CONFIDENCE-ORDERING REVERSAL UNDER CONTEXTUAL PRIORS IN NEURAL DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Contextual priors improve neural-to-language decoding by reshaping candidate scores. However, confidence is read from the same reshaped scores, so the errors a prior leaves behind can become more confident with no change in accuracy to reveal it. In this work, we ask how a prior shapes the confidence of these errors, studying speech retrieval on MEG-MASC and MOUS with local decoding scores and a contextual prior combined by additive shallow fusion, and the fused top-two margin as confidence. Among initially incorrect predictions, we find a **confidence-ordering reversal**: a larger margin makes a *repair*, an error that fusion corrects, more likely when the correct candidate starts near the top of the local ranking, but less likely when it starts lower. On MEG-MASC, pooled correctness AUROC is 0.87, yet the AUROC separating repairs from *residual errors*, which fusion leaves uncorrected, falls from 0.70 at initial ranks 2–3 to 0.39 at ranks 21–50. Errors starting beyond rank 20, inside the reversed region, make up 46.6% of all errors after fusion. We propose a score-level account: a repair must first close the correct candidate's initial deficit, which limits its final margin, whereas a residual error can build a large margin between two incorrect candidates. Through a causal intervention that changes only the fusion weight, we show that the reversal moves to deeper ranks, as the account predicts, and that under a word-level LM prior it keeps moving after accuracy gain peaks, so a weight chosen for accuracy does not settle confidence. Guided by this account, we read local and prior scores separately: read before fusion, the prior's own scores already separate repairs from residual errors where the fused margin reverses, and estimators built on local and prior scores let a selective decoder answer on 74.5% of windows instead of 56.7%, with 92% of its output sets still containing the correct candidate. Confidence after contextual fusion should retain the local and contextual evidence behind each prediction, not just the fused scores. Project Page Code

1 INTRODUCTION

Contextual priors can improve neural-to-language decoding by supplying linguistic structure and sequential context that local neural predictions can miss (Willett et al., 2023; d'Ascoli et al., 2025). They also change the remaining errors. Recent MEG work reports fluent sentences whose content differs from the intended text when neural evidence is weak (Zhang et al., 2026a), and invasive brain-to-text systems exhibit confident but incorrect neural predictions (Huang et al., 2026). In sustained BCI use, users must identify and correct decoded errors (Speier et al., 2016; Card et al., 2026); when confidence determines which outputs are flagged for review, a confidently scored error can pass unflagged and be left for the user to catch. Evaluation nonetheless centres on decoding accuracy or text similarity (Silva et al., 2024). Where confidence does guide selective output (Kim et al., 2026), the prior shapes both the prediction and the fused scores used to judge it.

We find that strong overall confidence discrimination can conceal opposite orderings of correction outcomes. We study this in offline speech retrieval (Défossez et al., 2023), retaining a fixed candidate pool and separate local and prior scores (Figure 1). The *fused margin*, the gap between the two highest fused scores, serves as confidence. Among initially incorrect predictions, *repairs* become correct after fusion and *residual errors* remain incorrect. Across MEG-MASC (Gwilliams et al., 2023) and MOUS (Schoffelen et al., 2019), larger margins favour repairs when the correct candidate starts near the top of the local ranking, but residual errors when it starts lower (Figure 2). On MEG-

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.

MASC, pooled correctness AUROC is 0.87, yet repair-separation AUROC falls from 0.70 at ranks 2–3 to 0.39 at ranks 21–50. We call this a **confidence-ordering reversal**. The reversed region is where errors concentrate: residual errors starting beyond rank 20 account for 46.6% of all post-fusion errors on MEG-MASC. The pre-fusion margin favours residual errors in every group, so the positive ordering near the top is produced by fusion itself. **How does a prior make some errors more confident, and what information can still help identify them?**

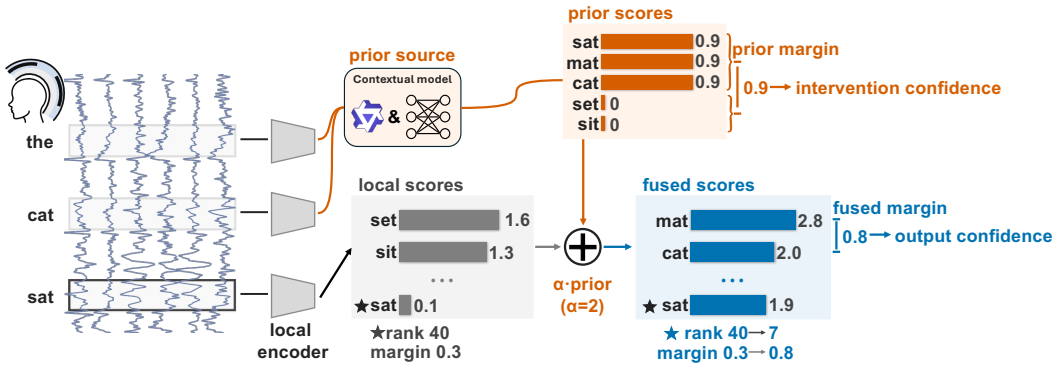

Figure 1: **Contextual fusion can make an error more confident.** A local encoder scores candidates from the current MEG window, while a contextual prior from preceding windows adjusts these scores through additive fusion with weight $\alpha = 2$. The prior operates at a coarser granularity, so several candidates share one score. Fusion advances the correct candidate ($\star$) from rank 40 to rank 7 without correcting the prediction, while increasing the top-two margin from 0.3 to 0.8. Brackets mark the prior and fused margins. Values are illustrative; real cases are in Table 6.

We trace the reversal to how repairs and residual errors form their margins. A repair must first close the correct candidate's initial score deficit, so its lead is bounded by the remaining contextual advantage, whereas a residual error can acquire a large lead between two incorrect candidates even as the correct candidate advances. Because this bound follows from additive fusion itself, it holds wherever a prior is added to local scores, and it predicts that increasing fusion weight extends positive repair ordering towards larger initial deficits.

We test this prediction with a causal intervention that varies only the fusion weight, keeping local and prior scores fixed. Under the history-aggregation prior, built from preceding windows, positive ordering extends from no rank group to ranks 11–20 and to larger score deficits, and the boundary never retreats in participant-bootstrap resamples (Figure 3). Under a word-level teacher-forced language-model prior, it also expands down the initial ranking and continues after net accuracy gain has peaked: **accuracy and confidence ordering are two different responses to the same prior.**

Grouping errors by initial rank exposes the reversal, and the margin-formation account locates its source: fusion merges local competition and prior support into one margin. Guided by this account, we read the two scores separately: the prior's own margin separates the outcomes that the fused margin reverses (Figure 2C), and learned estimators built on both recover positive repair ordering. Feature comparisons show that local features improve selective output beyond richer fused-score features, with further gains from prior features. At matched 92% coverage, the full estimator emits on 74.5% of test windows versus 56.7% for the fused margin. **Local and prior scores retain information for identifying confident errors that the final margin does not express.**

This paper proceeds in three steps.

Revealing where confident errors arise. In two MEG datasets, the fused margin orders repairs and residual errors oppositely by initial rank, even though pooled correctness AUROC stays high.

Explaining and testing causally. A score-level account of margin formation explains one way priors make errors more confident; a causal intervention on the fusion weight alone confirms the predicted extension of positive ordering, which accuracy gain does not track.

Identifying confident errors from retained information. Learned estimates recover positive repair ordering, and local and prior features improve selective output beyond richer fused-score features.

2 Related Work

Contextual priors and the errors they leave behind. Contextual priors support BCI communication (Speier et al., 2016; Herff et al., 2015), semantic reconstruction (Tang et al., 2023), and non-invasive decoding with sentence-level context or language-model assistance (d'Ascoli et al., 2025; Lévy et al., 2026). When neural evidence is weak, contextual support can dominate the final choice. Reviews and attribution studies already separate neural contributions from contextual gains (Silva et al., 2024; Zhang et al., 2026b; Gorenshtein et al., 2026a). Correctness alone does not establish how much neural evidence contributed, and the errors that remain take characteristic forms: non-invasive systems can produce fluent but incorrect text (Zhang et al., 2026a), while invasive decoders exhibit overconfident neural predictions (Huang et al., 2026). Evaluating contextual assistance therefore also requires assessing whether its remaining errors can be identified.

Confidence formation under contextual priors. Confidence already guides selective EEG output (Kim et al., 2026), language-model assistance (Huang et al., 2026), and error detection after P300 fusion (Gorenshtein et al., 2026b). In ASR, however, shallow fusion can improve recognition while degrading confidence (Kannan et al., 2018; Li et al., 2021; Futami et al., 2021). Beam analysis connects confidence to local and sequence-level competition (Jia & Van hamme, 2026), while selective classification reveals group differences concealed by aggregate performance (Jones et al., 2021). These observations share a source: under additive fusion, the prior changes both candidate competition and the scores that judge the result. What remains unclear is how confidence orders correction outcomes across groups of initial errors, and how that ordering changes with prior contribution.

Judging outputs and contextual revisions. In deployment, BCI users correct decoded outputs (Card et al., 2026), and shared-control systems arbitrate between neural commands and machine assistance (Lee et al., 2025; Deng et al., 2020). For such decisions, selective prediction and prediction sets formalise accepting, expanding, or withholding outputs (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017; Romano et al., 2020), while adaptive fusion adjusts language-model influence (Variani et al., 2022; Unni et al., 2022; Gong et al., 2022). For contextual revisions, retaining the local prediction provides the alternative, so confidence must support two judgements: final correctness and improvement over that prediction. Applying these decision rules requires knowing which information in local, prior, and fused scores supports each judgement after the prior has acted.

3 Method

3.1 Local scores and contextual fusion

Following the speech-retrieval framework of Défossez et al. (2023), each query is a word-aligned, 3-s MEG window paired with its correct audio segment in a fixed candidate pool $\mathcal{Y} = \{1, \dots, O\}$. At window index t, the local decoder maps the neural window x_t to scores $s_t \in \mathbb{R}^O$, while the contextual prior maps preceding history h_{t-1} to scores π_t over the same candidates. Let Y_t be the correct-candidate random variable and y_t its evaluation label; j and k index candidates. The distributions $q_t = \mathrm{softmax}(s_t)$ and $p_t = \mathrm{softmax}(\pi_t)$ define the local and prior model probabilities. Additive shallow fusion with weight $\alpha \geq 0$ (Kannan et al., 2018) gives

$$\tilde{s}_{t,j} = s_{t,j} + \alpha \pi_{t,j}, \tag{1}$$

$$\Pr_{\alpha}(Y_t = j \mid x_t, h_{t-1}) = \frac{q_t(j) p_t(j)^{\alpha}}{\sum_k q_t(k) p_t(k)^{\alpha}}. \tag{2}$$

Equation 2 is the softmax of Eq. 1, so fusion is a tempered product of the two distributions and reduces to the local one at $\alpha = 0$; for $\alpha > 0$ it gives the prior temperature $T_{\mathrm{prior}} = 1/\alpha$ while leaving local scores unchanged, so α controls the prior's contribution relative to neural evidence (Appendix A). The local and fused predictions are $\hat{y}_t = \arg\max_j s_{t,j}$ and $\tilde{y}_t = \arg\max_j \tilde{s}_{t,j}$.

For a score vector a with largest entries $a_{(1)} \geq a_{(2)}$, its top-two margin is $m(a) = a_{(1)} - a_{(2)}$, equal to the log odds between its two leading softmax candidates. The *pre-fusion margin* is $m(s_t)$ and the *fused margin* is $m(\tilde{s}_t)$. The *prior margin* is computed over the prior's own entries, distinct sentences or words, before their scores are shared by the candidates that map to them; entries with equal scores remain separate. All three margins are evaluated by how they order outcomes.

