NEAR-BENCH: BENCHMARKING PREDICTIVE PERFORMANCE AND EEG INPUT DEPENDENCE IN EEG-TO-TEXT DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Advances in electroencephalography (EEG) representation learning and pretrained language models have driven research on decoding text from noninvasive brain signals. However, differences in data splits and evaluation tasks hinder direct comparison across studies; even under identical evaluation conditions, higher text scores do not necessarily indicate greater use of EEG corresponding to the target text. We introduce NEAR-Bench, an evaluation and diagnostic benchmark that jointly assesses text prediction performance and dependence on EEG input. Built on ZuCo 1.0 and 2.0, NEAR-Bench standardizes data splits, candidate text identities, and reporting rules, jointly holding out participants and texts during downstream training. For sequence decoders, we fix the trained model, replace only its EEG input, and compare predictions on the same observations. We also evaluate candidate identification separately from text generation without candidate sentences or reference prefixes. Results show that better text prediction does not necessarily accompany stronger EEG input dependence: some sequence models outperform an independent language-model reference, yet their prediction scores and generation behavior change little when EEG values are zeroed or replaced with other real recordings. NEAR-Bench provides a common basis for comparing model performance under explicit conditions, testing the contribution of correctly paired EEG input, and interpreting progress in EEG-to-text.

1 INTRODUCTION

Translating neural activity into language is a goal of brain–computer interface research to restore or support communication. Electroencephalography (EEG) offers a noninvasive approach. The ZuCo natural-reading datasets provide synchronized EEG, text stimuli, and eye-tracking recordings for studying brain–language correspondence (Hollenstein et al., 2018; 2020). Building on these resources, EEG-to-text methods have progressed from connecting EEG representations to pretrained language models to semantic contrastive learning and discrete EEG representations (Wang & Ji, 2022; Feng et al., 2023; Duan et al., 2023). As studies report increasingly "better" text prediction, interpreting these scores accurately becomes more important.

Recent studies have therefore reexamined EEG-to-text evaluation reliability. Jo et al. (2024) found that the evaluated systems retained similar language scores after EEG was replaced with noise, while generation without reference prefixes remained limited. To examine which information predictions depend on, recent work uses semantic retrieval, autonomous generation, and test-time input controls (Liu et al., 2025; Wang et al., 2026). Addressing data acquisition and evaluation design, COFETT examines how recording and test conditions affect capability assessments (Zhang et al., 2026). Together, these studies motivate analysis of output information sources and conditional dependencies beyond final text metrics.

Existing evaluations, however, address different questions. Generalization to new participants differs from generalization to new texts; candidate identification differs from generation without those candidates. When these factors change together, performance differences may reflect EEG representations, language-model priors, candidate-space constraints, or data pairings. We therefore ask:

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.

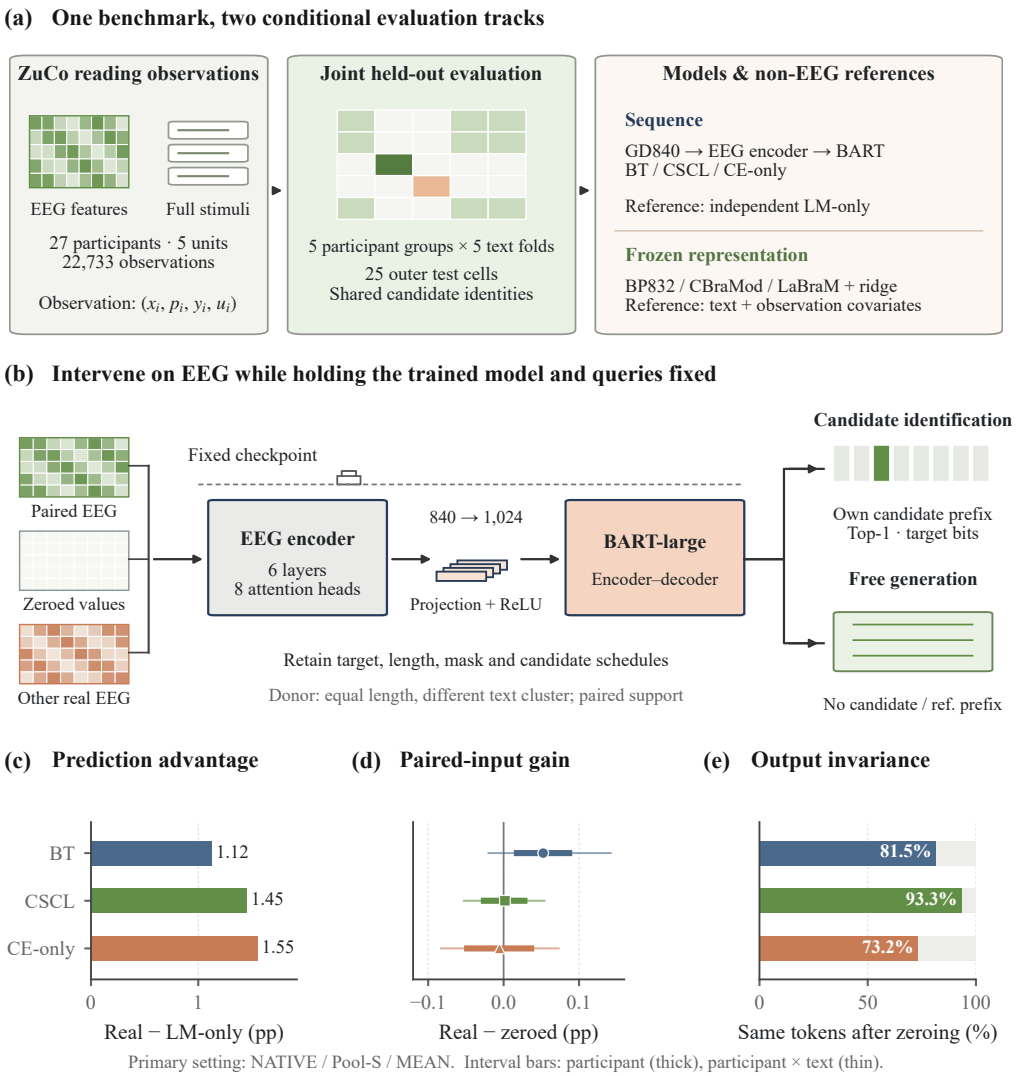


Figure 1: **NEAR-Bench overview.** (a,b) Two evaluation tracks and same-checkpoint input controls. (c–e) Higher predictive scores coexist with small paired-input gains and stable outputs under NATIVE/Pool-S/MEAN. Thick/thin bars: participant/crossed 95% intervals.

when predictive performance improves, does the improvement benefit from EEG information correctly paired with the target text?

To address this question, we introduce NEAR-Bench (Figure 1), with three main contributions.

Evaluation protocol. We combine joint participant–text holdout during downstream training, shared candidate identities, and explicit information conditions in an EEG-to-text information-attribution benchmark.

Controlled comparisons. For sequence models, we use same-checkpoint input replacements to distinguish model differences from contributions of the current recording, and report candidate identification and generation separately. For frozen representations, a separate adaptation procedure supports conditional analyses. Because these paradigms differ in prediction objectives, each is evaluated under its own scoring rules and references, without a cross-paradigm ranking.

Empirical findings. Under these conditions, some sequence methods outperform an independent language-model reference, yet replacing EEG at the same checkpoint barely changes candidate iden-

tification, and generated outputs remain highly consistent across input conditions. Their predictive advantages thus need not imply greater use of correctly paired EEG. Further learning checks and local generalization diagnostics show that artificial-signal learnability, training-text performance, and participant/text exposure do not automatically yield larger real-input gains.

2 Related Work

2.1 EEG-to-Text Modeling and Representations

EEG-to-text methods condition pretrained language models on EEG representations: word-level features connect to BART (Wang & Ji, 2022), curriculum contrastive learning organizes semantic relationships (Feng et al., 2023), and DeWave quantizes EEG into language-aligned discrete codes (Duan et al., 2023). LaBraM and CBraMod instead learn transferable EEG representations through masked pretraining (Jiang et al., 2024; Wang et al., 2025). Language-task adaptation must assess input-dependent contributions under its specific representation and prediction procedures.

2.2 EEG-to-Text Evaluation and Input Dependence

Text information available to the decoder changes what evaluation measures. Teacher forcing supplies correct preceding tokens; autonomous generation uses model outputs. Jo et al. (2024) found similar scores after replacing EEG with noise but limited generation without teacher forcing, motivating separation of reference-prefix information from paired EEG. Semantic evaluation uses retrieval and embedding comparisons. GLIM combines generation, EEG–text retrieval, and semantic classification (Liu et al., 2025); SemKey compares generated-text embeddings with candidates (Wang et al., 2026). Brain-CLIPLM uses EEG semantic anchors, task-corpus retrieval, and a language model to reconstruct sentences (Yang et al., 2026). Evaluation candidates and prediction-time retrieval corpora have different roles, so these tasks must be interpreted separately. GLIM and SemKey also replace EEG with Gaussian noise at inference. Because this shifts the input distribution, we additionally use real recordings from other texts while fixing weights, targets, and candidates. Noise responses and benefits of correct pairing require separate interpretation.

2.3 EEG Benchmarking and Generalization

MOABB standardizes BCI dataset interfaces (Jayaram & Barachant, 2018); EEG-FM-Bench compares EEG foundation models across adaptation settings with gradient and representation diagnostics (Xiong et al., 2026). EEG-to-text benchmarks must also distinguish participant and text generalization. GLIM and SemKey use text-disjoint splits (Liu et al., 2025; Wang et al., 2026); COFETT studies within-participant transfer across repeated reading and recall sessions (Zhang et al., 2026). NEAR-Bench combines joint participant–text holdout, separate identification and generation reports, and same-checkpoint input controls to test whether better prediction accompanies larger paired EEG gains.

3 Benchmark Design

3.1 Evaluation Tasks

NEAR-Bench evaluates EEG–text pairs collected during reading. Each observation contains EEG x_i, participant p_i, reference sentence y_i, and dataset–task unit u_i. **Candidate identification** requires scoring eight sentences—the reference and seven distractors—to measure discrimination among candidates. **Text generation** examines sequence models' generation behavior without candidate sentences or reference prefixes.

For EEG-conditioned generation methods, the **paired input gain** is the candidate identification performance difference between correctly paired EEG and a specified replacement input, with model weights and evaluation observations fixed. Changes in candidate rankings, generated outputs, and internal activations further characterize responses to input.

3.2 Data Construction

The benchmark uses ZuCo 1.0 and 2.0 (Hollenstein et al., 2018; 2020). After reserving three participants for development and auditing, the main cohort comprises 27 participants and 22,733 participant–sentence observations from five dataset–task units, before filtering by input requirements (Table F.1). Reference answers retain complete original stimulus sentences, without new manual annotations. Repeated readings remain distinct EEG observations, while stimulus identities retain each sentence's dataset and task.

Identical and identified near-duplicate texts are grouped before splitting, with each group assigned to one text fold to prevent leakage (Appendix A.2).

3.3 Candidate Construction

Each stimulus has eight fixed eight-way candidate sets per pool, shared across participants and systems, with the target appearing once in each position. All candidates share a dataset–task unit and text fold; distractors occupy distinct text clusters excluding the target cluster.

The **matched pool (Pool-S)** matches quartile bins of word count and mean language-model negative log-likelihood (NLL), expanding the neighborhood only if fewer than seven distractor clusters are available. The **broader pool (Pool-L)** uses all eligible sentences without matching, testing candidate-composition effects. Both have a 12.5% uniform Top-1 reference.

4 Evaluation Protocol

4.1 Joint Participant–Text Holdout

Crossing five participant groups with five text folds yields 25 outer test cells (Figure 2). For test cell (P_i, Q_j), validation intersects the next participant group and text fold in cyclic order; training uses all intersections of the remaining three participant groups and three text folds. Other regions are unused. Thus, neither test participants nor test texts appear in training or validation, and every valid observation receives one outer-test prediction. Estimated normalization parameters and adapters are fitted only on training data; hyperparameters and checkpoints are selected on validation data. This downstream-training holdout does not rule out overlap between pretraining data and evaluation texts.

4.2 Evaluation Tracks

The sequence and frozen-representation tracks use distinct inputs, scores, and references (Table F.2). Sequence implementations share an EEG encoder, projection, and BART-large decoder. BT uses SGD; CE-only uses Adam; CSCL adds 15 contrastive epochs to the same 30-epoch Adam cross-entropy budget. Validation token NLL selects checkpoints (Appendices A.3, C.5).

Both tracks share splits, candidates, and reporting rules. NATIVE is each track's valid support; COMMON is their intersection. Comparisons use identical observations on both sides. Sequence inputs retain fixated words, order, and masks, whereas references are complete sentences. The frozen track measures the increment of EEG scores over a richer covariate reference using a shared text-embedding adapter. This conditional auxiliary analysis differs from same-checkpoint sequence replacement; absolute scores do not support a cross-track ranking.

4.3 Evaluation Setting and Input Controls

The primary setting is sequence-track Top-1 on NATIVE with Pool-S and mean token log-likelihood (MEAN). Appendix B tests Pool-L, cumulative scoring (SUM), COMMON support, and continuous scores; Appendix A.5 specifies all settings.

Zeroing tests the contribution of EEG values beyond length and masks: it sets feature values to zero while preserving length and masks. **Donor replacement** selects a real EEG recording from a different text cluster with the same effective length, within the same outer cell and dataset–task unit, preferring recordings from the same participant. It tests the benefit of the original EEG–text pairing over a real recording of equal length. Both controls fix model weights, target texts, candidates,

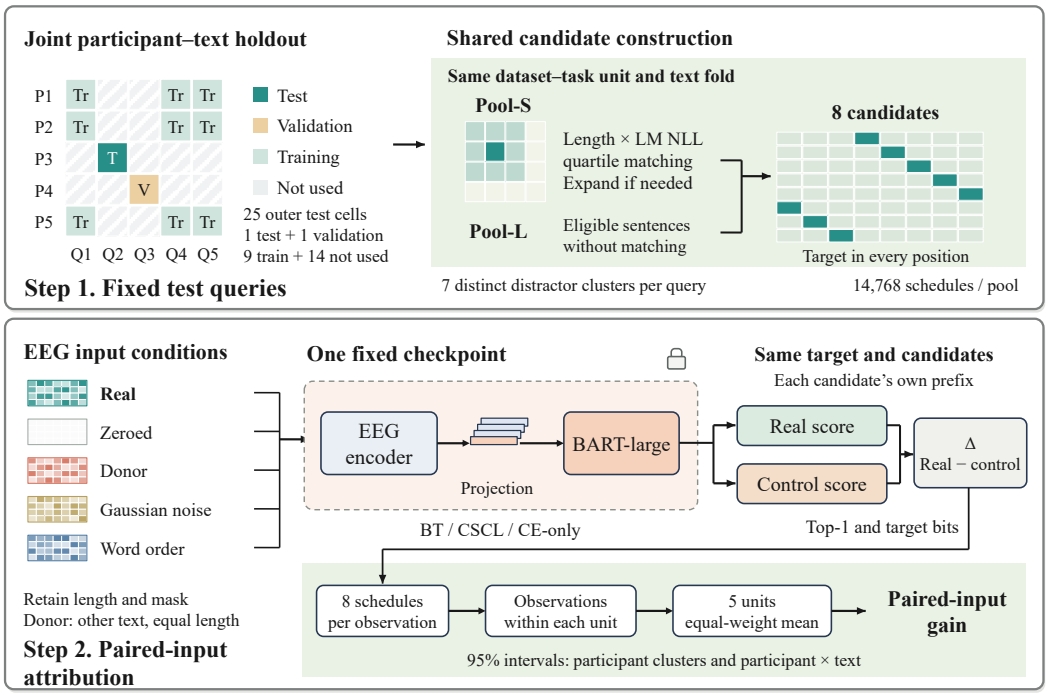


Figure 2: **Paired-input evaluation.** Joint participant–text holdouts share eight-way candidate schedules. Fixed-checkpoint input replacements preserve queries and scoring. Paired differences are averaged over schedules, observations, and five equally weighted units. Full construction: Appendix A.

and scoring rules. Donor analyses retain only observations with eligible replacements, recomputing real and control results on this shared support. Gaussian noise and word-order permutation provide additional sensitivity analyses. The independent LM-only reference contextualizes absolute prediction performance. Its different training history and input format require interpreting this comparison separately from input controls at the same checkpoint.

4.4 Metrics and Aggregation

A sequence model scores candidate y_c by mean token log-likelihood:

$$s_{ic} = \frac{1}{L_c} \sum_{t=1}^{L_c} \log p_\theta(y_{c,t} \mid y_{c,<t}, x_i). \tag{1}$$

Candidate likelihood conditions on the candidate's own preceding tokens; generation receives no reference prefix. L_c counts scored tokens, including tokenizer-added special tokens and excluding padding. Top-1 awards one point when the target alone scores highest, or $1/k$ when it is among k candidates exactly tied for the highest score.

Aggregation averages eight candidate resamples per observation, then observations per dataset–task unit, and weights the five units equally. Writing this aggregation as M, the real-minus-control difference is

$$\Delta_a = M\left(m^{\text{real}} - m^a; \mathcal{S}_a\right), \tag{2}$$

where m is the metric for a single candidate query and $\mathcal{S}_a$ is the paired observation set shared by real input and control a. Accuracy differences are expressed in percentage points (pp). Appendix B also reports SUM scoring and target bits, defined as the base-two logarithm of the target candidate's softmax score within the candidate set.

Generation uses deterministic single-beam decoding with at most 128 new tokens. The analysis retains empty outputs and reports distinct decoded string counts and exact token-sequence agreement

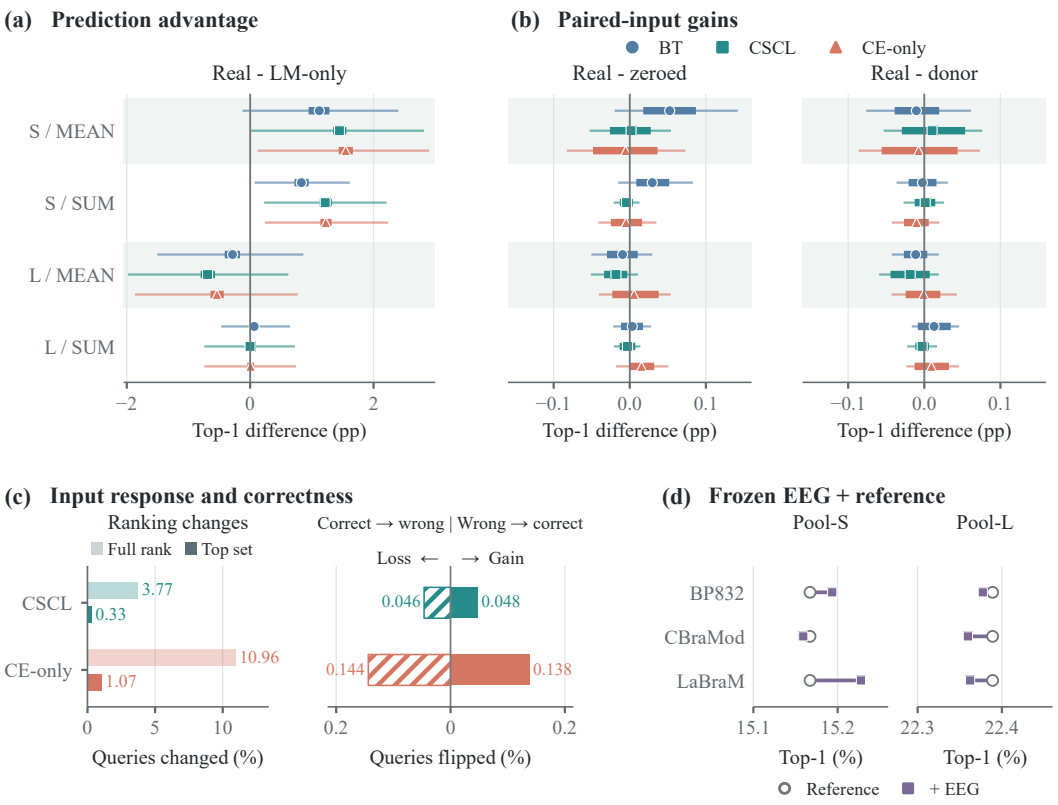


Figure 3: **Prediction and input dependence.** (a,b) Prediction contrasts and paired gains; thick/thin bars: participant/crossed 95% intervals. (c) Ranking/top-set changes are separate event rates; mirrored bars show correctness losses/gains. (d) Frozen-track reference-only versus reference-plus-EEG scores on COMMON. Pool axes each span 0.16 pp; tracks use distinct references.

rates between real and control inputs. These metrics describe output diversity and input responsiveness, respectively, without measuring semantic recovery accuracy.

4.5 Uncertainty and Validation

Uncertainty accounts for repeated participants and texts. Primary participant-clustered intervals resample participants within datasets; crossed intervals independently resample participants and text clusters and multiply their weights. Both use 5,000 resamples and pointwise 95% percentile intervals, preserving real–control pairing and equal unit weights. Models, candidates and their order, and donors remain fixed. Crossed intervals assess text sensitivity; their multiway-clustered coverage is uncalibrated. Appendix B adds text-only intervals and three-seed comparisons in five cells.

