WHOSE BRAIN IS IT? READER IDENTITY ACCOUNTS FOR ALMOST ALL OF EEG-TO-TEXT RETRIEVAL, AND A PER-READER ADAPTER RECOVERS IT

Anonymous authors
Paper under double-blind review

ABSTRACT

EEG-to-text systems are evaluated on splits that hold out sentences but keep every reader in training, so the reported numbers describe a new sentence read by a known person. We evaluate the setting a deployment would face, a reader absent from training. Tiling all 30 ZuCo readers into five disjoint folds under duration-matched candidate pools, retrieval on a held-out reader is 0.0458, $1.10\times$ chance and not distinguishable from it, against 0.1090, $2.62\times$ chance, for readers present in training on identical pools: 94% of the above-chance margin is reader identity rather than language. The same protocol on ChineseEEG, a different language, paradigm and laboratory, loses 96%, and no label-free normalisation moves it. We then close the gap. Freezing the reader-agnostic encoder and giving each new reader a residual linear remap of the channel axis, at most $16{,}384$ parameters and exactly the identity at initialisation, reaches the seen-reader ceiling on ZuCo from 400 labelled sentences and beats full fine-tuning at matched labels on ZuCo and on clause-level ChineseEEG. Four controls isolate the mechanism: matching the learning rate does not reproduce it, a shared readout with no per-reader parameters is the worst arm, the adapter on a random frozen encoder stays at chance, and a transplanted adapter loses the entire gain. Pre-training and calibration are separately near-worthless and jointly super-additive. Both halves survive a 4 Hz high-pass; the advantage of ZuCo's best electrode subset, which is entirely frontal, does not, a confound we report against our own recommendation.

1 INTRODUCTION

Decoding language from non-invasive EEG is reported as an open-vocabulary capability (Wang & Ji, 2022; Duan et al., 2023; Liu et al., 2025). A line of critiques has since shown that much of that performance is an artefact of how the systems are evaluated rather than of what they decode: scores collapse without teacher forcing (Jo et al., 2024), generation metrics reward fluent priors over anything read off the brain (Jo et al., 2026), and trial duration alone retrieves the right sentence above chance (Zhang et al., 2026). Those critiques leave a residue. Sentence *retrieval* from a duration-matched candidate pool survives all three, and under that protocol a small encoder does reach roughly $2.6\times$ chance. Something is being decoded. This paper asks what.

The answer is mostly the reader. Every split in this literature holds out sentences while keeping all subjects in the training set, so a reported number describes a new sentence read by a person the model already has data from. That is the easy direction and it is not the direction a deployment runs in, where a new user arrives with no data at all. We therefore tile all 30 ZuCo readers into five disjoint leave-six-out folds, so that each test trial is scored exactly once by a model that never saw its reader, and find that retrieval on an unseen reader is not distinguishable from chance. Ninety-four percent of the above-chance margin was reader identity, not language. The collapse is not a data-volume artefact, it reproduces on three unrelated encoder architectures, and it reproduces in direction, at lower signal-to-noise, on a published system (Liu et al., 2025) retrained under the same protocol, so it is a property of the problem, not of our model.

No label-free intervention we tried moves it. Per-reader spatial whitening (He & Wu, 2020) demonstrably does what it claims to the input statistics and changes retrieval by -0.0004. Whatever makes

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

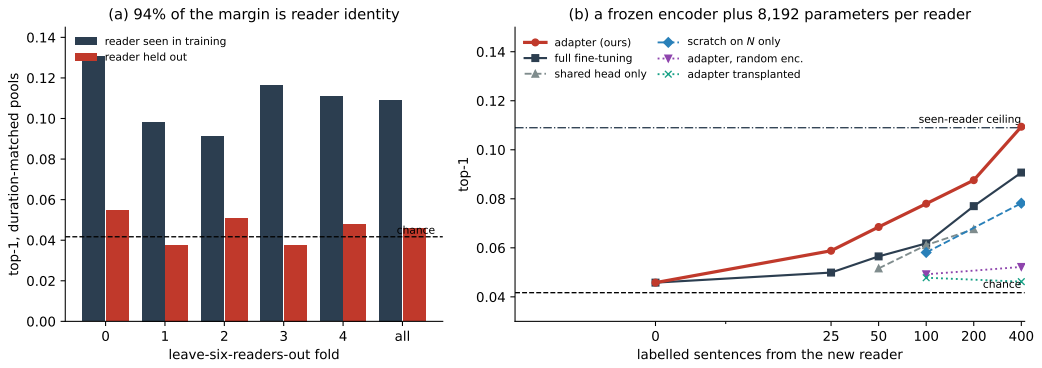

Figure 1: **(a)** Every one of the 30 readers is held out exactly once. Against readers present in training, on identical duration-matched pools, a held-out reader is at chance. **(b)** The gap is closeable with labels from the new reader, and where the parameters go decides how cheaply. The adapter reaches the seen-reader ceiling at 400 sentences; the three lower dotted arms are controls that remove the obvious alternative explanations.

a reader legible is not a stationary linear property of their signal, so the recovery has to be paid for with labels from that reader.

It can be. The second half of the paper closes the gap and shows that *where* the calibration parameters go matters more than how many there are. Freezing the reader-agnostic encoder and giving each new reader a residual linear remap of the channel axis, either the full $C \times C$ matrix or its rank-32 factorisation, $16{,}384$ or $8{,}192$ parameters and exactly the identity at initialisation, reaches the seen-reader ceiling from 400 labelled sentences, and on ZuCo beats fine-tuning the entire 1.0M-parameter encoder at every calibration budget we tested, on all three architectures. Four controls remove the obvious alternative readings: the same adapter on a randomly initialised frozen encoder stays at chance, so the adapter is not secretly the decoder; a shared readout with no per-reader parameters is the worst arm, so the win is per-reader rather than parameter-efficient; and transplanting a fitted adapter onto a different reader destroys the entire gain, so nothing it learns is shared.

That last set of controls forces a conclusion we did not expect and which we state against our own initial framing. Pre-training on other readers is worth almost nothing by itself, and training from scratch on a few hundred trials from the right reader is worth surprisingly much: $2{,}400$ trials from the target readers beat $14{,}000$ trials from other people. The two are not decoder-plus-index; they are super-additive, each roughly 70% larger together than the sum of their separate effects.

Contributions.