3.2 CORRECTION OUTCOMES, ORDERING, AND REACH

Among initially incorrect predictions ($\hat{y}_t \neq y_t$), a *repair* has $\tilde{y}_t = y_t$ and a *residual error* has $\tilde{y}_t \neq y_t$; among initially correct predictions, a *regression* has $\tilde{y}_t \neq y_t$. Let n_{repair} and $n_{\text{regression}}$ count repairs and regressions over N windows, so that retrieval accuracy changes by $\Delta\text{R@1} = (n_{\text{repair}} - n_{\text{regression}})/N$.

The correct candidate's initial rank is $r_t^0 = 1 + \left|\{j : s_{t,j} > s_{t,y_t}\}\right|$, the number of candidates scoring strictly above it plus one. We compare repairs with residual errors in fixed groups $\mathcal{B}$ of initially incorrect predictions, by the correct candidate's initial rank: 2–3, 4–5, 6–10, 11–20, and 21–50, adding 51–100 for the word-level sweeps. Within $\mathcal{B}$, draw indices r and e uniformly from its repairs and residual errors. Repair-separation AUROC measures whether confidence z ranks the repair above the residual error:

$$\text{AUROC}_{\text{repair},\mathcal{B}}(z) = \Pr(z_r > z_e) + \tfrac{1}{2}\Pr(z_r = z_e). \tag{3}$$

Values above 0.5 favour repairs; values below 0.5 favour residual errors. We call these positive and negative ordering. Pooled correctness AUROC applies the same expression with r and e drawn from all final correct and all final incorrect predictions. Repair rates across margin quartiles within each group give a complementary view of this ordering. Effective correction reach is the 90th percentile of the initial ranks among repairs:

$$R_{90} = Q_{0.90}\big(r_t^0 \mid \hat{y}_t \neq y_t,\ \tilde{y}_t = y_t\big)\,. \tag{4}$$

Reach summarises the repair distribution independently of confidence scores. Each AUROC curve is described by its deepest positively ordered initial-rank group. For a single downward crossing of 0.5, this locates the reversal boundary; Figure 3C also interpolates the crossing between group midpoints. Full curves describe settings with no observed crossing or multiple sign changes.

3.3 DATASETS AND CONTEXTUAL PRIORS

We evaluate MEG-MASC and MOUS (Gwilliams et al., 2023; Schoffelen et al., 2019) with three independently trained local decoders per dataset (Table 1). Their neural encoder maps each word-aligned MEG window to a time-resolved speech representation, and candidate scores are inner products with frozen wav2vec2-large-xlsr-53 audio representations (Conneau et al., 2021). Stimulus-identity splits evaluate zero-shot retrieval of segments unseen during local-decoder training. Both datasets support the reversal and prior-margin analyses; controlled sweeps, cross-intervention comparisons, and selective-output evaluation use MEG-MASC (Appendix B).

Table 1: MEG datasets and retrieval accuracy with the history-aggregation prior at $\alpha = 2.7$.

Dataset	Language	Participants	Sensors	Hours	Candidates	R@1 (%)	
						Local	Fused
MEG-MASC	English	27	208	56.2	1,464	44.42	58.82
MOUS	Dutch	96	272	80.9	825	21.72	29.98

Contextual priors differ in three respects that our interventions vary separately: the source of history (true preceding words, the decoder's own predictions, or accumulated local scores), the granularity at which prior scores are assigned (sentence or word, shared by all candidates that map to the same entry), and the form of the update (additive fusion or non-additive pruning). Every prior shares one property: candidates that map to the same entry receive the same prior score, so fusion cannot reorder them and the prior acts on competition between entries while local scores act within one.

Our main setting uses a history-aggregation prior that accumulates evidence only from preceding windows. Candidates from the same source sentence form a bucket $\mathcal{C}_b$, and stimulus timing identifies query sentence presentations and word onsets. Within each presentation, the prior integrates local-score support from earlier windows in word-onset order to infer which bucket matches the query. With $b(j)$ identifying candidate j's bucket, u_t the vector of bucket-size-normalised support from local scores, H_t the accumulated bucket scores, and η the update rate,

$$H_t = (1-\eta)H_{t-1} + \eta u_t, \qquad \pi_{t,j} = \mathcal{L}(H_{t-1}, b(j)), \tag{5}$$

where $\mathcal{L}$ retains up to the four strongest active bucket scores and broadcasts each to that bucket's candidates, giving the prior score vector π_t^{bucket} over distinct buckets. Prediction reads H_{t-1} before the current update; state resets at each presentation, and validation selects $\alpha = 2.7$ and $\eta = 0.1$.

Language-model (LM) priors replace accumulated support with Qwen3-4B-Base likelihoods (Yang et al., 2025), at sentence level (mean per-token log-likelihood per candidate sentence) or word level (log probability of each candidate's aligned word, shared by candidates carrying that word). Each comes in two history sources: free-running priors condition on preceding predicted words, and teacher-forced priors condition on true preceding words, probing correction when the history is right. For the word-level free-running prior, history is generated once from the frozen decoder's top-1 predictions, giving one prior-score vector per decoder for the whole sweep. Hard bucket pruning changes the update form: it uses the history-aggregation state to retain four buckets and discard the rest, without adding scores. Appendix C gives the constructions and validation selection.

Sweeps vary only α, holding both score streams, history, candidates, and test windows fixed: eight weights from 0.5 to 4.0 for history aggregation, sixteen from 0.05 to 16 for each word-level LM prior. Every weight determines its own repairs and residual errors, against which all three margins are evaluated; cross-intervention comparisons use validation-selected operating points.

3.4 Confidence estimation and evaluation

The analyses above use the correct candidate to label outcomes. A system predicting without it must judge two things: whether the output is correct, and whether the prior's revision improved on the local prediction. *Output confidence* targets the first, including whether an output set contains the correct candidate. *Intervention confidence* targets $B(I_t)$, the expected benefit of contextual revision over the local prediction given observable information I_t. Under top-1 error loss,

$$\begin{aligned} B(I_t) &= \Pr(\tilde{y}_t = y_t \mid I_t) - \Pr(\hat{y}_t = y_t \mid I_t) \\ &= \Pr(\text{repair} \mid I_t) - \Pr(\text{regression} \mid I_t). \end{aligned} \tag{6}$$

All probabilities use the same conditioning information. A repair and a prediction that remains correct share the final-correctness label, but only the repair improves on the local prediction. Among initially incorrect windows, final correctness is repair success; across all windows, $B(I_t)$ also depends on preserving correct local predictions.

For output confidence, multilayer perceptrons predict fused top-K success for $K \in \{1, 3, 5, 10, 20\}$. Decoder-score estimators combine two local and six fused score features; prior-feature estimators use seven features of prior state, current support, and sentence position; the full estimator combines both. Additional estimators use only the fused margin, the six fused features, or the two local features, with the same hidden-layer architecture and training procedure. Training and calibration use validation data, with cross-validation folds formed by presentation.

These estimates drive an output policy, which selects the smallest K meeting its validation-selected success-probability threshold, or abstains if none qualifies. Emission is the fraction of windows on which a set is output, and coverage the fraction of emitted sets containing the correct candidate. Validation requires at least 90% coverage and mean emitted set size at most 4.8. We compare decoder-score and full estimators with a margin-binned reference that applies conformal set calibration (Romano et al., 2020) in 20 fused-margin bins, reporting emission rate, mean set size, and coverage, also grouped by initial rank (Appendix I.3).

To compare selection at a fixed output size, fixed-K policies emit their top-K set or abstain, with validation thresholds maximising emission subject to at least 90% coverage (Appendix I.3).

Throughout, estimates are means and standard deviations across the three local decoders, with reach computed within each decoder before averaging; we report a repair-separation cell only when every decoder has at least five repairs, and initial rank serves only as an evaluation label. Uncertainty is quantified at three levels: Wilson intervals describe the quartile repair rates in Figure 2A, B, which pool evaluated predictions with fixed quartile membership; sentence-cluster bootstrap supports rate contrasts and within-sentence comparisons; and participant bootstrap, sharing each resampled participant set across weights, groups, confidence measures, and decoders, gives the core AUROC intervals and boundary stability (Appendix D).

4 RESULTS

4.1 THE FUSED MARGIN REVERSES ITS ORDERING ACROSS INITIAL RANKS

Confidence ordering reverses as the correct candidate starts lower in the local ranking. Pooled correctness AUROC is 0.872 on MEG-MASC (Figure 2), yet among initially incorrect predictions, repair-separation AUROC falls from 0.704 at ranks 2–3 to 0.392 at ranks 21–50. Pooled over all initially incorrect predictions, AUROC is 0.646, so grouping by initial rank exposes opposite orderings that this pooled comparison does not reveal. For the same outcomes, the pre-fusion margin remains negatively ordered across all five groups, with AUROCs of 0.30–0.33. Fusion changes the ordering to positive near the top, while deeper groups remain negative. Repairs concentrate at low initial ranks, leaving a substantial share of errors deeper in the ranking: on MEG-MASC, residual errors starting beyond rank 20 account for 46.6% of all post-fusion errors.

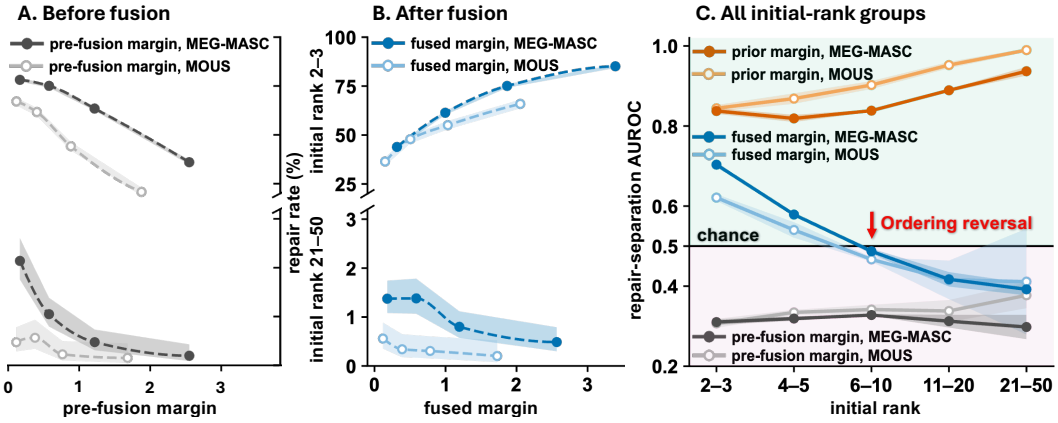

Figure 2: **Confidence ordering reverses as the correct candidate starts lower in the local ranking.** Initially incorrect windows on MEG-MASC (filled) and MOUS (hollow). Panels A and B use the same windows and outcomes, with one point per quartile at its mean margin; the upper and lower rows show initial ranks 2–3 and 21–50, respectively, with different vertical scales. **(A) Before fusion.** Repair rate decreases with margin in both groups. **(B) After fusion.** From the smallest to largest quartile, repair rate rises at ranks 2–3 and falls at ranks 21–50. **(C) All initial-rank groups.** Repair-separation AUROC of the fused, pre-fusion, and prior margins; above 0.5 favours repairs, below 0.5 residual errors. Bands are Wilson 95% intervals in A, B, and ± 1 SD across three local decoders in C; dashed curves interpolate quartile points. MOUS ranks 21–50 average 13 repairs per decoder; Appendix E reports full rates and participant-level intervals.

