets these studies use present each story to many participants—in MEG-MASC and the Brennan EEG corpus, the same stories to all of them (Gwilliams et al., 2023; Nastase et al., 2021; Brennan and Hale, 2019). Zhang et al. (2026) audit retrieval on MEG-MASC with a stimulus-identity split and a stimulus-derived nuisance, and Semenkov et al. (2026) trace retrieval largely to acoustic features and onsets. We reproduce the reference pipeline of Défossez et al. (2023) rather than propose a decoder, and add what these audits do not report: a brain-free model evaluated as a competitor, and the cost of the corrected split decomposed into volume and exposure.

Leakage, circularity and contamination. Neuroimaging has a long record of evaluation artefacts: circular selection of test data (Kriegeskorte et al., 2009), block-design confounds in EEG classification (Li et al., 2021), within-subject stimulus repetition (Kilgallen et al., 2025), and reconstructions manufactured by train–test stimulus overlap (Shirakawa et al., 2025); aggregation efforts now deduplicate stimuli as hygiene (Lahner et al., 2025). Closest to us, Yin et al. (2025) show brain-to-text generation collapsing under a stimulus-disjoint cross-subject split; the pretraining they examine is a task-specific encoder whose leakage they track by validation loss, not a cross-subject foundation model whose ranking they test, and their remedy also discards training data. In machine learning, leakage is a catalogued cause of irreproducible results (Kapoor and Narayanan, 2023), test-set contamination of pretraining corpora changes language-model evaluations (Magar and Schwartz, 2022; Schaeffer et al., 2026), and for time-series foundation models pretraining-corpus overlap has been catalogued, with a controlled test of indirect overlap through correlated series (Meyer et al., 2025). We measure the pretraining-level case for cross-subject decoding at matched volume.

Stimuli as a source of variance. Treating stimuli as fixed when they are a sample inflates confidence, a problem named for psycholinguistics by Clark (1973) and shown for task fMRI (Westfall et al., 2017) and multivoxel classification (Kajimura et al., 2023). Our intervals resample subjects conditional on the stimulus; we therefore report every withheld stimulus separately and do not generalise over stimuli from four narratives (Proposition 1(ii)).

3 Problem Setup

Protocol. MEG-MASC (Gwilliams et al., 2023) has 27 subjects who each hear four narratives $e \in \{0, 1, 2, 3\}$, each presented twice; we use ten of them (one recorded in a single session). For a held-out subject s and an encoder $m \in \{\mathrm{dyn}, \mathrm{stat}\}$, the model is pretrained on the other nine subjects' recordings of a story set P, calibrated on $k \in \{100, 500, 2000\}$ segments of s, and tested on a query set Q_s of held-out segments of s by retrieving each query segment's target t_q, the normalised output of a jointly trained head on its frozen GPT-2 (Radford et al., 2019) layer-6 embedding (App. I), from a candidate pool Π_q. An *adjacency guard* removes from Π_q the candidates within ± 10 s of q in the same recording (*recording guard*) or in any recording of the same story (*stimulus guard*), so retrieval cannot be won by returning temporal neighbours. The *recording split* (standard; like the block split of Défossez et al. (2023), it lets a story reach both sides) tests on half of s's recordings

and calibrates on the other half, which usually share stories with them; the *story split* tests on stories $\{0, 2\}$ and calibrates on $\{1, 3\}$. Arms that withhold a story from pretraining test on a single story (story 0 unless stated) and calibrate only on stories neither tested nor withheld: an UNSEEN arm withholds the tested story itself, a SEEN arm withholds story 3. The factorial's stimulus/story cell is the SEEN arm (pretraining $\{0, 1, 2\}$, calibration $\{1, 2\}$, test 0; Table 2). An interval that excludes zero is called *resolved*.

Score and estimands. With $\hat{z}_q$ the normalised output of f_m on MEG window x_q,

$$r_q = \frac{1}{|\Pi_q| - 1} \sum_{c \in \Pi_q \setminus \{q\}} \mathbf{1}[\langle \hat{z}_q, t_c \rangle < \langle \hat{z}_q, t_q \rangle], \qquad r_s^m = \frac{1}{|Q_s|} \sum_{q \in Q_s} r_q, \tag{1}$$

so chance is 0.5. We report the retained fraction $\rho_s = (r_s^{\mathrm{U}} - 0.5)/(r_s^{\mathrm{S}} - 0.5)$ between the UNSEEN and SEEN arms below, averaged over subjects, and the architecture advantage $\Delta_s = r_s^{\mathrm{dyn}} - r_s^{\mathrm{stat}}$. Conditions differ in pool size, and Lemma 1 (Appendix A) is why every cross-condition comparison uses r: if the distractor scores are i.i.d. given the target's score, $\mathbb{E}[r_q]$ does not depend on $|\Pi_q|$, while top-10 accuracy falls as the pool grows. The lemma covers pool *size*; a change in pool *composition* changes the distractor distribution and is not removed.

The 2×2 factorial. Write $\Delta_{g\sigma}$ for a subject's advantage under guard g and split σ, each R (recording) or S (stimulus guard; story split). The paired contrasts are, per subject,

$$\begin{aligned} M_{\mathrm{split}} &= \tfrac{1}{2}\big[(\Delta_{\mathrm{RS}} - \Delta_{\mathrm{RR}}) + (\Delta_{\mathrm{SS}} - \Delta_{\mathrm{SR}})\big], \\ M_{\mathrm{guard}} &= \tfrac{1}{2}\big[(\Delta_{\mathrm{SR}} - \Delta_{\mathrm{RR}}) + (\Delta_{\mathrm{SS}} - \Delta_{\mathrm{RS}})\big], \end{aligned} \qquad \Delta_{\mathrm{SS}} - \Delta_{\mathrm{RR}} = M_{\mathrm{split}} + M_{\mathrm{guard}}, \tag{2}$$

with three-way term $I = (\Delta_{\mathrm{SS}} - \Delta_{\mathrm{SR}}) - (\Delta_{\mathrm{RS}} - \Delta_{\mathrm{RR}})$, averaged over subjects and bootstrapped over them. Because Δ is already a difference between encoders, M_{split} is the architecture×split interaction. The identity on the right is exact: a two-condition comparison that changes guard and split together measures the sum of the two main effects, and each one-factor-at-a-time effect differs from its main effect by $\pm I/2$ (Appendix A). The factorial separates what the joint change mixes; whether the two main effects differ from each other is the further contrast $M_{\mathrm{split}} - M_{\mathrm{guard}}$.

What a subject-level split identifies. Let $\xi \in [0, 1]$ be the fraction of the evaluation stimulus's material in the data a model is fitted on (pretraining or calibration), and suppose that, with every other feature of the protocol held fixed,

$$r_s^m(e, \xi) = \theta^m + u_e^m + h_e^m(\xi) + \varepsilon_s^m, \qquad h_e^m(0) = 0, \qquad \mathbb{E}_e\big[u_e^m\big] = 0, \tag{3}$$

where ε_s^m has mean zero given (e, ξ) and a law that does not involve (θ, u, h). Here θ^m is the encoder's level on unheard stimuli, u_e^m how hard stimulus e is for it, and h_e^m what exposure adds. The target for generalisation to unheard stimuli is $\tau = \theta^{\mathrm{dyn}} - \theta^{\mathrm{stat}}$.

Proposition 1. *Under (3), write $\delta u_e = u_e^{\mathrm{dyn}} - u_e^{\mathrm{stat}}$ and $\delta h_e = h_e^{\mathrm{dyn}}(1) - h_e^{\mathrm{stat}}(1)$. (i) Under a subject-level split (ξ=1), $\mathbb{E}[\Delta_s \mid e] = \tau + \delta u_e + \delta h_e$, and data observed only at ξ=1 are consistent with values of τ of either sign whenever $\theta^m + u_e^m > |\tau|$ for both encoders and every stimulus e (as for rank percentiles near or above chance), even if each h_e^m is required to be non-negative, non-decreasing and continuous. (ii) With the stimulus withheld (ξ=0), $\mathbb{E}[\Delta_s \mid e] = \tau + \delta u_e$, which averages to τ over stimuli; δu_e is shared by all subjects, so a subject bootstrap conditional on e does not include its variance. (iii) Both arms on the same stimulus identify δh_e as their difference.*

The proof is in Appendix A. Part (i) is why the architecture comparison is rerun with the stimulus withheld; (iii) is the difference-in-differences of §4, which our data do not resolve. No real protocol holds "every other feature fixed": removing a story also changes pretraining volume or composition, which the designs below bracket.

The matched exposure contrast. **SEEN** pretrains on stories $\{0, 1, 2\}$ of the other nine subjects, so story 0 is in pretraining; **UNSEEN** pretrains on $\{1, 2, 3\}$; **both evaluate on story** 0, against the same pool, with the same number of pretraining stories. Within one encoder, the stimulus term u_0^m cancels from this contrast, which leaves the exposure term plus the change of replacement narrative, of pretraining volume (story 3 is the longest, so UNSEEN pretrains on 51,143 segments against SEEN's 30,270) and of calibration diversity (UNSEEN calibrates on stories $\{1, 2, 3\}$, SEEN on $\{1, 2\}$; k fixed); a composition-matched arm ($\{0, 1, 2\} \to \{1, 2\}$, 25,660) and a volume-pinned dose sweep probe these.

Table 1: **Withholding the evaluation story from pretraining removes roughly half the above-chance signal at $k \geq 500$.** Top: mean percentile rank; *both arms are scored on story* 0 (or story 1), three pretraining stories either way; retained fraction is the mean over subjects of ρ_s with a paired 95% subject bootstrap, conditional on narrative, corpus and one pretraining seed. One subject lacks a k=2000 curve. Bottom: retained fraction (static encoder unless noted) under the contrasts that bracket the confounds; point estimates, n=9–10 for the composition-matched and cross-story rows, n=5 for the story-2 row ($[56.2, 59.5]$ at k=500; no k=2000 SEEN arm) and the seed rows. At equal k, UNSEEN calibrates on three stories and SEEN on two; restricting UNSEEN to SEEN's two changes the static retained fraction by -0.8 and -1.6 points at k=500 and 2000 (App. H.1). At k=100, which the pre-set rule did not cover, it is -10.2 $[-18.7, -2.5]$.

encoder	k	n	SEEN	UNSEEN	above-chance signal retained
story 0, static conv	100	10	0.779	0.689	71.7% [60.2, 83.6]
	500	10	0.861	0.695	54.1% [51.4, 56.6]
	2000	9	0.884	0.707	53.9% [52.2, 55.8]
story 1, static conv	100	5	0.804	0.679	61.7% [52.1, 75.7]
	500	5	0.852	0.681	51.1% [43.4, 56.6]
	2000	5	0.881	0.706	54.0% [52.7, 55.5]
story 0, dynamic	100	10	0.799	0.677	60.7% [53.8, 67.9]
	500	10	0.855	0.689	53.5% [50.3, 56.6]
	2000	9	0.878	0.690	50.3% [48.5, 52.4]

retained fraction, alternative contrast	k=100	k=500	k=2000
composition-matched, $\{0,1,2\} \to \{1,2\}$	53.4%	45.2%	49.3%
cross-story, story 3 withheld vs story 0 seen	64.5%	48.3%	48.7%
same, dynamic encoder	55.5%	46.8%	46.6%
as top, story 2 evaluated	66.5%	57.6%	—
as top, seeds 1 / 2 (story 0)	59.4% / 58.8%	53.0% / 53.0%	52.0% / 53.7%

4 RESULTS

Withholding the story removes about half the signal. Withholding the evaluation story from pretraining removes about half of the above-chance signal at k=500 and k=2000 and less at k=100, where the intervals are widest (Table 1), with evaluation stimulus and pool fixed and pretraining story count matched. Segment counts are not matched: every UNSEEN arm pretrains on 32–69% *more* segments than its SEEN arm (§3). The two ways of removing a story are confounded in opposite directions. If the substituted story were poorer pretraining material, the composition-matched arm, which substitutes nothing, should lose *less*; it loses more. If less pretraining data were the cause, the story-count-matched arm, which has more of it, should lose *nothing*; it loses about half, as do the endpoints of the volume-pinned dose sweep below. Neither confound accounts for the effect alone, and the two designs bracket the retained fraction at roughly 45–54% at the larger budgets. There, changing the withheld stimulus, the evaluation narrative or the seed moves it by a few points, not by its sign or order of magnitude (three-seed means in App. H); at k=100 the rows of Table 1 spread more widely. A from-scratch encoder is not a useful control: at k=100 and k=500 it sits at chance (App. H), so its flatness across conditions is a floor effect.