Metadata and scoring checks verify split isolation, candidate identities, known answers, ties, and permutations. Artificial-signal, real-data micro-fitting, and relaxed-holdout diagnostics test learning and generalization (Section 5.4). Appendix C.6 details label-shifting, causal-mask, and generation-cache checks. The interface rejects incomplete, misordered, or nonfinite submissions; Appendix A gives preprocessing, training, and decoding settings.

5 Results and Analysis

Unless stated otherwise, results follow Section 4.3. Appendix B reports the full analysis grid, including settings that change the direction or uncertainty of a difference.

Table 1: Sequence-track Top-1 (%); Δ: real minus control (pp). Participant and crossed participant–text 95% intervals apply to Δ.

Scheme	Input	Real	Control	Δ	Participant 95%	Crossed 95%
BT	Zero	11.2811	11.2286	$+0.0525$	$[+0.0177, +0.0869]$	$[-0.0200, +0.1420]$
BT	Donor	11.3907	11.4011	-0.0104	$[-0.0390, +0.0195]$	$[-0.0763, +0.0613]$
CSCL	Zero	11.6081	11.6063	$+0.0017$	$[-0.0260, +0.0276]$	$[-0.0527, +0.0542]$
CSCL	Donor	11.7370	11.7268	$+0.0102$	$[-0.0295, +0.0536]$	$[-0.0534, +0.0762]$
CE-only	Zero	11.7089	11.7143	-0.0054	$[-0.0483, +0.0365]$	$[-0.0828, +0.0731]$
CE-only	Donor	11.8450	11.8525	-0.0075	$[-0.0560, +0.0438]$	$[-0.0863, +0.0730]$

5.1 Sequence Decoder Comparisons

Figure 3 summarizes comparisons; Table 1 gives Pool-S/MEAN results. Zero/donor comparisons use 22,609/21,847 observations, with real scores recomputed on each support. All three schemes show small paired gains. CE-only has the highest real-input Top-1, **11.71%**, but a real-minus-zero gain of -0.005 **pp**; higher predictive scores therefore need not indicate larger gains from paired input.

These small macro gains do not hide large opposing unit effects: CSCL and CE-only real-minus-zero gains stay within -0.074 to $+0.057$ pp across units. BT has a wider range; Figure B3 and Appendix B.9 provide all unit-level results.

Comparisons with the independent language-model reference also distinguish prediction performance from paired input gains. CSCL, for example, achieves **11.61%** Top-1, exceeding the LM-only reference of **10.16%**, although its point estimate is below the **12.5%** uniform reference. With EEG values zeroed, the same trained CSCL model still achieves **11.61%**. Thus, its advantage over LM-only does not primarily reflect gains from current EEG values; the LM-only comparison also changes task adaptation history and input structure.

Interval estimates also depend on including text variation. Both interval types include zero for CSCL and CE-only under zero and donor controls. BT has a positive real-minus-zero gain with a participant interval above zero, but its crossed participant–text interval spans both signs (Table 1). For CSCL real-input accuracy, the participant interval is [11.51%, 11.71%] and the crossed interval is [10.34%, 12.89%]; the latter includes 12.5%. These intervals constrain population-level conclusions for new texts; intervals containing zero do not establish equivalence between real and replacement inputs. Under the same Pool-S/MEAN setting, the real-minus-zero gains of CSCL and CE-only differ by only 0.007 pp, with both interval types including zero for this difference (Appendix B.7).

Continuous scores show the same distinction: under NATIVE/Pool-S/SUM, CSCL exceeds CE-only by 0.343 target bits (crossed interval [0.062, 0.616]), but paired zeroing gains differ by only 0.0027 bits ([-0.0148, 0.0209]; Appendix B.4). CSCL's extra training budget precludes attributing this contrast solely to contrastive learning.

Small accuracy differences do not imply that input leaves candidate scores unchanged. Switching from zero to real EEG changes the complete ranking in 3.77% and 10.96% of candidate queries for CSCL and CE-only, respectively. Far fewer queries change from incorrect to correct, nearly balanced by opposite changes (Appendix B.5). Input replacement can thus elicit responses without corresponding accuracy gains.

5.2 Sensitivity to Candidate Construction and Scoring

Changing candidates and score normalization changes identification results. For BT, switching from SUM to MEAN raises real-input accuracy from 9.73% to 11.28% in Pool-S but lowers it from 13.02% to 12.58% in Pool-L. Normalization thus has no consistent effect across candidate pools. Small input gains also change direction across these settings; the full grid shows this variation while retaining the main text's primary reporting endpoint (Appendices B.1–B.4).

The auxiliary frozen representation analysis shows candidate effects under another information condition. Its covariate reference alone scores 15.17% in Pool-S and 22.39% in Pool-L. On the common

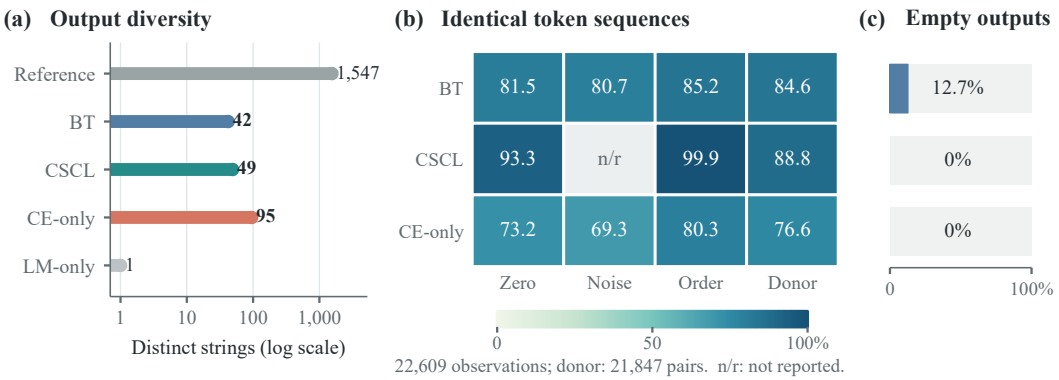


Figure 4: **Generation diversity and stability.** (a) Distinct strings (log axis). (b) Exact token-sequence agreement with real-input outputs. (c) Empty outputs. Non-donor/donor supports: 22,609/21,847 observations; CSCL/noise is unreported. Generation has no candidate or reference prefix.

support, adding an EEG score from band features, CBraMod, or LaBraM changes Top-1 by at most 0.061 pp in absolute value. These descriptive results concern the implemented frozen adapters and covariate reference; Appendices A–B provide full tables and the unresolved channel-correspondence limitation.

5.3 Text Generation under Input Replacement

Without candidates or reference prefixes, generation is concentrated (Figure 4; Table F.3). Across 22,609 observations, BT, CSCL, and CE-only produce 42, 49, and 95 distinct strings; BT is empty on 2,878 observations. These counts measure diversity, not correctly recovered sentences.

CSCL retains identical token sequences on 99.86% of word-order and 88.77% of donor replacements; donors use 21,847 pairs. CE-only changes more often, yet has no larger identification gains. Figure 4 and Appendix B.6 give all schemes' rates. Output agreement measures behavior, not semantic accuracy.

5.4 Learning and Generalization Diagnostics

Fitting training texts does not improve validation performance (Figure 5a). CSCL/CE-only each complete 30 cross-entropy epochs in 25 cells but select epoch 1. Median training NLL falls to about 0.24, while final validation NLL rises to 6.62/6.60 (Appendix C.5).

Artificial eight-class signals pass through the actual normalization and training pipeline, reaching 100% test accuracy at higher strengths versus 12.5% under zeroing. Yet micro-fitting 64 real observations gives 77.6%/78.0% real/zero accuracy, failing its accuracy and paired-gain targets. A post hoc text-membership reference reaches 74.6%, contextualizing training performance without identifying the model's mechanism (Appendices C.1–C.2).

Relaxing participant or text holdout in two CE-only/Pool-S/MEAN cells gives gains of -0.125 to $+0.750$ pp; four observations produce the largest gain. Each cell uses one seed and shares 200 test observations, 3,000 training observations, and the training budget across settings. Target-bits gains are slightly negative in all six settings. Text exposure lowers validation NLL without strong input differentiation; different validation pools prevent attributing NLL differences to EEG (Appendix C.3; Table C.2). Together, these checks separate learnability and exposure from paired EEG gains; sensitivity to weak real signals remains uncalibrated.

6 Discussion and Limitations

These results show that EEG-to-text comparisons should report both model performance and same-checkpoint input controls. An independent language model provides a reference for text prediction,

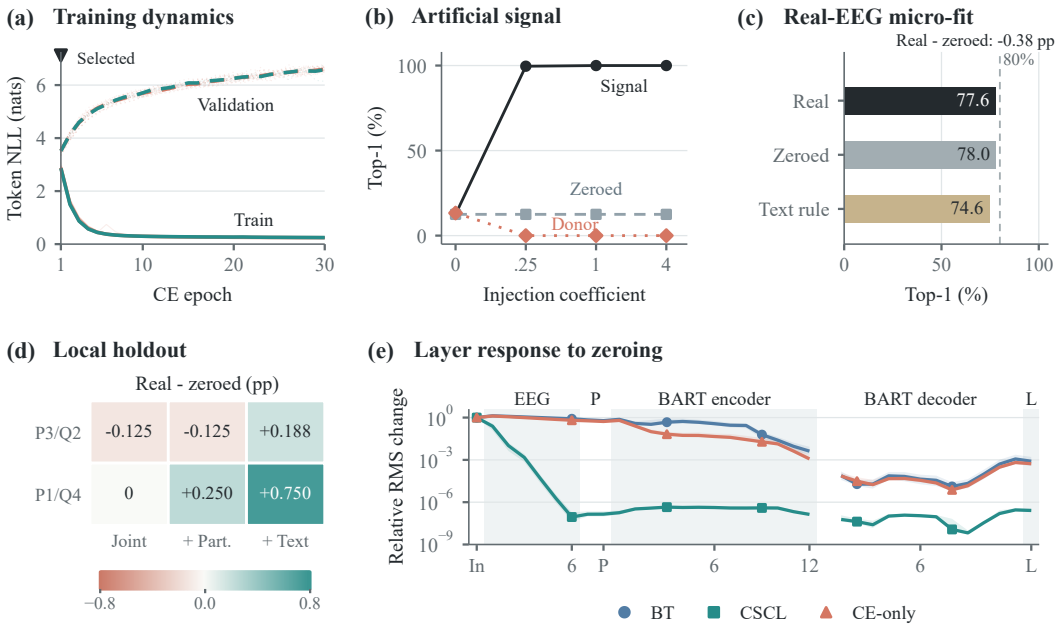


Figure 5: **Learning and generalization diagnostics.** (a) Training/validation NLL (solid/dashed). (b) Artificial-signal learnability. (c) Real-EEG micro-fitting versus zeroed input and a text-membership reference. (d) Local holdout gains. (e) Median layer responses to zeroing, with 10th–90th percentiles. Injection coefficients are not EEG SNR; activation responses are not information fractions. Details: Appendix C.

whereas input replacement tests dependence on the current recording. Reporting both avoids attributing all task-adaptation gains to EEG content. Zero and donor controls are complementary because they preserve different input aspects.

Interpreting these comparisons also requires specifying the task and generalization target. Candidate identification supplies text options, whereas generation must produce outputs without them. Likewise, participant exposure, text exposure, and joint holdout define different transfer settings. These tasks and settings should not be interchanged without explanation.

Our conclusions concern natural-reading EEG under the evaluated protocols. Sequence schemes share one backbone, and local holdout checks cover only two cells. Our generation analysis currently measures output diversity and agreement after input replacement; validated semantic accuracy metrics and human evaluation remain lacking. Input controls retain length, masks, and fixation-based alignment, while zeroing and noise also change the input distribution. Donor replacement uses real recordings but cannot fully separate linguistic contributions from other signal differences. Artificial-signal learning provides a task-level positive control, but does not establish calibrated sensitivity to weak information in real EEG.

7 Conclusion

NEAR-Bench combines explicit generalization conditions, shared candidates, and controlled EEG-to-text evaluation. In the main sequence track, predictive advantages over an independent language-model reference need not accompany larger paired input gains: prediction scores and generated outputs can remain largely unchanged after EEG replacement. These results support reporting task performance with real–control differences for a more complete assessment of EEG-to-text progress. Shared evaluation materials and supplementary tools support reproducible comparisons across candidate pools, scoring rules, and generalization settings. Future work should extend these comparisons to additional decoders and complement output stability with validated semantic evaluation, while calibrating sensitivity to weak EEG signals.

AI USE STATEMENT

Generative AI was used to assist with translating the manuscript from Chinese into English.

ETHICS STATEMENT AND POTENTIAL IMPACT

This study conducts a secondary analysis of publicly available ZuCo 1.0 and 2.0 natural-reading EEG data, without new participant recruitment, data collection, or human interventions. By standardizing evaluation conditions and using paired input controls, NEAR-Bench helps distinguish text prediction performance from gains due to correctly paired EEG input. Brain-signal analysis and text generation raise risks concerning individual privacy and misattribution; generated text should not be treated as a direct record of participants' actual intentions. Resource distribution will comply with the licensing and attribution requirements of the data, models, and code used.

REPRODUCIBILITY STATEMENT

Data construction, joint participant–text holdout, candidate construction, and input controls are described in Sections 3–4 and Appendix A. Model training and generation configurations are provided in Appendices A.3–A.4, and checkpoint selection in Appendix C.5. Statistical aggregation and uncertainty settings appear in Appendix A.5. The anonymous supplementary package provides fixed splits and candidate metadata, scoring and resampling code, and entry points for data preparation, training, and evaluation for the existing methods. Its CPU demo checks the protocol without pretrained models and recomputes the stored BP832 candidate scores. Full-data runs require users to obtain ZuCo data and pretrained weights from their original sources. Commands, dependencies, and coverage are documented in the package's `README.md` and `docs/REPRODUCTION.md`. Appendices B–C provide extended results and implementation checks; Appendix D describes the evaluation interface, and Appendix E indexes the numerical files.

REFERENCES

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, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1f2fd23309a5b2d2537d063b29ec1b52-Abstract-Conference.html.

Xiachong Feng, Xiaocheng Feng, Bing Qin, and Ting Liu. Aligning semantic in brain and language: A curriculum contrastive method for electroencephalography-to-text generation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:3874–3883, 2023. doi: 10.1109/TNSRE.2023.3314642. URL https://doi.org/10.1109/TNSRE.2023.3314642. Abstract.

Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Scientific Data*, 5:180291, 2018. doi: 10.1038/sdata.2018.291. URL https://doi.org/10.1038/sdata.2018.291.

Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. ZuCo 2.0: A dataset of physiological recordings during natural reading and annotation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 138–146, 2020. URL https://aclanthology.org/2020.lrec-1.18/.

Vinay Jayaram and Alexandre Barachant. MOABB: trustworthy algorithm benchmarking for BCIs. *Journal of Neural Engineering*, 15(6):066011, 2018. doi: 10.1088/1741-2552/aadea0. URL https://doi.org/10.1088/1741-2552/aadea0. Author preprint.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/47393e8594c82ce8fd83adc672cf9872-Abstract-Conference.html.

Hyejeong Jo, Yiqian Yang, Juhyeok Han, Yiqun Duan, Hui Xiong, and Won Hee Lee. Are EEG-to-text models working? *arXiv preprint arXiv:2405.06459v4*, 2024. URL https://arxiv.org/abs/2405.06459. Full text.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, 2020. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703/.

Xiaozhao Liu, Dinggang Shen, and Xihui Liu. Learning interpretable representations leads to semantically faithful EEG-to-text generation. *arXiv preprint arXiv:2505.17099v1*, 2025. URL https://arxiv.org/abs/2505.17099. Full text.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBraMod: A criss-cross brain foundation model for EEG decoding. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/bbbd6d915cb90be21c1254a82d45cedd-Abstract-Conference.html.

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:2603.03312v3*, 2026. URL https://arxiv.org/abs/2603.03312. Full text.

Zhenhailong Wang and Heng Ji. Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(5):5350–5358, 2022. doi: 10.1609/aaai.v36i5.20472. URL https://ojs.aaai.org/index.php/AAAI/article/view/20472.

Wei Xiong, Jiangtong Li, Jie Li, Kun Zhu, and Changjun Jiang. EEG-FM-Bench: A comprehensive benchmark for the systematic evaluation and diagnostic analyses of EEG foundation models. *arXiv preprint arXiv:2508.17742v3*, June 2026. URL https://arxiv.org/abs/2508.17742v3. 15 June 2026. Full text.

Xiaoli Yang, Huiyuan Tian, Yurui Li, Jianyu Zhang, Shijian Li, and Gang Pan. Brain-CLIPLM: semantic compression for EEG-to-text decoding. *Frontiers in Neuroscience*, 20:1899770, 2026. doi: 10.3389/fnins.2026.1899770. URL https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2026.1899770/full.

Zihan Zhang, Yu Bao, Xiao Ding, Tianyi Jiang, and Kai Xiong. Is EEG-to-text feasible in real-world scenarios? an in-depth analysis using a neuropsychology-inspired benchmark. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1378–1393, 2026. doi: 10.18653/v1/2026.acl-long.61. URL https://aclanthology.org/2026.acl-long.61/.

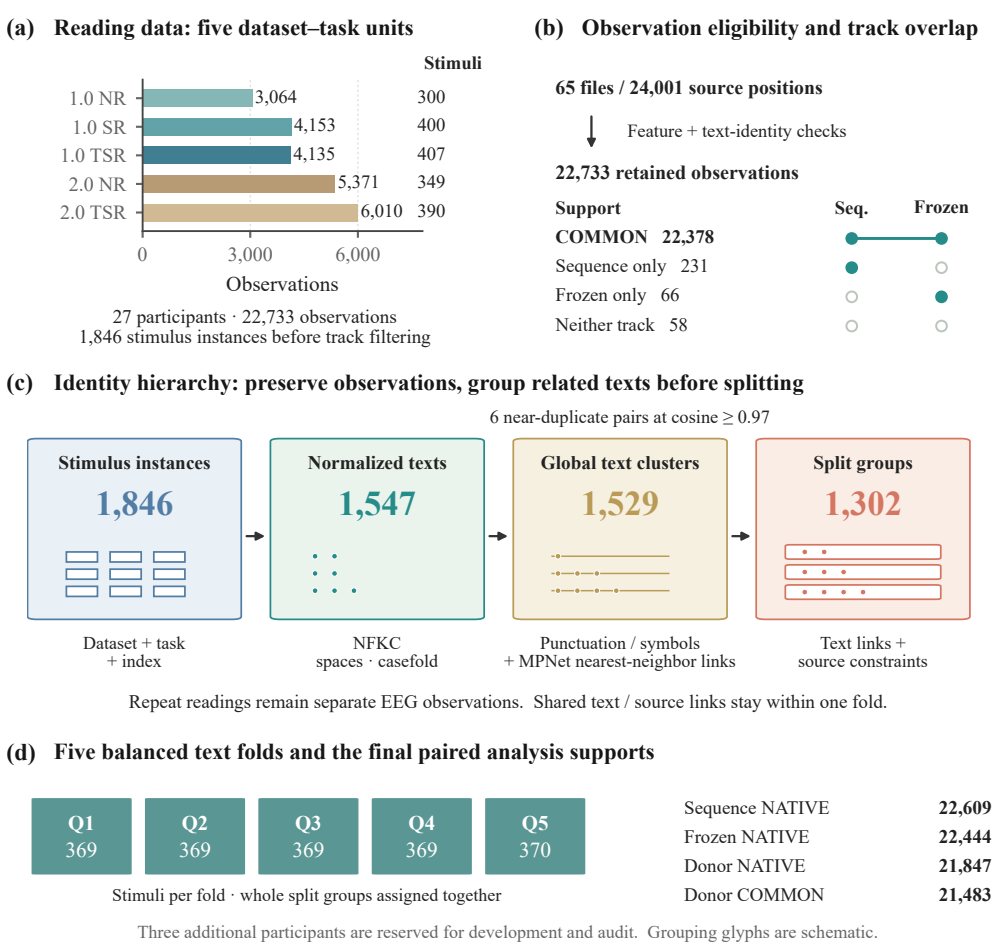


Figure A1: **Data construction and leakage-controlled text folds.** (a) Observations and stimuli by dataset–task unit. (b) Exclusive track memberships: COMMON 22,378, sequence-only 231, frozen-only 66, and neither 58. These are set differences, not individual exclusion reasons. (c) Stimulus identities retain dataset, task, and sentence index. Normalization, punctuation/symbol handling, and fixed MPNet nearest-neighbor checks form global text clusters; six nonidentical pairs meet cosine similarity ≥ 0.97. Source links impose additional split constraints. (d) The 1,302 resulting groups stay intact across five folds. Repeated readings remain separate observations. Group glyphs are schematic; grouping is not exhaustive semantic duplicate annotation.

A Implementation and Reproducibility

Figure A1 summarizes observation eligibility, text identities, split groups, and the resulting analysis supports.

A.1 Data Processing

Data sources and participant scope. The benchmark was assembled from 65 source files containing 24,001 source record positions. Feature-validity and text-identity checks retained 22,733 observations. The main analysis includes 27 participants; ZDM, YRH, and YAC were reserved for development and auditing and excluded from the main cohort.