Repair rates across margin quartiles show the same reversal. From the smallest to largest fused-margin quartile on MEG-MASC, repair rate rises from 42.4% to 84.4% at initial ranks 2–3, but falls from 1.37% to 0.55% at ranks 21–50. Sentence-cluster bootstrap supports both endpoint contrasts (Appendix E). Rates also rise at ranks 4–5 and fall at ranks 11–20 under this contrast; at ranks 6–10, the interval for the highest-to-lowest quartile ratio includes one.

The reversal replicates on MOUS and at the participant level. There, repair-separation AUROC falls from 0.62 at ranks 2–3 to 0.41 at ranks 21–50, and participant-level intervals support positive ordering near the top and negative ordering further down on both datasets. On MOUS, the 6–10 and 11–20 intervals lie below 0.5, while ranks 21–50 average 13 repairs per decoder and have an interval crossing 0.5 (Appendix E.4).

The reversal persists within sentences, among repairable errors, and for other confidence scores. Negative fused-margin ordering persists within sentence presentations and among errors whose correct candidate already leads its sentence bucket, the subset an additive prior can in principle repair (Table 5; Appendix G.3). Entropy, sentence-score mass gap, and concentration also order repairs negatively (0.36, 0.46, 0.43 at ranks 21–50; Appendix I.2).

4.2 A SCORE-LEVEL ACCOUNT OF MARGIN FORMATION

A prior can make an error more confident while moving its correct candidate up. In Table 6, the prior advances the correct candidate from rank 40 to rank 7 while the margin grows from 0.07 to 3.54 around an incorrect winner. We account for this by comparing how repairs and residual errors form their final leads.

A repair's margin shrinks as the deficit it must close grows. Let $G_t^0 = \max_{j \neq y_t} s_{t,j} - s_{t,y_t}$ be the correct candidate's initial score deficit and $\Delta_t^\pi = \max_j \pi_{t,j} - \min_j \pi_{t,j}$ the prior score range. On a repair, comparison with the strongest local competitor gives:

$$G_t^0 \leq \alpha \Delta_t^\pi, \tag{7}$$

$$m(\tilde{s}_t) \leq \alpha \Delta_t^\pi - G_t^0. \tag{8}$$

Increasing α relaxes both constraints, allowing repairs to close larger deficits with larger margins.

Deep residual errors gain margin between two wrong candidates. With post-fusion score deficit $\tilde{G}_t = \max_{j \neq y_t} \tilde{s}_{t,j} - \tilde{s}_{t,y_t}$, the fused margin equals $-\tilde{G}_t$ on a repair and is at most $\tilde{G}_t$ on a residual error. In the latter case, equality holds when the correct candidate is runner-up, in the absence of ties. The correct candidate's runner-up share among residual errors falls from 75.2% at initial ranks 2–3 to 2.1% at ranks 21–50 (Table 7). The margin at deep initial ranks therefore mostly compares two incorrect candidates, which can separate even while the correct candidate remains far behind; this is one way fusion leaves an error uncorrected yet more confident.

The bound certifies most reversals under history aggregation but few under the word-level LM. Negative ordering follows when a residual-error margin exceeds a repair's upper bound in more than half of within-group pairs (Appendix A). This range-based condition certifies 16 of 21 negative cells (weight and rank-group pairs) under history aggregation and 1 of 29 under the word-level teacher-forced prior, whose full score range can greatly exceed the correct candidate's prior advantage over the strongest local competitor. At $\alpha = 0.5$ under this prior, measured AUROC falls from 0.674 at ranks 2–3 to 0.419 at ranks 21–50, while the range bound does not certify this negative ordering (Appendix F.3), so the reversal is not implied by the bound alone.

These constraints motivate the hypothesis that increasing the prior's contribution extends positive repair ordering towards larger initial score deficits. We examine this response by varying the fusion weight while holding both score streams fixed.

4.3 CONTROLLED CHANGES IN PRIOR CONTRIBUTION MOVE CONFIDENCE ORDERING

We vary only the fusion weight, keeping local and prior scores fixed. Predictions and outcome labels are recomputed at each weight. Effective correction reach summarises where the resulting repairs begin. At the main history-aggregation setting, 84.8% of repairs start within the top five ranks and 96.3% within the top ten, giving $R_{90} = 7$ (Figure 3).

Raising the history-aggregation weight pushes positive ordering to deeper ranks. Reach grows from 3.0 to 9.3, and the deepest positively ordered group advances from none through 2–3, 4–5, and 6–10 to 11–20; every increase preserves or advances it, and the sequence is non-decreasing in all 1,000 participant-bootstrap resamples (Appendix G.1). On the same outcome labels, the pre-fusion margin stays negatively ordered in all 32 reported cells (0.111–0.378).

Positive ordering also extends to larger initial score deficits. In five equal-mass deficit groups fixed across weights, the deepest positively ordered group advances from none to the fourth group. The 90th percentile of initial deficits among repairs grows from 0.61 to 3.24 score units, while the fifth group remains negatively ordered at $\alpha = 4.0$ (Appendix G.2). Thus positive ordering expands in both initial-rank and score-deficit coordinates under this prior.

Under a teacher-forced word-level LM prior, positive ordering keeps expanding after accuracy gain peaks. Between $\alpha = 0.05$ and 1, reach grows from 3.0 to 19.3, and positive ordering extends from no group to every observed group through rank 100. Validation selects $\alpha = 0.5$ for accuracy; there, positive ordering reaches ranks 6–10 and the gain is 11.20 percentage points. At $\alpha = 1$, positive ordering extends further, but the gain falls to 9.16 points.

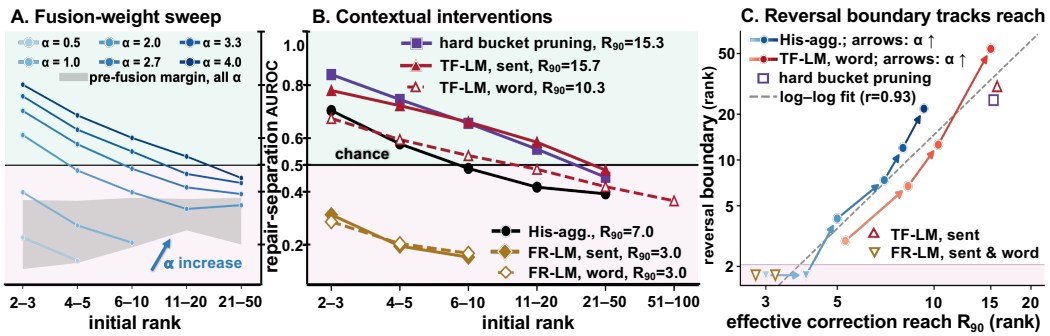

Figure 3: **Changing prior contribution moves the confidence-ordering structure.** MEG-MASC. His-agg.: history aggregation; TF/FR-LM: teacher-forced/free-running LM prior. **(A) Fusion-weight sweep.** Six weights from the eight-setting history-aggregation sweep, with local and prior scores fixed; grey shows pre-fusion-margin AUROC on the same outcomes. **(B) Contextual interventions.** Six validation-selected settings: filled markers are sentence-level interventions, hollow markers and dashed curves are word-level LM priors; hard bucket pruning is non-additive. **(C) Reversal boundary tracks reach.** Arrows follow increasing α for history aggregation and the word-level teacher-forced prior, the latter through $\alpha = 0.75$; at higher weights all observed groups through rank 100 are above chance. Open markers correspond to B; the bottom band denotes no positive ordering. Crossings interpolate between group midpoints; the dashed log–log fit summarises the displayed association. Full curves, boundaries, and event counts are in Appendix G.

With predicted history, positive ordering appears only after accuracy gain turns negative. At $\alpha = 0.15$, the word-level free-running prior has reach 3.0 and gain 1.25 points, with no positive group; at $\alpha = 0.5$, AUROC is 0.539 at ranks 2–3 and below 0.5 elsewhere, while gain is -0.69 points. The reversal therefore appears with predicted history as well, although positive ordering emerges only after gain turns negative. At larger weights, sign changes multiply and the deepest positive group retreats even as reach grows (Appendix G.4).

Priors with similar accuracy gains can differ in confidence ordering. Every prior in Section 3 shows negative repair ordering, but its extent varies. The sentence-level teacher-forced LM prior and the history-aggregation prior improve accuracy by 14.74 and 14.40 percentage points, respectively, yet retain positive ordering through ranks 11–20 and 4–5. Correction reach captures another aspect of this variation: hard bucket pruning and the sentence-level teacher-forced prior have similar reaches of 15.3 and 15.7, and both retain positive ordering through ranks 11–20. Similar reach can also accompany different ordering boundaries: history aggregation at $\alpha = 4.0$ reaches 9.3 with positive ordering through ranks 11–20, whereas the word-level teacher-forced prior at $\alpha = 0.5$ reaches 10.3 with positive ordering through ranks 6–10. Prior construction thus shapes both the extent of correction and the ordering of its outcomes.

Nearly half of the remaining errors start beyond rank 20, inside the reversed region; since fusion mixes local and prior support into one margin, we read these two scores separately to identify these errors.

4.4 Observable information for confidence estimation

Read before fusion, the prior's own margin separates the outcomes that the fused margin reverses. The final margin records the winner's lead without distinguishing how local competition and prior support produced it. On the same repairs and residual errors, the prior margin gives 0.937 at ranks 21–50 against 0.392 for the fused margin; under history aggregation it ranges from 0.819 to 0.937 across the five MEG-MASC groups and from 0.84 to 0.99 on MOUS (Figure 2C), stays above 0.7 in all 32 cells of the fusion-weight sweep while the fused margin reverses (Appendix G.1), and among initially correct MEG-MASC predictions it separates those that remain correct from regressions with AUROC 0.777 ± 0.022. Word-level LM priors discriminate less: 0.568–0.759 for repairs and 0.545 for preservation with teacher-forced history ($\alpha = 0.5$), and 0.532–0.701 (ranks 2–10) and 0.532 with predicted history ($\alpha = 0.15$; Tables 17 and 18).

Table 2: Confidence estimation and selective output with the history-aggregation prior on MEG-MASC; means over three local decoders, with estimator seeds averaged. **(a)** Top-1 repair and top-5 inclusion AUROC (Table 20). **(b)** Emission at matched 92.24% test-curve coverage; K is set size.

(a) Confidence-estimation AUROC

Initial rank	Success target	Decoder-score	Prior-feature
11–20	Top-1	0.716	0.848
	Top-5	0.552	0.752
21–50	Top-1	0.653	0.910
	Top-5	0.522	0.867

(b) Selective output

Features	Dim.	Emission (%)	Mean K
Fused margin	1	56.7 ± 0.5	6.3
Fused scores	6	66.6 ± 0.7	5.8
Decoder scores	8	72.5 ± 0.5	5.2
Full	15	74.5 ± 0.4	5.2

Learned estimators rank repairs above residual errors where the fused margin reverses. Under history aggregation, decoder-score and prior-feature estimators reach repair-separation AUROCs of 0.653 and 0.910 at ranks 21–50 (Table 2(a)). Among the 10% of these deep errors scored highest by a prior-state estimator (three prior features), repairs occur at 8.40%, against 1.80% for decoder scores and a base rate of 1.05% (Appendix I.2). For top-five inclusion, prior-feature discrimination remains strong while decoder-score discrimination is near chance. Across all test windows, the full estimator gives correctness AUROC 0.954, compared with 0.872 for the fused margin, unchanged under any strictly increasing recalibration (Appendices A and I.2).