The effect scales with exposure. Keeping a fraction of the other subjects' recordings of the evaluation story and backfilling every dose to the same pretraining segment count pins volume, the subject set and optimisation, so only exposure and the backfilled mix of other stories vary. Retrieval rises monotonically with exposure at both budgets (Appendix D; n=4 subjects per dose, reported descriptively). Without the backfill the same sweep loses the k=500 monotonicity, so volume does not produce the gradient (Appendix H).

What survives. On a narrative no brain in the training set has heard, the UNSEEN condition still scores percentile 0.66–0.71 against 0.5 chance. Above chance is not decoding meaning: in our own ten-subject pipeline a brain-free baseline of timing and word-length features also leads a MEG encoder (§5), so we claim transfer of retrievable signal, not of its content. Against a from-

Table 2: **The dynamic encoder's advantage is resolved at both low budgets only under the standard protocol.** dynamic − static rank percentile on the guard×split factorial, all four stories in pretraining except in the stimulus/story cell, which is Table 1's SEEN arm (story 3 withheld from pretraining and calibration, 54% of the other cells' pretraining segments, story 0 evaluated); n=10 held-out subjects paired across all cells; unadjusted paired 95% bootstrap intervals, * excludes zero. Contrasts follow Eq. (2): the split main effect is the architecture×split interaction. Pools differ up to sevenfold between cells, so ×chance is not shown.

	guard, split	k=100	k=500
cell	recording, recording (standard)	$+0.043$ $[+0.009, +0.081]^*$	$+0.014$ $[+0.001, +0.026]^*$
	stimulus, recording	$+0.034$ $[-0.016, +0.083]$	$+0.024$ $[+0.005, +0.045]^*$
	recording, story	$+0.013$ $[-0.011, +0.037]$	$+0.010$ $[-0.003, +0.024]$
	stimulus, story	$+0.020$ $[-0.019, +0.055]$	-0.006 $[-0.022, +0.011]$
contrast	split main M_{split}	-0.022 $[-0.071, +0.026]$	-0.017 $[-0.031, -0.001]^*$
	guard main M_{guard}	-0.002 $[-0.027, +0.023]$	-0.003 $[-0.014, +0.007]$
	three-way I	$+0.016$ $[-0.047, +0.078]$	-0.027 $[-0.061, +0.004]$

scratch encoder trained on the same k segments with learning rate and weight decay tuned on validation, pretraining roughly doubles ×chance at k=2000, both arms in the UNSEEN condition, so the comparison is untouched by the mechanism measured here (n=5; all budgets in Appendix H). On Narratives fMRI (Nastase et al., 2021), with no pretraining stage, an analogous held-out-material manipulation lowers retrieval in the same direction (Appendix E); that corroborates split-sensitivity and is silent on pretraining exposure.

The architecture comparison. The shared stimulus also reaches the model through calibration segments drawn from the evaluation stories. Under the protocol as originally run (recording guard, recording split), the dynamic encoder (App. I) beats a static encoder parameter-matched within 8.4% at both low budgets. The ×chance criterion preceded the deciding sweep, but the pair and input were chosen after pilots and all ten subjects had earlier zero-calibration and k=500 runs (App. I); it was met at k=100 (n=10; App. H). Every protocol change also changes the pool, so we analyse the comparison in r (Table 2; this reading, like the factorial, was chosen after both protocols' results); comparing ×chance across protocols whose pools differ sevenfold would have attributed roughly two fifths of the apparent collapse to the model when it was the pool moving (App. H). Making calibration story-disjoint leaves the advantage undetectable with all four stories still in pretraining, and so does combining it with the stimulus guard in a cell that also withholds story 3 from pretraining (Table 2). The stimulus/story cell's k=100 interval still contains the standard-protocol estimate (the recording/story cell's does not, though that is not a paired test), so these cells establish that nothing resolves positive, not that the difference is zero; a change in significance is not itself a significant change (Gelman and Stern, 2006), which is why the inference below rests on paired contrasts.

The factorial: an inference-sensitive split interaction. The four cells share the same ten subjects, so the contrasts of Eq. (2) are paired within subject (Table 2, Figure 2). They do not share one pretraining corpus or story set: the stimulus/story cell also withholds story 3, calibrates on $\{1, 2\}$ and scores story 0 only (§3), so every contrast that involves it carries those changes too, including the split simple effect under the stimulus guard, which carries most of the k=500 split main effect; with the recording guard and pretraining fixed, the split simple effect is near zero and unresolved at that budget (both in App. C). At k=500 the architecture×split interaction is resolved before correction and the guard main effect is bounded near zero; their difference is itself unresolved, so we do not order the two fixes. At k=100 both contrasts point the same way and neither resolves at n=10. Under a Bonferroni correction over the factorial's four contrasts (these three and $M_{\text{split}} - M_{\text{guard}}$) the k=500 interaction no longer resolves, its uncorrected upper limit sitting just below zero, and a post-hoc t_9 interval and sign-flip test agree (Appendix C). Both k=500 stars (standard cell, M_{split}) clear zero by less than plausible percentile-bootstrap undercoverage at n=10, so the standard-protocol advantage rests on k=100; and r is invariant to pool size, not composition, so the split main effect prices the protocol change as a whole. We report support for a low-budget comparison sensitive to the split at this sample size, not a confirmed change in the difference between encoders.

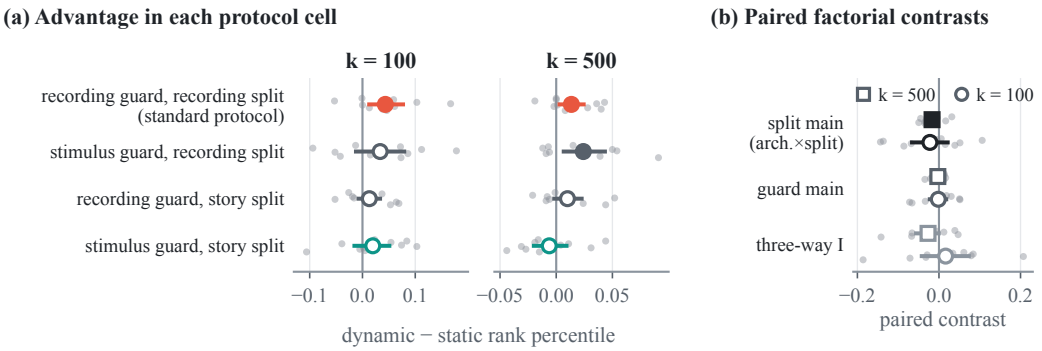

Figure 2: **Only the standard protocol gives the dynamic encoder a resolved advantage at both low budgets; the factorial contrasts are small and mostly unresolved.** (a) Advantage in each guard×split cell at k=100 and k=500. (b) Three paired contrasts of Eq. (2) ($M_{\text{split}}-M_{\text{guard}}$ is in App. C). Marks are subject means with unadjusted paired 95% bootstrap intervals (filled: excludes zero); grey dots are the ten held-out subjects. The k=500 interaction excludes zero before, not after, Bonferroni correction over the four contrasts (M_{split}, M_{guard}, I, $M_{\text{split}}-M_{\text{guard}}$). Guards and splits: §3. Values from `results/interaction_receipt.json` (cell intervals from `figures/fig_factorial_subjects.values.json`, same bootstrap).

Table 3: **With the story also withheld from pretraining, no narrative gives the dynamic encoder a resolved low-budget advantage.** Dynamic − static in rank percentile, * interval excludes zero (unadjusted paired 95% subject bootstrap). Each story row withholds that story from pretraining and calibration and evaluates on it; the stimulus/story cell of Table 2 instead withholds story 3 and evaluates story 0. The first and last rows have n=9 at k=2000.

withheld stimulus	n	k=100	k=500	k=2000
none (standard protocol)	10	+0.043*	+0.014*	−0.007*
story 0	10	−0.012	−0.006	−0.016*
story 1	5	−0.006	+0.010	−0.011*
story 2	5	−0.014	−0.020*	−0.024*
story 3	10	−0.008	−0.009	−0.011*

Across all four stimuli. Withholding the evaluation story from pretraining as well (Table 3), the dynamic advantage is a *low-budget* phenomenon of the *overlapping* protocol: at k=100 and k=500 it is resolved only in the standard row, and in none of the eight stimulus-held-out cells does the interval exclude zero in its favour. At k=2000 the static encoder leads everywhere, including under the standard protocol, so the largest-budget reversals are not purely an exposure effect; the advantage that disappears with the shared stimulus is the low-budget one, where a pretrained prior is supposed to matter and where the pre-specified criterion lived. Three of the four k=2000 reversals remain resolved under a Bonferroni correction over the four stimuli. The difference-in-differences of Proposition 1(iii), whether the pretraining holdout further changes the advantage, does not resolve at any budget and is wide enough to contain effects as large as any we report, so we draw no conclusion from it. At k=500 point estimates differ in sign across narratives, one of them resolved against the dynamic encoder at n=5 (Table 3); we do not test between-stimulus heterogeneity, and four narratives cannot say which stimulus properties matter.

5 Auditing a Published Reference Decoder

A brain-free clock outscores the decoder. We repeat the audit, and put the story back, inside a reproduction of the reference decoding pipeline (Défossez et al., 2023): a seed replicate reaches 41.4% top-1 with 7,140 candidates, against the published 41.3% with 1,363; these are different evaluation settings. We add one arm the pipeline does not contain, a *word-onset clock*: the same network and training with each MEG window replaced by a raster of the word onsets heard in it,

Table 4: **Reference pipeline: a brain-free clock beats the decoder; full-dose exposure is worth about half the story split's volume cost.** Evaluation narrative *The Black Willow*; guarded top-10 at a pinned 7,140-query pool. Arm rows: mean over runs with [min, max]; A–C three seeds. Contrast rows: conservative cross-seed bounds (not confidence intervals) or the cross-arm span. C′ and D share C's training stories but are scored on the evaluation narrative's block-hash test blocks, so their levels are not comparable with A–C; a D-type gain and $A-B$ are measured on different test blocks.

arm	train ex.	top-10 (%)
A block split (shipped)	203,152	57.80 [57.21, 58.41]
B block split, thinned to C	120,389	51.80 [51.58, 51.96]
C story split	118,310	47.28 [46.88, 47.69]
C′ dose 0 (3 runs)	118,310	45.86 [44.64, 46.60]
D dose 0.59 (5 runs)	119,502	48.26 [47.61, 49.23]
dose 0.30 / 1.0 (1 run each)	matched	46.63 / 48.86
clock, block split	203,152	90.36 [89.52, 92.00]
$A-B$: training volume		+6.01 [+5.25, +6.83]
$B-C$: exposure and composition		+4.51 [+3.89, +5.08]
$D-C'$: 0.59-dose exposure		+1.01 to +4.59
full dose − dose 0		+3.00
2nd narrative, full − dose 0		+2.00

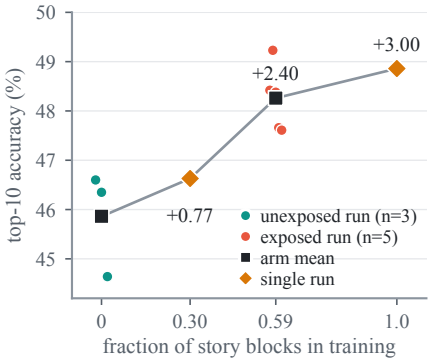

Figure 3: **Putting the story back: at matched volume, the story-split decoder's accuracy rises across four exposure doses; doses 0.30 and 1.0 are single runs, the first barely above the best unexposed run.** Labels: gain over the dose-0 mean. Runs from `onebit_wer_readouts.json` (stored as top-10 error), doses 0.30 and 1.0 from `fig_dose_ladder.values.json`.

and no brain data. It is not a rival decoder but a measure of how much of the metric word timing can win on its own. On the identical test set and pool the clock outscores the reproduced MEG decoder by more than 30 top-10 points in every seed (Table 4). In our own ten-subject pipeline a brain-free baseline of timing and word-length features also leads a MEG encoder in rank percentile, under a stimulus-scoped guard (one story) and under the recording guard; both intervals exclude zero and come from different designs (Appendix B.3).