Reference texts and record verification. Reference sentences come from a predefined stimulus list, with no new manual annotation. We checked each source record against this list, requiring finite sentence-level band features and a consistent mapping between the EEG record and stimulus sentence. Observations with nonempty source text that conflicted with the list were excluded, as

were observations affected by systematic stimulus-identity conflicts. Records with missing source text were retained only when the stored features were valid and the fixed stimulus index established the correspondence. These checks ensure consistent text identities and valid inputs.

Word-level EEG input. Sequence models use GD840, which directly reads the word-level band features stored in ZuCo for gaze-duration windows; power spectra are not reestimated from continuous signals. Each word has 105 coordinates in each of eight bands, concatenated into an 840-dimensional vector. The bands follow the order below.

Table A.1: Frequency bands.

Band	Original frequency range (Hz)
t1	4–6
t2	6.5–8
a1	8.5–10
a2	10.5–13
b1	13.5–18
b2	18.5–30
g1	30.5–40
g2	40–49.5

Each word vector is standardized across its 840 elements using the sample standard deviation with denominator 839. Moments are computed in FP64, and output features are stored in FP32. An entire observation is excluded if its input has an invalid shape, contains nonfinite values, has a standard deviation no greater than 10^{-12}, contains no valid words, or contains more than 56 valid words. Records exceeding the length limit are not truncated. Inputs follow the order of fixated words in the source text.

Sentence-level features and frozen-encoder inputs. BP832 concatenates the first 104 coordinates of each of the eight sentence-level band summaries. Frozen EEG encoders use recoverable sentence-level signals from these 104 source coordinates. Preprocessing comprises a 50 Hz notch filter with a quality factor of 30, a fourth-order bandpass filter, and polyphase resampling from 500 Hz to 200 Hz. The passband is 0.3–75 Hz for CBraMod and 0.1–75 Hz for LaBraM. After resampling, the first 800 samples are retained, and shorter signals are padded on the right. Standardization statistics are estimated only from the training folds.

Channel adaptation and representation dimensionality. To provide inputs to the frozen encoders, stored columns are first mapped to E1–E104 of GSN-HydroCel-128. For each target position, the nearest source position is then selected using MNE coordinates; a source position may be selected more than once. CBraMod uses 19 specified target channels. LaBraM uses channels among the first 128 entries in its official list for which standard-1020 coordinates are available. The correspondence between stored columns and anatomical locations has not been independently validated. Each encoder's pooled 200-dimensional output is mapped to 832 dimensions using a projection with orthonormal rows obtained by QR decomposition of a fixed Gaussian matrix.

Valid observation sets. The two tracks have different input requirements and therefore different numbers of eligible observations. Their respective valid sets are denoted NATIVE, comprising 22,444 observations for the frozen-representation track and 22,609 for the sequence track. Their intersection, COMMON, contains 22,378 observations. Each comparison is performed within its specified set, using identical observations on all sides of the comparison.

A.2 Splits and Candidates

Figure A2 illustrates candidate matching and compares the resulting length and language-model difficulty gaps.

Text identity and near-duplicate checks. To prevent identical or identified near-duplicate texts from appearing in different text folds, we first identify relationships between texts and then determine which stimuli must be assigned together. A stimulus instance is identified jointly by dataset,

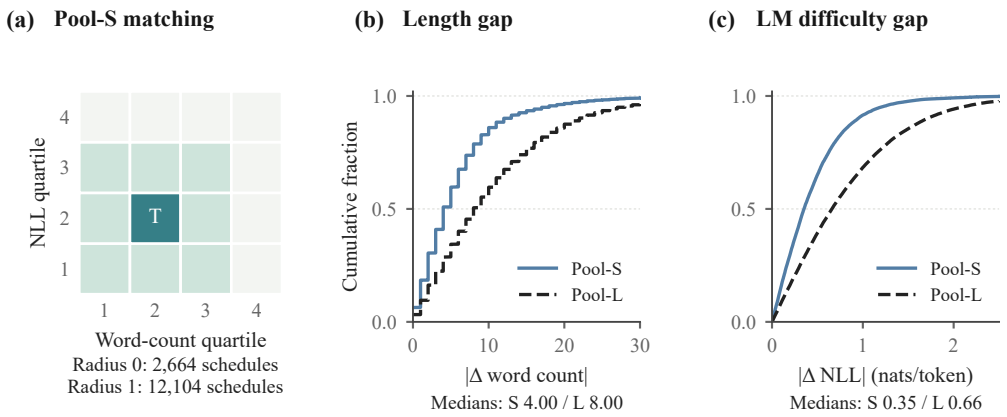


Figure A2: **Candidate matching and empirical gap distributions.** (a) Schematic quartile matching searches the target bin first and expands the Chebyshev neighborhood only when needed. (b,c) Exact empirical cumulative distributions from candidate schedules and text metadata, with 103,376 target–distractor gaps per pool. Display windows emphasize the distribution bulk; medians use all gaps. Word counts are recovered with expm1 from stored log1p counts, and GPT-2 mean NLL is the negated first text-feature column. The matching grid is not an empirical density map.

task, and sentence index. Thus, identical texts appearing in different datasets or tasks remain distinct stimulus instances.

Texts first undergo Unicode NFKC normalization, whitespace normalization, and case folding to define fine-grained text identities. Punctuation and symbols are then replaced with spaces, and whitespace is collapsed again to identify texts differing only in these surface forms. For near-duplicate checks, normalized MPNet sentence embeddings are used to retrieve the nearest other text for each distinct text. Among pairs that remain distinct after punctuation processing, six meet the cosine-similarity threshold of 0.97. These near-duplicate links are merged with normalized duplicate relations to form coarse-grained text clusters. This check uses a fixed nearest-neighbor rule; it does not exhaustively annotate semantic similarity across all sentence pairs.

Source constraints and split groups. Text clusters encode duplicate or near-duplicate relationships. Split groups additionally incorporate available source constraints to determine which stimuli must be assigned as a whole. For ZuCo 2.0, stimuli within a task sharing the source identifier `source_id1` are assigned to the same split group; this identifier comes from a fixed mapping of task materials. ZuCo 1.0 lacks usable source metadata, so only text-relation constraints are applied. Connected components formed jointly by text relations and source constraints define the final split groups. A split group may contain multiple text clusters.

The statistical units at each level are listed below. Text clusters retaining dataset–task identity are counted separately within each unit and therefore differ in number from global clusters merged across units.

Table A.2: Text identities and split groups.

Level	Count	Statistical meaning
Stimulus instances	1,846	Stimuli identified by dataset, task, and sentence index
Fine-grained text identities	1,547	Distinct texts after NFKC normalization, whitespace normalization, and case folding
Global coarse-grained text clusters	1,529	Clusters merging normalized duplicates and identified near-duplicate relations
Text clusters retaining dataset–task identity	1,802	Each global cluster counted separately within each dataset–task unit
Split groups	1,302	Sets of stimuli assigned together after additionally merging source constraints

Text-fold assignment. A greedy algorithm with random seed 20260809 assigns split groups to five text folds while balancing the size of each dataset–task unit across folds. Each group is assigned to one fold in its entirety and is never divided across folds. The five folds contain 369, 369, 369, 369, and 370 stimulus instances, respectively.

Crossing five text folds with five participant groups yields 25 outer test cells. Across these cells, each canonical observation is assigned to testing once, validation once, and training nine times; it falls in unused cross-regions in the remaining 14 cells. Section 4.1 specifies the training, validation, and test composition of each cell.

Shared candidate configuration. Each stimulus has eight fixed candidate sets per pool, each containing the reference sentence and seven distractors. Each pool therefore contains 14,768 candidate sets. Candidate membership and order are shared across participants and systems, and the target occupies each of the eight positions once. Candidates come from the same dataset–task unit and text fold. The seven distractors belong to mutually distinct text clusters and exclude the target cluster.

Two candidate pools. Within each dataset–task unit and text fold, Pool-S separately bins word count and GPT-2 mean negative log-likelihood (NLL) into quartiles, merging duplicate bin boundaries. Distractors are first selected from the target's two matching bins; the neighborhood is expanded only if fewer than seven eligible distractor clusters are available. Neighborhood distance is the maximum of the differences in the two bin indices, i.e., Chebyshev distance. In total, 2,664 candidate sets use radius zero and 12,104 use radius one. Pool-L retains the unit, text-fold, and text-cluster exclusion constraints but samples from all eligible sentences without matching word count or NLL.

Both pools use candidate random seed 2026080911. Sampling first randomly permutes eligible stimulus rows and then retains the first stimulus encountered from each text cluster until seven distractors have been selected. This ensures that distractors come from distinct clusters but does not sample text clusters uniformly.

Real-EEG donor selection. A donor is another real recording used to replace the current EEG input. Donors are determined by combining random seed 2026090939 with row identity. A donor must come from the same outer cell and dataset–task unit, have the same valid word length, and belong to a different coarse-grained text cluster. Recordings from the same participant are preferred when eligible. Evaluation retains the original observation's target text, candidate membership, and candidate order.

Donor analyses retain only observations for which an eligible replacement can be found. On COMMON, both the original observation and donor must belong to COMMON. Consequently, donor analyses have smaller eligible observation sets than non-donor sequence controls, as shown below.

Table A.3: Sequence-control support sets.

Sequence-control setting	NATIVE observations	COMMON observations
Donor replacement	21,847	21,483
Non-donor sequence controls	22,609	22,378

A.3 Models and Training

Evaluation-track overview. We define sequence-training and representation-adaptation tracks for EEG-conditioned generation and frozen representation adaptation, respectively (Table F.2). The three sequence schemes share the same word-level EEG encoder, projection layer, and BART-large decoder architecture. BT uses SGD for cross-entropy training, and CE-only uses Adam; CSCL adds curriculum contrastive training inspired by Feng et al. (2023) to Adam-based cross-entropy training. CSCL and CE-only each have a budget of 30 cross-entropy epochs, with checkpoints selected by validation token NLL. CSCL adds 15 contrastive epochs, giving different total training budgets. All three are local implementations. The training and adaptation settings below expand this overview; Appendix C.5 reports actual checkpoint selection.

Sequence inputs retain fixated words, word order, and explicit masks; reference texts remain complete sentences. The frozen track uses a shared text-embedding adapter to measure the increment over the covariate reference from EEG scores. Its input adaptation and richer non-EEG information define this conditional auxiliary analysis. Both tracks share splits, candidate identities, and reporting rules. NATIVE denotes observations available to each track; COMMON is their intersection. Both sides of each comparison use the same observations. The frozen representation track compares the full model with a covariate reference; the sequence track replaces input at the same checkpoint. Absolute scores are interpreted within their respective scoring and reference conditions.

Figure A3 summarizes frozen-track adaptation and its covariate reference.

The frozen track uses a ridge-regression adapter to predict normalized 768-dimensional all-mpnet-base-v2 text embeddings from EEG representations standardized using training-set statistics. Ridge regularization is selected from $\{0.1, 1, 10, 100, 1000\}$ by validation embedding mean squared error. Let $\hat{\mathbf{e}}_i$ denote the predicted embedding and $\mathbf{e}_c$ the embedding of candidate text c. The EEG score and final candidate score are

$$\begin{aligned} s_{\mathrm{EEG}}(i, c) &= \cos(\hat{\mathbf{e}}_i, \mathbf{e}_c), \\ s_{\mathrm{full}}(i, c) &= s_0(i, c) + \gamma s_{\mathrm{EEG}}(i, c). \end{aligned} \tag{3}$$

The shared reference s_0 receives no EEG representations. It combines candidate NLL, length, character, and word-frequency features with observation covariates, including reading duration, fixation count, pupil summaries, stimulus index, dataset/task/block indicators, and missingness indicators. These variables are part of this track's input conditions. With eight-dimensional candidate-text features $\boldsymbol{\phi}(c)$ and observation covariates $\mathbf{z}_i$, the reference score is

$$s_0(i, c) = \frac{\boldsymbol{\phi}(c)^\top (\mathbf{b} + W \mathbf{z}_i)}{T_0}. \tag{4}$$

Imputation and standardization parameters for continuous covariates are fitted only on training data; candidate-feature standardization uses deduplicated texts in the training candidate sets. Parameters $\mathbf{b}$ and W are trained on the training set using eight-way cross-entropy and AdamW with a learning rate of 0.03. Weight decay is selected from $\{10^{-4}, 10^{-3}, 10^{-2}\}$ by validation cross-entropy. Each configuration is optimized for at most 200 epochs, stopping after 20 consecutive epochs without improvement and retaining the parameters with the lowest validation loss. A positive temperature T_0 is then fitted on the validation set. With the reference and ridge adapter fixed, $\gamma = \mathrm{softplus}(a) + 10^{-4}$ is fitted using validation candidate cross-entropy. Reference fitting uses training and validation observations from the canonical row set; EEG alignment and γ fitting use the subset satisfying

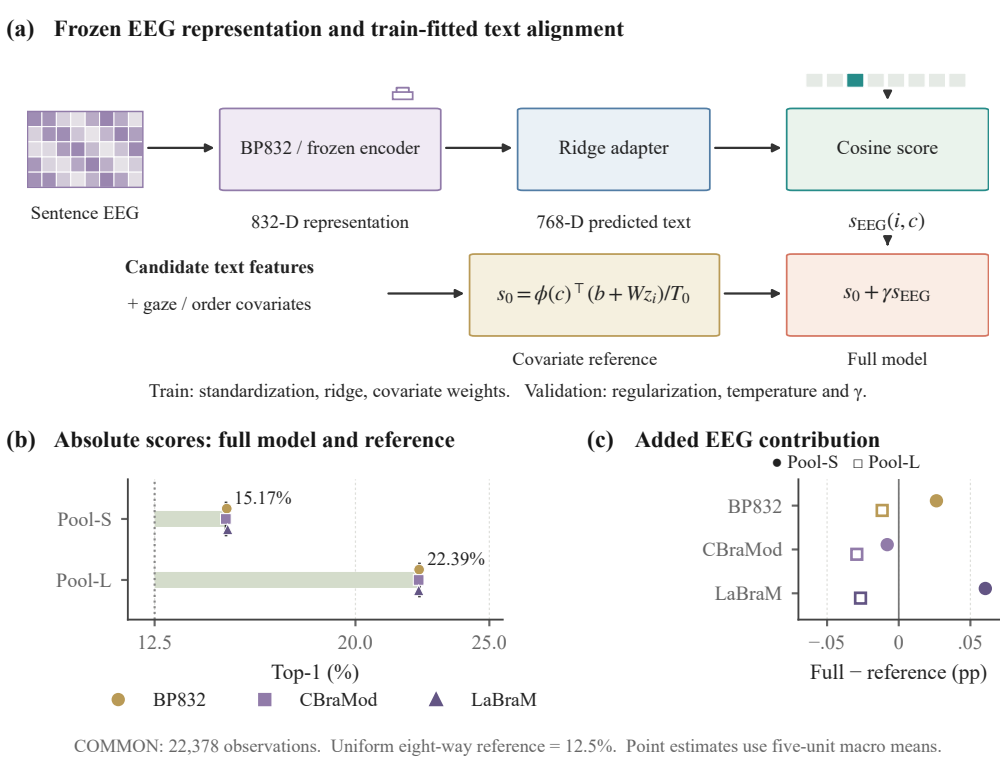


Figure A3: **Frozen-track adaptation and EEG increments.** (a) Training-standardized EEG features are mapped by ridge regression to 768-dimensional MPNet embeddings, giving a cosine score. A separately fitted reference uses candidate language features and observation covariates, including gaze and stimulus order; validation calibrates temperature and EEG weight. (b,c) COMMON five-unit macro Top-1 and full-minus-reference increments. The reference reaches 15.17% in Pool-S and 22.39% in Pool-L. These point estimates imply neither uncertainty bounds nor a cross-track ranking. This covariate reference differs from sequence-track LM-only; tensor glyphs are schematic.

the frozen-representation requirements. Candidate cross-entropy fitting uses a fixed schedule of 16 eight-candidate sets, whereas reported results use the eight-set schedule in Appendix A.2. All parameters remain fixed at test time, and s_{full} and s_0 are compared on identical candidate sets and support sets.

All three sequence schemes use a six-layer EEG Transformer with eight attention heads and 2,048-dimensional feedforward layers, followed by an 840-to-1,024-dimensional projection and ReLU, before the BART-large encoder–decoder (Lewis et al., 2020). BT uses SGD with momentum 0.9; CSCL and CE-only use Adam during cross-entropy training. The cross-entropy learning rate is 2×10^{-5}, the effective batch size is 32, and training runs for at most 30 epochs. CSCL and CE-only use zero weight decay and a gradient-norm clipping threshold of one. Checkpoints are selected by validation token-weighted NLL.

CSCL adds 15 epochs of contrastive training in three five-epoch curriculum stages. Samples come only from the current training set. Within the same dataset–task unit, positives come from different participants in the same text cluster; negatives come from different text clusters and different participants. Curriculum ordering uses cosine similarities between GD840 input vectors averaged over valid word positions. Positives are ordered by decreasing similarity and negatives by increasing similarity. Each ordering is divided into three stages of approximately equal size, used sequentially from easy to hard. Ties are broken by observation identifier. If fewer than three samples are available on either side, that side uses its full pool in all three stages. In each epoch, one positive and one negative are sampled from the current stage for each anchor. Anchors lacking a valid positive or negative are skipped during contrastive training.

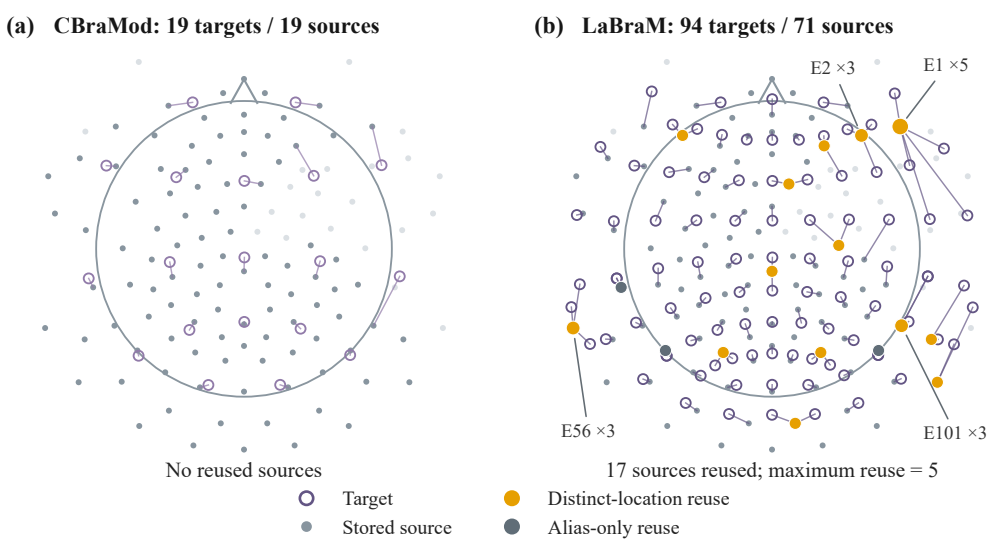


Figure A4: **Channel mapping and source reuse.** CBraMod maps 19 targets to 19 sources; LaBraM maps 94 targets to 71 sources. Seventeen LaBraM sources are reused, with up to five targets sharing E1. Coordinates use MNE 1.13.2 GSN-HydroCel-128 and standard_1020 montages in an azimuthal angular projection. Gray reuse marks aliases at the same target coordinate; orange marks reuse across distinct target coordinates. This is an input-adaptation diagram, not an EEG activation topography. Anatomical correspondence of the stored columns has not been independently verified.

The contrastive representation $\mathbf{h}_i \in \mathbb{R}^{840}$ is the mean output of the six-layer EEG encoder over valid word positions, taken before the 840-to-1,024-dimensional projection. For B anchors in a logical batch, their sampled positives and negatives form $2B$ comparison positions. For each anchor, valid positive and negative sets $\mathcal{P}_i$ and $\mathcal{N}_i$ are recomputed among these positions according to the same-unit, different-participant, and text-cluster conditions. The loss is

$$\mathcal{L}_{\mathrm{CL}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\sum_{j\in\mathcal{P}_i} \exp(\cos(\mathbf{h}_i, \mathbf{h}_j)/\tau)}{\sum_{j\in\mathcal{P}_i\cup\mathcal{N}_i} \exp(\cos(\mathbf{h}_i, \mathbf{h}_j)/\tau)}, \qquad \tau = 10^{-5}. \tag{5}$$

Positive terms are summed before taking the logarithm. Repeatedly sampled observations are retained at their respective batch positions, and positions failing the eligibility conditions are excluded from the denominator. The implementation computes the loss over the full logical batch using log-sum-exp. Contrastive training uses Adam with a learning rate of 10^{-3}. FP32 contrastive updates affect only the EEG encoder; subsequent BF16 cross-entropy updates affect the entire model.

Figure A4 documents the electrode mapping used by the frozen encoders; preprocessing and correspondence limitations are specified in Appendix A.1.

A.4 Runtime Settings

Candidate evaluation uses BF16 forward logits, FP32 token log-probabilities, and FP64 accumulation. Scoring includes tokenizer-added special tokens in the cache and excludes label-padding values of -100; MEAN normalizes by the number of tokens actually scored. Top-1 assigns fractional credit to exact ties for the highest score. Generation receives neither candidate texts nor reference prefixes.

