WHAT ARE WE ACTUALLY DECODING? RETRIEVAL-BASED SOURCE ATTRIBUTION FOR NON-INVASIVE NEURAL LANGUAGE DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Non-invasive neural language decoding has advanced rapidly, with end-to-end systems now turning MEG and EEG into fluent sentences. However, a high score may come from the decoder's language prior, the text metric, or the input structure rather than the brain, and the output alone cannot distinguish them. In this work, we ask where such scores come from. We find that in an end-to-end sentence decoding pipeline, input duration alone supports most of the score: scored before the decoder, Gaussian noise of the same duration keeps **66.3% sentence-retrieval R@1 against 90.8% for real MEG**, a shortcut absent from an earlier invasive ECoG control and missed by noise controls scored at the output. We propose **source attribution**, which reads the score out at the decoder's input, where evaluation becomes retrieval with a known chance level, and gives each of three sources—structure in the input, evidence within a window, and context across windows—its own control. With duration removed, word-locked MEG windows still identify their audio among 1,464 candidates for **44% of windows**, while noise stays at chance. We further state that a contextual gain can be credited to the brain only if it **falls with this local evidence and vanishes without it**. A training-free aggregation meets this criterion on two MEG datasets, raising R@1 by 7.7 and 6.9 points, and yields no spurious gain on EEG, where local evidence is very weak; a stage that uses sentence length in place of evidence fails it. Crucially, source attribution requires no change to a pipeline's task, architecture, or training, suggesting broad applicability across encoder–decoder pipelines. By turning whether a decoder reads the brain from an assumption into a measurement, it lays a foundation for trustworthy non-invasive language decoding. Project Page Code

1 INTRODUCTION

Decoding language from non-invasive brain recordings, a long-standing goal for both neuroscience and assistive communication, has advanced rapidly (Wang & Ji, 2022; Duan et al., 2023; Xi et al., 2023; Yang et al., 2024; 2026b; Li et al., 2025). End-to-end systems now pair a neural encoder with a pretrained language or speech model and turn MEG and EEG into fluent sentences with high reported accuracy. Because the text is produced by a pretrained decoder, its accuracy reflects the recording and the decoder's own prior together (Jo et al., 2026). **Evaluation still measures how well a system decodes, while a prior question remains open: does the progress in decoding performance come from the neural content itself?**

We find that it need not: **in a sentence-level pipeline, duration alone supports most of the score.** Scored before its Whisper-based decoder (Li et al., 2025), in a sentence-level design common to many systems (Wang & Ji, 2022; Duan et al., 2023; Yang et al., 2026b), with participants held out, real MEG reaches 90.8% R@1 and duration-matched Gaussian noise in its place still reaches 66.3% (Fig. 2a); with fixed-duration inputs, noise no longer scores high. Earlier answers point the other way: length has little or no effect. In invasive ECoG decoding, a network trained on noise of the correct length stayed near chance (Makin et al., 2020). In non-invasive decoding, noise tests have scored near zero at the output (Yang et al., 2026b; Li et al., 2025), and length has been described only as a possible aid (Li et al., 2025). Read before the decoder, the answer reverses: length is not an aid but the dominant cue (Fig. 1a).

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.

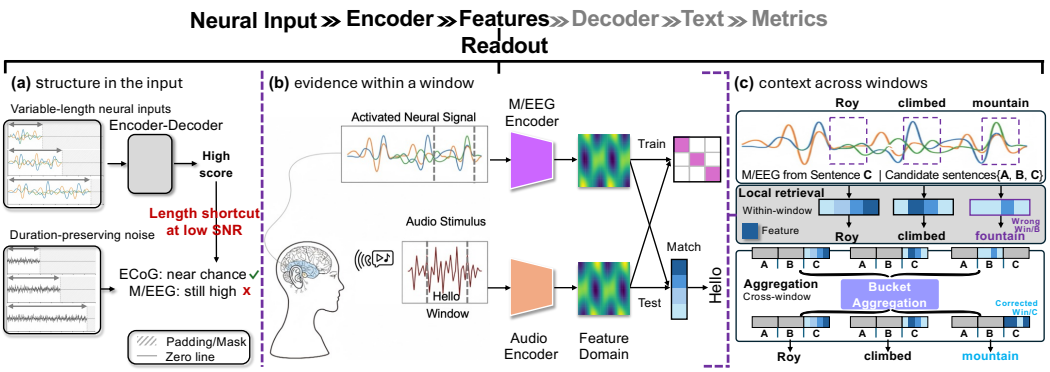


Figure 1: **Source attribution: where the score is read out, and the three sources it is traced to.** The readout sits before the decoder and ranks the correct stimulus among the test stimuli, so the decoder's prior and the text metric do not enter the score. (a) **Structure in the input.** Duration-preserving noise shows how much of the score non-neural cues support: near chance on invasive ECoG (Makin et al., 2020), most of it on MEG. Fixed-duration inputs remove the cue. (b) **Evidence within a window.** Fixed-duration windows are ranked against the time-aligned audio of held-out stimuli. (c) **Context across windows.** A contextual stage pools frozen window scores within a sentence and shifts whole candidate sentences; the readout before and after it attributes the change to that stage, which leaves the encoder and the candidate pool unchanged.

We move evaluation to the decoder's input, where what a score reads from the brain can be separated from what the input and the decoder supply. There, before the decoder's prior, its decoding strategy, and the text metric enter (Jo et al., 2026; Zhang et al., 2020; Hanna & Bojar, 2021), a representation can only be matched to the finite test stimuli, which gives a ranking with a known chance level (Sec. 3.1). Each source of the score can then be controlled: noise that replaces only the neural content exposes what the input's structure carries, and fixed-duration windows with held-out stimuli remove it. With length removed (Gwilliams et al., 2023; Défossez et al., 2023), real MEG ranks its time-aligned audio first among 1,464 candidates for 44% of windows, while noise stays at chance. This is the local evidence the length cue lets a sentence-level pipeline bypass.

Context then re-enters as a separate stage, and its gain is credited to the brain only if it builds on the local evidence. Scoring windows alone give up the context that sentence-level pipelines use legitimately, so we add it back on top of the frozen local scores, where we can measure its gain. Unlike the decoder's language prior, which yields fluent output even from noise, this stage should not score the target without local evidence: its gain must fall with that evidence and vanish without it. Sentence-bucket aggregation, which pools the window scores of each sentence (Sec. 3.4), raises R@1 from 44% to 52% on Gwilliams and from 22% to 29% on MOUS. Its gain shrinks when the grouping is broken and vanishes as local evidence is attenuated, and on EEG, where local evidence is near chance (Brennan & Hale, 2019), it yields none, as the criterion predicts; a stage that selects sentences by their length alone keeps its gain without evidence. These steps trace a score at the decoder's input to three sources: structure in the input, evidence within a window, and context across windows built on that evidence; we call this **source attribution** (Fig. 1).

Our contributions are as follows:

- **A dominant length shortcut.** In a sentence-level non-invasive pipeline, noise that keeps only the input's duration reaches 66.3% R@1 against 90.8% for real MEG. A length-only control that stays near chance for invasive recordings keeps most of the score here, and noise controls scored at the output do not.
- **Source attribution.** We evaluate a decoder by where its score comes from: read out at the decoder's input, a score is traced to structure in the input, evidence within a window, and context across windows, each measured with its own control (Sec. 3).
- **A criterion for context.** A contextual gain is credited to the brain only if it falls with the local evidence and vanishes without it. Our aggregation meets this criterion on two MEG datasets (+7.7 and +6.9 R@1), while a stage that uses sentence length in place of evidence fails it.

2 Related Work

Neural language decoding. Neural language decoding is dominated by encoder–decoder pipelines, in which a neural encoder maps the recording to a representation and a pretrained language or speech model turns it into text (Wang & Ji, 2022; Xi et al., 2023; Yang et al., 2026b; 2024; Duan et al., 2023; Li et al., 2025). These systems are evaluated on the output text (Jayalath et al., 2025a; Jo et al., 2026), which the decoder's prior shapes and can produce on its own (Jo et al., 2026). Retrieval-based decoding ranks the presented stimulus among candidates (Défossez et al., 2023; d'Ascoli et al., 2025; Jayalath et al., 2025b; Liu et al., 2025; Yang et al., 2026a). Our readout takes this form, but places it at the decoder input of generation pipelines, so that the same pipeline is scored before its prior enters.

Input length as a shortcut. Padding inputs to a common length is standard in speech recognition (Radford et al., 2023) and common in sentence-level neural decoding (Wang & Ji, 2022; Duan et al., 2023; Yang et al., 2026b; Li et al., 2025; Liu et al., 2025; Wang et al., 2026); at low signal-to-noise ratio, it can turn length into a shortcut (Geirhos et al., 2020; Baillargeon & Lamontagne, 2022). Earlier checks found little effect: in invasive ECoG decoding, noise of the correct length stays near chance (Makin et al., 2020); in non-invasive decoding, noise tests score near zero at the output (Yang et al., 2026b; Li et al., 2025), and length has been discussed only as a possible aid (Li et al., 2025). Other checks address teacher forcing (Jo et al., 2026), leakage between training and test (Brookshire et al., 2024; Yin et al., 2025), and the text similarity rewarded by embedding metrics (Zhang et al., 2020; Hanna & Bojar, 2021). Our setting differs from the invasive check in its signal-to-noise ratio, and from the non-invasive checks in where the score is read, which metric is used, and what the control replaces; with all three in place, the length cue becomes visible (Sec. 4.2).

Context in neural language decoding. Neural responses to speech are modulated by linguistic context carried by slower states that span words (Slaats et al., 2023; Brodbeck et al., 2022; Raghavan & Parra, 2025; Broderick et al., 2018), and current decoders model it with context Transformers trained end-to-end over local features (d'Ascoli et al., 2025; Jayalath et al., 2025a), which can also learn the statistics of the target labels (Radford et al., 2023; Jo et al., 2026). In speech recognition, external priors are added to acoustic scores by shallow fusion (Kannan et al., 2018). Our aggregation takes this additive form, but its prior comes only from local scores, with no trained parameters or labels, so we can test its gain against the criterion in Sec. 3.2; we can test learned context modules the same way.

3 Methods

3.1 Readout at the decoder input

We place the readout at the decoder input, before the decoder's prior enters the score. Encoder–decoder architectures are the prevailing paradigm in neural language decoding (Wang & Ji, 2022; Duan et al., 2023; Yang et al., 2026b; Li et al., 2025; Liu et al., 2025); beyond this point, other sources enter the pipeline, and the score becomes confounded.

Retrieval as the readout form. At the decoder input, no text has been generated yet, but the test stimuli form a finite, known set. One natural readout is therefore to rank the presented stimulus among them by similarity. Its chance level is known, $1/N$ for N unique candidates, and its rank shows how well the features supplied to the decoder single out the stimulus. Retrieval-based and generation-based decoding meet here: the former stops, and the latter continues, so any generation pipeline can be read out the same way. We use this readout in two places. In an end-to-end sentence decoding pipeline, the candidates are sentences or codebook entries, and the readout shows how much of the score input structure carries (Sec. 4.2). On fixed-duration M/EEG windows, the candidates are the test audio windows; there it measures local evidence and tests whether a contextual gain follows that evidence or sentence length (Secs. 4.3–4.4).