Retained scores let selective output emit more, including where errors concentrate. Emission at matched 92.24% test-curve coverage (Appendix I.4) rises from 56.7% with the fused margin to 66.6% with six fused features, then 72.5% and 74.5% as local and prior features are added (Table 2(b)). For validation-selected policies, adding prior features at ranks 21–50 raises emission from 32.5% to 38.2% and coverage from 36.4% to 50.8%. At output size one, the full estimator emits on 58.3% at 90.5% coverage, versus 48.7% and 89.1% for margin binning (Appendix I.3).

5 Discussion

A contextual prior sets both which prediction wins and how confident its errors look. Larger fused margins make repairs more likely near the top of the local ranking but less likely further down, where nearly half of remaining errors start, while pooled AUROC stays high. Raising the fusion weight moves this reversal to deeper ranks, and under a word-level LM prior it keeps moving after accuracy gain peaks. The prior's own margin, read before fusion, separates repairs from residual errors.

The prior can widen the gap between wrong candidates as easily as it lifts the right one. It adds to every margin only a weighted difference of its own scores, whether or not that difference favours the correct candidate. Confidence ordering therefore follows how the prior reshapes competition, not how much it improves accuracy, and the repair-margin bound holds for any additive fusion. Seeing this requires the local scores, which links confidence analysis to source attribution (Zhang et al., 2026b): identifying what context changed precedes judging whether the change succeeded.

Retained scores restore what recalibration cannot. Recalibrating the margin preserves its ordering, so no calibration can undo the reversal, and other summaries of the fused scores, such as entropy, reverse as well (Section 4.1); keeping local and prior scores separate restores the ordering (Table 2). A weight chosen for accuracy does not settle confidence, and a prior can be judged by how well its retained scores reveal its correction outcomes, alongside its accuracy gain.

Limitations and future work. Our offline retrieval setting keeps the candidate pool fixed and score streams separate, which makes each correction traceable and the prior's contribution controllable. Free-running histories come from local top-1 rather than fused predictions; with teacher-forced histories, they bracket what richer histories, such as N-best predictions, would provide. Generated candidates would add a changing pool and feedback from fused outputs, and fixed-set assistive interfaces such as BCI spellers (Speier et al., 2016) offer a direct test of revision benefit. Our estimators target final correctness; estimating revision benefit, and training objectives that make successful corrections identifiable, are natural next steps. **Confidence estimation after contextual fusion should therefore keep the local evidence and contextual support behind each prediction, rather than any summary of the fused scores alone.**

AI USE STATEMENT

The authors used ChatGPT and Claude to polish writing and LaTeX formatting. The authors produced and verified all analyses, results, and claims and take responsibility for the final content.

REFERENCES

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 2020.

Nicholas S Card, Tyler Singer-Clark, Hamza Peracha, Carrina Iacobacci, Xianda Hou, Maitreyee Wairagkar, Zachery Fogg, Elena C Offenberg, Leigh R Hochberg, Sergey D Stavisky, et al. Long-term independent use of an intracortical brain–computer interface for speech and cursor control. *Nature Medicine*, 2026.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition. In *Proc. Interspeech*, 2021.

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.

Xiaoyan Deng, Zhu Liang Yu, Canguang Lin, Zhenghui Gu, and Yuanqing Li. A Bayesian shared control approach for wheelchair robot with brain machine interface. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2020.

Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. Chapman and Hall/CRC, 1994.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010.

Hayato Futami, Hirofumi Inaguma, Masato Mimura, Shinsuke Sakai, and Tatsuya Kawahara. ASR rescoring and confidence estimation with ELECTRA. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Zhuo Gong, Daisuke Saito, and Nobuaki Minematsu. Entropy-based dynamic rescoring with language model in E2E ASR systems. *Applied Sciences*, 12(19):9690, 2022.

Alon Gorenshtein, Mahmud Omar, Eric L Jia, Yosef Adiniaev, Oved Daniel, Jonathan Kruskal, Muneeb Ahmed, Olga R Brook, Yiftach Barash, and Eyal Klang. Quantifying large language model influence in brain computer interface communication for amyotrophic lateral sclerosis. *medRxiv*, 2026a.

Alon Gorenshtein, Mahmud Omar, Eric L Jia, Yosef Adiniaev, Oved Daniel, Jonathan Kruskal, Muneeb Ahmed, Olga R Brook, Eyal Klang, and Yiftach Barash. Limits of trial-adaptive neural language fusion across large language models in P300 brain computer interfaces. *medRxiv*, 2026b.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, 2017.

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.

Christian Herff, Dominic Heger, Adriana De Pesters, Dominic Telaar, Peter Brunner, Gerwin Schalk, and Tanja Schultz. Brain-to-text: decoding spoken phrases from phone representations in the brain. *Frontiers in Neuroscience*, 9:217, 2015.

Jingya Huang, Sowmya Manojna Narasimha, Aashish N Patel, Ram Dyuthi Sristi, Gal Mishne, and Vikash Gilja. Probabilistic co-control in brain-computer interfaces: Uncertainty as a control signal in brain-to-text decoding. *bioRxiv*, 2026.

Yichen Jia and Hugo Van hamme. Leveraging beam search information for confidence estimation in E2E ASR. *IEEE Open Journal of Signal Processing*, 7, 2026.

Erik Jones, Shiori Sagawa, Pang Wei Koh, Ananya Kumar, and Percy Liang. Selective classification can magnify disparities across groups. In *International Conference on Learning Representations*, 2021.

Anjuli Kannan, Yonghui Wu, Patrick Nguyen, Tara N. Sainath, Zhifeng 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.

Soowon Kim, Byung-Kwan Ko, and Seo-Hyun Lee. Confidence-aware neural decoding of overt speech from EEG: Toward robust brain–computer interfaces. In *2026 14th International Conference on Brain-Computer Interface (BCI)*, 2026.

Johannes Y Lee, Sangjoon Lee, Abhishek Mishra, Xu Yan, Brandon McMahan, Brent Gaisford, Charles Kobashigawa, Mike Qu, Chang Xie, and Jonathan C Kao. Brain–computer interface control with artificial intelligence copilots. *Nature Machine Intelligence*, 2025.

Jarod Lévy, Mingfang Zhang, Svetlana Pinet, Jérémy Rapin, Hubert Banville, Stéphane d'Ascoli, and Jean-Rémi King. Noninvasive decoding of typed sentences from human brain activity. *Nature Neuroscience*, 2026.

Qiujia Li, David Qiu, Yu Zhang, Bo Li, Yanzhang He, Philip C Woodland, Liangliang Cao, and Trevor Strohman. Confidence estimation for attention-based sequence-to-sequence models for speech recognition. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

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.

Alexander B Silva, Kaylo T Littlejohn, Jessie R Liu, David A Moses, and Edward F Chang. The speech neuroprosthesis. *Nature Reviews Neuroscience*, 2024.

William Speier, C Arnold, and Nader Pouratian. Integrating language models into classifiers for BCI communication: a review. *Journal of Neural Engineering*, 13(3):031002, 2016.

Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G. Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866, 2023.

Vinit Unni, Shreya Khare, Ashish Mittal, Preethi Jyothi, Sunita Sarawagi, and Samarth Bharadwaj. Adaptive discounting of implicit language models in RNN-transducers. In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

Ehsan Variani, Michael Riley, David Rybach, Cyril Allauzen, Tongzhou Chen, and Bhuvana Ramabhadran. On adaptive weight interpolation of the hybrid autoregressive transducer. In *Proc. Interspeech 2022*, 2022.

Francis R. Willett, Erin M. Kunz, Chaofei Fan, Donald T. Avansino, Guy H. Wilson, Eun Young Choi, Foram Kamdar, Matthew F. Glasser, Leigh R. Hochberg, Shaul Druckmann, et al. A high-performance speech neuroprosthesis. *Nature*, 620(7976):1031–1036, 2023.

Edwin B Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 1927.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2002.

Mingfang Zhang, Jarod Lévy, Cedric Rommel, Jérémy Rapin, Corentin Bel, Julie Bonnaire, Daniel Nieto, Pierre Bourdillon, Svetlana Pinet, Stéphane d'Ascoli, et al. Accurate decoding of natural sentences from non-invasive brain recordings. *arXiv preprint arXiv:2608.18114*, 2026a.

Xinyu Zhang, Sichao Liu, Runhao Lu, Alexandra Woolgar, and Lihui Wang. What are we actually decoding? source attribution for non-invasive brain-to-language retrieval. *arXiv preprint arXiv:2605.24524*, 2026b.

A FORMAL PROPERTIES OF CONTEXTUAL FUSION

Equation 2 multiplies local candidate odds by the corresponding prior odds raised to α. For $\alpha > 0$, this gives the prior temperature $1/\alpha$, independently of calibration.

A.1 BOUNDED CORRECTIVE CAPACITY

Let $j^* = \arg\max_{j \neq y_t} s_{t,j}$ be the strongest local competitor on an initially incorrect window. Repair requires

$$G_t^0 = s_{t,j^*} - s_{t,y_t} \leq \alpha(\pi_{t,y_t} - \pi_{t,j^*}) \leq \alpha \Delta_t^\pi, \tag{9}$$

where $\Delta_t^\pi = \max_j \pi_{t,j} - \min_j \pi_{t,j}$. Comparison with the same competitor bounds the repaired prediction's final margin:

$$\begin{aligned} m(\tilde{s}_t) &\leq \tilde{s}_{t,y_t} - \tilde{s}_{t,j^*} \\ &= \alpha(\pi_{t,y_t} - \pi_{t,j^*}) - G_t^0 \\ &\leq \alpha \Delta_t^\pi - G_t^0. \end{aligned} \tag{10}$$

For a fixed fusion weight and prior score range, closing a larger deficit leaves less room for the repaired prediction's margin. Ordering additionally depends on the margins of residual errors. Write $z_t = m(\tilde{s}_t)$ and draw window indices r and e independently and uniformly from repairs and residual errors, respectively, in the same group $\mathcal{B}$. Then

$$\mathrm{AUROC}_{\mathrm{repair},\mathcal{B}}(z) \leq 1 - \Pr\left[z_e > \alpha \Delta_r^\pi - G_r^0\right]. \tag{11}$$

The event on the right makes the residual-error margin exceed the repair margin. If its probability exceeds one half, repair-separation AUROC is below one half under the tie convention in Eq. 3. The sufficient condition thus connects the individual repair bound to the distribution of residual-error margins.

A.2 MARGIN INTERPRETATION AND RECALIBRATION

On a repair, the correct candidate wins, so

$$m(\tilde{s}_t) = \tilde{s}_{t,y_t} - \max_{j \neq y_t} \tilde{s}_{t,j} = -\tilde{G}_t. \tag{12}$$

On a residual error, let k and ℓ index the two largest fused scores. Then

$$m(\tilde{s}_t) = \tilde{s}_{t,k} - \tilde{s}_{t,\ell} \leq \tilde{s}_{t,k} - \tilde{s}_{t,y_t} = \tilde{G}_t. \tag{13}$$

Without ties, equality holds exactly when the correct candidate is runner-up. When the correct candidate is below the top two, the margin compares two incorrect candidates.

For a strictly increasing function f, $z_r > z_e$ if and only if $f(z_r) > f(z_e)$. Recalibration by f thus preserves every pairwise ordering and leaves AUROC unchanged. A non-decreasing function can introduce ties but cannot reverse an individual pair. Re-estimating confidence from additional score features can change the ordering, as evaluated in Appendix I.

B DATASETS, RETRIEVAL TASK, AND LOCAL DECODER

MEG-MASC supplies 71,736 test windows in 6,027 sentence presentations, scored against 1,464 audio candidates from 123 sentences; MOUS uses an 825-candidate pool. Dataset sizes are in Table 1. Stimulus timing identifies query presentations, and source metadata identifies candidate sentence buckets. Presentations define prior resets and sentence-resampling units.