Table A.4: Generation settings.

Generation parameter	Value used
do_sample / num_beams	false / 1
max_new_tokens / use_cache	128 / true
no_repeat_ngram_size	3
min_length / min_new_tokens	0 / null
repetition_penalty	1.0
decoder_start_token_id	2
forced_bos_token_id / forced_eos_token_id	0 / 2
bos_token_id / eos_token_id / pad_token_id	0 / 2 / 1

A.5 Analysis Settings

The main text focuses on Top-1 in the sequence track under NATIVE, Pool-S, and MEAN, retaining valid observations for this track, matching candidate length and language-model difficulty, and using mean token likelihood. Zeroing tests the contribution of EEG values beyond length and masks; donor comparisons test the benefit of the original pairing relative to an equally long real recording. Each comparison uses its own paired support set. Pool-L, SUM, COMMON, noise, word-order permutation, and target bits provide sensitivity analyses. Frozen representations are reported separately on COMMON as Full minus the covariate reference, serving as a conditional auxiliary analysis.

Zeroing sets feature values to zero while preserving length and masks; noise replacement uses standard Gaussian features of the same length; word-order permutation reorders valid word vectors. The independent LM-only reference is BART without EEG-task fine-tuning, with a source containing only one EOS token. Its training history and input format differ from those of the fitted sequence decoders, so it is interpreted separately from same-checkpoint controls.

For candidate resample r, sequence scores and target bits are defined as

$$\begin{aligned} s_{irc}^{\mathrm{SUM}} &= \sum_{t=1}^{L_c} \log p_\theta(y_{c,t} \mid y_{c,<t}, x_i), \\ s_{irc}^{\mathrm{MEAN}} &= s_{irc}^{\mathrm{SUM}} / L_c, \\ b_{ir} &= \log_2 \frac{\exp(s_{ir,c_{ir}^*})}{\sum_{c=1}^{8} \exp(s_{irc})}. \end{aligned} \tag{6}$$

Here, c_{ir}^* is the target position. Target bits measure the target's relative score within the candidate set; they are neither EEG–text mutual information nor calibrated sentence probabilities. Uniform scores correspond to -3 bits. For eligible paired support $\mathcal{S}$, aggregation is defined as

$$M(m; \mathcal{S}) = \frac{1}{5} \sum_{u=1}^{5} \frac{1}{|\mathcal{S}_u|} \sum_{i \in \mathcal{S}_u} \frac{1}{8} \sum_{r=1}^{8} m_{ir}. \tag{7}$$

Participant resampling is the primary uncertainty analysis; text and crossed resampling are sensitivity analyses. The main text presents participant and crossed intervals side by side to make their differing conditioning explicit. Participant weights are shared across tasks within a dataset, and global text-cluster weights are shared across all observations of the same text. Crossed resampling uses the product of independently drawn participant and text weights. Each scheme uses 5,000 resamples, retaining pairing and equal unit weights. Multiway-cluster coverage has not been calibrated for the crossed scheme. All intervals are reported pointwise without multiplicity adjustment; uncertainty from training, model selection, and donor selection is not resampled. An interval containing zero does not establish equivalence.

A.6 Implementation and Computational Resources

Model resources and implementations. BT, CSCL, and CE-only initialize their text encoder–decoder with BART-large; the independent LM-only reference uses the same pretrained model. The BT and CSCL implementations respectively draw on NeuSpeech's corrected version of BrainTranslator and a third-party CSCL implementation. Frozen representations use CBraMod (pretrained weights) and LaBraM-Base. Text representations and candidate difficulty are computed using all-mpnet-base-v2 and GPT-2, respectively. Appendices A.1–A.4 describe the benchmark's data processing, model adaptation, training, and generation settings.

Computing environment. BT and CE-only training and the additional CSCL seed experiments use Python 3.11, PyTorch 2.4.1 (CUDA 12.1), and Transformers 4.44.2. Training and representation extraction used two NVIDIA GeForce RTX 4090 GPUs.

Code and supplementary resources. The implementation, fixed protocol metadata, and electronic supplementary materials S1–S12 accompany the manuscript in the anonymous archive `NEAR_Bench_Supplementary.zip`. After extraction, the common entry point is `NEAR_Supplementary/README.md`, which describes how to inspect results, run the benchmark, and integrate new models. Appendix E maps files to the paper's results.

B Extended Results

Figure B1 gives an overview of paired input gains across the complete content-replacement analysis grid.

B.1 Sequence Scoring

The following table reports both scoring rules, both candidate pools, and NATIVE/COMMON support sets. Real and zero scores are accuracy percentages; Δ_{Top1} is the real-minus-zero difference in percentage points (pp), and Δ_b is the corresponding target-bits difference. Equal-weight aggregation over five units is performed separately for each support set. This table reports point estimates. Intervals for CE-only real–control differences appear in Appendix B.8 and Appendix E; intervals for CSCL-minus-CE-only gain differences appear in Appendix B.4. These concern different statistical quantities.



Figure B1: **Paired input gains across 96 settings.** Three schemes, four controls, two supports, two scoring rules, and two pools are ordered by the Top-1 point estimate. The lower matrix identifies each setting; dark marks indicate NATIVE/Pool-S/MEAN. The largest absolute estimate is 0.054 pp. Participant intervals include zero in 84/96 settings and crossed intervals in 96/96; text-cluster intervals also include zero in all settings (not drawn). These are pointwise conditional intervals, not simultaneous inference or equivalence evidence. Donor comparisons use their eligible paired subset.

Table B.1: Sequence scores.

Support	Scoring	Pool	Obs.	Real (%)	Zero (%)	Δ Top-1 (pp)	Δ bits
BT							
COMMON	MEAN	L	22378	12.516784	12.523889	−0.007105	−0.000028
COMMON	MEAN	S	22378	11.208309	11.153862	+0.054446	−0.000032
COMMON	SUM	L	22378	12.813640	12.810387	+0.003253	+0.004300
COMMON	SUM	S	22378	9.599401	9.573485	+0.025916	+0.002728
NATIVE	MEAN	L	22609	12.577823	12.587250	−0.009427	−0.000031
NATIVE	MEAN	S	22609	11.281089	11.228622	+0.052467	−0.000038
NATIVE	SUM	L	22609	13.015433	13.012142	+0.003291	+0.004269
NATIVE	SUM	S	22609	9.728205	9.698659	+0.029546	+0.002702
CE-only							
COMMON	MEAN	L	22378	12.260985	12.257421	+0.003563	+0.000098
COMMON	MEAN	S	22378	11.653106	11.657587	−0.004481	+0.000034
COMMON	SUM	L	22378	12.767898	12.752348	+0.015550	+0.001290
COMMON	SUM	S	22378	10.010174	10.014306	−0.004132	−0.000801
NATIVE	MEAN	L	22609	12.324588	12.318982	+0.005606	+0.000110
NATIVE	MEAN	S	22609	11.708939	11.714329	−0.005390	+0.000049
NATIVE	SUM	L	22609	12.955984	12.940554	+0.015430	+0.001830
NATIVE	SUM	S	22609	10.124455	10.129267	−0.004811	−0.000329
CSCL							
COMMON	MEAN	L	22378	12.102257	12.119049	−0.016792	+0.000035
COMMON	MEAN	S	22378	11.537683	11.535863	+0.001819	+0.000032
COMMON	SUM	L	22378	12.768352	12.770954	−0.002602	+0.001305
COMMON	SUM	S	22378	10.003000	10.007300	−0.004299	+0.002374
NATIVE	MEAN	L	22609	12.172422	12.190581	−0.018158	+0.000035
NATIVE	MEAN	S	22609	11.608064	11.606338	+0.001727	+0.000031
NATIVE	SUM	L	22609	12.947431	12.950448	−0.003017	+0.001299
NATIVE	SUM	S	22609	10.114791	10.119061	−0.004270	+0.002338

B.2 Frozen Representations

Full-model and reference scores are accuracy percentages, and Δ_{Top1} is their difference in pp. The full model combines the fitted EEG adapter score with the covariate reference. The following table gives descriptive point estimates on both support sets; Top-1 intervals were not computed for the frozen track.

Table B.2: Frozen-representation scores.

Support	Pool	Obs.	Full (%)	Reference (%)	Δ Top-1 (pp)	Δ bits
BP832						
NATIVE	S	22444	15.198259	15.169374	+0.028886	−0.000113
NATIVE	L	22444	22.358437	22.370023	−0.011586	−0.000424
COMMON	S	22378	15.193544	15.167237	+0.026307	−0.000112
COMMON	L	22378	22.377396	22.388958	−0.011562	−0.000424
CBraMod-frozen						
NATIVE	S	22444	15.164507	15.169374	−0.004867	−0.000067
NATIVE	L	22444	22.341270	22.370023	−0.028753	−0.000178
COMMON	S	22378	15.159191	15.167237	−0.008045	−0.000075
COMMON	L	22378	22.359695	22.388958	−0.029263	−0.000186
LaBraM-frozen						
NATIVE	S	22444	15.231133	15.169374	+0.061759	−0.000020
NATIVE	L	22444	22.343583	22.370023	−0.026440	−0.000128
COMMON	S	22378	15.227649	15.167237	+0.060412	−0.000013
COMMON	L	22378	22.362281	22.388958	−0.026677	−0.000122

B.3 Input Controls

The following tables enumerate all 64 input-content replacement comparisons for BT/CSCL: two schemes, four replacements, two support sets, two scoring rules, and two candidate pools. The participant, text, and crossed columns give 95% percentile intervals under the three resampling schemes, each using 5,000 resamples and conditioning on the fitted model, candidate schedule, and donor assignments. N and C denote NATIVE and COMMON, respectively.

Intervals are pointwise, without multiplicity adjustment or equivalence testing; crossed intervals do not guarantee multiway coverage. Candidate resamples are averaged within each observation before resampling, and candidate queries are not treated as independent observations. Participant and text weights are shared across their corresponding repeated observations. If a resample leaves a unit empty, that resample is unavailable; it is not reweighted over fewer units. Fixed donors are not independently resampled. All Top-1 and bits changes are real minus control.

Table B.3: BT input replacements: Top-1.

Support	Scoring	Pool	Obs.	Δ Top-1 (pp)	Participant 95%	Text 95%	Crossed 95%
Donor							
COMMON	MEAN	L	21483	−0.011748	[−0.027008, +0.003370]	[−0.031408, +0.007059]	[−0.043447, +0.018468]
COMMON	MEAN	S	21483	−0.009863	[−0.038821, +0.020865]	[−0.053247, +0.032945]	[−0.076257, +0.062560]
COMMON	SUM	L	21483	+0.014699	[−0.006892, +0.036349]	[−0.001529, +0.032200]	[−0.015212, +0.047428]
COMMON	SUM	S	21483	+0.000866	[−0.017273, +0.019524]	[−0.017295, +0.018076]	[−0.033413, +0.034998]
NATIVE	MEAN	L	21847	−0.011289	[−0.026831, +0.003898]	[−0.030504, +0.007253]	[−0.042652, +0.019159]
NATIVE	MEAN	S	21847	−0.010407	[−0.038954, +0.019513]	[−0.053288, +0.032283]	[−0.076302, +0.061277]
NATIVE	SUM	L	21847	+0.012914	[−0.008555, +0.034743]	[−0.002707, +0.030172]	[−0.016472, +0.045729]
NATIVE	SUM	S	21847	−0.002528	[−0.020567, +0.015993]	[−0.019291, +0.013539]	[−0.036491, +0.030976]
Noise							
COMMON	MEAN	L	22378	+0.001003	[−0.016213, +0.016573]	[−0.022913, +0.025246]	[−0.036165, +0.037779]
COMMON	MEAN	S	22378	+0.017396	[−0.013331, +0.052417]	[−0.026099, +0.059045]	[−0.044563, +0.088058]
COMMON	SUM	L	22378	+0.017550	[+0.002435, +0.034170]	[−0.001009, +0.038154]	[−0.012223, +0.049943]
COMMON	SUM	S	22378	+0.008591	[−0.013917, +0.032418]	[−0.012944, +0.030759]	[−0.029922, +0.050469]
NATIVE	MEAN	L	22609	−0.001348	[−0.017918, +0.013534]	[−0.025166, +0.022860]	[−0.038001, +0.034673]
NATIVE	MEAN	S	22609	+0.014548	[−0.016039, +0.049946]	[−0.028795, +0.056064]	[−0.046883, +0.083773]
NATIVE	SUM	L	22609	+0.017624	[+0.002478, +0.034185]	[−0.000687, +0.037846]	[−0.011761, +0.049974]
NATIVE	SUM	S	22609	+0.008414	[−0.014112, +0.031431]	[−0.012296, +0.030018]	[−0.029949, +0.050245]
Word-order permutation							
COMMON	MEAN	L	22378	+0.002869	[−0.011364, +0.016432]	[−0.016591, +0.022069]	[−0.027238, +0.032022]
COMMON	MEAN	S	22378	+0.018932	[−0.024645, +0.070642]	[−0.033564, +0.076144]	[−0.055434, +0.111304]
COMMON	SUM	L	22378	+0.004844	[−0.008516, +0.019404]	[−0.010566, +0.022382]	[−0.020726, +0.032839]
COMMON	SUM	S	22378	−0.000114	[−0.015395, +0.014949]	[−0.018432, +0.017046]	[−0.033521, +0.028127]
NATIVE	MEAN	L	22609	+0.002741	[−0.011372, +0.016135]	[−0.016551, +0.021719]	[−0.027044, +0.031821]
NATIVE	MEAN	S	22609	+0.019384	[−0.024296, +0.071148]	[−0.032958, +0.076093]	[−0.054694, +0.111314]
NATIVE	SUM	L	22609	+0.005247	[−0.008137, +0.019787]	[−0.009900, +0.022942]	[−0.020064, +0.033439]
NATIVE	SUM	S	22609	−0.000204	[−0.015267, +0.014314]	[−0.018504, +0.016540]	[−0.033428, +0.027643]

Continued on next page

Table B.3: BT input replacements: Top-1. (continued)

Support	Scoring	Pool	Obs.	Δ Top-1 (pp)	Participant 95%	Text 95%	Crossed 95%
Zero							
COMMON	MEAN	L	22378	−0.007105	[−0.025779, +0.011547]	[−0.032102, +0.018664]	[−0.046607, +0.031173]
COMMON	MEAN	S	22378	+0.054446	[+0.019917, +0.088293]	[−0.000112, +0.116273]	[−0.018043, +0.143893]
COMMON	SUM	L	22378	+0.003253	[−0.011616, +0.017343]	[−0.010332, +0.016492]	[−0.021620, +0.027977]
COMMON	SUM	S	22378	+0.025916	[+0.004438, +0.048131]	[−0.001546, +0.056272]	[−0.016167, +0.076404]
NATIVE	MEAN	L	22609	−0.009427	[−0.029934, +0.010410]	[−0.034752, +0.016228]	[−0.050523, +0.029612]
NATIVE	MEAN	S	22609	+0.052467	[+0.017705, +0.086898]	[−0.001646, +0.114313]	[−0.019968, +0.141975]
NATIVE	SUM	L	22609	+0.003291	[−0.011469, +0.017243]	[−0.010368, +0.016638]	[−0.021773, +0.028035]
NATIVE	SUM	S	22609	+0.029546	[+0.008592, +0.051794]	[−0.000969, +0.065679]	[−0.015225, +0.082890]

Table B.4: BT input replacements: target bits.

Support	Scoring	Pool	Obs.	Δ bits	Bits crossed 95%
Donor					
COMMON	MEAN	L	21483	+0.000019	[−0.000142, +0.000182]
COMMON	MEAN	S	21483	+0.000020	[−0.000142, +0.000183]
COMMON	SUM	L	21483	+0.000176	[−0.004550, +0.004952]
COMMON	SUM	S	21483	+0.000818	[−0.004247, +0.005929]
NATIVE	MEAN	L	21847	+0.000021	[−0.000138, +0.000182]
NATIVE	MEAN	S	21847	+0.000023	[−0.000138, +0.000186]
NATIVE	SUM	L	21847	+0.000228	[−0.004438, +0.004956]
NATIVE	SUM	S	21847	+0.000917	[−0.004129, +0.005981]
Noise					
COMMON	MEAN	L	22378	+0.000277	[−0.000134, +0.000891]
COMMON	MEAN	S	22378	+0.000252	[−0.000136, +0.000835]
COMMON	SUM	L	22378	+0.029807	[−0.004550, +0.117264]
COMMON	SUM	S	22378	+0.027654	[−0.003447, +0.103813]
NATIVE	MEAN	L	22609	+0.000270	[−0.000136, +0.000880]
NATIVE	MEAN	S	22609	+0.000246	[−0.000138, +0.000818]
NATIVE	SUM	L	22609	+0.029589	[−0.004440, +0.116153]
NATIVE	SUM	S	22609	+0.027608	[−0.003406, +0.102936]
Word-order permutation					
COMMON	MEAN	L	22378	+0.000037	[−0.000123, +0.000248]
COMMON	MEAN	S	22378	+0.000038	[−0.000128, +0.000254]
COMMON	SUM	L	22378	−0.000053	[−0.004932, +0.005778]
COMMON	SUM	S	22378	−0.000103	[−0.005165, +0.005989]
NATIVE	MEAN	L	22609	+0.000034	[−0.000124, +0.000246]
NATIVE	MEAN	S	22609	+0.000036	[−0.000131, +0.000253]
NATIVE	SUM	L	22609	−0.000088	[−0.004946, +0.005737]
NATIVE	SUM	S	22609	−0.000105	[−0.005096, +0.005967]

Continued on next page

Table B.4: BT input replacements: target bits. (continued)

Support	Scoring	Pool	Obs.	Δ bits	Bits crossed 95%
Zero					
COMMON	MEAN	L	22378	−0.000028	[−0.000231, +0.000161]
COMMON	MEAN	S	22378	−0.000032	[−0.000229, +0.000145]
COMMON	SUM	L	22378	+0.004300	[−0.002266, +0.011118]
COMMON	SUM	S	22378	+0.002728	[−0.003178, +0.008484]
NATIVE	MEAN	L	22609	−0.000031	[−0.000235, +0.000157]
NATIVE	MEAN	S	22609	−0.000038	[−0.000236, +0.000138]
NATIVE	SUM	L	22609	+0.004269	[−0.002216, +0.011036]
NATIVE	SUM	S	22609	+0.002702	[−0.003188, +0.008390]

Table B.5: CSCL input replacements: Top-1.

Support	Scoring	Pool	Obs.	Δ Top-1 (pp)	Participant 95%	Text 95%	Crossed 95%
Donor							
COMMON	MEAN	L	21483	−0.018511	[−0.044341, +0.007035]	[−0.040863, +0.002756]	[−0.058759, +0.019687]
COMMON	MEAN	S	21483	+0.008968	[−0.031770, +0.053112]	[−0.025404, +0.044606]	[−0.054696, +0.075707]
COMMON	SUM	L	21483	−0.002801	[−0.012046, +0.006165]	[−0.016310, +0.010476]	[−0.022715, +0.016929]
COMMON	SUM	S	21483	+0.002052	[−0.011587, +0.015114]	[−0.015581, +0.018276]	[−0.026876, +0.027116]
NATIVE	MEAN	L	21847	−0.018587	[−0.044383, +0.007080]	[−0.040425, +0.002485]	[−0.059344, +0.019132]
NATIVE	MEAN	S	21847	+0.010214	[−0.029499, +0.053645]	[−0.024318, +0.045851]	[−0.053440, +0.076186]
NATIVE	SUM	L	21847	−0.002838	[−0.011810, +0.006115]	[−0.016153, +0.010192]	[−0.022493, +0.016723]
NATIVE	SUM	S	21847	+0.001197	[−0.012593, +0.014425]	[−0.016443, +0.017409]	[−0.027214, +0.025999]
Noise							
COMMON	MEAN	L	22378	−0.003078	[−0.015654, +0.008769]	[−0.022469, +0.017447]	[−0.032286, +0.028374]
COMMON	MEAN	S	22378	+0.029184	[+0.003254, +0.054739]	[−0.006739, +0.068799]	[−0.025797, +0.088543]
COMMON	SUM	L	22378	+0.000279	[−0.011980, +0.011793]	[−0.011055, +0.011495]	[−0.019278, +0.020000]
COMMON	SUM	S	22378	+0.010430	[−0.002010, +0.023976]	[−0.003593, +0.023722]	[−0.011346, +0.034616]
NATIVE	MEAN	L	22609	−0.004089	[−0.016500, +0.007606]	[−0.023250, +0.016216]	[−0.033034, +0.026794]
NATIVE	MEAN	S	22609	+0.028801	[+0.002996, +0.054246]	[−0.006818, +0.068309]	[−0.025996, +0.088001]
NATIVE	SUM	L	22609	−0.000207	[−0.012372, +0.011243]	[−0.011358, +0.010704]	[−0.019991, +0.019238]
NATIVE	SUM	S	22609	+0.010344	[−0.001936, +0.023759]	[−0.003518, +0.023497]	[−0.011191, +0.034324]

Continued on next page

Table B.5: CSCL input replacements: Top-1. (continued)

Support	Scoring	Pool	Obs.	Δ Top-1 (pp)	Participant 95%	Text 95%	Crossed 95%
Word-order permutation							
COMMON	MEAN	L	22378	+0.000130	[−0.001374, +0.001784]	[−0.001391, +0.001818]	[−0.002459, +0.003149]
COMMON	MEAN	S	22378	+0.000473	[−0.000966, +0.001935]	[−0.001002, +0.002045]	[−0.002218, +0.003480]
COMMON	SUM	L	22378	+0.000427	[+0.000000, +0.001317]	[+0.000000, +0.001354]	[+0.000000, +0.002493]
COMMON	SUM	S	22378	+0.000000	[+0.000000, +0.000000]	[+0.000000, +0.000000]	[+0.000000, +0.000000]
NATIVE	MEAN	L	22609	+0.000132	[−0.001368, +0.001783]	[−0.001386, +0.001818]	[−0.002456, +0.003143]
NATIVE	MEAN	S	22609	+0.000471	[−0.000961, +0.001923]	[−0.000998, +0.002035]	[−0.002210, +0.003468]
NATIVE	SUM	L	22609	+0.000417	[+0.000000, +0.001290]	[+0.000000, +0.001320]	[+0.000000, +0.002433]
NATIVE	SUM	S	22609	+0.000000	[+0.000000, +0.000000]	[+0.000000, +0.000000]	[+0.000000, +0.000000]
Zero							
COMMON	MEAN	L	22378	−0.016792	[−0.031534, −0.002486]	[−0.037946, +0.002021]	[−0.049207, +0.011994]
COMMON	MEAN	S	22378	+0.001819	[−0.025979, +0.027602]	[−0.034393, +0.037482]	[−0.052978, +0.054449]
COMMON	SUM	L	22378	−0.002602	[−0.012965, +0.007861]	[−0.012806, +0.007391]	[−0.020573, +0.014231]
COMMON	SUM	S	22378	−0.004299	[−0.013052, +0.004213]	[−0.014939, +0.006230]	[−0.021372, +0.013002]
NATIVE	MEAN	L	22609	−0.018158	[−0.033707, −0.003287]	[−0.039380, +0.000962]	[−0.051207, +0.010903]
NATIVE	MEAN	S	22609	+0.001727	[−0.026044, +0.027581]	[−0.034491, +0.037164]	[−0.052707, +0.054176]
NATIVE	SUM	L	22609	−0.003017	[−0.013238, +0.007277]	[−0.013177, +0.006882]	[−0.020805, +0.013954]
NATIVE	SUM	S	22609	−0.004270	[−0.013000, +0.004164]	[−0.014807, +0.006162]	[−0.021235, +0.012829]

Table B.6: CSCL input replacements: target bits.