1. A reader-holdout evaluation of EEG-to-text retrieval that tiles all 30 ZuCo readers, under which retrieval is at chance and 94% of the reported margin turns out to be reader identity. It reproduces on three encoder families, in direction on a published system, and on a second corpus in another language and presentation paradigm, where 96% is lost.
2. Evidence that the gap is not recoverable label-free: three input normalisation schemes, including Euclidean alignment, all leave it intact.
3. A per-reader spatial adapter fitted after training against a frozen encoder, which recovers the ceiling from 400 labelled sentences and beats full fine-tuning at matched labels on ZuCo and on clause-level ChineseEEG, never losing to it in any configuration, with four controls isolating why it works. On the second corpus the curve runs to $N = 1600$ and is still climbing, which settles that the ZuCo tie at 400 is the ceiling, not the adapter, saturating.
4. A quantitative account of what this costs to collect: the calibration budget a new user must supply, and a channel study finding that 16 well-chosen electrodes match a 128-channel research net, together with the control showing that the subset achieving this is entirely frontal and its advantage is confined to frequencies below 4 Hz, which disqualifies it as a recommendation and is why we report it.

5. An ocular control on the main result. The reader effect and the adapter recovery are unchanged under a 4 Hz high-pass, so they are not the low-frequency frontal component that the montage experiment exposes.

2 Related work

EEG-to-text systems and their critiques. Wang & Ji (2022) introduced open-vocabulary EEG-to-text generation on ZuCo (Hollenstein et al., 2018; 2020) by mapping word-level EEG features into a pre-trained sequence-to-sequence decoder; DeWave (Duan et al., 2023) added a discrete codebook, and GLIM (Liu et al., 2025) replaced the decoder with an instruction-tuned language model trained contrastively against sentence embeddings. All three report sentence-held-out splits with every subject present in training; none reports an unseen-subject number, and we are not aware of any EEG-to-text result that does. Jo et al. (2024) showed that the generation scores depend on teacher forcing and that pure-noise input scores comparably; Jo et al. (2026) showed the metrics are largely explained by the language model's prior; Zhang et al. (2026) showed that a candidate pool that is not duration-matched can be solved without reading the EEG. We adopt their duration-matched protocol throughout, which is why our numbers are lower than published ones and why they mean something. Those critiques concern the *measurement of language*; the confound we report concerns the *measurement of generality*, and the two are independent: our protocol is already duration-matched when the reader holdout removes the remaining signal.

Subject transfer and per-subject layers. That EEG does not transfer across subjects is old news outside this literature. Euclidean alignment (He & Wu, 2020) whitens each subject's trials by their reference covariance and is the standard label-free remedy in motor imagery and P300; we evaluate it here and it does not help. Défossez et al. (2023) decode perceived speech from M/EEG with a per-subject 1×1 convolution M_s over the channel axis, jointly trained with the encoder over all subjects. Our adapter is architecturally close to M_s; the difference is the regime. Those subject layers are trained jointly, with every subject present, to exploit inter-subject variability. Ours is fitted *after* training, for a subject the encoder has never seen, against a frozen encoder, from a few hundred labelled trials. The contribution is not the module. It is (i) the measurement that makes the module necessary, (ii) the demonstration that this is where the parameters must go rather than merely a cheap approximation to fine-tuning, and (iii) the controls showing that three label-free alternatives all fail. The residual low-rank parameterisation is LoRA's (Hu et al., 2022), applied to the input channel axis with the same asymmetric initialisation. That a small well-placed set of parameters can beat full fine-tuning is familiar from language models; what is not familiar is that the constraint here is not compute but *data*. A new reader can realistically supply a few hundred labelled sentences, and at that scale the extra freedom of full fine-tuning only buys overfitting.

3 Setup

Data. ZuCo 1.0 and 2.0 (Hollenstein et al., 2018; 2020): 22,335 sentence-level trials, 30 readers, 128-channel EGI HydroCel net resampled to 128 Hz and padded to 1280 samples. ZuCo retains 105 of the 128 sensors and the released preprocessing drops the flat reference row and zero-pads back to 128, so exactly 104 columns carry data and the column index is *not* the electrode number. This matters only for the montage study; we recover the mapping from the published homologous-pair list and validate it against the measured correlation-versus-distance falloff (Appendix G).

Task and protocol. Sentence retrieval from a candidate pool. Following the audit protocol, pools are matched on trial duration within task, 24 candidates per pool, chance 0.0417; without duration matching, trial length alone retrieves above chance. All comparisons are paired bootstraps over identical pools with 20,000 resamples.

Model and optimisation. A 1.0M-parameter encoder trained with InfoNCE (van den Oord et al., 2018) against a frozen 384-dimensional MiniLM sentence encoder (all-MiniLM-L6-v2; Reimers & Gurevych, 2019). A stem convolves the channel axis to width 128 with kernel 17 and stride 8, followed by batch normalisation and GELU; two pre-norm transformer encoder layers with 4 heads and a 512-wide feed-forward block follow, then masked mean+std pooling over valid frames and a

arm	top-1	× chance	
unseen reader	0.0458	1.10	($p = 0.074$)
seen reader, identical pools	0.1090		2.62

Table 1: Five folds, all 30 readers, 155 duration-matched pools, chance 0.0417. 94% of the above-chance margin is reader identity. Two of five folds fall below chance; fold 0 alone would have given $1.31\times$ ($p = 0.011$), which is why the full tiling is necessary.

two-layer projection to 384 dimensions, with dropout 0.3 throughout. The mask is the only route by which trial duration is visible to the model, and it is left exactly as recorded. The MLP and LSTM arms of Table 4 replace the two transformer layers with a two-layer MLP over the pooled features and with a single bidirectional LSTM of width 128; everything else is shared. AdamW, weight decay 10^{-2}, one-cycle schedule with 10% warmup, batch size 64, 40 epochs at peak learning rate $3 \cdot 10^{-4}$ for base models. Calibration runs train for 20 epochs, at $3 \cdot 10^{-3}$ for the adapter and 10^{-4} for full fine-tuning; both learning rates were tuned per arm, and the control in Table 3 rules out the resulting difference as an explanation. The InfoNCE temperature is learned, initialised at 0.07. Every arm reported here is three seeds. A base run is about five minutes on one GH200; the complete set of experiments behind this paper is roughly 1,000 runs.

Reader holdout. Five disjoint folds of six readers tile all 30, so each of the 2,227 test trials is scored exactly once by a model that never saw its reader; the 155 per-fold pools are concatenated. Three seeds per arm.