We split stimulus identities 70/10/20 into training, validation, and test sets, assigning every occurrence to the same split. MEG-MASC training and validation windows overlapping test windows from the same audio file are removed. We deduplicate candidates by stimulus identity and window index. Each neural window spans -0.35 to $+2.65$ seconds relative to word onset.

The neural encoder combines spatial sensor mixing, participant-specific adaptation, and temporal convolutions to predict frozen wav2vec2-large-xlsr-53 audio representations (Baevski et al., 2020;

Conneau et al., 2021). Scores are Frobenius inner products with normalised audio targets. Training uses matched-window positives and in-batch negatives, with early stopping on validation loss. The three independently trained local decoders provide cached full-pool scores shared across all interventions.

C Contextual interventions and operating-point selection

C.1 History aggregation and pruning

At the selected history-aggregation setting, let $\mathcal{T}_t$ contain the 64 highest-scoring local candidates and $\theta_t = Q_{0.94}(s_t)$. Each sentence bucket receives its largest positive score excess, normalised by the square root of its size:

$$u_t(b) = \frac{\max\big(\{(s_{t,j} - \theta_t)_+ : j \in \mathcal{T}_t \cap \mathcal{C}_b\} \cup \{0\}\big)}{\sqrt{|\mathcal{C}_b|}}. \tag{14}$$

Here $(v)_+ = \max(v, 0)$. Equation 5 accumulates this support with $\eta = 0.1$. A bucket is active when $H_{t-1}(b) > 0$, so $\mathcal{L}(H_{t-1}, b) = H_{t-1}(b)$ for the four highest-scoring active buckets and 0 otherwise; since state values are non-negative, $\Delta_t^\pi = \max_b H_{t-1}(b)$. Prior margin is the top-two gap of H_{t-1} over all buckets, with inactive buckets entering as zero. Fewer than four buckets are active only at the first window of a presentation, where no candidate is boosted, and hard pruning retains every candidate. Prediction reads the preceding state before the current update and uses $\alpha = 2.7$, selected on validation. State resets at each presentation; support comes from preceding windows in word-onset order.

Hard bucket pruning uses the same state but keeps only candidates in its four leading buckets:

$$\tilde{s}_{t,j}^{\text{mask}} = \begin{cases} s_{t,j}, & b(j) \in \mathrm{TopIdx}_4(H_{t-1}), \\ -\infty, & \text{otherwise}, \end{cases} \tag{15}$$

where TopIdx_4 returns the four highest-scoring bucket indices. The retained local scores and state update are unchanged.

C.2 Language-model priors

Qwen3-4B-Base scores either candidate sentences by mean per-token log-likelihood or aligned candidate words by

$$\pi_{t,j} = \log \Pr_{\text{LM}}(w_j \mid h_{t-1}).$$

Teacher-forced histories contain true preceding words; free-running histories contain decoded preceding words. Histories end before the current word and reset at presentation boundaries. In the word-level free-running sweep, each frozen local decoder supplies one predicted history and one prior-score vector per window for all weights. Changing α therefore changes only the combination of the fixed local and prior scores.

The sentence-level free-running and teacher-forced weights are 0.465 and 1.86. Word-level weights maximise mean validation accuracy gain across decoders, selecting 0.15 and 0.5, respectively. Both word-level priors are swept over sixteen weights from 0.05 to 16. Candidates carrying the same word share its score; we compute the prior margin over distinct words before assigning it to candidates.

D Statistical and selection protocols

Outcome labels are recomputed at every fusion weight, and all confidence measures use those same labels. Thus, fixed pre-fusion and prior scores can have changing AUROCs. Initially incorrect presentation-start windows, where no bucket is active and $\pi_t = 0$, are retained as residual errors with prior margin 0; excluding them lowers prior-margin AUROC by at most 0.05 (ranks 2–3) and changes fused-margin AUROC by at most 0.004. We compute estimates within each local decoder and then average them; learned estimators first average their five training seeds within a decoder. A repair-separation cell is omitted if any decoder has fewer than five repairs. Reach uses

Table 3: Repair rates across initial-rank groups when initially incorrect predictions are grouped by (a) fused margin or (b) pre-fusion margin. Q1 and Q4 contain the smallest and largest margins. Q4/Q1 divides their unrounded repair rates, and its intervals are sentence-cluster bootstrap 95% intervals. Panel (b) uses the same evaluated predictions and fusion-outcome labels as (a).

(a) Fused margin

Initial rank	Predictions	Repair rate (%)				Q4/Q1	
		Q1	Q2	Q3	Q4	Ratio	95% interval
2–3	32,277	42.44	59.75	72.61	84.39	1.988	[1.933, 2.047]
4–5	13,725	30.80	37.60	43.57	48.18	1.564	[1.467, 1.668]
6–10	16,861	20.61	22.85	22.16	19.15	0.929	[0.842, 1.022]
11–20	15,244	8.27	7.43	6.32	3.73	0.451	[0.357, 0.547]
21–50	16,613	1.37	1.47	0.77	0.55	0.404	[0.223, 0.654]

(b) Pre-fusion margin

Initial rank	Repair rate (%)				Q4/Q1	
	Q1	Q2	Q3	Q4	Ratio	95% interval
2–3	78.34	75.47	66.32	39.05	0.498	[0.481, 0.514]
4–5	56.44	48.70	37.02	17.98	0.319	[0.293, 0.345]
6–10	32.12	27.71	18.27	6.67	0.208	[0.181, 0.233]
11–20	11.34	8.53	4.38	1.50	0.132	[0.097, 0.172]
21–50	2.14	1.16	0.60	0.27	0.124	[0.061, 0.209]

the interpolated 90th percentile within each decoder. Boundary summaries report the deepest group with mean AUROC above 0.5; full curves retain non-contiguous positive groups.

Sentence-cluster bootstrap resamples presentations, and participant-cluster bootstrap resamples 27 MEG-MASC or 96 MOUS participants, each with 1,000 replicates and 95% percentile intervals. Each participant sample is shared across weights, rank groups, confidence measures, and decoders; its statistic is the three-decoder mean. Quartile analyses retain their original group membership and pool repeated decoder evaluations of each window. Wilson intervals (Wilson, 1927) describe quartile proportions; sentence bootstrap (Efron & Tibshirani, 1994) quantifies Q4/Q1 contrasts. Within-sentence AUROC pairs repairs with residual errors from the same presentation. Estimator targets and output-selection protocols are given in Appendix I.

E Evidence for confidence-ordering reversal

E.1 Quartile repair rates

Repair rates are higher in the largest than in the smallest fused-margin quartile at initial ranks 2–3 and 4–5, and lower at ranks 11–20 and 21–50 (Table 3). The Q4/Q1 interval contains one at ranks 6–10, placing the transition between groups 4–5 and 11–20 under this contrast. Pre-fusion-margin repair rates decrease across quartiles in all five groups on the same outcome labels.

E.2 Repair distributions and confidence ordering

The two datasets have similar repair distributions: 67.2% of MEG-MASC repairs and 65.3% of MOUS repairs begin at ranks 2–3, while each assigns 0.6% to ranks 21–50. On MEG-MASC, 84.8% of repairs begin within the top five and 96.3% within the top ten, giving $R_{90} = 7$. Table 4 reports the confidence measures with decoder-level variation and participant-level intervals.

E.3 Within-sentence comparison

Pairing repairs with residual errors from the same presentation retains the reversal (Table 5). Above initial rank 10, fused-margin AUROC changes from 0.415 across presentations to 0.344 within presentations; prior-margin AUROC remains positive in all five primary groups.

Table 4: Repair-separation AUROC at the main operating point. Point estimates are means across three local decoders. For the MEG-MASC margins, ± denotes across-decoder SD. Brackets give participant-cluster bootstrap 95% intervals, sharing each resample across decoders (27 participants on MEG-MASC, 96 on MOUS). Entropy-based confidence and repair counts are means.

(a) MEG-MASC

Initial rank	Fused margin	Pre-fusion margin	Prior margin	Fused top-16 entropy-based confidence
2–3	0.704 ± 0.006 [0.687, 0.718]	0.310 ± 0.007 [0.301, 0.321]	0.838 ± 0.007 [0.827, 0.847]	0.67
4–5	0.579 ± 0.001 [0.559, 0.596]	0.319 ± 0.006 [0.308, 0.329]	0.819 ± 0.009 [0.812, 0.826]	0.55
6–10	0.487 ± 0.007 [0.464, 0.505]	0.328 ± 0.002 [0.320, 0.335]	0.838 ± 0.003 [0.828, 0.849]	0.47
11–20	0.417 ± 0.018 [0.397, 0.439]	0.312 ± 0.016 [0.296, 0.329]	0.889 ± 0.004 [0.881, 0.898]	0.38
21–50	0.392 ± 0.015 [0.351, 0.435]	0.298 ± 0.031 [0.271, 0.327]	0.937 ± 0.011 [0.918, 0.956]	0.36

(b) MOUS

Initial rank	Fused margin	Pre-fusion margin	Prior margin	Repairs per local decoder
2–3	0.621 [0.600, 0.640]	0.307 [0.295, 0.320]	0.844 [0.831, 0.856]	1,438
4–5	0.540 [0.514, 0.565]	0.334 [0.317, 0.350]	0.868 [0.853, 0.882]	405
6–10	0.467 [0.440, 0.495]	0.342 [0.322, 0.363]	0.902 [0.888, 0.916]	267
11–20	0.416 [0.383, 0.450]	0.338 [0.305, 0.371]	0.952 [0.940, 0.963]	79
21–50	0.411 [0.304, 0.523]	0.377 [0.282, 0.457]	0.990 [0.985, 0.994]	13

Table 5: Within-sentence repair-separation AUROC. Chance is exactly 0.5; 11+ pools all initial ranks above 10.

Confidence measure	2–3	4–5	6–10	11–20	21–50	11+
Fused margin	0.672	0.520	0.404	0.357	0.284	0.344
Pre-fusion margin	0.284	0.276	0.256	0.242	0.204	0.268
Prior margin	0.757	0.686	0.644	0.658	0.719	0.694
Comparison pairs	6,457	1,279	1,308	326	58	1,097

E.4 Remaining errors and participant uncertainty

Across the three local decoders, 41,508 evaluated predictions have an initial rank above 20. Fusion repairs 179, leaving 41,329 residual errors: 46.6% of all post-fusion errors. On this fixed subset, fused-margin repair-separation AUROC is 0.385 with sentence-cluster bootstrap interval $[0.347, 0.425]$, compared with 0.950 $[0.934, 0.965]$ for prior margin. Pre-fusion margin and fused entropy-based confidence yield 0.305 and 0.365, respectively.

Participant-cluster intervals identify positively and negatively ordered groups on both datasets (Table 4). On MOUS, intervals for ranks 6–10 and 11–20 are below 0.5; ranks 21–50 contain 10–18 repairs per decoder and have a wider interval spanning 0.5. Pooling all initial ranks above 20, fused-

margin intervals are $[0.342, 0.430]$ for MEG-MASC and $[0.314, 0.538]$ for MOUS, while prior-margin intervals are $[0.933, 0.967]$ and $[0.992, 0.997]$.

F MARGIN FORMATION AND EMPIRICAL BOUND TIGHTNESS

F.1 REPRESENTATIVE OUTCOMES

Figure 4 and Table 6 provide score profiles and summary values for the four outcome types.

Table 6: Representative MEG-MASC windows from one local decoder.