Correcting the pipeline's own split. The pipeline assigns blocks to splits by hashing each block's word sequence; a story spans many blocks, so the model is fitted on parts of a narrative and scored on other parts of it. Story-level splitting keeps every narrative in one split. The released evaluation sets its pool to $\min(10{,}000, N_{\text{queries}})$, so a story split would also move the pool and chance; we pin pool size, not membership, at 7,140 in every arm (Appendix B.3). At best-validation checkpoints the story split costs guarded top-10 just under a fifth of its above-chance signal (Table 4; App. H), and the decoder stays far above chance: overstated, not collapsed. Pool size is pinned but composition is not, and the story-to-split assignment was selected (Appendix H) to give a non-degenerate partition with *The Black Willow* as evaluation narrative.

Point estimates put most of the gap on volume; full exposure is worth about half the volume cost. The story split also discards training data, so arm B keeps the block split and thins training to C's size. At three seeds per arm the ordering $A > B > C$ holds with no overlap (Table 4), and the larger share of the gap sits on the discarded data at every point estimate, while the cross-seed bound on that share still admits an even split (App. H). $B-C$ mixes the remaining exposure with composition and checkpoint-selection changes, so two further arms differ in one design bit, whether the evaluation narrative's other blocks are in training, at matched volume with byte-identical evaluation blocks. Volume matching thins the added narrative too, so the exposed arm D carries 59% of the narrative's blocks. Five D and three C′ runs separate completely (fifteen of fifteen cross-arm differences positive; span in Table 4); an exact permutation p value holds only if the eight runs are exchangeable, which the design does not guarantee (App. H). A dose ladder then measures exposure directly (Figure 3), with full dose (one run) worth about half the fix's volume cost ($A-B$), though the two are scored on different test blocks. Fixing *Easy Money* as evaluation narrative instead replicates the sign in single runs (Table 4). This exposure is calibration-level (other blocks of the same narrative, no pretraining), not the pretraining-level loss of §4.

6 RECOMMENDATIONS

Check the split first: the released tools report whether evaluation stimuli reach the fitting data (App. J); a subject-only holdout on MEG-MASC exposes every evaluation story, and where all listeners heard one stimulus no split separates transfer from exposure. **For unheard-stimulus generalisation, split by stimulus and report the guard separately.** A subject-level split cannot identify that target (Proposition 1(i)); the split effect has the larger estimate, but its difference from the guard effect remains unresolved. A story split is the expensive fix—§5 prices its volume cost at about twice the full-dose exposure it removes (one full-dose run, scored on other test blocks; App. H)—and it is worth the price because what it protects is the *conclusion*, not the score. **Hold the stimulus out of pretraining, not just the subject**, report both numbers, and attribute the gap to exposure only as far as the matched controls allow; a model pretrained on the evaluation corpus should list the evaluation stimuli it saw. **Report a stimulus-derived baseline** (timing or other stimulus-side features, no brain data) alongside every retrieval result (§5).

7 LIMITATIONS

- **One corpus carries the causal claims.** MEG-MASC supplies the factorial, the dose-response and the magnitude estimate; Narratives fMRI corroborates split-sensitivity in a second modality without a pretraining stage. In this corpus we report the direction and order of magnitude, without establishing an ordering of the guard and split effects; whether the fraction is near a half elsewhere is untested. Pretraining is small (at most $59{,}014$ segments from nine listeners; ten of the corpus's 27 subjects are used); whether the effect survives multi-corpus pretraining is untested.
- **Stimulus-level claims rest on four narratives.** Their outcomes differ (one narrative reverses at $k{=}500$, all four at $k{=}2000$, the remaining low-budget cells null), but we do not test that difference, and four are too few to say which properties of a stimulus decide it; the dose-response varies exposure within them, so diversity and exposure are inseparable here.
- **The architecture result is budget-resolved.** The *low-budget* advantage appears only under the overlapping protocol; the $k{=}2000$ reversals also hold under the standard protocol (§4), so they are not purely an exposure effect, and the pretraining-level difference-in-differences bounds nothing. All intervals price subject resampling only, conditional on this corpus and one pretraining seed (the three-seed check puts seed variance near subject variance at the largest budget); subject counts were extended $5 \to 10$ in some conditions after an interim look, and their starred intervals carry no sequential correction.
- **The selection sensitivity is an existence proof from one architecture pair of our own construction.** The two encoders (2.4M and 2.7M parameters, built to isolate the protocol, not published foundation models) differ in two ways—a residual-connection change in the temporal stack and the finite-difference branch (App. I)—so the sensitivity is an existence proof for the pair, not for either ingredient alone. The one published ingredient tested (Appendix F) retains its advantage; neither a general ranking reversal nor the contribution of either architectural ingredient is established.
- **The comparator is a tuned from-scratch baseline.** The from-scratch encoder was tuned on validation (§4) but sits at chance at $k{=}100$ and $k{=}500$, so only the largest budget compares two models that both learned; a stronger per-subject decoder or architecture could narrow the gap.
- **The decomposition and exposure arms have three to five runs each, which bound the shares rather than resolve them.** The ablated cells and both clock controls are single runs (the block-split clock has three seeds). The $A/B/C$ decomposition holds its ordering in every seed, but the volume-share bound still admits an even split, and cross-seed ranges are not confidence intervals. The D-versus-C$'$ contrast separates completely at five D vs three C$'$ runs but supports a range, not an interval; the ladder endpoints and the second narrative are single runs; the $B{-}C$ residual remains composition- and selection-confounded.
- **We audit a protocol, not anyone's conclusion.** Beyond the reproduced reference pipeline we do not re-run published models, so how far any published number would move is outside this design; the corpus auditor reports exposure, a precondition for inflation, not its measurement.

AI USE STATEMENT

We used generative AI tools (Anthropic Claude and OpenAI Codex agents) to implement experiment code and check it for correctness, to collect public datasets, to check references, and to polish the writing. The authors take full responsibility for all content.

ETHICS STATEMENT

This work re-analyses publicly released, de-identified neural recordings (MEG-MASC and the Narratives fMRI collection) under their existing licenses and collects no new human-subject data. It audits an evaluation protocol rather than any individual's decoder; where we reproduce a published pipeline we report our reproduction's numbers, not the authors'. The released split and audit tools are meant to reduce, not create, risk of over-claiming in brain decoding. We see no dual-use concern beyond those already attached to the public corpora.

REPRODUCIBILITY STATEMENT

Appendix I gives architectures, parameter counts, training and calibration settings, and compute; the generating scripts are named in Appendices J and I and in the supplement's `RELEASE.md`. The supplementary archive (inventory in its `RELEASE.md`) contains the two stdlib-only audit tools, the stimulus-disjoint MEG-MASC split with its manifests, the environment-gated patches to the reference pipeline in their required order, the model code, the per-table generators and the run configurations of every internal and external arm, the protocol note, and anonymised excerpts of the append-only lab notebook and program document that date the criterion and each pre-committed decision rule. It includes subject-level factorial contrasts and eight external exposure-arm readouts, with stdlib-only scripts that recompute their summaries off-cluster. Training logs are not shipped; the other generators need them and fail explicitly without them, and recomputing a summary does not rerun training. The original internal conv/dynamic sweep retained summary JSONs but not trained weights or per-query outputs, so common-pool rescoring cannot be recovered from those artifacts; a new controlled run is required.

REFERENCES

J. R. Brennan and J. T. Hale. Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PLOS ONE*, 14(1):e0207741, 2019.

P.-H. Chen, J. Chen, Y. Yeshurun, U. Hasson, J. Haxby, and P. Ramadge. A reduced-dimension fMRI shared response model. In *NeurIPS*, 2015.

H. H. Clark. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 12(4):335–359, 1973.

S. d'Ascoli, C. Bel, J. Rapin, H. Banville, Y. Benchetrit, C. Pallier, and J.-R. King. Towards decoding individual words from non-invasive brain recordings. *Nature Communications*, 16:10521, 2025.

A. Défossez, C. Caucheteux, J. Rapin, O. Kabeli, and J.-R. King. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5:1097–1107, 2023.

A. Gelman and H. Stern. The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, 60(4):328–331, 2006.

L. Gwilliams, G. Flick, A. Marantz, L. Pylkkänen, D. Poeppel, and J.-R. King. Introducing MEG-MASC a high-quality magneto-encephalography dataset for evaluating natural speech processing. *Scientific Data*, 10:862, 2023.

D. D. Han, Y. Gwon, et al. DIVER-1: Scaling intracranial EEG foundation models for transferable representations. arXiv:2512.19097, 2025.

U. Hasson, Y. Nir, I. Levy, G. Fuhrmann, and R. Malach. Intersubject synchronization of cortical activity during natural vision. *Science*, 303:1634–1640, 2004.

J. V. Haxby, J. S. Guntupalli, A. C. Connolly, Y. O. Halchenko, B. R. Conroy, M. I. Gobbini, M. Hanke, and P. J. Ramadge. A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, 72(2):404–416, 2011.

D. Jayalath and O. Parker Jones. MEG-XL: Data-efficient brain-to-text via long-context pre-training. In *ICML*, 2026. arXiv:2602.02494.

D. Jayalath, G. Landau, B. Shillingford, M. Woolrich, and O. Parker Jones. The brain's bitter lesson: Scaling speech decoding with self-supervised learning. In *ICML*, PMLR 267:27005–27020, 2025.

W.-B. Jiang, L.-M. Zhao, and B.-L. Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *ICLR*, 2024. arXiv:2405.18765.

S. Kajimura, T. Hoshino, and K. Murayama. Stimulus-specific random effects inflate false-positive classification accuracy in multivariate-voxel-pattern-analysis: A solution with generalized mixed-effects modelling. *NeuroImage*, 269:119901, 2023.

S. Kapoor and A. Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023.

K. N. Kay, T. Naselaris, R. J. Prenger, and J. L. Gallant. Identifying natural images from human brain activity. *Nature*, 452:352–355, 2008.

J. A. Kilgallen, B. A. Pearlmutter, and J. M. Siskind. The repeated-stimulus confound in electroencephalography. arXiv:2508.00531, 2025.

D. Kostas, S. Aroca-Ouellette, and F. Rudzicz. BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data. *Frontiers in Human Neuroscience*, 15:653659, 2021.

N. Kriegeskorte, W. K. Simmons, P. S. F. Bellgowan, and C. I. Baker. Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neuroscience*, 12(5):535–540, 2009.

B. Lahner, M. Deb, N. A. R. Murty, and A. Oliva. MOSAIC: A scalable framework for fMRI dataset aggregation and modeling of human vision. bioRxiv 10.64898/2025.11.28.690060, 2025.