Notation and metrics. For the window-level readout, let $\mathcal{Q}$ be the set of query neural windows and $\mathcal{C}$ the pool of N candidate audio windows, de-duplicated by stimulus and window index so that chance is well defined. For each query $w \in \mathcal{Q}$, $j^\star(w) \in \mathcal{C}$ is its time-aligned target and $\ell_{w,j}$ the local logit of candidate j (Eq. 5). Each window and candidate belong to a sentence $\sigma(\cdot)$, and the candidates of sentence s form a bucket $\mathcal{B}_s = \{j \in \mathcal{C} : \sigma(j) = s\}$; sentence identity is used by the controls and the contextual stage, never by the local logits. The rank $r(w)$ is the position of $j^\star(w)$ among the

candidates sorted by $\ell_{w,j}$ in descending order, and we report R@K, mean reciprocal rank (MRR), and median rank (MedR):

$$\mathrm{R@}K = \frac{1}{|\mathcal{Q}|}\sum_{w\in\mathcal{Q}} \mathbf{1}[r(w) \le K], \quad \mathrm{MRR} = \frac{1}{|\mathcal{Q}|}\sum_{w\in\mathcal{Q}} \frac{1}{r(w)}, \quad \mathrm{MedR} = \mathrm{median}\{r(w) : w \in \mathcal{Q}\}. \tag{1}$$

Signal-blind control. The signal-blind control replaces only the neural content of each input and keeps everything else, such as its length, padding, and masks:

$$x^{\varepsilon} = M \odot \varepsilon + (1 - M) \odot x_{\mathrm{pad}}, \qquad \varepsilon \sim \mathcal{N}(0, I), \tag{2}$$

where x_{pad} is the padded input and M is its valid-length mask (all ones for the fixed-duration windows of Sec. 3.2). Any retrieval above chance under Eq. 2 therefore comes from the structure of the input, not from the brain; Sec. 3.2 keeps this structure out of the local stage.

3.2 Controls for the three sources

We give each of the three sources of Fig. 1 a control that intervenes on that source alone, as Eq. 2 does: the source is removed while everything else is kept, and what the score loses is what it owes to that source. End-to-end, the sources enter together and cannot be removed one at a time, so we measure them at two stages. Fixed-duration windows and stimulus-identity splits remove structure from the local stage, so that its score rests on evidence within a window, which the oracle divides into within- and cross-sentence parts. Structure and context re-enter at the contextual stage, whose gain (Eq. 4) is tested by removing the evidence while keeping the grouping.

Fixed-duration windows. Fixed-duration windows fix the input duration rather than encoding it through padding or masks. For a word onset t_w in the continuous recording X, the neural input is

$$x_w = X[t_w + \delta - 0.5,\ t_w + \delta + 2.5] \in \mathbb{R}^{C\times T}, \tag{3}$$

in seconds, with lag $\delta = 150$ ms, C sensors, and $T = 360$ samples at 120 Hz (Défossez et al., 2023), as in event-related M/EEG language studies (Luck, 2014; Beres, 2017; Gwilliams et al., 2023).

Stimulus-identity splits. Stimulus-identity splits keep every test stimulus out of training. A stimulus identity is an audio segment with its onset and offset, and all windows sharing an identity, across participants and sessions, are assigned to the same partition.

Oracle within-bucket ranking. The oracle ranks each target only among the candidates of its own sentence, separating within-sentence discriminability from cross-sentence competition. For a query w with target sentence $s^\star = \sigma(j^\star(w))$, $j^\star(w)$ is ranked within $\mathcal{B}_{s^\star}$ with the model and local logits fixed; the gap to its full-pool rank bounds what a bucket-level correction can recover (Sec. 3.4).

Context controls. The contextual gain is measured on the same frozen local logits,

$$\Delta_{\mathrm{ctx}} = \mathrm{R@1}_{\mathrm{ctx}} - \mathrm{R@1}_{\mathrm{local}}. \tag{4}$$

A contextual stage takes two inputs, the grouping and the local evidence, and each control changes one while holding the other fixed. Grouping perturbations move a fraction p of query windows to another sentence, either a neighbouring one (boundary jitter) or a random one in the same story (reassignment). Evidence attenuation keeps the grouping and mixes each query window x_w with a stimulus-mismatched window $\tilde{x}_w$ from the same evaluation set, $x_w^{(\alpha)} = \sqrt{\alpha}\, x_w + \sqrt{1-\alpha}\, \tilde{x}_w$ with $\alpha \in [0, 1]$, so that the match to the stimulus weakens as α falls; Brennan EEG (Brennan & Hale, 2019) gives weak evidence without intervention, and further variants are in Appendix E.3. The criterion for context follows: a contextual gain is credited to the brain only if it falls with the evidence and vanishes without it. A stage that selects sentences by length alone is a positive control that should fail it (Appendix E.5).

3.3 Local neural-to-audio retrieval model

The local stage scores each window against every audio candidate with the M/EEG-to-audio retrieval model of Défossez et al. (2023), which serves as its instrument. We instantiate it with three temporal backbones; under the controls of Sec. 3.2, differences between them reflect the neural evidence each extracts from the same windows.

Retrieval scoring. Each audio candidate j is encoded by a frozen speech encoder (Baevski et al., 2020) into a target $y_j^{\text{aud}} \in \mathbb{R}^{D\times T}$ with D features, resampled to the T time steps of the neural window and shared across datasets (Appendix A.4). A neural encoder f_θ maps each window to $z_w^{\text{neu}} = f_\theta(x_w) \in \mathbb{R}^{D\times T}$, and the local logit is their Frobenius inner product,

$$\ell_{w,j} = \left\langle z_w^{\text{neu}},\ \frac{y_j^{\text{aud}}}{\|y_j^{\text{aud}}\|_F} \right\rangle_F, \tag{5}$$

which preserves alignment across the 3 s window. Keeping the audio targets frozen and shared fixes the space in which every encoder, dataset, and control is scored: the candidate side never changes, so any change in the ranking comes from the neural side, and the targets carry no text-label statistics.

Training objective. The neural encoder is trained with a contrastive retrieval loss, in which the matched logit competes against all audio candidates in a batch of B pairs:

$$\mathcal{L} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\ell_{i,i})}{\sum_{j=1}^{B}\exp(\ell_{i,j})}. \tag{6}$$

Encoder instantiations. The three encoders differ only in the temporal backbone: Dense-TCNN, the exp-dilated CNN of Défossez et al. (2023) with a matched receptive field, and a deeper variant of it (Appendix B.1). All share coordinate-conditioned spatial mixing, which maps each sensor layout to a common latent grid, and a subject-specific layer that adapts the representation across participants. The resulting local logits are frozen; the contextual stage of Sec. 3.4 operates on them without modifying the encoder, audio embeddings, or candidate pool.

3.4 Sentence-bucket aggregation as the contextual stage

The contextual stage needs an instance whose gain can be tested against the criterion of Sec. 3.2. We use sentence-bucket aggregation, which fuses each window's frozen local logits with a prior over candidate sentences formed from the local logits of all windows in its sentence. It uses the structure kept out of the local stage, which windows form a sentence and how many candidates each has, but forms its prior only from local scores. It has no trained parameters, never sees target labels, and changes only the competition between sentences, so its gain (Eq. 4) is bounded by the oracle.

Query groups and candidate buckets. Query windows are grouped by the presented sentence into groups $G \subseteq \mathcal{Q}$, which stimulus timing provides in evaluation; candidates form the sentence buckets $\mathcal{B}_s$ of Sec. 3.1. Fixing the grouping leaves the stage with a single quantity to infer from the local logits: which bucket a group corresponds to. The oracle of Sec. 3.2 is given this correspondence.

Shallow fusion with a bucket prior. For $w \in G$, the aggregated logits are obtained by additive shallow fusion (Kannan et al., 2018),

$$\ell_{w,j}^{\text{agg}} = \ell_{w,j} + \gamma_{\text{agg}}\, \pi_G(\sigma(j)), \tag{7}$$

where the prior π_G pools, for each bucket s, the local logits in $\mathcal{B}_s$ that rank highly for the windows of G, scales them by $|\mathcal{B}_s|^{-1/2}$ so that bucket size cannot stand in for evidence, and is zero outside the S best-supported buckets (Appendix D.1).

Probabilistic reading. The probabilistic form makes the contribution of context explicit. With $q_w = \text{softmax}(\ell_{w,\cdot})$ and $p_G(j) \propto \exp\big(\pi_G(\sigma(j))\big)$, Eq. 7 is exactly

$$P_{\text{agg}}(j \mid w, G) = \frac{q_w(j)\, p_G(j)^{\gamma_{\text{agg}}}}{\sum_k q_w(k)\, p_G(k)^{\gamma_{\text{agg}}}}. \tag{8}$$

Fusion thus multiplies the local odds between candidates i and j by $[p_G(i)/p_G(j)]^{\gamma_{\text{agg}}}$, with γ_{agg} an inverse temperature. Within a bucket, the factor is constant, so the local odds are unchanged and

$$\text{R@1}_{\text{agg}} \le \text{R@1}_{\text{oracle}}. \tag{9}$$

Across buckets, the factor depends only on the local logits of G and bucket size; the controls in Sec. 3.2 test whether the gain still requires that evidence.

Configuration. One configuration is used for all test evaluations, including MOUS: the pooling hyperparameters and the form of bucket-size normalisation were selected on Gwilliams validation (Appendix D.2), and $\gamma_{\text{agg}} = 0.7$ was fixed in advance. Component ablations are in Appendix D.3.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We use the three naturalistic listening datasets on which M/EEG-to-audio retrieval was established (Défossez et al., 2023), spanning two languages, two MEG systems, and EEG, with 156 participants and about 144 hours of recordings in total (Table 1(a)). Local evidence in EEG is much weaker than in MEG. Gwilliams is the primary setting; MOUS tests transfer across language and recording system without retuning, in the same multilingual target space; on Brennan EEG, the criterion of Sec. 3.2 predicts no contextual gain.

Table 1: **Datasets and embedding-metric diagnostic.** N: test candidates under the stimulus-identity split (no test stimulus is heard in training). Panel (b) evaluates fixed, signal-blind outputs.

(a) Dataset overview

Dataset	Language	Mode	Sensors	Participants	Duration	N
Brennan	EN	EEG	60	33	6.7 h	388
MOUS	NL	MEG	272	96	80.9 h	825
Gwilliams	EN	MEG	208	27	56.2 h	1,464

(b) Embedding-metric diagnostic

Prediction	**Category**	**tokens**	**WER**	**CER**	**BERT-F1**
"the the"	repetition	2	0.953	0.920	**0.868**
"the"	function word	1	0.965	0.962	0.826
"In the eye the"	fragment	4	0.964	0.871	0.814
"I don't know"	generic filler	3	0.997	0.894	0.800

Training and evaluation. For the length-shortcut diagnostic, we reproduce on Gwilliams the Whisper-based end-to-end pipeline of Li et al. (2025), representative of the sentence-level design in which a whole sentence's recording is padded to a common length (Wang & Ji, 2022; Duan et al., 2023; Yang et al., 2026b), and read out before its text decoder under its subject split (Appendix C.1). Retrieval encoders are trained as in Sec. 3.3, with checkpoints selected by validation loss and scaling statistics computed on the training split only (Appendices A.1 and A.3). Results are averaged over three seeds (one on Brennan EEG), and before–after contrasts are tested with a paired sentence-cluster bootstrap.