Support	Scoring	Pool	Obs.	Δ bits	Bits crossed 95%
Donor					
COMMON	MEAN	L	21483	−0.000024	[−0.000192, +0.000144]
COMMON	MEAN	S	21483	−0.000024	[−0.000194, +0.000145]
COMMON	SUM	L	21483	+0.002097	[−0.002618, +0.006816]
COMMON	SUM	S	21483	+0.001602	[−0.003506, +0.006713]
NATIVE	MEAN	L	21847	−0.000025	[−0.000196, +0.000145]
NATIVE	MEAN	S	21847	−0.000028	[−0.000201, +0.000142]
NATIVE	SUM	L	21847	+0.002063	[−0.002646, +0.006758]
NATIVE	SUM	S	21847	+0.001453	[−0.003603, +0.006466]

Continued on next page

Table B.6: CSCL input replacements: target bits. (continued)

Support	Scoring	Pool	Obs.	Δ bits	Bits crossed 95%
Noise					
COMMON	MEAN	L	22378	−0.000033	[−0.000197, +0.000126]
COMMON	MEAN	S	22378	−0.000027	[−0.000197, +0.000135]
COMMON	SUM	L	22378	+0.001761	[−0.003040, +0.006765]
COMMON	SUM	S	22378	+0.001418	[−0.004069, +0.006837]
NATIVE	MEAN	L	22609	−0.000032	[−0.000194, +0.000125]
NATIVE	MEAN	S	22609	−0.000026	[−0.000194, +0.000134]
NATIVE	SUM	L	22609	+0.001734	[−0.003021, +0.006722]
NATIVE	SUM	S	22609	+0.001417	[−0.003988, +0.006783]
Word-order permutation					
COMMON	MEAN	L	22378	−0.000003	[−0.000020, +0.000014]
COMMON	MEAN	S	22378	−0.000004	[−0.000021, +0.000013]
COMMON	SUM	L	22378	−0.000039	[−0.000749, +0.000605]
COMMON	SUM	S	22378	−0.000212	[−0.001029, +0.000502]
NATIVE	MEAN	L	22609	−0.000003	[−0.000020, +0.000014]
NATIVE	MEAN	S	22609	−0.000004	[−0.000021, +0.000013]
NATIVE	SUM	L	22609	−0.000039	[−0.000745, +0.000602]
NATIVE	SUM	S	22609	−0.000211	[−0.001027, +0.000499]
Zero					
COMMON	MEAN	L	22378	+0.000035	[−0.000093, +0.000164]
COMMON	MEAN	S	22378	+0.000032	[−0.000095, +0.000159]
COMMON	SUM	L	22378	+0.001305	[−0.002705, +0.005496]
COMMON	SUM	S	22378	+0.002374	[−0.001957, +0.006754]
NATIVE	MEAN	L	22609	+0.000035	[−0.000091, +0.000164]
NATIVE	MEAN	S	22609	+0.000031	[−0.000096, +0.000157]
NATIVE	SUM	L	22609	+0.001299	[−0.002669, +0.005412]
NATIVE	SUM	S	22609	+0.002338	[−0.001934, +0.006605]

The additional 16 LM-only comparisons are reported separately. LM-only changes both the model and its input format, so these comparisons measure the advantage over an independent reference.

Table B.7: BT versus LM-only: Top-1.

Support	Scoring	Pool	Obs.	Δ Top-1 (pp)	Participant 95%	Text 95%	Crossed 95%
LM-only							
COMMON	MEAN	L	22378	−0.229515	[−0.364413, −0.095257]	[−1.436728, +0.912950]	[−1.456446, +0.927422]
COMMON	MEAN	S	22378	+1.105555	[+0.936766, +1.266253]	[−0.130493, +2.362536]	[−0.140914, +2.380181]
COMMON	SUM	L	22378	+0.068298	[+0.003955, +0.125654]	[−0.463200, +0.629419]	[−0.470493, +0.649581]
COMMON	SUM	S	22378	+0.836315	[+0.724285, +0.954615]	[+0.109319, +1.611594]	[+0.078361, +1.614106]
NATIVE	MEAN	L	22609	−0.285671	[−0.411809, −0.165177]	[−1.491600, +0.849506]	[−1.506242, +0.861867]
NATIVE	MEAN	S	22609	+1.120529	[+0.947115, +1.284081]	[−0.122012, +2.378256]	[−0.125049, +2.403010]
NATIVE	SUM	L	22609	+0.065865	[+0.008565, +0.117771]	[−0.465132, +0.626951]	[−0.469893, +0.646749]
NATIVE	SUM	S	22609	+0.829593	[+0.718684, +0.947089]	[+0.098791, +1.610108]	[+0.072337, +1.613298]

Table B.8: BT versus LM-only: target bits.

Support	Scoring	Pool	Obs.	Δ bits	Bits crossed 95%
LM-only					
COMMON	MEAN	L	22378	+0.599182	[+0.508405, +0.692144]
COMMON	MEAN	S	22378	+0.416162	[+0.349642, +0.483470]
COMMON	SUM	L	22378	+67.546357	[+58.659319, +76.408530]
COMMON	SUM	S	22378	+51.089854	[+44.220599, +58.003665]
NATIVE	MEAN	L	22609	+0.594354	[+0.505497, +0.687422]
NATIVE	MEAN	S	22609	+0.414152	[+0.347258, +0.481293]
NATIVE	SUM	L	22609	+67.247384	[+58.407121, +76.123383]
NATIVE	SUM	S	22609	+50.884036	[+44.037551, +57.772520]

Table B.9: CSCL versus LM-only: Top-1.

Support	Scoring	Pool	Obs.	Δ Top-1 (pp)	Participant 95%	Text 95%	Crossed 95%
LM-only							
COMMON	MEAN	L	22378	−0.644042	[−0.774667, −0.503166]	[−1.932148, +0.649459]	[−1.929346, +0.676517]
COMMON	MEAN	S	22378	+1.434929	[+1.332418, +1.549620]	[+0.029811, +2.796375]	[+0.016129, +2.790907]
COMMON	SUM	L	22378	+0.023010	[−0.027999, +0.072691]	[−0.718164, +0.737712]	[−0.712497, +0.746461]
COMMON	SUM	S	22378	+1.239915	[+1.144848, +1.348917]	[+0.261120, +2.224569]	[+0.261355, +2.225610]
NATIVE	MEAN	L	22609	−0.691072	[−0.802506, −0.570533]	[−1.985199, +0.608200]	[−1.980389, +0.619205]
NATIVE	MEAN	S	22609	+1.447504	[+1.343761, +1.559116]	[+0.033079, +2.824530]	[+0.014278, +2.816388]
NATIVE	SUM	L	22609	−0.002137	[−0.048795, +0.040807]	[−0.754526, +0.720257]	[−0.743025, +0.725545]
NATIVE	SUM	S	22609	+1.216179	[+1.125734, +1.323220]	[+0.235545, +2.203236]	[+0.226208, +2.209678]

Table B.10: CSCL versus LM-only: target bits.

Support	Scoring	Pool	Obs.	Δ bits	Bits crossed 95%
LM-only					
COMMON	MEAN	L	22378	+0.486409	[+0.395128, +0.578204]
COMMON	MEAN	S	22378	+0.327775	[+0.258115, +0.398392]
COMMON	SUM	L	22378	+74.697153	[+68.411229, +81.303649]
COMMON	SUM	S	22378	+54.621536	[+49.289641, +60.230074]
NATIVE	MEAN	L	22609	+0.481923	[+0.389368, +0.573872]
NATIVE	MEAN	S	22609	+0.325751	[+0.255981, +0.396197]
NATIVE	SUM	L	22609	+74.430052	[+68.147821, +80.953387]
NATIVE	SUM	S	22609	+54.450548	[+49.146946, +60.057954]

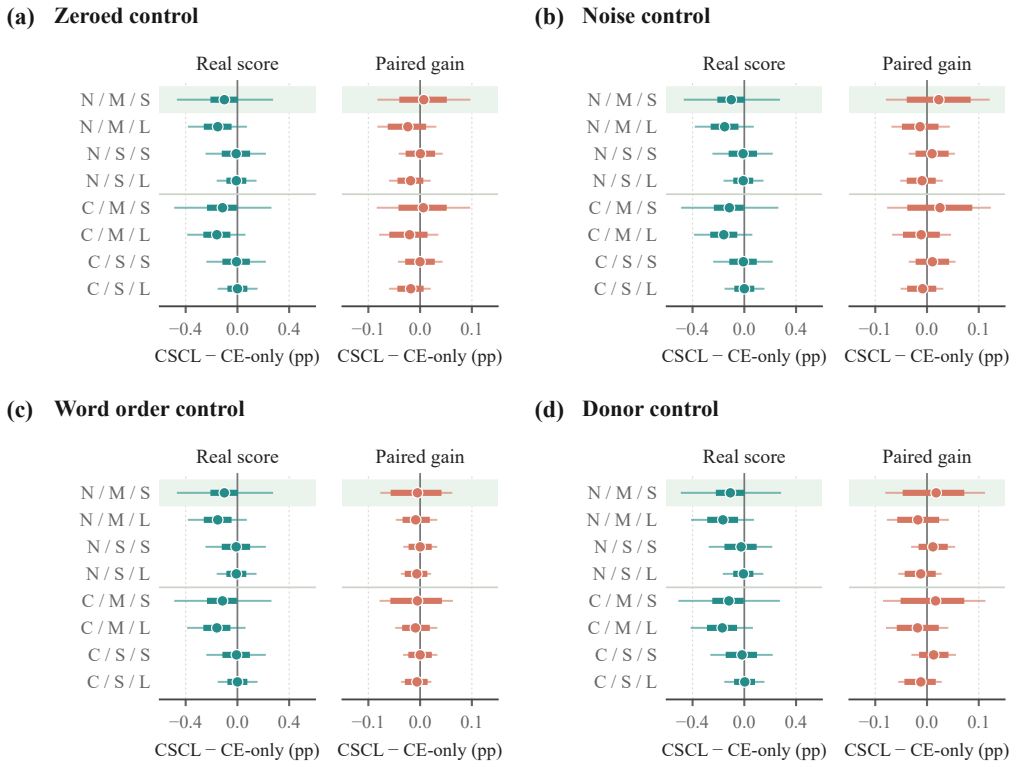


Figure B2: **Training-scheme contrasts by input control.** Each control includes eight support/scoring/pool settings. Left columns show CSCL-minus-CE-only real-input Top-1; right columns show the difference in paired input gains. The columns use different horizontal scales, and the primary setting is shaded. Thick/thin bars are participant/crossed 95% conditional intervals. Crossed paired-gain intervals include zero in all 32 settings. CSCL adds 15 contrastive epochs, so the contrast does not isolate a compute-matched training effect.

B.4 Training Schemes

Figure B2 separates between-scheme differences in real-input scores from differences in paired input gains.

Define $\Delta\Delta = (\text{Real} - \text{Control})_{\text{CSCL}} - (\text{Real} - \text{Control})_{\text{CE}}$. The 32 input-content replacement comparisons use matched observations, and the crossed intervals for both metrics all include zero. CSCL receives additional contrastive training, so the two schemes have different total optimization budgets.

Table B.11: CSCL versus CE-only input gains.

Support	Scoring	Pool	Obs.	ΔΔ Top-1 (pp)	Crossed 95%	ΔΔ bits	Crossed 95%
Donor							
COMMON	MEAN	L	21483	−0.018276	[−0.077261, +0.039343]	−0.000009	[−0.000343, +0.000301]
COMMON	MEAN	S	21483	+0.016551	[−0.083200, +0.111017]	−0.000021	[−0.000354, +0.000283]
COMMON	SUM	L	21483	−0.011933	[−0.053723, +0.026519]	+0.002314	[−0.007322, +0.011712]
COMMON	SUM	S	21483	+0.012845	[−0.028222, +0.054055]	+0.002386	[−0.007123, +0.011764]
NATIVE	MEAN	L	21847	−0.017642	[−0.075709, +0.040304]	−0.000009	[−0.000350, +0.000302]
NATIVE	MEAN	S	21847	+0.017664	[−0.078617, +0.110056]	−0.000026	[−0.000359, +0.000282]
NATIVE	SUM	L	21847	−0.012128	[−0.053567, +0.026425]	+0.002193	[−0.007437, +0.011540]
NATIVE	SUM	S	21847	+0.011433	[−0.028802, +0.052326]	+0.002001	[−0.007456, +0.011160]
Noise							
COMMON	MEAN	L	22378	−0.011285	[−0.065409, +0.044563]	−0.000042	[−0.000532, +0.000467]
COMMON	MEAN	S	22378	+0.025070	[−0.075254, +0.121478]	−0.000001	[−0.000447, +0.000501]
COMMON	SUM	L	22378	−0.008562	[−0.049108, +0.029743]	−0.031750	[−0.075365, +0.005245]
COMMON	SUM	S	22378	+0.010415	[−0.033400, +0.053246]	−0.019018	[−0.049429, +0.007421]
NATIVE	MEAN	L	22609	−0.013395	[−0.066879, +0.042122]	−0.000054	[−0.000535, +0.000443]
NATIVE	MEAN	S	22609	+0.022773	[−0.077408, +0.119139]	−0.000015	[−0.000458, +0.000484]
NATIVE	SUM	L	22609	−0.009533	[−0.049922, +0.028747]	−0.032067	[−0.075454, +0.004615]
NATIVE	SUM	S	22609	+0.009770	[−0.033500, +0.052047]	−0.019514	[−0.049435, +0.006674]
Word-order permutation							
COMMON	MEAN	L	22378	−0.009010	[−0.046191, +0.031231]	+0.000055	[−0.000148, +0.000265]
COMMON	MEAN	S	22378	−0.005121	[−0.076284, +0.061539]	+0.000039	[−0.000170, +0.000254]
COMMON	SUM	L	22378	−0.005918	[−0.035097, +0.019782]	+0.001179	[−0.004861, +0.007192]
COMMON	SUM	S	22378	+0.000129	[−0.031077, +0.031612]	+0.000512	[−0.006012, +0.007055]
NATIVE	MEAN	L	22609	−0.008601	[−0.045088, +0.031224]	+0.000044	[−0.000155, +0.000253]
NATIVE	MEAN	S	22609	−0.005163	[−0.075002, +0.060215]	+0.000027	[−0.000178, +0.000240]
NATIVE	SUM	L	22609	−0.006407	[−0.035208, +0.019140]	+0.000921	[−0.005034, +0.006888]
NATIVE	SUM	S	22609	+0.000104	[−0.030724, +0.031243]	+0.000211	[−0.006227, +0.006565]

Continued on next page

Table B.11: CSCL versus CE-only input gains. (continued)

Support	Scoring	Pool	Obs.	ΔΔ Top-1 (pp)	Crossed 95%	ΔΔ bits	Crossed 95%
Zero							
COMMON	MEAN	L	22378	−0.020355	[−0.077423, +0.033465]	−0.000063	[−0.000500, +0.000483]
COMMON	MEAN	S	22378	+0.006300	[−0.081930, +0.094875]	−0.000002	[−0.000428, +0.000566]
COMMON	SUM	L	22378	−0.018152	[−0.057988, +0.018736]	+0.000015	[−0.024031, +0.025037]
COMMON	SUM	S	22378	−0.000167	[−0.040826, +0.041537]	+0.003175	[−0.014381, +0.021694]
NATIVE	MEAN	L	22609	−0.023765	[−0.080928, +0.029388]	−0.000074	[−0.000507, +0.000457]
NATIVE	MEAN	S	22609	+0.007116	[−0.080576, +0.095174]	−0.000018	[−0.000437, +0.000535]
NATIVE	SUM	L	22609	−0.018447	[−0.058142, +0.018473]	−0.000531	[−0.024354, +0.024128]
NATIVE	SUM	S	22609	+0.000541	[−0.039362, +0.041883]	+0.002668	[−0.014784, +0.020946]

In a supplementary continuous-score comparison under Pool-S/SUM, the CSCL-minus-CE-only difference in real-input target bits is +0.343443, with a crossed interval of [+0.062476, +0.616300]. Their difference in real-minus-zero gains is +0.002668 bits, with an interval of [−0.014784, +0.020946]. The former compares absolute performance; the latter compares input gains.

Figure B3 complements these scheme-level contrasts with unit-level effects; full numerical results appear in Appendix B.9.

B.5 Prediction Changes and Seeds

The following table examines prediction changes when switching from zero input to real EEG under NATIVE/Pool-S/MEAN. Correctness transitions and ranking changes use five-unit macro-averages rather than direct proportions over pooled rows. Net Top-1 differences are in pp.

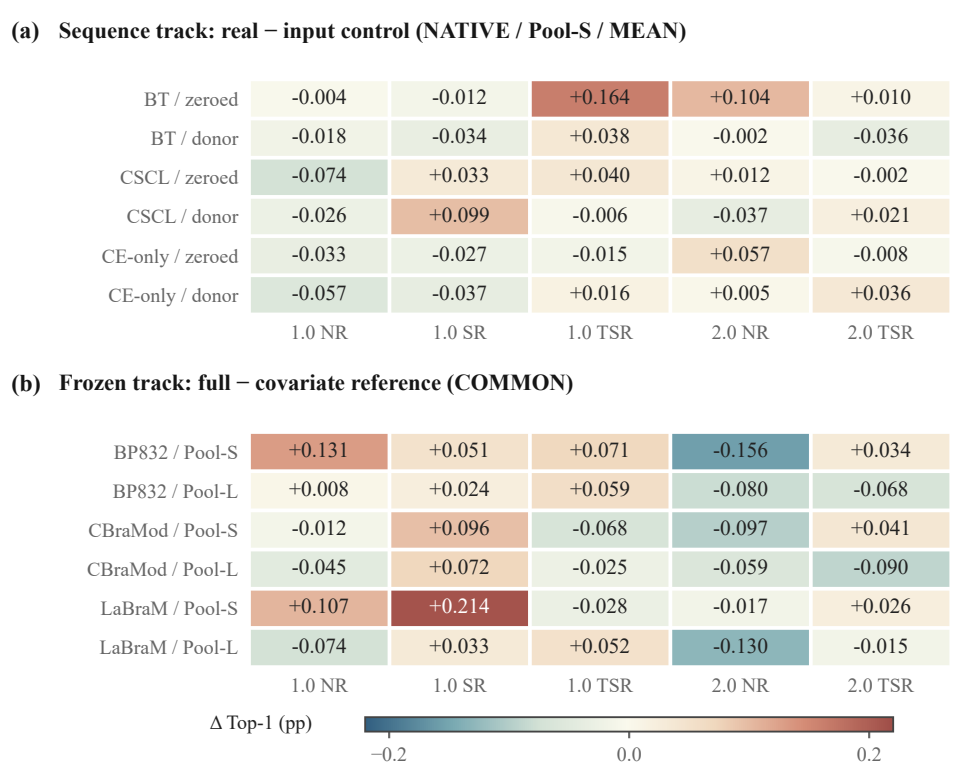


Figure B3: **Unit-level input gains within each track.** (a) Sequence-track real-minus-zeroed and real-minus-donor differences under NATIVE/Pool-S/MEAN, with donor-specific paired support. (b) Frozen-track full-minus-covariate-reference estimates on COMMON. A symmetric common color range shows magnitude while retaining the different track definitions. All displayed unit effects are within ± 0.22 pp: small macro estimates are not cancellation of large unit effects. These are descriptive estimates, not cross-track rankings.