4 Holding out the reader removes almost all of the signal

Not a data-volume artefact. The held-out model trains on 24 readers, 20% fewer trials. Scored on the 24 readers it *did* train on, it is indistinguishable from the full model (0.0958 vs. 0.0957, [-0.0058, +0.0059]). Dropping a fifth of the trials costs nothing; dropping the reader costs everything.

No label-free normalisation recovers it. Per-reader spatial whitening (He & Wu, 2020) does what it claims, since the between-reader spread of per-trial scale falls from $4.71\times$ to $1.86\times$, and retrieval on unseen readers still does not move: 0.0453 (-0.0004 [-0.0063, +0.0054] vs. unaligned). Per-reader z-scoring gives 0.0477 ($+0.0019$ [-0.0037, +0.0073]) and cross-subject view alignment 0.0465 ($+0.0007$ [-0.0042, +0.0056]); all three are inside each other's intervals and inside chance (Appendix A). Whatever makes a reader legible to the encoder is not a stationary linear property of their input, so the recovery has to be paid for with labels.

5 Method: a frozen encoder and a per-reader spatial residual

Figure 2 lays out the two stages. Write $x \in \mathbb{R}^{C\times T}$ for a trial from reader s. The adapter applies

$$x \;\leftarrow\; x + U_s V_s x, \qquad U_s \in \mathbb{R}^{C\times r},\;\; V_s \in \mathbb{R}^{r\times C}, \tag{1}$$

before a frozen encoder. Following LoRA (Hu et al., 2022) we initialise V_s randomly and U_s at zero, so the residual is exactly the identity at step 0 and the $N = 0$ point of the calibration curve is preserved bit-exactly, while the gradient is not. Initialising *both* factors at zero is a saddle ($\partial L/\partial U \propto V$, $\partial L/\partial V \propto U$) and the weights never move; we lost a sweep to this. Rows are indexed globally and the encoder is frozen, so readers in a batch are independent: one run calibrates six readers exactly as six separate runs would.

We use two instantiations. The full remap, $r = C$, is a single zero-initialised $C \times C$ matrix per reader, 16,384 parameters; it is the arm behind the ZuCo transformer curve and its controls below. The rank-32 factorisation is 8,192 parameters and is what we run on the other two architectures, under the ocular control and on ChineseEEG, having first checked on ZuCo that it tracks the full remap (Table 2, last column). Either way the adapter is under 2% of the encoder.

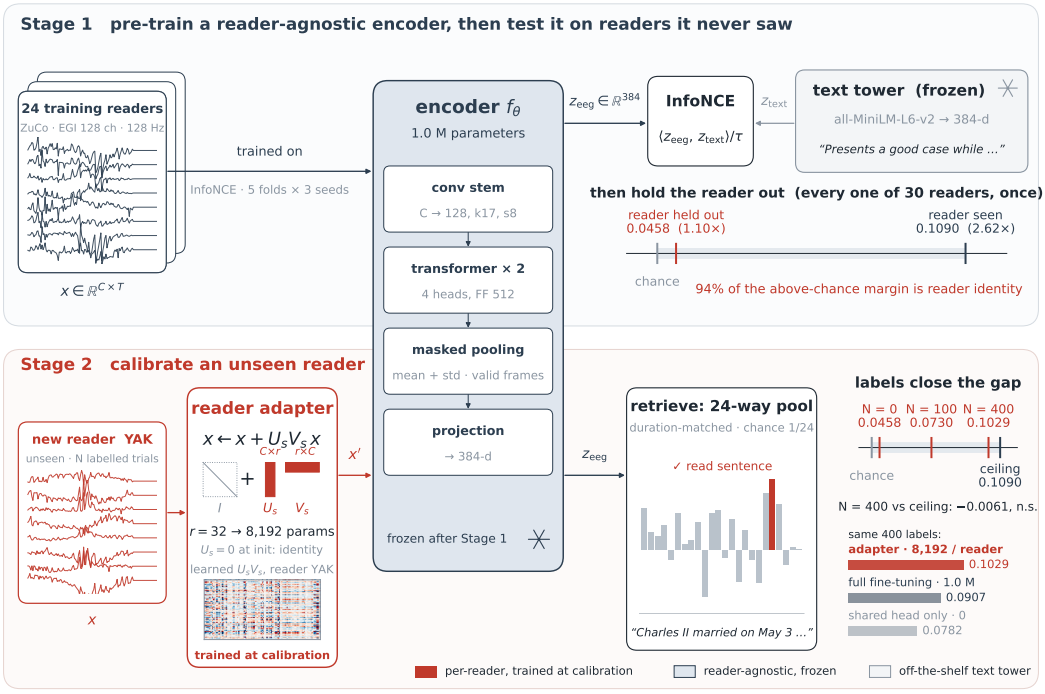

Figure 2: Overview. One encoder serves both stages. *Stage 1* trains it with InfoNCE against a frozen sentence encoder on the readers in the training folds, then scores it on readers it never saw: retrieval falls from 2.62× to 1.10× chance, so 94% of the margin was the reader. *Stage 2* gives an unseen reader a residual on the channel axis, $x \leftarrow x + U_s V_s x$ (drawn at rank 32), identity at initialisation, fitted from N labelled trials with the encoder frozen; at $N = 400$ it is indistinguishable from the seen-reader ceiling and beats fine-tuning the whole encoder on the same labels. Every element drawn from data is real: the traces are ZuCo trials, the heat-map is the learned $U_s V_s$ of a held-out reader after calibration, and the pool is an actual duration-matched 24-way pool scored by that checkpoint.

N / reader	FT, lr 10^{-4}	FT, lr $3{\cdot}10^{-3}$	adapter $r{=}C$	recovered	adapter $r{=}32$
0	0.0458	—	—	6%	—
25	0.0499 n.s.	—	0.0588	25%	0.0534
50	0.0565	0.0540	**0.0685**	40%	0.0641
100	0.0618	0.0624	**0.0780**	54%	0.0730
200	0.0770	0.0687	**0.0876**	68%	0.0866
400	0.0907	—	**0.1094**	101%	0.1029
seen-reader ceiling			0.1090	100%	

Table 2: Calibration curve, five folds, three seeds. At 400 sentences the full-rank adapter is indistinguishable from a model that had that reader from the start (+0.0004 [-0.0062, +0.0071]); so is the rank-32 one (−0.0061 [-0.0132, +0.0009], 91% of the margin). At 100 the adapter matches full fine-tuning at 200: the same accuracy for half the labels. Rank 32 beats full fine-tuning at every $N \geq 50$ (+0.0075, +0.0112, +0.0096, +0.0122, all significant) and is within 0.0065 of the full remap throughout.