N. Lee, K. Barmpas, Y. Panagakis, D. Adamos, N. Laskaris, and S. Zafeiriou. Are large brainwave foundation models capable yet? Insights from fine-tuning. In *ICML*, PMLR 267:32878–32888, 2025.

R. Li, J. S. Johansen, H. Ahmed, T. V. Ilyevsky, R. B. Wilbur, H. M. Bharadwaj, and J. M. Siskind. The perils and pitfalls of block design for EEG classification experiments. *IEEE TPAMI*, 43(1):316–333, 2021.

J.-Y. Lin, Y. C. Wu, and T.-P. Jung. The identity trap in EEG foundation models: A diagnostic audit. arXiv:2606.06647, 2026.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. arXiv:1711.05101.

I. Magar and R. Schwartz. Data contamination: From memorization to exploitation. In *ACL (Short Papers)*, pages 157–165, 2022.

M. Meyer, S. Kaltenpoth, K. Zalipski, and O. Müller. Rethinking evaluation in the era of time series foundation models: (Un)known information leakage challenges. arXiv:2510.13654v3, 2025.

S. A. Nastase, Y.-F. Liu, H. Hillman, A. Zadbood, L. Hasenfratz, N. Keshavarzian, J. Chen, C. J. Honey, Y. Yeshurun, M. Regev, M. Nguyen, C. H. C. Chang, C. Baldassano, O. Lositsky, E. Simony, M. A. Chow, Y. C. Leong, P. P. Brooks, E. Micciche, G. Choe, A. Goldstein, T. Vanderwal, Y. O. Halchenko, K. A. Norman, and U. Hasson. The "Narratives" fMRI dataset for evaluating models of naturalistic language comprehension. *Scientific Data*, 8:250, 2021.

M. Özdogan, G. Landau, G. Elvers, D. Jayalath, P. Somaiya, F. Mantegna, M. Woolrich, and O. Parker Jones. LibriBrain: Over 50 hours of within-subject MEG to improve speech decoding methods at scale. In *NeurIPS Datasets and Benchmarks Track*, 2025. arXiv:2506.02098.

A. Paszke, S. Gross, F. Massa, et al. PyTorch: An imperative style, high-performance deep learning library. In *NeurIPS*, volume 32, 2019. arXiv:1912.01703.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. OpenAI technical report, 2019.

M. Rosenblatt, L. Tejavibulya, R. Jiang, S. Noble, and D. Scheinost. Data leakage inflates prediction performance in connectome-based machine learning models. *Nature Communications*, 15:1829, 2024.

R. Schaeffer, J. Kazdan, B. Abbasi, et al. Quantifying the effect of test set contamination on generative evaluations. arXiv:2601.04301, 2026.

I. Semenkov, D. Kleeva, I. Dakhtin, Z. Maksudova, and A. Ossadtchi. Interpretable MEG decoding of perceived speech: cortical sources and the stimulus features that drive retrieval. arXiv:2608.01481, 2026.

K. Shirakawa, Y. Nagano, M. Tanaka, S. C. Aoki, Y. Muraki, K. Majima, and Y. Kamitani. Spurious reconstruction from brain activity. *Neural Networks*, 190:107515, 2025.

J. Tang, A. LeBel, S. Jain, and A. G. Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26:858–866, 2023.

A. van den Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. arXiv:1807.03748, 2018.

C. Wang, V. Subramaniam, A. U. Yaari, G. Kreiman, B. Katz, I. Cases, and A. Barbu. BrainBERT: Self-supervised representation learning for intracranial recordings. In *ICLR*, 2023. arXiv:2302.14367.

X. Wang, Y. Yang, and D. Coyle. EEG-FM-Audit: A systematic evaluation and analysis pipeline for EEG foundation models. arXiv:2605.26910, 2026.

J. Westfall, T. E. Nichols, and T. Yarkoni. Fixing the stimulus-as-fixed-effect fallacy in task fMRI. *Wellcome Open Research*, 1:23, 2017.

T. Wolf, L. Debut, V. Sanh, et al. Transformers: State-of-the-art natural language processing. In *EMNLP: System Demonstrations*, pages 38–45, 2020. DOI: 10.18653/v1/2020.emnlp-demos.6.

J. Wu, Z. Ren, J. Wang, et al. AdaBrain-Bench: Benchmarking brain foundation models for brain-computer interface applications. arXiv:2507.09882, 2025.

Q. Xiao, Z. Cui, C. Zhang, S. Chen, W. Wu, A. Thwaites, A. Woolgar, B. Zhou, and C. Zhang. BrainOmni: A brain foundation model for unified EEG and MEG signals. In *NeurIPS*, 2025. arXiv:2505.18185.

C. Yang, M. B. Westover, and J. Sun. BIOT: Biosignal transformer for cross-data learning in the wild. In *NeurIPS*, volume 36, pages 78240–78260, 2023.

C. Yin, Q. Yu, Z. Fang, C. Peng, and P. Li. Rethinking cross-subject data splitting for brain-to-text decoding. In *EMNLP*, pages 5675–5689, 2025.

A. Zahorodnii, C. Wang, G. Chau, et al. Neuroprobe: Evaluating intracranial brain responses to naturalistic stimuli. arXiv:2509.21671, 2025.

X. Zhang, S. Liu, R. Lu, A. Woolgar, and L. Wang. What are we actually decoding? Source attribution for non-invasive brain-to-language retrieval. arXiv:2605.24524, 2026.

A PROOFS

Lemma 1 (pool-size invariance of the rank percentile)**.** *Fix a query and let T be the target's score and $D_1, \dots, D_{N-1}$ the distractor scores, where $N = |\Pi_q|$. Suppose that, given T, the D_i are i.i.d. with a continuous distribution function G that does not depend on N, and the law of T does not depend on N either. Then $\mathbb{E}[r_q] = \mathbb{E}[G(T)]$ for every N, whereas top-10 accuracy $A_N = \Pr\left(\#\{i : D_i > T\} \le 9\right)$ is non-increasing in N, and strictly decreasing for $N \ge 10$ whenever $\Pr(0 < G(T) < 1) > 0$.*

Proof. By Eq. (1), $r_q = \frac{1}{N-1}\sum_i \mathbf{1}[D_i < T]$. Given T each indicator has mean $G(T)$, so $\mathbb{E}[r_q \mid T] = G(T)$ for every N; take expectations. For accuracy, given T the count $S_{N-1} = \#\{i : D_i > T\}$ is $\mathrm{Bin}(N-1, p)$ with $p = 1 - G(T)$ (ties have probability zero). Adding one distractor gives $\Pr(S_{n+1} \le 9) = \Pr(S_n \le 9) - p\,\Pr(S_n = 9)$, so $\Pr(S_n \le 9)$ is non-increasing in n and strictly decreasing for $n \ge 9$ when $0 < p < 1$; take expectations over T. □

The assumption is the one the body states: pools that differ in *size* but draw distractors from the same distribution. A story-disjoint pool also changes G (it removes cross-story candidates), and the lemma says nothing about that change; this is why the split main effect prices the protocol change as a whole.

The factorial identities of Eq. (2). Write $a = \Delta_{\mathrm{RS}} - \Delta_{\mathrm{RR}}$ and $b = \Delta_{\mathrm{SS}} - \Delta_{\mathrm{SR}}$ (split simple effects under each guard) and $c = \Delta_{\mathrm{SR}} - \Delta_{\mathrm{RR}}$, $d = \Delta_{\mathrm{SS}} - \Delta_{\mathrm{RS}}$ (guard simple effects under each split). Then $M_{\mathrm{split}} = (a+b)/2$, $M_{\mathrm{guard}} = (c+d)/2$ and $I = b - a = d - c$. Hence $M_{\mathrm{split}} + M_{\mathrm{guard}} = \frac{1}{2}(a+b+c+d) = \Delta_{\mathrm{SS}} - \Delta_{\mathrm{RR}}$, $a = M_{\mathrm{split}} - I/2$, $b = M_{\mathrm{split}} + I/2$, and likewise $c, d = M_{\mathrm{guard}} \mp I/2$. All identities hold per subject, so they hold for subject means and for every bootstrap draw. In `results/interaction_receipt.json` at k=500, $\Delta_{\mathrm{SS}} - \Delta_{\mathrm{RR}} = -0.0197 = M_{\mathrm{split}} + M_{\mathrm{guard}}$. □

Proof of Proposition 1. (i) Subtracting the two encoders' versions of Eq. (3) at ξ=1 and taking expectations given e gives $\mathbb{E}[\Delta_s \mid e] = \tau + \delta u_e + \delta h_e$. For non-identification, take any $a > 0$ and replace $(\theta^{\mathrm{stat}}, h_e^{\mathrm{stat}}(\xi))$ by $(\theta^{\mathrm{stat}} - a,\ h_e^{\mathrm{stat}}(\xi) + a\xi)$ for every e. The new h still vanishes at 0 and stays non-negative, non-decreasing and continuous if the old one was; $\theta + h(1)$ is unchanged for every stimulus, so, since the residual law does not involve the parameters, the joint law of all observations at ξ=1 is unchanged, while τ increases by a. For $a < 0$ apply the same change, with $-a$, to the dynamic encoder: $(\theta^{\mathrm{dyn}} + a,\ h_e^{\mathrm{dyn}}(\xi) - a\xi)$. The shifts only lower levels, so the binding constraint is $\theta^m + u_e^m \ge 0$ at ξ=0: every τ' in $[\tau - \min_e(\theta^{\mathrm{dyn}} + u_e^{\mathrm{dyn}}),\ \tau + \min_e(\theta^{\mathrm{stat}} + u_e^{\mathrm{stat}})]$ is therefore reachable from the same observable law, and this interval contains both signs under the stated condition. (ii) At ξ=0, $h_e^m(0) = 0$ gives $\mathbb{E}[\Delta_s \mid e] = \tau + \delta u_e$, and $\mathbb{E}_e[\delta u_e] = 0$ gives τ on average over stimuli. Conditional on e, δu_e is a constant added to every subject's contrast, so resampling subjects leaves it fixed and the interval omits $\mathrm{Var}_e(\delta u_e)$. (iii) Subtracting (ii) from (i) for the same e gives δh_e. □

The proposition is about identification, not about magnitudes: it says which contrasts can carry the unheard-stimulus comparison, not how large δh_e is. Its "everything else fixed" assumption fails in practice, because removing a story also changes pretraining volume or composition; §4 brackets those changes with the story-count- and composition-matched arms and the volume-pinned dose sweep, and the difference-in-differences that estimates δh_e in our data does not resolve.

B SUPPORTING LITERATURE AND REFERENCE-PIPELINE DETAILS

B.1 PRIOR WORK ON STIMULUS OVERLAP

The overlap itself is documented—for generation, not for transfer. Yin et al. (2025) identify that subjects sharing stories carries the stimulus into training in cross-subject brain-to-text decoding and show performance collapsing under a stimulus-disjoint split, which we take as established. Four gaps separate their result from ours, two of them named in their own limitations: text *generation* scored by BLEU, with cross-subject foundation-model *pretraining* unexamined (their pretrained encoder is task-specific); fMRI and EEG, leaving ECoG and MEG to future work; a remedy that

discards data, confounding the collapse with training-set size—our story-count- and composition-matched arms bracket that confound and the volume-pinned dose sweep removes it by construction (§4), and our external pipeline prices it (§5); and no question of whether the overlap changes which model wins. Retrieval metrics have also been audited with stimulus-side controls: Zhang et al. (2026) audit brain-to-language retrieval on the same MEG corpus, identifying sentence *duration* as a stimulus-derived nuisance and enforcing a stimulus-identity split covering repetitions across subjects and sessions—such splitting on this corpus is not new, and we do not claim it. But their design has no pretraining stage; they report the corrected number rather than what the correction *costs*; and their strongest stimulus-side result is a signal-blind control collapsing, where ours is a stimulus-derived model with no brain data *winning*. AdaBrain-Bench (Wu et al., 2025) establishes that a pretrained prior helps most when target data is scarce; we treat that regime as the backdrop and ask what varies inside it.