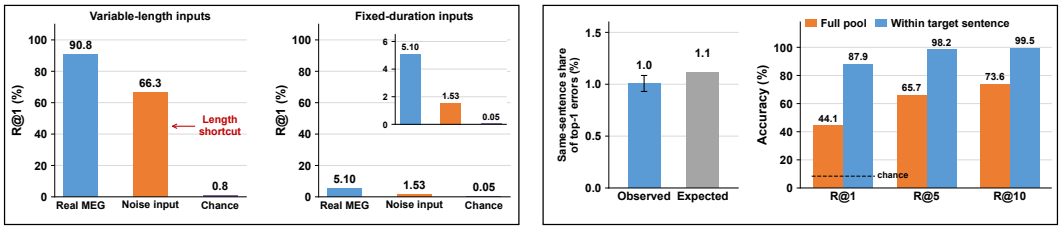
Results follow the stages of Sec. 3.2: in an end-to-end pipeline, structure enters with the recording and dominates the score (Sec. 4.2); with structure kept out of the local stage, the evidence that remains is measured (Sec. 4.3); structure and context then re-enter on top of that evidence and are tested (Sec. 4.4).

4.2 EXPOSING STRUCTURE IN THE INPUT

Scored before its text decoder, the end-to-end pipeline obtains most of its accuracy without neural content. The readout ranks the correct sentence among the evaluation sentences: real MEG reaches 90.8% R@1, and **Gaussian noise within the valid length, with padding and masks unchanged, still reaches 66.3%** (chance 0.8%; Fig. 2a). The noise keeps only the input's duration, so duration alone supports most of the score. An end-to-end readout counts that share as neural decoding. With fixed-duration inputs, noise no longer scores high (Fig. 2a; Appendix C.1).

A stricter split lowers the score but leaves the cue in the input. Padding a weak signal to a common length puts duration in the input under any split; the split decides whether it pays at test. Under the subject split, every evaluation sentence was heard in training, and duration serves as a key. With whole stories held out, the key no longer fits, and real MEG falls from 90.8% to 7.1% (Appendix C.1). Generation shows the same drop: BLEU-1 falls to 6.9 when stories are held out (Yang et al., 2026b).

The shortcut is exposed before the decoder. Noise tests reported for such pipelines replace the whole input and score the output text, and they yield BLEU-1 near zero (Yang et al., 2026b; Li et al., 2025). With only the values within the valid length replaced and the score read before the decoder, the reproduced pipeline keeps 66.3%. Before the decoder, each target is a unique stimulus with a known chance level. At the output, the score also responds to the text itself: as Table 1(b) shows, embedding metrics give fixed, signal-blind outputs such as "the the" or "I don't know" with a BERT-F1 of 0.80–0.87 despite WER above 0.95, and automatically aligned text labels add errors of their own (Appendix C.3). The cue can therefore be both exposed and removed at the decoder input: with duration fixed, what remains of the ranking can be traced to neural content (Sec. 4.3).



(a) Length-shortcut diagnostic. **(b) Local errors and the within-sentence oracle.**

Figure 2: **Diagnostics for the length shortcut and for local errors.** (a) Left: with variable-length sentence inputs, ranking the correct sentence among the evaluation sentences succeeds even with duration-preserving noise in place of MEG (chance 0.8%). Right: with fixed-duration inputs and a token-level readout over the 2,048-entry codebook, noise no longer scores high (chance 0.05%). The two panels use different readouts with different chance levels and are not directly comparable. Both use a subject split. (b) Left: share of top-1 errors that fall within the target's sentence, observed (mean±std over three seeds) and expected from sentence lengths. Right: the same frozen local scores ranked in the full pool and among the candidates of the target's sentence (Dense-TCNN, one checkpoint); the dashed line marks chance within a sentence.

4.3 Measuring evidence within a window

With fixed-duration windows and stimulus-identity splits, above-chance ranking requires neural content. Gaussian inputs stay at chance (0.072% R@1 against 0.068% on Gwilliams, 0.116% against 0.121% on MOUS), whereas **real MEG ranks its time-aligned audio first for 44.4% of windows among 1,464 candidates**, about 650 times chance, and for 21.7% among 825 on MOUS, about 180 times chance (Table 2); every participant is well above chance, the lowest at 14.0% on Gwilliams and 2.4% on MOUS (Appendix E.2). Differences between encoders are then attributable to the neural evidence each extracts: Dense-TCNN exceeds the receptive-field-matched exp-dilated encoder of Défossez et al. (2023) by 2.5 points on Gwilliams and 5.6 on MOUS with all other components shared, and is higher for all 27 Gwilliams participants and for 91 of 96 MOUS participants (Appendix E.2).

Table 2: **Local retrieval under fixed-duration windows and stimulus-identity splits, and the contextual gain at the fixed $\gamma_{\text{agg}} = 0.7$.** Chance is the value for a uniform ranking over the pool of size N. Gaussian inputs replace each test window by $\mathcal{N}(0, I)$ noise through the trained Dense-TCNN encoder (Appendix C.2). Exp-dilated $(L{=}5)$ is the receptive-field-matched encoder of Défossez et al. (2023) and $(L{=}10)$ the deeper variant; all other components are shared. R@1, MRR, and R@10 are percentages, mean±std over three seeds; MedR is the median rank, averaged over seeds. Δ columns give the change after sentence-bucket aggregation on the same frozen logits; its size depends on γ_{agg} (Appendix D.2), and full metrics after aggregation are in Table 9. †: significant in all seed-level paired sentence-cluster bootstraps at $p < 0.05$ (Table 10).

Setting	Input / encoder	Local retrieval				Gain	
		R@1 (%)↑	MRR (%)↑	R@10 (%)↑	MedR↓	ΔR@1 (pp)	ΔMRR (pp)
Gwilliams $N{=}1{,}464$	Chance	0.07	0.54	0.68	732	–	–
	Gaussian input	0.07±0.01	0.55±0.01	0.69±0.01	731	−0.01±0.01	−0.02±0.00
	Dense-TCNN	44.4±0.3	54.4±0.2	73.6±0.1	2.0	+7.7±0.9†	+6.4±0.7†
	Exp-dilated ($L = 5$)	41.9±0.9	51.9±0.9	71.1±0.7	2.0	+8.7±0.6†	+7.4±0.6†
	Exp-dilated ($L = 10$)	39.6±0.3	49.7±0.2	69.2±0.1	3.0	+6.6±0.8†	+5.7±0.7†
MOUS $N{=}825$	Chance	0.12	0.88	1.21	413	–	–
	Gaussian input	0.12±0.01	0.89±0.01	1.21±0.04	410	0.00±0.01	0.00±0.01
	Dense-TCNN	21.7±0.4	30.2±0.5	46.6±0.5	13.3	+6.9±0.2†	+6.9±0.1†
	Exp-dilated ($L = 5$)	16.1±0.8	24.0±0.7	39.5±1.0	21.7	+5.4±0.1†	+5.3±0.1†
	Exp-dilated ($L = 10$)	13.4±0.9	20.8±1.1	35.1±1.3	30.0	+4.5±0.3†	+4.4±0.2†

The local score shows no preference for the target's sentence. Windows of the same sentence share its linguistic context, so a local score that drew on this context would, when wrong, favour other windows of the target's sentence. However, the wrong windows that the local score ranks first belong to the target's sentence only 1.0% of the time, close to the 1.1% expected from sentence lengths alone

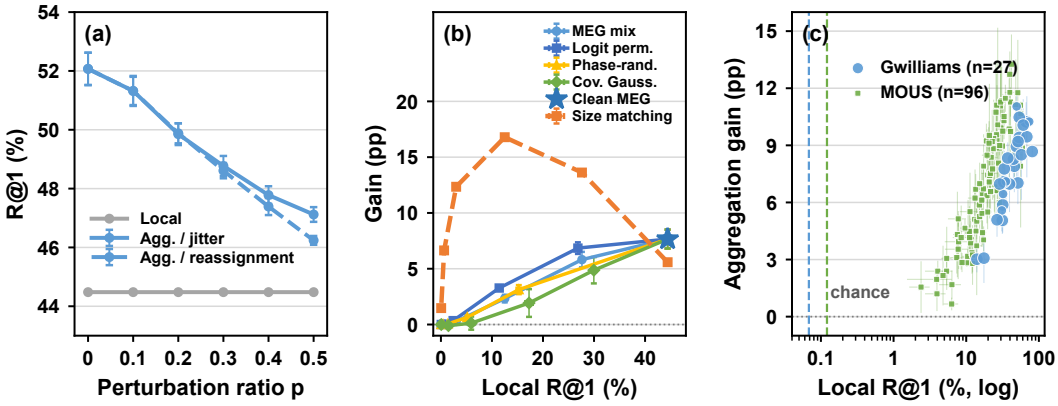


Figure 3: **Grouping and evidence controls.** (a) Moving a fraction p of query windows to a neighbouring sentence (jitter) or to a random sentence in the same story (reassignment) lowers aggregation R@1 as p grows; local logits are fixed, so local retrieval is unchanged. Three seeds; error bars are the standard deviation over seeds. (b) Gain (pp) versus local R@1 (%) as evidence is attenuated, three seeds: aggregation under four attenuation families (solid; scores and bucket support recomputed at each level; the star marks clean MEG) and a stage that selects sentences by length alone (dashed; real-MEG mixing; Appendix E.5); error bars are the standard deviation over seeds. (c) Aggregation gain versus local R@1, one point per participant (Dense-TCNN; Gwilliams and MOUS). Marker size is proportional to the number of test windows; error bars are the standard deviation over three seeds; dashed lines mark chance.

(Fig. 2b; Appendix E.6). The local readout thus resolves the stimulus content of each window, not the sentence it belongs to, so any sentence-level gain must come from a later stage.

For at least half of the windows, the target is ranked first or second. On Gwilliams, R@1 is 44.4%, the median rank is 2, and R@10 is 73.6%; among the misses, 38.6% place the target second to fifth (Appendix E.6). Ranking within the correct sentence with logits fixed raises R@1 from 44% to 88%, against 8.4% chance within a sentence (Fig. 2b). This is the room shallow fusion (Sec. 3.4) is built for: a prior pooled from a sentence's windows shifts whole sentences and keeps the order within each, so selecting the right sentence recovers the within-sentence ranking.

4.4 Testing context across windows

Pooling a sentence's windows raises R@1 by 7.7 points on Gwilliams and 6.9 on MOUS. On the same frozen logits, with γ_{agg} fixed in advance and the configuration selected on Gwilliams validation applied unchanged to MOUS, R@1 rises from 44.4% to 52.1% and from 21.7% to 28.6% (Table 2; Appendix D.2 varies γ_{agg}). The gain is positive for all three encoders and, with Dense-TCNN, for every participant (Appendix E.2). The controls of Sec. 3.2 test what it rests on: the grouping, the local evidence, and structure alone.

The gain depends on coherent grouping. With each query group reduced to a single window, the gain disappears (R@1 42.3% vs 44.1% locally on one checkpoint); pooling across windows produces it (Appendix D.3). Moving a fraction p of query windows to a neighbouring sentence or to a random one in the same story lowers aggregated R@1 as p grows: at $p = 0.5$, R@1 falls from 52.1% to 47.1% and 46.2%, and a quarter to a third of the gain remains (Fig. 3a).