6 Results

The curve and the controls. Table 2 is the calibration curve and Table 3 the four controls on it. Rank 32 tracks the full $C \times C$ remap: −0.0054, −0.0044, −0.0050, −0.0011, −0.0065 at $N = 25$ to 400, significant only at the two ends and never by more than 0.0065, while beating rank 16 at every N tested (+0.0065, +0.0069, +0.0092, all significant). The factorisation halves the parameters at a cost of at most a tenth of the margin, which is why it is the operating point everywhere except the ZuCo transformer curve itself.

control	alternative explanation it removes	result
full FT at the adapter's LR	"you raised the learning rate"	-0.0082 [-0.0139, -0.0026] at N=200
shared head, no per-reader params	"you froze the encoder"	worst arm at every N; -0.015 to -0.021
adapter on a *random* frozen encoder	"the adapter is the decoder"	$+0.0034$ n.s. vs. zero-shot, N=100
adapter transplanted to another reader	"it found a generic filter"	0.0462 / 0.0435 vs. 0.0458 uncalib.

Table 3: Each row removes one reading of the result. Raising full fine-tuning to the adapter's learning rate makes it *worse*. A shared readout with zero per-reader parameters is the weakest arm on all five folds, so the win is per-reader, not parameter-efficient. Transplanting a calibrated adapter onto a different reader destroys the whole gain, so nothing it learns is shared.

arm on unseen readers	transformer	MLP	LSTM
no calibration	0.0458 (1.10×)	0.0405 (0.97×)	0.0453 (1.09×)
adapter, N=50	0.0685	0.0584	0.0603
adapter, N=100	0.0780	0.0700	0.0623
adapter, N=200	0.0876	0.0824	0.0719
adapter, N=400	**0.1094**	**0.0998**	**0.0849**
full FT, N=400	0.0907	0.0785	0.0791
seen-reader ceiling	0.1090	0.0902	0.0851
margin lost by holdout	94%	102%	91%

Table 4: Same folds, same pools. The transformer column is the full-rank adapter of Table 2; the MLP and LSTM columns use the rank-32 adapter. Each ceiling is that architecture trained on all 30 readers. All three collapse to chance without the reader, climb monotonically in N, and return to their own ceiling at $N = 400$: a tie on the transformer ($+0.0004$, n.s.) and the LSTM (-0.0002, n.s.), and a pass on the MLP ($+0.0097$ [+0.0031, +0.0162]). The adapter beats full fine-tuning at matched labels on all three ($+0.0187$, $+0.0213$, $+0.0058$).

Transfer and labels are separately useless and jointly super-additive. Training the whole encoder from scratch on nothing but the new reader's N trials reaches $1.88\times$ chance at $N = 400$, with six times less data than the 14,272 trials the zero-shot model saw but from the right readers, and it wins by 0.0324. Above-chance margins decompose as follows: at $N = 100$, pre-training alone is worth $+0.0041$, labels alone $+0.0164$, both together $+0.0363$, which is 77% more than the sum of parts; at $N = 400$ the three numbers are $+0.0041$, $+0.0365$ and $+0.0677$, an excess of 67% (Appendix B). So the claim is not that pre-training carries the language while the adapter merely re-indexes it. Pre-training buys nothing on its own; it multiplies what calibration buys. The method also beats scratch-on-N at equal labels ($+0.0312$ at N=400, $+0.0200$ at N=100) using 16,384 parameters per reader instead of 1.0M.

Three encoder families, one curve. Nothing above is a property of attention. Repeating the whole protocol on a two-layer MLP over channel-pooled features and on a bidirectional LSTM, with the same text tower, pooling, head and objective (Table 4), all three collapse to chance without the reader (94%, 102% and 91% of the margin lost), climb monotonically in N, and return to their own ceiling at $N = 400$: a tie on the transformer ($+0.0004$, n.s.) and the LSTM (-0.0002, n.s.), and a pass on the MLP ($+0.0097$ [+0.0031, +0.0162]). The adapter beats full fine-tuning at matched labels on all three ($+0.0187$, $+0.0213$, $+0.0058$).

A published system collapses in the same direction. If the collapse were specific to our encoder it would be a curiosity. We therefore retrained GLIM (Liu et al., 2025) from scratch with each fold's six readers removed, in its published configuration, and scored its exported embeddings on our pools (Appendix C, Table 8). Its architecture, objective and decoder are all different from ours, and its published configuration drops the subject prompt with probability 1.0, so it carries no per-subject parameters and applies to an unseen reader unchanged. GLIM's margin is small to begin with ($1.44\times$ chance on readers it trained on, against $2.62\times$ for our transformer), so it has little to lose and the measurement is correspondingly noisy: no single fold is significant, four of five go the same way as our encoders, one fold's readers GLIM cannot retrieve for even when they are in the

N / unseen reader	adapter	recovered
0	0.0442 (1.06×)	4%
50	0.0491 (1.18×)	11%
100	0.0535 (1.28×)	18%
200	0.0617 (1.48×)	31%
400	**0.0765** (1.84×)	54%
800	**0.0921** (2.21×)	78%
1600	**0.0984** (2.36×)	88%
seen-reader ceiling	0.1065 (2.56×)	100%

Table 5: ChineseEEG, five folds, three seeds, chance 0.0417. Holding out the reader removes 96% of the above-chance margin (94% on ZuCo). Every step in N is significant against the uncalibrated arm (+0.0049, +0.0093, +0.0175, +0.0323, +0.0479, +0.0542; all CIs clear of zero) and the curve is monotone to $N = 1600$. Full fine-tuning of the whole encoder was run at $N = 400$ and $N = 1600$: 0.0587 and 0.0822, so the adapter leads by +0.0178 [+0.0151, +0.0205] and +0.0162 [+0.0133, +0.0191], against +0.0187 on ZuCo, and the adapter at $N = 800$ (0.0921) is above full fine-tuning at $N = 1600$. The two curves are overlaid in Figure 3.

training set, and the pooled effect is −0.0086 with a 95% interval [-0.0172, +0.0000] whose upper end is exactly zero, 47% of GLIM's above-chance margin. We read this as the same phenomenon at lower signal-to-noise, not as an exception to it, but the claim it licenses is weaker than the one we make for our own encoders: five folds of GLIM establish that reader dependence is not an artefact of our architecture or training recipe, not its magnitude in a system whose seen-reader performance is itself close to chance.