Outcome	Correct-candidate rank	Margin before → after
Repair	$22 \to 1$	$0.05 \to 0.61$
Regression	$1 \to 2$	$0.04 \to 2.41$
Remained correct	$1 \to 1$	$6.09 \to 20.34$
Residual error	$40 \to 7$	$0.07 \to 3.54$

F.2 CORRECT-CANDIDATE PARTICIPATION IN THE FUSED MARGIN

As the initial rank increases, the fused margin increasingly compares two incorrect candidates (Table 7). The difference $\tilde{G}_t - m(\tilde{s}_t)$ is the correct candidate's deficit below the runner-up and equals zero when the correct candidate is runner-up.

Table 7: Correct-candidate participation on residual errors. Values are mean ± SD across local decoders. The deficit is measured below the fused runner-up.

Initial rank	Runner-up share (%)	Median $\tilde{G}_t - m(\tilde{s}_t)$
2–3	75.22 ± 0.90	0.000 ± 0.000
4–5	41.18 ± 1.30	0.259 ± 0.038
6–10	26.32 ± 0.49	0.770 ± 0.040
11–20	9.78 ± 0.49	1.614 ± 0.104
21–50	2.13 ± 0.13	2.876 ± 0.150

F.3 EMPIRICAL EVALUATION OF THE REPAIR-MARGIN BOUNDS

For repair r, let $z_r = m(\tilde{s}_r)$ and let j_r^* be the strongest local incorrect competitor. Substituting the range bound $\alpha\Delta_r^{\pi} - G_r^0$ or the competitor bound $c_r = \alpha(\pi_{r,y_r} - \pi_{r,j_r^*}) - G_r^0$ into Eq. 11 gives AUROC upper bounds U_{range} and U_{comp}. With measured AUROC A, we have $A \le U_{\text{comp}} \le U_{\text{range}}$; an upper bound below 0.5 certifies negative ordering.

Table 8 summarises all populated cells, each a weight and rank group. Statistics are averaged across decoders, then equally across cells. Under history aggregation, the range bound certifies 16 of 21 negative cells. The other five have AUROCs of 0.434–0.487 and upper bounds of 0.504–0.558, leaving their signs unresolved by this bound.

Table 8: Empirical tightness of repair-margin bounds across prior constructions. Gaps average all included cells, which index fusion weights and initial-rank groups. The final column reports cells with mean $U_{\text{range}} < 0.5$ among those with mean $A < 0.5$.

Prior	Cells	Mean upper-bound gap		Reversals certified by range bound
		$U_{\text{range}} - A$	$U_{\text{comp}} - A$	
History aggregation	32	0.041	0.028	16/21
Word-level teacher-forced LM	83	0.403	0.223	1/29

For the word-level teacher-forced prior, the score range spans the 709 distinct words represented in the candidate pool and can greatly exceed the correct candidate's prior advantage over its strongest

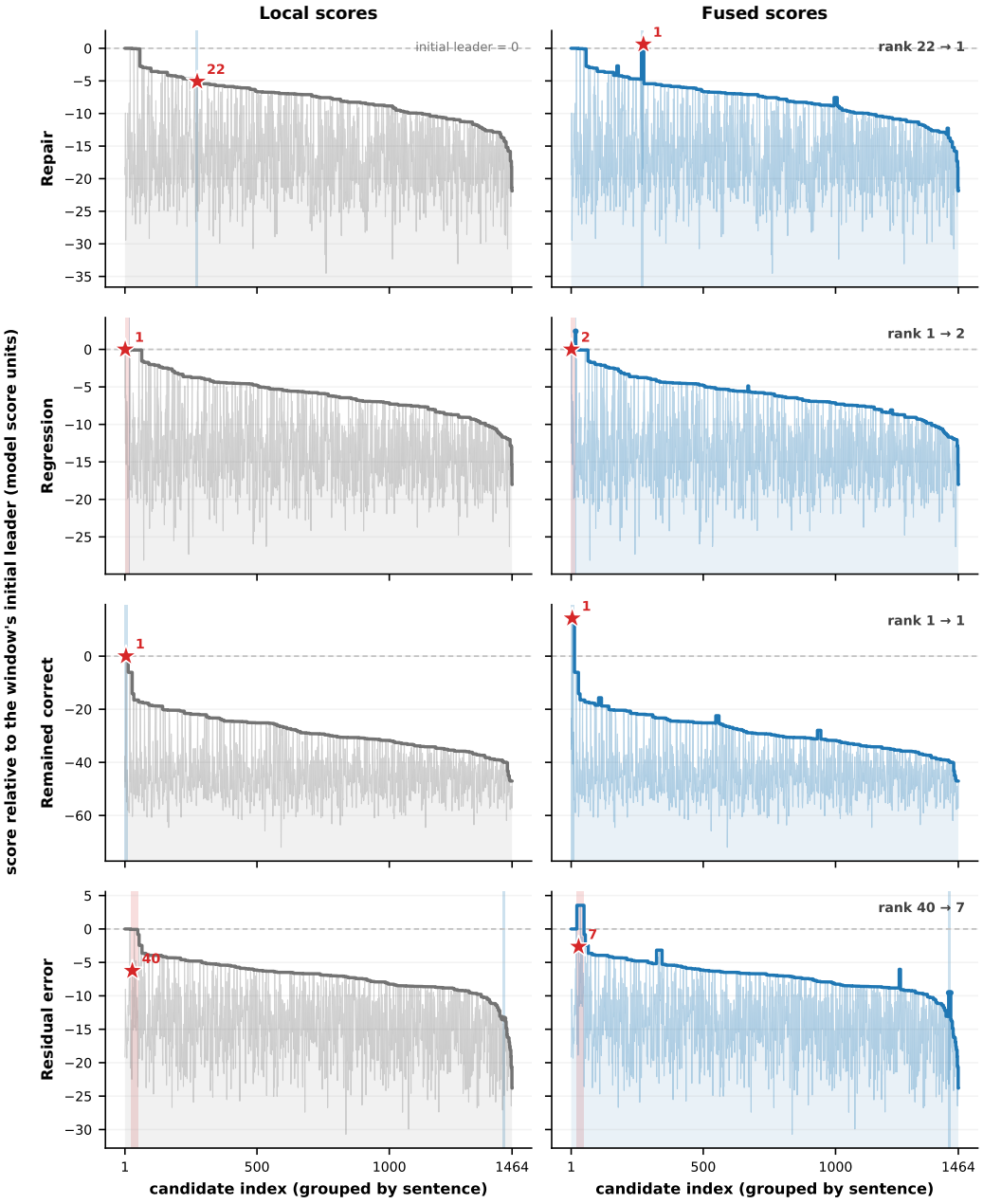

Figure 4: **Candidate scores before and after contextual fusion.** Rows show a repair, a regression, a prediction that remains correct, and a residual error from local decoder seed 0 on MEG-MASC. The left and right columns show local and fused scores. Candidates are grouped by sentence and retain the same horizontal positions within each pair; stars mark the correct candidate. Both score profiles in each row are shifted by the initial local winner's score, preserving ranks and margins. Shaded blocks highlight the contextual and correct-candidate sentence buckets. In the residual example, the correct candidate rises from rank 40 to rank 7 while the prediction remains incorrect.

local competitor. The range bound certifies 1 of the 29 negative cells, while the competitor-based bound certifies 11, all at $\alpha \leq 0.5$. The slack of the competitor-based bound is

$$c_r - z_r = \max_{k \neq y_r} \tilde{s}_{r,k} - \tilde{s}_{r,j_r^*}.$$

The competitor-based bound is attained when j_r^* remains the strongest incorrect competitor after fusion. It is attained more often under history aggregation than under the word-level teacher-forced prior. The empirical tightness therefore depends on both the prior score range and the competitors that determine the final margin.

G Correction-reach interventions

G.1 History-aggregation fusion-weight sweep

Table 9 reports all eight weights with fixed local scores, prior scores, candidate pool, prior-state trajectories, and test windows. The 32 sufficiently populated cells retain negative pre-fusion ordering (0.111–0.378) and positive prior-margin ordering (0.724–0.937), while the fused margin becomes positive across progressively higher initial-rank groups. Reach increases from 3.0 to 9.3, accompanied by additional repairs and regressions (Table 10). Figure 5 shows the control margins, and Figure 6 the score-deficit analysis.

The history-aggregation boundary sequence is non-decreasing in all 1,000 participant resamples and matches the point-estimate sequence at every weight in 903 of the 1,000 resamples.

Table 9: History-aggregation sweep: fused-margin repair-separation AUROC (three-decoder means) and deepest positive group. All eight weights are shown. Dashes denote fewer than five repairs in at least one decoder. Control curves are in Figure 5.

α	2–3	4–5	6–10	11–20	21–50	Deepest positive group
0.5	0.229	0.143	–	–	–	none
0.7	0.303	0.200	0.180	–	–	none
1	0.399	0.275	0.210	–	–	none
1.5	0.520	0.379	0.319	0.265	–	2–3
2	0.613	0.480	0.399	0.336	0.350	2–3
2.7	0.704	0.579	0.487	0.417	0.392	4–5
3.3	0.758	0.633	0.551	0.468	0.434	6–10
4	0.802	0.687	0.602	0.533	0.452	11–20

Table 10: History-aggregation reach and intervention outcomes at six fusion weights. Counts are per-decoder means.

α	Repairs	R_{90}	ΔR@1 (pp)	Regressions
0.5	2,376	3.0	3.30	7.3
1.0	4,554	4.0	6.33	16.7
2.0	8,278	5.0	11.49	34.3
2.7	10,381	7.0	14.40	51.3
3.3	11,910	8.0	16.52	65.3
4.0	13,410	9.3	18.58	88.0

G.2 Score-deficit groups

Grouping by the initial score deficit $G_t^0 = \max_{j \neq y_t} s_{t,j} - s_{t,y_t}$ gives the same direction of boundary movement (Table 11; Figure 6). Five equal-mass groups of initially incorrect predictions remain fixed across weights, with cutoffs 1.24, 2.51, 3.97, and 6.03. Score-deficit reach, $D_{90} = Q_{0.90}(G_t^0 \mid \text{repair})$, is computed within each decoder and averaged. It rises from 0.61 to 3.24 while the deepest positive group advances from none to the fourth group.

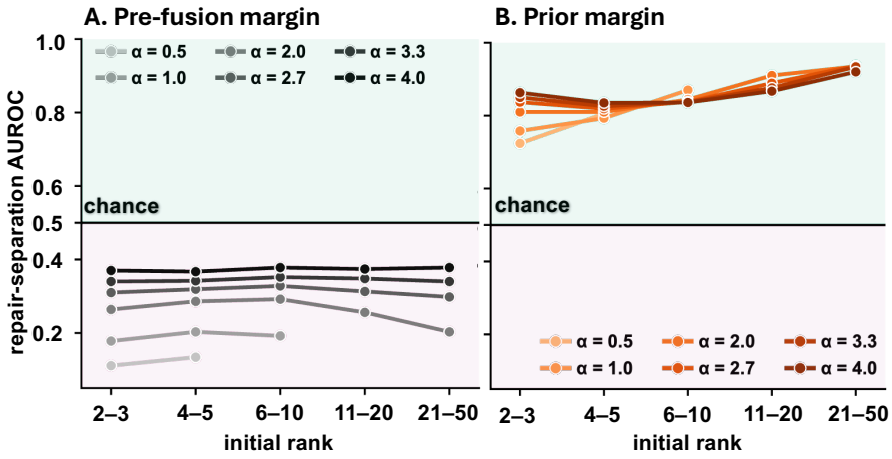

Figure 5: **Pre-fusion and prior margins across the fusion-weight sweep.** Both panels use the history-aggregation prior on MEG-MASC and the same repairs and residual errors at each α. Darker curves indicate larger weights; 0.5 marks chance. **(A) Pre-fusion margin** remains below chance in every reported initial-rank group. **(B) Prior margin** remains above chance. Groups with fewer than five repairs in any decoder are omitted, as in Table 9.