The gain falls with local evidence and vanishes without it. As query windows are mixed with stimulus-mismatched MEG, the gain falls from $+7.7$ points on clean MEG to $+2.3$ at $\alpha = 0.5$ and to zero at $\alpha \leq 0.10$; across four attenuation families, it approaches zero as local R@1 approaches chance (Fig. 3b; Appendix E.3). The same dependence holds across participants without intervention: the gain rises with local R@1 (Spearman $\rho = 0.88$ on Gwilliams, 0.90 on MOUS; Fig. 3c; Appendix E.2). Where there is no local evidence, the stage yields no gain: on Gaussian inputs R@1 changes by -0.01 points on Gwilliams and 0.00 on MOUS (Table 2), and on Brennan EEG, where local R@1 is 0.72% among 388 candidates (chance 0.26%), aggregation lowers it to 0.21%, and the gain stays negative at every smaller candidate-pool size (Appendix E.4).

Structure can support local evidence or replace it, and the criterion tells the two apart. Because query groups and candidate sentences are cut at the same sentence boundaries, sentence length alone points to the right sentence: choosing the three sentences whose number of windows best matches the query group keeps the correct one for 65.0% of queries, against 2.4% by chance (Appendix E.5). A stage that chooses sentences this way relies on structure rather than evidence. Under attenuation, its gain peaks at 16.8 points at $\alpha = 0.5$, where aggregation adds 2.3 (Fig. 3b, dashed), and at $\alpha = 0$ it still reaches about 1.5% R@1 under all four attenuation families, some 22 times chance, while aggregation adds nothing. Both stages use structure; only aggregation's gain needs the recording.

Query by query, the gain follows the local evidence. At one checkpoint, aggregation corrects 6,724 local errors and overturns 549 correct windows, 1.7% of those correct locally. It shifts only the sentences that a group's local logits select: the correct sentence is among the three selected for 57.0% of queries, and where it is not, no error is corrected. Selection and correction both follow the local rank: for errors at initial ranks 2–5, 6–10, and beyond 10, the correct sentence is selected for 56.0%, 46.1%, and 30.2%, and aggregation repairs 38.4%, 11.2%, and 0.7% (Appendix E.6). What limits the gain is therefore sentence selection, which the local evidence makes: queries whose sentence is not selected account for about two-thirds of the gap to the oracle, which is given the correct sentence.

Context can therefore enter on top of local evidence without standing in for it. Aggregation uses sentence grouping and sentence length, yet its gain comes from pooling a sentence's windows and falls with their evidence, as the criterion of Sec. 3.2 requires. Structure kept out of the local stage thus returns at the contextual stage, where we can measure its contribution.

5 Discussion

Our results separate how well a system scores from how much of that score it reads from the brain. Separating the two requires controlling where the information comes from, and comparing final outputs cannot do so by itself: the score is taken after the decoder's prior, the decoding strategy, and the text metric have entered. Read out at the decoder input instead, an end-to-end pipeline scored 66.3% R@1 on noise against 90.8% on MEG, carried by input duration. With duration removed from the local stage, noise fell to chance, while word-locked MEG windows still ranked their time-aligned audio first for 44% of windows among 1,464 candidates. Context pooled from those windows added 7.7 points that fell with the local evidence and vanished without it, whereas a stage that read sentence length kept its gain without evidence. These checks can be placed at the decoder input of any encoder–decoder pipeline without changing its task or architecture.

Three conditions kept the shortcut out of sight. A length-only control stayed near chance in invasive ECoG decoding (Makin et al., 2020); in the MEG pipeline here it keeps most of the score, and on EEG, output scores and retrieval diverge further: under a subject split, generation on Brennan EEG reaches about 90% BLEU-1 (Li et al., 2025), whereas held-out EEG segments are retrieved at 5.2% (Défossez et al., 2023). A check passed in one decoding regime therefore does not carry over to another. The score must also be read before the decoder, where the prior and the text metric have not entered, and the control must replace only the neural content rather than the padded input as a whole (Sec. 4.2). Each earlier check lacked at least one of these; the readout of Sec. 3 meets all three.

The same test extends to other forms of context. Sentence-bucket aggregation is one way to pool local evidence; sequence models over window scores, learned context modules, and language-model rescoring are others, and each can follow the same frozen local readout. Each also meets the structure that aggregation makes explicit: in a closed set with known segmentation, any sentence-level module can match on sentence length, and a language model's total log-likelihood over a candidate sentence already depends on it. Whether other forms gain more, with gains that fall with the local evidence rather than with sentence length, is the natural next question.

Scope and future work. We analyse aggregation offline, with the candidate pool, the grouping, and the local scores fixed so that each source can be controlled and intervened on, and with an explicit structured prior, the grouping of windows into sentences. A natural next step is to build methods, including language models, that bring in context only through the local evidence. Systems that pair a weak neural signal with strong priors can be useful assistants, and their task scores measure that use; source attribution adds a second measure of how much of the score is read from the recording. Reporting both lays a foundation for trustworthy non-invasive language decoding.

AI USE STATEMENT

Generative AI tools assisted with language editing and developing some analysis code. The authors are responsible for the experimental design, validation of the analyses, interpretation of the results, and the manuscript's final content.

REFERENCES

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12449–12460, 2020.

Jean-Thomas Baillargeon and Luc Lamontagne. Assessing the impact of sequence length learning on classification tasks for transformer encoder models. *arXiv preprint arXiv:2212.08399*, 2022.

Anna M. Beres. Time is of the essence: A review of electroencephalography (EEG) and event-related brain potentials (ERPs) in language research. *Applied Psychophysiology and Biofeedback*, 42(4): 247–255, 2017. doi: 10.1007/s10484-017-9371-3.

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

Christian Brodbeck, Shohini Bhattasali, Aura A. L. Cruz Heredia, Philip Resnik, Jonathan Z. Simon, and Ellen Lau. Parallel processing in speech perception with local and global representations of linguistic context. *eLife*, 11:e72056, 2022.

Michael P. Broderick, Andrew J. Anderson, Giovanni M. Di Liberto, Michael J. Crosse, and Edmund C. Lalor. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology*, 28(5):803–809.e3, 2018. doi: 10.1016/j.cub.2018.01.080.

Geoffrey Brookshire, Jake Kasper, Nicholas M. Blauch, Yunan Charles Wu, Ryan Glatt, David A. Merrill, Spencer Gerrol, Keith J. Yoder, Colin Quirk, and Ché Lucero. Data leakage in deep learning studies of translational EEG. *Frontiers in Neuroscience*, 18:1373515, 2024.

Stéphane d'Ascoli, Corentin Bel, Jérémy Rapin, Hubert Banville, Yohann Benchetrit, Christophe Pallier, and Jean-Rémi King. Towards decoding individual words from non-invasive brain recordings. *Nature Communications*, 16(1):10521, 2025. doi: 10.1038/s41467-025-65499-0.

Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5(10): 1097–1107, 2023.

Yiqun Duan, Jinzhao Zhou, Zhen Wang, Yu-Kai Wang, and Chin-Teng Lin. DeWave: Discrete encoding of EEG waves for EEG-to-text translation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 9907–9918, 2023.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Laura Gwilliams, Graham Flick, Alec Marantz, Liina Pylkkänen, David Poeppel, and Jean-Rémi King. Introducing meg-masc a high-quality magneto-encephalography dataset for evaluating natural speech processing. *Scientific data*, 10(1):862, 2023.

Michael Hanna and Ondřej Bojar. A fine-grained analysis of BERTScore. In *Proceedings of the Sixth Conference on Machine Translation*, pp. 507–517, 2021.

Dulhan Jayalath, Gilad Landau, and Oiwi Parker Jones. Unlocking non-invasive brain-to-text. In *2nd Generative AI for Biology Workshop at ICML 2025*, 2025a. Poster.

Dulhan Jayalath, Gilad Landau, Brendan Shillingford, Mark Woolrich, and Oiwi Parker Jones. The brain's bitter lesson: Scaling speech decoding with self-supervised learning. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 27005–27020. PMLR, 13–19 Jul 2025b.

Hyejeong Jo, Yiqian Yang, Juhyeok Han, Yiqun Duan, Hui Xiong, and Won Hee Lee. Evaluating EEG-to-text models through noise-based performance analysis. *Scientific Reports*, 16(1):350, 2026. doi: 10.1038/s41598-025-29587-x.

Anjuli Kannan, Yonghui Wu, Patrick Nguyen, Tara N Sainath, Zhijeng Chen, and Rohit Prabhavalkar. An analysis of incorporating an external language model into a sequence-to-sequence model. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

Jilong Li, Zhenxi Song, Jiaqi Wang, Meishan Zhang, Honghai Liu, Min Zhang, and Zhiguo Zhang. BrainECHO: Semantic brain signal decoding through vector-quantized spectrogram reconstruction for whisper-enhanced text generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 2762–2778, 2025. doi: 10.18653/v1/2025.findings-acl.142.

Xiaozhao Liu, Dinggang Shen, and Xihui Liu. Learning interpretable representations leads to semantically faithful eeg-to-text generation. *arXiv preprint arXiv:2505.17099*, 2025.

Steven J. Luck. *An Introduction to the Event-Related Potential Technique*. MIT Press, Cambridge, MA, 2 edition, 2014. ISBN 9780262525855.

Joseph G Makin, David A Moses, and Edward F Chang. Machine translation of cortical activity to text with an encoder–decoder framework. *Nature neuroscience*, 23(4):575–582, 2020.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 2023.

Vinay S. Raghavan and Lucas C. Parra. Neural encoding of linguistic features during natural sentence reading. *iScience*, 28(7):112798, 2025. doi: 10.1016/j.isci.2025.112798.

Jan-Mathijs Schoffelen, Robert Oostenveld, Ngoc H. L. Lam, Julia Uddén, Annika Hultén, and Peter Hagoort. A 204-subject multimodal neuroimaging dataset to study language processing. *Scientific Data*, 6(1):17, 2019. doi: 10.1038/s41597-019-0020-y.

Sophie Slaats, Hugo Weissbart, Jan-Mathijs Schoffelen, Antje S. Meyer, and Andrea E. Martin. Delta-band neural responses to individual words are modulated by sentence processing. *Journal of Neuroscience*, 43(26):4867–4883, 2023.

Yuchen Wang, Haonan Wang, Yu Guo, Honglong Yang, and Xiaomeng Li. Escaping the BLEU trap: A signal-grounded framework with decoupled semantic guidance for EEG-to-text decoding. *arXiv preprint arXiv:2603.03312*, 2026.

Zhenhailong Wang and Heng Ji. Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 5350–5358, 2022.

Nuwa Xi, Sendong Zhao, Haochun Wang, Chi Liu, Bing Qin, and Ting Liu. UniCoRN: Unified cognitive signal reconstruction bridging cognitive signals and human language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023.

Xiaoli Yang, Huiyuan Tian, Yurui Li, Jianyu Zhang, Shijian Li, and Gang Pan. Brain-cliplm: Semantic compression for eeg-to-text decoding. *arXiv preprint arXiv:2604.16370*, 2026a.