The effect is not ocular. ZuCo is natural reading with simultaneous eye tracking, and the critiques of this literature are about exactly this class of confound. Our own montage experiment (Section 7) shows the concern is not hypothetical: the electrode subset that best preserves retrieval is entirely frontal and its advantage vanishes above 4 Hz. ZuCo's nine designated EOG electrodes are not among the 104 retained columns, so frequency is the available handle: ocular artefact is concentrated below about 4 Hz, so we rerun the spine of the paper on a 4 Hz-high-passed signal, applied identically at training, validation and export (Appendix D, Table 9). The filter costs the seen-reader ceiling 0.0130 (0.1090 → 0.0960); holding out the reader still removes 92% of the margin (0.0461, 1.11×), and the adapter at $N = 400$ still reaches the ceiling (0.0936 vs. 0.0960, −0.0023 [-0.0085, +0.0039], n.s.). Both findings therefore stand, and they are separable: a real part of ZuCo's *spatial* structure is a low-frequency frontal component of probably ocular origin, which is what invalidates the electrode recommendation below, and the reader effect is not that component.

6.1 A SECOND CORPUS, ANOTHER LANGUAGE

Everything above is one corpus, and a reader could reasonably ask whether the collapse is a property of ZuCo, of its 30 readers, its English, or its self-paced presentation. We therefore ran the same protocol on ChineseEEG (Mou et al., 2024): 10 participants silently reading two novels while each character is highlighted at a fixed rate, recorded in a different laboratory on the same 128-channel EGI net. The released segmentation is display lines of at most ten characters, so we rebuild the character stream from the row markers and cut it into clauses on Chinese clause punctuation; clauses average 4.1 s against ZuCo's 4.43 s, so the two corpora have nearly the same trial-duration distribution and duration matching means the same thing in both. Everything else is held fixed: the encoder, the objective, the rank-32 residual, the scoring code, and a 384-d text tower from the same MiniLM family, the multilingual paraphrase model. Five disjoint folds of two readers tile all ten, giving 9,454 test trials in 761 duration-matched pools within session, and because each reader contributes about 10,000 trials the calibration curve can be traced to $N = 1600$, four times further than ZuCo allows.

Table 5 gives the result. The collapse replicates at the same magnitude: 96% of the above-chance margin is lost with the reader, against 94% on ZuCo. So does the repair, in shape and in its relation to the control. Every step in N buys a significant gain, the curve is monotone with no plateau to $N = 1600$, and the adapter beats full fine-tuning of the whole encoder at matched labels by +0.0178 at $N = 400$, where the ZuCo contrast was +0.0187. The gap to full fine-tuning does not close with more labels (+0.0162 at $N = 1600$), so the adapter's advantage is not a small-data regime that

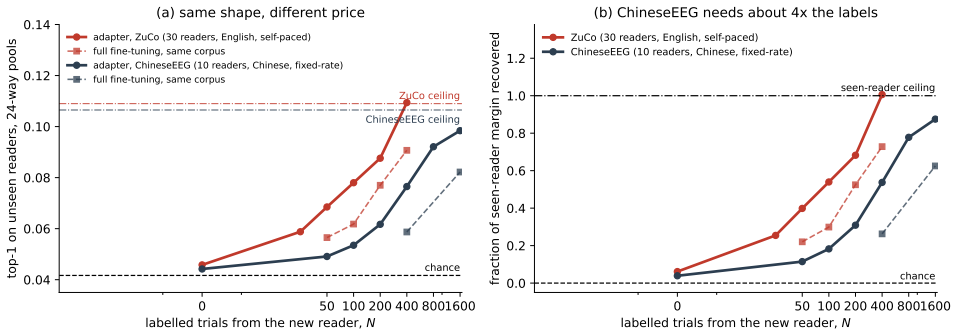

Figure 3: The calibration curve on two corpora. **(a)** Top-1 on unseen readers against labelled trials, each corpus with its own seen-reader ceiling (dash-dot) and full fine-tuning at matched labels (dashed). **(b)** The same curves as a fraction of the seen-reader margin. The shapes agree; the price does not. ChineseEEG needs roughly four times the labels for the same fraction of the margin, its $N = 400$ sitting on ZuCo's $N = 100$, and at $N = 1600$ it has not reached the ceiling.

the full model grows out of. The controls transfer as well (Appendix E.1, Table 10): the shared readout is the worst trained arm at both budgets, the same adapter on a randomly initialised frozen encoder does not leave chance at $N = 400$, and the adapter calibrated on one held-out reader is worth nothing to the other. Going past $N = 400$ also settles a question the ZuCo curve had to leave open: the rank-32 adapter's 8,192 parameters keep buying accuracy to $N = 1600$, so the ZuCo tie at 400 was the ceiling being low, not the adapter saturating.

What does differ is the price. ChineseEEG needs about four times the labels for the same fraction of the margin, and at $N = 1600$ the adapter is still 0.0081 [-0.0112, -0.0050] below the seen-reader ceiling. One candidate explanation is the trial unit: clause trials are cut from a continuous fixed-rate stream at commas, and a comma is not an event the reader experiences in the way a self-paced ZuCo sentence is. We tested it by rebuilding the cache at sentence level (terminal punctuation only; the 36% of sentences over the 10 s window are dropped, leaving 18,893 trials of mean 5.5 s, about 1,900 per reader) and rerunning the collapse and the curve to $N = 400$ (Appendix E.2, Table 11). Segmentation is part of the answer and not all of it. At sentence level the seen-reader ceiling rises to ZuCo's (0.1098, 2.63×), the adapter at $N = 400$ recovers 70% of the margin against 54% on clauses, and the collapse itself is less complete, 80% of the margin lost rather than 96%, with a zero-shot arm that stays above chance (1.33×, CI clear of it). The cleaner unit helps full fine-tuning more than it helps the adapter (26% → 63% against 54% → 70%), so the adapter's lead at $N = 400$ shrinks to $+0.0042$ [-0.0020, +0.0103] and is no longer significant; the claim that survives every configuration is adapter ≥ full fine-tuning at matched labels, with the strict inequality significant on ZuCo and on clause-level ChineseEEG. But 70% is not 101%: the remaining gap survives the change of unit and belongs to something else about the corpus: the fixed-rate paradigm, the language, the multilingual text tower, or the smaller reader pool, which this design cannot separate. We therefore state the conclusion conservatively. The failure is not a property of ZuCo, its readers, its language or its paradigm: it reappears in a different language, reading paradigm and laboratory, and the per-reader residual recovers it in every configuration from a few hundred labelled trials with under 2% of the encoder's parameters. How many labels buy a given recovery depends on the corpus and on how it is segmented, and should not be quoted as a universal number.