Table B.12: Prediction changes under EEG replacement.

Scheme	Incorrect to correct (%)	Correct to incorrect (%)	Top-score set changed (%)	Full ranking changed (%)	Net Top-1 (pp)
CSCL	0.047990	0.046263	0.328725	3.767075	+0.001727
CE-only	0.138451	0.143841	1.066342	10.958151	−0.005390

The three-seed CSCL experiment covers the five Q2 cells. The following table reports real-minus-zero Top-1 differences in pp for each fixed cell under MEAN scoring. These are 15 cell–seed runs, not three complete replications of the 25-cell benchmark. Seed ranges are interpreted separately from the conditional intervals above.

Figure B4 visualizes the three saved CSCL seed estimates in each of the five Q2 cells.

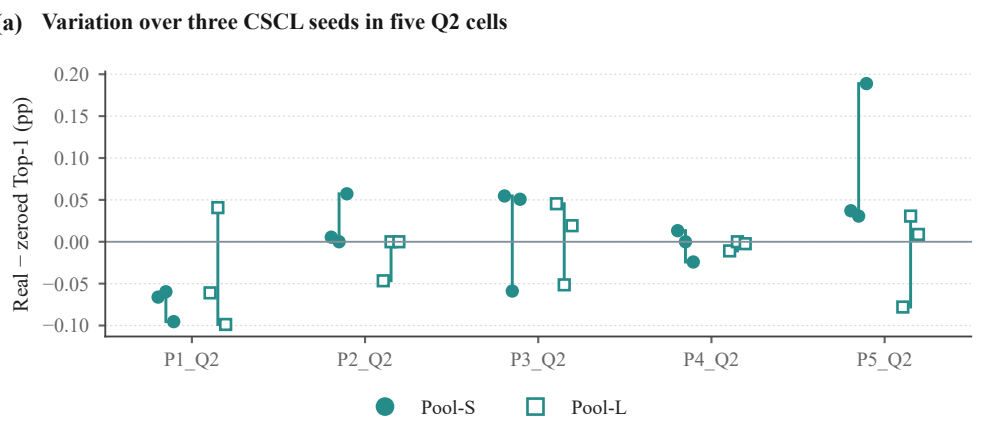


Figure B4: **Local seed variability.** Three CSCL estimates per Q2 cell under NATIVE/MEAN and zeroing. Horizontal offsets distinguish seeds and pools; vertical segments span the minimum and maximum. These 15 cell–seed runs are not three complete 25-cell benchmark replications. In P5_Q2/Pool-S, estimates range from +0.031 to +0.189 pp. Training variability is separate from the main conditional bootstrap intervals.

Table B.13: CSCL seed sensitivity.

Cell	Pool	Obs.	Seed 0	Seed 1	Seed 2	Mean	Range
P1_Q2	S	1176	−0.066081	−0.059680	−0.095380	−0.073714	[−0.095380, −0.059680]
P1_Q2	L	1176	−0.060999	+0.040768	−0.098745	−0.039659	[−0.098745, +0.040768]
P2_Q2	S	823	+0.005480	+0.000000	+0.057292	+0.020924	[+0.000000, +0.057292]
P2_Q2	L	823	−0.046584	+0.000000	+0.000000	−0.015528	[−0.046584, +0.000000]
P3_Q2	S	865	+0.054697	−0.058852	+0.050748	+0.015531	[−0.058852, +0.054697]
P3_Q2	L	865	+0.045406	−0.051432	+0.019175	+0.004383	[−0.051432, +0.045406]
P4_Q2	S	822	+0.013228	+0.000000	−0.024145	−0.003639	[−0.024145, +0.013228]
P4_Q2	L	822	−0.010917	+0.000000	−0.002310	−0.004409	[−0.010917, +0.000000]
P5_Q2	S	830	+0.037037	+0.030654	+0.188889	+0.085527	[+0.030654, +0.188889]
P5_Q2	L	830	−0.077936	+0.030654	+0.008758	−0.012841	[−0.077936, +0.030654]

B.6 Generation Results

Generation counts pool observations from 25 fitted cells. The number of distinct decoded strings measures diversity, not the number of correctly recovered reference sentences. Mean new-token counts exclude the initial decoder-start token and include subsequently generated special tokens. Outputs decoded as empty strings remain in the denominator. Donor conditions use their eligible support set; other conditions use the full sequence-track NATIVE support set.

Table B.14: Generation diversity and length.

Scheme	Input	Obs.	Distinct strings	Empty outputs	Mean new tokens
BT	Real	22609	42	2878	21.152019
BT	Zero	22609	25	2878	20.930337
BT	Noise	22609	92	2878	21.230926
BT	Word-order permutation	22609	45	2878	21.153479
BT	Donor	21847	42	2770	21.199158
BT	LM-only	22609	1	0	128.000000
CSCL	Real	22609	49	0	24.058826
CSCL	Zero	22609	48	0	24.138131
CSCL	Noise	22609	52	0	24.063913
CSCL	Word-order permutation	22609	48	0	24.058959
CSCL	Donor	21847	48	0	23.952030
CSCL	LM-only	22609	1	0	128.000000
CE-only	Real	22609	95	0	24.533504
CE-only	Zero	22609	50	0	24.931222
CE-only	Noise	22609	116	0	24.941926
CE-only	Word-order permutation	22609	98	0	24.515105
CE-only	Donor	21847	92	0	24.481531
CE-only	LM-only	22609	1	0	128.000000

For each BT condition, distinct strings are counted after pooling and deduplicating outputs from all 25 cells. Real, zero, noise, and word-order-permutation conditions have the same number of empty outputs, but these need not arise from the same observations. All three schemes use the same BART/EOS-source reference without task fine-tuning. Under the current generation settings, this reference produces only one string, with mean generation length reaching the cap of 128 new tokens. This generation behavior is interpreted separately from candidate likelihood scoring.

The following table lists observation-level generated-token agreement rates for controls that have been aggregated. Outputs agree only if their actual lengths match and all token IDs within the valid range are identical. This definition retains generated special tokens; equality of strings after removing special tokens is not a substitute. Rates are aggregated over observations, unlike the five-unit macro-average used for candidate metrics.

Table B.15: Generated-token agreement.

Scheme	Control versus real input	Paired observations	Identical pairs	Agreement (%)
BT	Zero	22609	18428	81.5074
BT	Noise	22609	18253	80.7333
BT	Word-order permutation	22609	19261	85.1917
BT	Donor	21847	18484	84.6066
BT	LM-only	22609	0	0.0000
CSCL	Zero	22609	21104	93.3434
CSCL	Word-order permutation	22609	22577	99.8585
CSCL	Donor	21847	19394	88.7719
CE-only	Zero	22609	16549	73.1965
CE-only	Noise	22609	15677	69.3396
CE-only	Word-order permutation	22609	18154	80.2955
CE-only	Donor	21847	16742	76.6329
CE-only	LM-only	22609	0	0.0000

B.7 Accuracy and Gains

The following table reports accuracy and intervals for CSCL and the independent LM-only reference under NATIVE/Pool-S/MEAN. All values are percentages. Intervals for the corresponding paired differences appear in Appendix B.3.

Table B.16: CSCL and independent-reference accuracy.

Input or reference	Top-1 (%)	Participant 95%	Text 95%	Crossed sensitivity 95%
CSCL real EEG	11.608064	[+11.512181, +11.705705]	[+10.338727, +12.874156]	[+10.335030, +12.891470]
CSCL zero	11.606338	[+11.510812, +11.702616]	[+10.333572, +12.874897]	[+10.335679, +12.888370]
Independent LM-only	10.160560	[+10.073092, +10.241174]	[+8.983535, +11.403522]	[+8.982940, +11.405792]

Under the same setting, the CSCL-minus-CE-only difference in real-minus-zero gains is +0.007116391 pp, with a participant interval of [−0.034286519, +0.045646578] and a crossed sensitivity interval of [−0.080575619, +0.095173629]. CSCL receives 15 contrastive epochs before the shared 30-epoch cross-entropy budget.

B.8 CE-only Input Controls

The following table reports four input-replacement controls for CE-only across 32 settings; eight comparisons with the independent LM-only reference are listed separately. Differences are computed on the observation sets specified in the tables; real and control inputs in donor comparisons use the same eligible observations. Intervals are estimated using the participant, text, and crossed resampling schemes in Appendix A.5, with 5,000 resamples per scheme and random seed 20260910. Intervals for the primary setting appear in Table 1; complete point estimates and intervals appear in Appendix E.

These are pointwise conditional intervals for fixed models and fixed evaluation materials. Training is not resampled, and no multiplicity adjustment is applied. They directly estimate CE-only real–control differences, a different statistical quantity from the CSCL-minus-CE-only gain differences in Appendix B.4.

Table B.17: CE-only input replacements: Top-1.

Support	Scoring	Pool	Obs.	Real (%)	Control (%)	Δ Top-1 (pp)
Zero						
NATIVE	SUM	S	22609	10.124455	10.129267	−0.004811
NATIVE	SUM	L	22609	12.955984	12.940554	+0.015430
NATIVE	MEAN	S	22609	11.708939	11.714329	−0.005390
NATIVE	MEAN	L	22609	12.324588	12.318982	+0.005606
COMMON	SUM	S	22378	10.010174	10.014306	−0.004132
COMMON	SUM	L	22378	12.767898	12.752348	+0.015550
COMMON	MEAN	S	22378	11.653106	11.657587	−0.004481
COMMON	MEAN	L	22378	12.260985	12.257421	+0.003563

Continued on next page

Table B.17: CE-only input replacements: Top-1. (continued)

Support	Scoring	Pool	Obs.	Real (%)	Control (%)	Δ Top-1 (pp)
Noise						
NATIVE	SUM	S	22609	10.124455	10.123882	+0.000574
NATIVE	SUM	L	22609	12.955984	12.946658	+0.009326
NATIVE	MEAN	S	22609	11.708939	11.702911	+0.006028
NATIVE	MEAN	L	22609	12.324588	12.315282	+0.009306
COMMON	SUM	S	22378	10.010174	10.010158	+0.000016
COMMON	SUM	L	22378	12.767898	12.759057	+0.008841
COMMON	MEAN	S	22378	11.653106	11.648992	+0.004114
COMMON	MEAN	L	22378	12.260985	12.252777	+0.008207
Word-order permutation						
NATIVE	SUM	S	22609	10.124455	10.124560	−0.000104
NATIVE	SUM	L	22609	12.955984	12.949160	+0.006824
NATIVE	MEAN	S	22609	11.708939	11.703304	+0.005635
NATIVE	MEAN	L	22609	12.324588	12.315855	+0.008733
COMMON	SUM	S	22378	10.010174	10.010303	−0.000129
COMMON	SUM	L	22378	12.767898	12.761553	+0.006345
COMMON	MEAN	S	22378	11.653106	11.647512	+0.005595
COMMON	MEAN	L	22378	12.260985	12.251845	+0.009140
Donor						
NATIVE	SUM	S	21847	10.253429	10.263665	−0.010236
NATIVE	SUM	L	21847	13.173356	13.164066	+0.009290
NATIVE	MEAN	S	21847	11.845019	11.852469	−0.007450
NATIVE	MEAN	L	21847	12.531134	12.532079	−0.000944
COMMON	SUM	S	21483	10.104766	10.115559	−0.010793
COMMON	SUM	L	21483	12.902686	12.893554	+0.009132
COMMON	MEAN	S	21483	11.784661	11.792243	−0.007583
COMMON	MEAN	L	21483	12.457429	12.457664	−0.000235

Table B.18: CE-only input replacements: target bits.

Support	Scoring	Pool	Obs.	Real bits	Control bits	Δ bits
Zero						
NATIVE	SUM	S	22609	−60.885754	−60.885425	−0.000329
NATIVE	SUM	L	22609	−82.922347	−82.924176	+0.001830
NATIVE	MEAN	S	22609	−3.221765	−3.221814	+0.000049
NATIVE	MEAN	L	22609	−3.367708	−3.367817	+0.000110
COMMON	SUM	S	22378	−61.054050	−61.053249	−0.000801
COMMON	SUM	L	22378	−83.211672	−83.212961	+0.001290
COMMON	MEAN	S	22378	−3.223716	−3.223750	+0.000034
COMMON	MEAN	L	22378	−3.370638	−3.370736	+0.000098
Noise						
NATIVE	SUM	S	22609	−60.885754	−60.906685	+0.020931
NATIVE	SUM	L	22609	−82.922347	−82.956148	+0.033802
NATIVE	MEAN	S	22609	−3.221765	−3.221754	−0.000011
NATIVE	MEAN	L	22609	−3.367708	−3.367730	+0.000022
COMMON	SUM	S	22378	−61.054050	−61.074486	+0.020436
COMMON	SUM	L	22378	−83.211672	−83.245183	+0.033511
COMMON	MEAN	S	22378	−3.223716	−3.223690	−0.000026
COMMON	MEAN	L	22378	−3.370638	−3.370648	+0.000009

Continued on next page

Table B.18: CE-only input replacements: target bits. (continued)

Support	Scoring	Pool	Obs.	Real bits	Control bits	Δ bits
Word-order permutation						
NATIVE	SUM	S	22609	−60.885754	−60.885332	−0.000422
NATIVE	SUM	L	22609	−82.922347	−82.921387	−0.000960
NATIVE	MEAN	S	22609	−3.221765	−3.221734	−0.000031
NATIVE	MEAN	L	22609	−3.367708	−3.367661	−0.000047
COMMON	SUM	S	22378	−61.054050	−61.053326	−0.000724
COMMON	SUM	L	22378	−83.211672	−83.210454	−0.001218
COMMON	MEAN	S	22378	−3.223716	−3.223674	−0.000042
COMMON	MEAN	L	22378	−3.370638	−3.370580	−0.000058
Donor						
NATIVE	SUM	S	21847	−57.426384	−57.425836	−0.000548
NATIVE	SUM	L	21847	−78.257374	−78.257243	−0.000131
NATIVE	MEAN	S	21847	−3.214858	−3.214856	−0.000002
NATIVE	MEAN	L	21847	−3.359952	−3.359936	−0.000016
COMMON	SUM	S	21483	−57.671183	−57.670399	−0.000784
COMMON	SUM	L	21483	−78.684030	−78.683814	−0.000217
COMMON	MEAN	S	21483	−3.217714	−3.217711	−0.000003
COMMON	MEAN	L	21483	−3.364240	−3.364226	−0.000014

The independent LM-only reference has a different training history and source format. The following differences describe the performance of the complete scheme relative to this reference and are not interpreted as same-checkpoint EEG-content gains.

Table B.19: CE-only versus LM-only: Top-1.

Support	Scoring	Pool	Obs.	Real (%)	Control (%)	Δ Top-1 (pp)
LM-only						
NATIVE	SUM	S	22609	10.124455	8.898612	+1.225844
NATIVE	SUM	L	22609	12.955984	12.949569	+0.006416
NATIVE	MEAN	S	22609	11.708939	10.160560	+1.548379
NATIVE	MEAN	L	22609	12.324588	12.863494	−0.538906
COMMON	SUM	S	22378	10.010174	8.763085	+1.247089
COMMON	SUM	L	22378	12.767898	12.745342	+0.022556
COMMON	MEAN	S	22378	11.653106	10.102754	+1.550353
COMMON	MEAN	L	22378	12.260985	12.746299	−0.485314

Table B.20: CE-only versus LM-only: target bits.

Support	Scoring	Pool	Obs.	Real bits	Control bits	Δ bits
LM-only						
NATIVE	SUM	S	22609	−60.885754	−114.992859	+54.107105
NATIVE	SUM	L	22609	−82.922347	−157.056932	+74.134585
NATIVE	MEAN	S	22609	−3.221765	−3.544038	+0.322272
NATIVE	MEAN	L	22609	−3.367708	−3.845361	+0.477653
COMMON	SUM	S	22378	−61.054050	−115.334383	+54.280333
COMMON	SUM	L	22378	−83.211672	−157.614558	+74.402886
COMMON	MEAN	S	22378	−3.223716	−3.547975	+0.324259
COMMON	MEAN	L	22378	−3.370638	−3.852682	+0.482044

B.9 Results by Unit

Equal-weight aggregation over five units gives each dataset–task unit the same weight, irrespective of its observation count. The following tables disaggregate the primary reporting setting, NATIVE/Pool-S/MEAN, into real scores, control scores, and paired differences for BT, CSCL, and CE-only. Within each unit, scores are first averaged over the eight candidate sets for each observation and then over observations. These are point estimates; unit-level intervals were not separately computed.

Real minus zero.

Table B.21: Unit-level real-minus-zero results.

Dataset–task unit	Obs.	Real (%)	Control (%)	Δ Top-1 (pp)	Δ bits
BT					
ZuCo 1.0 NR	3049	10.273860	10.277960	−0.004100	−0.000021
ZuCo 1.0 SR	4148	11.909354	11.921408	−0.012054	−0.000082
ZuCo 1.0 TSR	4105	11.961023	11.796590	+0.164434	−0.000003
ZuCo 2.0 NR	5307	11.604956	11.501319	+0.103637	−0.000004
ZuCo 2.0 TSR	6000	10.656250	10.645833	+0.010417	−0.000081
CSCL					
ZuCo 1.0 NR	3049	10.897015	10.970810	−0.073795	+0.000127
ZuCo 1.0 SR	4148	12.084137	12.050988	+0.033149	+0.000006
ZuCo 1.0 TSR	4105	12.539586	12.500000	+0.039586	+0.000008
ZuCo 2.0 NR	5307	11.244583	11.232806	+0.011777	−0.000051
ZuCo 2.0 TSR	6000	11.275000	11.277083	−0.002083	+0.000065
CE-only					
ZuCo 1.0 NR	3049	10.921614	10.954411	−0.032798	+0.000204
ZuCo 1.0 SR	4148	12.286041	12.313163	−0.027122	+0.000022
ZuCo 1.0 TSR	4105	12.685749	12.700974	−0.015225	−0.000157
ZuCo 2.0 NR	5307	11.284624	11.228095	+0.056529	+0.000291
ZuCo 2.0 TSR	6000	11.366667	11.375000	−0.008333	−0.000115

Real minus donor. Real scores here are recomputed on donor-eligible observations.

Table B.22: Unit-level real-minus-donor results.

Dataset–task unit	Obs.	Real (%)	Control (%)	Δ Top-1 (pp)	Δ bits
BT					
ZuCo 1.0 NR	2851	10.470011	10.487548	−0.017538	+0.000050
ZuCo 1.0 SR	4041	11.893714	11.927741	−0.034026	−0.000108
ZuCo 1.0 TSR	3938	12.258761	12.220670	+0.038090	+0.000103
ZuCo 2.0 NR	5135	11.565239	11.567673	−0.002434	−0.000097
ZuCo 2.0 TSR	5882	10.765896	10.802023	−0.036127	+0.000168
CSCL					
ZuCo 1.0 NR	2851	11.101368	11.127675	−0.026307	−0.000002
ZuCo 1.0 SR	4041	12.187577	12.088592	+0.098985	−0.000112
ZuCo 1.0 TSR	3938	12.839639	12.845988	−0.006348	−0.000100
ZuCo 2.0 NR	5135	11.148978	11.185492	−0.036514	+0.000005
ZuCo 2.0 TSR	5882	11.407684	11.386433	+0.021251	+0.000069

Continued on next page

Table B.22: Unit-level real-minus-donor results. (continued)

Dataset–task unit	Obs.	Real (%)	Control (%)	Δ Top-1 (pp)	Δ bits
CE-only					
ZuCo 1.0 NR	2851	11.140828	11.197825	−0.056998	−0.000208
ZuCo 1.0 SR	4041	12.354615	12.391735	−0.037120	+0.000043
ZuCo 1.0 TSR	3938	12.995175	12.979304	+0.015871	−0.000277
ZuCo 2.0 NR	5135	11.243914	11.239046	+0.004869	+0.000253
ZuCo 2.0 TSR	5882	11.490564	11.454437	+0.036127	+0.000178

CSCL's zero-input gains vary in sign across units, but their absolute magnitudes are all below 0.08 pp; those for CE-only are all below 0.06 pp. Thus, their near-zero macro-averages do not result from cancellation of large positive and negative unit-level gains. Donor results apply only to observations with eligible real recordings of the same length.

BT's zero-input gains range from −0.012054 to +0.164434 pp, including +0.164434 pp for ZuCo 1.0 TSR and +0.103637 pp for ZuCo 2.0 NR; the macro-average is +0.052467 pp. Its donor gains range from −0.036127 to +0.038090 pp across the five units, with a macro-average of −0.010407 pp. All five BT zero-input comparisons have negative target-bits differences, indicating that Top-1 and continuous scores should be reported separately rather than substituted for each other.