Yiqian Yang, Hyejeong Jo, Yiqun Duan, Qiang Zhang, Jinni Zhou, Won Hee Lee, Renjing Xu, and Hui Xiong. MAD: Multi-alignment MEG-to-text decoding, 2024.

Yiqian Yang, Yiqun Duan, Qiang Zhang, Hyejeong Jo, Jinni Zhou, Won Hee Lee, Renjing Xu, and Hui Xiong. Neuspeech: Decode neural signal as speech. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6636–6640. IEEE, 2026b.

Congchi Yin, Qian Yu, Zhiwei Fang, Changping Peng, and Piji Li. Rethinking cross-subject data splitting for brain-to-text decoding. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5675–5689, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.289.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *Proceedings of the 8th International Conference on Learning Representations*, 2020.

A Data and preprocessing

A.1 Stimulus identities and data splits

Each dataset manifest associates a sentence identifier with its audio file, stimulus onset and offset, and neural recording interval. Gwilliams and Brennan provide sentence and word timings in their released event metadata. MOUS sentence blocks are constructed from BIDS auditory events, with word timings obtained as described in Appendix A.3.

A stimulus identity is the tuple (audio path, stimulus onset, stimulus offset). All presentations of an identity belong to the same split across participants and sessions. We assign identities to training, validation, and test sets in proportions of 70%, 10%, and 20%, respectively. For Gwilliams, we remove training and validation windows that overlap a test window on the same audio file, with the test set taking priority.

A.2 Data and model access

The analyses use the published Gwilliams/MEG-MASC, MOUS, and Brennan datasets (Gwilliams et al., 2023; Schoffelen et al., 2019; Brennan & Hale, 2019), released under CC0 1.0, RU-DI-HD-1.0, and CC-BY 4.0, respectively. Consent and ethics procedures are documented by the original studies. The fixed pretrained components are wav2vec2-large-xlsr-53 (Baevski et al., 2020) and Whisper large-v2 (Radford et al., 2023), under Apache-2.0 and MIT licences. We obtain the data and model weights from their original repositories. The accompanying code covers preprocessing, splits, training, scoring, and aggregation analyses.

A.3 Neural preprocessing and MOUS alignment

Neural windows undergo pre-onset baseline correction, channel-wise robust scaling using training-set medians and interquartile ranges, and value clamping. We retain dataset-provided filtering and add no additional band-pass or notch filters. Inputs have shape $C \times 360$ at 120 Hz. All encoder and aggregation comparisons within a dataset use this preprocessing and the same stimulus timings.

MOUS alignment. Sentence blocks are identified from story and word-list segments in BIDS `events.tsv`. Whisper large-v2, with Dutch language and word timestamps enabled, supplies word onsets within each stimulus audio block. These onsets define the 3 s neural windows of Eq. 3, with delay $\delta = 150$ ms. The MOUS transfer results evaluate the fixed Gwilliams configuration using this alignment.

A.4 Audio representations

The frozen wav2vec2-large-xlsr-53 encoder receives 3 s audio windows resampled to 16 kHz and waveform-normalised. We average hidden activations from layers 14–18 across layers while retaining their temporal structure. We resample the resulting sequence to $T = 360$, giving each target a shape of 1024×360. This configuration is shared across all three datasets.

B Retrieval model and training

B.1 Encoder configurations

All backbones use the coordinate-conditioned spatial frontend of Défossez et al. (2023). A small MLP maps 2D Fourier features of sensor coordinates to mixing weights, producing 270 virtual channels from the native sensors (208 for Gwilliams, 272 for MOUS, and 60 for Brennan). A subject-specific layer and temporal backbone map this grid to a $1024 \times T$ neural representation.

The temporal hidden width is 320, with a final 1×1 projection to $D = 1024$. Dense-TCNN uses five residual blocks, each with two kernel-3 convolutions at dilations 2 and 4. The five-block exp-dilated encoder uses DilatedGLU blocks with kernel size 3 and dilations $[1, 2, 4, 8, 16]$.

Receptive fields and depth. For unstrided convolutions, $R = 1 + \sum_i (k-1)d_i$. At $k = 3$, Dense-TCNN has $R = 61$ time steps and the five-block exp-dilated model has $R = 63$. The ten-block

exp-dilated variant changes temporal depth while retaining preprocessing, audio targets, loss, training protocol, and aggregation settings. Table 2 reports both depths. Brennan uses the same encoder configurations with its native input channel count.

B.2 Training and compute

Training updates the neural encoder while keeping the audio encoder frozen. Scores use the time-resolved neural representation and Frobenius-L_2-normalised audio targets, as in Eq. 5. The one-directional neural-to-audio contrastive loss uses in-batch audio negatives. Adam uses a learning rate of 3×10^{-4}, zero weight decay, and an effective batch size of 256. A 10% linear warm-up precedes cosine decay; training runs for up to 100 epochs with validation-loss early stopping (patience 15) and gradient-norm clipping at 1.0.

Training uses PyTorch on one NVIDIA A100-40GB GPU in full precision. Retrieval encoders typically train in about 3 hours; the length-shortcut diagnostic runs within 24 hours. Aggregation analyses operate on cached local logits.

C Structural and text-side diagnostics

C.1 Length-shortcut diagnostic in a Whisper-based pipeline

To test whether input duration supports decoding without neural content, we reproduce the representation stage of a Whisper-based neural-to-text pipeline (Li et al., 2025) on Gwilliams. Vector quantisation maps the Whisper encoder representations to a 2,048-entry codebook, and we evaluate the representation before the text decoder. Both conditions use a subject split: held-out participants are unseen in training, while every evaluation sentence has been heard by training participants, since all participants hear the same stories. In the variable-length condition, sentence-aligned inputs retain natural duration through padding and valid-time masks, and the readout ranks the correct sentence among the evaluation sentences (chance 0.8%). In the fixed-duration condition, inputs are 3 s windows and the readout ranks the correct next code among the 2,048 codebook entries (chance 0.05%). The two conditions are therefore separate diagnostics with their own chance levels; the fixed-duration condition tests only whether noise still scores high once duration is fixed. In the noise condition, we replace the neural values within each input's valid length with standard Gaussian noise at evaluation; padding and masks remain unchanged. Figure 2a reports real-neural and Gaussian-input results for both, and Table 3 gives the full results of the reproduction, including further splits. The noise tests reported for such pipelines (Yang et al., 2026b; Li et al., 2025) replace the whole input with Gaussian noise and score the generated text; this description follows their public code, in which the padded input tensor is replaced by noise of the same shape, so the padding that marks each input's duration is replaced as well.

C.2 Gaussian inputs in fixed-duration retrieval

For the audio-retrieval control, we replace each Gwilliams test window with Gaussian noise before it enters the trained Dense-TCNN encoder. Participant metadata, sensor coordinates, candidate pool, and evaluation split remain fixed. We compare standard Gaussian inputs with Gaussian inputs matched to each window's mean and standard deviation. We score both before and after aggregation. Retrieval is near chance in both conditions, with no consistent aggregation gain (Table 4). This table reports seed 0; Table 2 reports standard Gaussian inputs over three seeds on both datasets, where aggregation changes R@1 by -0.01 ± 0.01 points on Gwilliams and 0.00 ± 0.01 on MOUS.

C.3 Audio–text label mismatch

Table 5 compares stimulus content with an automatically aligned transcript used as a training label. Dropping the negation changes the meaning, while repeated words replace part of the clause. Such mismatches affect text-target evaluation even when a prediction agrees with the audio; in this study, the retrieval target is the aligned audio segment itself.

Table 3: **Reproduction of the Whisper-based pipeline: retrieval accuracy on held-out participants.** Top: variable-length inputs, ranking the correct sentence among the evaluation sentences (chance 0.79%). Splits follow Li et al. (2025): the subject split holds out participants, and the sentence split holds out 10% of sentences per participant, so that the training set may still cover every sentence; the story split holds out whole stories. These are the reproduced pipeline splits and differ from the stimulus-identity split used for attribution (Sec. 3.2). Bottom: fixed-duration inputs, ranking the next code among 2,048 codebook entries (chance 0.05%). The two panels use different readouts and are not directly comparable. Values are mean±std where available.

Variable-length inputs	**R@1 (%)**	**R@5 (%)**	**R@10 (%)**
Real MEG, subject split (Li et al., 2025)	90.79 ± 3.89	99.05 ± 1.40	99.42 ± 1.08
Valid-length Gaussian noise	66.34 ± 3.03	74.27 ± 2.78	79.90 ± 2.11
Real MEG, sentence split (Li et al., 2025)	56.31 ± 23.13	59.02 ± 22.97	60.61 ± 20.52
Real MEG, story split	7.06 ± 0.68	11.62 ± 1.16	24.18 ± 1.67
Random chance	0.79 ± 0.02	3.93 ± 0.09	7.85 ± 0.18
Fixed-duration inputs	**R@1 (%)**	**R@5 (%)**	**R@10 (%)**
Real MEG	5.10	9.81	11.28
Valid-length Gaussian noise	1.53	3.07	3.56
Random chance	0.05	0.24	0.49

Table 4: **Gaussian-input retrieval with fixed-duration windows on Gwilliams.** The trained encoder is kept fixed, and each test MEG window is replaced by signal-blind Gaussian noise. Chance is 0.068% for R@1, 0.342% for R@5, and 0.683% for R@10 with $N = 1{,}464$.

Input	**Method**	**R@1 (%)**	**R@5 (%)**	**R@10 (%)**
Gaussian $N(0, 1)$	Local	0.079	0.335	0.700
Gaussian $N(0, 1)$	+ aggregation	0.063	0.316	0.669
Mean/std-matched Gaussian	Local	0.054	0.321	0.676
Mean/std-matched Gaussian	+ aggregation	0.060	0.311	0.673
Random chance	–	0.068	0.342	0.683

Table 5: **Illustrative label noise from forced alignment.** The automated transcript drops the negation "didn't" (inverting meaning) and collapses a clause into a repetitive loop.

Actual audio content:
"...the short-eared people didn't really know the secrets of the long ears and had pulled down all the moai statues and destroyed some of the tablets."
Force-aligned transcript (training label):
*"...the short eared people really know the secrets **the the the**"*

D SENTENCE-BUCKET AGGREGATION

D.1 GROUPING AND SCORE-UPDATE ORDER

Query groups specify which neural windows contribute to a shared estimate. Candidate buckets specify which audio windows receive their bias. Queries are grouped by presented sentence, and candidates by source sentence. Stimulus-side timings align each fixed-duration neural query with its audio target. The local model first scores every query–candidate pair. Aggregation then infers the correspondence between query groups and candidate buckets from those scores. The oracle diagnostic supplies the correct candidate bucket and ranks within it.

Grouping perturbations. A story is one audio file heard by one participant in one session. Boundary jitter moves each selected query window to the neighbouring sentence of the same story, and

reassignment moves it to a random other sentence of the same story; In each sentence group G, $\lfloor p\,|G|\rfloor$ windows are moved, so the moved share of the 71,736 test windows is 6.3%, 16.6%, 25.9%, 36.3%, and 47.7% for $p = 0.1, \ldots, 0.5$ (34,251 windows at $p = 0.5$); with the same p, both perturbations move the same windows, and every moved window changes sentence. Local logits are unchanged throughout.