7 What this means for data collection

Report the unseen-reader number, and budget a calibration block. A seen-reader score on its own does not bound what the system will do for a new user, and the two differ by the whole effect; the five-fold tiling used here costs five training runs and needs no extra data. A new reader needs on the order of 100–400 labelled sentences before they are legible, and that block has to be designed into the session, because the calibration sentences must be disjoint from the evaluation sentences to mean anything. In ZuCo terms 400 trials is roughly a fifth of what a reader contributes: a real but not prohibitive tax.

Electrodes: placement beats count, and the best placement is an artefact. Restricting the input to candidate montages under the seen-reader protocol (Appendix F, Table 12) gives a result more useful than "more channels are better". Doubling from 16 to 32 electrodes is worth $+0.0027$ [+0.0000, +0.0055], on the boundary of significance, whereas *which* 16 are chosen moves top-1 over a range of 0.0107 [+0.0050, +0.0165] across five random placements, so placement is worth about four times count. The standard 10-20 positions sit at or below the bottom of the random range at both sizes, and not because they land on dead electrodes: their mean amplitude is higher, because 10-20 deliberately samples the scalp periphery where ocular and muscular artefact lives. Ranking electrodes by a group lasso on the stem's input-channel weights, one group per electrode, then retraining from scratch on the selection, 16 chosen electrodes reach 0.0977 and 32 reach 0.1032 against 0.1035 for all 104, both indistinguishable from the full research cap, and $+0.0137$ [+0.0079, +0.0197] and $+0.0174$ [+0.0121, +0.0227] above random placements of the same size. Then the control. All 16 selected electrodes lie anterior to the C_z coronal plane, including Fp_1/Fp_2 and F_7/F_8, the belt where blinks, saccades and temporalis artefact live. The same ranking restricted to the 54 live columns at or behind C_z gives 0.0781, *below* pooled random (-0.0059 [-0.0108, -0.0009]). Under a 4 Hz high-pass the anterior advantage over posterior falls from $+0.0196$ [+0.0116, +0.0276] to $+0.0029$ [-0.0033, +0.0090], and the posterior montage gets *better* by $+0.0078$ [+0.0019, +0.0136]. The entire placement advantage was a sub-4 Hz frontal component, and we retract the recommendation it implies: what remains is that 16 channels suffice and that, once that band is removed, placement barely matters (anterior 0.0887, posterior 0.0859, n.s., against 0.0918 for all 104). Two consequences for the next corpus, both cheap: record dedicated EOG and keep it in the release, and report a high-passed number beside the unfiltered one, because the filter costs 19% of the margin here and that is precisely the part a sceptical reader will attribute to the eyes.

8 Limitations

- Two corpora, both on the same EGI net, and the second is not a full replication of the first: the learning-rate control, the rank ablation, the label-free baselines and the other two encoder families were run on ZuCo only. The corpora differ in language, paradigm, laboratory, reader count and trial unit at once; of these we could vary only the trial unit, which explains about a third of the difference in label cost.
- The GLIM comparison is borderline, a pooled -0.0086 with an interval ending at zero, limited by GLIM's own small seen-reader margin.
- Retrieval from a 24-way pool is a weaker claim than generation; we adopt it because it is the formulation that survives the existing critiques (Jo et al., 2024; 2026).
- The ZuCo curve stops at $N = 400$ and the ChineseEEG curve is still climbing at $N = 1600$, so we do not know what the adapter converges to, or why the two corpora price a unit of recovery so differently.
- The montage study restricts a research net in software, and the ocular control separates ocular from cortical by frequency because ZuCo does not release its EOG channels. Why reader identity carries the signal is not answered here; the controls establish only that it is neither a stationary linear property of the input nor shared across readers.

9 Conclusion

Under the evaluation this field uses, EEG-to-text retrieval reaches $2.6\times$ chance; under the evaluation a deployment would face, where the reader is new, it reaches chance. The difference is 94% of the reported margin on ZuCo and 96% on ChineseEEG, it survives three encoder architectures and a change of language and reading paradigm, it reproduces in direction on a published system, and no label-free normalisation touches it. The gap is closeable and closing it is cheap: 8,192 to 16,384 parameters on the channel axis of a frozen encoder recover the seen-reader ceiling from 400 labelled sentences on ZuCo and never lose to fine-tuning the whole encoder at matched labels, beating it significantly on ZuCo and on clause-level ChineseEEG. Pre-training on other readers and calibration on the target reader are each worth little alone and are super-additive together, so the question a collection protocol has to answer is not how many readers to record but how much of each. The most striking number we produced, 16 electrodes matching a 128-channel net, is the one we retracted, because a 4 Hz high-pass removed its advantage while leaving the reader effect and the adapter untouched.

Reproducibility statement

All numbers in this paper come from the authors' own runs; none is transcribed from prior work. Section 3 gives the data preparation, model, optimisation and fold construction; Appendix G documents the ZuCo column-to-electrode map, including the error it corrects and the validation that catches it; Section 6.1 and Appendix E.2 give the ChineseEEG segmentation at both trial units. Every reported comparison is a paired bootstrap over identical duration-matched pools with $20{,}000$ resamples and three training seeds per arm. Code for the cache builders, training, the adapter, all controls, the GLIM retraining harness and the scoring scripts, with the exact commands behind every table, is provided as anonymised supplementary material. Both corpora are public (ZuCo 1.0/2.0; ChineseEEG, OpenNeuro ds004952).

Ethics statement

This work uses two publicly released EEG corpora recorded under the original authors' ethics approvals with participant consent; no new human data were collected and no participant is identifiable from the released signals or from anything reported here. The paper's central finding is that a published capability is largely reader-specific and does not hold for a new user without calibration; we regard reporting this as a safety-relevant correction for any assistive or clinical use of EEG-to-text decoding, not as a capability gain. The electrode-placement result is reported together with the control that disqualifies it, precisely so that it is not adopted as a hardware recommendation. We declare no conflicts of interest.