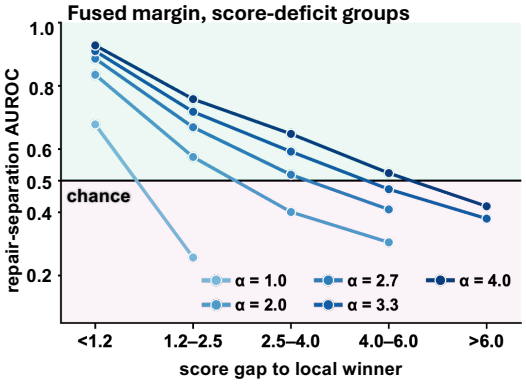

Figure 6: **Fused-margin ordering in score-deficit groups.** History-aggregation prior on MEG-MASC, with the same repairs and residual errors at each α as in Figure 5. Darker curves indicate larger weights; 0.5 marks chance. Five equal-mass groups of initial score deficit are fixed across weights, with cutoffs 1.24, 2.51, 3.97, and 6.03; the boundary moves toward larger deficits as α increases. The $\alpha = 0.5$ curve is omitted because only one group qualifies. Omitted cells follow the event-count rule of Table 11.

Table 11: Fused-margin repair-separation AUROC in fixed score-deficit groups on MEG-MASC. Entries are means across three local decoders; dashes denote fewer than five repairs in at least one decoder. Groups 1–5 are ordered by increasing initial score deficit.

α	Score-deficit group					Deepest group
	1	2	3	4	5	mean AUROC > 0.5
0.5	0.441	–	–	–	–	none
0.7	0.561	0.148	–	–	–	1
1.0	0.678	0.257	–	–	–	1
1.5	0.777	0.447	0.306	–	–	1
2.0	0.835	0.575	0.401	0.305	–	2
2.7	0.886	0.669	0.519	0.409	–	3
3.3	0.910	0.718	0.592	0.473	0.380	3
4.0	0.928	0.758	0.648	0.524	0.419	4

G.3 Structurally repairable errors

Shared bucket scores preserve the local order within each bucket: $\tilde{s}_{t,i} - \tilde{s}_{t,j} = s_{t,i} - s_{t,j}$ for $i, j \in \mathcal{C}_b$. An additive repair therefore requires the correct candidate to have the highest local score within its bucket. With $\mathbf{1}[\cdot]$ denoting the indicator function, the structurally repairable subset is

$$\text{repairable}_t = \mathbf{1}\left[s_{t,y_t} = \max_{j \in \mathcal{C}_{b(y_t)}} s_{t,j} \right]. \tag{16}$$

This subset contains 98.6% of initial errors at ranks 2–3 and 72.5% at ranks 21–50, retaining every repair. Across the eight-weight sweep, restricting comparisons to this subset changes fused-margin AUROC by at most 0.004 and preserves every boundary in each decoder. At the main operating point, pre-fusion ordering remains negative and prior-margin ordering positive. The reversal therefore persists among errors satisfying the within-bucket condition for repair.

G.4 Word-level LM sweeps

Tables 12 and 13 report all sixteen weights for each word-level prior with both score channels fixed. The teacher-forced prior extends positive fused-margin ordering from no group at $\alpha = 0.05$ to all six observed groups at $\alpha = 1$, as reach increases from 3.0 to 19.3. All observed groups remain positive at larger weights, so no crossing is observed through rank 100 for $\alpha \geq 1$. Figure 3C plots the observed crossings at $\alpha \leq 0.75$.

This extension continues after net accuracy gain peaks: between $\alpha = 0.5$ and 1, gain declines from 11.20 to 9.16 percentage points while positive ordering expands from ranks 6–10 to all observed groups. At $\alpha \geq 3$, regressions outnumber repairs even though fused-margin repair ordering remains positive throughout the observed range.

The free-running prior produces a different trajectory. At its validation-selected weight, $\alpha = 0.15$, reach is 3.0 and gain is 1.25 percentage points. Positive ordering first appears at ranks 2–3 when $\alpha = 0.5$, where gain is -0.69 points. At $\alpha = 2$ and 3, ranks 21–50 remain negative while ranks 51–100 are positive; at $\alpha = 4$, the deepest positive group moves back to ranks 11–20. Fusion weight therefore changes both the shape and extent of positive ordering.

Pre-fusion and prior margins. At each weight, we evaluate the pre-fusion and prior margins on the same repair and residual-error labels as the fused margin, showing how ordering changes for fixed input scores as outcomes change. Through $\alpha = 1$, pre-fusion ordering is negative and prior-margin ordering positive in every reported group for both word-level priors (Table 15). At larger weights, pre-fusion AUROC approaches 0.5, and some groups become positive. For the teacher-forced prior, ranks 11–20 reach 0.503 at $\alpha = 1.5$, with a net accuracy gain of 5.05 percentage points. Teacher-forced prior-margin ordering stays positive throughout the sweep; the free-running prior margin has one negative group, ranks 51–100 at $\alpha = 16$ (0.479).

Table 12: Word-level teacher-forced LM sweep. Fused-margin AUROCs, reach, and accuracy gain are three-decoder means; validation selects $\alpha = 0.5$. We show every tested weight and all six rank groups. Dashes follow the repair-count rule. Event counts are in Table 14.

α	Initial rank						Deepest positive group	R_{90}	ΔR@1 (pp)
	2–3	4–5	6–10	11–20	21–50	51–100			
0.05	0.173	0.112	–	–	–	–	none	3.0	+2.21
0.1	0.301	0.212	0.172	–	–	–	none	3.0	+4.20
0.15	0.399	0.305	0.248	0.205	–	–	none	4.0	+5.88
0.2	0.471	0.378	0.301	0.257	–	–	none	5.0	+7.29
0.25	0.526	0.432	0.356	0.291	0.260	–	2–3	5.3	+8.47
0.3	0.571	0.475	0.394	0.335	0.278	–	2–3	6.3	+9.45
0.4	0.633	0.547	0.480	0.417	0.356	0.267	4–5	8.3	+10.65
0.5	0.674	0.595	0.535	0.484	0.419	0.366	6–10	10.3	+11.20
0.75	0.732	0.664	0.623	0.567	0.517	0.486	21–50	15.0	+10.80
1	0.762	0.701	0.667	0.617	0.571	0.557	51–100	19.3	+9.16
1.5	0.785	0.733	0.705	0.671	0.640	0.625	51–100	25.7	+5.05
2	0.800	0.747	0.717	0.686	0.667	0.647	51–100	30.0	+1.31
3	0.806	0.765	0.733	0.703	0.687	0.682	51–100	35.7	-4.19
4	0.810	0.772	0.739	0.707	0.695	0.695	51–100	39.3	-7.72
8	0.808	0.768	0.741	0.708	0.695	0.695	51–100	47.7	-14.41
16	0.796	0.750	0.727	0.690	0.680	0.677	51–100	52.7	-18.55

Table 13: Word-level free-running LM sweep. Fused-margin AUROCs, reach, and accuracy gain are three-decoder means; validation selects $\alpha = 0.15$. We show every tested weight and all six rank groups. Dashes follow the repair-count rule. Event counts are in Table 14.

α	Initial rank						Deepest positive group	R_{90}	ΔR@1 (pp)
	2–3	4–5	6–10	11–20	21–50	51–100			
0.05	0.124	0.082	–	–	–	–	none	2.0	+0.64
0.1	0.214	0.169	–	–	–	–	none	3.0	+1.03
0.15	0.288	0.206	0.170	–	–	–	none	3.0	+1.25
0.2	0.347	0.264	0.225	–	–	–	none	3.0	+1.28
0.25	0.394	0.305	0.253	–	–	–	none	4.0	+1.20
0.3	0.431	0.347	0.290	0.259	–	–	none	4.0	+1.00
0.4	0.492	0.395	0.348	0.330	–	–	none	4.0	+0.27
0.5	0.539	0.433	0.381	0.358	0.211	–	2–3	5.0	-0.69
0.75	0.607	0.506	0.452	0.419	0.338	–	4–5	5.3	-3.71
1	0.645	0.537	0.482	0.435	0.420	–	4–5	6.0	-6.78
1.5	0.691	0.584	0.534	0.474	0.427	0.486	6–10	7.0	-12.15
2	0.714	0.608	0.564	0.496	0.442	0.520	51–100	8.0	-16.23
3	0.739	0.644	0.592	0.527	0.451	0.519	51–100	9.0	-21.49
4	0.755	0.668	0.606	0.548	0.463	0.462	11–20	9.7	-24.66
8	0.782	0.691	0.632	0.547	0.491	0.454	11–20	11.3	-30.04
16	0.787	0.690	0.626	0.529	0.487	0.463	11–20	13.0	-33.13

Table 14: Repairs and regressions in the word-level sweeps, shown as means per local decoder.

α	Teacher-forced		Free-running	
	Repairs	Regressions	Repairs	Regressions
0.05	1,748.0	161.0	832.3	373.3
0.1	3,358.3	343.3	1,536.7	796.3
0.15	4,771.7	553.7	2,121.0	1,223.3
0.2	6,027.0	796.7	2,597.3	1,680.0
0.25	7,120.7	1,048.0	3,000.3	2,137.7
0.3	8,086.0	1,308.7	3,323.0	2,606.0
0.4	9,542.7	1,899.3	3,783.0	3,586.7
0.5	10,575.0	2,542.3	4,058.7	4,554.0
0.75	11,972.0	4,222.0	4,303.0	6,962.0
1	12,434.0	5,865.0	4,269.0	9,131.3
1.5	12,378.0	8,756.0	3,908.7	12,627.7
2	12,016.7	11,078.3	3,540.3	15,181.3
3	11,251.7	14,260.7	3,035.3	18,449.3
4	10,659.0	16,195.3	2,712.3	20,399.7
8	9,402.0	19,742.0	2,119.3	23,671.7
16	8,625.0	21,934.7	1,783.3	25,551.0

Table 15: Pre-fusion and prior-margin repair-separation AUROC in the word-level sweeps. Each range spans the reported three-decoder means over all sufficiently populated initial-rank groups and tested weights in the indicated interval. Both measures use the same cells and outcome labels.

Word-level LM prior	Weights	Cells	Pre-fusion margin	Prior margin
Teacher-forced	$\alpha \leq 1$	47	0.102–0.494	0.554–0.913
	$\alpha > 1$	36	0.477–0.509	0.566–0.706
Free-running	$\alpha \leq 1$	36	0.098–0.475	0.522–0.711
	$\alpha > 1$	36	0.450–0.525	0.479–0.596

G.5 Reach and ordering across interventions

At the selected operating points, effective correction reach spans 3.0–15.7 across six interventions (Table 16). Both free-running priors have no positively ordered group; the history-aggregation and word-level teacher-forced priors extend positive ordering through ranks 4–5 and 6–10; hard bucket pruning and the sentence-level teacher-forced prior extend it through ranks 11–20 (Table 17).

Configurations with similar reach can have different ordering boundaries. History-aggregation fusion at $\alpha = 4$ has $R_{90} = 9.3$ and positive ordering through ranks 11–20; the word-level teacher-forced prior at $\alpha = 0.5$ has $R_{90} = 10.3$ and positive ordering through ranks 6–10. Thus, reach locates the repair distribution, while the AUROC curves additionally describe how repair and residual-error margins are arranged.