Pooling in full. This paragraph gives the pooling behind the prior π_G of Eq. 7. For a query group G and each $w \in G$, let $\mathcal{I}_w$ be the K_{top} highest-scoring candidates and τ_w the κ-quantile of $\ell_{w,\cdot}$ over the full candidate set; the retained set and excess scores are

$$\mathcal{J}_w = \{j \in \mathcal{I}_w : \ell_{w,j} \geq \tau_w\}, \qquad e_{w,j} = \ell_{w,j} - \tau_w. \tag{10}$$

The retained evidence assigned to bucket s is

$$\mathcal{E}_G(s) = \{e_{w,j} : w \in G,\ j \in \mathcal{J}_w,\ \sigma(j) = s\}, \tag{11}$$

and its support is

$$U_G(s) = |\mathcal{B}_s|^{-1/2}\ \text{MeanTop}_m\left(\mathcal{E}_G(s)\right), \tag{12}$$

where MeanTop_m averages the m largest values and $U_G(s) = 0$ when $\mathcal{E}_G(s)$ is empty. With $\mathcal{S}_G$ the S buckets of largest support, the prior of Eq. 7 is $\pi_G(s) = U_G(s)\,\mathbf{1}[s \in \mathcal{S}_G]$.

For each query group, aggregation therefore collects above-quantile excess scores among each window's top-K_{top} candidates, aggregates the top-m values per bucket with bucket-size normalisation, and selects the top-S supported buckets. It adds their shared biases to the local logits using Eq. 7. We compute support from the original logits before applying any update. We retain the full candidate pool and within-bucket ordering.

Removing cross-window pooling. Single-window aggregation replaces every query group with a singleton $G = \{w\}$. Evidence extraction, normalisation, bucket selection, and bias strength remain the same. Comparing the two aggregation conditions measures pooling's contribution across windows.

D.2 Aggregation settings and sensitivity to bias strength

The configuration uses $K_{\text{top}} = 128$, $\kappa = 0.95$, $m = 3$, $S = 3$, and square-root bucket-size normalisation (Eq. 12), selected on Gwilliams validation; $\gamma_{\text{agg}} = 0.7$ was fixed in advance. We hold this configuration fixed for test evaluation and transfer to MOUS. Table 6 reports representative aggregation configurations and their sentence-bucket retrieval rates.

Table 6: **Representative aggregation validation configurations.** Pool: how the top-m excess scores of a bucket are combined (LSE: log-sum-exp); Norm: the quantity by which bucket support is divided (kept count: the number of retained scores in the bucket). Bucket@K: share of query groups whose correct sentence is among the K best-supported buckets. The reported test configuration uses $K_{\text{top}} = 128, \kappa = 0.95, m = 3, S = 3$, square-root bucket-size normalisation, and $\gamma_{\text{agg}} = 0.7$ (bold).

K_{top}	κ	Pool	m	Norm	S	Bucket@1 (%) ↑	Bucket@5 (%) ↑
64	0.95	mean	3	$\sqrt{\lvert\mathcal{B}_s\rvert}$	3	89.0	95.0
128	0.95	LSE	1	$\sqrt{\lvert\mathcal{B}_s\rvert}$	3	90.0	96.0
256	0.98	mean	3	$\sqrt{\lvert\mathcal{B}_s\rvert}$	3	89.0	94.0
128	0.95	mean	3	$\sqrt{\lvert\mathcal{B}_s\rvert}$	3	**92.0**	**95.0**
128	0.95	mean	3	none	3	82.0	87.0
128	0.95	mean	3	kept count	3	75.0	85.0
128	0.95	mean	3	$\lvert\mathcal{B}_s\rvert$	3	41.0	75.0

By Eq. 8, γ_{agg} is the inverse temperature of the contextual factor p_G, with local logits held at their original scale.

A subsequent sensitivity analysis varies only γ_{agg} on one Gwilliams validation checkpoint, holding local scores, bucket support, and the candidate pool fixed (Table 7). At the fixed scale of 0.7, aggregation gives a 9.3-point R@1 gain with 3,605 corrections and 213 regressions. Neighbouring

scales of 0.6 and 0.8 yield gains of 8.3 and 10.3 points. At 2.2, the gain rises to 15.3 points, with 1,050 regressions and changes to 35.0% of top-1 decisions. Stronger weighting changes the extent of correction while leaving the evidence used to select buckets unchanged.

Table 7: **Sensitivity to aggregation bias strength on Gwilliams validation.** One Dense-TCNN checkpoint; 36,358 queries, 742 candidates, and 2,891 query sentence groups. Only the aggregation weight γ_{agg} varies. The local R@1 is 52.6%; gains are percentage points and other rates are percentages; Corr./Regr. is the ratio of corrections to regressions. Bold marks the scale fixed in advance.

γ_{agg}	**ΔR@1 (pp)**	**Corrections**	**Regressions**	**Corr./Regr.**	Regressions / local correct (%)	top-1 changed (%)
0.0	+0.0	0	0	–	0.0	0.0
0.4	+5.9	2,260	112	20.2	0.6	8.8
0.6	+8.3	3,186	184	17.3	1.0	12.7
0.7	**+9.3**	**3,605**	**213**	**16.9**	**1.1**	**14.6**
0.8	+10.3	3,974	244	16.3	1.3	16.3
1.1	+12.6	4,952	358	13.8	1.9	21.3
1.4	+14.1	5,644	529	10.7	2.8	25.6
2.2	+15.3	6,606	1,050	6.3	5.5	35.0

D.3 Aggregation components and pruning

Component ablations. Table 8 examines the components of aggregation using a fixed Dense-TCNN retrieval model and its frozen Gwilliams test logits, so every variant receives the same local evidence. Removing cross-window pooling (single-window aggregation, $G = \{w\}$) drops R@1 to 42.3%, slightly below local retrieval (44.1%). With this rule, pooling within a single window does not improve retrieval; restoring pooling adds $+10.4$ pp R@1 and $+8.6$ pp MRR (paired sentence-cluster bootstrap, $p < 0.001$). Removing the quantile gate drops R@1 to 37.8%, showing that selecting high-evidence logits matters for this aggregation rule. Removing bucket-size normalisation raises R@1 to 71.2%, and this increase correlates with bucket size (Spearman $\rho = 0.27$, $p < 0.001$). Without normalisation, pooled support grows with the number of candidates in a sentence bucket, making the correction sensitive to sentence length. Bucket-size normalisation is chosen to remove this dependence (Sec. 3.4); Appendix E.5 tests whether any structural use remains.

Table 8: **Aggregation-rule interventions with frozen local scores.** Variants use the same Dense-TCNN checkpoint on the Gwilliams test set. Table 2 reports three-seed means. Retrieval rates are percentages; MedR is in rank units. The w/o norm† row removes bucket-size normalisation. For hard pruning, a target outside the retained set contributes zero to R@K and reciprocal rank.

Scoring rule	**R@1 (%)**	**R@5 (%)**	**R@10 (%)**	**MRR (%)**	**MedR**
Component ablations					
Local only	44.1	65.7	73.6	54.2	2
w/o cross-window	42.3	65.0	73.4	52.8	2
w/o gate	37.8	64.0	72.8	49.8	3
Full aggregation	52.7	71.4	77.7	61.4	1
w/o norm†	71.2	85.6	88.5	77.9	1
Pruning comparison (S=3)					
Hard pruning	50.5	56.5	56.9	53.2	1

Aggregation vs. hard pruning. Hard pruning improves R@1 over local retrieval but lowers R@5, R@10, and MRR because it excludes targets outside the selected buckets. Aggregation retains every candidate and yields higher values on all four metrics. The same cross-window support therefore has different effects when it shifts bucket scores versus discards candidates.

E CONTEXTUAL GAIN AND ITS DEPENDENCE ON LOCAL EVIDENCE

The following analyses relate changes in retrieval rank to bucket selection. They also examine the aggregation rule under reduced local evidence. Local retrieval and aggregation use the same local logits within each comparison.

E.1 FULL METRICS AND SEED-LEVEL INTERVALS

Table 9 gives the full metrics before and after aggregation that Table 2 summarises.

Table 9: **Retrieval before and after sentence-bucket aggregation, in full.** Each backbone is evaluated before and after applying aggregation to the same frozen retrieval logits, so the contrast gives the contextual gain while holding the encoder, candidate pool, and local logits fixed. Exp-dilated (L=5) is the receptive-field-matched encoder of Défossez et al. (2023) and (L=10) the deeper variant. Arrows indicate the preferred direction. R@1, MRR, and R@10 are reported as percentages; Δ columns report absolute percentage-point gains. R@1, MRR, R@10, and Δ are mean±std over three seeds, while MedR is the mean over three seeds. Bold: the best value within each dataset block and column.

Setting	Backbone	Before aggregation				After aggregation				Change	
		R@1 (%)↑	MRR (%)↑	R@10 (%)↑	MedR↓	R@1 (%)↑	MRR (%)↑	R@10 (%)↑	MedR↓	ΔR@1 (pp)↑	ΔMRR (pp)↑
Gwilliams N=1,464	Dense-TCNN	**44.4±0.3**	**54.4±0.2**	**73.6±0.1**	**2.0**	**52.1±0.6**	**60.8±0.5**	**77.3±0.3**	**1.0**	7.7±0.9†	6.4±0.7†
	Exp-dilated (L = 5)	41.9±0.9	51.9±0.9	71.1±0.7	**2.0**	50.6±0.7	59.2±0.7	75.4±0.6	1.3	**8.7±0.6**†	**7.4±0.6**†
	Exp-dilated (L = 10)	39.6±0.3	49.7±0.2	69.2±0.1	3.0	46.2±0.6	55.3±0.6	72.7±0.5	2.0	6.6±0.8†	5.7±0.7†
MOUS N=825	Dense-TCNN	**21.7±0.4**	**30.2±0.5**	**46.6±0.5**	**13.3**	**28.6±0.6**	**37.0±0.5**	**52.8±0.3**	**8.3**	**6.9±0.2**†	**6.9±0.1**†
	Exp-dilated (L = 5)	16.1±0.8	24.0±0.7	39.5±1.0	21.7	21.5±0.8	29.3±0.8	44.0±1.1	16.7	5.4±0.1†	5.3±0.1†
	Exp-dilated (L = 10)	13.4±0.9	20.8±1.1	35.1±1.3	30.0	17.9±1.2	25.2±1.3	38.8±1.6	24.7	4.5±0.3†	4.4±0.2†

†On Δ columns, the dagger indicates that the corresponding improvement is significant in all seed-specific paired sentence-cluster bootstrap tests at $p < 0.05$; full intervals are reported in Table 10.

Table 10 reports window-weighted point estimates and paired sentence-cluster bootstrap intervals for aggregation gains, separately for each dataset, backbone, and seed. Each bootstrap sample resamples sentences and computes the difference in retrieval rates over their query windows. The three-seed means reproduce Table 2, and every 95% interval is above zero.

E.2 CONTEXTUAL GAIN ACROSS PARTICIPANTS

Per-participant statistics use the frozen local logits and aggregation outputs of Table 2, computed within each seed and averaged over the three seeds; no participant is excluded. Window-weighted means over participants reproduce Table 2. Every participant is above chance locally (lowest Dense-TCNN local R@1: 14.0% on Gwilliams, 2.4% on MOUS, against 0.07% and 0.12% chance), and with Dense-TCNN the aggregation gain is positive for every participant on both MEG datasets and rises with local R@1 (Table 11; Figure 3c). Dense-TCNN exceeds the five-block exp-dilated encoder in local R@1 for all 27 Gwilliams participants and for 91 of 96 MOUS participants.