AI use statement

Generative AI tools were used throughout this work: to draft and refactor the data-processing, training, evaluation and analysis code; to help design and run experiments under the authors' direction; to perform statistical analyses; for retrieval and discovery of related work; and to draft and edit the text, including sections of the paper. We have not used generative AI tools to generate data or to prove mathematical claims. All code was tested by the authors, including regression checks where existing results were reused; all numbers were regenerated from the authors' own runs; all citations were verified against the cited documents; and all claims were checked by the authors against the underlying results. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

References

Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5: 1097–1107, 2023. doi: 10.1038/s42256-023-00714-5.

Yiqun Duan, Jinzhao Zhou, Zhen Wang, Yu-Kai Wang, and Chin-Teng Lin. DeWave: Discrete EEG waves encoding for brain dynamics to text translation. In *Advances in Neural Information Processing Systems*, 2023. arXiv:2309.14030.

He He and Dongrui Wu. Transfer learning for brain-computer interfaces: A Euclidean space data alignment approach. *IEEE Transactions on Biomedical Engineering*, 67(2):399–410, 2020. doi: 10.1109/TBME.2019.2913914.

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.

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, Marseille, France, 2020. European Language Resources Association.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. arXiv:2106.09685.

Hyejeong Jo, Yiqian Yang, Juhyeok Han, Yiqun Duan, Hui Xiong, and Won Hee Lee. Are EEG-to-Text models working? *arXiv preprint arXiv:2405.06459*, 2024.

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

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

Xinyu Mou, Cuilin He, Liwei Tan, Junjie Yu, Huadong Liang, Jianyu Zhang, Yan Tian, Yu-Fang Yang, Ting Xu, Qing Wang, Miao Cao, Zijiao Chen, Chuan-Peng Hu, Xindi Wang, Quanying Liu, and Haiyan Wu. ChineseEEG: A chinese linguistic corpora EEG dataset for semantic alignment and neural decoding. *Scientific Data*, 11:550, 2024. doi: 10.1038/s41597-024-03398-7.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019. arXiv:1908.10084.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Zhenhailong Wang and Heng Ji. Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022. arXiv:2112.02690.

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.

A Label-free normalisation arms

Table 6 lists the three label-free arms of Section 4, scored on the same held-out folds and pools as the main result.

arm on unseen readers	top-1	× chance	vs. unaligned
unaligned	0.0458	1.10	—
per-reader spatial whitening	0.0453	1.09	−0.0004 [-0.0063, +0.0054]
per-reader z-scoring	0.0477	1.14	+0.0019 [-0.0037, +0.0073]
cross-subject view alignment	0.0465	1.12	+0.0007 [-0.0042, +0.0056]

Table 6: Three label-free arms, all inside each other's intervals and inside chance. Spatial whitening is Euclidean alignment (He & Wu, 2020) computed per reader from that reader's own trials, so it uses no labels.

B Super-additivity of pre-training and calibration

Table 7 decomposes the unseen-reader margin at $N{=}100$ and $N{=}400$ into what pre-training alone and labels alone buy, and what the two buy together.

N	pre-training alone	labels alone	both	sum of parts	excess
100	+0.0041	+0.0164	+0.0363	0.0205	+77%
400	+0.0041	+0.0365	+0.0677	0.0406	+67%

Table 7: Above-chance margins on unseen readers. "Pre-training alone" is the zero-shot held-out model; "labels alone" is the whole encoder trained from scratch on the new reader's N trials; "both" is the adapter on the frozen pre-trained encoder. Neither ingredient works alone.

C GLIM UNDER READER HOLDOUT

GLIM (Liu et al., 2025) was retrained from scratch in its published configuration (60 epochs, batch 72) with each fold's six readers removed from training and validation, and its exported test embeddings were scored on the same duration-matched pools as every other arm. The reference is GLIM trained on all 30 readers at the same budget, so the only difference between the two arms is which readers were available. Row alignment between the exported embeddings and the pool table is asserted, not assumed.

GLIM, five folds	seen	held out	held out − seen	pools
fold 0	0.0703	0.0508	−0.0195 [-0.0352, -0.0039]	32
fold 1	0.0574	0.0462	−0.0113 [-0.0293, +0.0068]	37
fold 2	0.0597	0.0556	−0.0042 [-0.0278, +0.0208]	30
fold 3	0.0417	0.0517	+0.0100 [-0.0100, +0.0300]	25
fold 4	0.0672	0.0538	−0.0134 [-0.0309, +0.0027]	31
pooled	0.0599 (1.44×)	0.0513 (1.23×)	−0.0086 [-0.0172, +0.0000]	155

Table 8: GLIM retrained without each fold's six readers, scored on the same pools as Table 1. Four folds of five go the same direction as our encoders; fold 3's readers are the ones GLIM trained on all 30 cannot retrieve for either way (seen arm at chance, 1.00×). Pooled, holding out the reader removes 47% of GLIM's above-chance margin, with the interval ending at zero.

D THE 4 HZ HIGH-PASS CONTROL

Table 9 reports the seen-reader ceiling, the reader holdout, the adapter recovery and the montage ranking after a 4 Hz high-pass applied at training, calibration and export.

arm on unseen readers	unfiltered	4 Hz high-pass
held out, no calibration	0.0458 (1.10×)	0.0461 (1.11×)
adapter, N=400	0.1094 (2.63×)	0.0936 (2.25×)
seen-reader ceiling	0.1090 (2.62×)	0.0960 (2.30×)
margin lost by holdout	94%	92%
adapter vs. ceiling	+0.0004 n.s.	−0.0023 [-0.0085, +0.0039] n.s.

Table 9: Both halves of the paper survive the control. The filter is applied identically at training, validation and export; the high-passed adapter arm is rank 32. The filter costs the ceiling 0.0130 and moves the collapse by two points; the adapter still reaches the ceiling at $N = 400$.

E CHINESEEEG: THE CONTROLS AND THE TRIAL UNIT

E.1 MECHANISM CONTROLS ON CHINESEEEG

Table 10 reruns the shared-head, random-encoder and transplant controls of Table 3 on the second corpus.

unseen readers, ChineseEEG	N=400	vs. adapter	N=1600	vs. adapter
adapter, rank 32	**0.0765**		**0.0984**	
shared head only	0.0504	−0.0261 [-0.0291, -0.0230]	0.0606	−0.0377 [-0.0411, -0.0345]
adapter, random frozen encoder	0.0433	−0.0332 [-0.0360, -0.0304]	0.0497	−0.0487 [-0.0521, -0.0454]
adapter transplanted to other reader	0.0447	−0.0318 [-0.0348, -0.0288]	0.0452	−0.0532 [-0.0570, -0.0495]
uncalibrated	0.0442		0.0442	

Table 10: Three of the four controls of Table 3, rerun on ChineseEEG at the two budgets of Table 5. The shared head is the worst trained arm at both N; the adapter on a random encoder is indistinguishable from zero-shot at $N = 400$ (-0.0009 [-0.0031, +0.0014]) and reaches $1.19\times$ at $N = 1600$, as it reached $1.25\times$ at ZuCo's $N = 400$; the transplanted adapter is indistinguishable from no calibration at both N ($+0.0005$, $+0.0010$, both n.s.). Each fold holds out two readers, so the transplant is the one possible swap.