B.2 Adjacent evaluation audits

Adjacent failures are documented at other scales. Within-subject stimulus repetition inflates EEG decoding (Kilgallen et al., 2025; Li et al., 2021)—the cross-subject case is the same mechanism at a different scale: the other subjects hear the stimulus before the held-out one is evaluated—and Schaeffer et al. (2026) sweep contamination *dose* in language models, where we hold volume fixed and remove one stimulus. Audits find subject-identity shortcuts (Lin et al., 2026) and catalogue pretraining-corpus overlap in time-series foundation models, testing only indirect overlap in a controlled design (Meyer et al., 2025); iEEG benchmarks guard temporal leakage within a session and add a cross-session split on a different movie, but evaluate baselines pretrained on the benchmark's own corpus (Zahorodnii et al., 2025; Han et al., 2025); DIVER-1 (Han et al., 2025) reports in-corpus pretraining beating external at matched size; stimulus overlap can manufacture fMRI reconstructions (Shirakawa et al., 2025), and MOSAIC (Lahner et al., 2025) deduplicates stimuli as hygiene. Appendix G engages each; none measures what the overlap is worth at the pretraining level. Concurrent work on the same benchmark reaches for the opposite frame: Semenkov et al. (2026) rework Défossez-style MEG retrieval for interpretability (39.75% top-1 among $1{,}005$ candidates) and show, by feature occlusion, that retrieval is driven largely by silence, sound intensity, vowels and acoustic onsets; they read a narrative-versus-word-list substitution as narrative-evoked activity carrying more recoverable information. Every one of those driving features is *stimulus-derived*: their decomposition and our word-onset clock (§5) are two routes to the same finding—this metric is largely winnable from the stimulus side—and their probe addresses within-training shortcut use, not the pretraining-level overlap here. Finally, independent audits report tuned supervised baselines matching brain foundation models (Lee et al., 2025; Wang et al., 2026); ours is not another such audit: we find a transfer effect a tuned from-scratch encoder does *not* reach. It is smaller than advertised, and it is real.

B.3 The reference candidate pool

The reference implementation's candidate pool is set by the split, not by the config. Reproducing the pipeline surfaced §3's metric problem inside the released evaluation code itself. The evaluation draws its candidate pool as the first `wer_negatives` entries of a random permutation of the retained queries, with `wer_negatives` defaulting to $10{,}000$; the pool is therefore $\min(10{,}000,\ N_{\text{queries}})$, and on this corpus the cap does *not* bind: the run retains $7{,}140$ queries, so the pool is $7{,}140$, fixed by how the split sized the test partition, not by the experimenter. A protocol change then alters the pool as a side effect: story-level splits grow the WER evaluation from 56 to 453 batches—about $58{,}000$ retained queries—the cap binds, the pool becomes $10{,}000$, and chance moves from 0.140% to 0.100%: top-10 moves at unchanged skill, in the direction that flatters the corrected protocol. We therefore pin the query set to $7{,}140$ in both conditions and release that change as a patch. Reproduction surfaced a third determinant while we were pinning the second: pool *membership* also depends on each run's own fitted normalization (App. H, qualification (ii)), so only pool *size* can be pinned.

The mechanism is documented in the pipeline's own lineage: word length is linearly predictable from their word-embedding targets ($r{=}0.71$ over the $5{,}000$ most frequent English words), incorrect predictions share the true word's length more often than chance, and the protocol presupposes known onsets (d'Ascoli et al., 2025). Unreported is that the redundancy suffices for a brain-free model

to *win*—a question noise and shuffle controls cannot ask: they test whether the recording carries information, not whether it is redundant with what the experimenter already knows. Nor is the gap seed noise: three runs of the brain arm give top-10 of 58.41, 57.79 and 57.21%, a spread of 1.2 points; three of the timing arm give 89.52, 89.57 and 92.00%, a spread of 2.5 points; the arms are about 31 points apart, more than twelve times the larger spread. We quote ranges; at n=3 a standard deviation is estimated only to within about half its value. One caveat cuts against us: only the two reference runs were early-stopped on validation loss, so the additional seeds are floors, and a fully converged brain arm would narrow this gap rather than widen it. The reproduction trains on every MEG-MASC recording as released; the rank-percentile checks come from our own pipeline (the ten subjects of §3, 3-s segments), where a brain-free encoder of nine per-segment timing and word-length features leads a MEG encoder trained on the same segments by $+0.122$ under a stimulus-scoped guard ($[+0.109, +0.135]$, one story), and by $+0.144$ under the recording guard with all four stories (crossed cluster bootstrap $[+0.127, +0.160]$; a different design, not an interval for $+0.122$).

C Guard-versus-Split Decomposition

The comparison in §4 changes the adjacency guard and the calibration/evaluation split together. Crossing the two on the same ten held-out subjects separates them, with one caveat: three cells pretrain on all four stories, while the stimulus-guard/story-split cell is the SEEN arm of Table 1, which withholds story 3 from pretraining (30,270 against 55,753 segments) and calibration and evaluates story 0 only, so contrasts involving it also carry those pretraining, calibration and test-story changes (Table 2; dynamic − static in rank percentile, n=10 held-out subjects paired across all four cells). The pools differ up to sevenfold between cells, so ×chance is not comparable across them and is not shown. The inference in the text is carried by the paired main-effect contrasts, not by which cells cross a threshold: at k=500 the split main effect is -0.017 $[-0.031, -0.001]$ and the guard main effect -0.003 $[-0.014, +0.007]$; at k=100 neither resolves at n=10. Since each cell holds dynamic − static, the split main effect is the architecture×split interaction (per-subject difference of the advantage between the two splits, averaged over guards; `scripts/interaction.py`, receipt `results/interaction_receipt.json`). The three-way guard×split term on that advantage is unresolved at both budgets ($+0.016$ $[-0.047, +0.078]$ at k=100; -0.027 $[-0.061, +0.004]$ at k=500). Per guard, the split simple contrast at k=500 is -0.004 $[-0.017, +0.010]$ with the recording guard (interval includes zero) and -0.030 $[-0.058, -0.001]$ with the stimulus guard (excludes zero); at k=100 both include zero (-0.030 $[-0.078, +0.011]$; -0.014 $[-0.081, +0.055]$). Three further caveats in full: the split-minus-guard *difference* of main effects (-0.014 as a point value) is itself an unresolved contrast, so we do not order the two fixes; that the guard-only cell retains a resolved advantage while no split cell resolves positive is a comparison of a star with a non-star and is suggestive only; both k=500 stars (the standard-protocol premise $+0.014$ $[+0.001, +0.026]$ and the split main effect) exclude zero by margins smaller than plausible small-sample undercoverage of a percentile bootstrap at n=10, and the pre-specified criterion's per-budget family carries no correction, so the premise rests on k=100, where the rank and as-specified ×chance intervals both survive a t_9-width check; and mean percentile is invariant to pool *size*, not *composition*—a story-disjoint cell also loses the cross-story shortcut—so the split main effect prices the protocol change as a whole, not exposure separately from the pool it reshapes.

The grid suggests a larger split contribution, but the split-minus-guard contrast does not resolve. With the stricter guard but the original split, the point estimates remain close to the standard recipe rather than reduced by it ($+0.034$ and $+0.024$ against $+0.043$ and $+0.014$; significant at k=500, unresolved at k=100). With the original guard but a story-level split, the advantage is no longer distinguishable from zero at either budget. We say "no longer detectable" rather than "removed": these are intervals covering zero at n=10, bounding the residual advantage below a few hundredths of rank percentile without establishing that it is absent. The paired main-effect contrasts in Table 2 are the quantitative form of this paragraph; at k=100 they do not resolve, and the sentence above about the guard rests on k=500.

The point estimate under the stricter guard is somewhat higher than under the original one at k=500 ($+0.024$ versus $+0.014$), but those intervals overlap and we do not read a mechanism into that. Making the split story-level while leaving the guard scoped to the recording moves the point estimates toward zero at both budgets without resolving either cell; only the joint change, whose cell also withholds story 3 from pretraining, yields a point estimate at or below zero at k=500.

Post-hoc small-sample sensitivity. From the same ten subject contrasts, a t_9 interval for the k=500 split interaction is $[-0.0355, +0.0017]$; enumerating all 1,024 sign flips gives two-sided p=0.0625. For split minus guard the interval is $[-0.0304, +0.0022]$, p=0.0859. At k=100 the split interval is $[-0.0815, +0.0372]$, p=0.4219. The t calculation treats subject contrasts as independent; sign flips additionally require a symmetric null. These are sensitivity checks, not replacement confirmatory tests, and do not establish absence. `scripts/interaction_sensitivity.py` reads the shipped `results/interaction_receipt.json` and regenerates `results/interaction_sensitivity.json`. The two fixes enforce different constraints; these data do not establish that one is statistically superior to the other.

D Per-Dose Values of the Exposure Dose-Response

fraction retaining the evaluation story	0.00	0.25	0.50	0.75	1.00
k=500	0.684	0.805	0.808	0.832	0.854
k=2000	0.700	0.825	0.846	0.877	0.889

Percentile rank, static encoder, n=4 held-out subjects per dose, paired. Column heads are nominal doses: the number of the nine pretraining subjects that keep the story is rounded to a whole subject, so the retained fractions are $2/9$, $4/9$ and $7/9$ (round-half-to-even takes 4.5 to 4). Every dose is backfilled to exactly 55,753 pretraining segments, so volume is pinned by construction rather than controlled after the fact. Endpoint ranges quoted in Appendix H ($+0.170$, $+0.188$) are computed from unrounded percentiles; the rounded cells above imply 0.189 at k=2000.

E Split Sensitivity in Narratives fMRI

On Narratives fMRI (Nastase et al., 2021), under a plain cross-subject design with no pretraining stage, an analogous held-out-material manipulation lowers retrieval in the same direction. Training on the evaluation segments themselves gives $\times$1.94 chance (pool 50.8); holding those segments out but keeping the story, $\times$1.25 (pool 47.9)—these two share training-segment counts and pools within 6%, so holding out the evaluation segments retains roughly a quarter of the above-chance signal. Withholding the narrative entirely leaves an interval covering chance ($\times$1.10, CI $[0.85, 1.41]$; its pool is 166, threefold larger, so we read only the covered-chance interval from it, not its $\times$chance against the others). That arm clears chance at exactly one of seven lags—lag 0, zero haemodynamic delay, which is what a favourable multiple comparison looks like. With no pretraining stage this says nothing about pretraining exposure; it establishes that cross-subject scores move with whether the evaluation material was fitted through other people's brains.

F Additional Reference-Pipeline Controls

Two controls on the reference pipeline of §5, in full.

A published ingredient under the same audit. Removing the reference decoder's subject-specific input layer (`subject_layers`, on in the shipped configuration) costs 26.7 points under the shipped protocol (57.80 vs 31.07), 26.3 at matched volume (51.80 vs 25.54), and 18.4 under the story split (47.28 vs 28.89; one run per ablated cell against three-seed arm means, pool 7,140). A published ingredient we tested therefore does *not* flip under the corrected protocol—its advantage deflates by roughly a third, and the matched-volume point comparison does not support volume as the explanation. Protocol correction can remove a resolved advantage (§4) or merely deflate one; which published comparisons flip is an open question.