E.3 WITHIN-MEG EVIDENCE ATTENUATION

We weaken target-specific evidence on Gwilliams while holding the encoder, candidate pool, sentence groups, and aggregation parameters fixed. For a query x_w, a donor $\tilde{x}_w$ is sampled from another sentence, preferring the same participant, and mixed after preprocessing:

$$x_w^{(\alpha)} = \sqrt{\alpha}\, x_w + \sqrt{1-\alpha}\, \tilde{x}_w, \qquad \alpha \in \{0.75, 0.50, 0.25, 0.10, 0.00\}. \tag{13}$$

The clean query corresponds to $\alpha = 1$; at zero, the input contains only the stimulus-mismatched donor. Phase-randomised MEG and spatial-covariance-matched Gaussian surrogates use the same mixing rule. The former preserves channel-wise spectral magnitude; the latter matches sensor covariance and query RMS amplitude.

A fourth perturbation acts directly on local logits. Clean and independently candidate-permuted rows are standardised, mixed, and returned to the clean row's scale:

$$\tilde{L}^{(\alpha)} = \mu_L + \eta_L \left(\sqrt{\alpha}\, Z_L + \sqrt{1-\alpha}\, Z_{\text{perm}} \right), \tag{14}$$

Table 10: **Window-weighted aggregation gains with paired sentence-cluster bootstrap intervals.** Point estimates and 95% confidence intervals are in percentage points. Each comparison uses the same frozen local logits. Query windows have equal weight within each resampled set of sentences; seeds 1–3 correspond to run seeds 0–2.

Backbone	Seed	ΔR@1	95% CI	ΔMRR	95% CI
Gwilliams					
Dense-TCNN	1	8.61	[7.62, 9.64]	7.21	[6.40, 8.02]
Dense-TCNN	2	7.40	[6.48, 8.41]	6.23	[5.46, 7.04]
Dense-TCNN	3	6.96	[6.08, 7.90]	5.86	[5.16, 6.62]
Exp-dilated ($L = 5$)	1	8.91	[7.88, 9.94]	7.54	[6.70, 8.40]
Exp-dilated ($L = 5$)	2	8.04	[7.07, 9.09]	6.72	[5.90, 7.58]
Exp-dilated ($L = 5$)	3	9.15	[8.19, 10.16]	7.81	[7.02, 8.63]
Exp-dilated ($L = 10$)	1	5.79	[4.85, 6.81]	4.96	[4.20, 5.78]
Exp-dilated ($L = 10$)	2	6.58	[5.68, 7.57]	5.64	[4.89, 6.46]
Exp-dilated ($L = 10$)	3	7.36	[6.39, 8.40]	6.35	[5.55, 7.21]
MOUS					
Dense-TCNN	1	6.70	[6.20, 7.21]	6.84	[6.39, 7.30]
Dense-TCNN	2	6.92	[6.37, 7.49]	6.95	[6.51, 7.42]
Dense-TCNN	3	7.00	[6.49, 7.50]	6.85	[6.40, 7.29]
Exp-dilated ($L = 5$)	1	5.50	[4.98, 6.00]	5.38	[4.91, 5.83]
Exp-dilated ($L = 5$)	2	5.27	[4.82, 5.72]	5.22	[4.80, 5.66]
Exp-dilated ($L = 5$)	3	5.43	[4.93, 5.91]	5.21	[4.75, 5.67]
Exp-dilated ($L = 10$)	1	4.63	[4.10, 5.16]	4.56	[4.09, 5.03]
Exp-dilated ($L = 10$)	2	4.16	[3.65, 4.68]	4.21	[3.73, 4.70]
Exp-dilated ($L = 10$)	3	4.65	[4.21, 5.09]	4.49	[4.13, 4.86]

Table 11: **Per-participant local retrieval and aggregation gain.** Medians over participants; $\Delta > 0$ is the share of participants with a positive gain; Spearman ρ correlates each participant's local R@1 with the participant's gain.

Dataset	Backbone	Participants	Median local R@1 (%)	Median ΔR@1 (pp)	$\Delta > 0$ (%)	Spearman ρ (p)
Gwilliams	Dense-TCNN	27	44.6	+7.9	100	0.88 (1.9×10^{-9})
MOUS	Dense-TCNN	96	19.7	+6.8	100	0.90 (2.5×10^{-35})
Gwilliams	Exp-dilated ($L = 5$)	27	42.9	+9.1	100	0.85 (1.4×10^{-8})
MOUS	Exp-dilated ($L = 5$)	96	14.3	+5.3	100	0.93 (1.6×10^{-43})
Gwilliams	Exp-dilated ($L = 10$)	27	40.9	+6.8	100	0.83 (6.3×10^{-8})
MOUS	Exp-dilated ($L = 10$)	96	11.8	+4.0	99	0.93 (4.3×10^{-43})

where μ_L and η_L are the mean and standard deviation of the clean row, and Z_L and Z_{perm} are the standardised clean and permuted rows. Neural-space perturbations precede the frozen encoder; logit permutation follows it.

Figure 3 reports both controls. Figure 4 adds MRR under grouping perturbations and the real-MEG attenuation trajectory.

Under real-MEG mixing, query-weighted aggregation gains fall from 8.6 points on clean MEG to 6.6, 2.7, and 0.3 points at $\alpha = 0.75, 0.50, 0.25$, respectively. They round to zero at $\alpha \leq 0.10$ (Table 12). The same α causes different degradation across perturbations, so Figure 3b compares all four families against achieved local R@1. In each family, the gain approaches zero as local retrieval approaches chance.

Sentence-weighted uncertainty. Table 13 gives each of the 123 sentences equal weight, whereas Table 12 gives each query equal weight. At $\alpha = 0.10$, sentence-weighted gains remain nonzero but negligible (0.21 R@1 points and 0.31 MRR points).

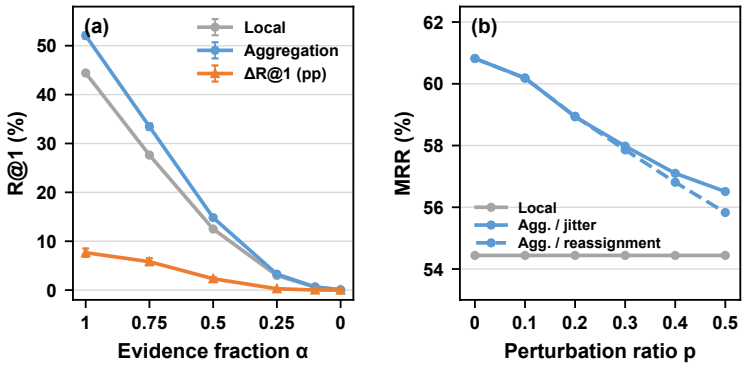


Figure 4: **Additional readouts of the grouping and evidence controls.** (a) Local and aggregated R@1 and their difference under real-MEG mixing as the evidence fraction α decreases (three seeds; error bars are the standard deviation over seeds). (b) MRR under boundary jitter and within-story reassignment, with local logits fixed (three-seed means). The corresponding R@1 results appear in Figure 3.

Table 12: **Within-Gwilliams real-MEG evidence attenuation.** Each attenuated query mixes the original MEG window with a stimulus-mismatched real MEG window from the same evaluation set. We fix the encoder, candidate pool, sentence grouping, and aggregation hyperparameters. Values are query-weighted over Q=71,736 windows from one checkpoint. All retrieval metrics are reported as percentages; deltas denote absolute percentage-point differences computed before rounding.

Condition	Local R@1 (%)	+ aggregation R@1 (%)	ΔR@1 (pp)	ΔMRR (pp)
Clean MEG	44.1	52.7	+8.6	+7.2
Real-MEG mix α=0.75	27.5	34.2	+6.6	+6.1
Real-MEG mix α=0.50	12.4	15.1	+2.7	+2.8
Real-MEG mix α=0.25	3.0	3.3	+0.3	+0.4
Real-MEG mix α=0.10	0.6	0.6	+0.0	+0.0
Real-MEG mix α=0.00	0.1	0.1	+0.0	+0.0

Table 13: **Sentence-weighted gains under real-MEG attenuation.** Paired bootstrap over 123 sentence buckets, with equal sentence weights. Deltas and 95% intervals are percentage points; $^*p < 0.001$.

Condition	Metric	Δ (pp)	95% CI (pp)
Clean MEG	R@1	+10.78*	[9.99, 11.59]
Clean MEG	MRR	+8.94*	[8.31, 9.57]
Real-MEG mix α=0.75	R@1	+9.34*	[8.43, 10.30]
Real-MEG mix α=0.75	MRR	+8.49*	[7.73, 9.28]
Real-MEG mix α=0.50	R@1	+5.06*	[4.27, 5.93]
Real-MEG mix α=0.50	MRR	+5.07*	[4.35, 5.84]
Real-MEG mix α=0.25	R@1	+1.06*	[0.74, 1.44]
Real-MEG mix α=0.25	MRR	+1.28*	[0.93, 1.68]
Real-MEG mix α=0.10	R@1	+0.21*	[0.08, 0.38]
Real-MEG mix α=0.10	MRR	+0.31*	[0.17, 0.48]
Real-MEG mix α=0.00	R@1	+0.02	[0.00, 0.05]
Real-MEG mix α=0.00	MRR	+0.02	[0.00, 0.05]

E.4 BRENNAN EEG AND CANDIDATE-POOL SIZE

The Brennan evaluation applies the same retrieval and aggregation procedure to a different modality, participant group, and stimulus set. At $N = 388$, local R@1 is 0.38% for the exp-dilated model and 0.72% for Dense-TCNN, against 0.26% chance. Aggregation reduces these to 0.31% and 0.21%, respectively (Table 14). Smaller candidate pools improve absolute retrieval, but aggregation gains remain negative for both backbones (Figure 5). This pattern is consistent with bucket biases reinforcing incorrect hypotheses when the local scores provide weak target-specific support. Earlier speech retrieval also found EEG substantially weaker than MEG (Défossez et al., 2023). Sentence grouping remains available in Brennan but yields no positive contextual gain. The within-Gwilliams attenuation experiment examines dependence on local evidence in the dataset, with the encoder held fixed.

Table 14: **Brennan EEG retrieval with 388 candidates.** Retrieval metrics are percentages; parentheses give the multiple of the corresponding random baseline. Single seed.