E.2 Clauses against sentences

The sentence-level cache is derived from the clause-level one by tiling consecutive clauses back into their source sentence, so the two share every sample of EEG and differ only in where trials begin and end. Sentences longer than the 10 s window (36%) are dropped, leaving 18,893 trials, 2,174 distinct sentences, and 1,865 test trials in 126 duration-matched pools – a fifth of the clause cache's statistical power.

fraction of seen-reader margin	ZuCo	ChineseEEG, clauses	ChineseEEG, sentences
seen-reader ceiling	0.1090 (2.62×)	0.1065 (2.56×)	0.1098 (2.63×)
lost by holding out the reader	94%	96%	80%
recovered, adapter N=100	54%	18%	38%
recovered, adapter N=200	68%	31%	44%
recovered, adapter N=400	101%	54%	70%
adapter − full FT, N=400	+0.0187	+0.0178	+0.0042 [-0.0020, +0.0103]
test trials / pools	2,227 / 155	9,454 / 761	1,865 / 126

Table 11: Changing the trial unit moves ChineseEEG about a third of the way to ZuCo. At sentence level the seen-reader ceiling is ZuCo's ($2.63\times$), the adapter at $N = 400$ recovers 70% of the margin rather than 54%, and less of the margin is lost to begin with (80%; 0.0556 [0.0503, 0.0610] against 0.0442 [0.0423, 0.0461]). The remaining gap to ZuCo's 101% is not the segmentation. The sentence-level test set is five times smaller, so its adapter-vs-full-fine-tuning contrast has the same sign but is not significant.

F Electrode count and placement

montage	top-1	× chance	share of 104-channel margin
all 104 live channels	0.1035	2.48	100%
group lasso 32	0.1032	2.48	100%
group lasso 16	0.0977	2.34	91%
random 32, 5 placements	0.0867	2.08	73%
random 16, 5 placements	0.0840	2.01	68%
10-20 32	0.0813	1.95	64%
group lasso 16, posterior only	0.0781	1.87	59%
10-20 16	0.0764	1.83	56%
language-targeted 16	0.0742	1.78	53%
4-channel headband	0.0578	1.39	26%

Table 12: Seen-reader protocol, three training seeds per placement; the random rows pool five electrode draws, 15 runs each. Random beats designed at matched count, by $+0.0075$ [+0.0032, +0.0118] at 16 and $+0.0053$ [+0.0008, +0.0098] at 32; the pooled arm averages more runs than the designed arm, so read those magnitudes as upper bounds and the ordering as the result. The two group-lasso rows and the posterior-only restriction of the same ranking are the experiment that both produces a good montage and disqualifies it.

The placement spread is not training noise. Across the five draws at 16 electrodes the best and worst differ by $+0.0107$ [+0.0050, +0.0165], and two mid-ranked draws differ by $+0.0082$ [+0.0024, +0.0141]; at 32 electrodes the extremes differ by $+0.0129$ [+0.0070, +0.0188]. Each of these exceeds the effect of doubling the electrode count. The group-lasso ranking has a built-in check: the 24 zero-padded columns come out as exactly the bottom 24 of 128 in all three seeds.

An amortised evaluator has to be as good as what it replaces. We reached the group lasso only after a more obvious approach failed, and the failure generalises. Greedy forward selection needs candidates scored without retraining, so we trained encoders that tolerate an arbitrary channel subset – a fresh random subset of 8 to 64 live channels every training batch – and scored candidates by masked forward passes. Selection was decisive and worthless: the chosen 16 beat pooled random by $+0.065$ under the surrogate and by -0.0018 [-0.0062, +0.0025] after retraining. Scoring 2,832 subsets on one validation set and taking the argmax put the chosen 16 *above* all 104 channels under the surrogate, which cannot be a real effect. Worse, the surrogate's ranking is anti-correlated with the truth on the one contrast where ground truth exists: it ranks 10-20 above random, retraining ranks it below. The masking that made masked evaluation in-distribution also crippled the model – best validation top-1 0.13 against 0.20–0.22 for normal training – so it ranked montages while being worse than every montage-specific model it stood in for. The trick that makes amortisation cheap is often what destroys its validity.

G Recovering the channel map

The montage study needs to know which scalp position each cache column corresponds to, and that information is not in the release. ZuCo retains 105 of the 128 HydroCel sensors, the released preprocessing drops the flat reference row, and the survivors close ranks, so column i is the ith surviving electrode rather than E_{i+1}. An earlier version of this study assumed the identity map and spent three of every sixteen slots on the zero-padded columns 104–127; every number in that sweep was wrong, and the error was invisible because a montage with three dead channels still trains and still beats chance.

We recover the map from the 48 left–right homologous sensor pairs published with the corpus. Those pairs constrain the mirror symmetry of the retained set; the remaining degrees of freedom are fixed by requiring the midline electrodes to be the eight retained sensors closest to $x = 0$ and by excluding the nine sensors the corpus documentation designates as EOG. Including the EOG channels as candidate midline sensors is the one mistake worth recording: it places $E17$ on the midline, displaces Oz, and degrades the validation below.

The reconstruction is validated against the data rather than assumed. If the assignment is right, the correlation between two columns should fall off with the geodesic distance between their assigned positions. Spearman ρ between inter-column correlation and inter-electrode distance is -0.832 with the EOG channels excluded, -0.770 with them allowed, and -0.001 ± 0.020 over random permutations of the assignment.

Turning 10-20 names into columns needs a second step, because `mne`'s `standard_1005` template and `GSN-HydroCel-129` are different head geometries and raw nearest-neighbour matching disagrees with the manufacturer's published correspondence on 5 of its 19 entries. A Umeyama similarity calibration between the two coordinate systems does not fix this – HydroCel sensors sit roughly 2 cm apart, which is the scale of the residual misalignment – so we use the published correspondence where it exists and interpolate only the positions it does not cover.