Frozen-representation track. The following table uses COMMON. Full combines EEG scores with the covariate reference, and Δ is Full minus the covariate reference. This track differs from the sequence track in scores, references, and input adaptation; the table examines differences between units within the frozen track.

Table B.23: Unit-level frozen-representation results.

Pool	Dataset–task unit	Obs.	Full (%)	Reference (%)	Δ Top-1 (pp)	Δ bits
BP832						
S	ZuCo 1.0 NR	3045	13.830049	13.698686	+0.131363	+0.000127
S	ZuCo 1.0 SR	4146	15.439580	15.388326	+0.051254	−0.000060
S	ZuCo 1.0 TSR	4050	14.805556	14.734568	+0.070988	+0.000145
S	ZuCo 2.0 NR	5281	15.195986	15.352206	−0.156220	−0.000430
S	ZuCo 2.0 TSR	5856	16.696551	16.662398	+0.034153	−0.000340
L	ZuCo 1.0 NR	3045	22.738095	22.729885	+0.008210	−0.000007
L	ZuCo 1.0 SR	4146	23.371925	23.347805	+0.024120	−0.000100
L	ZuCo 1.0 TSR	4050	20.660494	20.601852	+0.058642	−0.000646
L	ZuCo 2.0 NR	5281	23.177429	23.257906	−0.080477	−0.000683
L	ZuCo 2.0 TSR	5856	21.939037	22.007343	−0.068306	−0.000683
CBraMod-frozen						
S	ZuCo 1.0 NR	3045	13.686371	13.698686	−0.012315	−0.000137
S	ZuCo 1.0 SR	4146	15.484805	15.388326	+0.096479	+0.000069
S	ZuCo 1.0 TSR	4050	14.666667	14.734568	−0.067901	+0.000010
S	ZuCo 2.0 NR	5281	15.255160	15.352206	−0.097046	−0.000236
S	ZuCo 2.0 TSR	5856	16.702954	16.662398	+0.040557	−0.000082
L	ZuCo 1.0 NR	3045	22.684729	22.729885	−0.045156	−0.000098
L	ZuCo 1.0 SR	4146	23.420164	23.347805	+0.072359	+0.000183
L	ZuCo 1.0 TSR	4050	20.577160	20.601852	−0.024691	−0.000598
L	ZuCo 2.0 NR	5281	23.198731	23.257906	−0.059174	−0.000277
L	ZuCo 2.0 TSR	5856	21.917691	22.007343	−0.089652	−0.000142

Continued on next page

Table B.23: Unit-level frozen-representation results. (continued)

Pool	Dataset–task unit	Obs.	Full (%)	Reference (%)	Δ Top-1 (pp)	Δ bits
LaBraM-frozen						
S	ZuCo 1.0 NR	3045	13.805419	13.698686	+0.106732	−0.000015
S	ZuCo 1.0 SR	4146	15.602388	15.388326	+0.214062	+0.000092
S	ZuCo 1.0 TSR	4050	14.706790	14.734568	−0.027778	+0.000058
S	ZuCo 2.0 NR	5281	15.335637	15.352206	−0.016569	−0.000141
S	ZuCo 2.0 TSR	5856	16.688012	16.662398	+0.025615	−0.000057
L	ZuCo 1.0 NR	3045	22.655993	22.729885	−0.073892	−0.000078
L	ZuCo 1.0 SR	4146	23.380970	23.347805	+0.033164	+0.000141
L	ZuCo 1.0 TSR	4050	20.654321	20.601852	+0.052469	−0.000488
L	ZuCo 2.0 NR	5281	23.127722	23.257906	−0.130184	−0.000079
L	ZuCo 2.0 TSR	5856	21.992401	22.007343	−0.014942	−0.000109

Disaggregating BT's 40 control settings across five dataset–task units yields 200 unit-level point estimates. Appendix E provides the complete results and unit-level results for the other schemes.

C Validation and Diagnostics

C.1 Artificial Signals

Eight target classes are represented by 840-dimensional class codes, with independent noise added before the actual per-word standardization step. All inputs have a valid length of 12. Training uses 1,024 class-balanced observations, and validation uses 128 observations. The model is fixed at update 1,000 rather than selected using test results. Each test coefficient uses 256 new noise realizations. Coefficients specify artificial signal strength; they do not estimate real-EEG signal-to-noise ratios.

Table C.1: Artificial-signal learnability.

Injection coefficient	Signal-input Top-1 (%)	Zero (%)	Wrong-class donor (%)
0	12.5000	12.5000	13.2813
0.25	99.6094	12.5000	0.0000
1	100.0000	12.5000	0.0000
4	100.0000	12.5000	0.0000

At higher signal strengths, results meet the engineering criteria of at least 90% accuracy and a 60 pp real-minus-zero gain. These criteria assess learnability of the artificial task; they do not calibrate statistical power for real EEG or test generalization to new target classes.

C.2 Real-data Micro-fitting

Micro-fitting uses 64 training observations from distinct text clusters, with 13/13/13/13/12 observations across the five units. A reinitialized model is trained for 1,000 updates and evaluated on those same training observations with equal unit weights. Real/zero accuracy is 77.6122%/77.9968%, a difference of -0.3846 pp. The target-bits difference is +0.007505, and 21.2019% of candidate queries change their full ranking. Of 16 generation pairs, 13 are exactly identical. On a separate subset of 29 donor-eligible observations, real/donor accuracy is 82.0685%/79.9851%, a difference of +2.0833 pp.

Real-data micro-fitting does not meet the prespecified criteria of 80% real-input accuracy and a 30 pp real-minus-zero gain. Subsequent local holdout comparisons are therefore interpreted as diagnostics, not as evidence that a real-data positive control has passed. Gradient checks find finite, nonzero

gradients in the EEG encoder, projection, and BART, showing that gradients reach these components. This alone does not establish that the model has learned correct EEG–text relationships.

A post hoc training-text-membership reference further clarifies the meaning of high training performance. Candidates whose coarse-grained text clusters appear among the 64 training observations receive the same high score; all others receive the same low score. Scoring uses neither EEG nor target indices. This reference achieves 74.5780% Top-1. In 53.4615% of queries, only one candidate has been seen during training, and it is the target. This is an explanatory analysis, not proof that the neural decoder implements the same rule. Under the text-exposure condition in the next section, both target and distractor texts occur in training, so this binary rule alone cannot distinguish them.

C.3 Holdout Conditions

Local comparison overview. To examine the effects of participant or text exposure, we relax the corresponding holdout in two cells using CE-only and Pool-S/MEAN, with one initialization seed per cell. Within each cell, settings share 200 test observations, 3,000 training observations, and the cross-entropy budget. Participant exposure uses recordings of test participants reading other texts; text exposure uses recordings of other participants reading the target and distractor text clusters. Table C.2 reports real-input Top-1 accuracy (%) and real-minus-zero differences (pp). Real-minus-zero gains remain between -0.125 and $+0.750$ pp. Four observations produce the largest positive gain; continuous target-bits differences are slightly negative in all six settings.

Table C.2: Local holdout diagnostics.

Training condition	**P3_Q2 real**	**P3_Q2 Δ**	**P1_Q4 real**	**P1_Q4 Δ**
Joint holdout	9.5000	-0.1250	11.5625	$+0.0000$
Participant exposure	10.4375	-0.1250	12.6875	$+0.2500$
Text exposure	9.5000	$+0.1875$	9.6250	$+0.7500$

Text exposure also combines low validation loss with weak input gains. Both text-exposure models achieve best validation NLLs of approximately 0.337, selecting epochs 30 and 24, respectively. Yet each produces only one token sequence across its fixed 16 generation-check observations, with identical real and zero outputs. Joint-holdout and participant-exposure models all select epoch 1. Because the three protocols use different validation pools, NLL differences describe their learning conditions and cannot be interpreted directly as EEG contributions. These diagnostics show that, under the tested conditions, artificial-signal learnability, decreasing training loss, or greater text exposure does not automatically yield larger gains from paired EEG; measurement sensitivity to weak real signals remains uncalibrated.

Figure C1 resolves local paired gains and ranking responses by dataset–task unit.

For test cell $P_i \times Q_j$, let P_v, Q_v denote the cyclic successor groups and P_R, Q_R the remaining three groups on each axis. This section denotes joint holdout, participant exposure, and text exposure by D, P, and Q, respectively.

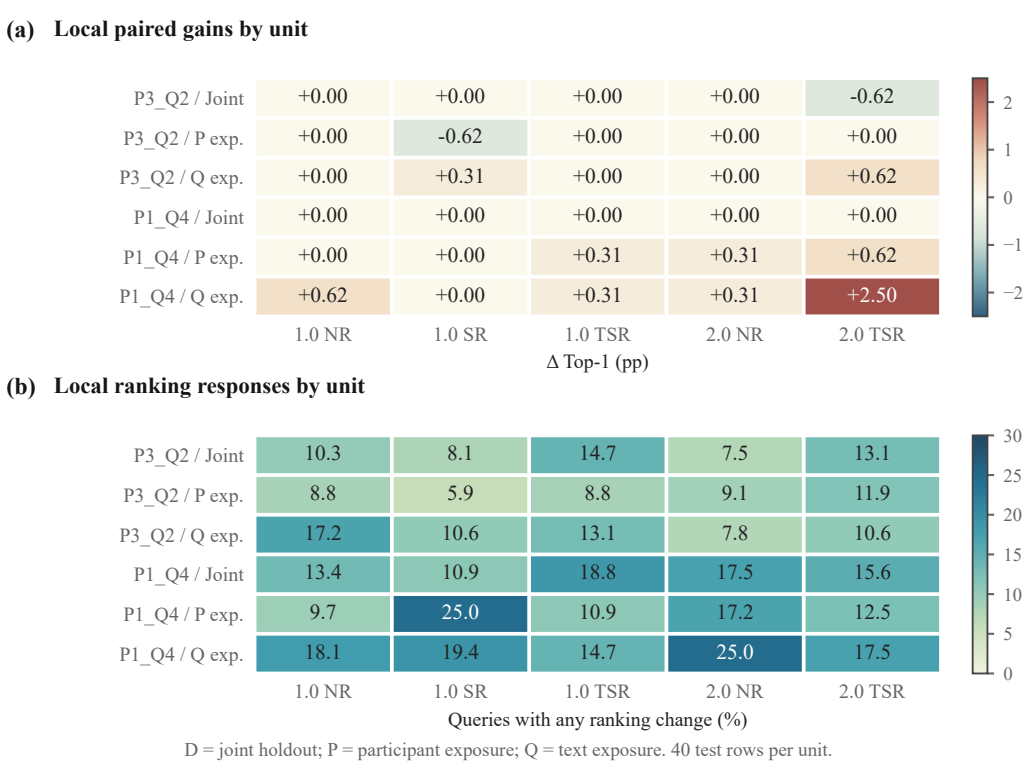


Figure C1: **Local holdout diagnostics by unit.** CE-only/Pool-S/MEAN under joint holdout, participant exposure, and text exposure in two cells. Each cell uses the same 200 test observations (40 per unit) across protocols. (a) Real-minus-zeroed accuracy in pp. (b) Queries with any complete-ranking change. The largest macro gain, +0.75 pp for P1_Q4/text exposure, comes from four observations; one contributes +0.50 pp. Six models use one initialization seed per setting. No uncertainty intervals are estimated, and validation pools differ across protocols.

Table C.3: Local holdout protocols.

Protocol	Training pool	Validation pool	Test pool
D	$P_R \times Q_R$	$P_v \times Q_v$	$P_i \times Q_j$
P	$(P_R \cup P_i) \times Q_R$	$P_i \times Q_v$	$P_i \times Q_j$
Q	$P_R \times (Q_R \cup Q_j)$	$P_v \times Q_j$	$P_i \times Q_j$

For each model, 600/40/40 training/validation/test observations are sampled per unit. The three protocols within a cell share test observation identities, and these EEG observations are excluded from training and validation. Both target and distractor text clusters have training coverage under text exposure. Exact equality of initial weights across training conditions has not been verified. All six models complete 2,820 updates but may select different checkpoints.

All six models complete 30 cross-entropy epochs. The following table distinguishes the final executed training epoch from the checkpoint selected by validation; exact ties select the earliest epoch.

Table C.4: Local training and checkpoint selection.

Cell/protocol	Updates	Chosen epoch	Training NLL		Validation NLL	
			First epoch	Final epoch	Best	Final epoch
P3_Q2_D	2820	1	3.672535	0.280932	3.187902	5.610200
P3_Q2_P	2820	1	3.615332	0.284975	3.190358	5.442082
P3_Q2_Q	2820	30	3.904819	0.309671	0.337384	0.337384
P1_Q4_D	2820	1	3.626647	0.274879	3.249482	5.730027
P1_Q4_P	2820	1	3.360603	0.276363	3.183022	5.679353
P1_Q4_Q	2820	24	3.530003	0.300853	0.337044	0.341945

Joint-holdout and participant-exposure models all reach their lowest validation NLL at epoch 1; training loss then continues to decrease while validation loss increases. The two text-exposure models select epochs 30 and 24, with minimum validation NLL near 0.337. Nevertheless, each produces only one real-input sequence in its fixed generation check, and all 16/16 real–zero output pairs are identical. Thus, lower validation token loss on seen texts does not accompany substantial EEG-conditioned differentiation. Validation pools differ across protocols, so NLL differences cannot be directly attributed to additional EEG information.

The supplementary metrics below use Pool-S/MEAN. Donor support contains 195 observations for P3_Q2 and 192 for P1_Q4, fixed across the three protocols within each cell. Each model uses 16 fixed observations for generation checks. Only point estimates are reported; confidence intervals and hypothesis tests were not computed for this local experiment.

Table C.5: Local holdout diagnostics.

Cell/protocol	Donor observations	Real minus donor (pp)	Real minus zero bits	Identical real/zero generations	Distinct real-input sequences
P3_Q2_D	195	−0.128205	−0.00029531	12/16	3
P3_Q2_P	195	−0.064103	−0.00013637	8/16	2
P3_Q2_Q	195	−0.067476	−0.00034992	16/16	1
P1_Q4_D	192	−0.062500	−0.00247208	1/16	3
P1_Q4_P	192	+0.182432	−0.00001187	3/16	7
P1_Q4_Q	192	+0.635135	−0.00047335	16/16	1

The largest real-minus-zero Top-1 gain, +0.7500 pp for P1_Q4/Q, comes from the following four observations. The other 196 have zero change in row-mean Top-1. Each unit has 40 test observations, so each row contributes its change divided by 200 to macro-average accuracy. One ZuCo 2.0 TSR observation contributes 0.5 pp.

Table C.6: Observations contributing to the largest local gain.

Row ID	Unit	Row-mean Top-1 change (0–1)	Macro-average contribution (pp)
2275	zuco1:NR	+0.250	+0.1250
2820	zuco1:TSR	+0.125	+0.0625
16220	zuco2:NR	+0.125	+0.0625
17485	zuco2:TSR	+1.000	+0.5000

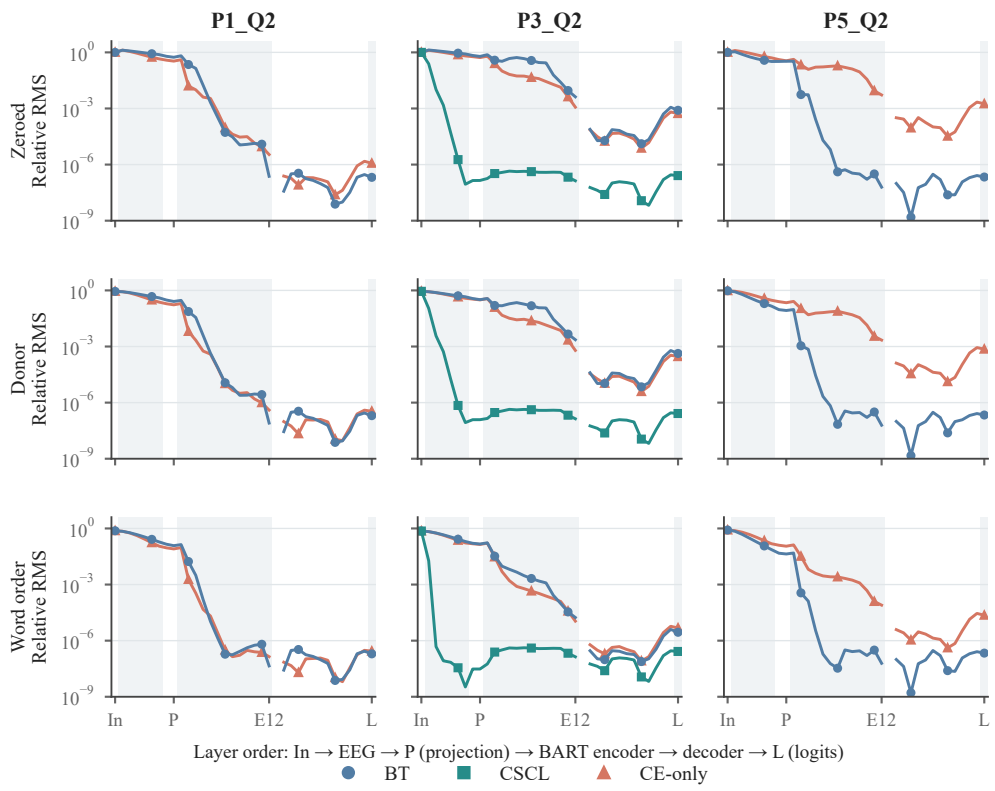


Figure C2: **Layer responses across cells and controls.** Median row-level relative RMS changes under zeroing, donor replacement, and word-order permutation; columns show P1_Q2, P3_Q2, and P5_Q2. CE-only and BT are available in all columns; CSCL only in P3_Q2. All panels share a logarithmic scale and omit exact zeros. The 36 taps run from EEG input to first-step logits. Separate model summaries do not reconstruct row-level pairing across models. Responses are not retained semantic-information fractions, and no numerical noise floor was established.

The experiment contains 400 distinct test observation identities, each scored by three models. The resulting 1,200 model–observation records do not represent 1,200 independent EEG samples. Validation-set composition differs across protocols, and their NLL values describe their respective learning conditions.

C.4 Layerwise Responses

Figure C2 compares layer-response profiles across cells and input controls.

For activation h, relative response is defined over valid positions as $\mathrm{RMS}(h_{\mathrm{real}} - h_{\mathrm{zero}}) / \mathrm{RMS}(h_{\mathrm{real}})$. It is undefined when the reference RMS is no greater than 10^{-12}. The following table uses the same 500 P3_Q2 observations with FP32 inference, TF32 and autocast disabled, no target labels, and a fixed decoder-start token.

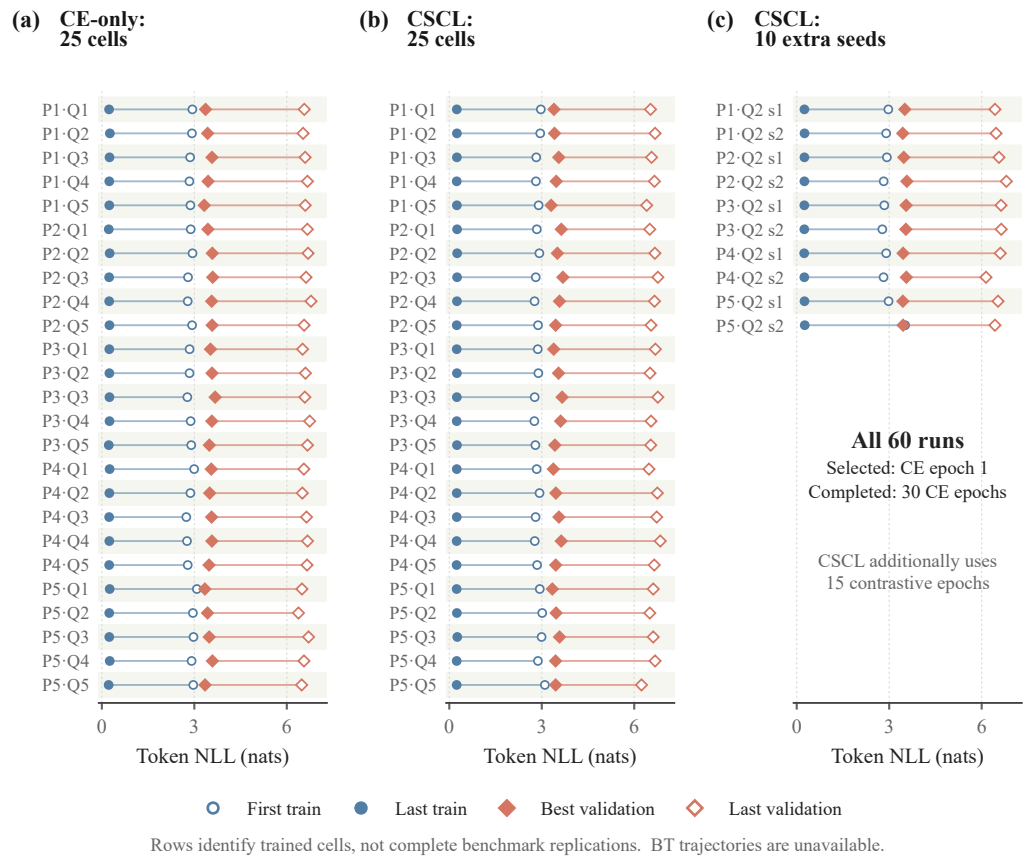


Figure C3: **Training and checkpoint selection across 60 runs.** First/last training NLL and best/last validation NLL for 25 CE-only runs, 25 primary CSCL runs, and ten extra CSCL seed runs in five Q2 cells. Panels share the horizontal scale. All runs select CE epoch 1 and complete 30 CE epochs; CSCL additionally uses 15 contrastive epochs. These are local training runs, not 60 complete benchmark replications. Only 29 full histories are available in S10; this figure uses all 60 S09 summary records. BT histories are unavailable.