H Prior information and revision outcomes

For the history-aggregation prior, the highest-scoring sentence bucket is correct on 98% of windows in the highest prior-margin quintile and 20% in the lowest. Among initially correct predictions, the highest prior-margin region places 99.9% of correct candidates in the prior's leading bucket. Shuffling candidate-to-sentence assignments changes net additional correct predictions from $10{,}329 \pm 190$ to -198 ± 23 per decoder. This contrast links the gain to aligning accumulated support with candidate identity.

Table 18 evaluates prior margin among initially correct test predictions, using those that remain correct as positives and regressions as negatives. At the tested settings, prior margin distinguishes

Table 16: Reach and intervention outcomes at the selected operating points on MEG-MASC. Reach and accuracy gain are mean ± SD across three local decoders; repair and regression counts are per-decoder means.

Intervention	R_{90}	Repairs	Regressions	ΔR@1 (pp)
Sentence LM, free-running	3.0 ± 0.0	2,529.0	1,579.7	1.32 ± 0.13
Word LM, free-running	3.0 ± 0.0	2,121.0	1,223.3	1.25 ± 0.07
History aggregation	7.0 ± 0.0	10,380.7	51.3	14.40 ± 0.26
Word LM, teacher-forced	10.3 ± 0.6	10,575.0	2,542.3	11.20 ± 0.23
Hard bucket pruning	15.3 ± 0.6	16,452.7	992.7	21.55 ± 0.24
Sentence LM, teacher-forced	15.7 ± 0.6	14,148.3	3,573.7	14.74 ± 0.41

Table 17: Fused- and prior-margin repair-separation AUROC at the six selected settings (three-decoder means). Dashes follow the repair-count rule; blank cells were not assessed. The 51–100 group is used only for word-level priors.

Prior	Margin	2–3	4–5	6–10	11–20	21–50	51–100
Sentence LM, free-running	Fused	0.314	0.197	0.156	–	–	
	Prior	0.603	0.663	0.782	–	–	
History aggregation	Fused	0.704	0.579	0.487	0.417	0.392	
	Prior	0.838	0.819	0.838	0.889	0.937	
Hard bucket pruning	Fused	0.838	0.745	0.656	0.559	0.454	
	Prior	0.773	0.713	0.675	0.638	0.621	
Sentence LM, teacher-forced	Fused	0.778	0.722	0.661	0.585	0.481	
	Prior	0.816	0.846	0.861	0.869	0.873	
Word LM, free-running	Fused	0.288	0.206	0.170	–	–	–
	Prior	0.532	0.560	0.701	–	–	–
Word LM, teacher-forced	Fused	0.674	0.595	0.535	0.484	0.419	0.366
	Prior	0.568	0.600	0.617	0.659	0.752	0.759

preserved correct predictions from regressions more strongly for history aggregation than for the word-level LM priors.

Table 18: Prior-margin AUROC among initially correct MEG-MASC predictions at validation-selected weights. Positives remain correct; negatives are regressions. Means are across three decoders, with SD where available.

Prior	α	AUROC
History aggregation	2.7	0.777 ± 0.022
Word-level teacher-forced LM	0.5	0.545
Word-level free-running LM	0.15	0.532

I Direct confidence measures and learned estimators

I.1 Features and prediction targets

Table 19 separates local, fused, and prior information. The prior state summarises preceding windows; current support comes from the present window. Sentence-position features use the window index and the annotated presentation length. Removing position features in separately fitted top-1 estimators changes repair-separation AUROC over local errors from 0.775 to 0.783; removing support entropy as well gives 0.783.

Two-layer MLPs with 32 units per hidden layer predict fused top-K inclusion for $K \in \{1, 3, 5, 10, 20\}$. Training uses five-fold, presentation-grouped validation data; per-head temperature

Table 19: Observable inputs to the confidence estimators. Decoder scores combine local and fused features (8); prior features combine prior state and additional prior features (7); the full estimator uses all 15.

Feature group	Count	Inputs
Local	2	Pre-fusion margin; local top-16 entropy-based confidence.
Fused	6	Fused-score gaps at $K \in \{1, 3, 5, 10, 20\}$; fused top-16 entropy-based confidence.
Prior state	3	History indicator; prior-state entropy-based confidence; prior-state margin.
Additional prior	4	Current-support entropy-based confidence; leading-bucket agreement between prior state and current support; two sentence-position features.

calibration (Guo et al., 2017) uses out-of-fold logits. Repair-focused evaluations use test windows where the local top-K set excludes the correct candidate; at $K = 1$, success is repair. Tied candidates share a rank, so rank-based inclusion resolves ties optimistically. Pooled correctness uses all test windows and fused top-1 correctness. Results average five estimator seeds within each of three local decoders.

I.2 Outcome discrimination

At initial ranks 21–50, negative repair ordering extends beyond the fused margin: fused top-16 entropy-based confidence, sentence-score mass gap, and sentence-score concentration give AUROCs of 0.36, 0.46, and 0.43. The last two summarise fused score mass between and across sentence buckets. Table 20 shows the positive ordering recovered by learned estimates. Over all test windows, the full estimator achieves an AUROC of 0.954, compared with 0.872 for the fused margin; among local errors, the latter achieves 0.646, falling to 0.415 above rank 10 and 0.344 when comparisons are restricted within presentations.

Table 20: AUROC for fused top-K success within initial-rank groups on MEG-MASC. Values are means across three local decoders, averaged over estimator seeds within each decoder. The $K = 1$ columns measure repair-separation AUROC, and n_{repair} counts top-1 repairs per local decoder.

(a) Top-1 success

Initial rank	n_{repair}	Local scores	Decoder scores	Prior features
2–3	6,971	0.399	0.891	0.720
4–5	1,832	0.318	0.833	0.728
6–10	1,191	0.284	0.779	0.775
11–20	327	0.240	0.716	0.848
21–50	58	0.271	0.653	0.910

(b) Top-5 success

Initial rank	Decoder scores	Prior features	Full
6–10	0.663	0.681	0.676
11–20	0.552	0.752	0.676
21–50	0.522	0.867	0.771

Table 21 complements ranking with rare-repair precision. At ranks 21–50, the prior-state estimator's highest-scoring 10% of predictions have an 8.40% repair rate, compared with 1.80% for decoder scores and a population rate of 1.05%.

Table 21: Rare-repair prediction with learned decoder-score and prior-state estimators on MEG-MASC. P@10% is the repair rate among the 10% highest-scoring predictions. The prior-state estimator uses the prior-state group in Table 19.

Initial rank	Repair rate (%)	AUPRC		P@10% (%)	
		Decoder-score	Prior-state	Decoder-score	Prior-state
11–20	6.44	0.129	0.396	13.70	36.40
21–50	1.05	0.018	0.242	1.80	8.40

I.3 Selective-output policies

At the history-aggregation setting $\alpha = 2.7$, adaptive policies emit the smallest set whose calibrated success probability exceeds its threshold, or abstain. Thresholds are selected on validation under coverage $\geq 90\%$ and mean $K \leq 4.8$. The margin-binned reference applies conformal calibration in 20 fused-margin bins; isotonic calibration (Zadrozny & Elkan, 2002) of that margin finds no point satisfying both constraints. Table 22 reports the selected policies on test windows; validation constraints are not guaranteed on test, where mean K slightly exceeds 4.8, and fixed-K coverage falls below 90% for $K \geq 3$ with learned estimators and at every K for margin-based policies (Table 24). Emission uses all windows as its denominator; coverage and mean K use emitted sets.

Table 22: Test performance of validation-selected selective-output policies with the history-aggregation prior. Means and SDs are across three local decoders, after averaging estimator seeds within each decoder. The stratified analysis uses these adaptive policies; fixed-K thresholds are selected separately.

Policy	Emission (%)	Coverage (%)	Mean K
Margin-binned	55.9 ± 8.0	91.7 ± 3.0	4.90
Decoder-score	73.6 ± 0.1	91.8 ± 0.2	5.02 ± 0.05
Full	74.5 ± 0.8	92.2 ± 0.3	5.18 ± 0.08

The same policies are applied to all test windows before grouping by initial rank (Table 23). Adding prior features raises both emission and coverage at ranks 21–50; beyond rank 50, coverage remains low for both policies.

Table 23: Selective output by initial rank on MEG-MASC with the history-aggregation prior. Test results at each policy's selected operating point, averaged across local decoders. Emission uses all windows in the group as its denominator; coverage and mean K use emitted sets. Rates are percentages.

	Decoder-score			Full		
Initial rank	Emission	Coverage	Mean K	Emission	Coverage	Mean K
1	96.8	100.0	2.36	96.9	100.0	2.33
2–3	89.1	99.7	5.48	86.9	99.5	5.92
4–5	72.3	97.8	9.15	70.0	97.1	8.80
6–10	51.7	93.1	11.20	55.6	93.5	10.32
11–20	35.6	73.9	12.35	43.8	82.5	12.02
21–50	32.5	36.4	12.61	38.2	50.8	12.86
51+	30.2	3.4	12.76	30.9	7.3	13.75

For fixed-K output, validation instead chooses a separate threshold for each policy and K, maximising emission at coverage $\geq 90\%$. All searches are feasible; no additional mean-set-size constraint applies. Table 24 reports test emission and coverage at these thresholds for every assessed K.

I.4 Feature comparison at matched operating points

The six feature sets in Table 25 share cached inputs, a hidden-layer architecture, and the same training procedure. For each run, shifting its five thresholds together traces a test operating curve.

Table 24: Fixed-size selective output on MEG-MASC with the history-aggregation prior. Thresholds are selected on validation separately for each policy and K. Entries are test percentages averaged across local decoders, with estimator seeds averaged where applicable. Emission uses all test windows as its denominator; coverage uses emitted sets.

Policy	Rate	$K = 1$	$K = 3$	$K = 5$	$K = 10$	$K = 20$
Full	Emission	58.3	71.2	77.1	86.3	99.5
	Coverage	90.5	89.2	88.5	87.4	86.0
Decoder-score	Emission	56.8	69.3	74.8	84.2	99.2
	Coverage	90.7	89.6	88.9	87.7	86.0
Margin-binned	Emission	48.7	55.8	63.0	76.3	94.1
	Coverage	89.1	88.8	87.7	86.5	86.6
Margin isotonic	Emission	49.7	57.9	64.7	75.9	98.4
	Coverage	88.4	87.8	87.0	86.7	86.0

Linear interpolation matches emission at 74.52% or coverage at 92.24%, the reference full policy's operating values. Runs without a feasible validation point start from the search's penalised optimum. We match per run, then average within and across decoders. These are test-curve comparisons; Table 22 separately evaluates validation-selected policies.

Table 25: Feature comparison with the history-aggregation prior. Matched emission and coverage are percentages; parentheses give mean emitted set size. Validation counts report runs with a feasible point under coverage $\geq 90\%$ and mean $K \leq 4.8$, before the test-curve sweep. Values are matched per run and then averaged over seeds within decoder; SDs are across decoders.

Features	Coverage at 74.52% emission	Emission at 92.24% coverage	Validation feasible
Fused margin	87.2 ± 0.1 (7.1)	56.7 ± 0.5 (6.3)	0/15
Fused scores	89.3 ± 0.2 (5.5)	66.6 ± 0.7 (5.8)	6/15
Local scores	86.4 ± 0.1 (7.7)	51.9 ± 0.9 (8.2)	0/15
Prior features	87.9 ± 0.3 (6.4)	61.5 ± 1.4 (6.3)	0/15
Decoder scores	91.4 ± 0.2 (5.1)	72.5 ± 0.5 (5.2)	15/15
Full	92.2 ± 0.2 (5.2)	74.5 ± 0.4 (5.2)	15/15