Backbone	R@1 (%)	R@5 (%)	R@10 (%)	MRR (%)
Random chance	0.26	1.29	2.58	1.69
Exp-dilated	0.38 (1.5×)	1.72 (1.3×)	3.47 (1.3×)	2.04 (1.2×)
Exp-dilated + aggregation	0.31 (1.2×)	1.69 (1.3×)	3.39 (1.3×)	1.98 (1.2×)
Dense-TCNN	0.72 (2.8×)	3.59 (2.8×)	7.21 (2.8×)	3.88 (2.3×)
Dense-TCNN + aggregation	0.21 (0.8×)	1.18 (0.9×)	2.37 (0.9×)	1.57 (0.9×)

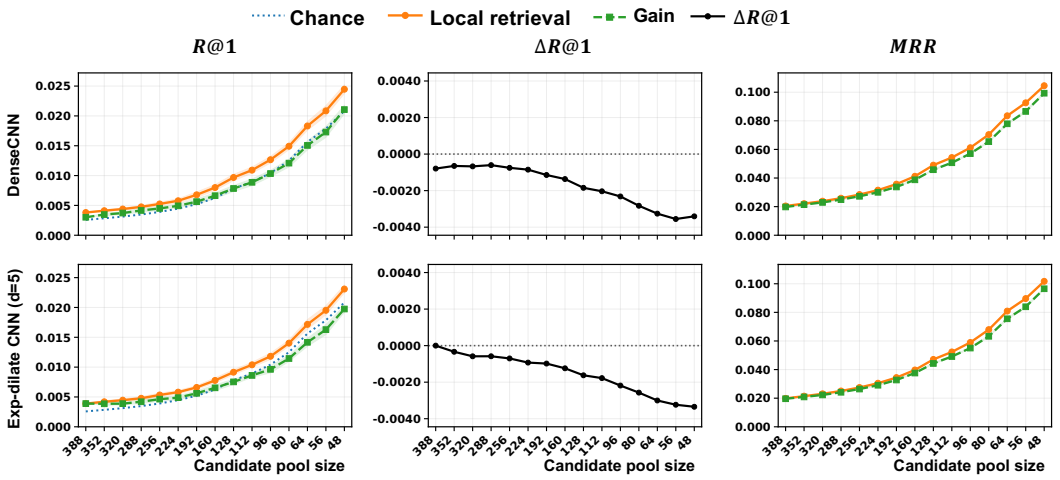


Figure 5: **Brennan EEG across candidate-pool sizes.** Before and After denote retrieval before and after aggregation. Shrinking the pool raises absolute R@1 and MRR, while the aggregation gain remains negative for both backbones. Single seed.

E.5 SENTENCE LENGTH AS AN IDENTIFYING CUE

Window counts expose sentence structure even when neural evidence is weakened. A size-matching control selects three candidate sentence buckets from their window counts relative to the query-group size, without using neural values or local scores to select buckets. It then ranks the retained candidates by the same perturbed local scores as the aggregation comparison. A target outside the retained set contributes zero to R@K and reciprocal rank. Score-based hard pruning uses the three buckets selected by aggregation with this same update and scoring rule.

Across the 123 test sentence buckets, size matching places the target bucket among the selected three for 65.0% of query windows, compared with 2.44% for uniform selection. Its sentence-cluster bootstrap interval is [58.3, 71.2]%. This selection depends only on window counts, so it does not

change with attenuation or decoder seed. Table 15 reports three-seed, window-weighted results under real-MEG mixing. With the update form matched to hard pruning, size-based and score-based selection give gains of 16.8 and 1.2 points at $\alpha = 0.5$. The score-based bucket hit rate falls as neural evidence weakens, while size-based selection retains the structural cue.

Table 15: **Bucket selection and retrieval under real-MEG attenuation.** Three decoder seeds; query windows have equal weight. Bucket hit rates are percentages; gains over local retrieval are percentage points. Size matching has a 65.0% bucket hit rate at every attenuation level. The diagnostic in Table 12 uses one checkpoint; this table averages all three. At $\alpha = 1$, Score + hard varies across seeds ($+6.4$, $+2.5$, $+0.5$): a missed bucket counts as an error, so the spread in score-based bucket hit (50.4–57.0%) moves its R@1 by several points.

	Bucket hit (%)		ΔR@1 (pp)		
α	Score-based	Size-based	Aggregation	Score + hard	Size + hard
1.00	53.4	65.0	7.7	3.1	5.6
0.75	38.0	65.0	5.8	3.7	13.6
0.50	19.3	65.0	2.3	1.2	16.8
0.25	4.7	65.0	0.28	-0.6	12.3
0.10	1.3	65.0	0.02	-0.3	6.6
0.00	0.50	65.0	0.00	-0.02	1.5

At $\alpha = 0$, size matching reaches 1.52–1.61% R@1 across the four attenuation families, about 22–24 times uniform full-pool chance (Table 16). For hard selection, $\Pr(\text{correct}) = \Pr(\text{target retained})\Pr(\text{correct} \mid \text{target retained})$: bucket selection can preserve an advantage even when ranking within the retained set has little target-specific information. Both rules preserve within-bucket order; the gain remaining after attenuation distinguishes their use of structural information.

Table 16: **Retrieval at $\alpha = 0$.** R@1 values are three-seed means in percentage. Size matching uses hard selection; its paired gain over Local is mean±std across seeds, in percentage points. Uniform full-pool chance is $1/1{,}464 = 0.068\%$.

Perturbation	Local	Aggregation	Size matching	ΔR@1 (pp)
Real-MEG mixing	0.050	0.052	1.52	1.47 ± 0.03
Phase randomisation	0.075	0.077	1.58	1.51 ± 0.05
Covariance Gaussian	0.068	0.066	1.61	1.54 ± 0.02
Logit permutation	0.072	0.072	1.61	1.54 ± 0.02

The mean aggregation gains remain near zero. Two of the twelve seed-specific 95% intervals for aggregation exclude zero, both under logit permutation: $+0.007$ points $[0.001, 0.014]$ and -0.006 points $[-0.011, -0.001]$. These small effects differ in scale from the size-matching gains. This control and the duration diagnostic expose two length cues of different origin. Input duration in the end-to-end pipeline is learned in training and pays off when evaluation sentences are seen in training. Sentence length in contextual retrieval needs no training: it follows from cutting query groups and candidates at the same sentence boundaries in a closed set, so any sentence-level context module under these conditions can match on it, whether through buckets or through scores over candidate sentences. A clean split removes the first cue but not the second; the criterion detects the second. The two diagnostics also differ in task, candidate set, and outcome measure. The present control demonstrates that window counts are identifying; it does not numerically explain the 66.3% result.

E.6 Bucket selection and the oracle gap

The component diagnostics use one fixed Dense-TCNN checkpoint with local R@1 of 44.1% and aggregation R@1 of 52.7%; the main table averages three seeds (44.4% and 52.1%). Among 71,736 queries, aggregation corrects 6,724 local errors and overturns 549 correct windows, giving 6,175 net corrections. Table 18 groups these corrections by the target's initial rank.

At the same diagnostic checkpoint, ranking within the correct sentence bucket yields an oracle R@1 of 87.9% (R@5 98.2%; R@10 99.5%). The correct bucket is selected among aggregation's top-S buckets for 57.0% of queries and 42.4% of local errors (Table 17).

Among local errors where the correct bucket is selected, aggregation reaches 39.6% R@1, hard pruning to the selected buckets reaches 72.7%, and the oracle reaches 87.2%. When the correct bucket is missed, aggregation corrects none of the local errors. The gap to the oracle therefore reflects missed correct buckets and remaining cross-bucket competition: of the 35.2 points between aggregation and oracle R@1, queries whose correct bucket is missed account for 23.8 and bucket-hit queries for 11.4 (Table 17, weighted by subset size). Within-bucket errors limit the oracle itself, since both comparisons preserve local order within each bucket. Increasing the bias scale changes only cross-bucket competition, leaving bucket selection and within-bucket ranks unchanged.

Table 17: **Decomposition of the gap between aggregation and oracle on Gwilliams (Dense-TCNN).** Bucket hit denotes whether the ground-truth sentence bucket is included in the top-S buckets selected by aggregation ($S = 3$). Conditional metrics are computed on the specified subset. All retrieval rates are reported as percentages.

Subset	# queries	Bucket hit (%)	Local R@1 (%)	Aggregation R@1 (%)	Hard-prune R@1 (%)	Oracle R@1 (%)
All queries	71,736	57.0	44.1	52.7	50.5	87.9
Bucket hit	40,894	100.0	58.5	74.7	88.7	94.7
Bucket miss	30,842	0.0	25.1	23.6	0.0	78.9
Local errors	40,079	42.4	0.0	16.8	30.8	78.3
Local errors, bucket hit	16,980	100.0	0.0	39.6	72.7	87.2
Local errors, bucket miss	23,099	0.0	0.0	0.0	0.0	71.8

Corrections concentrate near the top of the initial ranking: aggregation repairs 38.4% of rank 2–5 errors, 11.2% of rank 6–10 errors, and 0.7% beyond rank 10. The correct-bucket selection rate also falls with initial rank (Table 18), linking the correction pattern to both local candidate competition and bucket support.

Table 18: **Top-1 correction by initial local rank on Gwilliams (Dense-TCNN).** For local errors, we report how often aggregation promotes the target to the top. Bucket hit is the fraction of queries whose ground-truth sentence bucket is included in the top-S buckets selected by aggregation. All rates are reported as percentages.

Initial local rank	# queries	Corrected	Correction rate (%)	Bucket hit (%)	Oracle R@1 (%)
2–5	15,491	5,956	38.4	56.0	98.1
6–10	5,656	636	11.2	46.1	93.7
> 10	18,932	132	0.7	30.2	57.6

Where local errors fall. Over the three Dense-TCNN seeds, $1.01 \pm 0.08\%$ of top-1 errors fall on a candidate from the target's sentence (seed 0: 372 of 40,079). If the erroneous candidate were unrelated to sentence identity, the expected share would be the mean of $(|\mathcal{B}_{s^\star}| - 1)/(N - 1)$ over the errors, 1.1% for the 123 test sentences (11.9 candidates each on average). Chance R@1 within a sentence, the mean of $1/|\mathcal{B}_{s^\star}|$ over queries, is 8.4%.

E.7 Sentence length and contextual gain

Figure 6 groups sentences by their number of evaluation windows. Aggregation gains are larger for short-to-medium sentences and smaller in the longest bins. Sentence length varies together with query-group size, candidate-bucket size, and within-bucket difficulty. The figure therefore shows how contextual gain varies across these factors.

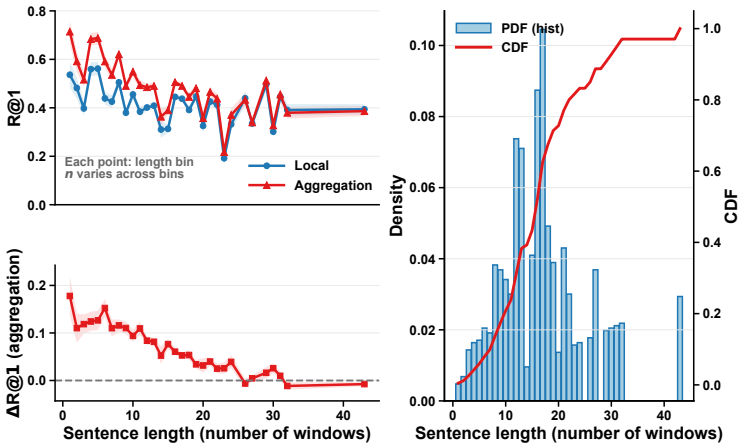


Figure 6: **Retrieval and contextual gain by sentence length on Gwilliams.** Left: local and aggregation R@1 by window-count bin, with their difference below. Right: sentence-length distribution. Each bin combines variation in query-group size, candidate-bucket size, and sentence content.