The brain-free arm does not move. The same manipulation on the word-onset clock—same corpus, code and 7,140 pool, each arm at its best-validation checkpoint—gives 89.52% $\rightarrow$ 91.79%, or 102.5% of its above-chance top-10 signal: no loss observed (one story-split run against three block-split seeds, 89.52–92.00%). The decoder loses and the clock does not, as the mechanism predicts: timing is not story-specific, so a clean split takes nothing from a clock; a decoder does gain, and loses it. Correcting the protocol therefore widens the clock's lead from 31.1 points to 44.1. A matched-volume clock arm on arm B's realized 120,389-example subset scores 90.31%

(one run; 0.06 below the three-seed mean 90.36): no volume cost is observed, consistent with—though one run inside a 2.5-point seed spread cannot establish—the story-split gain not being a volume loss masked by a split gain. This bears on the convergence caveat only weakly: a task at 91.8% has less headroom than one at 47.7%, and a clock's insensitivity near ceiling says nothing about how a harder decoder responds; for the decoder that separation is arm B.

G Extended Discussion of Adjacent Audits

Kilgallen et al. (2025) show that repeating stimuli within a subject inflates EEG decoding accuracy and identify a set of affected publications; Li et al. (2021) is the canonical case, where block design rather than stimulus content drove a celebrated EEG decoding result. The cross-subject case is the same mechanism at a different scale: instead of one subject seeing a stimulus twice, the other subjects hear it before the held-out one is evaluated. Schaeffer et al. (2026) give the design precedent from the language-model side, pretraining models on mixtures while sweeping the number of test-set replicas; they vary contamination *dose*, whereas we hold pretraining volume fixed and remove a single stimulus. Related audits of biosignal foundation models identify adjacent but distinct failures: subject-identity shortcuts (Lin et al., 2026) and, for time-series foundation models generally, pretraining-corpus overlap catalogued, with only indirect overlap tested in a controlled design (Meyer et al., 2025).

In iEEG benchmarks the guards are temporal and cross-session, not pretraining-level. Neuroprobe (Zahorodnii et al., 2025) cross-validates its within-session split on non-overlapping movie segments to avoid *temporal* leakage—the adjacency constraint—and its cross-session split tests on a different movie, which is stimulus-disjoint within a subject; but its pretrained baselines were pretrained on the benchmark's own corpus, a point the DIVER-1 authors make (Han et al., 2025) and the benchmark does not discuss. DIVER-1 (Han et al., 2025) pretrains on external recordings and reports, with architecture fixed and corpus size matched, that in-corpus pretraining scores higher than external pretraining; it frames this as distributional alignment, not contamination, and it is corpus-level—same subjects, same recordings—where ours is stimulus-level within a corpus whose subjects are shared by design. In fMRI reconstruction, Shirakawa et al. (2025) show that stimulus-content overlap between training and test can manufacture apparently successful decoding, and corpus aggregations such as MOSAIC (Lahner et al., 2025) now deduplicate stimuli across subjects and datasets as hygiene; neither measures what the overlap is worth, and neither touches pretraining.

H Extended Results and Caveats

This appendix keeps the full numerical detail behind §4 and §5; the body tables carry the headline values.

Headline table, estimator notes. The retained fraction is the mean over subjects of $(\text{unseen}_i - 0.5)/(\text{seen}_i - 0.5)$; the ratio of averages would differ only at k=100 (67.7% for 71.7%; within 0.2 points elsewhere). n varies: story 0 went from five to ten subjects after an interim look (estimates moved under a point); one subject lacks a k=2000 curve (n=9). At k=100 the retained-fraction intervals are on average three-and-a-half times wider than at the larger budgets (under twice to over eight, row by row). The story-count-matched pair across three narratives gives 54.1%, 51.1% and 57.6% retained at k=500 (mean 54.3%); the cross-story contrast (story 3 withheld against story 0 seen) uses n=9–10. Over three pretraining seeds the story-0 static retained fraction averages 53.4% and 53.2% at k=500 and k=2000. The from-scratch encoder sits at percentile 0.500 and 0.504 at k=100 and k=500.

Dose-response endpoints. The comparisons above are endpoints, so we also vary the *amount* of exposure: keep a fraction of the other subjects' recordings of the evaluation story, resample the rest up to a fixed count—every dose backfilled to exactly 55,753 pretraining segments, so volume, the subject set and optimisation are pinned and only exposure and the backfilled mix of the other stories vary. Retrieval rises monotonically with exposure at both budgets (per-dose values in Appendix D)—subject-level rank correlation +0.80 at k=500 and +0.92 at k=2000 (n=4 subjects per dose, reported descriptively), over an endpoint range of +0.170 and +0.188. Because every dose is backfilled to the same segment count, volume cannot explain the gradient—by measurement, not

assertion: the identical sweep *without* the backfill (segment count falling 8.3% from full exposure to none) gives $+0.184$ at k=2000 against $+0.188$ pinned, and loses the k=500 curve's monotonicity, which the pinned sweep keeps.

What survives, all budgets. The unseen condition still scores percentile 0.66–0.71 against 0.5 chance on a narrative no brain in the training set has heard, so above-chance signal remains. Against a from-scratch encoder trained on the same k segments, with learning rate (3e-5 to 3e-3) and weight decay (1e-2 to 1) selected on validation, pretraining raises $\times$chance from $\times$0.95, $\times$1.09 and $\times$2.47 to $\times$3.42, $\times$4.15 and $\times$5.02 at k=100/500/2000—*both* in the UNSEEN condition, same story and pool, so the comparison is untouched by the mechanism this paper measures (n=5, the subjects with both arms); the selected rates fall in the grid's interior, so the search brackets the optimum. Cross-subject pretraining transfers something per-subject training does not recover—at roughly half the magnitude the standard protocol reports.

Calibration-level overlap, as originally run. The shared stimulus enters the pipeline twice: at *calibration*, when the segments used to adapt to a held-out subject come from the same stories as the evaluation segments, and at *pretraining*, through the other subjects. The magnitude result above concerns the second; the architecture comparison concerns the first. Under the protocol as originally run—guard scoped to the recording, calibration and evaluation drawn from the same stories—the dynamic encoder (residual-free, endpoint-augmented after pooling; App. I) beats a parameter-matched (within 8.4%) static encoder at both low budgets. The criterion was fixed before the deciding sweep (App. I)—in $\times$chance, and read $+8.26$ (95% CI $[+2.40, +14.19]$, n=10) at k=100; because every protocol change also changes the pool, we analyse the comparison in the pool-size-invariant rank metric throughout, where the same advantage is $+0.043$ $[+0.009, +0.081]$ at k=100 and $+0.014$ $[+0.001, +0.026]$ at k=500. (Comparing $\times$chance values across protocols would mislead: the pools differ sevenfold between these conditions, so roughly two fifths of the apparent $+8.26 \rightarrow +0.58$ collapse is the pool moving rather than the model—the same class of error §5 reports in the reference implementation. All cross-protocol numbers below are rank differences.) Scoping the guard to the stimulus and making calibration story-disjoint leaves the advantage undetectable: $+0.020$ $[-0.019, +0.055]$ at k=100 and -0.006 $[-0.022, +0.011]$ at k=500, in a cell that also withholds story 3 from pretraining; the story split alone, with every story in pretraining, gives $+0.013$ $[-0.011, +0.037]$ and $+0.010$ $[-0.003, +0.024]$. The first cell's k=100 interval contains the standard-protocol $+0.043$ (the story-split cell's does not; not a paired test), and no corrected cell excludes an advantage merely shrinking with the removed signal: these cells establish that nothing resolves positive, not that the difference is zero. That cell changes guard, split and pretraining together; the factorial separates the guard from the split.

Multiplicity for the four-stimulus grid. Under a Bonferroni correction over the four withheld stimuli, three of the four k=2000 reversals remain resolved; two rows have five subjects, where the 98.75% bootstrap tails are thinly supported.

The reference candidate pool is split-dependent. The implementation uses $\min(10{,}000, N_{\text{queries}})$ candidates: the original pool is 7,140, while story splitting can reach the 10,000 cap and change chance. We pin pool size to 7,140, but fitted normalization still makes membership run-dependent, so this is not yet a common-membership comparison. The patch, retained-query counts, timing-arm seed ranges and their early-stopping limitation are reported in Appendix B.3.

Correcting the pipeline's own number. The pipeline assigns blocks to splits by hashing each block's word sequence; on MEG-MASC a story spans many blocks, which hash independently and land in different splits, so the model is fitted on parts of a narrative and scored on other parts of the same narrative. Blocks are disjoint—what "stimulus-disjoint splits" is usually taken to mean—but the story is shared. Story-level splitting keeps every narrative wholly within one split. Holding the candidate-pool *size* at 7,140 in every condition—verified in each run's log—and reading each at its best-validation checkpoint, guarded top-10 falls from 58.41% to 47.69%: 10.72 points, an 18.4% loss of the above-chance signal. Chance is 0.140%, so what remains is a working decoder—overstated, not collapsed. Two disclosures travel with the number: pool size is pinned but pool *composition* is not—A and B are scored on the block-level test partition spanning all four narratives,

C on one whole narrative, and a single-narrative pool removes a story-identification shortcut a mixed pool admits (§4 scores both arms against the same pool for exactly this reason; here it can only be stated)—and with four narratives the story-to-split assignment is a discrete choice: the split seed was selected, and disclosed as selected, to yield a non-degenerate 1/1/2 partition, fixing *The Black Willow* as the evaluation narrative. The corrected number is for that assignment.

Point estimates put most of that gap on training volume, not the shared story. The corrected split does two things at once, so rather than bound the confound we refit: a third arm keeps the block-level split and thins only the training set to the story split's size, validation and test untouched so the pool cannot move. Arm A (block split as shipped, 203,152 train examples) reaches 58.41%; arm B (block split, train thinned to 120,389, 1.8% above C) 51.96%; arm C (story-level, 118,310) 47.69%. The contrasts are not perfectly clean: B's random thinning keeps 58% of *all* training blocks, including the same-narrative ones, so $A{-}B$ is the data cut plus the removed 42% of exposure, and $B{-}C$ the remaining exposure plus a composition change (B trains on blocks of four narratives, C on two whole ones) plus a selection change (A/B pick checkpoints on validation sharing stories with their test, C on a story-pure one)—the latter two, if anything, inflate $B{-}C$. With that labelling: $A{-}B = 6.45$ points and $B{-}C = 4.27$ of the 10.72-point gap on the original seeds. At three seeds per arm the *ordering* $A > B > C$ holds in every observed seed with no overlap between arms (ranges 1.20/0.38/0.80; means 57.80/51.80/47.28); as conservative cross-seed bounds, $A{-}B = +6.01$ $[+5.25, +6.83]$ and $B{-}C = +4.51$ $[+3.89, +5.08]$, so the volume share of the gap is 57% with bound $[46\%, 72\%]$—the larger share sits on the discarded data at every point estimate while the bound still admits an even split, and we say both. Attributing the 10.72 points to shared stimulus alone would roughly double the estimates, which put the shared story's share at about two fifths (43%, bound $[34\%, 53\%]$, computed separately, so not the volume bound's complement)—an upper bound, the direct full-dose point being 28% ($3.00/10.72$), and likely an overestimate, $B{-}C$ still carrying the composition and selection-criterion changes.