Table C.7: Layerwise input responses.

Location	CE-only median	CSCL median
EEG encoder layer 6	0.64627914	8.8785722×10^{-8}
Projection after ReLU	0.52841445	1.4179888×10^{-7}
BART encoder layer 12	0.0011747685	1.3600645×10^{-7}
First-step logits	0.00053352072	2.5774283×10^{-7}

Values are medians of row-level relative responses, not proportions of semantic information retained across layers. For responses near machine precision, a numerical noise floor has not been established through independent repeated forward passes, so these values are not treated as exact zeros.

C.5 Training and Checkpoint Selection

Figure C3 displays training and validation NLL summaries for all 60 saved runs.

Checkpoints are selected by validation token-weighted NLL within each cell, without using outer-test performance. The following table separately summarizes the 25 main cells and 10 additional

seed runs over the five Q2 cells. Together with the five seed-0 runs in the main grid, the latter form the 15 local seed runs in Appendix B.5. Each CSCL run completes 15 contrastive and 30 cross-entropy epochs; CE-only completes 30 cross-entropy epochs. NLL values are medians across each group of runs, not pooled NLL over all validation tokens.

Table C.8: Training-run summaries.

Run group	Runs	Chosen CE epoch	Training NLL		Validation NLL	
			First epoch	Final epoch	Best	Final epoch
CSCL: 25 main cells	25	1 (all)	2.869870	0.244037	3.465639	6.616832
CE-only: 25 main cells	25	1 (all)	2.880700	0.241746	3.524450	6.598425
CSCL: 10 additional seed runs	10	1 (all)	2.902608	0.248999	3.489767	6.546725

All checkpoints in the three run groups are selected at cross-entropy epoch 1, although every run continues through epoch 30. These records show improving training fit alongside deteriorating held-out validation performance and identify the checkpoint-selection stage underlying the main candidate and generation results. This summary covers only the listed Adam training schemes; it does not imply a corresponding epoch-by-epoch trajectory for BT.

The following table lists checkpoint selections and NLL summaries for 60 training runs: 35 CSCL and 25 CE-only runs. CSCL comprises 25 main-cell runs and 10 additional seed runs on the five Q2 cells. These records therefore do not represent 60 complete benchmark replications.

Table C.9: Checkpoint selection by run.

Run	CL/CE epochs	Chosen CE epoch	Training NLL		Validation NLL	
			First epoch	Final epoch	Best	Final epoch
CE_P1_Q1_s0	0/30	1	2.935732	0.233950	3.364284	6.566572
CE_P1_Q2_s0	0/30	1	2.922824	0.260156	3.432053	6.528225
CE_P1_Q3_s0	0/30	1	2.863225	0.248320	3.573776	6.594726
CE_P1_Q4_s0	0/30	1	2.841062	0.240820	3.446653	6.669949
CE_P1_Q5_s0	0/30	1	2.872723	0.240772	3.321220	6.598425
CE_P2_Q1_s0	0/30	1	2.883141	0.224330	3.441061	6.667176
CE_P2_Q2_s0	0/30	1	2.941259	0.252918	3.583175	6.689954
CE_P2_Q3_s0	0/30	1	2.795329	0.235993	3.596429	6.623655
CE_P2_Q4_s0	0/30	1	2.785658	0.237589	3.561959	6.786140
CE_P2_Q5_s0	0/30	1	2.927216	0.245505	3.577988	6.561342
CE_P3_Q1_s0	0/30	1	2.846337	0.241746	3.524450	6.512557
CE_P3_Q2_s0	0/30	1	2.844364	0.243642	3.574690	6.604852
CE_P3_Q3_s0	0/30	1	2.772879	0.246121	3.675560	6.585389
CE_P3_Q4_s0	0/30	1	2.880700	0.248344	3.567754	6.741646
CE_P3_Q5_s0	0/30	1	2.900022	0.224503	3.486896	6.671134
CE_P4_Q1_s0	0/30	1	2.994957	0.251756	3.549639	6.553662
CE_P4_Q2_s0	0/30	1	2.872747	0.243164	3.498462	6.503601
CE_P4_Q3_s0	0/30	1	2.740112	0.236639	3.560524	6.638922
CE_P4_Q4_s0	0/30	1	2.762415	0.241807	3.570603	6.671490
CE_P4_Q5_s0	0/30	1	2.785886	0.237287	3.476621	6.656472
CE_P5_Q1_s0	0/30	1	3.082153	0.253849	3.341687	6.491460
CE_P5_Q2_s0	0/30	1	2.954000	0.239485	3.428784	6.378269

Continued on next page

Table C.9: Checkpoint selection by run. (continued)

Run	CL/CE epochs	Chosen CE epoch	Training NLL		Validation NLL	
			First epoch	Final epoch	Best	Final epoch
CE_P5_Q3_s0	0/30	1	2.972999	0.240613	3.483293	6.705955
CE_P5_Q4_s0	0/30	1	2.911281	0.255539	3.586915	6.556801
CE_P5_Q5_s0	0/30	1	2.966180	0.220300	3.347740	6.482289
CSCL_P1_Q1_s0	15/30	1	2.969859	0.244037	3.396184	6.528033
CSCL_P1_Q2_s0	15/30	1	2.950339	0.247938	3.413020	6.675640
CSCL_P1_Q2_s1	15/30	1	2.971526	0.249211	3.506226	6.431784
CSCL_P1_Q2_s2	15/30	1	2.902179	0.253579	3.440248	6.466860
CSCL_P1_Q3_s0	15/30	1	2.821338	0.246874	3.552074	6.562518
CSCL_P1_Q4_s0	15/30	1	2.811034	0.241366	3.469863	6.655988
CSCL_P1_Q5_s0	15/30	1	2.898233	0.246141	3.303353	6.403281
CSCL_P2_Q1_s0	15/30	1	2.839487	0.252696	3.632771	6.503540
CSCL_P2_Q2_s0	15/30	1	2.922190	0.255074	3.506977	6.669700
CSCL_P2_Q2_s1	15/30	1	2.924921	0.249648	3.473309	6.561473
CSCL_P2_Q2_s2	15/30	1	2.820931	0.248787	3.573294	6.801304
CSCL_P2_Q3_s0	15/30	1	2.805447	0.253091	3.685315	6.766830
CSCL_P2_Q4_s0	15/30	1	2.766182	0.248775	3.575363	6.657600
CSCL_P2_Q5_s0	15/30	1	2.879251	0.241716	3.450289	6.545281
CSCL_P3_Q1_s0	15/30	1	2.869870	0.241502	3.387740	6.685571
CSCL_P3_Q2_s0	15/30	1	2.888822	0.247520	3.543434	6.512067
CSCL_P3_Q2_s1	15/30	1	2.841438	0.246166	3.549431	6.628300
CSCL_P3_Q2_s2	15/30	1	2.776922	0.246193	3.541503	6.634730
CSCL_P3_Q3_s0	15/30	1	2.769775	0.244571	3.660724	6.764561
CSCL_P3_Q4_s0	15/30	1	2.759281	0.240810	3.612614	6.544458
CSCL_P3_Q5_s0	15/30	1	2.793079	0.248340	3.425863	6.535590
CSCL_P4_Q1_s0	15/30	1	2.834470	0.242882	3.374963	6.478400
CSCL_P4_Q2_s0	15/30	1	2.927519	0.241800	3.457706	6.750214
CSCL_P4_Q2_s1	15/30	1	2.903036	0.242011	3.451406	6.604683
CSCL_P4_Q2_s2	15/30	1	2.816770	0.246421	3.557799	6.140263
CSCL_P4_Q3_s0	15/30	1	2.800038	0.242462	3.557416	6.729555
CSCL_P4_Q4_s0	15/30	1	2.772972	0.238818	3.631898	6.844996
CSCL_P4_Q5_s0	15/30	1	2.848186	0.237615	3.453802	6.649390
CSCL_P5_Q1_s0	15/30	1	2.935005	0.239250	3.342860	6.616832
CSCL_P5_Q2_s0	15/30	1	3.014580	0.246472	3.465639	6.504984
CSCL_P5_Q2_s1	15/30	1	2.980324	0.251780	3.443474	6.531976
CSCL_P5_Q2_s2	15/30	1	3.517527	0.259932	3.450435	6.430853
CSCL_P5_Q3_s0	15/30	1	2.994327	0.259688	3.576414	6.615737
CSCL_P5_Q4_s0	15/30	1	2.876018	0.239586	3.447961	6.676452
CSCL_P5_Q5_s0	15/30	1	3.100242	0.239914	3.449940	6.235130

Epoch-level NLL data cover all 25 CE-only runs and four CSCL runs with seeds 1 and 2 on P2_Q2 and P4_Q2, totaling 870 cross-entropy epoch records. The other 31 CSCL runs provide summaries for the first epoch, final epoch, and best validation checkpoint. Appendix C.3 reports checkpoint selection for the six local holdout models.

C.6 Implementation Checks

Scoring checks and neural learning checks validate different components. The former construct known answers for the scorer; the latter test input learnability through the actual training pipeline. The following scoring checks use 14,768 candidate sets from the real candidate schedule.

Table C.10: Scoring checks.

Check	Result	Component checked
Target given the unique highest score	Top-1 is 100%	Target indexing and hit scoring
All candidates given equal scores	Top-1 is 12.5%; target bits is -3	Fractional credit for exact ties and candidate normalization
Highest score deliberately assigned to a non-target	Top-1 is 0%	Incorrect targets are not scored as correct
Candidates and target indices jointly permuted	Metrics unchanged	Binding of candidate order to target indices

Decoding diagnostics cover three BT checkpoints selected by validation on P1_Q2, P3_Q2, and P5_Q2, plus two diagnostic models trained for 300 updates using SGD and AdamW, respectively, on eight training observations from P3_Q2. The latter examine small-sample training behavior and are not among the three main sequence-training schemes. Each model uses eight fixed training observations. BT observations are sampled across dataset–task units, whereas the two diagnostic models use their own training samples. Label-shift, causal-mask, and same-prefix cache comparisons in the following table use the first three observations for each model. These checks concern only the listed models and inputs, not all 25 outer cells.

Cache comparisons use raw logits before generation-rule processing, comparing full and incremental computation with identical prefixes. Elementwise tolerance is $|a-b| \leq \mathrm{atol}+\mathrm{rtol}|b|$. FP32 absolute and relative tolerances are 10^{-5} and 10^{-4}, respectively; both are 0.03 for BF16.

Table C.11: Decoding diagnostics.

Check	Scope and results
Automatic/explicit label shifting	Five models times three observations times FP32/BF16: 30 computations, with exactly matching results
Future-token perturbation	Logits at past positions remain exactly unchanged in the same 30 checks
Same-prefix/incremental caching	No Top-1 disagreement in FP32, but the maximum elementwise difference is approximately 1.11×10^{-4}, and some comparisons fail the stated tolerance; numerical differences are larger in BF16
Generation cache disabled	Of 40 output pairs, 36 have identical full token sequences and 38 have identical decoded strings
Repeated n-gram restriction disabled	All 24 inputs across three BT checkpoints repeatedly generate BOS up to the 128-new-token limit; outputs are empty after removing special tokens

Custom and library greedy decoding differ in two comparisons, with the first divergence occurring at an exact tie for the highest score in both cases. The implementations select tokens using `topk(2).indices[0]` and `argmax`, respectively; these differences therefore cannot be attributed solely to the generation cache. These diagnostics apply only to the tested models and inputs and do not rule out other preprocessing, numerical, or implementation issues.

D Evaluation Interface

The scoring interface binds scores to fixed row identities and candidate schedules. It requires complete coverage and correct ordering on the declared support set, finite arrays of shape $N \times 8 \times 8$, and the prescribed control pairings. Invalid or incomplete submissions fail validation rather than

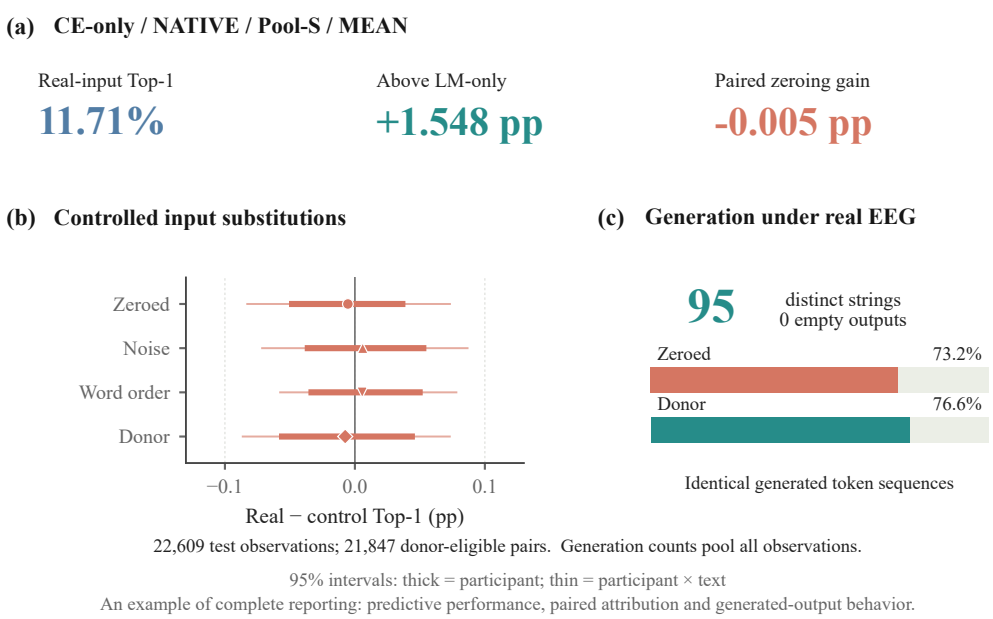


Figure D1: **A complete result profile.** CE-only under NATIVE/Pool-S/MEAN: predictive Top-1, advantage over a separately evaluated LM-only reference, same-checkpoint input gains, and generation behavior. Bars report exact token-sequence identity. Candidate estimates use five-unit macro means; generation fractions pool observations. Participant and crossed intervals are conditional on fitted models. No training or inference was rerun for this report. Near-zero gains do not establish the absence of EEG information.

being filled with zero scores. Empty generated strings remain valid observation outputs. Candidate metrics introduce no generic abstention or timeout penalty.

The supplementary package provides `code/benchmark.py` as the entry point for protocol checks, candidate scoring, and conditional-interval computation. Submission fields, shapes, and reference declarations are specified in `docs/SUBMISSION_FORMAT.md`. The scorer reads target positions from fixed metadata and aggregates over the eight candidate schedules, then observations within units, and finally the five equally weighted units. Complete sharded sequence results are aggregated by `code/postprocess.py`, with donor comparisons retaining their own eligible paired support sets.

Figure D1 illustrates a complete result profile combining prediction, paired input gains, and generation behavior.

E Numerical Results and Electronic Supplementary Materials

The main text and Appendices B–C present the result tables needed to interpret the principal findings. The complete analysis grid accompanies the anonymous supplementary archive as electronic supplementary materials S1–S12, indexed in Table E.1. After extraction, the root `README.md` maps each identifier to its files. Data reside in `data/`, with filename prefixes `S01_`–`S12_` for S1–S12. Readable tables support result inspection, while machine-readable CSV/JSON files support numerical verification. Each record represents a statistical item; record counts do not denote independent experiments.

Electronic supplementary material index. S1 and S2 each include readable tables and full-precision JSON; the remaining entries correspond to machine-readable result tables.

Table E.1: Electronic supplementary material index.

ID	Material	Records	Coverage
S1	BT unit-level results and full-precision generation statistics	200	Five-unit breakdown of 40 control settings; also includes full-precision macro-averages and generation statistics
S2	CE-only conditional intervals	40	Three resampling intervals for control differences across observation sets, scoring rules, and pools; full-precision interval data included
S3	Sequence-scheme control summaries	120	40 macro-averages each for BT, CSCL, and CE-only, including real and reference Top-1 and target bits
S4	CSCL and CE-only unit-level results	400	Two observation sets, two scoring rules, two candidate pools, and all controls
S5	Frozen-representation unit-level results	60	Unit-level scores for three representations, two observation sets, and two candidate pools
S6	BT and CSCL conditional intervals	1,920	Intervals for each metric under participant, text, and crossed resampling
S7	CSCL-minus-CE-only difference intervals	480	Three resampling intervals for between-scheme differences in absolute performance and input gains
S8	Five-cell, three-seed results	1,200	Three-seed results and dispersion across conditions for the five Q2 cells
S9	Training and checkpoint selection	60	Training and validation summaries for 35 CSCL and 25 CE-only runs
S10	Epoch-level training NLL	870	Epoch-level cross-entropy training records for 25 CE-only and four CSCL seed runs
S11	Local-holdout unit-level results	36	Five-unit results and equal-weight macro-averages for six local models
S12	Layerwise response statistics	17,280	Summary statistics by unit, input condition, and network layer for four comparison groups; descriptive comparisons only

Top-1 fields `real_top1`, `reference_top1`, and `full_top1` use proportions on a 0–1 scale. The suffix `_pp` denotes percentage points, and `_pct` on accuracy intervals denotes percentages. Other units are specified in field names; seed results retain the units of their respective metrics. In local-holdout files, `delta_top1` uses a 0–1 scale, and `MACRO_EQUAL_UNITS` denotes the five-unit equal-weight macro-average, not an additional data unit. Results from different observation sets cannot be directly subtracted.

Empty fields indicate unavailable statistics and are not treated as zeros. Macro-average records in local-holdout files do not contain unit-level proportions of positive and negative observations. Layerwise-response files provide finite-value sample counts and summary statistics for each side but lack the detail needed to reconstruct row-level finite-value pairs. Differences between the two sides' medians therefore cannot be interpreted as medians of row-level paired differences. Appendix A.5 specifies the conditioning and interpretation of intervals.

The supplementary package's `docs/FIELD_GUIDE.md` and `docs/STATISTICAL_NOTES.md` explain fields, units, support sets, and interval interpretation. From the extracted directory, `pythonscripts/verify.py` checks file integrity and numerical correspondence across tables; `pythonscripts/show_main_results.py` extracts the six comparisons and intervals in main-text Table 1. Both tools use only stored package values, require neither network access nor a GPU, and perform no model training, inference, or interval resampling.

Code and fixed protocol assets are included in the same archive under `code/`, without a separate external download. After installing NumPy, `python -B code/benchmark.py demo` checks fixed splits, candidate identities, and scoring rules and recomputes the 22,444 stored BP832 candidate scores. `docs/QUICKSTART.md` provides offline-check commands, and `docs/REPRODUCTION.md` provides commands for real-data preparation, BT/CSCL/CE-only training, fixed-checkpoint evaluation, and result aggregation. For the frozen-representation track, `code/frozen.py` provides entry points for data preparation, representation extraction, text embedding, and covariate-reference construction; `code/run_grid.py` schedules adapter fitting, scoring, and aggregation across 25 cells. Detailed usage appears in `docs/INDEPENDENT_USE.md`. `code/frozen_adapter.py` additionally exposes an alignment-fitting and prediction interface for prepared arrays. Real-data runs require users to supply the original data and local pretrained models; raw EEG, complete stimulus texts, and model weights are not distributed with the package.

F Reference Tables

Table F.1: Dataset composition. NR, SR, and TSR denote normal reading, sentiment reading, and task-specific reading, respectively.

Dataset–task unit	Stimulus instances	Observations
ZuCo 1.0 NR	300	3,064
ZuCo 1.0 SR	400	4,153
ZuCo 1.0 TSR	407	4,135
ZuCo 2.0 NR	349	5,371
ZuCo 2.0 TSR	390	6,010
Total	1,846	22,733

Table F.2: Evaluation tracks.

Track	Sequence: BT, CSCL, CE-only	Frozen representations: band features, CBraMod, LaBraM
EEG input	Word-level band features (GD840), length, and masks	Sentence-level band features (BP832), or adapted sentence-level signals
Trainable components	EEG encoder, projection, BART	Ridge-regression text-embedding adapter and score combination; EEG encoder frozen
Prediction score	Candidate token log-likelihood	Candidate cosine scores combined with the covariate reference
Non-EEG reference	Independent BART without task fine-tuning; a single EOS token as source	Candidate linguistic features and observation covariates, including eye-tracking measures and stimulus index
Main reporting support	NATIVE: 22,609; eligible donors: 21,847	COMMON: 22,378

Table F.3: Generation diversity and input agreement.

Scheme	Distinct strings	Empty outputs	Real/zero token-identical (%)
BT	42	2,878	81.51
CSCL	49	0	93.34
CE-only	95	0	73.20