The one-bit exposure arm. The decomposition's residual can be isolated: two further arms share the validation story, byte-identical evaluation blocks, a 7,140-query pool and training volume matched to within about 1% (119,502 vs 118,310 examples), and differ in one design bit: whether the evaluation narrative's other blocks are in the training set. Three qualifications define what the contrast measures. (i) The volume match thins D's whole training set uniformly—*including* the added narrative—so D carries about 59% of the narrative's train blocks ($\sim$49,600 of 84,050): the contrast prices a 0.59-dose exposure; the dose ladder below measures the full-dose exposure directly ($+3.00$). (ii) The pool cannot be identity-matched within the pipeline's own evaluation path: the retained-query set lies downstream of each run's own fitted normalization (the amplitude-reject mask applies the run-fitted scaler)—a third silent pool determinant beyond the split and the cap. Re-scoring the six seed-matched checkpoints under an independent pool draw moved no value by more than 0.02 points (three of six by 0.00), so the observed draw sensitivity—one independent redraw per checkpoint—is at most 0.02 points, where the prior assumption was near half a point. (iii) Checkpoints are selected on the third story's validation loss; in all twelve mixed-protocol logs (six seed-matched runs, two thinning realizations, two dose arms, the second-narrative pair) the selected checkpoint is also the best logged evaluation value, so no selection loss is observed. At three model seeds per arm plus two further thinning realizations of D (five D, three C$'$), the arms separate completely: fifteen of fifteen cross-arm differences positive. An exact one-sided permutation test would give $p = 1/56 \approx 0.018$, but only if the eight runs were exchangeable under no effect, and the design does not guarantee that: the three original D runs share thinning realization 0, the two added D runs change model seed and realization together, no C$'$ run was added to match them, and the realization visibly moves D (below). We therefore rest the claim on the separation and the range and read the p value as conditional on that assumption. The realization-0 chain pairs give $+1.83$/$+2.89$/$+3.75$; the new realizations land at the low end of the D range (47.61/47.66 vs 48.38–49.23), so the thinning lottery contributes visibly to spread—measured, no longer assumed—and the cross-arm span is $+1.01$ to $+4.59$, mean $+2.40$ (no stable t interval at these n; we quote separation and range). At the measured dose the exposure is 9–43% of the raw 10.72-point gap across runs (22% at the mean), consistent with $B{-}C = 4.51$ mixing exposure with the composition change; every observed run falls short of arm B's 6.45-point price, and a two-arm dose ladder settles the full-dose question by measurement: at dose 0.30 the exposure is $+0.77$, at dose 1.0 it is $+3.00$ (volume matched to within 0.1%; single runs), the external dose-response is monotone at four doses ($45.86 \to 46.63 \to 48.26 \to 48.86$; Figure 3), and the full-dose exposure is about half the fix's vol-

ume cost (2.15× against the original-seed $A-B$ of 6.45, 2.0× against the three-seed mean of 6.01; 1.4–6.4× the individual 0.59-dose runs; the two differences are scored on different test blocks). A second assignment—fixing *Easy Money* as the evaluation narrative instead—replicates the pricing's sign: its full-dose exposure arm gains +2.00 over its own dose-0 arm (55.133 vs 53.137; single runs, pool 7,140, volume matched to −0.39%), against +3.00 on *The Black Willow*—same sign, a different observed magnitude, and absolute levels not comparable across narratives (different single-narrative pools), which is why we read the sign first. Two things the corrected number is not: it is not an estimate of the exposure alone—volume falls together with the split (one test, one validation, two training stories at four stimuli), and arm B prices that at 6.45 of the 10.72 points—and it is not §4's loss of about half the signal: that figure is *pretraining*-level—the evaluation narrative reaching the model through nine other people's recordings—where this one is *calibration*-level, parts of a narrative reaching the model through the same recordings it is scored on, in a pipeline with no pretraining stage; the ordering—a story seen through nine other brains is worth more than other blocks of the same story—is what one would expect. Five D and three C′ runs support this one too: separation and range, no formal interval.

Two further controls (in full in Appendix F). The reference decoder's subject-specific input layer does *not* flip under the corrected protocol (its advantage deflates from 26.7 to 18.4 points): correction can remove a resolved advantage or merely deflate one. And the word-onset clock does not lose under the story split (89.52% → 91.79%; matched-volume arm 90.31%): timing is not story-specific, so a clean split takes nothing from a clock.

H.1 Calibration-diversity control

At equal k the UNSEEN arm of Table 1 calibrates on stories $\{1, 2, 3\}$ and SEEN on $\{1, 2\}$. A MATCHED arm keeps UNSEEN's pretraining ($\{1, 2, 3\}$) and calibrates on SEEN's stories only; with the same seed it draws SEEN's validation segments and first-k calibration subsets, and every subject's evaluation set and pool are identical across the three arms. Its pretraining is a fresh run with a different 1,500-segment pretraining-validation split, a realisation difference of the size of the seed rows of Table 1; the historical recipe itself reproduced the published SEEN files exactly in a rerun (below). D is the matched minus the published retained fraction, per subject, in points; intervals are paired subject bootstraps (20,000 draws). The decision rule (a 10-point margin on D for the static encoder at k=500 and 2000) was fixed before any run; verdict: NO MATERIAL CHANGE.

row	k	n	SEEN	UNSEEN	MATCHED	published	matched	D (points)
story 0, static	100	10	0.779	0.689	0.662	71.7%	61.5% [50.4, 72.3]	−10.2 [−18.7, −2.5]
story 0, static	500	10	0.861	0.695	0.693	54.1%	53.3% [51.2, 55.5]	−0.8 [−3.8, +2.1]
story 0, static	2000	9	0.884	0.707	0.701	53.9%	52.3% [50.1, 54.7]	−1.6 [−3.7, +0.3]
story 0, dynamic	100	10	0.799	0.677	0.675	60.7%	60.7% [51.6, 71.3]	0.0 [−6.6, +6.7]
story 0, dynamic	500	10	0.855	0.689	0.691	53.5%	53.8% [51.4, 56.1]	+0.3 [−2.9, +3.6]
story 0, dynamic	2000	9	0.878	0.690	0.685	50.3%	48.9% [46.4, 51.4]	−1.3 [−5.2, +2.3]
story 1, static	100	5	0.804	0.679	0.659	61.7%	53.9% [39.4, 70.4]	−7.8 [−19.4, +0.2]
story 1, static	500	5	0.852	0.681	0.698	51.1%	56.4% [53.6, 59.3]	+5.4 [−2.9, +15.6]
story 1, static	2000	5	0.881	0.706	0.701	54.0%	52.8% [50.9, 54.9]	−1.2 [−3.7, +1.4]
story 1, dynamic	100	5	0.825	0.673	0.662	55.6%	52.6% [36.3, 69.1]	−3.0 [−13.5, +5.9]
story 1, dynamic	500	5	0.872	0.691	0.674	51.5%	47.1% [41.3, 51.9]	−4.4 [−9.8, +0.9]
story 1, dynamic	2000	5	0.876	0.695	0.684	52.0%	49.1% [47.1, 50.8]	−2.9 [−4.7, −0.9]

Story-1 rows (descriptive, n=5) use the same construction with SEEN calibrating on $\{0, 2\}$ and UNSEEN on $\{0, 2, 3\}$. Pre-registration: `notes/PREREG_calibmatch_20260924.md` (commit `8994b23`); receipts: `results/calmatch_20260924/analysis.json` (SHA-256 `a90df8e428c4`) with per-run preflight and receipt files. Reused runs from an earlier calibration-matched design (static encoder, k=100/500/2000; descriptive): the calibration-matched cross-story contrast (story 3 withheld vs story 0 seen, both calibrated on $\{1, 2\}$) retains 49.5% [47.5, 51.5], 47.2% [44.4, 50.5], 47.1% [44.9, 49.3] (n=5); a rerun of the SEEN configuration against the published SEEN arm retains 100.0% [100.0, 100.0], 100.0% [100.0, 100.0], 100.0% [100.0, 100.0] (n=5); its result files are identical to the published SEEN files for all five subjects and both encoders.

I MODELS, TRAINING AND COMPUTE

Every internal result (§3–4) is produced by one script, `src/train_mb.py`; the two encoders live in `src/tokenizers.py`, the shared target head and the retrieval metrics in `src/train_infonce.py` and `src/eval_metrics.py`. These four files, the segment builder `src/train_m1a_seg.py` and `src/extract_contextual.py` ship in the supplement's `src/` in the versions that produced the reported numbers. The hyperparameters below are that script's defaults. The launch scripts (`scripts/slurm/mb1*.sbatch`) override only the input resolution (`--decim 1`, `--seg_s 1.0`, `--min_words 1`), the budget list, the random seed in the seed replicates, and the protocol flags (withheld story, evaluated story, guard scope, split level, dose); no launch script changes the optimiser, learning rates, batch size, schedule lengths or architecture, with one exception noted below (the tuned from-scratch grid).

Input representation. MEG-MASC word-locked epochs: 208-channel KIT MEG, epoched $[-0.2, +0.8]$ s around each word onset at 250 Hz (250 samples), baseline-corrected and z-scored, following the MEG-MASC release conventions (Gwilliams et al., 2023). The model input is samples [50:250] of each epoch—word onset to $+0.8$ s—at full resolution in every reported run, so one word contributes a 208×200 (channels $\times$ samples) array x (the script's *default* decimates time by 2; every reported run, and every `mb1` launch from `mb1d` on, passes `--decim 1`; the three earlier pilot scripts use the defaults). Words are grouped into segments by flooring onset time into 1.0 s bins within a recording (default 3.0 s / $\geq$ 3 words; all reported runs use 1.0 s / $\geq$ 1 word); a segment's MEG input is the mean of its member words' windows, and its target is the mean of their embeddings. Targets are contextual GPT-2 embeddings: each sentence is reconstructed from the metadata and forwarded through frozen HuggingFace `gpt2` (Radford et al., 2019; Wolf et al., 2020) (124M, d=768), and the layer-6 hidden state at the last sub-token of each word is stored in fp16 (`src/extract_contextual.py`). GPT-2 is frozen throughout; the retrieval space is not—see the target head below.

The two encoders. Paper name $\rightarrow$ code variant: "static conv" = `conv` (class `BrainEncoder`); "dynamic" = `dynamic_big` (class `DynamicTokenEncoder`). Both share the same trunk shape: input BatchNorm over the 208 channels, a 1×1 convolution $208 \rightarrow 320$ with GELU (learned spatial mixing), five dilated temporal convolutions at 320 channels with kernel 3 and dilations $1, 2, 4, 8, 16$ (sequence length preserved), a mean-pool over time, and an identical projection head $\mathrm{Linear}(320\rightarrow768)$–GELU–$\mathrm{Linear}(768\rightarrow768)$. They differ in exactly two ways:

- **Static conv.** The five temporal convolutions are residual blocks, $h \mapsto h + \mathrm{GELU}(\mathrm{BN}(\mathrm{Conv}_{3,d}(h)))$. The pooled feature is the time-averaged latent state.
- **Dynamic.** The five temporal convolutions are stacked without residual connections (Conv–BN–GELU), giving a latent trajectory z of 320 channels by T steps. The encoder then forms the one-step finite difference $\Delta z_t = z_t - z_{t-1}$ (zero-padded at the first step), concatenates $[z; \Delta z]$ (640 channels), projects back to 320 channels with a 1×1 convolution, and mean-pools. The explicit difference branch reduces to an endpoint contrast: with the first difference padded by zero, $T^{-1}\sum_t \Delta z_t = (z_T - z_1)/T$. Since the flow projection is linear and its output is pooled before the head's GELU, the pooled feature is exactly $W_z\bar{z} + W_\Delta(z_T - z_1)/T + b$. Thus "dynamics-aware" is the code variant's historical label, not a claim to retain a full trajectory: its nonlinear temporal trunk is augmented by endpoint displacement. The residual-block change and this augmentation are the two differences from the comparator.

The codebase also contains a vector-quantised variant (`dynamic_big_vq`: a 256-code VQ-VAE codebook of dimension 320 with straight-through estimator and commitment loss β=0.25, inserted before pooling) and a velocity-removed control (`dynamic_big_noflow`); neither appears in this paper's tables. An early exploratory run also used a smaller patch-transformer named `static` in the code (1.75M parameters); the "static" encoder of every *reported* comparison is the conv encoder above, not that model. Both arms share the target head `LLMHead`—LayerNorm(768), $\mathrm{Linear}(768\rightarrow768)$, GELU, $\mathrm{Linear}(768\rightarrow768)$; 1.18M parameters—applied to the frozen GPT-2 vector and trained jointly with the encoder in every phase, plus one learnable temperature (initialised to $1/0.07$, clamped at 100).

How "matched-parameter" is achieved, and the counts. The first dynamic encoder was 0.97M parameters against conv's 2.45M, so the initial comparison confounded architecture with capacity. The comparison in this paper uses `dynamic_big`, whose width and depth are set equal to conv's (d_{lat}=320 matching hidden=320; five temporal layers matching five blocks). Exact counts at n_{ch}=208, derived from the layer definitions (and printed per run by `train_mb.py`, logged in each result JSON as `params_M`):

component	static conv	dynamic
input BatchNorm (2×208)	416	416
spatial 1×1 conv ($208\cdot320 + 320$)	66,880	66,880
$5\times$ temporal conv ($320\cdot320\cdot3+320$ + BN 640)	1,540,800	1,540,800
flow projection 1×1 conv ($640\cdot320 + 320$)	—	205,120
head ($320\cdot768+768$; $768\cdot768+768$)	837,120	837,120
total	2,445,216	2,650,336

The two totals differ only by the flow projection: +205k parameters, +8.4%, in favour of the dynamic encoder—i.e. the mismatch runs *against* the direction of this paper's conclusion. The shared target head and temperature are identical in both arms and excluded from these totals.

Pretraining. Objective: symmetric InfoNCE (van den Oord et al., 2018). Within a batch, encoder outputs and head-projected targets are L2-normalised; the loss is the mean of the two cross-entropies over the scaled similarity matrix (brain→text and text→brain), plus the VQ commitment term where a VQ variant is used (identically zero for both reported encoders, so their losses are the same function). Data: all segments of the nine non-held-out subjects (ten of MEG-MASC's 27 subjects are used), minus the withheld story where the protocol says so: 55,753 segments with all four stories (the dose arms backfill by resampling to exactly that count); with one story withheld, 30,270 (story 3), 39,850 (story 2), 45,999 (story 1) or 51,143 (story 0); 25,660 with stories 0 and 3 withheld (first held-out subject; other held-out subjects differ by at most 6%). Optimiser: AdamW (Loshchilov and Hutter, 2019), lr 3e-4, weight decay 1e-2, cosine annealing over 80 epochs, batch 256 (leftover batches of fewer than 8 items are dropped), fp32 throughout, no mixed precision or augmentation. Model selection: after each epoch, unguarded top-10 retrieval on a validation set of 1,500 segments held out of the pretraining pool (never the held-out subject); the best-validation checkpoint is restored at the end. All 80 epochs always run; "early stopping" is checkpoint selection, not early exit. Seed 0 everywhere except the seed replicates (seeds 1, 2), where the seed drives initialisation, batch order and the calibration draw together.

Calibration at k. The held-out subject's recordings are split into a calibration side and an evaluation side—by story under the corrected protocol, by recording under the standard one; in the stimulus-held-out arms the evaluation side is the evaluated story and calibration draws only on stories that are neither evaluated nor withheld. From the calibration side, 300 segments are removed as the subject-level validation set; the remaining pool is permuted once (seeded) and the first k segments form the budget, so budgets are nested across k and identical for both encoders. A budget exceeding a subject's pool is dropped, which is why one subject leaves the k=2000 rows (n=9). Fine-tuning (the reported transfer arm): reload the pretrained encoder, head and temperature, and fine-tune *all* parameters on the k segments—AdamW, lr 5e-5, weight decay 1e-2, 120 epochs, cosine schedule, batch 256, best-validation checkpoint on the 300 segments. From-scratch baseline: a freshly initialised encoder and head trained on the same k segments with the same 120-epoch recipe at lr 3e-4; in the tuned-baseline runs the grid lr $\in$ {3e-5, 1e-4, 3e-4, 1e-3, 3e-3} $\times$ wd $\in$ {1e-2, 1e-1, 1} is swept, the configuration is selected on validation top-10 only, and that configuration is evaluated once. The script also scores the pretrained encoder before any calibration; those numbers do not enter this paper's tables.

Evaluation. The candidate pool is every evaluation-side segment of the held-out subject—nothing is sampled and no cap is applied, which is why pool size moves with the protocol (§4). Similarity is the cosine between the L2-normalised encoder output and the head-projected frozen GPT-2 layer-6 target. Adjacency guard: for query i, candidates within ±10 s of i's mean onset in the same *recording* are masked; under the stimulus-scoped guard the mask is keyed instead on story identity and time-within-story (onsets re-zeroed per presentation), so the twin segment in the

other session of the same story is masked too. The true target is never masked. From the resulting ranks: top-10 accuracy, ×chance (top-10 × pool/10), median rank, and mean percentile rank $1 - (\text{rank} - 1)/(\text{pool} - 1)$—the pool-invariant statistic used for every cross-protocol comparison (chance $= 0.5$). Each run reports a 1,000-draw query bootstrap and a shuffled-target control (targets permuted; sits at chance throughout). The intervals quoted in the paper are paired subject-level bootstraps computed downstream by `scripts/retained_fraction.py`, `scripts/reversal_inference.py` and `scripts/mb_conditions.py` from the per-run JSONs.

The word-onset clock arm. The clock replaces the brain input with a word-onset raster of the same shape as the MEG tensor: zeros everywhere, with the sample-rate-aligned onset index of each word marked, built from the same stimulus alignment the protocol already assumes. Architecture, features, training procedure, splits and evaluation are byte-identical to the reproduced decoder (`BM_TIMING_ONLY`, an env-gated block inserted by hand into the pipeline's data loader; its text is in the supplement, `RELEASE.md` §3); the decoder therefore sees stimulus timing and nothing else. Its 89.5% top-10 is the reproduction of the released pipeline scored with that one substitution. The matched-volume clock arm applies the same substitution inside arm B's environment (block split, training thinned to 120,389 examples—1.8% above arm C's 118,310 after per-recording rounding—with the thinning seed arm B used, pool 7,140); it reached its best validation loss at epoch 7 and patience-stopped at epoch 17, scoring 90.31% at that checkpoint.

Compute. Every internal run used a single NVIDIA H200 (Slurm `--gres=gpu:1`, 8 CPU cores, 200 GB RAM; two H200 partitions of a university HPC cluster). One subject×condition run—two encoders pretrained for 80 epochs plus, at each of three budgets, a fine-tuned and a from-scratch arm at 120 epochs each—completes within a 2-hour wall-clock cap; the tuned-baseline run with its 15-configuration grid takes about 45 minutes. The MB-1 series comprises roughly 195 such single-GPU runs across all conditions, subjects and seeds; `sacct` accounting over the named production jobs totals 41 GPU-hours, excluding exploratory runs before 2026-07-28. The reference-pipeline reproduction of §5 ran on the same cluster with 2 H200s per job; `sacct` accounting over all reproduction, story-split, decomposition and one-bit arms totals approximately 200 GPU-hours. Software: Python 3.11, PyTorch 2.6.0+cu124 (Paszke et al., 2019), HuggingFace Transformers 4.57.1 (Wolf et al., 2020); the environment file itself is not in the supplement.

Criterion fixed in advance. The architecture hypothesis (a dynamics-aware encoder transfers better than a static one in low-shot cross-subject transfer) and its gate (the dynamic model clearly beating the static one at small k, in ×chance) were recorded in the program document on 2026-07-08, where the static arm was a patch transformer. Single-subject pilots (subject 01, 2026-07-08) then excluded that arm (it did not optimise). On the script's default input (3-s segments, time decimated by 2) the reported pair showed no consistent zero-calibration difference on subjects 01–03 (dynamic ahead on two of three, 2026-07-08); at the full-resolution 1-s input of every reported run the dynamic encoder led on all three (2026-07-09). The pair and that input were fixed after these pilots (2026-07-09), before the deciding 10-subject sweep was launched (2026-07-26). Before that sweep the same pair, on the same input, had also been run on the same ten subjects at zero calibration and at k=500 only (2026-07-09/10): the dynamic encoder was ahead at zero calibration for all ten and at k=500 for seven of ten (mean percentile 0.880 against 0.861), and the three pilot subjects are among the ten; the budget curve on which the gate is judged, including k=100, was first run in the deciding sweep, which is why the standard-protocol premise rests on k=100 (§4). The notebook entry of 2026-07-30 applies the gate to the deciding sweep's job record as the paired per-subject dynamic−static contrast at each calibration budget, judged by whether its bootstrap interval excludes zero, and records that an interim single-subject reading had pointed the other way (the entry's interval, $[+2.05, +14.46]$, came from an earlier bootstrap run; the reported $[+2.40, +14.19]$ is the seeded 20,000-draw recomputation by the released script). The stimulus-scoped guard used here was itself adopted under a pre-committed decision rule for an earlier analysis in this project (notebook entry of 2026-07-29 and a decision record), before any job under that guard was submitted. Anonymised excerpts of these entries, the program document's hypothesis and gate, and the decision record are in the supplement's `prespecification/`; the full notebook is not released.

J RELEASED ARTIFACTS

An anonymised subset for review accompanies the submission as supplementary material: both audit tools, the corrected split with its manifests, the pool-pinning and other pipeline patches, the model code, the per-table regeneration scripts and run configurations, the receipts behind Table 2 and Figures 2–3, and anonymised excerpts of the lab notebook and program document that fix each criterion in advance (`RELEASE.md` is the inventory). A public code repository will be released upon publication.

Two tools, both stdlib-only and both usable without our models or results.

`stimulus_leakage_check.py` audits a split you already have: given a manifest of (subject, stimulus, split) it reports what fraction of the evaluation material appears in the fitting partition, distinguishes *held-out subjects with shared stimuli* from *no shared stimuli*—the distinction a subject-level split hides—and exits non-zero on overlap so it can run in CI.

`leakage_report.py` audits a *corpus*, before any model exists. From the metadata alone it reports what a conventional cross-subject protocol would share between fitting and evaluation, and what the corrected protocol costs. On MEG-MASC it recovers the design constraints this paper had to discover experimentally: $4/4$ evaluation stimuli seen in fitting under a subject-only holdout, a second overlap inside the held-out subject because each story is presented twice, and a 25% loss of pretraining diversity as the price of the fix—small enough corpora that it recommends bracketing the estimate with volume- and composition-matched variants, which is what §4 does. Run on a corpus with eight stimuli instead of four, the overlap is still total but the fix costs about 12% rather than 25%: the exposure does not depend on corpus size, the remedy's price does.

Applied to example manifests with the shapes of corpora the field uses (MEG-MASC, Gwilliams et al., 2023; Brennan and Hale, 2019; and a synthetic single-subject corpus of 30 stimuli standing in for LibriBrain-like designs, Özdogan et al., 2025), it separates three regimes:

corpus shape	example	exposure	cost of the fix
many subjects, few stimuli	27×4 (MEG-MASC)	100%	25% of pretraining diversity
many subjects, one stimulus	33×1 (Brennan)	100%	*not fixable*
one subject, many stimuli	1×30 (synthetic)	0%	3%

The middle row is the one worth naming. A corpus with a single stimulus admits no stimulus-disjoint cross-subject protocol at all: withholding the evaluation stimulus leaves nothing to pretrain on. Such designs can support within-subject claims, but a cross-subject transfer number computed on them cannot be separated from shared exposure by any split. The bottom row is the reassuring case—a single subject hearing many stimuli has no cross-subject overlap to begin with—and the top row is the one this paper measures.

The tool measures *exposure*, not inflation: it says the scores could be inflated, not by how much. That is what §4 measures, on one corpus, and the audit does not substitute for it. Its use is to say whether the question applies to your data, and what answering it will cost. We also release the corrected MEG-MASC split, the pool-pinning patch of §5, and a note stating the protocol we recommend.